NEPS SURVEY PAPERS

Luise Fischer, Sven Rieger, Nicolas Hübner, Kerstin Oschatz, Jochen Kramer, and Wolfgang Wagner

NEPS TECHNICAL REPORT FOR SCIENTIFIC THINKING: SCALING RESULTS OF STARTING COHORTS 3 (WAVE 9) AND 4 (WAVE 7) IN 12TH GRADE

NEPS Survey Paper No. 66
Bamberg, January 2020

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 Scientific Thinking: Scaling Results of Starting Cohorts 3 (Wave 9) and 4 (Wave 7) in 12th Grade

Luise Fischer[1], Sven Rieger[2], Nicolas Hübner[2], Kerstin Oschatz[2], Jochen Kramer[2] & Wolfgang Wagner[2]

[1]*Leibniz Institute for Educational Trajectories, Bamberg, Germany*

[2]*Hector Research Institute of Education Sciences and Psychology,*
University of Tübingen, Germany

E-mail address of lead author:

luise.fischer@lifbi.de

Bibliographic data:
Fischer, L., Rieger, S., Hübner, N., Oschatz, K., Kramer, J. & Wagner, W. (2020). *NEPS Technical Report for Scientific Thinking: Scaling Results for the Starting Cohorts 3 (Wave 9) and 4 (Wave 7) in 12th Grade* (NEPS Survey Paper No. 66). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP66:1.0

Acknowledgments:

Various parts of this report (e.g., regarding the analytic strategy) are reproduced *verbatim* from previous working papers to facilitate the understanding of the presented results. We would like to thank Eve Sarah Müller for her excellent support and valuable feedback.

NEPS Technical Report for Scientific Thinking: Scaling Results of Starting Cohorts 3 (Wave 9) and 4 (Wave 7) in 12th Grade

Abstract

The National Educational Panel Study (NEPS) investigates 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 the competence tests, a range of analyses based on item response theory (IRT) are performed. This paper describes the data and scaling procedures for the scientific thinking competence test that was administered in wave 9 of Starting Cohort 3 (Grade 5) and wave 7 of Starting Cohort 4 (Grade 9) to individuals attending 12th Grade. The items were specifically developed for young adults graduating from secondary school in Germany. The scientific thinking competence test contained 32 subtasks referring to 5 short vignettes of controversial science claims. These items were designed to measure a common dimension reflecting the competence to engage in metascientific reflection. The test was finished by 5,668 individuals (55% women) from Starting Cohort 3 (*N* = 1,775) and Starting Cohort 4 (*N* = 3,893). The participants' responses were scaled using the 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 an acceptable reliability and that the items fitted the model in a satisfactory way. The subtasks covered a wide range of the ability distribution of the participants and the variance implies good differentiation between respondents. Furthermore, test fairness could be confirmed for different subgroups. Analyses of missing values revealed no shortcomings of the test. A limitation of the test is related to the dimensionality analyses based on the five vignettes. Overall, the scientific thinking test had satisfactory psychometric properties that allowed for an estimation of reliable competence scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the R code for estimating the manifest ability score.

Keywords

item response theory, scaling, scientific thinking competence, scientific use file

Contents

1 Introduction

Within the National Educational Panel Study (NEPS) different competences are measured coherently across the life span. These include, among others, reading competence, mathematical competence, scientific literacy, information and communication technologies literacy, metacognition, vocabulary, and domain general cognitive functioning. An overview of the competences measured in the NEPS is given by Weinert and colleagues (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 thinking competence test that was administered in wave 9 of Starting Cohort 3 (Grade 5) and wave 7 of Starting Cohort 4 (Grade 9) to individuals attending 12th Grade. First, the main concepts of the scientific thinking test and the test design are introduced. Then, the competence data of the two starting cohorts (SC) 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 scientific use file 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 Thinking Competence

2.1 Conceptual Framework

The framework and item development for the scientific thinking competence test are described in Oschatz, Kramer, and Wagner (2018). In the following, there will be a brief description of specific aspects of the scientific thinking competence test that are necessary for understanding the scaling results presented in this paper.

The aim of the scientific thinking competence test is to measure “Wissenschaftspropädeutik”, one of the educational aims for higher secondary school in Germany. “Wissenschaftspropädeutik” could be translated as the preparation of young individuals for a sophisticated handling of science itself as well as the preparation for a lifetime of learning and operating in a society, deeply dependent on science and its outcomes (Huber, 2000). In a theoretical approach by Huber (1997) the construct has been divided up into three subsequent tiers. The third and most central dimension entails “thinking and reflecting about science” from a meta-scientific perspective. It comprises the critical reflection of scientific knowledge in regard to its formation, development, potential and limitations. This dimension was chosen as the core of the NEPS assessment of scientific thinking competence as “metascientific reflection”.

The popular three tiers of Huber (2000) were translated into a normative structural model with the dimensions by Müsche (2009). Her framework served as the basis for the construction of a test of a meta-scientific reflection competence in NEPS. It was used to differentiate five different aspects of meta-scientific reflection: (1) Contextualization of scientific ways of research and knowledge, conclusions and results, (2) reflection of scientific ideas and results with regard to their foundation, potential, the circumstances of their development and consequences, (3) evaluation of scientific processes of knowledge generation and potential using methodological knowledge, (4) evaluation of the validity and explanatory power, and importance of scientific conclusions, (5) contrasting and integration of inconsistent results or contradictory theoretical approaches. As aspects (1) to (4) often have to be applied during the "contrasting or integration of inconsistent results or contradictory theoretical approaches" this aspect was chosen as the corner stone of this one-dimensional test. Furthermore, students encounter conflicting scientific evidence in the media every day. Conflicting science claims (5) can therefore be conceived as a typical and authentic problem of everyday life.

The test construction was based on five vignettes regarding scientific controversies on problems of a wider interest for society. Each vignette describes a scientific debate in one of the disciplines chemistry, medical science, biology (2 controversies), or sport science. The vignette of approximately 300 to 400 words is followed by 5 to 7 items regarding central aspects of the controversy. Therefore, students have to read one text and subsequently answer multiple test items related to it. Across the five controversies, all aspects of meta-scientific reflection are addressed equally.

There is one type of response format on the scientific thinking test (complex multiple choice; CMC). For CMC tasks, a number of subtasks with two response options (correct/incorrect) are presented. Examples for CMC tasks are given in Pohl and Carstensen (2012).

Table 1

Content Areas of the Items on the Scientific Thinking test

Content area	**Number of Subtasks**
Chemistry	7
Medical science	7
Biology	12
Sport science	6
Total number of subtasks	32

The scientific thinking competence test that was administered in wave 9 of Starting Cohort 3 (Grade 5) and wave 7 of Starting Cohort 4 (Grade 9) to individuals attending 12th Grade included 5 items comprising 32 subtasks overall. Extensive preliminary analyses were conducted to evaluate the quality of these subtasks and items, resulting in a satisfactory fit for all subtasks and items. Therefore, all 32 subtasks and items were included in the final scaling procedure. Table 1 shows the distribution of items on the 4 domains and the response formats.

3 Data

3.1 Design of the Study

The studies in wave 9 of Starting Cohort 3 (Grade 5) and wave 7 of Starting Cohort 4 (Grade 9) assessed different competence domains including technological and information literacy, reading competence, mathematical competence, English as a foreign language as well as scientific thinking. In order to control for test position effects, the tests were administered to participants in different sequence. For each participant the scientific thinking test was either administered as the fourth or fifth test (i.e., after English as a foreign language). There was no multi-matrix design regarding the order of the items *within* a specific test. All students received the test items in the same order. A detailed description of the study design is available on the NEPS website (http://www.neps-data.de).

3.2 Sample

A total of 5,668[1] participants (55% women), graduating from secondary school in Germany, answered at least one item on the scientific thinking competence test and, thus, were used for the psychometric analyses (cf. Pohl & Carstensen, 2012). Of these, *N* = 1,775 (53% women) were from Starting Cohort 3 and *N* = 3,893 (56% women) were from Starting Cohort 4, all attending grade 12. Basic sociodemographic information of the two subsamples is summarized in Table 2.

Table 2.

Number of Participants and Basic Sociodemographic Information

	Starting Cohort 3	**Starting Cohort 4**
Sample size	1,775	3,893
Women	53%	56%
Migration background	11%	11%
Test position: "Scientific thinking" before "English as a foreign language"	50%	50%

[1] Note that these numbers may differ from those found in the SUF. This is due to still ongoing data protection and data cleansing issues.

4 Analyses

This section briefly describes the analyses that were conducted to evaluate the test. These included inspecting various types of missing responses, scaling the data, and examining the quality of the test.

4.1 Missing Responses

There are different types of missing responses in competence test data. These include missing responses due to a) omitted items, b) invalid responses, c) items that test takers did not reach, and, finally, f) multiple kinds of missing responses within CMC items that are not determined.

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 were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. When one subtask contained a missing response, the CMC item was coded as missing. If just one kind of missing response occurred, the item was coded according to the corresponding missing response. If the subtasks contained different kinds of missing responses, the item was labeled as a not-determinable missing response.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions, handling of different response formats). Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the persons were coping with the test. Missing responses per item were examined in order to evaluate how well each of the items functioned.

4.2 Scaling Model

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

The 32 dichotomous subtasks were aggregated to five CMC items. As such, 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 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 of the five CMC items, categories were collapsed (see Appendix A).

Scientific thinking competences were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in section 7 of the present report.

4.3 Checking the Quality of the Scale

The scientific thinking competence test was specifically constructed to be implemented in the NEPS (Oschatz, Kramer, & Wagner, 2028). In order to ensure appropriate psychometric properties, the quality of the test was examined in several 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 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 dichotomous subtasks of CMC items consisted of one correct response option and one distractor (i.e., incorrect response option). The quality of the distractors 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 distractors with correlations above .05 are viewed as problematic (Pohl & Carstensen, 2012).

After aggregating the subtasks to polytomous variables, the fit of the 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 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 thinking competence test should measure the same construct for all respondents. If some items favored certain subgroups (e.g., items were easier for males than for females, although being equally able), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., males and females) would be biased and, thus, unfair. For the present studies, test fairness was investigated for the variables test position, gender, migration background, the number of books at home (as a proxy for socioeconomic status; see Pohl & Carstensen, 2012, for a description of these variables). Moreover, measurement invariance was analysed between the starting cohorts. Differential item functioning (DIF) was examined 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 considerable and noteworthy of further investigation, differences between 0.4 and 0.6 as small but not severe, and differences smaller than 0.4 as negligible DIF. 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 thinking competence 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 dimensionality of the test was evaluated by a multidimensional analysis. The different subdimensions of the multidimensional model represented the five vignettes regarding scientific controversies on problems of a wider interest for society. The correlations among the dimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the test. Moreover, we examined whether the residuals of the one-dimensional model exhibited approximately zero-order correlations as indicated by Yen's (1984) Q_3. Because in case of locally independent items, the Q_3 statistic tends to be slightly negative, we report the corrected Q_3 that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) values of Q_3 falling below .20 indicate essential unidimensionality. Moreover, to account for possible dependencies between the subtasks within an item, a Rasch testlet model (Wang & Wilson, 2005) was specified.

4.4 Software

The item response models were estimated with the *TAM* package version 3.3-10 (Robitzsch, Kiefer, & Wu, 2019) in *R* version 3.6.1 (R Core Team, 2019).

5 Results

5.1 Missing Responses

5.1.1 Missing responses per person

The amount of missing values was very similar in Starting Cohorts 3 and 4. Almost none of the participants (i.e., about 0.3%) produced an invalid response. As displayed in Figure 1, most respondents (i.e., about 93%) did not skip any item, about 6% omitted one item and less than two percent omitted more than one item.

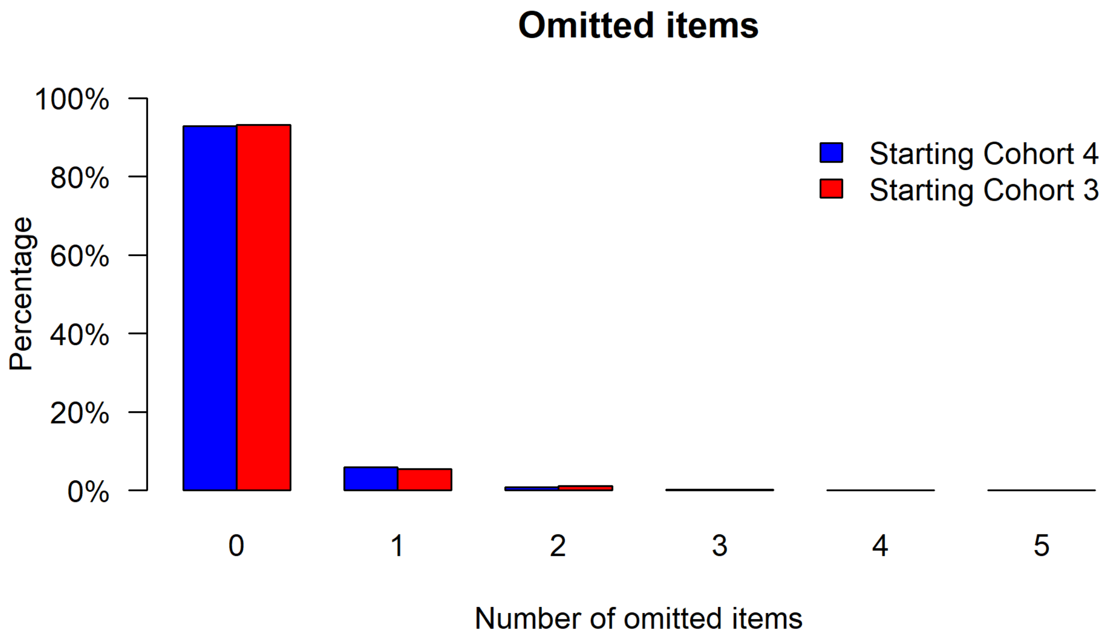

Figure 1. Number of omitted items by starting cohort.

Another source of missing responses are items that were not reached by the respondents because they ran out of time; these are all missing responses after the last valid response. About 99% of the respondents of Starting Cohorts 3 and 4 finished the entire test (see Figure 2). Thus, testing time did not seem to be an issue for the respondents.

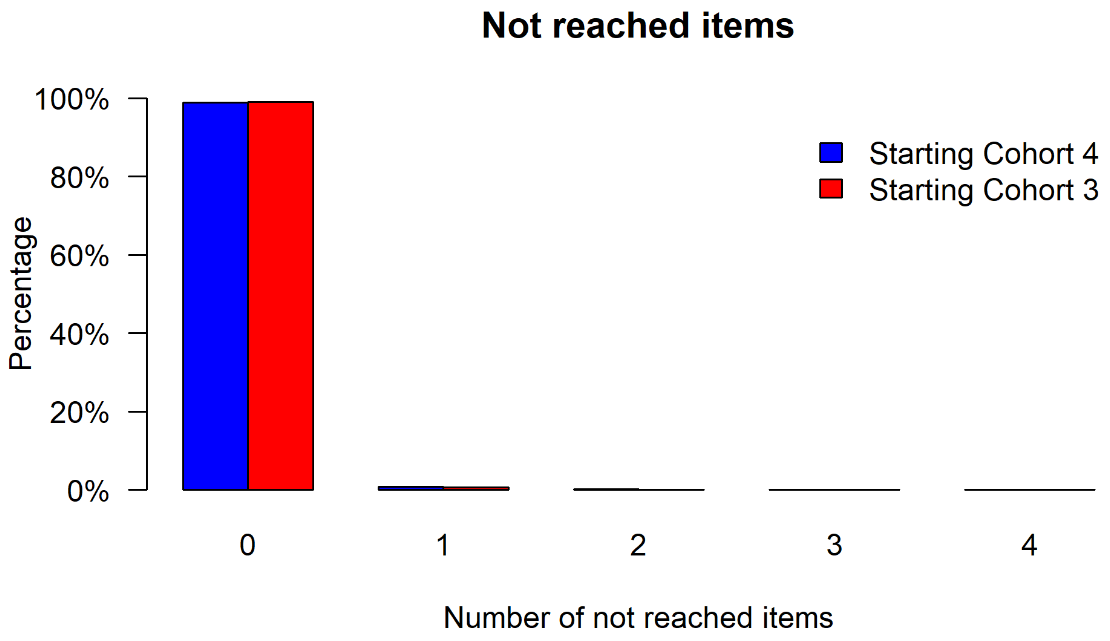

Figure 2. Number of not-reached items by starting cohort.

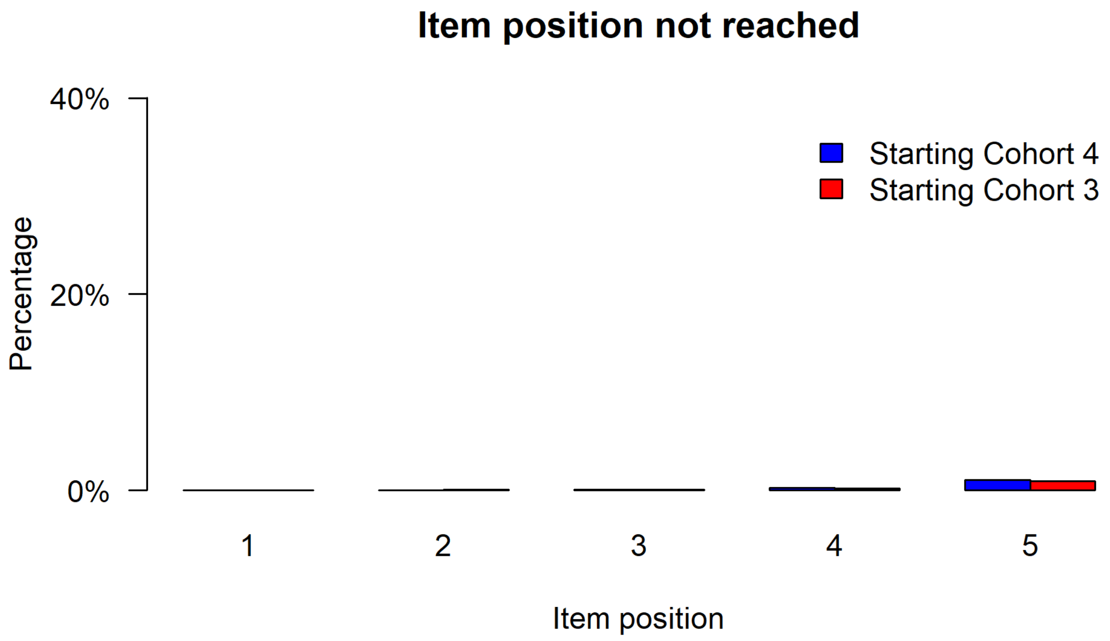

Figure 3. Item position not reached by starting cohort. Note that the scale on the x-axis was adapted (i.e., the scale was cut off at 40%).

With an item's progressing position in the test, the number of persons that did not reach the item rose in both starting cohorts (see Figure 3). However, only about 1% of the respondents did not reach the last item.

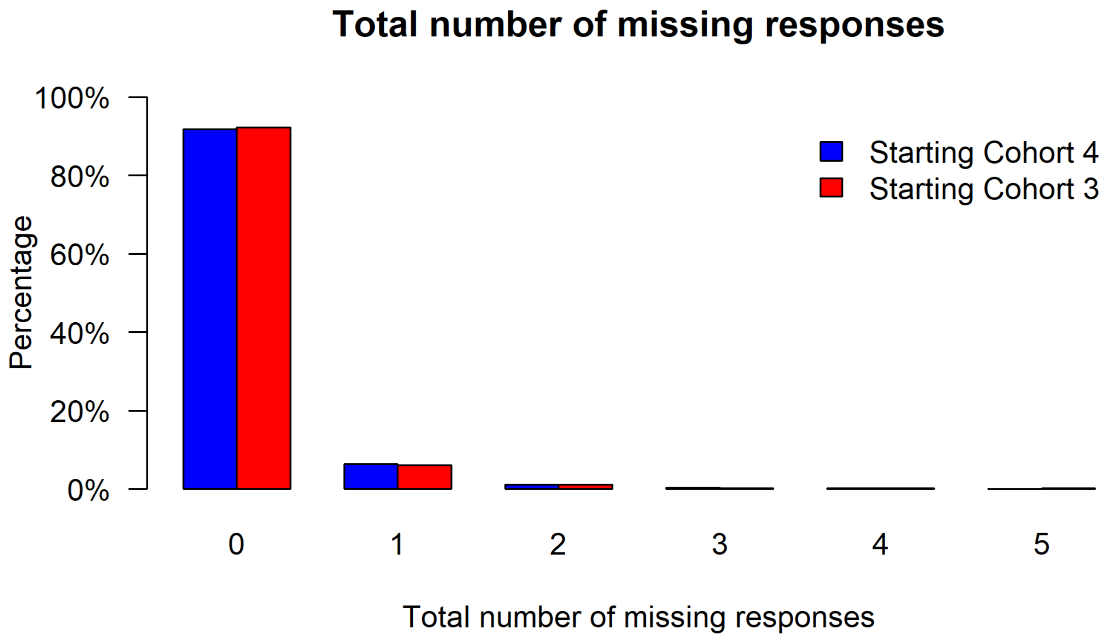

Figure 4. Total number of missing responses by starting cohort.

The total number of missing responses, aggregated over omitted, not-reached, not valid, and not-determinable missing responses per person, is illustrated in Figure 4. Respondents of Starting Cohort 3 had *M* = 1.57 (*SD* = 2.52) and respondents of the Starting Cohort 4 had *M* = 1.64 (*SD* = 2.73) missing responses. About 92% of the test takers of both starting cohorts had no missing response at all and only about 2% had more than one missing response.

5.1.2 Missing responses per item

Table 3 provides information on the occurrence of different kinds of missing responses per item. Overall, the number of missing values per item was very low and, thus, negligible. A maximum of 1.10% (*Mdn* = 0.17%) of the participants failed to reach items due to time constraints. The number of omitted, invalid and not-determinable responses varied across items between 0.84% and 3.15% (*Mdn* = 1.61%), 0.00% and 0.17% (*Mdn* = 0.08%) as well as 0.00% and 0.18% (Mdn = 0.02%), respectively.

Table 3

Percentage of Missing Values by Item and by Starting Cohort

		Starting Cohort 3					*Starting Cohort 4*				
Item	**Position**	***N***	**NR**	**OM**	**NV**	**ND**	***N***	**NR**	**OM**	**NV**	**ND**
stg12nhs_c	1	1747	0.06	1.52	0.17	0.00	3843	0.00	1.31	0.08	0.00
stg12egs_c	2	1739	0.17	1.91	0.11	0.00	3827	0.00	1.69	0.08	0.03
stg12mts_c	3	1751	0.17	1.35	0.00	0.00	3820	0.08	1.82	0.05	0.03
stg12cws_c	4	1715	0.28	3.15	0.11	0.00	3769	0.26	2.93	0.08	0.03
Stg12pds_c	5	1743	1.01	0.84	0.06	0.06	3799	1.10	1.15	0.08	0.18

Note. Position = Item position within test, *N* = Number of valid responses, NR = Percentage of respondents that did not reach the item, OM = Percentage of respondents that omitted the item, NV = Percentage of respondents who gave an invalid answer, ND = Percentage of not-determinable missings due to different kinds of missing responses contained in the subtasks. Item names refer to Starting Cohort 4; the corresponding variable names for Starting Cohort 3 are given in Appendix B.

5.2 Parameter Estimates

5.2.1 Item parameters

The data of Starting Cohorts 3 and 4 was analyzed concurrently. The estimated location parameters are given in Table 4, whereas the respective step parameters are summarized in Table 5. The location and step parameters were estimated by constraining the mean of the ability distribution to be zero. The estimated location parameters ranged from -0.34 (item stg12cws_c) to 0.12 (item stg12egs_c) with an average difficulty of -0.19 (*Mdn* = -0.24). Overall, the location parameters were distributed in a rather narrow section around the samples' mean. Due to the large sample size the standard errors (*SE*) of the estimated item location parameters (column 3 in Table 4) were rather small (all *SE*s = 0.01).

Table 4

Item Parameters

Item	**location parameter**	***SE***	**WMNSQ**	***t***	r_{it}	**Discr.**	Q_3
stg12nhs_c	-0.24	0.01	0.99	-0.71	0.38	0.55	0.04
stg12egs_c	0.12	0.01	0.99	-0.39	0.39	0.52	0.04
stg12mts_c	-0.19	0.01	1.07	3.79	0.29	0.32	0.05
stg12cws_c	-0.34	0.01	0.95	-2.57	0.45	0.74	0.06
Stg12pds_c	-0.30	0.01	1.00	0.01	0.36	0.48	0.02

Note. *SE* = Standard error of location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, r_{it} = Corrected item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model, Q_3 = Average absolute residual correlation for item (Yen, 1983). Estimated parameters are based on *N* = 5,674 (Starting Cohorts 3 and 4). The item-total correlation corresponds to the product-moment correlation between the corresponding categories and the total score. Item names refer to Starting Cohort 4; the corresponding variable names for Starting Cohort 3 are given in Appendix B.

Table 5

Step Parameters (with Standard Errors)

Item	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**
stg12nhs_c	-0.80 (0.03)	-0.38 (0.03)	0.30 (0.03)	0.88	
stg12egs_c	-0.83 (0.04)	-0.64 (0.03)	-0.02 (0.03)	0.36 (0.04)	1.12
stg12mts_c	-1.08 (0.04)	-0.66 (0.03)	0.01 (0.03)	0.67 (0.04)	1.06
stg12cws_c	-0.51 (0.03)	-0.45 (0.03)	-0.08 (0.03)	0.03 (0.03)	1.01
stg12pds_c	-0.56 (0.03)	-0.43 (0.03)	0.38 (0.03)	0.61	

Note. The last step parameter for each item is not estimated and has, thus, no standard error because it is a constrained parameter for model identification. Estimated parameters are based on *N* = 5,668 (Starting Cohorts 3 and 4). Item names refer to Starting Cohort 4; the corresponding variable names for Starting Cohort3 is given in Appendix B.

5.2.2 Test targeting and reliability

Test targeting focuses on comparing the item location parameters with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. Because all items in the scientific thinking test were polytomous, we calculated Thurstonian thresholds for each response category (Wu, Adams, Wilson, & Haldane, 2007). These indicate the location at the latent dimension at which the probability of achieving a score above the respective threshold is 50%. Thus, it is similar to the item difficulties of dichotomous items. In Figure 5, the category thresholds of the scientific thinking 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 category thresholds. The respective thresholds ranged from -3.40 (stg12mts_c) to 3.18 (stg12egs_c) and, thus, spanned a rather broad range. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 1.00, which implies good differentiation between respondents. The reliability of the test (EAP/PV reliability =.62, WLE reliability = .58) was acceptable. The mean of the item threshold distribution was about 0.36 logits below the mean person ability distribution. Thus, although the items covered a wide range of the ability distribution, the items were slightly too easy. As a consequence, person ability in medium- and low-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors of measurement.

5.3 Quality of the Test

5.3.1 Fit of the subtasks of complex multiple choice items

Before the subtasks of the CMC items were aggregated and analyzed via a partial credit model, the fit of the 32 subtasks was checked by analyzing the single subtasks in a Rasch model. The probability of a correct response ranged from 31% to 98% across all subtasks (*M* = 68%). Thus, the range of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.94 to 1.09, the respective *t*-value from -7.13 to 8.75, and there were no noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. Due to the good model fit of the subtasks, their aggregation to polytomous variables seemed justified.

5.3.2 Item fit

The evaluation of the item fit was performed based on the final scaling model, the partial credit model, using the polytomous CMC items. Altogether, item fit can be considered good (see Table 4). Values of the WMNSQ ranged from 0.95 (item stg12cws_c) to 1.07 (item stg12mts_c). None of the items exhibited a *t*-value of the WMNSQ greater than 8. Thus, there was no indication of severe item over- or underfit. Product-moment correlations between the corresponding categories and the total score ranged from .29 (item stg12mts_c) to .45 (item stg12cws_c) and had a median of .38. All item characteristic curves showed a good fit of the items.

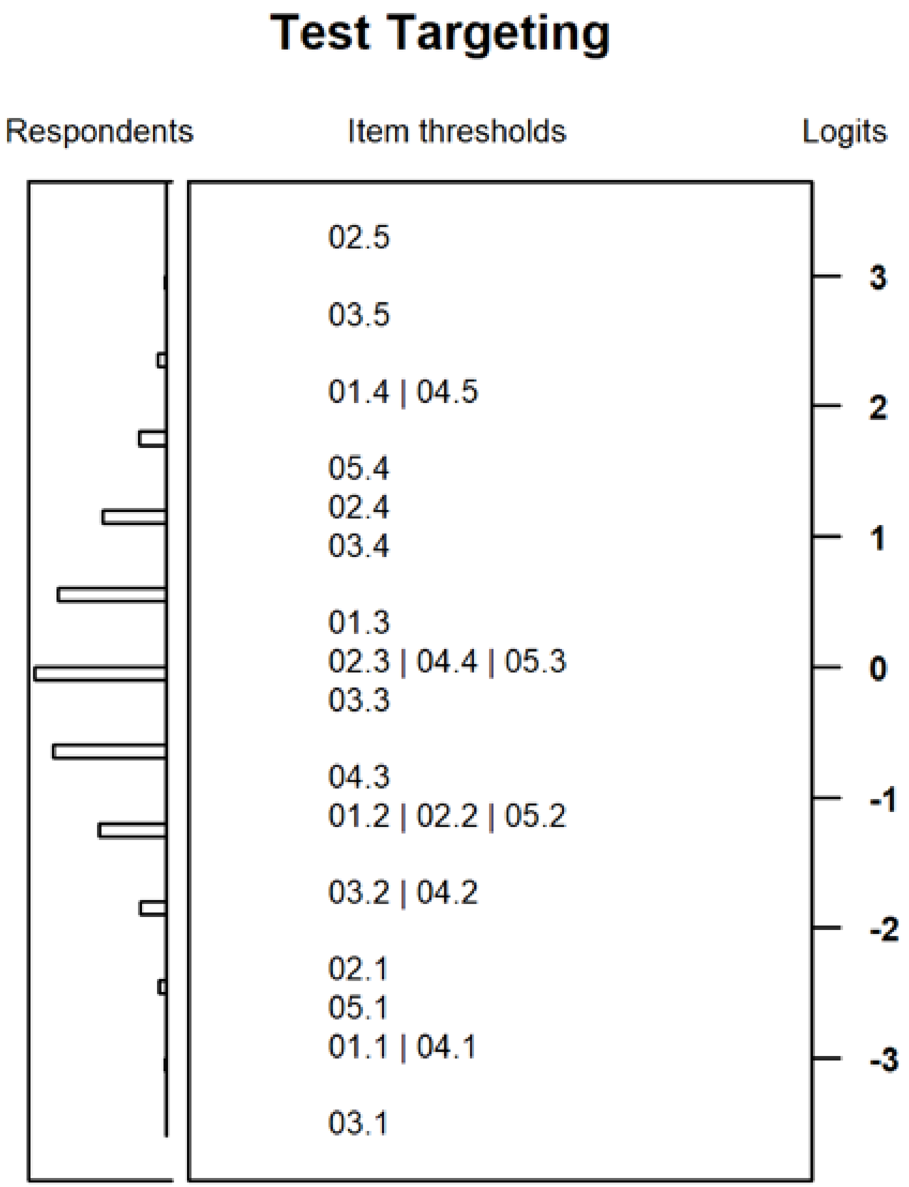

Figure 5. The distribution of person ability in the sample is given on the left-hand side of the graph. The category thresholds of the items are given on the right-hand side of the graph. Each number represents one threshold with the first part (before the dot) corresponding to the sequential position in Table 3 and the second part indicating the threshold.

5.3.3 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 sex, the number of books at home (as a proxy for socioeconomic status), migration background, and test position (see Pohl & Carstensen, 2012, for a description of these variables). In addition, the effect of the two starting cohorts was also studied. The differences between the estimated item location parameters in the various groups are summarized in Table 6. For example, the column “Male vs. female” reports the differences in item location parameters between men and women; a positive value would indicate that the test was more difficult for males, whereas a negative value would highlight a lower difficulty for males as opposed to females. 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 7).

Table 6

Differential Item Functioning

Item	Sex	Books	Migration	Position	Starting cohort
	male vs. female	≤ 100 vs. > 100	without vs. with	fourth vs. fifth	SC 4 vs. SC 3
stg12nhs_c	0.02 (0.04)	0.07 (0.15)	-0.02 (-0.05)	-0.03 (-0.06)	-0.03 (-0.07)
stg12egs_c	-0.08 (-0.17)	0.02 (0.04)	0.01 (0.02)	0.03 (0.06)	0.01 (0.02)
stg12mts_c	-0.13 (-0.27)	-0.11 (-0.22)	0.09 (0.19)	-0.03 (-0.06)	0.01 (0.02)
stg12cws_c	0.04 (0.07)	0.00 (0.00)	-0.02 (-0.02)	0.06 (0.13)	-0.05 (-0.09)
stg12pds_c	0.16 (0.32)	0.02 (0.04)	-0.07 (-0.13)	-0.04 (-0.07)	0.06 (0.12)
Main effect (DIF model)	0.01 (0.01)	-0.28 (-0.58)	0.31 (0.63)	-0.01 (-0.02)	-0.00 (-0.00)
Main effect (Main effect model)	0.01 (0.03)	-0.28 (-0.59)	0.31 (0.62)	-0.01 (-0.02)	-0.00 (-0.00)

Note. Raw differences between item location parameters with standardized differences (Cohen's *d*) in parentheses. No absolute standardized difference is significantly, p < .05, greater than 0.40 (see Fischer et al., 2016). Item names refer to Starting Cohort 4; the corresponding variable names for Starting Cohort 3 are given in Appendix B.

Sex: The sample included 2,561 (45%) males and 3,113 (55%) females. On average, male participants had a comparable estimated scientific thinking ability to females (main effect = 0.01 logits, Cohen's d = 0.01). There was no considerable DIF comparing male and female participants (highest DIF = 0.16 logits for item stg12pds_c). An overall test for DIF (see Table 7) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). However, model comparisons using Akaike's (1974) information criterion (AIC) and the Bayesian information criterion (BIC; Schwarz, 1978) both favored the model estimating DIF. Nevertheless, the deviation was small in both cases. Thus, overall, there was no pronounced DIF regarding the gender of the participants.

Books: The number of books at home was used as a proxy for socioeconomic status. There were 1,277 (23%) test takers with 0 to 100 books at home, 4,247 (75%) test takers with more than 100 books at home, and 148 (3%) test takers without a valid response. There were considerable average differences between the two groups. Participants with 100 or less books at home performed on average 0.28 logits (Cohen's d = -0.58) lower in scientific thinking than participants with more than 100 books. There was no considerable DIF comparing participants with many or fewer books (highest |DIF| = 0.11 logits for item stg12mts_c). Whereas the AIC favored the model estimating DIF, the BIC favored the main effects model (Table 7).

Migration background: There were 5,044 participants (89%) with no migration background, 602 respondents (11%) with a migration background and 28 (0.5%) test takers without respective information. In comparison to participants with migration background, participants without migration background had, on average, a higher scientific thinking ability (main effect = 0.31 logits, Cohen's d = 0.63). There was no noteworthy item DIF due to migration background; differences in estimated location parameters did not exceed 0.4 logits. The overall test for DIF using the BIC favored the main effects model, while the AIC favored the model estimating DIF.

Position: The scientific thinking competence test was administered in two different positions (see section 3 for the design of the study). A sample of 2,846 (50%) persons received the scientific thinking test on fourth position (before the English as a foreign language test) and 2,827 (50%) respondents took the scientific thinking test after having completed the English as a foreign language test. Differential item functioning of the position of the test may, for example, occur if there are differential fatigue effects for certain items. The results show negligible effects of item position[2]. In this study, persons who received the scientific thinking test first performed on average -0.01 logits (Cohen's d = -0.02) worse than respondents who received the scientific thinking test second. There was no DIF due to the position of the test in the booklet. The largest difference in difficulty between the two design groups was 0.06 logits (item stg12cws_c). The overall test for DIF using the BIC favored the main effects model, while the AIC favored the model estimating DIF.

Starting cohort: The scientific thinking test was administered in Starting Cohorts 3 and 4. To ensure test fairness and comparable person ability estimates among the starting cohorts we examined potential DIF. There were 3,897 participants (67%) in Starting Cohort 4 and 1,777 participants (31%) in Starting Cohort 3. On average, there was no difference in mean ability of scientific thinking (main effect = -0.00 logits, Cohen's d = -0.00) among participants of Starting Cohorts 3 and 4. Moreover, there was no noteworthy item DIF due to starting co-

[2] Note that this main effect does not indicate a threat to measurement invariance. Instead, it may be an indication of fatigue effects that are similar for all items.

hort association; differences in estimated location parameters did not exceed 0.4 logits. The overall test for DIF using the BIC favored the main effects model, while the AIC favored the model estimating DIF.

Table 7

Comparison of Models with and without DIF

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	main effect	5,668	88,331	25	88,381	88,547
	DIF	5,668	88,238	29	88,296	88,489
Books	main effect	5,519	85,807	25	85,857	86,022
	DIF	5,519	85,783	29	85,841	86,033
Migration	main effect	5,640	87,775	25	87,825	87,991
	DIF	5,640	87,765	29	87,823	88,015
Position	main effect	5,667	88,314	25	88,364	88,530
	DIF	5,667	88,300	29	88,358	88,551
Starting cohort	main effect	5,668	88,332	25	88,382	88,548
	DIF	5,668	88,321	29	88,379	88,572

5.3.4 Rasch homogeneity

An essential assumption of the Rasch (1960) model 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 were all lower than expected by the GPCM, ranging from 0.32 (item stg12mts_c) to .74 (item stg12cws_c). The median discrimination parameter fell at *Mdn* = 0.52. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 88,250, BIC = 88,436) as compared to the PCM model (AIC = 88,380, BIC = 88,540). Despite the empirical preference for the GPCM, the PCM more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012; and 2013 for a discussion of this issue). For this reason, the PCM was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

5.3.5 Unidimensionality

The unidimensionality of the test was investigated by specifying a multidimensional model and comparing it to a unidimensional model. In the multidimensional model, the 32 subtasks loaded each on one of five dimensions, representing the five vignettes of controversial science claims. The multidimensional model was estimated using Quasi Monte Carlo method with 10,000 nodes. The estimated variances and correlations between the five dimensions representing the five items are reported in Table 8. The correlations among the five dimensions were moderate and ranged from .43 to .71, and, thus, deviated from a perfect correlation (i.e., they were lower than *r* = .95, see Carstensen, 2013). According to model fit indices, the five-dimensional Rasch model fitted the data better (AIC = 194,231, BIC = 194,543, number of parameters = 47) than the unidimensional Rasch model (AIC = 195,469, BIC = 195,689, number of parameters = 33).

Table 8

Results of Five-Dimensional Scaling

Item	**Dim 1**	**Dim 2**	**Dim 3**	**Dim 4**	**Dim 5**
stg12nhs_c (Dim 1)	(0.72)				
stg12egs_c (Dim 2)	0.67	(0.60)			
stg12mts_c (Dim 3)	0.46	0.43	(0.82)		
stg12cws_c (Dim 4)	0.71	0.71	0.46	(0.97)	
stg12pds_c (Dim 5)	0.54	0.57	0.44	0.63	(0.90)

Note. Variances of the dimensions are given in the diagonal and correlations are presented in the off-diagonal. Item names refer to Starting Cohort 4; the corresponding variable names for Starting Cohort 3 are given in Appendix B.

Additionally, a testlet model (Wang & Wilson, 2005) was estimated using the dichotomous subtasks. As such, all items were modelled to load on a general factor while simultaneously accounting for testlet-specific effects. The testlet model was estimated using Quasi Monte Carlo method with 10,000 nodes. The estimated variances and EAP reliabilities of each of the six dimensions are reported in Table 9. According to model fit indices, the testlet model fitted the data better (AIC = 194,359, BIC = 194,612, number of parameters = 38) than the unidimensional Rasch model and a little worse than the five-dimensional model (see above). However, a unidimensional scientific thinking competence score was estimated based on a unidimensional PCM (AIC = 88,380, BIC = 88,540, number of parameters = 24) that similarly accounts for item-specific effects.

Table 9

Results of the Testlet Model

Subtasks of …	**Variance (EAP reliability)**
All five items (Dim 1) (32 subtasks)	0.45 (0.63)
stg12nhs_c (Dim 2) (5 subtasks)	0.26 (0.17)
stg12egs_c (Dim 3) (7 subtasks)	0.19 (0.17)
stg12mts_c (Dim 4) (6 subtasks)	0.54 (0.31)
stg12cws_c (Dim 5) (7 subtasks)	0.35 (0.24)
stg12pds_c (Dim 6) (7 subtasks)	0.45 (0.26)

Note. Dimension 1 includes the subtasks of all items while the dimensions 2 to 6 include the subtasks of one item each. Item names refer to Starting Cohort 4; the corresponding variable names for Starting Cohort 3 are given in Appendix B.

6 Discussion

The analyses in the previous sections reported information on the quality of the scientific thinking test that was administered in Starting Cohorts 3 and 4 to participants attending grade 12 in secondary school in Germany. Furthermore, the estimation of the respective scientific thinking competence scores was described. Different kinds of missing responses were examined, item fit statistics and item characteristic curves were evaluated, and item discriminations were investigated. Further quality inspections were conducted by examining differential item functioning and testing Rasch-homogeneity. Various criteria indicated a good fit of the items and measurement invariance across various subgroups. The number of missing responses was low. The test had a satisfactory reliability and distinguished well between test takers. However, the test was slightly better targeted at mediocre- and low-performing students and covered the high ability spectrum less well. As a consequence, ability estimates will be precise for low-performing respondents but less precise for high performing respondents. Furthermore, some degree of multidimensionality is present for the five items. In summary, the test had acceptable psychometric properties that allowed the estimation of a unidimensional scientific thinking competence score.

7 Data in the Scientific Use File

7.1 Naming conventions

The data in the Scientific Use File contains 5 CMC variables in Starting Cohort 3 (Wave 9) and Starting Cohort 4 (Wave 7). CMC items are marked with a ‘s_c’ at the end of the variable name. For further details on the naming conventions of the variables see Fuß and colleagues (2019).

7.2 Linking of competence scores

In Starting Cohorts 3 and 4 the participants attending grade 12 in secondary school were administered an identical form of the scientific thinking test under standardized conditions. As measurement invariance among Starting Cohorts 3 and 4 was verified, a concurrent calibration of the two data sets seemed justified. As such, the data of the two starting cohorts were placed on a common scale. Consequently, the competence scores derived in Starting Cohorts 3 and 4 are directly comparable.

7.3 Scientific thinking competence scores

In the SUF manifest scientific thinking competence scores are provided in the form of WLEs. In Starting Cohorts 3 and 4, the respective variable is called “stg12_sc1”, including its respective standard error, “stg12_sc2”. The estimated WLE scores were corrected for differences in the test position as the scientific thinking test was either presented as the first or the second test within the test battery (see page 7). To correct for differences in the test position, we added the main effect related to the test position (see Table 6) to the WLE scores of respondents that received the scientific thinking test after working on another test. The R code for estimating the WLE is provided in Appendix C. For persons who did not give enough valid responses (SC3: *N* = 3, SC4: *N* = 4) or where the test position was unknown (SC4: *N* = 1), no WLE was estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–722. https://doi.org/10.1007/BF02296272

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.

Huber, L. (1997). Fähigkeit zum Studieren- Bildung durch Wissenschaft. Zum Problem der Passung zwischen Gymnasialer Oberstufe und Hochschule [Ability to study - education through science. With respect to the fit between academic track schools and higher education institutions]. In E. Liebau, W. Mack & C. T. Scheilke (Eds.), *Das Gymnasium. Alltag, Reform, Geschichte, Theorie* (pp. 333-351). Weinheim, Germany: Juventa.

Huber, L. (2000). Wissenschaftspropädeutik, allgemeine Studierfähigkeit und ihre unterrichtliche Umsetzung in Grundkursen [Wissenschaftspropädeutik, general ability to study and the instructional implementation in basic courses]. In Hessisches Landesinstitut für Pädagogik (Eds.), *Bildung braucht guten Grund: Beiträge zur Reform der Grundkurse* (pp. 17–46). Wiesbaden, Germany: HeLP.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Müsche, H. (2009). Wissenschaftspropädeutik aus psychologischer Perspektive – Zur Dimensionierung und Konkretisierung eines bildungstheoretischen Konzeptes [Wissenschaftpropedeutik from a psychological perspective - Regarding the dimension and ascertainment of the theoretical based concept]. *TriOS, 4*, 61-109.

Oschatz, K., Kramer, J., & Wagner, W. (2018). *The Assessment of Scientific Thinking as Meta-Scientific Reflection.* Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical report – Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg: 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. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: Mesa Press.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Robitzsch, A., Kiefer, T., & Wu, M. (2019). *TAM: Test analysis modules. R package version 2.7-56*. Retrieved from https://CRAN.R-project.org/package=TAM

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

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126-149. https://doi.org/10.1177/0146621604271053

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450. https://doi.org/10.1007/BF02294627

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of competencies across the life span. In H. P. Blossfeld, H. G. Roßbach, & J. von Maurice & (Eds.), *Education as a lifelong process: The German National Education Panel Study (NEPS)* (pp. 67-86). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalized item response modelling software*. Camberwell, Australia: ACER Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145. https://doi.org/10.1177/014662168400800201

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

Appendix

Appendix A: Response Categories of all Five CMC Items

Table 10

Original and Collapsed Response Categories of CMC Items

item	Response category	0	1	2	3	4	5	6	7
stg12nhs_c	Original RC	0	1	2	3	4	5	-	-
	Collapsed RC	0	0	1	2	3	4	-	-
stg12egs_c	Original RC	0	1	2	3	4	5	6	7
	Collapsed RC	0	0	0	1	2	3	4	5
stg12mts_c	Original RC	0	1	2	3	4	5	6	-
	Collapsed RC	0	0	1	2	3	4	5	-
stg12cws_c	Original RC	0	1	2	3	4	5	6	7
	Collapsed RC	0	0	0	1	2	3	4	5
stg12pds_c	Original RC	0	1	2	3	4	5	6	7
	Collapsed RC	0	0	0	0	1	2	3	4

Note. RC = Response category. Response Categories were collapsed when a cell contained less than 200 individuals. In these cases, the lower categories were collapsed into one category. Item names refer to Starting Cohort 4; the corresponding variable names for Starting Cohort 3 are given in Appendix B.

Appendix B: Item Names in Starting Cohorts 3 and 4

Table 11

Item Names in Starting Cohorts 3 and 4

Position	Starting Cohort 3	Starting Cohort 4
1	stg12nhs_sc3g12_c	stg12nhs_c
2	stg12egs_sc3g12_c	stg12egs_c
3	stg12mts_sc3g12_c	stg12mts_c
4	stg12cws_sc3g12_c	stg12cws_c
5	stg12pds_sc3g12_c	stg12pds_c

Note. Repeatedly administered items retain their original names in all test administrations and, thus, reflect their very first application. Consequently, items that have been administered before are supplemented with a suffix that represents the present test application.

Appendix C: R Code for WLE Estimation

```
# Load packages
library(dplyr)
library(haven)
library(TAM)

# Load competence data of Starting Cohort 4
dat <- read_spss('SC4_xTargetCompetencies_D_10-0-0.sav')

# Items of the test
items <- paste0('stg12', c('nhs', 'egs', 'mts', 'cws', 'pds'), '_c')

# Select variables
dat <- select(dat, ID_t, tx80211_w7, one_of(items))

# collapse response categories with N < 200
dat$stg12nhs_c <- recode(as.numeric(dat$stg12nhs_c),
                         c('0' = 0, '1' = 0, '2' = 1,
                           '3' = 2, '4' = 3, '5' = 4))
dat$stg12egs_c <- recode(as.numeric(dat$stg12egs_c),
                         c('0' = 0, '1' = 0, '2' = 0,
                           '3' = 1, '4' = 2, '5' = 3,
                           '6' = 4, '7' = 5))
dat$stg12mts_c <- recode(as.numeric(dat$stg12mts_c),
                         c('0' = 0, '1' = 0, '2' = 1,
                           '3' = 2, '4' = 3, '5' = 4,
                           '6' = 5))
dat$stg12cws_c <- recode(as.numeric(dat$stg12cws_c),
                         c('0' = 0, '1' = 0, '2' = 0,
                           '3' = 1, '4' = 2, '5' = 3,
                           '6' = 4, '7' = 5))
dat$stg12pds_c <- recode(as.numeric(dat$stg12pds_c),
                         c('0' = 0, '1' = 0, '2' = 0,
                           '3' = 0, '4' = 1, '5' = 2,
                           '6' = 3, '7' = 4))

# Select respondents with valid position information
dat$pos <- recode(as.numeric(dat$tx80211_w7),
                  '296' = 0, '297' = 0, '300' = 0, '301' = 0, # 4th position
                  '298' = 1, '299' = 1, '302' = 1, '303' = 1, # 5th position
                  .default = NA_real_)
pos <- filter(dat, !is.na(pos)) %>% select(pos)
resp <- filter(dat, !is.na(pos)) %>% select(one_of(items))

# Estimate model with main effect of test position
frmA <- ~ 0 + item + item:step + pos
mod <- tam.mml.mfr(resp = resp, irtmodel = "PCM2", facets = pos,
                   formulaA = frmA, constraint = "cases")

# Estimate WLEs
wle <- tam.wle(mod, Msteps = 100)
```