NEPS *SURVEY PAPERS*

Jana Kähler

# NEPS TECHNICAL REPORT FOR SCIENCE: SCALING RESULTS OF STARTING COHORT 3 FOR GRADE 9

NEPS *Survey Paper* No. 79
Bamberg, November 2020

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

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

# NEPS Technical Report for Science: Scaling Results of Starting Cohort 3 for Grade 9

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

**Email address of the lead author:**
jkaehler@leibniz-ipn.de

**Bibliographic data:**

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

# NEPS Technical Report for Science: Scaling Results of Starting Cohort 3 for Grade 9

## Abstract

The National Educational Panel Study (NEPS) examines the development of competencies across the life span and develops tests for the assessment of different competence domains. To evaluate the quality of these competence tests various analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedures for the scientific literacy test that was administered in Grade 9 of Starting Cohort 3. The scientific literacy test contained 39 items (distributed among three booklets with a low, medium, or high level of difficulty) with different response formats representing different contexts as well as different areas of knowledge. The test was administered to 4,882 students. Their responses were scaled using a partial credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, the test's dimensionality, and local item independence were evaluated to ensure the quality of the test. These analyses showed that the test exhibited a good reliability and that all items but one satisfactorily fitted the model. Furthermore, test fairness could be confirmed for different subgroups. As the correlations between the two knowledge domains were very high, the assumption of unidimensionality seems adequate. A limitation of the test was the lack of very difficult items. However, the results revealed good psychometric properties of the scientific literacy test, thus supporting the estimation of a reliable scientific literacy score. Besides the scaling results, this paper also describes the data available in the scientific use file and provides the ConQuest syntax for scaling the data. Additionally, the design and results of the linking study for the competence scores in grades 6 and 9 are presented.

**Keywords:** scientific literacy, 9th grade, linking grade 6 and 9, differential item functioning, item response theory, scaling, scientific use file

## Content

1 Introduction ..... 4

2 Testing Scientific Literacy ..... 4

3 Data ..... 5

3.1 The design of the study ..... 5

3.2 Sample ..... 7

4 Analyses ..... 7

4.1 Missing responses ..... 7

4.2 Scaling model ..... 8

4.3 Checking the quality of the test ..... 8

4.4 Software ..... 9

5 Results ..... 9

5.1 Descriptive statistics of the responses ..... 9

5.2 Missing Responses ..... 10

5.2.1 Missing responses per person ..... 10

5.2.2 Missing responses per item ..... 13

5.3 Parameter estimates ..... 17

5.3.1 Item parameters ..... 17

5.3.2 Person parameters ..... 19

5.3.3 Test targeting and reliability ..... 19

5.4 Quality of the test ..... 21

5.4.1 Fit of the subtasks of complex multiple-choice items ..... 21

5.4.2 Distractor analyses ..... 21

5.4.3 Item fit ..... 21

5.4.4 Differential item functioning ..... 21

5.4.5 Rasch-homogeneity ..... 27

5.4.6 Unidimensionality of the test ..... 27

6 Discussion ..... 28

7 Data in the Scientific Use file ..... 28

7.1 Naming conventions ..... 28

7.2 Linking of competence scores ..... 29

7.2.1 Samples ..... 29

7.2.2 The design of the link study ..... 29

7.2.3 Results ..... 29

7.3 Scientific literacy scores ..... 32

8 References ..... 33

# 1 Introduction

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

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

In this paper, the results of these analyses are presented for a scientific literacy test that was administered in Grade 9 of Starting Cohort 3. First, the main concepts of the scientific literacy test are introduced. Then, the scientific literacy data of Starting Cohort 3 and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the scientific use file is presented.

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

# 2 Testing Scientific Literacy

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

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

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

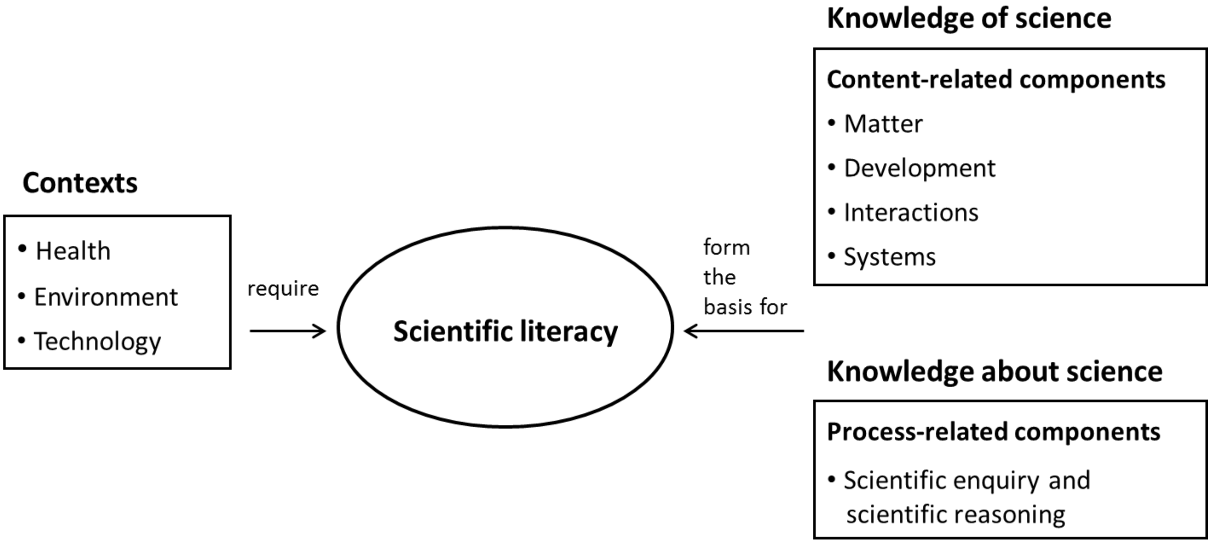

*Figure 1.* Assessment framework for scientific literacy (Hahn et al., 2013).

In the scientific literacy test for Grade 9 of Starting Cohort 3 (fifth grade), there were two types of response formats. These were simple multiple-choice (MC), and complex multiple-choice (CMC) in the special form of true-false items. In MC items the test-taker had to identify the correct answer out of four response options. The three incorrect response options functioned as distractors. In CMC items four subtasks with two response options each (e.g., yes/ no) were presented.

# 3 Data

## 3.1 The design of the study

The study assessed different competence domains including, among others, scientific literacy and computer literacy. The competence tests for these two domains were always presented first within the test battery. To control for test position effects, the tests were administered to participants in a different sequence (see Table 1). For each participant, the scientific literacy test was either administered as the first or the second test (i.e., after the computer literacy test). The test position in grade 9 was identical to the test position in grade 6, thus, all participants received the tests in the same sequence in the two grades. The test time for the scientific literacy test was 28 minutes.

To measure participants' scientific literacy with great accuracy, the difficulty of the administered tests should adequately match the participants' abilities. Therefore, the study adopted the principals of longitudinal multistage testing (Pohl & Carstensen, 2013). Based on preliminary studies three different versions of the scientific literacy test were developed that differed in their average difficulty (i.e., a test with a low level of difficulty, a test with a medium level of difficulty, and a test with a high level of difficulty). Each of the three tests included 28 items that represented the two knowledge domains (see Table 2) and the three contexts of the scientific literacy framework (Hahn et al., 2013; see Table 3).

Seventeen items were identical in all three test versions, twenty-two items were identical in the tests with a low and medium level of difficulty, and twenty-three items were identical in the tests with a medium and high level of difficulty (see Tables 2 and 3).

Six items were unique to the test with a low level of difficulty and five items to the test with a high level of difficulty (see Appendix B for the detailed assignment of the test items to each test version). The different response formats of the items are summarized in Table 4. Participants were assigned to the test version based on their scientific literacy competence in the previous assessment (grade 6).

Table 1: Number of Participants by the (Quasi-)Experimental Conditions for all Test Versions

| **Test position** | **Easy Test Version** | **Medium Test Version** | **Difficult Test Version** | **Total** |
|---|---|---|---|---|
| First Position | 486 | 1182 | 756 | 2424 |
| Second Position | 651 | 1166 | 639 | 2456 |
| Total | 1137 | 2348 | 1395 | 4880 |

Table 2: Number of Items by Knowledge Domains by Test Version

| **Knowledge domains** | **Easy Test Version** | **Medium Test Version** | **Difficult Test Version** | **Total** |
|---|---|---|---|---|
| Knowledge of Science (KOS) | 19 | 19 | 19 | 27 |
| Knowledge about Science (KAS) | 9 | 9 | 9 | 12 |
| Total number of items | 28 | 28 | 28 | 39 |

Table 3: Number of Items by Different Contexts by Test Version

| **Context** | **Easy Test Version** | **Medium Test Version** | **Difficult Test Version** | **Total** |
|---|---|---|---|---|
| Health | 6 | 6 | 6 | 8 |
| Environment | 11 | 11 | 10 | 17 |
| Technology | 11 | 11 | 12 | 14 |
| Total number of items | 28 | 28 | 28 | 39 |

Table 4: Number of Items by Response Formats by Difficulty of the Test

| Response format | Easy Test Version | Medium Test Version | Difficult Test Version | Total |
|---|---|---|---|---|
| Simple Multiple-Choice | 19 | 19 | 23 | 27 |
| Complex Multiple-Choice (True false items) | 9 | 9 | 5 | 12 |
| Total number of items | 28 | 28 | 28 | 39 |

## 3.2 Sample

A total of 4,882 individuals received the scientific literacy test. For two participants less than three valid item responses were available. Because no reliable ability scores can be estimated based on such few valid responses, these cases were excluded from further analyses (Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 4,880 individuals (49.6% girls). A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

# 4 Analyses

A total of 38 items (including all subtasks for the polytomous items) were included in the analyses. One item (scg9611s_c) had to be excluded due to insufficient item quality.

## 4.1 Missing responses

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

Invalid responses occurred, for example, when two response options were selected in simple MC items where only one was required, or when numbers or letters that were not within the range of valid responses were given as a response. Omitted items occurred when test-takers skipped some items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response were coded as not-reached. As CMC items are aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses may be found in these items. A CMC item was coded as missing if at least one subtask contained a missing response.

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

## 4.2 Scaling model

To estimate item and person parameters for scientific literacy, a partial credit model was used (PCM; Masters, 1982) that estimates item difficulties for dichotomous variables and location parameters for polytomous variables. Ability estimates for scientific literacy were estimated as weighted maximum likelihood estimates (WLEs). Item and person parameter estimation in NEPS is described in Pohl and Carstensen (2012), whereas the data available in the SUF are described in Section 7. CMC items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly solved subtasks within that item. If at least one of the subtasks contained a missing response, the whole CMC item was scored as missing. When categories of the polytomous variables had less than $N = 200$, the categories were collapsed to avoid any possible estimation problems. This usually occurred for the lower categories of polytomous items. For seven of the twelve CMC items categories were collapsed (see Appendix A). To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items were scored dichotomously as 0 for an incorrect and as 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats).

## 4.3 Checking the quality of the test

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

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

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

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

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

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

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

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

## 4.4 Software

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

# 5 Results

## 5.1 Descriptive statistics of the responses

To a) get a first rough descriptive measure of the item difficulties and b) check for possible estimation problems, before performing IRT analyses we evaluated the relative frequency of the responses given for all items. The percentage of persons correctly responding to an item (relative to all valid responses) ranged from 17.7% to 85.0% for the MC items. For the CMC items, the percentage of persons who correctly answered all subtasks varied between 6.0% and 49.6%. From a descriptive point of view, the items covered a rather wide range of difficulties.

## 5.2 Missing Responses

### 5.2.1 Missing responses per person

Figure 2 shows the number of invalid responses per person by test version. Overall, there were very few invalid responses. Between 83.4% and 89.5% of the respondents did not have any invalid response at all. For the three test versions, less than 2.2% (low level), less than 0.9% (medium level), and less than 0.4% (high level) had more than three invalid responses. There was a slight difference in the number of invalid responses between the different test versions.

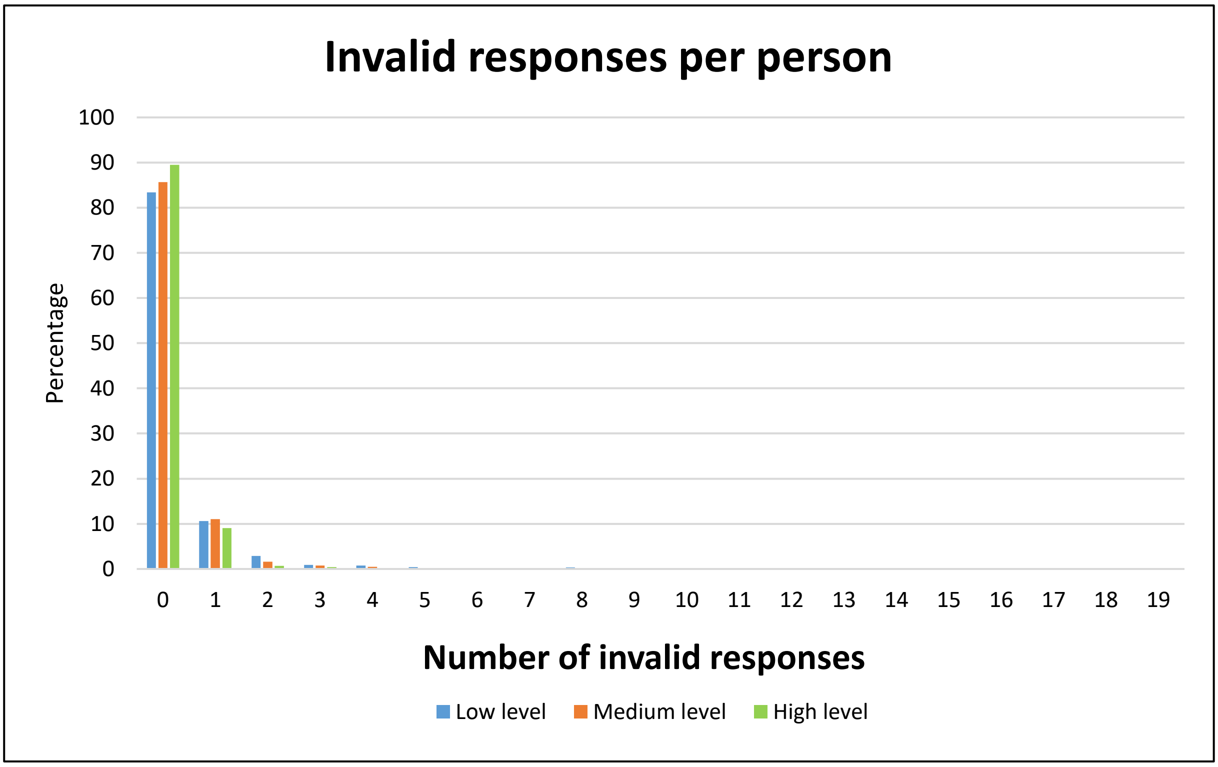

*Figure 2.* Number of invalid responses per person by test difficulty.

Missing responses may also occur when respondents omit items. As illustrated in Figure 3 most respondents, 77.8% to 85.0%, did not skip any item, and less than 2.9% (low level), less than 3.4% (medium level), and less than 2.7% (high level) omitted more than three items. There was only a slight difference in the number of omitted items between the different test versions.

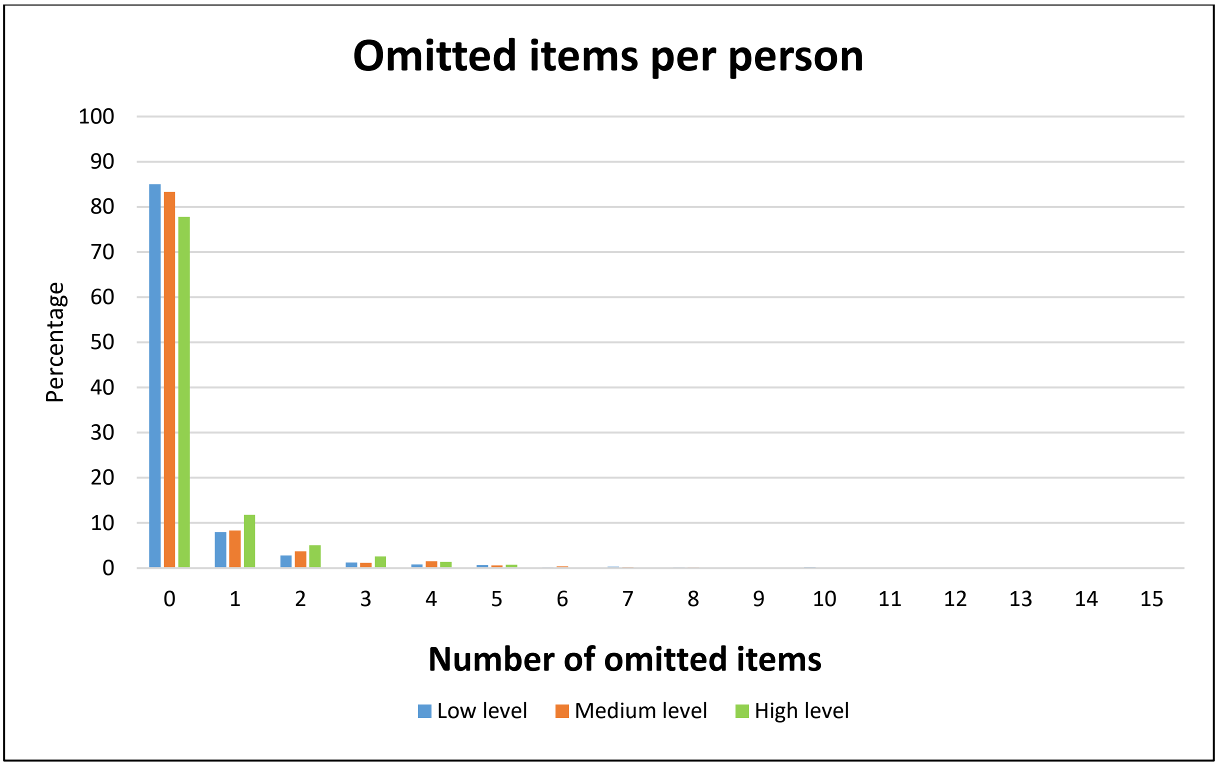

*Figure 3.* Number of omitted responses per person by test difficulty.

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 also rather low, between 74.3% and 85.6% of the respondents were able to finish the test within the allocated time limit (Figure 4). Less than 0.2% (low level), less than 0.4% (medium level), and less than 0.5% did not finish more than half of the items.

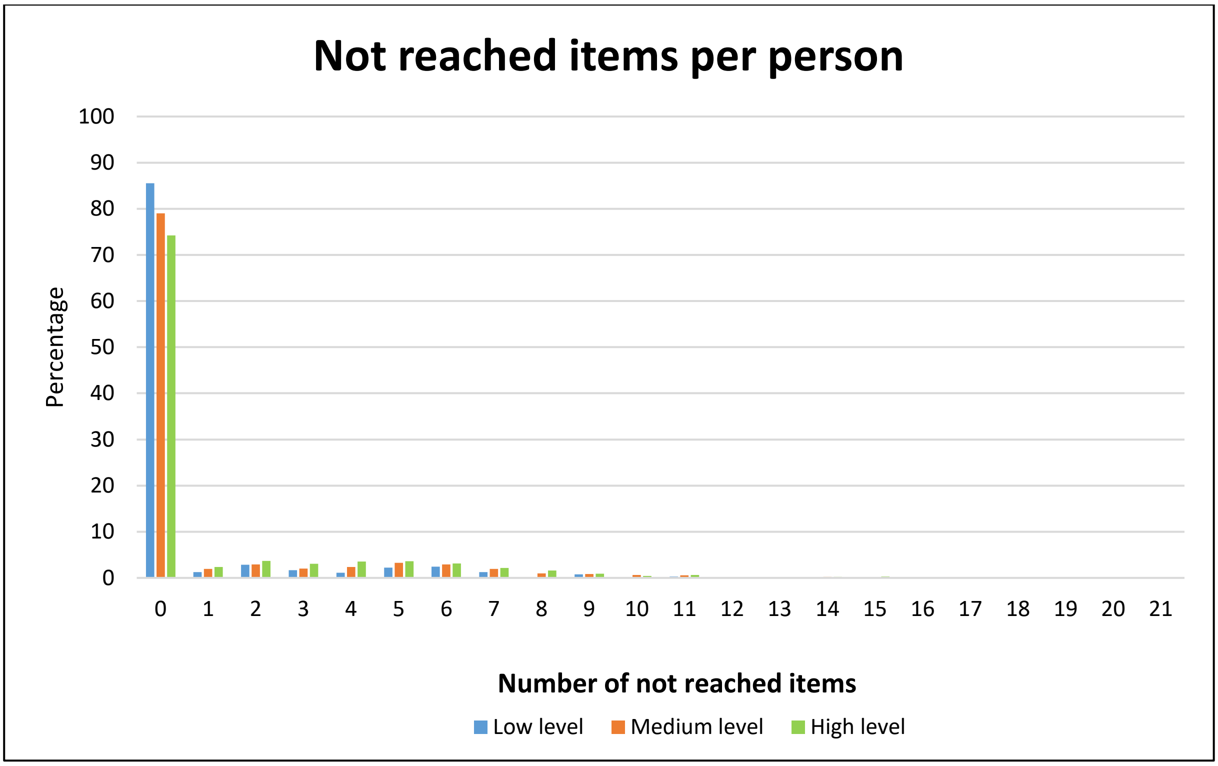

*Figure 4.* Number of not reached items per person by test difficulty.

The total number of missing responses, aggregated over invalid, omitted and not-reached missing responses, is illustrated in Figure 5. About 54.8% to 62.8% of the respondents had no missing response at all and about 14.5% to 21.5% of the participants had four or more missing responses. Particularly, respondents receiving the test with a medium or high level of difficulty showed more missing responses (19.5% and 21.5%) than respondents receiving the test with a low level of difficulty (14.5%). Overall, the amount of invalid and omitted items is small, whereas a reasonable part of missing responses occurred due to not-reached items.

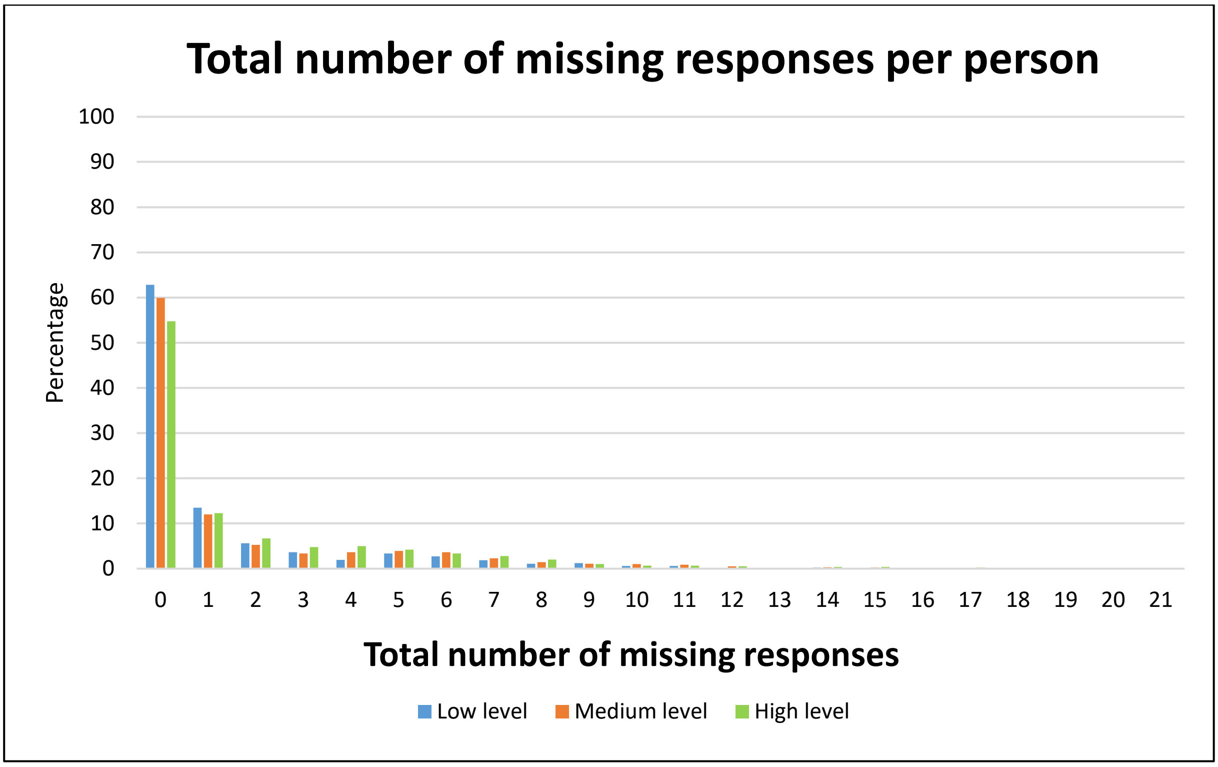

*Figure 5.* Total number of missing responses per person by test difficulty.

### 5.2.2 Missing responses per item

Tables 5 to 7 provide information on the occurrence of different kinds of missing responses per item by test difficulty. In all of the three tests, the omission rates were rather low, varying across items between 0.0 % and 6.4%. Overall, the missing rates correlated with the item difficulties at about $r$ = .297 ($p$ > .05), indicating that test-takers often missed difficult items. Generally, the percentage of invalid responses per item (column 6 Tables 5 to 7) was rather low with the maximum rate being 3.3%. With an item's progressing position in the test, the number of persons that did not reach the item (column 4 in Tables 5 to 7) rose to a reasonable amount of 14.4% to 25.7% for the different test versions.

Table 5: Percentage of Missing Values for Easy Test Version

| Item | Position in the test | Number of valid responses | Not reached items (%) | Omitted items (%) | Invalid responses (%) |
|---|---|---|---|---|---|
| scg90110_c | 1 | 1131 | 0.0 | 0.4 | 0.1 |
| scg9012s_c | 2 | 1101 | 0.0 | 0.1 | 3.1 |
| scg90510_c | 3 | 1121 | 0.0 | 1.1 | 0.4 |
| scg9052s_c | 4 | 1098 | 0.0 | 1.1 | 2.4 |
| scg90920_c | 5 | 1114 | 0.0 | 0.5 | 1.5 |
| scg90930_c | 6 | 1118 | 0.0 | 0.9 | 0.8 |
| scg9611s_c | 7 | 1105 | 0.0 | 0.3 | 2.6 |
| scg96120_c | 8 | 1121 | 0.0 | 0.4 | 1.0 |
| scg96410_c | 9 | 1121 | 0.0 | 0.6 | 0.8 |
| scg96420_c | 10 | 1110 | 0.0 | 1.9 | 0.4 |
| scg9061s_c | 11 | 1079 | 0.1 | 2.7 | 2.3 |
| scg90630_c | 12 | 1099 | 0.1 | 1.8 | 1.4 |
| scg90810_c | 13 | 1125 | 0.1 | 0.4 | 0.5 |
| scg9083s_c | 14 | 1097 | 0.1 | 1.8 | 1.7 |
| scg91030_c | 15 | 1106 | 0.2 | 2.2 | 0.4 |
| scg91040_c | 16 | 1119 | 0.2 | 0.9 | 0.5 |
| scg91050_c | 17 | 1106 | 0.3 | 2.0 | 0.4 |
| scg9042s_c | 18 | 1084 | 0.5 | 0.8 | 3.3 |
| scg9043s_c | 19 | 1089 | 0.7 | 1.1 | 2.5 |
| scg9651s_c | 20 | 1081 | 1.5 | 1.9 | 1.5 |
| scg96530_c | 21 | 1094 | 1.7 | 1.8 | 0.4 |
| scg90320_c | 22 | 1057 | 2.9 | 3.3 | 0.8 |
| scg90330_c | 23 | 1033 | 5.4 | 3.4 | 0.4 |
| scg9621s_c | 24 | 1006 | 7.6 | 1.7 | 2.3 |
| scg96220_c | 25 | 1010 | 8.7 | 1.8 | 0.6 |
| scg91110_c | 26 | 1001 | 10.4 | 0.8 | 0.8 |
| scg91120_c | 27 | 961 | 13.2 | 1.2 | 1.1 |
| scg91130_c | 28 | 969 | 14.4 | 0.0 | 0.4 |

Table 6: Percentage of Missing Values For Medium Test Version

| Item | Position in the test | Number of valid responses | Not reached items (%) | Omitted items (%) | Invalid responses (%) |
|---|---|---|---|---|---|
| scg90110_c | 1 | 2333 | 0.0 | 0.6 | 0.0 |
| scg9012s_c | 2 | 2307 | 0.0 | 0.1 | 1.7 |
| scg90510_c | 3 | 2322 | 0.0 | 0.9 | 0.2 |
| scg9052s_c | 4 | 2282 | 0.0 | 0.5 | 2.3 |
| scg90920_c | 5 | 2316 | 0.0 | 0.9 | 0.4 |
| scg90930_c | 6 | 2309 | 0.0 | 1.2 | 0.4 |
| scg9611s_c | 7 | 2252 | 0.0 | 1.4 | 2.6 |
| scg96120_c | 8 | 2328 | 0.0 | 0.5 | 0.3 |
| scg96410_c | 9 | 2219 | 0.1 | 5.2 | 0.2 |
| scg96420_c | 10 | 2308 | 0.1 | 1.3 | 0.3 |
| scg9061s_c | 11 | 2246 | 0.1 | 2.5 | 1.8 |
| scg90630_c | 12 | 2295 | 0.1 | 0.8 | 1.4 |
| scg90810_c | 13 | 2330 | 0.1 | 0.6 | 0.1 |
| scg9083s_c | 14 | 2280 | 0.2 | 1.2 | 1.5 |
| scg91030_c | 15 | 2284 | 0.4 | 2.1 | 0.2 |
| scg91040_c | 16 | 2312 | 0.5 | 0.7 | 0.3 |
| scg91050_c | 17 | 2283 | 0.6 | 2.2 | 0.0 |
| scg9042s_c | 18 | 2239 | 1.1 | 1.7 | 1.8 |
| scg9043s_c | 19 | 2215 | 1.8 | 2.3 | 1.6 |
| scg9651s_c | 20 | 2244 | 2.6 | 0.8 | 1.1 |
| scg96530_c | 21 | 2226 | 3.6 | 1.1 | 0.5 |
| scg90320_c | 22 | 2159 | 5.5 | 2.3 | 0.2 |
| scg90330_c | 23 | 2069 | 8.4 | 3.1 | 0.3 |
| scg9621s_c | 24 | 2007 | 11.7 | 1.2 | 1.6 |
| scg96220_c | 25 | 1955 | 14.1 | 1.9 | 0.8 |
| scg91110_c | 26 | 1936 | 16.1 | 1.1 | 0.4 |
| scg91120_c | 27 | 1873 | 19.0 | 0.7 | 0.5 |
| scg91130_c | 28 | 1850 | 21.0 | 0.0 | 0.3 |

Table 7: Percentage of Missing Values For Difficult Test Version

| Item | Position in the test | Number of valid responses | Not reached items (%) | Omitted items (%) | Invalid responses (%) |
|---|---|---|---|---|---|
| scg90110_c | 1 | 1386 | 0.0 | 0.5 | 0.1 |
| scg9012s_c | 2 | 1345 | 0.0 | 2.7 | 0.9 |
| scg90510_c | 3 | 1390 | 0.0 | 0.4 | 0.0 |
| scg9052s_c | 4 | 1314 | 0.0 | 5.2 | 0.6 |
| scg90920_c | 5 | 1382 | 0.0 | 0.6 | 0.3 |
| scg90930_c | 6 | 1373 | 0.0 | 1.2 | 0.4 |
| scg9611s_c | 7 | 1339 | 0.0 | 1.0 | 3.0 |
| scg96120_c | 8 | 1387 | 0.0 | 0.4 | 0.2 |
| scg96410_c | 9 | 1346 | 0.0 | 3.3 | 0.2 |
| scg96420_c | 10 | 1374 | 0.0 | 1.2 | 0.3 |
| scg9061s_c | 11 | 1355 | 0.0 | 1.7 | 1.1 |
| scg90630_c | 12 | 1377 | 0.0 | 0.6 | 0.6 |
| scg90810_c | 13 | 1363 | 0.0 | 2.1 | 0.2 |
| scg9083s_c | 14 | 1333 | 0.3 | 4.1 | 0.1 |
| scg91030_c | 15 | 1369 | 0.5 | 1.3 | 0.1 |
| scg91040_c | 16 | 1376 | 0.6 | 0.6 | 0.1 |
| scg91050_c | 17 | 1369 | 0.6 | 1.2 | 0.0 |
| scg9042s_c | 18 | 1319 | 1.3 | 0.9 | 3.3 |
| scg9043s_c | 19 | 1338 | 1.7 | 1.4 | 1.0 |
| scg9651s_c | 20 | 1346 | 2.7 | 0.3 | 0.6 |
| scg96530_c | 21 | 1329 | 4.2 | 0.4 | 0.1 |
| scg90320_c | 22 | 1287 | 6.4 | 1.1 | 0.2 |
| scg90330_c | 23 | 1225 | 9.5 | 2.4 | 0.2 |
| scg9621s_c | 24 | 1121 | 13.1 | 6.4 | 0.1 |
| scg96220_c | 25 | 1134 | 16.6 | 1.8 | 0.3 |
| scg91110_c | 26 | 1107 | 19.7 | 0.8 | 0.1 |
| scg91120_c | 27 | 1052 | 23.4 | 1.0 | 0.2 |
| scg91130_c | 28 | 1035 | 25.7 | 0.0 | 0.1 |

## 5.3 Parameter estimates

### 5.3.1 Item parameters

Because preliminary analyses identified satisfactory measurement models for each test version, the following analyses report the results of the concurrent scaling of all three test versions. Column 3 in Table 8 shows the percentage of correct responses in relation to all valid responses for each item. Note that since there was a non-negligible amount of missing responses, this probability cannot be interpreted as an index for item difficulty. The percentage of correct responses within items varied between 6.0% and 85.0% with an average of 46.3% (*SD* = 18.1) correct responses.

The estimated item difficulties (for dichotomous items, MC items) and location parameters (for polytomous variables, CMC items) are given in Table 8. The step parameters (for polytomous variables) are depicted in Table 9.

For three of the CMC items (scg9052s_c, scg9083s_c, scg9752s_c) the two lowest categories were collapsed, thus, these items were scaled using a scoring of 0, 0.5, 1, and 1.5. For three of the CMC items (scg9012s_c, scg9042s_c, scg9043s_c) the three lowest categories were collapsed, thus, these items were scaled using a scoring of 0, 0.5, and 1. One of the CMC items (scg9061s_c) was treated as a MC-item using a scoring of 0 and 1 (right answer on all subtasks).

The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties (or location parameters for polytomous variables) ranged between −2.13 (scg90810_c) and 2.17 (scg97910_c). In total, the estimated item difficulties had a mean of −0.20 (*SD* = 0.96). Due to the large sample size, the standard errors of the estimated item difficulties were very small (SE(*ß*) ≤ 0.08). Overall, the item difficulties were rather low; the test did not include many items with high difficulty.

Table 8: Item parameters

| No. | Item | Correct (%) | Item difficulty | *SE* | WMNSQ | *t* | $r_{\text{it}}$ | Discrimination (GPCM) | Q3 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | scg90110_c | 63.0 | −0.61 | 0.03 | 1.00 | 0.4 | 0.37 | 0.76 | 0.14 |
| 2 | scg9012s_c | n.a. | −0.94 | 0.04 | 0.97 | −1.6 | 0.38 | 1.11 | 0.07 |
| 3 | scg90510_c | 66.3 | −0.78 | 0.03 | 1.01 | 1.0 | 0.36 | 0.80 | 0.07 |
| 4 | scg9052s_c | n.a. | −0.84 | 0.04 | 0.96 | −1.7 | 0.41 | 1.88 | 0.05 |
| 5 | scg90920_c | 44.5 | 0.26 | 0.03 | 1.09 | 8.3 | 0.29 | 0.48 | 0.06 |
| 6 | scg90930_c | 54.4 | −0.73 | 0.06 | 1.03 | 1.5 | 0.31 | 0.56 | 0.09 |
| 7 | scg96120_c | 67.3 | −0.82 | 0.03 | 1.00 | 0.2 | 0.37 | 0.81 | 0.09 |
| 8 | scg96410_c | 66.6 | −1.30 | 0.07 | 0.98 | −0.9 | 0.38 | 0.99 | 0.08 |
| 9 | scg96420_c | 57.9 | −0.36 | 0.03 | 0.92 | -7.7 | 0.49 | 1.31 | 0.07 |
| 10 | scg9061s_c | 46.5 | 0.17 | 0.03 | 0.93 | -7.1 | 0.49 | 1.26 | 0.10 |
| 11 | scg90630_c | 69.6 | −0.94 | 0.03 | 0.96 | -3.1 | 0.42 | 1.10 | 0.09 |
| 12 | scg90810_c | 85.0 | −2.13 | 0.05 | 0.92 | −2.5 | 0.41 | 1.62 | 0.07 |
| 13 | scg9083s_c | n.a. | −1.31 | 0.04 | 0.91 | -4.8 | 0.51 | 2.17 | 0.10 |
| 14 | scg91030_c | 47.1 | 0.14 | 0.03 | 1.01 | 0.7 | 0.40 | 0.82 | 0.09 |
| 15 | scg91040_c | 66.3 | −1.29 | 0.07 | 1.01 | 0.6 | 0.32 | 0.70 | 0.06 |
| 16 | scg91050_c | 66.6 | −0.79 | 0.03 | 1.00 | 0.2 | 0.37 | 0.86 | 0.10 |
| 17 | scg9042s_c | n.a. | −0.75 | 0.07 | 0.96 | −1.7 | 0.42 | 1.31 | 0.07 |
| 18 | scg9043s_c | n.a. | −0.98 | 0.08 | 1.03 | 1.2 | 0.23 | 0.54 | 0.10 |
| 19 | scg9651s_c | n.a. | −1.18 | 0.03 | 1.02 | 0.8 | 0.46 | 1.56 | 0.07 |
| 20 | scg96530_c | 53.8 | −0.18 | 0.03 | 0.96 | -3.6 | 0.44 | 1.03 | 0.08 |
| 21 | scg90320_c | 59.1 | −0.42 | 0.03 | 0.87 | -12.1 | 0.56 | 1.78 | 0.07 |
| 22 | scg90330_c | 44.3 | 0.27 | 0.03 | 1.02 | 1.4 | 0.36 | 0.75 | 0.16 |
| 23 | scg9621s_c | n.a. | −1.13 | 0.03 | 0.96 | −2.0 | 0.51 | 2.33 | 0.05 |
| 24 | scg96220_c | 54.4 | −0.22 | 0.03 | 1.03 | 2.5 | 0.36 | 0.72 | 0.11 |
| 25 | scg91110_c | 47.9 | 0.08 | 0.03 | 1.07 | 6.2 | 0.31 | 0.56 | 0.07 |
| 26 | scg91120_c | 23.7 | 1.30 | 0.04 | 1.02 | 1.0 | 0.31 | 0.69 | 0.06 |
| 27 | scg91130_c | 32.6 | 0.80 | 0.04 | 1.10 | 6.1 | 0.25 | 0.40 | 0.06 |
| 28 | scg97410_c | 22.3 | 1.57 | 0.04 | 1.00 | −0.1 | 0.32 | 0.80 | 0.07 |
| 29 | scg9771s_c | n.a. | 0.84 | 0.03 | 1.17 | 7.7 | 0.20 | 0.37 | 0.06 |
| 30 | scgb6320_c | 33.7 | 0.95 | 0.04 | 1.06 | 3.7 | 0.28 | 0.52 | 0.16 |
| 31 | scg98910_c | 32.4 | 0.10 | 0.04 | 1.05 | 3.4 | 0.28 | 0.52 | 0.06 |
| 32 | scg9751s_c | n.a. | −0.33 | 0.03 | 1.06 | 2.8 | 0.34 | 1.02 | 0.06 |
| 33 | scg9752s_c | n.a. | −0.84 | 0.03 | 0.91 | -4.0 | 0.55 | 2.75 | 0.10 |
| 34 | scg98010_c | 70.2 | −0.46 | 0.06 | 1.00 | −0.1 | 0.37 | 0.90 | 0.11 |
| 35 | scg97910_c | 17.7 | 2.17 | 0.08 | 1.00 | 0.1 | 0.31 | 0.91 | 0.07 |
| 36 | scg98210_c | 40.1 | 0.93 | 0.06 | 0.99 | −0.3 | 0.36 | 0.73 | 0.10 |
| 37 | scg98310_c | 63.6 | −0.14 | 0.06 | 1.04 | 1.9 | 0.31 | 0.57 | 0.09 |
| 38 | scgb6210_c | 30.8 | 1.38 | 0.07 | 1.00 | 0.0 | 0.32 | 0.72 | 0.08 |

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

Table 9: Step parameters for the CMC items

| Item | Step 1 (*SE*) | Step 2 (*SE*) | Step 2 (*SE*) | Step 3 |
|---|---|---|---|---|
| scg9012s_c | 0.19 (0.04) | −0.19 | | |
| scg9052s_c | −0.45 (0.04) | −1.01 (0.04) | 1.46 | |
| scg9083s_c | −0.56 (0.04) | 0.33 (0.04) | 0.23 | |
| scg9042s_c | 0.42 (0.07) | −0.42 | | |
| scg9043s_c | −0.07 (0.07) | 0.07 | | |
| scg9651s_c | 0.70 (0.03) | −1.29 (0.03) | 1.43 (0.05) | −0.84 |
| scg9621s_c | −0.22 (0.04) | −0.47 (0.04) | 0.40 (0.04) | 0.29 |
| scg9771s_c | −0.80 (0.04) | −0.71 (0.03) | 0.84 (0.05) | 0.67 |
| scg9751s_c | −0.62 (0.04) | −0.93 (0.04) | 0.94 (0.04) | 0.61 |
| scg9752s_c | −0.06 (0.04) | −0.45 (0.03) | 0.02 (0.04) | 0.49 |

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

### 5.3.2 Person parameters

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

### 5.3.3 Test targeting and reliability

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

The mean of the ability distribution was constrained to be zero. The variance was estimated to be 0.693, indicating a somewhat limited variability between subjects. The reliability of the test (EAP/PV reliability = .775; WLE reliability = .803) was acceptable. Although the items covered a wide range of the ability distribution, few items were covering the lower and upper peripheral ability areas. As a consequence, person ability in medium ability regions will be measured relative precisely, whereas lower and higher ability estimates will have larger standard errors of measurement.

| Scale in logits | Person ability | Item difficulty |
|---|---|---|
| | | 35 |
| 2 | | |
| | | |
| | X | |
| | XX | |
| | XX | 28 |
| | X | |
| | XX | 38 |
| | XXX | 26 |
| | XXXX | |
| 1 | XXX | |
| | XXXX | 30 31 36 |
| | XXXXXX | 27 29 |
| | XXXXXX | |
| | XXXXXXX | |
| | XXXXXXXX | |
| | XXXXXXXXX | |
| | XXXXXXXX | 5 22 |
| | XXXXXXXXX | 10 14 |
| | XXXXXXXX | 25 |
| 0 | XXXXXX | |
| | XXXXXXX | 20 24 37 |
| | XXXXXXX | 32 |
| | XXXXXXXX | 9 21 |
| | XXXXXXXXXX | 34 |
| | XXXXXXX | 1 |
| | XXXXXX | 3 6 17 |
| | XXXX | 4 7 16 33 |
| | XXXXX | 2 11 18 |
| -1 | XXXXX | |
| | XXX | 19 23 |
| | XXX | 8 13 15 |
| | XX | |
| | XX | |
| | X | |
| | X | |
| | X | |
| | X | |
| -2 | | |
| | | 12 |

*Figure 6.* Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 29.5 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see Table 8).

## 5.4 Quality of the test

### 5.4.1 Fit of the subtasks of complex multiple-choice items

Before the subtasks of the CMC item were aggregated and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the single subtasks together with the MC items in a Rasch model. Counting the subtasks of the CMC item separately, there were 78 items. The percentage of a correct response ranged from 17.7% to 88.7% across all items (*Mdn* = 66.5%). Thus, the number of correct and incorrect responses was reasonably large. All subtasks of the CMC items showed a satisfactory item fit. WMNSQ ranged from 0.88 to 1.18, the respective *t*-value from −7.4 to 16.0, and there were no noticeable deviations of the empirically estimated probabilities from the model-implied item characteristic curves. Due to the good model fit of the subtasks, their aggregation to a polytomous variable seemed justified.

### 5.4.2 Distractor analyses

In addition to the overall item fit, we specifically investigated how well the distractors performed in the test by evaluating the point-biserial correlation between each incorrect response (distractor) and the students' total score. All accept two distractors had a point-biserial correlation with the total scores below zero. One distractor had a point-biserial correlation of 0.10 (scg91120), and one had a point-biserial correlation of 0.00 (scg97410). Besides that, the results indicate that the distractors worked well.

### 5.4.3 Item fit

The evaluation of the item fit was performed based on the final scaling model, the partial credit model, using the MC items and the CMC items. Altogether, the item fit can be considered to be very good (see Table 8). Values of the WMNSQ ranged from 0.87 (item scg90320_c) to 1.17 (scg9771s_c). Four items exhibited a *t*-value of the WMNSQ greater than 6 (scg90920_c, scg91110_c, scg91130_c, and scg9771s_c). The highest t-value was 8.3 (scg90920_c). Thus, there was no indication of a severe item over- or underfit. Point-biserial correlations between the item scores and the total scores ranged from .09 (item scg9771s_c) to .45 (item scg9621s_c) and had a mean of .37. All item characteristic curves showed a good fit of the items to the PCM.

### 5.4.4 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables gender, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Table 10 shows the difference between the estimated item difficulties in different groups. Male vs. female, for example, indicates the difference in difficulty ß(male) – ß(female). A positive value indicates a higher difficulty for males, a negative value a lower difficulty for males as opposed to females. Also, Table 8 shows the main effect for the examined subgroups (inclusive Cohen's d).

Table 10: Differential item functioning (differences between difficulties)

| Item | Gender | Books | | | Migration status | | | School type | Rotation |
|---|---|---|---|---|---|---|---|---|---|
| | Male vs. female | <100 vs. >100 | <100 vs. missing | >100 vs. missing | Without vs. With | Without vs. Missing | With vs. Missing | Other vs. Gymnasium | ICT first vs. Science first |
| scg90110_c | −0.242 | 0.044 | 0.006 | −0.034 | 0.024 | −0.092 | −0.126 | −0.322 | 0.026 |
| scg9012s_c | 0.122 | 0.232 | −0.010 | −0.228 | −0.284 | −0.136 | 0.130 | 0.160 | 0.054 |
| scg90510_c | −0.220 | −0.118 | 0.022 | 0.144 | 0.134 | 0.070 | −0.072 | −0.156 | 0.052 |
| scg9052s_c | −0.128 | 0.094 | 0.118 | 0.018 | 0.012 | −0.018 | −0.038 | 0.300 | 0.024 |
| scg90920_c | −0.208 | −0.186 | 0.018 | 0.210 | −0.034 | 0.208 | 0.234 | −0.352 | −0.028 |
| scg90930_c | −0.212 | 0.098 | 0.054 | −0.040 | 0.024 | 0.070 | 0.034 | −0.092 | 0.326 |
| scg96120_c | 0.056 | −0.020 | −0.120 | −0.096 | 0.188 | −0.008 | −0.204 | 0.028 | 0.152 |
| scg96410_c | 0.320 | 0.384 | 0.240 | −0.140 | 0.048 | 0.040 | −0.022 | 0.132 | 0.026 |
| scg96420_c | 0.288 | 0.032 | −0.270 | −0.298 | −0.236 | −0.302 | −0.076 | 0.288 | 0.054 |
| scg9061s_c | 0.294 | 0.266 | 0.028 | −0.234 | 0.192 | −0.052 | −0.252 | 0.374 | −0.150 |
| scg90630_c | 0.590 | 0.338 | 0.054 | −0.280 | −0.200 | −0.142 | 0.050 | 0.348 | 0.172 |
| scg90810_c | −0.448 | 0.254 | −0.058 | −0.308 | −0.336 | −0.300 | 0.024 | 0.444 | 0.208 |
| scg9083s_c | −0.840 | 0.344 | 0.002 | −0.356 | −0.276 | −0.234 | 0.052 | 0.378 | −0.180 |
| scg91030_c | 0.086 | −0.064 | −0.040 | 0.028 | −0.086 | 0.046 | 0.122 | −0.098 | −0.052 |
| scg91040_c | 0.110 | 0.178 | 0.284 | 0.108 | −0.164 | 0.110 | 0.264 | −0.124 | 0.286 |
| scg91050_c | −0.366 | −0.030 | −0.036 | −0.002 | −0.414 | 0.018 | 0.422 | −0.192 | −0.044 |
| scg9042s_c | 0.260 | 0.046 | 0.190 | 0.140 | −0.158 | 0.168 | 0.348 | 0.014 | 0.354 |
| scg9043s_c | 0.348 | −0.120 | −0.300 | −0.180 | 0.254 | −0.074 | −0.332 | 0.590 | −0.094 |
| scg9651s_c | 0.188 | −0.026 | −0.054 | −0.022 | −0.062 | −0.098 | −0.046 | −0.050 | −0.028 |
| scg96530_c | 0.172 | 0.014 | −0.102 | −0.110 | 0.036 | −0.048 | −0.092 | 0.164 | 0.118 |
| scg90320_c | −0.074 | 0.254 | −0.046 | −0.298 | −0.286 | −0.198 | 0.078 | 0.306 | −0.034 |
| scg90330_c | 0.112 | −0.016 | −0.102 | −0.082 | 0.240 | 0.032 | −0.218 | 0.066 | −0.070 |
| scg9621s_c | 0.110 | 0.026 | 0.082 | 0.036 | −0.082 | −0.028 | 0.074 | 0.182 | −0.066 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| scg96220_c | 0.260 | −0.014 | 0.010 | 0.028 | 0.080 | −0.002 | −0.090 | 0.062 | −0.058 |
| scg91110_c | 0.000 | −0.194 | 0.088 | 0.284 | 0.240 | 0.188 | −0.060 | −0.142 | 0.022 |
| scg91120_c | −0.024 | 0.020 | 0.244 | 0.228 | −0.038 | 0.180 | 0.210 | −0.114 | −0.114 |
| scg91130_c | 0.000 | −0.214 | −0.002 | 0.218 | 0.116 | 0.222 | 0.098 | −0.328 | −0.116 |
| scg97410_c | 0.150 | −0.148 | −0.198 | −0.048 | −0.046 | −0.128 | −0.090 | 0.002 | 0.028 |
| scg9771s_c | 0.150 | −0.358 | 0.108 | 0.468 | 0.004 | 0.294 | 0.282 | −0.462 | −0.118 |
| scgb6320_c | −0.280 | −0.368 | 0.000 | 0.372 | 0.370 | 0.240 | −0.138 | −0.214 | 0.106 |
| scg98910_c | −0.102 | −0.168 | 0.016 | 0.186 | −0.034 | 0.146 | 0.172 | −0.294 | −0.158 |
| scg9751s_c | 0.088 | 0.020 | 0.114 | 0.112 | 0.172 | 0.084 | −0.088 | −0.198 | 0.070 |
| scg9752s_c | −0.258 | 0.284 | 0.044 | −0.224 | 0.008 | −0.152 | −0.176 | 0.264 | 0.022 |
| scg98010_c | −0.260 | −0.018 | 0.194 | 0.212 | 0.268 | 0.148 | −0.134 | −0.268 | 0.006 |
| scg97910_c | −0.050 | 0.132 | 0.144 | 0.016 | 0.038 | −0.080 | −0.130 | 0.632 | −0.076 |
| scg98210_c | −0.402 | −0.144 | −0.146 | 0.000 | 0.008 | 0.016 | −0.004 | −0.216 | −0.080 |
| scg98310_c | −0.246 | −0.124 | 0.126 | 0.252 | 0.202 | 0.154 | −0.060 | −0.306 | 0.138 |
| scgb6210_c | 0.204 | −0.232 | −0.238 | −0.004 | 0.466 | −0.058 | −0.536 | 0.022 | 0.018 |

### *Gender*

The sample included 2,458 (50.4%) male test-takers (coded 0) and 2,422 (49.6%) female test-takers (coded 1). On average, male students had slightly higher scores in scientific literacy than female students (main effect = 0.120 logits, Cohen's *d* = 0.144). However, there was one item with considerable gender DIF (scg9083s_c). The difference in difficulties between the two groups was −0.840 logits.

### *Books*

The number of books at home was used as a proxy for socioeconomic status. There were 1,050 (21.5%) test takers with 0 to 100 books at home (coded 0), 2,030 (41.6%) test takers with more than 100 books at home (coded 1), and 1,800 (36.9%) test-takers did not give a valid response (coded 9). DIF was investigated using these three groups. There were considerable average differences between these three groups. Participants with 100 or fewer books at home performed showed lower scientific literacy scores than participants with more than 100 books (main effect = −0.628 logits, Cohen's *d* = −0.813). Participants without a valid response on the variable 'books at home' performed higher than participants with up to 100 (main effect = −0.066 logits, Cohen's *d* = −0.083) and also lower than participants with more than 100 books at home (main effect = −0.558 logits, Cohen's *d* = −0.714). There was no considerable DIF comparing participants with many or fewer books (highest DIF = 0.384). Comparing the group without valid responses to the two groups with valid responses, DIF occurred up to −0.356 logits (Participants with 100 or fewer books at home vs. Participants without a valid response).

### *Migration background*

There were 2,408 (49.3%) participants without a migration background (coded 0) and 584 (12.0%) participants with a migration background (coded 1; for 0.9% students neither their mother, father or themselves were born in Germany, for 4.7% only the participants were born in Germany and both of their parents were born abroad, and for 6.5% of the participants only one of their parents was born abroad). A total of 1,888 (38.7%) students could not be allocated to either group (coded 9). These groups were used for investigating DIF of migration. There was a considerable difference in the average performance of participants with or without migration background. Participants without a migration background showed higher scientific literacy scores than participants with a migration background (main effect = 0.448 logits, Cohen's *d* = 0.547) and also higher scores than students with an unknown background on migration (main effect = 0.414 logits, Cohen's *d* = 0.523). Furthermore, students with a migration background performed lower than those with an unknown background on migration (main effect = −0.030 logits, Cohen's *d* = −0.037). There was no considerable DIF comparing participants with and without a migration background (highest DIF = −0.414). Comparing the group without valid responses to the two groups with valid responses, DIF occurred up to 0.422 logits.

### *Type of School*

DIF was also investigated for the type of secondary school. At the end of primary school, children in Germany will be mainly allocated for secondary school to one of the following types: “Hauptschule”, a secondary general school for grades five through nine or ten, “Realschule”, a more practical secondary school for grades five through ten, or “Gymnasium”, a more academic secondary school for grades five through twelve/thirteen. There were 2,317 (47.5%) students visiting “Gymnasium” (coded 1), and 2,563 (52.5%) students from lower schools (coded 0), such as “Hauptschule” or “Realschule”. On average, students visiting “Gymnasium” had distinctly higher scores in scientific literacy than students from other school types (main effect = –0.856 logits, Cohen’s d = −1.197). But there was no item with a considerable DIF. The highest difference in difficulties between the two groups was 0.590 logits.

### *Rotation – Test order*

The scientific literacy test was administered in two different positions (see section 3.1 for the design of the study). A total of 2,456 (50.3%) received the ICT literacy test before completing the scientific literacy test (coded 0), while 2,424 (49.6%) of the test takers received the scientific literacy test first and then the ICT literacy test (coded 1). The students were randomly assigned to either of the two design groups. Differential item functioning of the position of the test may, for example, occur if the different certain parts or items of the test are more or less tiring for the participants. There was a small difference between the two test positions (main effect = –0.088 logits, Cohen’s d = −0.106), indicating a lower difficulty for test-takers, who started with the scientific literacy test. Also, the highest difference in difficulties between the two groups is 0.354 logits.

Table 11: Main effects and Cohen’s d of the examined subgroups

| Variables | Subgroups | Main effect | Cohen`s d |
|---|---|---|---|
| **Gender** | Male (0)<br>Female (1) | 0.120 | 0.144 |
| **Books** | 0 to 100 books at home (0)<br>More than 100 books at home (1) | −0.628 | −0.813 |
| | 0 to 100 books at home (0)<br>Invalid response (9) | −0.066 | −0.083 |
| | More than 100 books at home (1)<br>Invalid response (9) | 0.558 | 0.714 |
| **Migration background** | Without migration background (0)<br>With migration background (1) | 0.488 | 0.547 |
| | Without migration background (0)<br>Invalid response (9) | 0.414 | 0.523 |
| | With migration background (1)<br>Invalid response (9) | −0.030 | −0.037 |
| **School type** | Other school type (0)<br>Gymnasium (1) | −0.856 | -1.197 |
| **Rotation** | Science first (0)<br>Science second (1) | −0.088 | −0.106 |

*Note. The numbers behind the subgroups display their coding.*

Besides investigating DIF for every single item, an overall test for DIF was performed by comparing models that allow for DIF with those that allow only for main effects. In Table 12, the models including only the main effects are compared with those that additionally estimate DIF. For these models, we used the valid responses from the participants. For example, the variable books represents the comparison of the participants with less than 100 books and those with more than 100 books. Akaike (1974) information criterion (AIC) and the Bayesian information criterion (BIC, Schwarz, 1978) were used for comparing the models. The AIC favored the model considering DIF for four DIF variables (gender, books, migration background, and school type). Only for rotation, the AIC favored the model which only allows for main effects. The BIC takes the number of estimated parameters into account and, thus, prevents from overparameterization of models. Using BIC, the more parsimonious model including only the main effect is preferred over the more complex DIF model for three DIF variables (books, migration background, and rotation).

Table 12: Comparison of models with and without DIF

| DIF variable | Model | Deviance | N | Number of parameters | AIC | BIC |
|---|---|---|---|---|---|---|
| **Gender** | main effect | 195275.11 | 4880 | 62 | 195399.11 | 195801.67 |
| | DIF | 194810.86 | 4880 | 100 | 195010.86 | 195660.15 |
| **Books** | main effect | 121031.14 | 3080 | 62 | 121155.14 | 121529.17 |
| | DIF | 120913.60 | 3080 | 100 | 121113.60 | 121716.87 |
| **Migration background** | main effect | 117874.04 | 2992 | 62 | 117998.04 | 118370.27 |
| | DIF | 117788.01 | 2992 | 100 | 117988.01 | 118588.38 |
| **School type** | main effect | 194154.13 | 4880 | 62 | 194278.13 | 194680.69 |
| | DIF | 193765.42 | 4880 | 100 | 193965.42 | 194614.71 |
| **Rotation** | main effect | 195284.29 | 4880 | 62 | 195408.29 | 195810.85 |
| | DIF | 195213.05 | 4880 | 100 | 195413.05 | 196062.34 |

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

### 5.4.5 Rasch-homogeneity

An essential assumption of the Rasch (1980) model is that all item-discrimination parameters are equal. To test this assumption, a generalized partial credit model (GPCM; Muraki, 1992) that estimates discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 8), ranging from 0.37 (item scg9771s_c) to 2.75 (item scg9752s_c). The average discrimination parameter fell at 1.02. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 193,885.47, BIC = 194,521.78) as compared to the PCM model (AIC = 195,417.20, BIC = 195,813.26). Despite the empirical preference for the GPCM, the PCM model matches the theoretical conceptions underlying the test construction more adequately (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the partial credit model was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

### 5.4.6 Unidimensionality of the test

The dimensionality of the test was investigated by specifying a one- and a two- dimensional model. The first model is based on the assumption that scientific literacy is a one-dimensional construct that measures one distinct competence whereas the second model distinguishes between the two sub-competencies: the process-related components (knowledge about science – KAS) and the content-related components (knowledge of science – KOS; for more details see Hahn et al., 2013). For estimating a two-dimensional model Gauss' Hermite quadrature estimation in ConQuest was used (nodes were chosen in such a way that stable parameter estimation was obtained). The two-dimensional model (BIC = 195,802.81, number of parameters = 63) fitted the data slightly better than the unidimensional model (BIC = 195,813.26, number of parameters = 61). As the correlation between the two dimensions was *r* = .95, the one-dimensional measurement model was used to estimate a single competence score for scientific literacy.

# 6 Discussion

The analyses in the previous sections aimed at providing detailed information on the quality of the science test administered in Grade 9 of Starting Cohort 3 and at describing how scientific literacy was estimated.

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

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

Indicated by various fit criteria – WMNSQ, t-value of the WMNSQ – the items exhibited a good item fit. Also, discrimination values of the items (either estimated in a GPCM or as a correlation of the item score with total score) were acceptable. Different variables were used for testing measurement invariance across various subgroups. No considerable DIF except one for gender became evident for any of these variables, indicating that the test was mainly fair to the considered subgroups.

Fitting a two-dimensional partial credit model (the dimensions being the "content-related components" and the "process-related components") yielded no better model fit than the unidimensional partial credit model. Moreover, the high correlation between the two dimensions indicates that a unidimensional model describes the data reasonably well.

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

# 7 Data in the Scientific Use file

## 7.1 Naming conventions

There are 38 items in the data set that are either scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response or scored as a polytomous variable (CMC items) indicating the (partial) credit. The dichotomous variables are marked with a '_c' at the end of the variable name, the CMC items are marked with a 's_c' at the end of the variable name. Note that the value of the polytomous variable does not necessarily indicate the number of correctly responded subtasks (see section 4.2 aggregation of CMC items). In the scaling model, each category of CMC items was scored with 0.5 points. Manifest scale scores are provided in form of WLE estimates (scg9_sc1) including the respective standard error (scg9_sc2). Please note that when categories of the polytomous variables had less than 200 valid responses, the categories were collapsed. For the science test this concerned the two lowest categories of three CMC items (scg9052s_c, scg9083s_c, scg9752s_c), and the three lowest categories of three CMC items (scg9012s_c, scg9042s_c, scg9043s_c; see section 5.3.1)**Fehler! Verweisquelle konnte nicht gefunden werden.**. In the scaling model, the collapsed polytomous item was scored in steps of 0.0, 0.5, 1.0, and 1.5 (denoting the highest) for items with the two lowest categories collapsed, and steps of 0.0, 0.5, and 1.0 (denoting the highest) for items with the three lowest categories collapsed. One of the CMC items (scg9061s_c) was treated as a MC-item using a scoring of 0 and 1 (right answer on all subtasks). The ConQuest Syntax for estimating the WLE scores from the items is provided in Appendix A. Students who did not take part in the test or those who did not have enough valid responses to estimate a scale score have a non-determinable missing value on the WLE score for scientific literacy.

## 7.2 Linking of competence scores

In Starting Cohort 3, the scientific literacy tests which were administered in grades 6 and 9, included different items that were constructed in such a way as to allow for an accurate measurement of scientific literacy within each age group. As a consequence, the competence scores derived in the different grades cannot be directly compared. Differences in observed scores would reflect differences in competences as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competencies across grades, I adopted the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016). Following an anchor-group-design, all items from the grade 6 and the grade 9 scientific literacy tests were administered in an independent link sample – including students from Grade 9 that were not part of Starting Cohort 3 – within a single measurement occasion. These responses were used to link the two tests administered in Starting Cohort 3 across the two grades.

### 7.2.1 Samples

In Starting Cohort 3, a subsample of 3,169 students participated at both measurement occasions, in grade 6 and also in grade 9. Consequently, N = 3,169 students were used to link the two tests across both grades (Fischer et al., 2016). Moreover, an independent link sample of *N* = 399 students (53.1% female) from Grade 9 received both tests within a single measurement occasion.

### 7.2.2 The design of the link study

The test administered in the linking sample included 25 items for grade 6, whereas the test administered for grade 9 included 28 items from the easy test version (see above). One item (scg9611s_c) had to be excluded due to insufficient item quality (see section 4). Thus, this item was also excluded from the linking. The science tests were administered in random order. Half of the sample received the grade 6 test before working on the grade 9 test, whereas the other half received the grade 9 test before the grade 6 test. No multi-matrix design regarding the selection and order of the items within a test was established. Thus, all test takers were given the science items in the same order.

### 7.2.3 Results

To examine whether the two tests administered in the link sample measured the same construct, we compared a one-dimensional model that specified a single latent factor for all items to a two-dimensional model that specified separate latent factors for the two tests.

The information criteria favored the two-dimensional model (AIC = 22,675.52, BIC = 22,958.73), over the one-dimensional model (AIC = 22,717.61, BIC = 22,992.85). However, an examination of the residual correlations for the one-dimensional model using the corrected $Q_3$ statistic (Yen, 1984) indicated a largely unidimensional scale – the average absolute residual correlation was $M$ = 0.00 ($SD$ = 0.06). This indicates that the scientific literacy tests administered in grades 6 and 9 were essentially unidimensional.

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal subsample from the starting cohort. The differences in item difficulties between the link sample and Starting Cohort 3 and the respective tests for measurement invariance based on the Wald statistic (Fischer et al., 2016) are summarized in Table 13.

Measurement invariance for grade 6 and grade 9 showed no items with $F$-statistics exceeding the critical value of $F_{.0154}(1, 3{,}568) = 83.32$. Consequently, no item had to be excluded from the estimation of the correction term.

Moreover, analyses of differential item functioning between the link sample and Starting Cohort 3 showed in grade 6 for 21 items of the test no DIF greater than 0.40 (difference in logits: *Min* = −0.32, *Max* = 0.22). But, four items (scg66320_c, scg6144s_c, scg6661s_c and scg61310_c) showed a DIF greater than 0.40. These items were therefore excluded from the estimation of the correction term. For Grade 9 (difference in logits: *Min* = −0.19, *Max* = 0.39) there were no items with a DIF greater than 0.40. Therefore, the scientific literacy tests administered in the two grades were linked using the “mean/mean” method for the anchor-group design (Fischer et al., 2016).

The correction term was calculated as $c$ = 0.8782 (with a link error of 0.088). This correction term was subsequently added to each difficulty parameter estimated in Grade 9 (see Table 8) to derive the linked item parameters.

Table 13: Differential Item Functioning Analyses between the Starting Cohort and the Link Sample

| | **Grade 6** | | | | **Grade 9** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Item** | **Δσ** | $SE_{\Delta\sigma}$ | $F$ | **Item** | **Δσ** | $SE_{\Delta\sigma}$ | $F$ |
| 1. | scg6103s_c | 0.11 | 0.13 | 0.65 | scg90110_c | 0.19 | 0.12 | 2.35 |
| 2. | scg61050_c | −0.02 | 0.15 | 1.16 | scg9012s_c | 0.18 | 0.15 | 1.46 |
| 3. | scg60120_c | −0.06 | 0.15 | 14.16 | scg90510_c | −0.37 | 0.12 | 9.58 |
| 4. | scg60410_c | −0.06 | 0.14 | 20.69 | scg9052s_c | 0.07 | 0.15 | 0.21 |
| 5. | scg60430_c | −0.01 | 0.17 | 0.14 | scg90920_c | −0.02 | 0.12 | 0.03 |
| 6. | scg66310_c | −0.01 | 0.13 | 0.76 | scg90930_c | −0.35 | 0.13 | 7.45 |
| 7. | scg66320_c | 0.87 | 0.22 | 15.32 | scg96120_c | 0.17 | 0.13 | 1.66 |
| 8. | scg66340_c | 0.45 | 0.17 | 6.68 | scg96410_c | 0.08 | 0.15 | 0.28 |
| 9. | scg61410_c | 0.28 | 0.24 | 1.33 | scg96420_c | −0.06 | 0.12 | 0.26 |
| 10. | scg6142s_c | −0.03 | 0.13 | 5.06 | scg9061s_c | −0.11 | 0.12 | 0.89 |
| 11. | scg61430_c | 0.07 | 0.18 | 0.16 | scg90630_c | −0.04 | 0.13 | 0.08 |
| 12. | scg6144s_c | −0.08 | 0.13 | 42.07 | scg90810_c | 0.61 | 0.21 | 8.01 |
| 13. | scg60510_c | −0.02 | 0.22 | 1.12 | scg9083s_c | 0.67 | 0.13 | 26.61 |
| 14. | scg60530_c | −0.04 | 0.20 | 3.24 | scg91030_c | −0.16 | 0.12 | 1.99 |
| 15. | scg6661s_c | 0.64 | 0.15 | 18.89 | scg91040_c | 0.17 | 0.15 | 1.26 |
| 16. | scg66620_c | −0.01 | 0.15 | 0.82 | scg91050_c | −0.22 | 0.12 | 3.25 |
| 17. | scg66630_c | 0.32 | 0.17 | 3.71 | scg9042s_c | 0.07 | 0.15 | 0.21 |
| 18. | scg6664s_c | 0.39 | 0.15 | 6.91 | scg9043s_c | 0.48 | 0.17 | 8.32 |
| 19. | scg6111s_c | 0.42 | 0.20 | 4.24 | scg9651s_c | −0.29 | 0.10 | 9.03 |
| 20. | scg6113s_c | −0.01 | 0.19 | 0.60 | scg96530_c | 0.28 | 0.12 | 5.43 |
| 21. | scg66040_c | 0.40 | 0.22 | 3.29 | scg90320_c | −0.10 | 0.12 | 0.66 |
| 22. | scg61310_c | −0.10 | 0.12 | 59.24 | scg90330_c | −0.42 | 0.12 | 11.44 |
| 23. | scg61330_c | −0.10 | 0.13 | 0.66 | scg9621s_c | −0.14 | 0.10 | 1.82 |
| 24. | scg6061s_c | 0.30 | 0.15 | 3.97 | scg96220_c | −0.05 | 0.13 | 0.15 |
| 25. | scg60620_c | 0.32 | 0.18 | 3.33 | scg91110_c | −0.29 | 0.12 | 5.37 |
| 26. | | | | | scg91120_c | −0.15 | 0.14 | 1.07 |
| 27. | | | | | scg91130_c | −0.20 | 0.13 | 2.15 |

*Note*. Δσ = Difference in item difficulty parameters between the longitudinal subsample in Grade 6 and 9 and the link sample (positive values indicate easier items in the link sample); $SE_{\Delta\sigma}$ = Pooled standard error; $F$ = Test statistic for the minimum effects hypothesis test (Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is $F_{.0154}$ (1, 3,568) = 83.32. A non−significant test indicates measurement invariance.

## 7.3 Scientific literacy scores

In the SUF manifest scientific literacy scores are provided in the form of two different WLEs (scg9_sc1 and scg9_sc1u), including their respective standard error (scg9_sc2 and scg9_sc2u).

For scg9_sc1u, person abilities were estimated using the linked item difficulty parameters. As a result, the WLE scores provided in scg9_sc1u can be used for longitudinal comparisons between grades 6 and 9. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points.

In contrast, the WLE scores in scg9_sc1 are not linked to the underlying reference scale of kindergarten. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions. The non-longitudinal WLE scores (scg9_sc1) were corrected for differences in the test position because, in grade 9, the science test was either presented as the first or the second test within the test battery.

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

# 8 References

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

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

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

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (2016). *NEPS Survey Paper No. 1, 2016 Linking the Data of the Competence Tests* (NEPS Survey Paper No. 1). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

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

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

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

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

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

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

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

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

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

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

# Appendix

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

```
Title G9 Science analysis, Partial Credit Model;
data filename.dat;
format id 1–7 responses 8–45;

labels << filename_with_labels.txt;

recode (0,1,2,3,4)     (0,0,0,1,2)          !item (2,17,18);
recode (0,1,2,3,4)     (0,0,1,2,3)          !item (4,13);

codes 0,1,2,3,4;
score (0,1)            (0,1)                !item (1,3,5-12,14-16,20-22,24-28,30,31,34-38);
score (0,1,2)          (0,0.5,1)            !item (2,17,18);
score (0,1,2,3)        (0,0.5,1,1.5)        !item (4,13);
score (0,1,2,3,4)      (0,0.5,1,1.5,2)      !item (19,23,29,32,33);

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

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

Appendix B: Assignment of the test items to each test version, the content and process-related components, and the contexts

| Items | Low Level | Medium Level | High Level | Component | Context |
|---|---|---|---|---|---|
| **scg90110_c** | X | X | X | KOS | Health |
| **scg9012s_c** | X | X | | KOS | Health |
| **scg90510_c** | X | X | X | KOS | Environment |
| **scg9052s_c** | X | X | | KOS | Environment |
| **scg90920_c** | X | X | X | KOS | Environment |
| **scg90930_c** | X | | | KOS | Environment |
| **scg96120_c** | X | X | X | KAS | Health |
| **scg96410_c** | X | | | KAS | Technology |
| **scg96420_c** | X | X | X | KAS | Technology |
| **scg9061s_c** | X | X | X | KOS | Health |
| **scg90630_c** | X | X | X | KOS | Health |
| **scg90810_c** | X | X | | KOS | Technology |
| **scg9083s_c** | X | X | | KOS | Technology |
| **scg91030_c** | X | X | X | KOS | Technology |
| **scg91040_c** | X | | | KOS | Technology |
| **scg91050_c** | X | X | X | KOS | Technology |
| **scg9042s_c** | X | | | KOS | Environment |
| **scg9043s_c** | X | | | KOS | Environment |
| **scg9651s_c** | X | X | X | KAS | Environment |
| **scg96530_c** | X | X | X | KAS | Environment |
| **scg90320_c** | X | X | X | KOS | Technology |
| **scg90330_c** | X | X | X | KOS | Technology |
| **scg9621s_c** | X | X | | KAS | Environment |
| **scg96220_c** | X | X | X | KAS | Environment |
| **scg91110_c** | X | X | X | KOS | Technology |
| **scg91120_c** | X | X | X | KAS | Technology |
| **scg91130_c** | X | X | X | KOS | Environment |
| **scg97410_c** | | X | X | KOS | Technology |
| **scg9771s_c** | | X | X | KAS | Health |
| **scgb6320_c** | | X | X | KAS | Technology |
| **scg98910_c** | | X | X | KOS | Environment |
| **scg9751s_c** | | X | X | KOS | Environment |
| **scg9752s_c** | | X | X | KOS | Environment |
| **scg98010_c** | | | X | KOS | Technology |
| **scg97910_c** | | | X | KOS | Health |
| **scg98210_c** | | | X | KOS | Technology |
| **scg98310_c** | | | X | KOS | Technology |
| **scgb6210_c** | | | X | KAS | Environment |

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