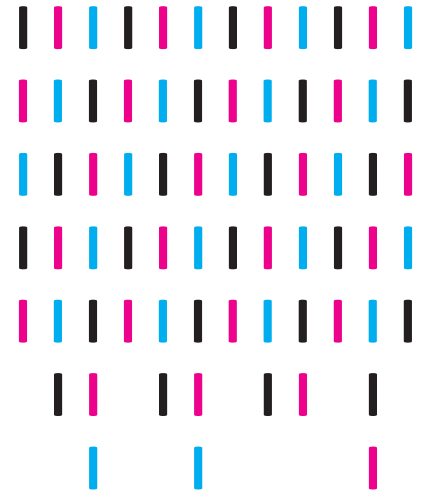




NEPS *SURVEY PAPERS*

Timo Gnambs

NEPS TECHNICAL REPORT FOR GENERAL KNOWLEDGE: SCALING RESULTS OF STARTING COHORT 8 FOR GRADE 6

Survey Papers of the German National Educational Panel Study (NEPS)

at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS *Survey Paper* series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

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

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

The NEPS *Survey Papers* are available at www.neps-data.de (see section "Publications") and at www.lifbi.de/publications.

Editor-in-Chief: Thomas Bäumer, LIfBi

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

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



This work is licensed under CC BY 4.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

NEPS Technical Report for General Knowledge: Scaling Results of Starting Cohort 8 for Grade 6

Timo Gnambs 

Leibniz Institute for Educational Trajectories

E-mail address of the lead author:

timo.gnambs@lifbi.de

Bibliographic data:

Gnambs, T. (2026). *NEPS technical report for general knowledge: scaling results of starting cohort 8 for grade 6* (NEPS Survey Paper No. 134). Leibniz Institute for Educational Trajectories. <https://doi.org/10.5157/NEPS:SP134:1.0>

NEPS Technical Report for General Knowledge: Scaling Results of Starting Cohort 8 for Grade 6

Abstract

The National Educational Panel Study (NEPS) examines the development of competencies across the lifespan. Therefore, the NEPS assesses different domains of competence in different age cohorts. To evaluate the quality of these tests, several analyses based on item response theory are conducted. This paper describes the data and scaling procedures for a test of general knowledge that was administered in Grade 6 of Starting Cohort 8. The general knowledge test consisted of 64 items representing different content domains. The test was administered to 4,313 participants (50% girls). A Rasch model was used to scale the data. Item fit statistics, differential item functioning, Rasch-homogeneity, test dimensionality, and local item independence were evaluated to ensure the quality of the test. The analyses showed that the items had good fit to the item response model. Although some items exhibited differential item functioning across different subgroups, approximate measurement invariance could be established. In addition, the test showed a high reliability and the different content domains supported an essentially unidimensional construct. Overall, the test had good psychometric properties that allowed the estimation of reliable general knowledge scores. In addition to the scaling results, this report also describes the data available in the scientific use file.

Keywords

item response theory, scaling, general knowledge, scientific use file

Contents

1	Introduction	4
2	Testing General Knowledge	4
2.1	Conceptual Framework	4
2.2	Study Design	6
3	Sample	6
4	Psychometric Analyses	6
4.1	Missing Responses	6
4.2	Scaling Model	6
4.3	Software	8
5	Results	8
5.1	Missing Responses	8
5.1.1	Missing Responses per Person	8
5.1.2	Missing Responses per Item	10
5.2	Parameter Estimates	12
5.2.1	Item Parameters	12
5.2.2	Test Targeting and Reliability	15
5.3	Quality of the test	16
5.3.1	Distractor Analyses	16
5.3.2	Item Fit	17
5.3.3	Differential Item Functioning	17
5.3.4	Rasch-homogeneity	23
5.3.5	Dimensionality	23
6	Discussion	25
7	Data in the Scientific Use Files	26
7.1	Naming Conventions	26
7.2	General Knowledge Scores	26
	References	27
	Appendix A: Content Areas and Domains	29

1. Introduction

The National Educational Panel Study (NEPS)¹ measures different competencies coherently across the lifespan. These include, among others, reading competence, mathematical competence, and domain-general cognitive functioning. An overview of the competencies measured in the NEPS is given by Weinert et al. (2019) and Fuß et al. (2025). Most of the administered competence tests are developed specifically for use in the NEPS and are therefore routinely evaluated using psychometric models based on item response theory (IRT, see Pohl & Carstensen, 2012).

This report presents the results of these analyses for a test of general knowledge that was administered in Grade 6 of Starting Cohort 8 (Grade 5). The following sections first introduce the main concepts of the administered test and the test design. Then, the data are described and the results of various psychometric analyses are reported that were conducted to estimate knowledge scores and to check the quality of the test. Finally, an overview of the data that are available for public use in the scientific use file (SUF) is given.

The analyses summarized in this report are based on the data available at some point prior to the public data release. Due to ongoing data protection and data cleansing issues, the data in the SUF may differ slightly from the data used for the analyses in this report. However, pronounced differences in the presented results are not expected for the final data available in the SUF.

2. Testing General Knowledge

2.1 Conceptual Framework

According to the Cattell-Horn-Carroll (CHC) model of intelligence (Schneider & McGrew, 2018), human cognitive abilities can be described by a hierarchical structure that classifies abilities into different strata. While a single general ability is at the apex of the hierarchy (Stratum III), several broad abilities such as fluid and crystallized intelligence, together with various additional abilities (e.g., memory, cognitive speed), make up Stratum II. In contrast to fluid intelligence, which involves basic reasoning and problem-solving abilities, crystallized intelligence includes acquired skills and knowledge and thus largely reflects learning, education, and acculturation.

Crystallized intelligence can be measured with items that capture the depth and breadth of the dominant knowledge in a culture (Horn & Noll, 1997). Accordingly, the Berlin test of fluid and crystallized intelligence (BEFKI, Schroeders et al., 2020), which was administered in the present study, measures crystallized intelligence with items on factual knowledge in 16 content areas, each with four items (see Table 1). The content areas can be further grouped into three major domains, that is, humanities, natural sciences, and social sciences. The items cover topics that are part of formal education, but also reflect shared cultural knowledge, that is, they describe what children and adolescents are expected to learn and know in a society. Consequently, the test items cover both formal and non-formal knowledge domains. The items do not refer to specialized knowledge, trivial content, or highly time-dependent information (e.g., current sports results or regional train schedules) but focus on general knowledge that is widely considered important and relevant in a society.

¹<https://www.neps-data.de>

Table 1*Content Areas of General Knowledge Test*

Domain	Content Area	Description
Humanities	Literature	Literary work and text types
	Art	Artistic techniques and famous artists
	Music	Musical instruments and musical styles
	Religion	Concepts of different world religions
	Philosophy	Famous philosophers and questions of philosophy
Natural sciences	Physics	Physical quantities and phenomena
	Chemistry	Organic and inorganic chemistry
	Biology	Biological processes, metabolism, and the human body
	Medicine	Human diseases and their treatment
	Geography	Knowledge about geography and geology
	Technology	Knowledge about technical devices and the Internet
Social sciences	History	Historical events and personalities
	Law	Legal institutions and laws in Germany
	Politics	Political institutions and elections in Germany
	Economics	Terms and participants in an economy
	Finances	Payment transactions and financial terms

Table 2*Number of Items by Domain*

Domain	Number of items
Humanities	20
Natural sciences	24
Social sciences	20
Total number of items	64

The general knowledge test consisted of 64 multiple-choice items. Each item was accompanied by four response options, one of which was a correct solution and the other three acted as distractors (i.e., they were incorrect). The distribution of the items across the three domains of the assessment framework is shown in Table 2. The allocation of each item to the content areas and domains is provided in Appendix A.

2.2 Study Design

The study was conducted in small groups at the respective schools of the students. The tests were presented on computer by trained test administrators. The study assessed different competence domains including general knowledge, digital competence (Kutscher, 2026), and information and communication technology literacy (Senkbeil, 2026). The general knowledge test was always presented third after the other tests. There was no multi-matrix design with respect to the order of the items within the test. All respondents received the test items in the same order. The testing time was limited to 20 minutes. A comprehensive description of the study design is available on the NEPS website².

3. Sample

A total of 4,313 participants (50% girls) were administered the general knowledge test. However, 9 participants had less than 16 valid responses to the test items. This suggests that these students did not follow the instructions correctly (see Schroeders et al., 2020). Therefore, these cases were excluded from the psychometric analyses and the analyses presented in this report are based on 4,304 respondents.

4. Psychometric Analyses

4.1 Missing Responses

Competence data contain different types of missing responses. These are missing responses due to a) omitted items and b) items that test takers did not reach. Omitted items occurred when respondents skipped some items. Due to time constraints or lack of motivation, not all participants completed the test within the allotted time. All missing responses after the last valid response were coded as not reached. Missing responses provide information about how well the test worked (e.g., time limits, understanding of instructions). Therefore, the occurrence of missing responses in the test was evaluated to get an idea of how well respondents coped with the test. Missing responses per item were examined to evaluate how well each of the items functioned.

4.2 Scaling Model

Item and person parameters were estimated using a Rasch (1960) model with Gauss-Hermite quadrature (21 nodes). In these analyses, missing responses were scored as incorrect to align with the scoring scheme proposed by the test authors (Schroeders et al., 2020). In order to ensure appropriate psychometric properties of the test, the quality of the test was examined in several analyses.

The items consisted of one correct response option and several distractors (i.e., incorrect response options). The quality of the distractors within the multiple-choice items was examined to evaluate whether the distractors were predominantly chosen by respondents with a lower ability rather than by those who gave a correct response. We therefore calculated the point-biserial correlation between selecting an incorrect response option for a given item and the

²<https://www.neps-data.de/Data-Center/Overview-and-Assistance/Plausible-Values>

total correct score for the remaining items. Negative correlations indicate good distractors, while correlations between .00 and .05 were considered acceptable and correlations above .05 were considered problematic distractors (Pohl & Carstensen, 2012).

The fit of the items to the Rasch (1960) model was evaluated using the WMNSQ statistic, the respective t -value, and the item characteristic curves (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (t -value > |6|) were considered to have a noticeable item misfit, and items with a WMNSQ > 1.20 (t -value > |8|) were considered to have a considerable item misfit and their performance was investigated further. Correlations of the item score with the corrected total score greater than .30 were considered as good, those greater than .20 as acceptable, and those less than .20 as problematic. The overall judgment of item fit was based on all fit indicators.

The general knowledge test should measure the same construct for all respondents. If some items favored certain subgroups (e.g., they were easier for boys than for girls), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., boys and girls) would be biased. In the present study, differential item functioning (DIF) was investigated for the variables gender, the number of books at home (as a proxy for cultural status), migrant background, and school type. DIF was examined using a multigroup item response model, in which the main effects of the subgroups as well as the differential effects of the subgroups on item difficulty were modeled. Based on experience 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 worthy of further investigation, differences between 0.4 and 0.6 as small but not severe, and differences less than 0.4 as negligible DIF. In addition, measurement invariance was examined by comparing the fit of a model that acknowledged DIF to a model that included only main effects and no DIF.

The general knowledge test was scaled using the Rasch (1960) model because it preserves the weighting of the different aspects of the framework as intended by the test developers (Schroeders et al., 2020). Nevertheless, Rasch-homogeneity is an assumption that may not hold for empirical data. To test the assumption of equal item discrimination parameters, a two-parametric item response model (2PL, Birnbaum, 1968) was also fitted to the data and compared to the Rasch model.

The dimensionality of the test was evaluated by examining the residuals of the Rasch model. Approximately zero-order correlations, as indicated by Yen's (1984) Q_3 , suggest unidimensionality. Because the Q_3 tends to be slightly negative in case of locally independent items, we report the corrected Q_3 , which has an expected value of 0. According to Yen (1993), values of Q_3 falling below .20 indicate essential unidimensionality. In addition, we estimated two multidimensional models using quasi-Monte Carlo integration (with 2,000 nodes) that specified different dimensions based on the construction rationale for the test (see Section 2.1), that is, a model with 16 dimensions representing the 16 content areas and a model with three dimensions based on the three domains (see Appendix A). The correlations between the dimensions and the differences in model fit between the unidimensional and multidimensional models were used to evaluate the unidimensionality of the test.

4.3 Software

The item response models were estimated with the *TAM* package (Robitzsch et al., 2024) in *R* (R Core Team, 2025).

5. Results

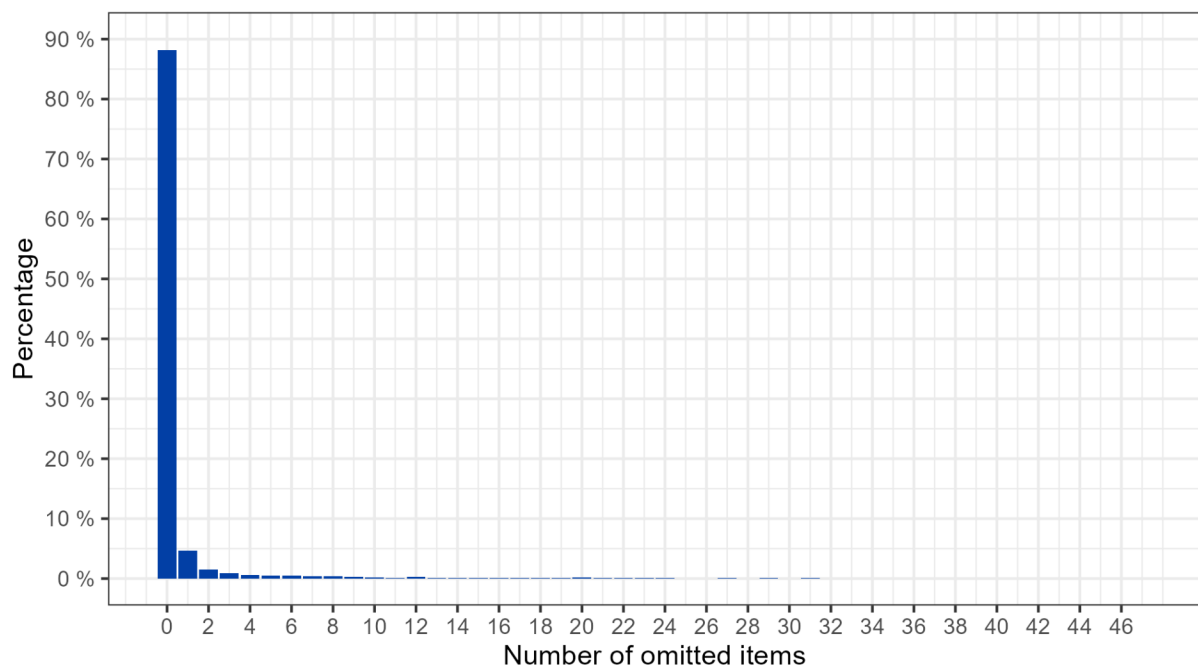
5.1 Missing Responses

5.1.1 Missing Responses per Person

The number of omitted responses per person is shown in Figure 1. The figure shows that only few items were omitted. Approximately 88% of respondents had no omitted items, while 5% exhibited a single omitted item. However, less than 4% of respondents had five or more omitted items.

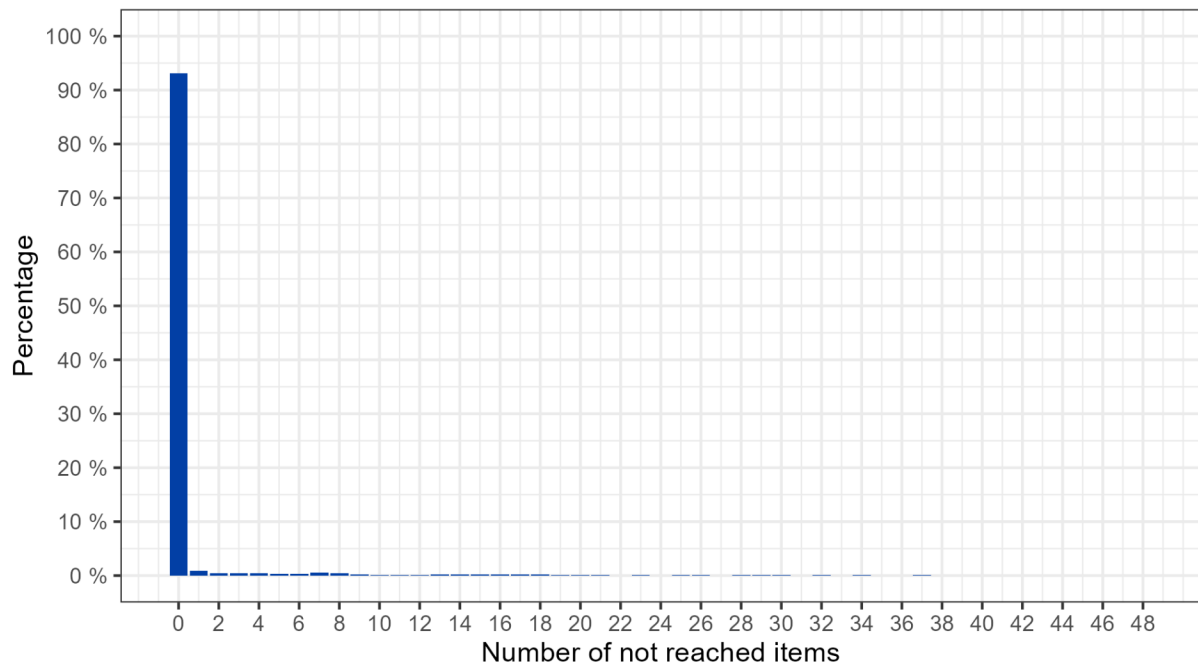
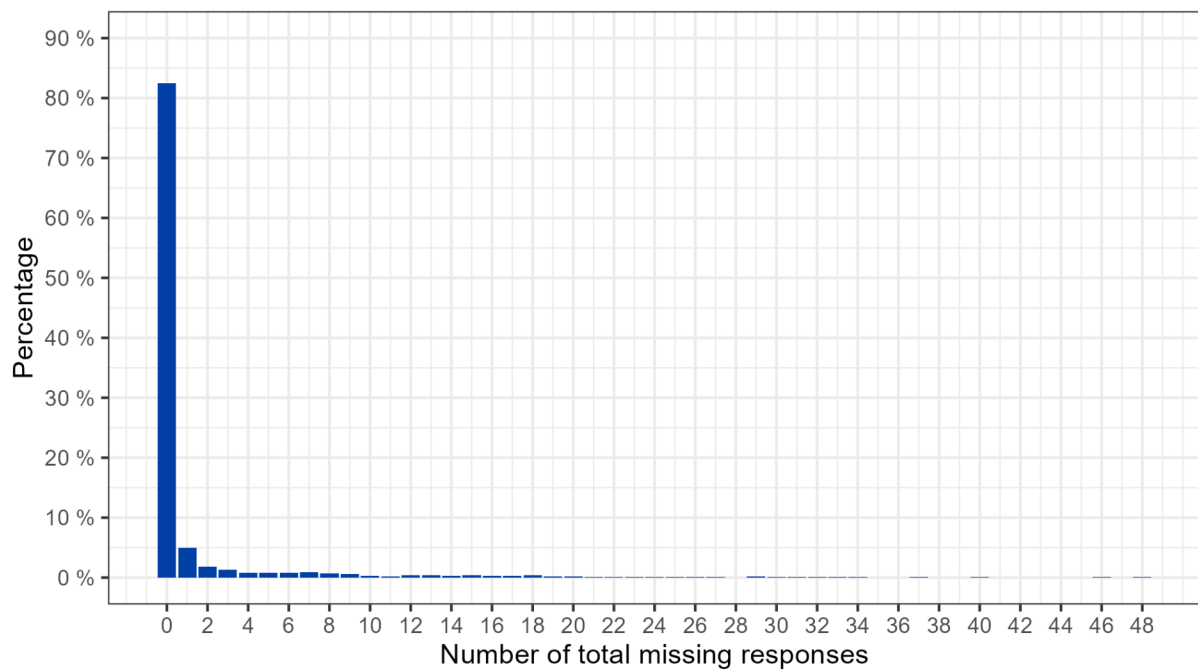
Figure 1

Number of Omitted Items



Another source of missing responses is items that respondents did not reach because they aborted the test, for example, due to time constraints or lack of motivation. These missing values refer to items after the last valid response. As shown in Figure 2, about 93% of respondents did not abort the test and were administered all items. Only 4% of the sample received less than 90% of the entire item set.

The total number of missing responses, aggregated across omitted and not reached missing responses for each respondent, is shown in Figure 3. Because most respondents reached the end of the test and had few omitted items, the total number of missing values was small. The median number of missing responses was 0; about 82% of the respondents had no missing

Figure 2*Number of Not-Reached Items***Figure 3***Total Number of Missing Responses*

response at all. Only 8% of respondents had more than five missing responses, while 3% had missing responses for 16 or more items (i.e., about a quarter of the administered test).

Overall, there was a small amount of omitted and not-reached missing responses. Also, the total number of missing responses was reasonable and did not indicate substantial problems with the test.

5.1.2 Missing Responses per Item

Table 3 provides information on the occurrence of different types of missing responses per item. Overall, the number of respondents that omitted an item was small. The percentage of omitted responses per item varied between 0% and 3% ($Mdn = 1\%$). In addition, most items had relatively few items that were not reached. The percentage of not-reached items varied across items between 0% and 7% ($Mdn = 0\%$). The percentage of respondents not reaching an item slightly increased with the position of the item in the test (see Figure 4).

Table 3

Percentage of Missing Responses by Item

Nr.	Item	Pos.	N	OM	NR	ALL
1	gkg6nch1_c	1	4,300	0.09	0.00	0.09
2	gkg6hph1_c	2	4,289	0.35	0.00	0.35
3	gkg6nge1_c	3	4,299	0.12	0.00	0.12
4	gkg6sec1_c	4	4,296	0.19	0.00	0.19
5	gkg6shi1_c	5	4,280	0.56	0.00	0.56
6	gkg6nbi1_c	6	4,281	0.53	0.00	0.53
7	gkg6hli1_c	7	4,273	0.72	0.00	0.72
8	gkg6nte1_c	8	4,274	0.70	0.00	0.70
9	gkg6shi2_c	9	4,289	0.35	0.00	0.35
10	gkg6hmu1_c	10	4,283	0.49	0.00	0.49
11	gkg6nch2_c	11	4,285	0.44	0.00	0.44
12	gkg6spo1_c	12	4,297	0.16	0.00	0.16
13	gkg6shi3_c	13	4,266	0.88	0.00	0.88
14	gkg6hli2_c	14	4,231	1.70	0.00	1.70
15	gkg6npy1_c	15	4,278	0.60	0.00	0.60
16	gkg6hre1_c	16	4,288	0.37	0.00	0.37
17	gkg6har1_c	17	4,265	0.88	0.02	0.91
18	gkg6nge2_c	18	4,287	0.37	0.02	0.39

Nr.	Item	Pos.	N	OM	NR	ALL
19	gkg6sla1_c	19	4,246	1.32	0.02	1.35
20	gkg6nge3_c	20	4,278	0.56	0.05	0.60
21	gkg6nte2_c	21	4,237	1.49	0.07	1.56
22	gkg6hli3_c	22	4,278	0.53	0.07	0.60
23	gkg6nme1_c	23	4,266	0.81	0.07	0.88
24	gkg6npy2_c	24	4,254	1.09	0.07	1.16
25	gkg6spo2_c	25	4,236	1.51	0.07	1.58
26	gkg6hmu2_c	26	4,242	1.35	0.09	1.44
27	gkg6nch3_c	27	4,238	1.44	0.09	1.53
28	gkg6sla2_c	28	4,223	1.74	0.14	1.88
29	gkg6hph2_c	29	4,234	1.46	0.16	1.63
30	gkg6sfi1_c	30	4,275	0.49	0.19	0.67
31	gkg6nme2_c	31	4,197	2.23	0.26	2.49
32	gkg6har2_c	32	4,250	0.98	0.28	1.25
33	gkg6nte3_c	33	4,258	0.74	0.33	1.07
34	gkg6nch4_c	34	4,260	0.70	0.33	1.02
35	gkg6hli4_c	35	4,253	0.81	0.37	1.18
36	gkg6hmu3_c	36	4,238	1.07	0.46	1.53
37	gkg6sec2_c	37	4,185	2.23	0.53	2.76
38	gkg6hre2_c	38	4,215	1.51	0.56	2.07
39	gkg6sfi2_c	39	4,203	1.65	0.70	2.35
40	gkg6hph3_c	40	4,178	2.11	0.81	2.93
41	gkg6nbi2_c	41	4,239	0.67	0.84	1.51
42	gkg6sla3_c	42	4,180	1.95	0.93	2.88
43	gkg6nme3_c	43	4,227	0.84	0.95	1.79
44	gkg6spo3_c	44	4,221	0.88	1.05	1.93
45	gkg6sla4_c	45	4,179	1.81	1.09	2.90
46	gkg6nme4_c	46	4,194	1.37	1.18	2.56
47	gkg6npy3_c	47	4,207	0.88	1.37	2.25
48	gkg6hre3_c	48	4,106	2.97	1.63	4.60

Nr.	Item	Pos.	N	OM	NR	ALL
49	gkg6nge4_c	49	4,175	1.14	1.86	3.00
50	gkg6hph4_c	50	4,153	1.39	2.11	3.51
51	gkg6hmu4_c	51	4,162	0.95	2.35	3.30
52	gkg6sec3_c	52	4,094	2.28	2.60	4.88
53	gkg6nbi3_c	53	4,142	1.05	2.72	3.76
54	gkg6sec4_c	54	4,094	2.04	2.83	4.88
55	gkg6np4_c	55	4,127	1.14	2.97	4.11
56	gkg6shi4_c	56	4,146	0.53	3.14	3.67
57	gkg6hre4_c	57	4,094	1.32	3.55	4.88
58	gkg6sfi3_c	58	4,037	2.14	4.07	6.20
59	gkg6har3_c	59	4,087	0.63	4.41	5.04
60	gkg6spo4_c	60	4,072	0.60	4.79	5.39
61	gkg6nte4_c	61	4,066	0.33	5.20	5.53
62	gkg6sfi4_c	62	4,040	0.49	5.65	6.13
63	gkg6har4_c	63	4,024	0.46	6.04	6.51
64	gkg6nbi4_c	64	4,006	0.00	6.92	6.92

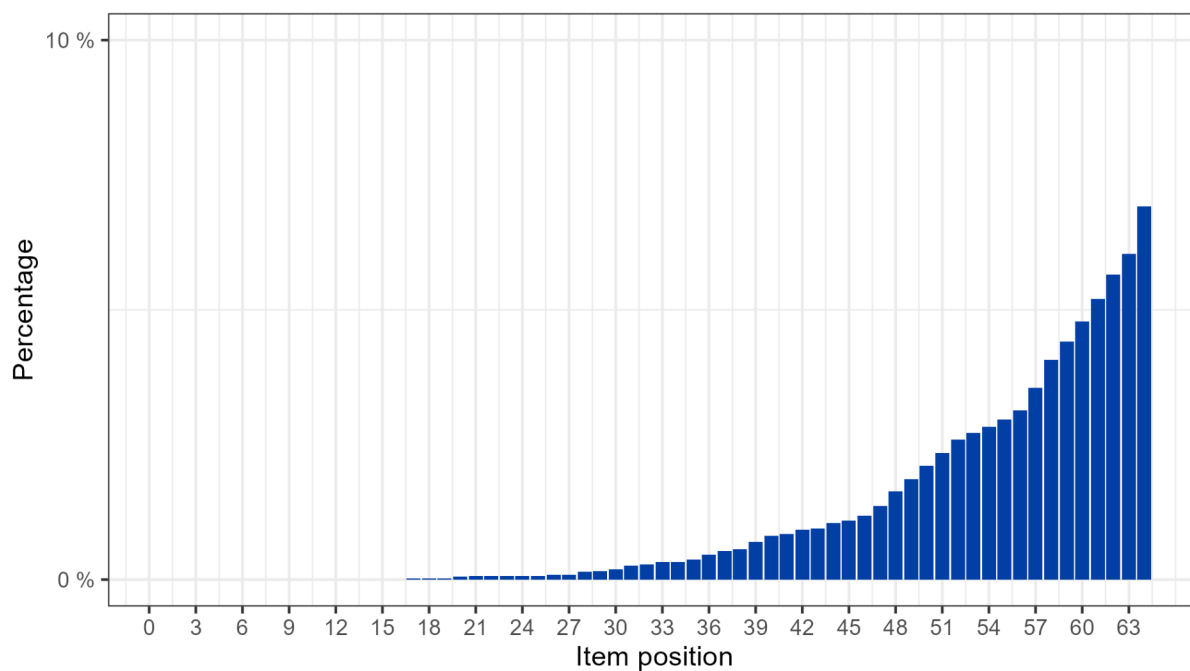
Note. Pos. = Item position within the test. N = Number of valid responses. NR = Percentage of respondents that did not reach an item. OM = Percentage of omitted responses. ALL = Percentage of all types of missing responses.

Because a relatively small number of items was omitted or not reached by respondents, the total number of missing responses per item was also quite small. These varied between 0% and 7% (*Mdn* = 2%).

5.2 Parameter Estimates

5.2.1 Item Parameters

The percentage of correct responses out of all valid responses for each item administered varied between 9.90% and 94.56%, thus covering a broad range (see Table 4). The estimated item difficulties are given in Table 4. Item difficulties were estimated by constraining the mean of the ability distribution to zero. The estimated item difficulties ranged from -3.06 (gkg6nge1_c) to 2.40 (gkg6nch2_c) with a median of 0.09. Thus, the test included easy as well as difficult items. The standard errors (*SE*) of the estimated parameters were rather small for most items (*SE* ≤ 0.07).

Figure 4*Item Position not Reached***Table 4***Item Parameters*

Nr.	Item	<i>N</i>	Percentage correct	Difficulty	<i>SE</i>	WMNSQ	<i>t</i>	<i>r</i> _{it}	<i>Q</i> ₃	Discr.
1	gkg6nch1_c	4,304	45.75	0.19	0.03	1.01	1.02	0.27	0.02	0.65
2	gkg6hph1_c	4,304	55.95	-0.27	0.03	0.93	-7.12	0.40	0.03	1.05
3	gkg6nge1_c	4,304	94.56	-3.06	0.07	0.96	-0.70	0.24	0.02	1.42
4	gkg6sec1_c	4,304	64.57	-0.67	0.03	0.94	-5.21	0.38	0.03	1.07
5	gkg6shi1_c	4,304	47.58	0.11	0.03	1.02	1.73	0.27	0.01	0.62
6	gkg6nbi1_c	4,304	36.57	0.61	0.03	1.05	3.95	0.20	0.02	0.47
7	gkg6hli1_c	4,304	31.88	0.84	0.03	1.09	6.05	0.13	0.02	0.30
8	gkg6nte1_c	4,304	42.96	0.31	0.03	1.01	0.84	0.27	0.02	0.64
9	gkg6shi2_c	4,304	27.02	1.10	0.04	1.15	8.35	0.02	0.03	0.04
10	gkg6hmu1_c	4,304	53.46	-0.16	0.03	1.00	0.16	0.29	0.02	0.68
11	gkg6nch2_c	4,304	9.90	2.40	0.05	1.04	1.06	0.09	0.01	0.32
12	gkg6spo1_c	4,304	82.57	-1.70	0.04	0.97	-1.22	0.29	0.02	0.97
13	gkg6shi3_c	4,304	32.71	0.80	0.03	1.14	9.66	0.06	0.03	0.13

Nr.	Item	N	Percentage correct	Difficulty	SE	WMNSQ	t	r_{it}	Q_3	Discr.
14	gkg6hli2_c	4,304	32.37	0.81	0.03	1.09	6.13	0.13	0.02	0.30
15	gkg6npy1_c	4,304	64.47	-0.66	0.03	1.00	0.16	0.28	0.02	0.71
16	gkg6hre1_c	4,304	70.42	-0.96	0.03	1.00	0.02	0.27	0.02	0.71
17	gkg6har1_c	4,304	50.84	-0.04	0.03	1.01	1.17	0.27	0.02	0.62
18	gkg6nge2_c	4,304	66.61	-0.77	0.03	1.02	1.64	0.24	0.02	0.57
19	gkg6sla1_c	4,304	46.17	0.17	0.03	0.99	-1.06	0.31	0.02	0.74
20	gkg6nge3_c	4,304	54.48	-0.20	0.03	0.97	-3.60	0.35	0.02	0.88
21	gkg6nte2_c	4,304	33.55	0.76	0.03	1.08	6.15	0.14	0.02	0.33
22	gkg6hli3_c	4,304	63.17	-0.60	0.03	0.97	-2.56	0.34	0.02	0.86
23	gkg6nme1_c	4,304	68.05	-0.84	0.03	1.04	2.82	0.20	0.02	0.48
24	gkg6npy2_c	4,304	48.61	0.06	0.03	0.92	-8.53	0.42	0.03	1.10
25	gkg6spo2_c	4,304	47.98	0.09	0.03	0.93	-7.49	0.40	0.03	1.05
26	gkg6hmu2_c	4,304	48.49	0.07	0.03	1.04	4.32	0.23	0.03	0.51
27	gkg6nch3_c	4,304	21.03	1.45	0.04	1.14	6.11	-0.02	0.03	-0.04
28	gkg6sla2_c	4,304	37.96	0.54	0.03	1.07	5.94	0.18	0.02	0.40
29	gkg6hph2_c	4,304	40.80	0.41	0.03	1.03	2.98	0.23	0.01	0.54
30	gkg6sfi1_c	4,304	77.46	-1.36	0.04	0.92	-4.12	0.38	0.03	1.29
31	gkg6nme2_c	4,304	30.13	0.93	0.03	1.11	6.83	0.10	0.03	0.22
32	gkg6har2_c	4,304	17.17	1.72	0.04	1.04	1.41	0.15	0.02	0.44
33	gkg6nte3_c	4,304	33.22	0.77	0.03	1.00	0.32	0.26	0.02	0.65
34	gkg6nch4_c	4,304	67.50	-0.81	0.03	0.91	-7.20	0.43	0.04	1.30
35	gkg6hli4_c	4,304	72.24	-1.06	0.04	0.94	-4.00	0.36	0.04	1.08
36	gkg6hmu3_c	4,304	50.81	-0.04	0.03	1.07	7.31	0.18	0.02	0.40
37	gkg6sec2_c	4,304	33.71	0.75	0.03	1.11	7.92	0.11	0.02	0.23
38	gkg6hre2_c	4,304	58.62	-0.39	0.03	0.93	-6.87	0.40	0.03	1.08
39	gkg6sfi2_c	4,304	52.51	-0.11	0.03	0.91	-10.10	0.44	0.04	1.24
40	gkg6hph3_c	4,304	46.56	0.15	0.03	0.99	-0.53	0.30	0.01	0.71
41	gkg6nbi2_c	4,304	26.79	1.11	0.04	1.04	2.53	0.19	0.02	0.47
42	gkg6sla3_c	4,304	38.48	0.52	0.03	1.00	-0.30	0.29	0.02	0.70

Nr.	Item	<i>N</i>	Percentage correct	Difficulty	<i>SE</i>	WMNSQ	<i>t</i>	<i>r_{it}</i>	<i>Q₃</i>	Discr.
43	gkg6nme3_c	4,304	50.49	-0.02	0.03	1.06	5.91	0.20	0.02	0.45
44	gkg6spo3_c	4,304	37.99	0.54	0.03	0.98	-1.49	0.31	0.02	0.75
45	gkg6sla4_c	4,304	46.49	0.15	0.03	0.94	-6.29	0.38	0.02	0.98
46	gkg6nme4_c	4,304	55.65	-0.25	0.03	0.98	-2.49	0.33	0.02	0.82
47	gkg6npy3_c	4,304	49.51	0.02	0.03	0.99	-1.13	0.31	0.02	0.76
48	gkg6hre3_c	4,304	39.17	0.49	0.03	1.03	2.92	0.23	0.02	0.53
49	gkg6nge4_c	4,304	32.99	0.78	0.03	1.02	1.39	0.25	0.02	0.60
50	gkg6hph4_c	4,304	40.45	0.43	0.03	0.99	-1.32	0.31	0.02	0.75
51	gkg6hmu4_c	4,304	55.18	-0.23	0.03	0.95	-5.52	0.37	0.03	0.97
52	gkg6sec3_c	4,304	26.93	1.10	0.04	1.01	0.74	0.24	0.02	0.64
53	gkg6nbi3_c	4,304	64.41	-0.66	0.03	0.90	-8.13	0.43	0.03	1.30
54	gkg6sec4_c	4,304	47.84	0.09	0.03	0.94	-6.30	0.39	0.03	1.02
55	gkg6npy4_c	4,304	41.84	0.36	0.03	0.90	-10.15	0.45	0.04	1.29
56	gkg6shi4_c	4,304	46.75	0.14	0.03	0.97	-2.98	0.34	0.02	0.83
57	gkg6hre4_c	4,304	40.73	0.41	0.03	0.96	-3.35	0.34	0.02	0.85
58	gkg6sfi3_c	4,304	33.78	0.74	0.03	0.96	-3.38	0.35	0.03	0.92
59	gkg6har3_c	4,304	71.03	-0.99	0.04	0.94	-4.24	0.37	0.03	1.14
60	gkg6spo4_c	4,304	48.21	0.08	0.03	1.08	8.68	0.16	0.03	0.36
61	gkg6nte4_c	4,304	82.83	-1.72	0.04	0.89	-4.52	0.43	0.05	1.99
62	gkg6sfi4_c	4,304	68.77	-0.87	0.03	0.91	-6.83	0.42	0.05	1.36
63	gkg6har4_c	4,304	71.26	-1.00	0.04	0.93	-4.78	0.38	0.04	1.23
64	gkg6nbi4_c	4,304	52.35	-0.11	0.03	1.04	4.46	0.23	0.02	0.53

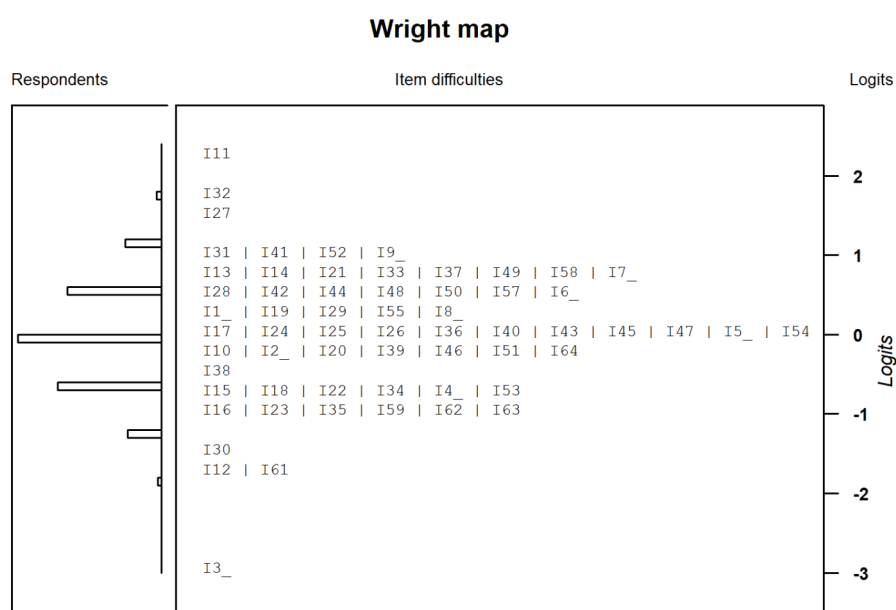
Note. *N* = Number of observed responses. Difficulty = Item difficulty. *SE* = Standard error of item parameter. WMNSQ = Weighted mean square error. *t* = *t*-value for WMNSQ. Discr. = Discrimination parameter of a two-parametric item response model. *r_{it}* = Corrected item-total correlation. *Q₃* = Average absolute residual correlation for item (Yen, 1983).

5.2.2 Test Targeting and Reliability

Test targeting focuses on comparing the item difficulties with the person abilities estimated as weighted likelihood estimates (WLE, Warm, 1989) to evaluate the appropriateness of the test for the specific target population. In Figure 5, the item difficulties of the general knowledge items and the ability of the respondents are plotted on the same scale. The distribution of respondents' estimated abilities is plotted on the left, while the distribution of the item difficulties is plotted

on the right.

Figure 5
Test Targeting



Note. The distribution of the person abilities in the sample is given on the left-hand side of the graph. The item difficulties are given on the right-hand side of the graph, with each number corresponding to the item numbers given in Table 4.

The item difficulties ranged from -3.06 (gkg6nge1_c) to 2.40 (gkg6nch2_c) and thus covered a rather wide range. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 0.48, which implies a somewhat restricted differentiation between respondents. However, the reliability of the test (EAP/PV reliability = 0.86, WLE reliability = 0.86) was good. The median of the item difficulty distribution was about 0.09 logits above the mean person ability distribution. Thus, the items covered a wide range of the ability distribution, closely matching the ability distribution of the respondents.

5.3 Quality of the test

5.3.1 Distractor Analyses

To investigate how well the distractors of the multiple-choice items performed, point-biserial correlations were calculated between each incorrect response (distractor) and the respondents' total correct scores for the remaining items. The median point-biserial correlation for the distractors fell at -.13 (*Min* = -.31, *Max* = .16). Only 5 out of 192 distractors showed correlations greater than .05. Overall, these results indicate that the distractors worked well.

5.3.2 Item Fit

The correlations of the correct responses with the total-rest scores varied between -0.02 (gkg6nch3_c) and 0.45 (gkg6npy4_c), with a median of 0.28. For about a quarter of the items, the point-biserial correlations were somewhat small and fell below .20; six of them had correlations smaller than .10.

The evaluation of item fit was based on the scaling model presented above. Overall, the item fit was good (see Table 4). The WMNSQ values ranged from 0.89 (gkg6nte4_c) to 1.15 (gkg6shi2_c), with a median of 1.00. There was no item with a WMSNQ greater than 1.15. However, 3 items (gkg6shi2_c, gkg6shi3_c, gkg6spo4_c) had *t*-values exceeding 8. These exhibited slightly flatter ICCs than expected by the Rasch model. Furthermore, visual inspections of the empirical ICCs revealed pronounced deviations from the expected ICCs for two additional items (gkg6shi2_c, gkg6nch3_c). These results show that most of the items exhibited a good fit to the Rasch (1960) model.

5.3.3 Differential Item Functioning

DIF was used to evaluate whether the measurement models were comparable for different subgroups. For this purpose, DIF was examined for the variables gender, migrant background, number of books at home (as a proxy for cultural capital), and school type. A description of these variables can be found in Pohl and Carstensen (2012). The differences between the estimated item difficulties (on the logit scale) in the different groups are summarized in Table 5. For example, the column “boys vs. girls” reports the differences in item difficulties between boys and girls; a positive value would indicate that the test was more difficult for boys, whereas a negative value would indicate that the item was less difficult for boys than for girls. In addition to examining DIF for each item, an overall test for DIF was performed by comparing models that allow for DIF with those that only estimate main effects. In Table 6, models including only main effects were compared with those additionally estimating DIF using Akaike’s (1974) information criterion (AIC) and Bayesian information criterion (BIC, Schwarz, 1978).

Gender: The sample included 2,134 boys and 2,169 girls. There were no gender differences in general knowledge as indicated by the main effect of 0.07 logits (Cohen’s $d = 0.10$) that reflected higher general knowledge for boys than girls. Four items (gkg6hmu2_c, gkg6nch4_c, gkg6hli4_c, gkg6har3_c) showed DIF effects of more than 0.60 logits that were more difficult for boys than for girls. Furthermore, two items (gkg6nte1_c, gkg6nte3_c) with considerable DIF were more difficult for girls than for boys. Consequently, an overall test for DIF (see Table 6) also suggested a better fit for the model with DIF compared to a model that only estimated the main effect (but ignored potential DIF). However, these DIF effects had no noticeable impact on the main effect for gender, which was rather similar whether DIF effects were estimated (0.07) or not (0.08). These results suggest that for most items there was no pronounced DIF concerning gender that might have biased the parameter estimates.

Table 5*Differential Item Functioning*

Item	Gender	Migration	Books	School type		
	boys vs. girls	without vs. with	<100 vs. >100	lower vs. mixed	lower vs. upper	mixed vs. upper
gkg6nch1_c	0.11 (0.16)	0.07 (0.10)	0.15 (0.23)	-0.02 (-0.03)	0.02 (0.03)	0.04 (0.06)
gkg6hph1_c	0.13 (0.19)	-0.09 (-0.13)	0.28 (0.42)	0.09 (0.15)	0.42 (0.68)	0.32 (0.52)
gkg6nge1_c	0.52 (0.74)	-0.19 (-0.28)	0.49 (0.74)	-0.08 (-0.13)	0.62 (1.00)	0.71 (1.15)
gkg6sec1_c	-0.38 (-0.54)	-0.49 (-0.72)	0.11 (0.17)	-0.07 (-0.11)	0.19 (0.31)	0.26 (0.42)
gkg6shi1_c	0.00 (0.00)	0.05 (0.07)	-0.08 (-0.12)	0.07 (0.11)	0.08 (0.13)	0.00 (0.00)
gkg6nbi1_c	0.25 (0.36)	-0.15 (-0.22)	-0.15 (-0.23)	0.05 (0.08)	-0.10 (-0.16)	-0.15 (-0.24)
gkg6hli1_c	0.26 (0.37)	0.28 (0.41)	-0.26 (-0.39)	-0.07 (-0.11)	-0.34 (-0.55)	-0.27 (-0.44)
gkg6nte1_c	-0.79 (-1.13)	-0.06 (-0.09)	-0.06 (-0.09)	0.07 (0.11)	-0.30 (-0.48)	-0.37 (-0.60)
gkg6shi2_c	-0.06 (-0.09)	0.65 (0.95)	-0.40 (-0.60)	-0.13 (-0.21)	-0.71 (-1.15)	-0.58 (-0.94)
gkg6hmu1_c	0.43 (0.62)	-0.44 (-0.64)	0.14 (0.21)	0.02 (0.03)	0.36 (0.58)	0.34 (0.55)
gkg6nch2_c	-0.12 (-0.17)	0.60 (0.88)	-0.51 (-0.77)	-0.04 (-0.06)	-0.38 (-0.61)	-0.34 (-0.55)
gkg6spo1_c	-0.14 (-0.20)	0.18 (0.26)	0.27 (0.41)	-0.09 (-0.15)	0.35 (0.57)	0.44 (0.71)
gkg6shi3_c	0.17 (0.24)	0.38 (0.56)	-0.44 (-0.66)	-0.26 (-0.42)	-0.75 (-1.21)	-0.49 (-0.79)
gkg6hli2_c	0.14 (0.20)	0.39 (0.57)	-0.18 (-0.27)	-0.02 (-0.03)	-0.40 (-0.65)	-0.38 (-0.61)
gkg6npy1_c	-0.46 (-0.66)	-0.68 (-1.00)	0.12 (0.18)	-0.02 (-0.03)	-0.05 (-0.08)	-0.03 (-0.05)
gkg6hre1_c	0.21 (0.30)	0.82 (1.20)	-0.18 (-0.27)	-0.01 (-0.02)	0.16 (0.26)	0.18 (0.29)
gkg6har1_c	-0.03 (-0.04)	-0.25 (-0.37)	0.02 (0.03)	0.02 (0.03)	0.10 (0.16)	0.07 (0.11)

Item	Gender	Migration	Books		School type	
	boys vs. girls	without vs. with	<100 vs. >100	lower vs. mixed	lower vs. upper	mixed vs. upper
gkg6nge2_c	-0.46 (-0.66)	0.39 (0.57)	-0.01 (-0.02)	-0.16 (-0.26)	-0.06 (-0.10)	0.10 (0.16)
gkg6sla1_c	-0.20 (-0.29)	-0.01 (-0.01)	0.07 (0.11)	0.05 (0.08)	-0.01 (-0.02)	-0.05 (-0.08)
gkg6nge3_c	-0.35 (-0.50)	-0.51 (-0.75)	0.29 (0.44)	0.10 (0.16)	0.11 (0.18)	0.01 (0.02)
gkg6nte2_c	-0.28 (-0.40)	0.60 (0.88)	-0.21 (-0.32)	-0.01 (-0.02)	-0.51 (-0.82)	-0.51 (-0.82)
gkg6hli3_c	0.09 (0.13)	-0.28 (-0.41)	0.16 (0.24)	0.09 (0.15)	0.35 (0.57)	0.27 (0.44)
gkg6nme1_c	0.14 (0.20)	0.47 (0.69)	-0.30 (-0.45)	0.02 (0.03)	-0.34 (-0.55)	-0.36 (-0.58)
gkg6npy2_c	-0.32 (-0.46)	-0.43 (-0.63)	0.30 (0.45)	-0.05 (-0.08)	0.28 (0.45)	0.32 (0.52)
gkg6spo2_c	0.12 (0.17)	-0.25 (-0.37)	0.24 (0.36)	0.16 (0.26)	0.36 (0.58)	0.21 (0.34)
gkg6hmu2_c	1.34 (1.92)	-0.01 (-0.01)	-0.21 (-0.32)	0.19 (0.31)	0.10 (0.16)	-0.09 (-0.15)
gkg6nch3_c	-0.30 (-0.43)	0.67 (0.98)	-0.32 (-0.48)	-0.10 (-0.16)	-0.95 (-1.53)	-0.85 (-1.37)
gkg6sla2_c	0.09 (0.13)	0.39 (0.57)	-0.21 (-0.32)	-0.04 (-0.06)	-0.40 (-0.65)	-0.37 (-0.60)
gkg6hph2_c	-0.20 (-0.29)	0.79 (1.16)	-0.03 (-0.05)	-0.05 (-0.08)	-0.16 (-0.26)	-0.12 (-0.19)
gkg6sfi1_c	0.23 (0.33)	-0.59 (-0.86)	0.00 (0.00)	0.03 (0.05)	0.32 (0.52)	0.29 (0.47)
gkg6nme2_c	0.07 (0.10)	0.23 (0.34)	-0.37 (-0.56)	0.04 (0.06)	-0.58 (-0.94)	-0.63 (-1.02)
gkg6har2_c	0.48 (0.69)	0.44 (0.64)	-0.08 (-0.12)	-0.06 (-0.10)	-0.31 (-0.50)	-0.25 (-0.40)
gkg6nte3_c	-0.93 (-1.33)	-0.33 (-0.48)	0.01 (0.02)	0.04 (0.06)	-0.20 (-0.32)	-0.24 (-0.39)
gkg6nch4_c	0.69 (0.99)	-1.06 (-1.55)	0.37 (0.56)	0.23 (0.37)	0.60 (0.97)	0.38 (0.61)
gkg6hli4_c	0.90 (1.29)	-1.00 (-1.47)	0.50 (0.75)	0.25 (0.40)	0.53 (0.86)	0.27 (0.44)
gkg6hmu3_c	0.29 (0.42)	-0.13 (-0.19)	-0.16 (-0.24)	0.25 (0.40)	-0.03 (-0.05)	-0.28 (-0.45)

Item	Gender	Migration	Books		School type	
	boys vs. girls	without vs. with	<100 vs. >100	lower vs. mixed	lower vs. upper	mixed vs. upper
gkg6sec2_c	0.08 (0.11)	0.42 (0.62)	-0.23 (-0.35)	-0.04 (-0.06)	-0.30 (-0.48)	-0.25 (-0.40)
gkg6hre2_c	-0.30 (-0.43)	-0.39 (-0.57)	-0.05 (-0.08)	-0.10 (-0.16)	0.15 (0.24)	0.25 (0.40)
gkg6sfi2_c	0.15 (0.21)	-0.73 (-1.07)	0.23 (0.35)	0.33 (0.53)	0.60 (0.97)	0.27 (0.44)
gkg6hph3_c	0.19 (0.27)	0.04 (0.06)	-0.13 (-0.20)	0.06 (0.10)	0.02 (0.03)	-0.03 (-0.05)
gkg6nbi2_c	0.51 (0.73)	0.08 (0.12)	-0.20 (-0.30)	-0.04 (-0.06)	-0.07 (-0.11)	-0.03 (-0.05)
gkg6sla3_c	0.17 (0.24)	0.07 (0.10)	0.06 (0.09)	0.14 (0.23)	0.04 (0.06)	-0.10 (-0.16)
gkg6nme3_c	0.00 (0.00)	-0.07 (-0.10)	-0.49 (-0.74)	-0.02 (-0.03)	-0.30 (-0.48)	-0.28 (-0.45)
gkg6spo3_c	-0.37 (-0.53)	-0.03 (-0.04)	0.31 (0.47)	-0.07 (-0.11)	0.16 (0.26)	0.23 (0.37)
gkg6sla4_c	0.01 (0.01)	-0.52 (-0.76)	0.05 (0.08)	0.01 (0.02)	0.14 (0.23)	0.13 (0.21)
gkg6nme4_c	0.11 (0.16)	0.17 (0.25)	-0.09 (-0.14)	-0.07 (-0.11)	-0.06 (-0.10)	0.01 (0.02)
gkg6npy3_c	-0.48 (-0.69)	0.10 (0.15)	-0.17 (-0.26)	0.10 (0.16)	-0.18 (-0.29)	-0.28 (-0.45)
gkg6hre3_c	0.03 (0.04)	-0.02 (-0.03)	-0.15 (-0.23)	0.07 (0.11)	-0.30 (-0.48)	-0.37 (-0.60)
gkg6nge4_c	-0.33 (-0.47)	0.22 (0.32)	-0.11 (-0.17)	-0.04 (-0.06)	-0.23 (-0.37)	-0.19 (-0.31)
gkg6hph4_c	-0.10 (-0.14)	-0.02 (-0.03)	-0.02 (-0.03)	-0.20 (-0.32)	-0.17 (-0.27)	0.03 (0.05)
gkg6hmu4_c	0.35 (0.50)	-0.86 (-1.26)	0.27 (0.41)	0.07 (0.11)	0.39 (0.63)	0.32 (0.52)
gkg6sec3_c	-0.16 (-0.23)	0.21 (0.31)	-0.01 (-0.02)	-0.01 (-0.02)	-0.27 (-0.44)	-0.26 (-0.42)
gkg6nbi3_c	0.13 (0.19)	-0.51 (-0.75)	0.14 (0.21)	0.00 (0.00)	0.48 (0.78)	0.48 (0.78)
gkg6sec4_c	-0.29 (-0.42)	-0.08 (-0.12)	0.14 (0.21)	-0.16 (-0.26)	0.04 (0.06)	0.20 (0.32)
gkg6npy4_c	-0.38 (-0.54)	-0.14 (-0.21)	0.27 (0.41)	-0.03 (-0.05)	0.17 (0.27)	0.19 (0.31)

Item	Gender	Migration	Books	School type		
	boys vs. girls	without vs. with	<100 vs. >100	lower vs. mixed	lower vs. upper	mixed vs. upper
gkg6shi4_c	-0.38 (-0.54)	0.56 (0.82)	0.02 (0.03)	-0.11 (-0.18)	0.07 (0.11)	0.18 (0.29)
gkg6hre4_c	-0.07 (-0.10)	-0.04 (-0.06)	0.13 (0.20)	-0.18 (-0.29)	0.04 (0.06)	0.22 (0.36)
gkg6sfi3_c	-0.53 (-0.76)	0.20 (0.29)	0.16 (0.24)	0.10 (0.16)	-0.04 (-0.06)	-0.14 (-0.23)
gkg6har3_c	0.68 (0.97)	0.14 (0.21)	0.24 (0.36)	-0.05 (-0.08)	0.42 (0.68)	0.47 (0.76)
gkg6spo4_c	-0.09 (-0.13)	0.91 (1.33)	-0.36 (-0.54)	-0.08 (-0.13)	-0.32 (-0.52)	-0.24 (-0.39)
gkg6nte4_c	-0.02 (-0.03)	0.04 (0.06)	0.31 (0.47)	-0.02 (-0.03)	0.64 (1.03)	0.67 (1.08)
gkg6sfi4_c	-0.51 (-0.73)	-0.15 (-0.22)	0.31 (0.47)	-0.10 (-0.16)	0.28 (0.45)	0.38 (0.61)
gkg6har4_c	0.07 (0.10)	-0.30 (-0.44)	0.19 (0.29)	-0.03 (-0.05)	0.38 (0.61)	0.41 (0.66)
gkg6nbi4_c	0.12 (0.17)	-0.30 (-0.44)	0.16 (0.24)	-0.01 (-0.02)	0.13 (0.21)	0.14 (0.23)
Main effect (DIF model)	0.07 (0.10)	0.55 (0.81)	-0.48 (-0.73)	-0.03 (-0.05)	-0.69 (-1.11)	-0.66 (-1.06)
Main effect (Main effect model)	0.08 (0.11)	0.57 (0.84)	-0.48 (-0.73)	-0.03 (-0.05)	-0.69 (-1.12)	-0.65 (-1.07)

Note. Raw differences between item difficulties (in logits) with standardized differences (Cohen's *d*) in parentheses. lower = lower secondary modern school ("Haupt-/Realschule"); upper = upper secondary modern school ("Gymnasium"); mixed = secondary modern schools without differentiation ("Gesamtschule").

Migrant background: There were 3,589 participants without a migrant background and 132 participants with a migrant background. There was a moderate difference between the average performance of participants with and without a migrant background. Participants without a migrant background had a higher general knowledge than participants with a migrant background (main effect = 0.55 logits, Cohen's $d = 0.81$). There were 10 items with DIF greater than 0.60 between respondents with and without a migrant background. The largest difference in item difficulties between these groups was 1.06 logits (gkg6nch4_c). Although this appears to be a rather large difference, it may be a consequence of the uncertainty in the estimation due to the small number of respondents with a migrant background. Accordingly, an overall test for DIF using the BIC (see Table 6) suggested a better fit for the main effect model that did not acknowledge DIF effects.

Number of books: The number of books at home was used as a proxy for cultural capital. There were 1,120 participants with up to 100 books at home and 1,396 participants with more than 100 books at home. There was a considerable average difference between the two groups. Participants with up to 100 books at home performed on average -0.48 logits (Cohen's $d = -0.73$) lower in general knowledge than participants with more than 100 books. There were no items with DIF greater than 0.60 between respondents with many or few books at home. The largest difference in item difficulties between these groups was 0.51 logits (gkg6nch2_c). Therefore, the overall test for DIF using the BIC (see Table 6) favored the model without DIF.

Table 6

Comparisons of Models with and without DIF

DIF variable	Model	<i>N</i>	Deviance	Number of parameters	AIC	BIC
Gender	Main effect	4,303	331,497	66	331,629	332,049
	DIF	4,303	329,525	129	<u>329,783</u>	<u>330,604</u>
Migration	Main effect	3,721	286,343	66	286,475	<u>286,885</u>
	DIF	3,721	286,035	129	<u>286,293</u>	287,096
Books	Main effect	2,516	192,317	66	192,449	<u>192,834</u>
	DIF	2,516	191,915	129	<u>192,173</u>	192,925
School type	Main effect	4,150	318,618	67	318,752	319,177
	DIF	4,150	317,300	193	<u>317,686</u>	<u>318,908</u>

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

School type: The students attended different types of schools. There were 799 respondents from lower secondary modern schools ("Haupt-/Realschule"), 2,355 respondents from upper secondary modern schools ("Gymnasium"), and 996 respondents from secondary modern schools without differentiation ("Gesamtschule"). While respondents from lower secondary modern schools and schools without differentiation performed comparably on average -0.03 logits (Cohen's $d = -0.05$), students from upper secondary modern schools had higher general

knowledge of -0.69 logits (Cohen's $d = -1.11$) and -0.66 logits (Cohen's $d = -1.06$), respectively. There were 9 items with noticeable DIF between the three groups. The largest difference in item difficulty was 0.95 logits (gkg6nch3_c). Also, the overall test for DIF using the BIC (see Table 6) showed a better fit for the model with DIF. However, this hardly affected the main effects for the comparisons involving upper secondary modern schools. These were similar regardless whether DIF was acknowledged or not.

5.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item discrimination parameters are equal. To test this assumption, a two-parametric item response model (2PL, Birnbaum, 1968) was fitted to the data to estimate the discrimination parameters. The estimated discrimination parameters differed moderately between items (see Table 4). The median discrimination parameter fell at 0.71 ($Min = -0.04$, $Max = 1.99$). Model fit indices suggested a better model fit of the 2PL (AIC = 327,612, BIC = 328,427, number of parameters = 128) compared to the Rasch model (AIC = 331,717, BIC = 332,131, number of parameters = 65). However, an inspection of the respective item characteristic curves indicated an adequate fit of the Rasch model for most items (see Section 5.3.2). Despite the empirical preference for the 2PL, the Rasch model is more in line with the theoretical concepts underlying the test construction and the scoring approach recommended by the test authors (see Schroeders et al., 2020).

5.3.5 Dimensionality

The dimensionality of the test was investigated by evaluating the correlations between the residuals of the Rasch (1960) model. The adjusted Q_3 statistics (see Table 4) were quite low ($Mdn = 0.02$) - the largest absolute residual correlation was 0.05 (gkg6nte4_c, gkg6sfi4_c). Overall, these results indicate an essentially unidimensional test.

The dimensionality of the test was also examined by specifying two multidimensional models based on the 16 content areas and the three domains (see Section 2.1). Each item was assigned to one of 16 or three latent factors (between-item multidimensionality). The multidimensional models were estimated using Quasi Monte Carlo method with 5,000 nodes.

The variances and correlations of the 16 dimensions reflecting the content areas are summarized in Table 7. Most dimensions exhibited somewhat restricted variances. The largest variances were observed for the content areas *Finances* and *Physics*. As expected, the correlations between the content areas were rather high, ranging between .54 and .92. Because many of them were substantially smaller than $r = .95$, the different content areas seem to capture some unique variance in addition to their shared variances (see Carstensen, 2013). Accordingly, the 16-dimensional model did fit the data better (AIC = 330,749, BIC = 332,023, number of parameters = 200) than the unidimensional model (AIC = 331,716, BIC = 332,129, number of parameters = 65).

Table 7*Results of 16-Dimensional Scaling for Content Areas*

Dimension	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8	Dim 9	Dim 10	Dim 11	Dim 12	Dim 13	Dim 14	Dim 15	Dim 16
Dim 1: Art	0.78															
Dim 2: Biology	.88	0.50														
Dim 3: Chemistry	.85	.82	0.45													
Dim 4: Economy	.82	.86	.76	0.60												
Dim 5: Finances	.88	.87	.78	.92	1.59											
Dim 6: Geography	.87	.86	.73	.85	.90	0.71										
Dim 7: History	.79	.74	.67	.70	.75	.78	0.24									
Dim 8: Law	.82	.81	.81	.82	.82	.75	.77	0.62								
Dim 9: Literature	.81	.76	.80	.68	.71	.70	.74	.81	0.52							
Dim 10: Medicine	.76	.80	.70	.79	.80	.81	.65	.69	.57	0.33						
Dim 11: Music	.81	.73	.82	.64	.68	.72	.73	.78	.79	.67	0.65					
Dim 12: Philosophy	.84	.85	.77	.81	.81	.84	.77	.85	.76	.75	.78	0.71				
Dim 13: Physics	.82	.81	.76	.90	.92	.83	.63	.78	.64	.74	.64	.75	1.18			
Dim 14: Politics	.83	.82	.74	.86	.89	.82	.79	.86	.75	.73	.68	.78	.82	0.63		
Dim 15: Religion	.78	.78	.71	.81	.84	.82	.80	.86	.72	.77	.77	.84	.76	.85	0.79	
Dim 16: Technology	.77	.75	.63	.82	.85	.79	.70	.73	.54	.71	.57	.72	.80	.83	.74	0.73

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

The variances and correlations of the three dimensions reflecting the domains of the general knowledge test are summarized in Table 8. All dimensions exhibited somewhat restricted variances. The correlations between the three dimensions were large and fell around .90. Consequently, the unidimensional model provided a better fit than the three-dimensional model (AIC = 331,755, BIC = 332,201, number of parameters = 70). This suggests that the three domains measured an essentially unidimensional construct.

Table 8

Results of Three-Dimensional Scaling for Domains

Dimension	Dim 1	Dim 2	Dim 3
Dim 1: Humanities	0.54		
Dim 2: Natural sciences	.90	0.49	
Dim 3: Social sciences	.91	.90	0.54

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

These results suggest that the 16 content areas and three domains measure a common construct. However, the test is not entirely unidimensional. Rather, the content areas also capture some unique variance components beyond the common general knowledge factor.

6. Discussion

The analyses in the previous sections provided information on the quality of the general knowledge test that was administered in Starting Cohort 8 of the NEPS. Different types 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.

The number of missing values was low suggesting that the test was appropriate for the examined sample and the available testing time. Also, several criteria indicated a good fit of the items to Rasch model. However, some items exhibited low discriminations and lacked measurement invariance across different subgroups. Nevertheless, the test was adequately targeted at the ability level of the respondents and exhibited a good reliability.

Although strict unidimensionality could not be corroborated for the different content areas, the three domains of the general knowledge test were substantially correlated showing little unique variance. This suggests that despite some multidimensionality the items capture an essentially unidimensional construct.

In conclusion, the test had satisfactory psychometric properties that allowed the estimation of a unidimensional general knowledge score for the participants.

7. Data in the Scientific Use Files

7.1 Naming Conventions

The SUF for the starting cohort contains 64 items that are scored dichotomously with 0 indicating an incorrect response and 1 indicating a correct response. Details on the naming conventions of the variables can be found in Fuß et al. (2025).

7.2 General Knowledge Scores

In the SUF, general knowledge scores are provided in the form of sum scores (gkg6_sc3) as recommended by the test authors (Schroeders et al., 2020). Additionally, also domain scores are available that measure students' knowledge in humanities (gkg6_sc3a), natural sciences (gkg6_sc3b), and social sciences (gkg6_sc3c).

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–722. <https://doi.org/10.1109/TAC.1974.1100705>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.
- Carstensen, C. H. (2013). Linking PISA competencies over three cycles – results from Germany. In M. Prenzel, M. Kobarg, Schöps, & S. Rönnebeck (Eds.), *Research outcomes of the PISA Research Conference 2009* (pp. 199–213). Springer. https://doi.org/10.1007/978-94-007-4458-5_12
- Fuß, D., Gnambs, T., Lockl, K., Attig, M., & Nusser, L. (2025). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 8)*. Leibniz Institute for Educational Trajectories.
- Horn, J. L., & Noll, J. (1997). Human cognitive abilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53–91). Guilford Press.
- Kutscher, T. (2026). *NEPS Technical Report for digital competence: Scaling results of Starting Cohort 8 in Grade 6* [Unpublished manuscript]. Leibniz Institute for Educational Trajectories.
- Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the data of the competence tests* (NEPS Working Paper 14). University of Bamberg, National Educational Panel Study. <https://doi.org/10.5157/NEPS:WP14:1.0>
- R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.0) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
- Robitzsch, A., Kiefert, T., & Wu, M. (2024). *TAM: Test analysis modules* (Version 4.2-21) [Computer software]. <https://doi.org/10.32614/CRAN.package.TAM>
- Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 73–163). Guilford Press.
- Schroeders, U., Schipolowski, S., & Wilhelm, O. (2020). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 5. und 7. Jahrgangsstufe* (1st ed.). Hogrefe.
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464. <https://doi.org/10.1214/aos/1176344136>
- Senkbeil, M. (2026). *NEPS Technical Report for computer literacy: Scaling results of Starting Cohort 8 in Grade 6* [Unpublished manuscript]. Leibniz Institute for Educational Trajectories.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(2), 427–450. <https://doi.org/10.1007/BF02294627>
- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., Carstensen, C. H., & Lockl, K. (2019). Development of competencies across the life course. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 57–82). Springer. https://doi.org/10.1007/978-3-658-23162-0_4
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125–145. <https://doi.org/10.1177/014662168400800201>
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item depen-

dence. *Journal of Educational Measurement*, 30(3), 187–213. <https://doi.org/10.1111/j.1745-3984.1993.tb00423.x>

Appendix A: Content Areas and Domains

No.	Item	Content area	Domain
1	gkg6nch1_c	Chemistry	Natural sciences
2	gkg6hph1_c	Philosophy	Humanities
3	gkg6nge1_c	Geography	Natural sciences
4	gkg6sec1_c	Economy	Social sciences
5	gkg6shi1_c	History	Social sciences
6	gkg6nbi1_c	Biology	Natural sciences
7	gkg6hli1_c	Literature	Humanities
8	gkg6nte1_c	Technology	Natural sciences
9	gkg6shi2_c	History	Social sciences
10	gkg6hmu1_c	Music	Humanities
11	gkg6nch2_c	Chemistry	Natural sciences
12	gkg6spo1_c	Politics	Social sciences
13	gkg6shi3_c	History	Social sciences
14	gkg6hli2_c	Literature	Humanities
15	gkg6npy1_c	Physics	Natural sciences
16	gkg6hre1_c	Religion	Humanities
17	gkg6har1_c	Art	Humanities
18	gkg6nge2_c	Geography	Natural sciences
19	gkg6sla1_c	Law	Social sciences
20	gkg6nge3_c	Geography	Natural sciences
21	gkg6nte2_c	Technology	Natural sciences
22	gkg6hli3_c	Literature	Humanities
23	gkg6nme1_c	Medicine	Natural sciences
24	gkg6npy2_c	Physics	Natural sciences
25	gkg6spo2_c	Politics	Social sciences
26	gkg6hmu2_c	Music	Humanities
27	gkg6nch3_c	Chemistry	Natural sciences
28	gkg6sla2_c	Law	Social sciences

No.	Item	Content area	Domain
29	gkg6hph2_c	Philosophy	Humanities
30	gkg6sfi1_c	Finances	Social sciences
31	gkg6nme2_c	Medicine	Natural sciences
32	gkg6har2_c	Art	Humanities
33	gkg6nte3_c	Technology	Natural sciences
34	gkg6nch4_c	Chemistry	Natural sciences
35	gkg6hli4_c	Literature	Humanities
36	gkg6hmu3_c	Music	Humanities
37	gkg6sec2_c	Economy	Social sciences
38	gkg6hre2_c	Religion	Humanities
39	gkg6sfi2_c	Finances	Social sciences
40	gkg6hph3_c	Philosophy	Humanities
41	gkg6nbi2_c	Biology	Natural sciences
42	gkg6sla3_c	Law	Social sciences
43	gkg6nme3_c	Medicine	Natural sciences
44	gkg6spo3_c	Politics	Social sciences
45	gkg6sla4_c	Law	Social sciences
46	gkg6nme4_c	Medicine	Natural sciences
47	gkg6npy3_c	Physics	Natural sciences
48	gkg6hre3_c	Religion	Humanities
49	gkg6nge4_c	Geography	Natural sciences
50	gkg6hph4_c	Philosophy	Humanities
51	gkg6hmu4_c	Music	Humanities
52	gkg6sec3_c	Economy	Social sciences
53	gkg6nbi3_c	Biology	Natural sciences
54	gkg6sec4_c	Economy	Social sciences
55	gkg6npy4_c	Physics	Natural sciences
56	gkg6shi4_c	History	Social sciences
57	gkg6hre4_c	Religion	Humanities
58	gkg6sfi3_c	Finances	Social sciences

No.	Item	Content area	Domain
59	gkg6har3_c	Art	Humanities
60	gkg6spo4_c	Politics	Social sciences
61	gkg6nte4_c	Technology	Natural sciences
62	gkg6sfi4_c	Finances	Social sciences
63	gkg6har4_c	Art	Humanities
64	gkg6nbi4_c	Biology	Natural sciences