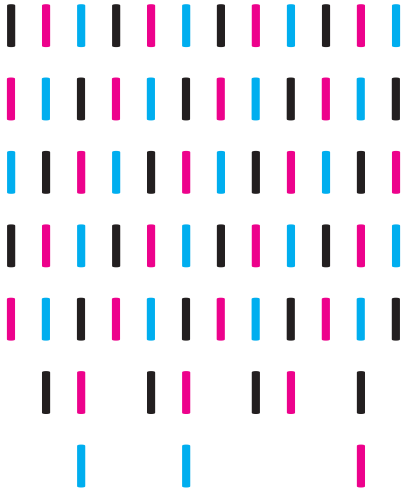

NEPS SURVEY PAPERS

Inga Hahn

NEPS TECHNICAL REPORT FOR SCIENCE: SCALING RESULTS OF STARTING COHORT 1 FOR FIVE-YEAR-OLD CHILDREN

NEPS Survey Paper No. 59
Bamberg, November 2019

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).

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

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

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

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

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

NEPS Technical Report for Science: Scaling Results of Starting Cohort 1 for Five-Year-Old Children

Inga Hahn

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

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

Bibliographic data:

Hahn, I. (2019). *NEPS Technical Report for Scientific Literacy: Scaling Results of Starting Cohort 1 for Five-Year-Old Children* (NEPS Survey Paper No. 59). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP59:1.0

NEPS Technical Report for Scientific Literacy:

Scaling Results of Starting Cohort 1 for Five-Year-Old Children

Abstract

The National Educational Panel Study (NEPS) examines the development of competencies across the life span and develops tests for the assessment of different competence domains. In order to evaluate the quality of these competence tests, various analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedures for the scientific literacy test that was administered to five-year-olds in starting cohort 1. The scientific literacy test contained 20 items with different response formats representing different contexts as well as different areas of knowledge. The test was administered to 2,080 children. Their responses were scaled using a partial credit model. Item fit statistics, differential item functioning (DIF), 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 an acceptable reliability and that all items fitted the model well. Furthermore, test fairness could be confirmed for different subgroups. Only one item showed a substantial gender DIF. The analysis of the test's dimensionality supported a one-dimensional model. A limitation of the test was the lack of very difficult items. Overall, the scientific literacy test showed good psychometric properties that allowed for the estimation of reliable scientific literacy scores. 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.

Key words: scientific literacy, 5-year-olds, differential item functioning, item response theory, scaling, scientific use file

Content

1 Introduction

Within the National Educational Panel Study (NEPS), different competencies are measured coherently across the life span. Tests have been developed for different competence domains. These include, among other things, reading competence, mathematical competence, scientific literacy, information and communication literacy (computer literacy), metacognition, vocabulary, and domain-general cognitive functioning. An overview of the competences measured in the NEPS is given by Weinert et al. (2011) as well as 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 wave 6 to a sample of 5-year-olds in starting cohort 1 (Newborns). First, the main concepts of the scientific literacy test are introduced. Then, the scientific literacy data of starting cohort 1 and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the scientific use file (SUF) is presented.

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

2 Testing Scientific Literacy

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

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

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

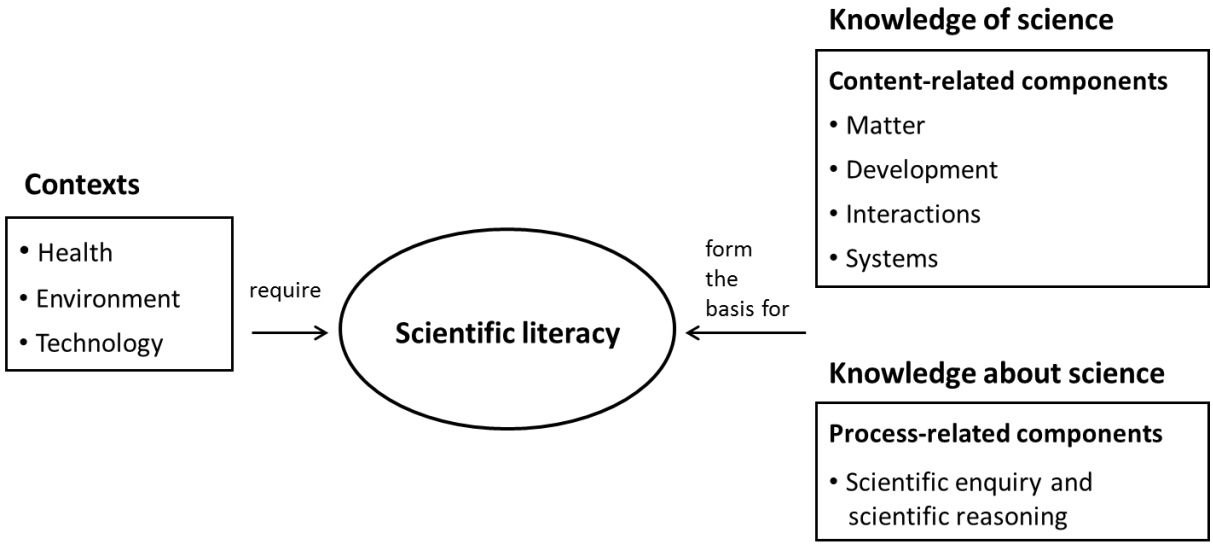

Figure 1. Assessment framework for scientific literacy (Hahn et al, 2013).

In the scientific literacy test for five-year-olds in starting cohort 1 (Newborns) there are two types of response formats. These are simple multiple choice (MC) and complex multiple choice (CMC) in the special form of true false items. In MC items the test taker 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 (true/ false) were presented.

3 Data

3.1 The design of the study

In this study, two domain-specific measures and one stage-specific measure were administered. The domain-specific competence tests measured receptive German vocabulary and scientific literacy. The stage-specific measure targeted the ability of delayed gratification as part of the executive control of the children. The tests were administered without any rotation design. First, the children took the German vocabulary test followed by the scientific literacy test and the delayed gratification test. There was no multi-matrix design regarding the order of the items within the scientific literacy test. All participants received the test items in the same order. The scientific literacy test was conducted as an individual tablet-based test. The test consisted of 20 items which were administered in a testing time of 20 minutes. The allocation of the 20 items to the content areas (KOS and KAS) is depicted in Table 1. Table 2 shows how the items cover the different contexts of the science framework (Hahn et al., 2013) whereas Table 3 gives an overview of the response formats.

Table 1

Classification of Items into Knowledge Domains

Knowledge domains	**Number of Items**
Knowledge of Science (KOS)	13
Knowledge about Science (KAS)	7
Total number of items	20

Table 2

Number of Items by Different Contexts

Context	**Number of Items**
Health	1
Environment	10
Technology	9
Total number of items	20

Table 3

Number of Items by Response Format

Response format	**Number of Items**
Simple Multiple-Choice	17
Complex Multiple-Choice (true / false items)	3
Total number of items	20

3.2 Sample

The science test was administered to 2,080 children. However, 20 children had less than three valid responses on the test. Because no reliable ability scores can be estimated based on such few valid responses, these cases were excluded from further analyses (c.f. Pohl & Carstensen, 2012). Therefore, the analyses presented in this paper are based on a sample of

2,060 individuals (49.8 % girls). A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

4 Analyses

4.1 Missing responses

There are different kinds of missing response. These are a) invalid responses, b) omitted items, c) items that test takers did not reach, d) items that have not been administered, and finally, e) multiple kinds of missing responses within CMC items that are not determined. In this study, all subjects received the same set of items so there are no missing responses due to items not being administered.

Invalid responses occurred, for example, when less than four answers were given in a CMC item (which consisted of four subtasks). Omitted items occurred when test takers skipped some items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response given were coded as not-reached. As CMC items are aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses may be found in these items. A CMC item was coded as missing if at least one subtask contained a missing response. When one subtask contained a missing response, the CMC item was coded as missing. When just one kind of missing response occurred, the item was coded according to the corresponding missing response. When the subtasks contained different kinds of missing responses, the item was labeled as a not-determinable missing response.

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

4.2 Scaling model

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

CMC items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly responded subtasks within that item. If at least one of the subtasks contained a missing response, the whole CMC item was scored as missing. Categories of polytomous variables with less than $N = 200$ responses were collapsed 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. For all three 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 scale

The scientific literacy test for five-year-olds was specifically constructed to be implemented in the NEPS. In order to ensure appropriate psychometric properties, the quality of the test was evaluated in several pretests and analyses.

Before aggregating the subtasks of CMC items to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the MC items in a Rasch model (Rasch, 1960). 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 option and three distractors (i.e., incorrect response options). The quality of the distractors within MC items was examined using the point-biserial correlation between selecting an incorrect response option and the total correct score. Negative correlations indicate good distractors, whereas correlations between .00 and .05 are considered acceptable and correlations above .05 are viewed as problematic distractors (Pohl & Carstensen, 2012).

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

The scientific literacy test should measure the same construct for all children. If any items favored certain subgroups (e.g., 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 socio-economic status), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Differential item functioning (DIF) analyses were estimated using a multigroup IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties between the

subgroups that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as noteworthy of further investigation, differences between 0.4 and 0.6 as considerable but not severe, and differences smaller than 0.4 as negligible DIF. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

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

The science test was constructed to measure a unidimensional scientific literacy score (Hahn et al., 2013). The assumption of unidimensionality was, nevertheless, tested by specifying a two-dimensional model with process-related items representing one dimension and content-related items representing the other dimension. The correlation between the subdimensions as well as differences in model fit between the unidimensional model and the two dimensional model were used to evaluate the unidimensionality of the scale.

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

4.4 Software

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

5 Results

All but one of the 20 items (including all subtasks for the polytomous items) were included in the following analyses. Item scn66000_c was excluded from the analyses due to an insufficient discrimination. For item scn6130s_c only the first and the last subitem were aggregated (s. section 5.4.1 for further information) and were used in this way for estimating the person abilities.

5.1 Descriptive statistics of the responses

In order to a) get a first rough descriptive measure of the item difficulties and b) check for possible estimation problems, before performing IRT analyses we evaluated the relative frequency of the responses given. The percentage of persons correctly responding to an item (relative to all valid responses) ranged from 36.6 % to 87.9 % for the MC items. For the CMC items, the percentage of persons who correctly answered all subtasks varied between 10.2 % and 40.2 %.

5.2 Missing responses

5.2.1 Missing responses per person

Figure 2 shows the number of invalid responses per person. Overall, the number of invalid responses is very small. Over 96 % of the children did not have any invalid response at all. Less than 4.0 % had one or more invalid responses.

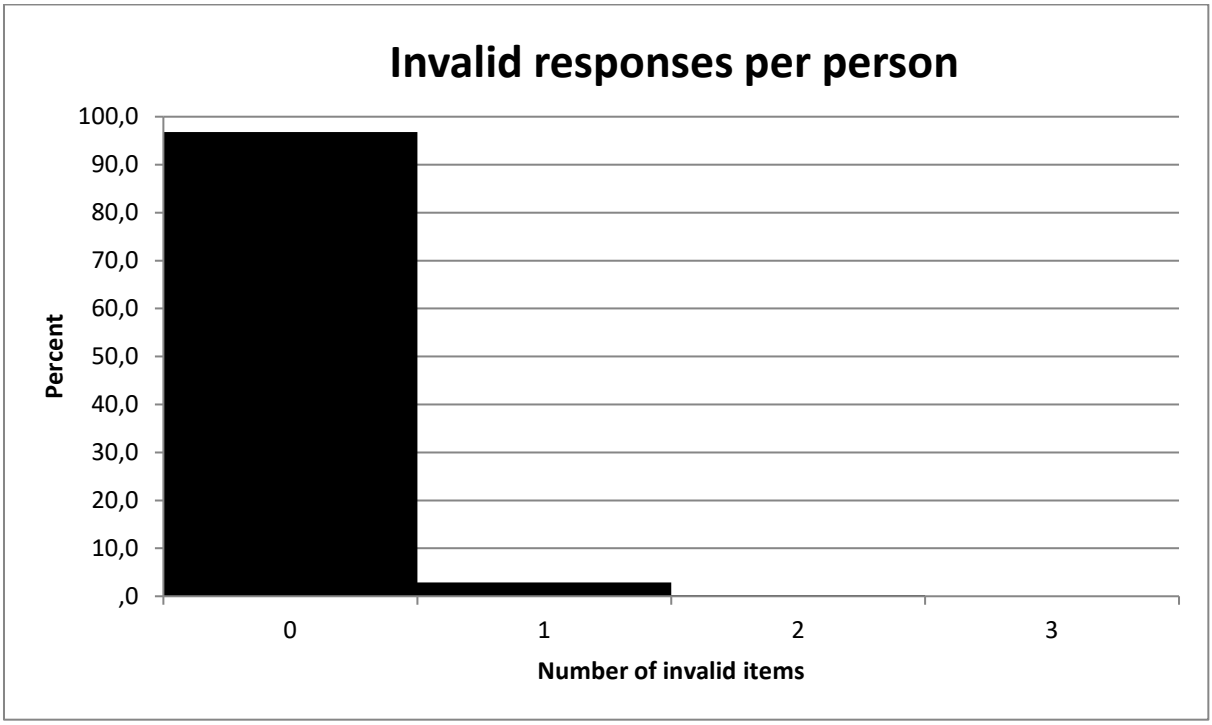

Figure 2. Number of invalid responses

Missing responses may also occur when respondents omit items. As illustrated in Figure 3 most of the respondents (89.5 %) did not skip any item, and less than 1.0 % omitted more than three items.

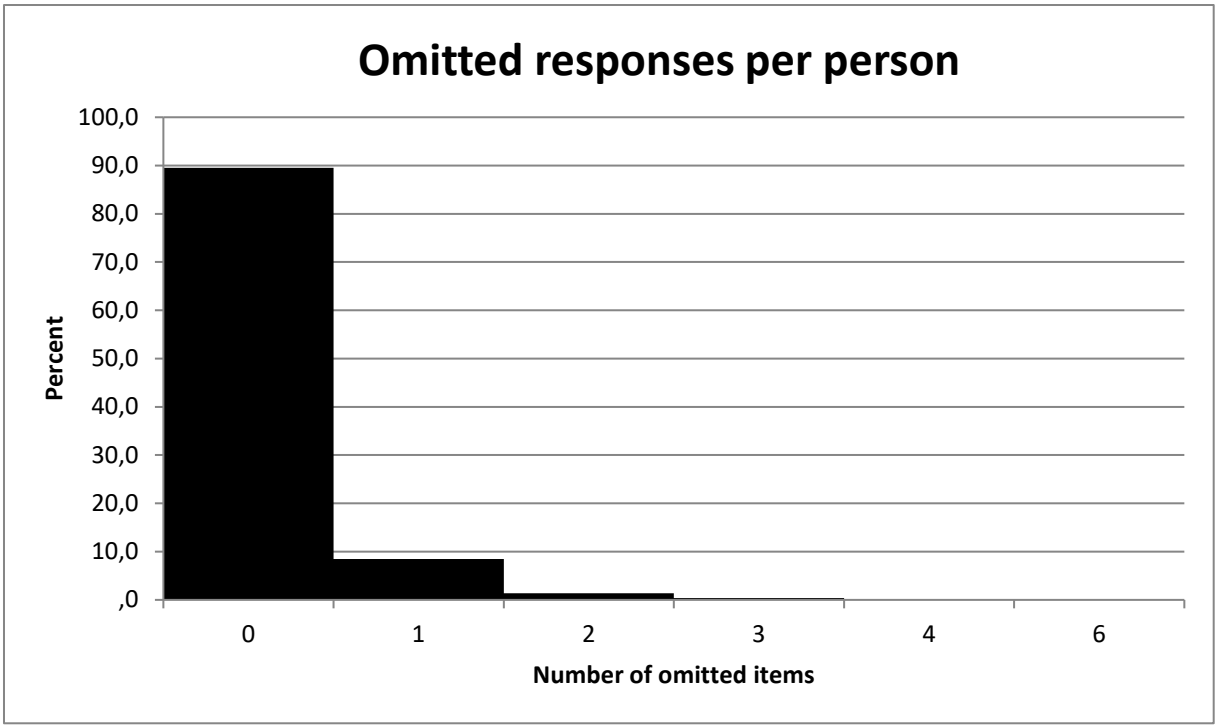

Figure 3. Number of omitted items

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 very low. Most children reached the end of the test (over 99.1 %) and only a very small proportion (0.5 %) did not manage to finish at least two thirds of the test (Figure 4).

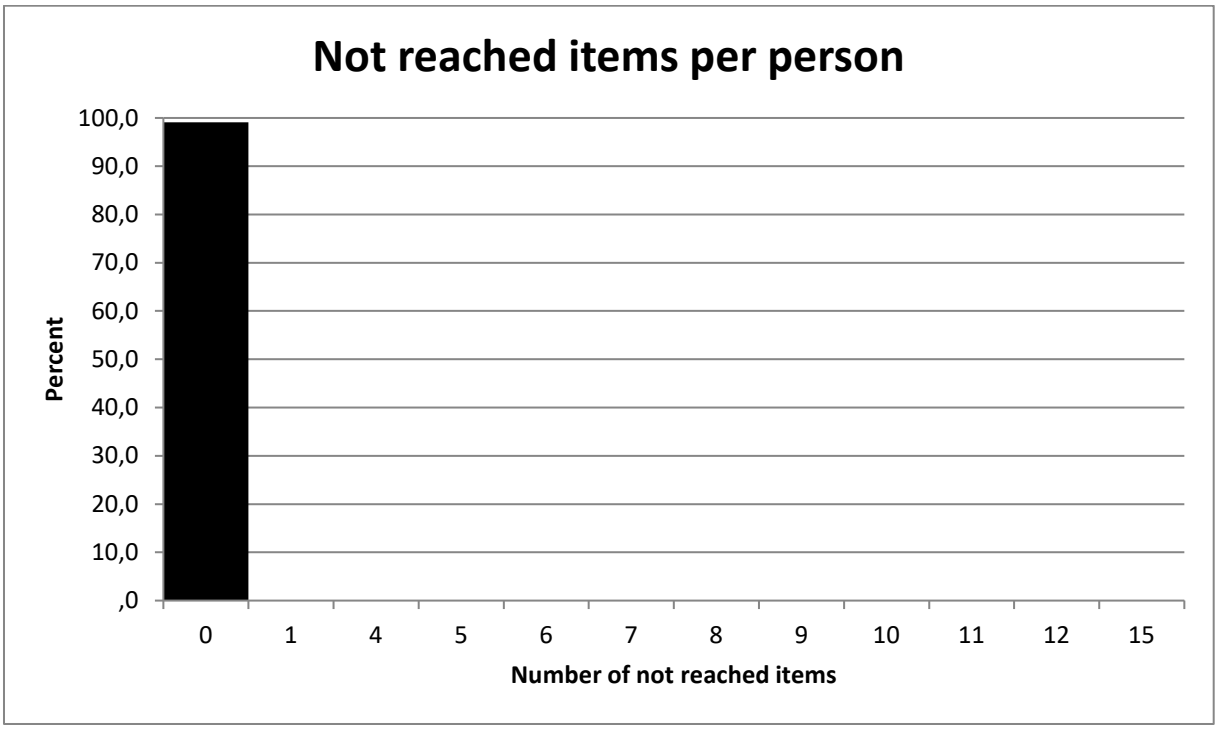

Figure 4. Number of not reached items

The total number of missing responses, aggregated over invalid, omitted and not reached items per person, is illustrated in Figure 5. Overall, the total number of missing responses was very small. 87.0 % of the children answered all questions and consequently had no missing responses. Only 0.3 % of the children show missing responses on more than half of the items.

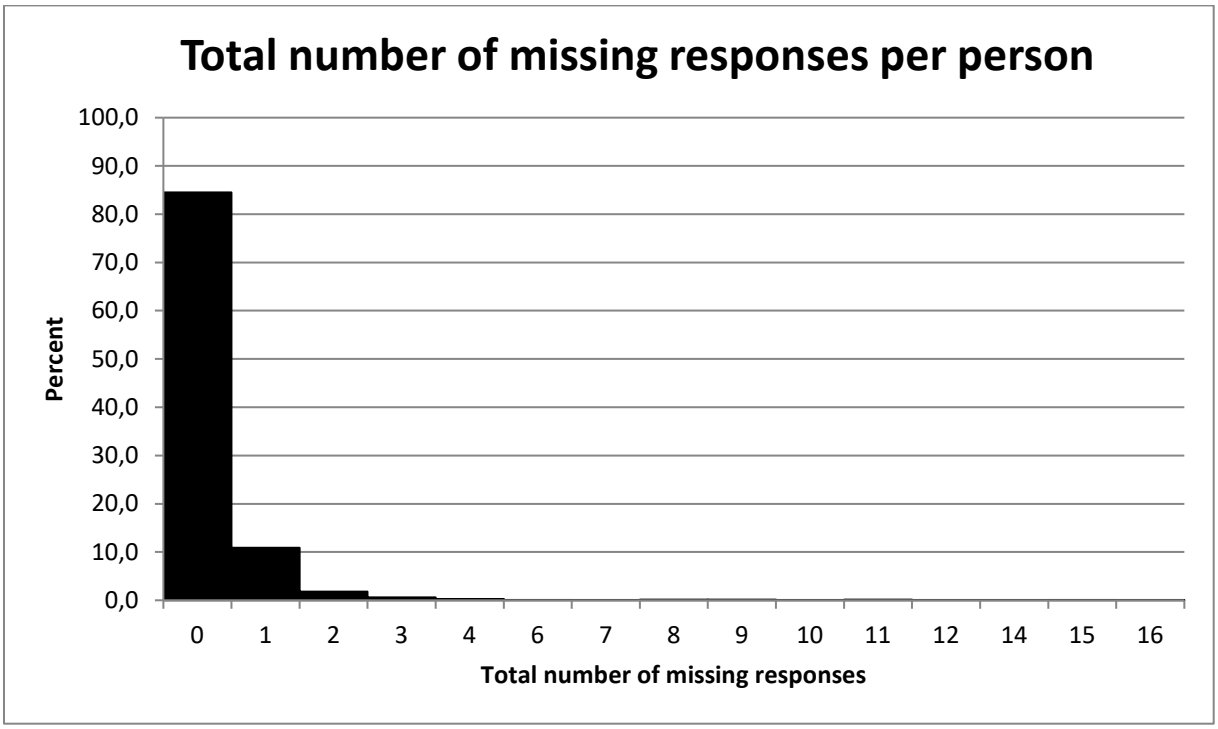

Figure 5. Total number of missing responses

5.2.2 Missing responses per item

Table 4 shows the number of valid responses for each item as well as the percentage of missing responses. Overall, omission rates were rather low, varying across items between 0.0 % and 1.7 %. The omission rate was correlated at $r = -.03$ ($p = .89$) with the difficulty of the item. This result indicates that the test takers did not omit items that are more difficult. Generally, the percentage of invalid responses per item was rather low with the maximum rate being 1.4 % (item scn6130s_c). The relative frequency of not reached items slightly increased towards the end of the test. Eventually only 0.9 % of the children did not reach the last item and thus did not complete the test. The total percentage of missing responses per item varied between 0.2 % and 2.5 %.

Table 4
Valid Responses and Missing Values

Item	Position in the test	Number of valid responses	Not reached items (%)	Omitted items (%)	Invalid responses (%)
sck10420_sc1n6_c	1	2035	0.0	1.2	0.0
scn6130s_c	2	2032	0.0	0.0	1.4
sck16120_sc1n6_c	4	2056	0.0	0.2	0.0
sck1102s_sc1n6_c	5	2041	0.0	0.0	0.9
sck11030_sc1n6_c	6	2033	0.1	1.2	0.0

sck11110_sc1n6_c	7	2041	0.1	0.8	0.0
sck11120_sc1n6_c	8	2048	0.2	0.4	0.0
sck16010_sc1n6_c	9	2045	0.2	0.5	0.0
sck16020_sc1n6_c	10	2030	0.3	1.2	0.0
sck10510_sc1n6_c	11	2031	0.3	1.1	0.0
sck10530_sc1n6_c	12	2035	0.4	0.8	0.0
sck1162s_sc1n6_c	13	2023	0.5	0.0	1.3
sck10710_sc1n6_c	14	2025	0.7	1.0	0.0
sck10720_sc1n6_c	15	2024	0.7	1.0	0.0
scn60100_c	16	2034	0.8	0.5	0.0
sck11330_sc1n6_c	17	2008	0.8	1.7	0.0
sck10910_sc1n6_c	18	2028	0.9	0.8	0.0
scn61800_c	19	2033	0.8	0.5	0.0
sck16210_sc1n6_c	20	2028	0.9	0.7	0.0

Note. The item scn66000_c on position 3 was excluded from the analyses due to insufficient item quality (see section 5.3.1).

5.3 Parameter estimates

5.3.1 Item parameters

Column 2 in table 5 shows the percentage of correct responses in relation to all valid responses for each item. The percentage of correct responses within items varied between 10.2 % (for CMC items) and 88.0 % (for MC items) with an average of 59.9 % correct responses.

The estimated item difficulties for dichotomous variables (MC items) and location parameters for polytomous variables (CMC items) are listed in Table 5. The step parameters (for polytomous variables) are depicted in Table 6. For all CMC items categories were collapsed. As these items were CMC items with a maximum score of 2, they were scaled using the following intervals: 0, 0.5, 1.0 (for item scn6130s_c) and 0, 0.5, 1.0 and 1.5 (for items sck1102s_sc1n6_c and sck1162s_sc1n6_c). 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) varied between −2.219 (sck16120_sc1n6_c) and 0.605 (scn60100_c) with a mean difficulty of −0.753 (SD = 0.06). From a descriptive point of view, the items covered a rather wide range of difficulties. However, there were no very difficult items as the majority of items showed low or medium difficulties.

Table 5

Item Parameters

Item	Percentage correct	Difficulty/location parameter	*SE* (difficulty/ location parameter)	Weighted MNSQ	Weighted t-value	Pt.-bis. Corr. of correct response	Discrimination (2PL)	Yens Q_3
sck10420_sc1n6_c	84.0	−1.852	0.066	1.04	1.0	.27	0.55	.05
scn6130s_c	n.a.	−0.695	0.064	0.97	−1.5	.41	1.14	.10
sck16120_sc1n6_c	88.0	−2.219	0.073	0.99	−0.1	.31	0.85	.09
sck1102s_sc1n6_c	n.a.	0.215	0.050	0.98	−0.7	.47	1.48	.10
sck11030_sc1n6_c	59.4	−0.433	0.051	1.04	2.5	.36	0.58	.08
sck11110_sc1n6_c	80.0	−1.558	0.061	1.04	1.2	.31	0.54	.28
sck11120_sc1n6_c	65.6	−0.733	0.053	1.04	2.2	.34	0.50	.28
sck16010_sc1n6_c	86.0	−2.022	0.069	0.97	−0.7	.37	1.09	.06
sck16020_sc1n6_c	37.2	0.593	0.052	0.99	−0.8	.41	0.85	.06
sck10510_sc1n6_c	79.7	−1.535	0.061	1.00	−0.0	.38	0.87	.20
sck10530_sc1n6_c	58.0	−0.367	0.051	0.98	−1.5	.46	0.96	.20
sck1162s_sc1n6_c	n.a.	0.163	0.067	0.98	−0.7	.40	1.45	.08
sck10710_sc1n6_c	76.0	−1.300	0.058	0.98	−0.8	.41	0.93	.06
sck10720_sc1n6_c	58.5	−0.387	0.051	0.99	−0.8	.44	0.86	.08
scn60100_c	37.1	0.605	0.052	1.11	5.9	.24	0.23	.10
sck11330_sc1n6_c	54.7	−0.212	0.051	0.96	−2.7	.48	1.01	.08
sck10910_sc1n6_c	65.3	−0.714	0.053	1.00	−0.2	.42	0.87	.07
scn61800_c	65.4	−0.720	0.053	1.04	2.0	.35	0.58	.07
sck16210_sc1n6_c	73.2	−1.130	0.056	0.94	−2.5	.48	1.30	.09

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

Due to the large sample size, the standard error of the estimated item difficulties were rather small (SE(ß) ≤ 0.073). Overall, the item difficulties were rather low.

Table 6

Step Parameters for the CMC Items

Item	Step 1 (SE)	Step 2 (SE)	Step 3 (SE)
scn6130s_c	−0.351 (0.047)	0.351	-
sck1102s_sc1n6_c	−0.484 (0.051)	0.371 (0.065)	0.112
sck1162s_sc1n6_c	−1.460 (0.059)	0.197 (0.060)	1.264

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

5.3.2 Person parameters

Person parameters are estimated as WLEs (Pohl & Carstensen, 2012a). A description of the data in the SUF can be found in section 7. An overview of how to work with competence data is given in Pohl and Carstensen (2012a).

5.3.3 Test Targeting and Reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. In Figure 6, 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 0.627, indicating somewhat limited variability between subjects. The reliability of the test (EAP/PV reliability = .673; WLE reliability = .639) was sufficient. The figure shows that the items cover a limited range of the ability distribution of the persons. There is a lack of items covering persons with high science ability. Instead, the majority of items are easy or of medium difficulty. As a consequence, persons with a medium and low ability will be measured relatively precisely with a low standard error while ability estimates for children with high science ability will have a larger standard error.

Scale in logits	Person ability	Item difficulty
3		Each ‚X' represents 39.7 cases
	X	
2		
	X	
	X	
	X	
	XX	
	XX	
	XXX	
	XX	
1	XXX	
	XXX	
	XXXXX	
	XXXX	
	XXXXXX	
	XXXXXXXX	9 15
	XXXXXXXXX	
	XXXXXXXX	
	XXXXXXXXX	4
0	XXXXXXXXXX	12
	XXXXXXXXX	
	XXXXXXXXXX	
	XXXXXXXXX	
	XXXXXXXX	16
	XXXXXXXX	11 14
	XXXXXXXX	5
	XXXXXXX	
−1	XXXXXXX	2 7 17 18
	XXXXXX	
	XXXXXX	
	XXXX	
	XXX	19
	XXX	
	XXX	13
	XX	
−2	X	6 10
	X	
	X	
		1
		8
		3
−3		

Figure 6. Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 11.8 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see table 4).

5.4 Quality of the test

5.4.1 Fit of the subtasks of complex multiple-choice items

Before the responses on the subtasks of CMC items are aggregated and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the single subtasks together with the simple MC items in a Rasch model. Counting the subtasks of the CMC items separately, there were 29 items. The probability of a correct response ranged from 25.0 % to 87.0 % (*Mdn* = 59 %). All except two subtasks showed a satisfactory item fit. WMNSQ ranged from 0.92 to 1.17, the respective *t*-value between −5.0 and 9.7. Two of the four subtasks of item scn6130s_c showed noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. We dealt with this problem by aggregating only the first and the last subitem which exhibited good quality. This aggregated item snc6130s_c showed a good item fit and was used in this way for estimating the person abilities. The remaining subtasks showed a good model fit, so their aggregation to a polytomous variable seemed to be 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 childrens' total score. Most distractors had a point biserial correlation with the total score below zero with the exception of five items with a point-biserial-correlation between .00 and .15. The results indicate that the majority of the distractors worked well.

5.4.3 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the partial credit model, using the MC items and the CMC items. Altogether, the item fit can be considered to be very good (see Table 5). Values of the WMNSQs ranged from 0.94 (item sck16210_sc1n6_c) to 1.11 (item scn60100_c). There were no items with a t-value above 6. Point-biserial correlations between the item scores and the total scores ranged from .24 to .48 and had a mean of .38. Hence no indications for a pronounced misfit of these items could be detected and therefore they were kept in the analysis for estimating the scientific literacy scores. All item characteristic curves showed a good fit of the items to the PCM.

5.4.4 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables gender, the number of books at home (as a proxy for socioeconomic status) and migration background (see Pohl & Carstensen, 2012, for a description of these variables).

The differences between the estimated item difficulties in the various groups are summarized in Table 7. For example, the column "Male vs. female" reports the differences in item difficulties between boys and girls; a positive value would indicate the item was more difficult for males, whereas a negative value would highlight a lower difficulty for boys as opposed to girls.

Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF to those that only estimate main effects (see Table 8). Akaike's (1974) information criterion (AIC) and the Bayesian information criterion (BIC; Schwarz, 1978) were used for comparing the models.

Gender

The sample included 1,026 (49.8 %) girls and 1,034 (50.2 %) boys. On average, boys showed slightly lower scientific literacy scores than girls (main effect = −0.024 logits, Cohen's *d* = −0.030). There was one item with a large gender DIF (scn61800_c) of −1.282. We dealt with this item by splitting it into two unique items, one including the responses for girls (and missing values for boys) and another one including the responses for boys (and missing values for girls), and estimating the person abilities for the scientific use file using the new items. For the remaining items the largest difference in difficulties between the two groups was 0.558 logits which is still acceptable. Model comparisons using the AIC and the BIC both favoured the model estimating DIF which might be due to the item scn61800. None of the other items showed substantial DIF.

Books

The number of books at home was used as a proxy for socio-economic status. There were 575 (27.9 %) test takers with 0 to 100 books at home, 1,367 (66.4 %) test takers with more than 100 books at home, and 118 (5.7 %) test takers did not give a valid response. DIF was investigated using these three groups. There were considerable average differences between the three groups. Participants with 100 or less books at home on average showed lower scientific literacy scores than participants with more than 100 books (main effect = −0.532 logits, Cohen's *d* = −0.705). Participants without a valid response on the variable 'books at home' performed better than participants with up to 100 (main effect = 0.298 logits, Cohen's *d* = 0.409) and lower than participants with more than 100 books at home, respectively (main effect = −0.230 logits, Cohen's *d* = −0,297). There was only one item with a medium DIF comparing participants with many and fewer books (item scn60100_c, DIF = −0.690). Comparing the group without valid responses to the two groups with valid responses, DIF occurred up to 0.672 logits which was deemed acceptable. The AIC favoured the model allowing for DIF while the BIC favoured the model estimating the main effect.

Migration Background

There were 1,467 (71.2 %) participants without a migration background, 495 (24.0 %) participants with a migration background and 98 (4.8 %) children without respective information. Children without a migration background on average showed a higher scientific literacy than children with a migration background (main effect = 0.504 logits, Cohen's *d* = 0.673). Children without a migration background also showed a higher scientific literacy than children whose background was not indicated (main effect = 0.478 logits, Cohen's *d* = 0.617). The difference between children with a migration background and those with an unknown background was small (main effect = −0.026 logits, Cohen's *d* = −0.034). There was no considerable DIF on the item level comparing the groups with and without a migration background. Comparing the group without valid responses to the two groups with valid responses, there were only two items with a medium DIF larger than 0.6 logits. Item sck16010_sc1n6_c showed a DIF of 0.732 logits and the DIF of item scn60100_c was 0.668

logits. The AIC favoured the model allowing for DIF while the BIC favoured the model estimating the main effect.

Table 7

Differential item functioning (absolute differences between difficulties)

Item	Gender	Books			Migration status		
	Male vs. female	<100 vs. >100	<100 vs. missing	>100 vs. missing	Without vs. With	Without vs. Missing	With vs. Missing
sck10420_sc1n6_c	0.032	−0.056	0.672	0.610	0.168	−0.160	−0.334
scn6130s_c	−0.062	−0.014	−0.254	−0.266	0.090	−0.304	−0.402
sck16120_sc1n6_c	0.384	0.220	0.058	0.276	−0.078	−0.156	−0.084
sck1102s_sc1n6_c	0.094	0.100	0.392	0.512	−0.034	−0.038	−0.008
sck11030_sc1n6_c	−0.150	−0.230	−0.034	−0.262	0.190	0.234	0.040
sck11110_sc1n6_c	0.418	−0.414	0.024	−0.390	0.336	0.024	−0.318
sck11120_sc1n6_c	0.558	−0.030	0.128	0.096	0.052	−0.096	−0.152
sck16010_sc1n6_c	0.164	0.104	−0.214	−0.110	−0.454	0.284	0.732
sck16020_sc1n6_c	−0.170	−0.038	−0.316	−0.350	−0.158	0.168	0.320
sck10510_sc1n6_c	0.430	−0.006	−0.144	−0.150	−0.032	0.424	0.450
sck10530_sc1n6_c	0.256	0.262	−0.182	0.080	−0.166	−0.312	−0.150
sck1162s_sc1n6_c	−0.316	−0.022	0.190	0.198	0.150	0.276	0.120
sck10710_sc1n6_c	−0.080	0.272	−0.388	−0.114	−0.150	0.148	0.294
sck10720_sc1n6_c	0.186	0.082	−0.012	0.070	−0.194	−0.490	−0.300
scn60100_c	−0.324	−0.690	0.226	−0.464	0.542	0.668	0.118
sck11330_sc1n6_c	−0.262	0.184	−0.028	0.154	−0.212	−0.030	0.176
sck10910_sc1n6_c	0.144	0.114	−0.042	0.070	−0.204	−0.186	0.012
scn61800_c	−1.282	−0.102	0.244	0.140	0.268	0.078	−0.194
sck16210_sc1n6_c	0.414	0.376	−0.494	−0.116	−0.210	−0.518	−0.312

Table 8

Comparison of models with and without DIF

DIF variable	Model	Deviance	N	Number of parameters	AIC	BIC
Gender	main effect	50,982.25	2060	26	51,034.25	51180.64
	DIF	50,694.81	2060	45	50,784.81	51038.18
Books	main effect	47,944.55	1942	26	47,996.55	48141.41
	DIF	47,858.96	1942	45	47,948.96	48199.68
Migration	main effect	48,385.68	1962	26	48,437.68	48582.80
	DIF	48,317.05	1962	45	48,407.05	48658.22

5.4.5 Rasch-homogeneity

An essential assumption of the Rasch model (1960) is that all item-discrimination parameters are equal. In order 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 between 0.23 and 1.48 (see Table 5). The average discrimination parameter fell at 0.88. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 50,839.22, BIC = 50,895.72) as compared to the PCM model (AIC = 51,032.56, BIC = 51,065.41). 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 unidimensionality of the test was investigated by specifying a one- and a two-dimensional model. For these analyses we only used the 19 items that met the selection criteria for estimating the person parameters. 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 two subscales: the process related components (KAS) and the content related components (KOS; for more details see Hahn et al.. 2013). For estimating the two-dimensional model Gauss' estimation in ConQuest was used (nodes were chosen in such a way that stable parameter estimation was obtained). The unidimensional model (BIC= 50,900.51, number of parameters = 25) fitted the data better than the two-dimensional model (BIC= 51,056.26, number of parameters = 27). The correlation of the two dimensions was 0.925 which supports the decision to use the one-dimensional construct and to consequently use a single competence score for scientific literacy.

Moreover, an examination of the residual correlations for the one-dimensional model using the corrected Q_3 statistic (Yen, 1984) indicated a largely unidimensional scale. Only two items showed Q3-values above 0.20 which – in this case – is quite understandable because these items belong to a unit and thus share the same context and item stem. However, the average absolute residual correlation was M = .00 (SD = .04). This indicates that the test was essentially unidimensional.

6 Discussion

The analyses in the previous sections aimed at providing information on the quality of the science test for five-year-old children and at describing how the scientific literacy score is estimated.

We investigated different kinds of missing responses finding that the amount of invalid responses, not-reached items and omitted responses is very low. The relative frequency of missing values was *not* correlated significantly with the difficulty of the items.

Apart from that we examined the item and test parameters and thoroughly checked item fit statistics for simple MC items, subtasks of CMC items, as well as the aggregated 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 and investigating the tests' dimensionality as well as local item dependence.

The various criteria (WMNSQ, t-value of the WMNSQ, ICC) indicated a good fit of the items. Also, discrimination values of the items (either estimated in a GPCM or as a correlation of the item score with total score) are acceptable. Different variables were used for testing measurement invariance across various subgroups. Except for one item in the gender category no considerable DIF became evident indicating that the test is fair to the considered subgroups. We explained how we dealt with this item when estimating the person abilities (see section 5.3.4).

The test has a sufficient reliability, but had a somewhat limited variance. As a consequence, children with a medium and low scientific literacy will be measured relatively precisely with a low standard error while ability estimates for children with high science ability will have a larger standard error.

Fitting a unidimensional partial credit model (the dimensions being the "content related components" and the "process related components") yielded a better model fit than the two-dimensional partial credit model. This result and the high correlation of 0.925 between the two dimensions indicate that a unidimensional model describes the data reasonably well.

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

7 Data in the Scientific Use file

7.1 Naming conventions and scientific literacy scores

There are 20 items in the data set that are either scored as dichotomous variables (MC or SCR items) with 0 indicating an incorrect response and 1 indicating a correct response, or scored as a polytomous variable (CMC items) indicating the (partial) credit. The dichotomous variables are marked with a '_c' at the end of the variable name, the CMC items are marked with a 's_c' at the end of the variable name. Note that the value of the polytomous variable does not necessarily indicate the number of correctly responded subtasks (see section 4.2

aggregation of CMC items). When categories of the polytomous variables had less than N=200, the categories were collapsed. In the scaling model, the collapsed polytomous item is scored in steps of 0, 0.5, 1.0 and 1.5 (denoting the highest).

Manifest scale scores are provided in form of WLE estimates (scn6_sc1) including the respective standard error (scn6_sc2). The ConQuest Syntax for estimating the WLE scores from the items is provided in Appendix A. For persons who either did not take part in the scientific literacy test or who did not give enough valid responses, no WLE is estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values. Alternatively, users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values. A description of these approaches can be found in Pohl and Carstensen (2012).

References

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

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

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

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

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

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

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

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

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

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *Technical Report of Reading– Scaling Results of Starting Cohort 3 in Fifth Grade* (NEPS Working Paper No. 15). Bamberg, Germany: Otto-Friedrich-Universität. Nationales Bildungspanel.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C., & Hamann, M. (2007). Naturwissenschaftliche Kompetenz im internationalen Vergleich. In PISA-Konsortium Deutschland (Hrsg.), *PISA 2006 – Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 63–105). Münster, Germany: Waxmann.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Kopenhagen, Denmark: Paedagogiske Institut.

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

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. In H. P. Blossfeld, H. G. Roßbach, & J. v. Maurice & (Eds.), *Education as a Lifelong Process: The German National Educational Panel Study (NEPS). (Zeitschrift für Erziehungswissenschaft. Sonderheft 14 .* Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

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

Appendix

Appendix A: ConQuest-Syntax for estimating WLE in starting cohort I (five-year-old children)

```
Title SC1 five-year-olds Scientific Literacy partial credit model;

/*load data*/
datafile filename.dat;
format pid 1-7 responses 8-26;
labels <<filename.txt;
model item;
set warnings=yes;

/*scoring*/
codes 0,1,2,3;
score (0,1)          (0,1)          !item (1,3,5-11,13-19);
score (0,1,2)        (0,0.5,1)      !item (2);
score (0,1,2,3)      (0,0.5,1,1.5) !item (4,12);

/*model specification*/
set constraint=cases;
model item + item*step;

/*estimate model*/
estimate ! method=gauss, nodes=45;

/*save results to file*/
show cases !estimates=wle >> filename.wle;
show !  estimates=latent >> filename.shw;
itanal!  estimates=latent >> filename.ita;

quit;
```

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

Variable name	Position in the test	Component	Context
sck10420_sc1n6_c	1	KOS	Health
scn6130s_c	2	KOS	Technology
scn66000_c	3	KAS	Technology
sck16120_sc1n6_c	4	KAS	Technology
sck1102s_sc1n6_c	5	KAS	Environment
sck11030_sc1n6_c	6	KAS	Environment
sck11110_sc1n6_c	7	KOS	Environment
sck11120_sc1n6_c	8	KOS	Environment
sck16010_sc1n6_c	9	KAS	Technology
sck16020_sc1n6_c	10	KAS	Technology
sck10510_sc1n6_c	11	KOS	Environment
sck10530_sc1n6_c	12	KOS	Environment
sck1162s_sc1n6_c	13	KOS	Environment
sck10710_sc1n6_c	14	KOS	Environment
sck10720_sc1n6_c	15	KOS	Environment
scn60100_c	16	KOS	Environment
sck11330_sc1n6_c	17	KOS	Technology
sck10910_sc1n6_c	18	KOS	Technology
scn61800_c	19	KOS	Technology
sck16210_sc1n6_c	20	KAS	Technology

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