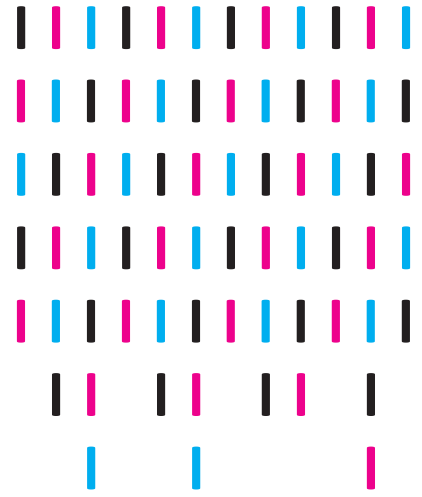




NEPS *SURVEY PAPERS*

Timo Gnambs and
Martin Senkbeil

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

NEPS *Survey Paper* No. 106
Bamberg, September 2023

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

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

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

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

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

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

Editor-in-Chief: Thomas Bäumer, LIfBi

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

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

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

Timo Gnambs¹ and Martin Senkbeil²

¹Leibniz Institute for Educational Trajectories, Bamberg, Germany

²Leibniz Institute for Science and Mathematics Education, Kiel, Germany

E-mail address of lead author:

timo.gnambs@lifbi.de

Bibliographic data:

Gnambs, T., & Senkbeil, M. (2023). *NEPS Technical Report for Computer Literacy: Scaling Results for Starting Cohorts 4 and 6 (Wave 14)* (NEPS Survey Paper No. 106). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP106:1.0>

Acknowledgements:

This report is an extension to NEPS *Working Paper 61* (Senkbeil & Ihme, 2015) and NEPS *Survey Paper 25* (Senkbeil & Ihme, 2017) that present the scaling results for computer literacy of Starting Cohorts 4 and 6 in the previous waves. Therefore, various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced verbatim to facilitate the understanding of the presented results.

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

Abstract

The National Educational Panel Study (NEPS) examines the development of competencies across the life span. Therefore, the NEPS develops tests for the assessment of various competence domains in different age cohorts. To evaluate the quality of these competence tests, several analyses based on item response theory (IRT) are performed. This paper describes the data and scaling procedures for a test of computer literacy that was administered in Wave 14 of Starting Cohorts 4 (ninth grade) and 6 (adults). The computer literacy test included 34 static and interactive items with different response formats that were administered in a computerized multi-stage test design. The adaptive design allowed the administration of different items to each respondent tailored to the individual competence level. The test was administered to a total of 6,215 respondents (50% women) from Starting Cohort 4 ($N = 2,493$) and Starting Cohort 6 ($N = 3,722$). The responses of the participants were scaled using a unidimensional partial credit model. Item fit statistics and differential item functioning were evaluated to ensure the quality of the test. These analyses showed that the test differentiated well between respondents, exhibited good reliability, and showed a satisfactory fit to the item response model. Furthermore, comparable measurement models could be confirmed for different subgroups. Limitations of the test pertained to an increased number of missing values at the end of the test and in the adult cohort. Overall, the computer literacy test had good psychometric properties that allowed for an estimation of reliable competence scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the R syntax for scaling the data.

Keywords

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

1. Introduction

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

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

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

2. Testing Computer Literacy

2.1 Conceptual Framework

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

Computer literacy is conceptualized as a unidimensional construct comprising different facets of technological and information literacy. In line with the literacy concepts of international large-scale assessments, computer literacy is defined from a functional perspective. That is, functional literacy is understood to include the knowledge and skills that people need to live satisfying lives in terms of personal and economic satisfaction in modern-day societies. This leads to an assessment framework that relies heavily on everyday problems, which are more or less distant to school curricula. As a basis for the construction of the instrument assessing computer literacy in the NEPS, we use a framework that identifies four process components (*access, create, manage, and evaluate*) of computer literacy representing the knowledge and skills needed for a problem-oriented use of modern information and communication technology (see Figure 1). Apart from the process components, the test construction of the *Test of Technological and Information Literacy* (TILT) is guided by a categorization of software applications (*word processing, spreadsheet / presentation software, e-mail / communication tools, and internet / search engines*) that are used to locate, process, present, and communicate information.

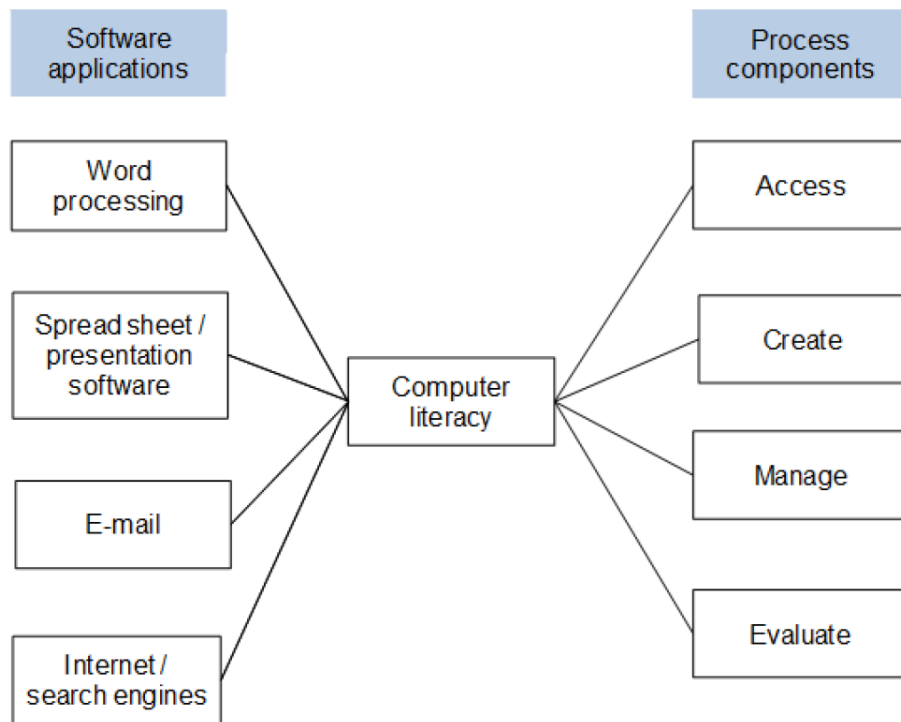


Figure 1. Assessment Framework for Computer Literacy.

The distribution of the items across the two dimensions of the assessment framework are summarized in Tables 1 and 2. Most items in the test referred to a single process component and a single software application. However, a few items also represented a combination of two process components (see Appendix A).

Table 1.

Number of Items by Different Software Applications

Software application	Number of Items
Word processing	6
Spreadsheet / presentation software	8
E-mail / communication tools	6
Internet / search engines	12
Total number of items	32

Table 2.

Number of Items by Different Process Components

Process components	Number of Items
Access	8
Create	10
Manage	11
Evaluate	8
Total number of items	32

Note. Five interactive items referred to a combination of two process components (*access, create*). The assignment of each item to the process components is summarized in Appendix A.

2.2 Item Format

The assessment of computer literacy contains two types of tasks represented in static and interactive items (see Table 3).

Table 3.

Number of Items by Different Response Formats

Response format	Number of Items
<i>Static items</i>	
Simple multiple-choice items	5
Complex multiple-choice items	13
<i>Interactive items</i>	14
Total number of items	32

The first type of tasks, that is, knowledge-based, *static items*, accounted for slightly more than half of the tasks (see Table 3) and were pencil-and-paper questions. These tasks addressed factual (e.g., computer terminology) and conceptual knowledge (classifications and principles) to examine whether respondents can deal appropriately with certain computer-based tasks. To do so, participants were presented with realistic problems embedded in a range of authentic situations. About half of the items used screenshots, for example, of an internet browser, an electronic database, or a spreadsheet as prompts (see Senkbeil et al., 2013). The static items of the computer literacy test of Starting Cohorts 4 and 6 administered in Wave 14 included two types of response formats (see Table 3). These were simple multiple choice (MC) and complex multiple choice (CMC) items. In MC items the test taker had to identify the

correct answer out of four to six response options with one option being correct and three to five response items functioning as distractors (i.e., they were incorrect). In CMC items several subtasks with two response options each (true / false) were presented. The number of subtasks of CMC items varied between two and ten. Examples of the different response formats are given in Pohl and Carstensen (2012).

The second type of tasks, which accounted for slightly less than half of the tasks (see Table 3), were *interactive items* of simulations of generic software or universal applications to complete an action. These interactive items additionally addressed procedural and strategic knowledge (e.g., planning, executing, and monitoring the problem-solving process). Respondents were required to solve specific tasks and problems by using and interacting with these simulations: These may be single-action tasks (such as opening a web-browser) or may contain a sequence of steps (such as “Save As” with a specific file name, sorting or filtering a web-based database according to one or more criteria, or navigation through a menu structure). The interactive items contained linear and nonlinear tasks. Linear tasks required the execution of at least two commands executed in a necessary prescribed sequence (e.g., opening a file from the desktop, “Save As” with a specific file name, and moving the file to another drive). Nonlinear tasks required respondents to reach a desired outcome by executing subcommands in a number of different sequences.

Neuwagen

Frau Zyglinski möchte sich einen möglichst umweltfreundlichen Neuwagen kaufen und hat eine Tabelle mit für sie wichtigen Fahrzeugdaten erstellt. Sie möchte sich diejenigen Fahrzeuge anzeigen lassen, die sowohl den kleinsten CO₂-Ausstoß haben als auch am wenigsten Kraftstoff verbrauchen. Diese Fahrzeuge möchte sie sich dann nach dem Anschaffungspreis geordnet ansehen.

Ordnen Sie die Tabelle bitte so, dass die gesuchten Fahrzeuge in der Tabelle ganz oben stehen. Markieren Sie die Fahrzeuge, die aufgrund ihrer Umweltfreundlichkeit für einen Kauf in Frage kommen, in der ganz rechten Spalte, indem Sie dort ein Häkchen setzen.

	A	B	C	D	E	F	G
	Sortieren (Werte aufst.)	Sortieren (Werte aufst.)	Sortieren (Werte aufst.)	Sortieren (Werte aufst.)	Sortieren (Werte aufst.)	Sortieren (Werte aufst.)	
	Modell	CO ₂ -Ausstoß (l/100km)	Leistung (kW)	Hubraum (cm ³)	Verbrauch (g/km)	Preis (in €)	
1	Audi A3 1.9 TDI	144	77	1896	5,2	22950	☐
2	BMW 118d	146	105	1995	6,1	24550	☐
3	BMW 740i	162	235	2979	7,9	85000	☐
4	Citroen C1	122	50	993	5,2	9390	☐
5	Citroen C3	116	81	1199	6,1	14450	☐
6	Daihatsu Cuore 1.0	118	43	989	4,8	9190	☐
7	Ford Fiesta Econetic	90	74	998	3,7	15000	☐
8	Mercedes E 220 CDI	173	135	1991	7,9	41055	☐
9	Mini Cooper D	107	85	1598	4,8	19900	☐
10	Opel Corsa 1.3	144	44	997	5,6	13800	☐
11	Peugeot 107	120	55	995	5,2	9650	☐
12	Renault Clio 1.5	148	72	1390	5,6	11550	☐
13	Renault Laguna 1.5	164	83	1461	6,1	22750	☐
14	Skoda Superb 1.9 TDI	173	103	1968	5,6	23990	☐
15	Smart Coupe CDI	90	30	799	3,7	11360	☐
16	Toyota Aygo 1.0	107	58	991	3,7	9350	☐
17	Toyota iQ	90	53	999	4,8	12700	☐
18	VW Golf 2.0 TDI	107	110	1969	4,8	20625	☐
19							
20							
21							
22							
23							
24							
25							

Figure 2. Example item of an interactive task

The example item presented in Figure 2 illustrates a nonlinear interactive task. This task required respondents to use the filtering functions of a web-based database and interpret some simple text to locate a vehicle that matches a given set of characteristics (low fuel consumption, low carbon dioxide emissions, and low cost). The web-based database contained too many objects and information for a respondent to search it manually with ease.

Therefore, the automatic scoring gave the highest level of credit to respondents who made use the filtering features (in any order) to support their search.

In summary, interactive items required respondents to execute given software commands, while others required respondents to execute commands along with some information processing (e.g., sorting or filtering a web-based database according to one or more criteria). The interactive items required following several steps to solve a specific problem in the simulated application and were scored automatically. The number of successful task completions were summed for each item to create a composite score.

2.3 Test Design

The computer literacy test administered in the present study adopted a multi-stage design that included a total of 34 items (see Figure 3). However, each respondent received only a subset of 20 items depending on his or her estimated computer literacy. The multi-stage test (MST) included two distinct stages that were sequentially administered to each respondent. Each stage consisted of different levels that included items of different difficulty. Lower levels (e.g., Level 1) included easier items, whereas higher levels (e.g., Level 2) included more difficult items. The difficulty of each item was estimated from previously published competence data in different starting cohorts and unpublished data from various developmental studies. Stage 1 included two levels (*easy*, *difficult*) with 12 static items in each block, whereas Stage 2 had three levels (*easy*, *medium*, *difficult*) with 8 interactive items in each block (see Figure 3). The level at the first stage was assigned to each respondent based on his or her estimated computer literacy in the previous wave (see Senkbeil & Ihme, 2015, 2017). Respondents with a competence score falling below the mean competence of the sample received the easy level in Stage 1, whereas respondents with competence scores exceeding the mean of the sample were administered the difficult level in Stage 1. Depending on their responses in Stage 1, each respondent received one of three levels from Stage 2.

The MST tried to respect the theoretical testing framework (see above) by distributing the different process components and software applications approximately uniformly across the different levels within each stage. To ensure a common scale for all test takers, we linked the different levels within each stage by assigning some items to two adjacent levels. For example, in Stage 1 item *ica5017s_sc4a14_c* was assigned to Level 1 and to Level 2. In this case, item *ica5017s_sc4a14_c* acts as link between the two levels and allows for the estimation of a common score for respondents receiving different levels. There was no multi-matrix design regarding the order of the items within a specific level. All respondents assigned to a specific level received the test items in the same order.

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

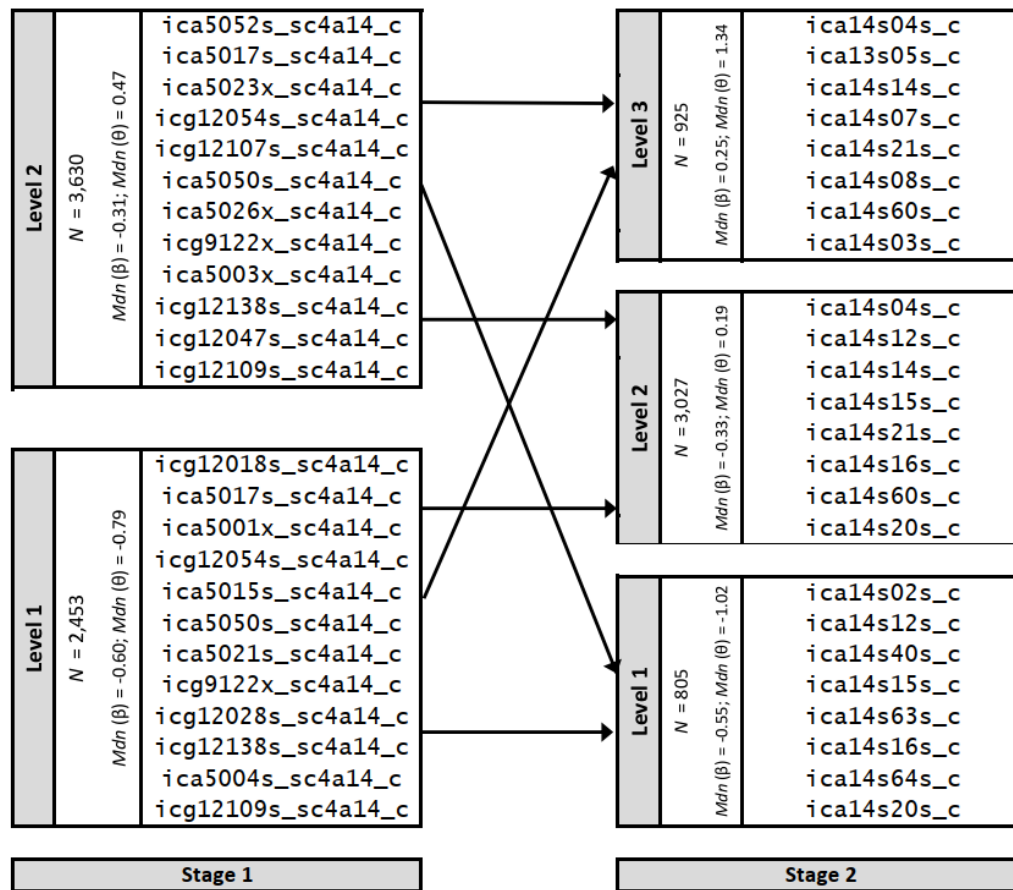


Figure 3. Multi-Stage Test Design with Items, Sample Sizes (N), Median Item Difficulties (β), and Median Respondent Proficiencies (θ) for Each Block.

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

3. Data

A total of 6,083 students (50% women) had at least three valid responses on the computer literacy test and, thus, were used for the psychometric analyses (see Pohl & Carstensen, 2012). Of these, $N = 2,491$ (51% women) were from Starting Cohort 4 and $N = 3,592$ (49% women) were from Starting Cohort 6. The age of the respondents ranged from 25 to 76 years. Basic sociodemographic information of the two samples is summarized in Table 4.

Table 4.

Sample Descriptions

	Starting Cohort 4	Starting Cohort 6
Sample size	2,491	3,592
Percentage women	51%	49%
Percentage migration background	9%	7%
Mean age (<i>SD</i>)	26.56 (1.34)	57.90 (10.02)

4. Analyses

4.1 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) omitted items, b) items that test takers did not reach within a stage, c) items that were not presented in the second stage because test takers prematurely aborted the test in the first stage, d) items that have not been administered, and, finally, e) multiple kinds of missing responses within complex multiple-choice items that are not determined. Omitted items occurred when test takers skipped some items. Due to time limits or lack of motivation, not all persons finished the test. All missing responses after the last valid response within a level of the current stage were coded as not reached, whereas the remaining items in the following stages were coded as missing due to test abortion. Because of the MST design, many items were not administered to any given participant. These items were missing by design. Because complex multiple-choice and interactive items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A complex multiple choice or interactive item was coded as missing if at least one subtask contained a missing response. If just one kind of missing response occurred, the item was coded according to the corresponding missing response. If the subtasks contained different kinds of missing responses, the item was labeled as a not determinable missing response.

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

4.2 Scaling Model

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

Complex multiple-choice and interactive items consisted of a set of subtasks that were aggregated to a polytomous variable for each item, indicating the number of correctly solved subtasks within that item. If at least one of the subtasks contained a missing response, the partial credit item was scored as missing. Response categories of polytomous variables with less than $N = 200$ responses were collapsed in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items; in these cases, the lower categories were collapsed into one category.

Respondents' computer literacies were estimated as weighted likelihood estimates (WLE; Warm, 1989). To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while dichotomous items were scored as 0 for an incorrect and 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in Section 3.

4.3 Checking the Quality of the Test

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

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

Before aggregating the subtasks of a complex multiple-choice and interactive item to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the multiple-choice items in a Rasch (1960) model. The fit of the subtasks was evaluated based on the weighted mean square (WMNSQ), the respective t -value, and the item characteristic curves. Only if the subtasks exhibited a satisfactory item fit, they were used to construct polytomous variables that were included in the final scaling model.

After aggregating the subtasks to polytomous variables, the fit of the dichotomous and polytomous items to the PCM (Masters, 1982) was evaluated using the weighted mean square (WMNSQ) statistic, the respective t -value, and the item characteristic curves (see Pohl & Carstensen, 2012). Items with a $WMNSQ > 1.15$ (t -value $> |6|$) were considered as having a noticeable item misfit, and items with a $WMNSQ > 1.20$ (t -value $> |8|$) were judged as having a considerable item misfit and their performance was further investigated. Correlations of the item score with the corrected total score 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 computer literacy test should measure the same construct for all students. If some items favored certain subgroups (e.g., they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., males and females) would be biased. For the present study, measurement invariance was investigated for the variables sex, age, the number of books at home (as a proxy for cultural status), migration background (see Pohl & Carstensen, 2012, for a description of these variables), starting cohort (4 or 6), and test rotation (first versus second). Differential item functioning (DIF) was examined using a multigroup item response model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties between the subgroups that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as considerable and noteworthy of further investigation, differences between 0.4 and 0.6 as small but not severe, and differences smaller than 0.4 as negligible DIF. Minimum hypothesis tests (see Fischer, Rohm, Gnambs, & Carstensen, 2016) were used to statistically test whether the observed differences were significantly larger than 0.4 and, thus, were at least small in size. Additionally, measurement invariance was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

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

The dimensionality of the test was evaluated by examining the residuals of the PCM. Approximately zero-order correlations as indicated by Yen's (1984) Q3 indicate unidimensionality. Because in case of locally independent items, the Q3 statistic tends to be slightly negative, we report the corrected Q3 that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) values of Q3 falling below .20 indicate essential unidimensionality. Moreover, we examined a multidimensional model based on the five process components and software applications (see Appendix A) using quasi-Monte Carlo integration (with 2,000 nodes). The correlations between the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the test.

4.4 Software

The item response models were estimated with the TAM package version 4.1-4 (Robitzsch, Kiefer, & Wu, 2022) in R version 4.3.1 (R Core Team, 2023).

5. Results

Preliminary analyses showed that one subtask of the CMC item `icg12109a_sc4a14_c` and three subtasks of the interactive item `ica14s16s_c` exhibited a poor fit to the item response model. Therefore, these subtasks were removed from the item scoring. Moreover, the interactive item `ica14s63s_c` exhibited substantial misfit in Starting Cohort 6 and, consequently, also

differential item functioning between the two starting cohorts. Therefore, this item was excluded from Starting Cohort 6 and only considered in Starting Cohort 4. Because a technical error in the testing system prevented the proper recording of responses to two items (ica14s02s_c, ica14s04s_c), these items had to be entirely excluded from the scaling procedure. Thus, the analyses presented in the following sections and the competence scores derived for the respondents are based on the remaining 32 items.

5.1 Missing Responses

5.1.1 Missing responses per person

The number of missing responses when participants omitted items is illustrated in Figure 4. The figure highlights that there were pronounced differences between the two starting cohorts, with the adult sample (Starting Cohort 6) exhibiting substantially more omitted items. Only about 27% of respondents in Starting Cohort 6 had no omitted item, whereas 18% exhibited a single omitted item. In contrast, in Starting Cohort 4 about 59% of respondents had no omitted item, while about 22% had a single omitted item.

Another source of missing responses is items that were not reached within a stage by the respondents because they aborted the test, for example, because the time limit was reached or a lack of motivation. These missing values refer to items after the last valid response. As illustrated in Figure 5, only about 49% of the respondents did not abort the test and were administered all 19 items. However, more than 90% of the sample received at least 10. This indicates that some items might have been too complex for some respondents. The results given in Figure 3 also show that this missing rates were substantially higher in the older sample as compared to the younger sample. In Starting Cohort 4 about 63% of the respondents did not prematurely abort the test, whereas in Starting Cohort 6 this was the case for only about 39% of the respondents.

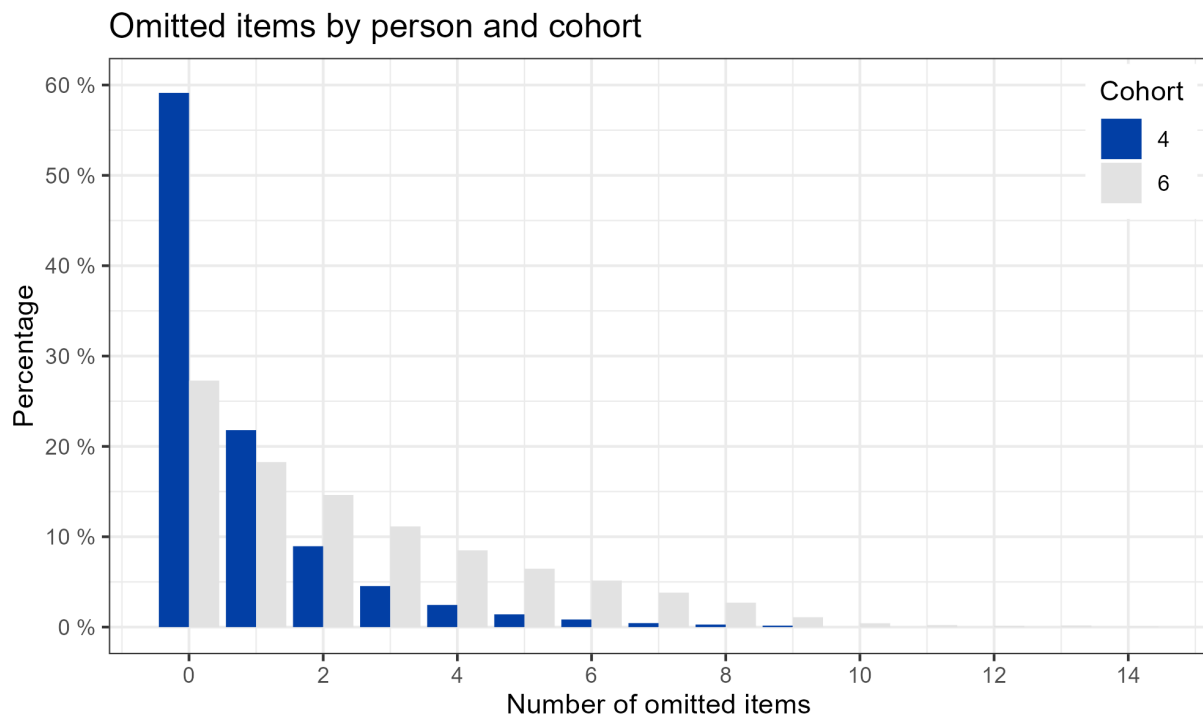


Figure 4. Number of Omitted Items by Starting Cohort.

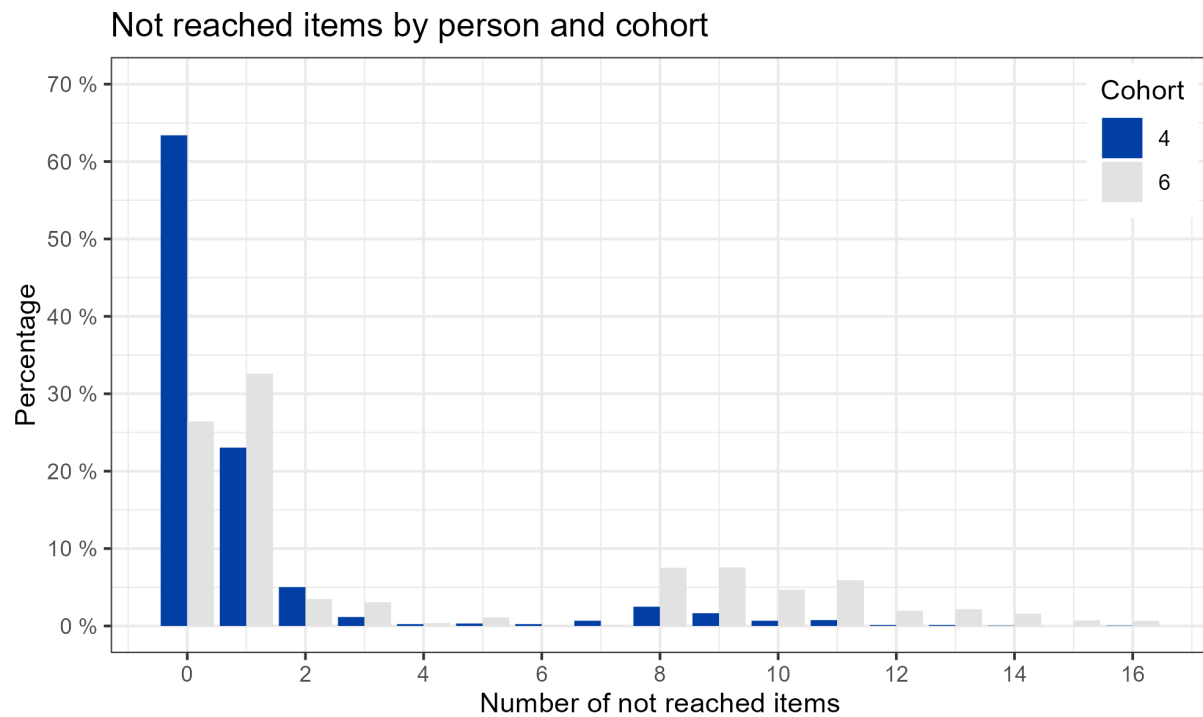


Figure 5. Number of not reached items by sample

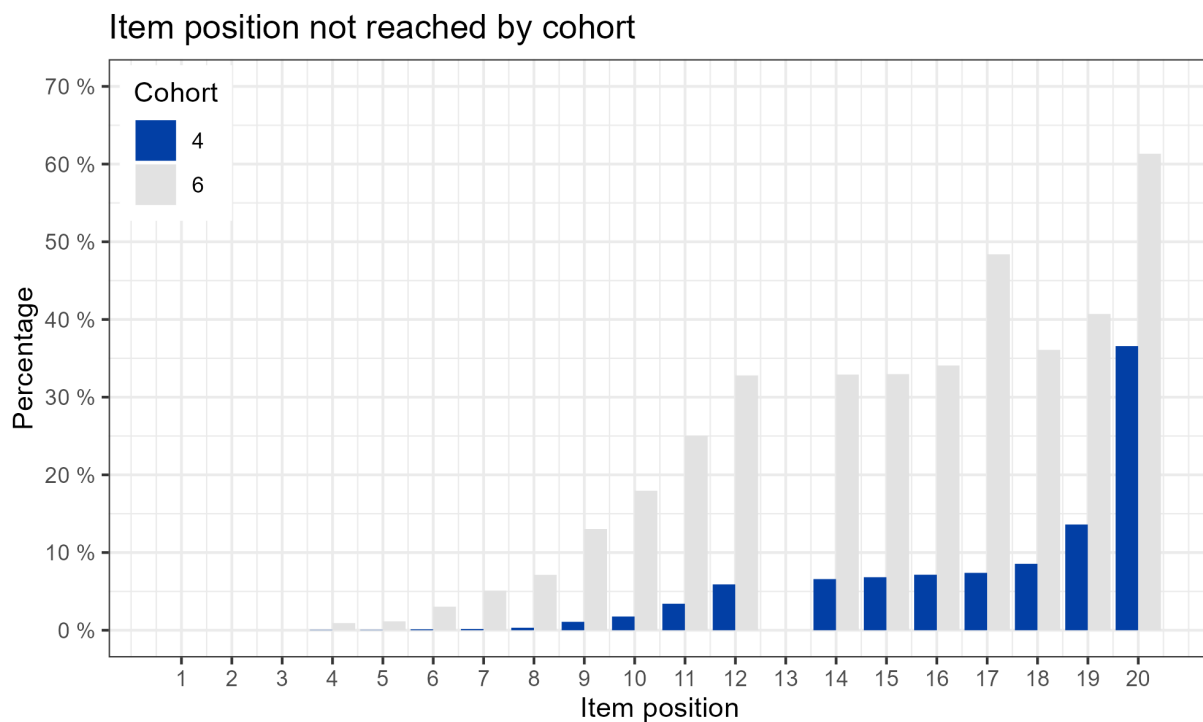


Figure 6. Item position not reached by sample.

Note. The items at position 13 were excluded from the analyses because of a technical error in the testing system.

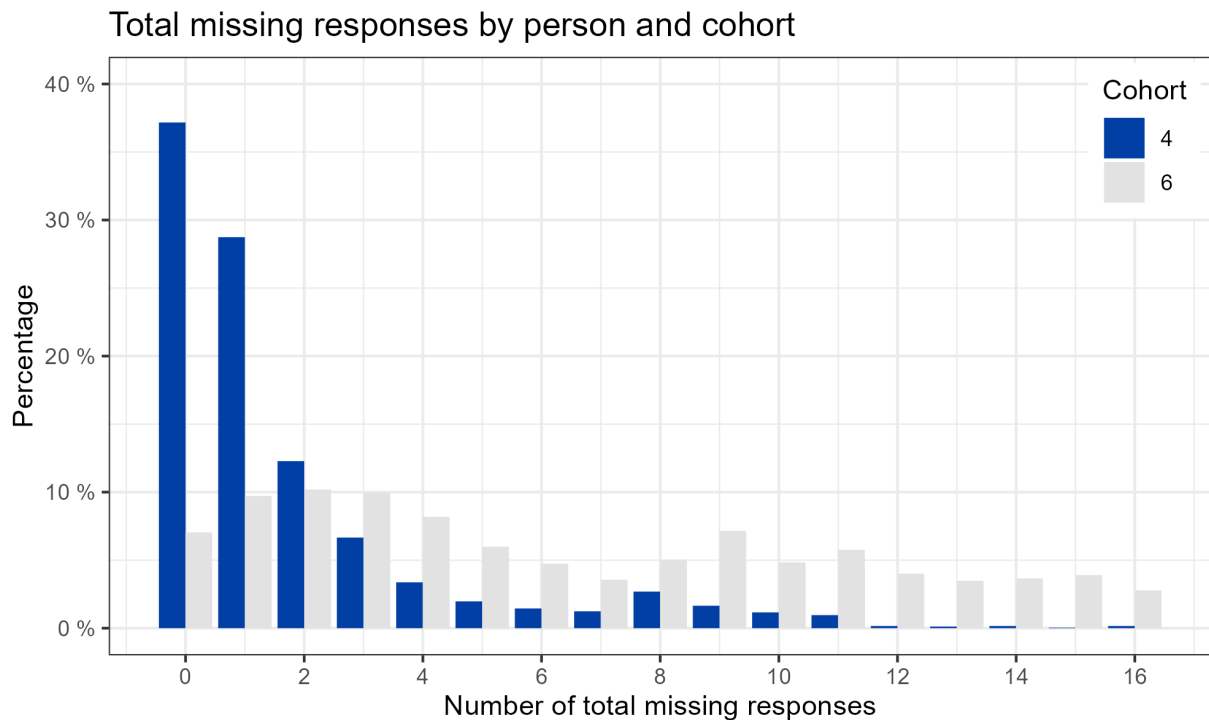


Figure 7. Total number of missing responses by sample.

With an item's progressing position in the test, the amount of persons that did not reach an item steadily increased. Particularly, the last item of the test was reached by only few respondents. As illustrated in Figure 6, in the adult sample about a third of the respondents did not reach the second half of the test. This might indicate that the interactive items that were presented in the second stage of the MST might have been too challenging for them. In contrast, respondents in Starting Cohort 4 seemed to have less problems and also finished most of the interactive items.

The total number of missing responses, aggregated over omitted and not reached missing responses per person, is illustrated in Figure 7. Because the majority of the sample did not reach the end of the test, there was a substantial number of missing values. The median number of missing responses was 3; only about 20% of the respondents had no missing response at all. Again, in Starting Cohort 6 fewer participants had no missing value (about 8%) as compared to Starting Cohort 4 (about 37%).

In sum, the amount of missing responses was rather large in Starting Cohort 6 because many respondents did not reach the end of the test. In contrast, respondents from Starting Cohort 4 exhibited more valid responses.

5.1.2 Missing responses per item

Table 5 provides information on the occurrence of different kinds of missing responses per item for the two samples. The percentage of omitted responses varied across items between 0% and 40% ($Mdn = 2\%$) in Starting Cohort 4 and between 0% and 66% ($Mdn = 12\%$) in Starting Cohort 6. In contrast, the percentage of missing responses because participants did not reach the item within a stage (i.e., excluding missing responses resulting from test abortion) were $Mdn = 1\%$ and 2% in the two starting cohorts.

Table 5.

Percentage of Missing Values by Item.

Item	Pos.	Stage	Starting Cohort 4				Starting Cohort 6			
			<i>N</i>	<i>N_v</i>	OM	NR	<i>N</i>	<i>N_v</i>	OM	NR
ica5052s_sc4a14_c	1	1	1788	1695	5.20	0.00	1842	1568	14.88	0.00
icg12018s_sc4a14_c	1	1	703	689	1.99	0.00	1750	1587	9.31	0.00
ica5017s_sc4a14_c	2	1	2491	2360	5.26	0.00	3592	3032	15.59	0.00
ica5001x_sc4a14_c	3	1	703	701	0.28	0.00	1750	1686	3.66	0.00
ica5023x_sc4a14_c	3	1	1788	1788	0.00	0.00	1842	1827	0.81	0.00
icg12054s_sc4a14_c	4	1	2491	2356	5.38	0.04	3592	2513	29.12	0.92
ica5015s_sc4a14_c	5	1	703	666	5.12	0.14	1750	1331	21.60	2.34
icg12107s_sc4a14_c	5	1	1788	1727	3.41	0.00	1842	1557	15.47	0.00
ica5050s_sc4a14_c	6	1	2491	2430	2.33	0.12	3592	2986	13.84	3.04
ica5021s_sc4a14_c	7	1	703	658	5.83	0.57	1750	1221	20.00	10.23
ica5026x_sc4a14_c	7	1	1788	1775	0.73	0.00	1842	1715	6.68	0.22
icg9122x_sc4a14_c	8	1	2491	2454	1.16	0.32	3592	3102	6.51	7.13
ica5003x_sc4a14_c	9	1	1788	1728	3.36	0.00	1842	1474	18.78	1.19
icg12028s_sc4a14_c	9	1	703	643	4.69	3.84	1750	1020	16.23	25.49
icg12138s_sc4a14_c	10	1	2491	2303	5.78	1.77	3592	2225	20.10	17.96
ica5004s_sc4a14_c	11	1	703	610	3.41	9.82	1750	795	6.29	48.29
icg12047s_sc4a14_c	11	1	1788	1436	18.79	0.90	1842	762	55.70	2.93
icg12109s_sc4a14_c	12	1	2491	1985	14.41	5.90	3592	1346	29.73	32.80
ica14s05s_c	14	2	708	699	0.00	1.27	217	215	0.00	0.92
ica14s12s_c	14	2	1636	1628	0.00	0.49	2196	2195	0.00	0.05
ica14s14s_c	15	2	2060	2040	0.00	0.97	1892	1889	0.00	0.16
ica14s40s_c	15	2	284	169	39.44	1.06	521	177	65.64	0.38
ica14s07s_c	16	2	708	696	0.00	1.70	217	214	0.00	1.38
ica14s15s_c	16	2	1636	1286	20.23	1.16	2196	1278	39.89	1.91
ica14s21s_c	17	2	2060	2030	0.00	1.46	1892	1854	0.00	2.01
ica14s63s_c	17	2	284	272	1.76	2.47	521	436	14.20	2.11
ica14s08s_c	18	2	708	687	0.00	2.97	217	211	0.00	2.77
ica14s16s_c	18	2	1636	1591	0.00	2.75	2196	2085	0.00	5.06
ica14s60s_c	19	2	2060	1900	0.05	7.72	1892	1680	0.00	11.21
ica14s64s_c	19	2	284	226	8.80	11.62	521	327	23.61	13.63
ica14s03s_c	20	2	708	404	0.00	42.94	217	111	0.00	48.85
ica14s20s_c	20	2	1636	1176	0.00	28.12	2196	1278	0.00	41.80

Note. Pos. = Item position within the test. *N* = Number of respondents the item was administered to. *N_v* = Number of valid responses. NR = Percentage of respondents that did not reach an item. OM = Percentage of respondents that omitted the item. The items on position 13 were excluded from the analyses because of a technical error in the testing system. Item names refer to Starting Cohort 4; the corresponding variable names for Starting Cohort 6 are given in Appendix A.

5.2 Parameter Estimates

5.2.1 Item Parameters

The fourth column in Table 6 presents the percentage of correct responses in relation to all valid responses for each multiple-choice item. The percentage of correct responses varied between 58% and 70% correct responses and, thus, was rather high. Because there was a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index of item difficulty. The estimated item difficulties (for dichotomous items) and location parameters (for polytomous items) are given in the fifth column of Table 6. The step parameters for polytomous variables are summarized in Table 7. The item difficulties and location parameters were estimated by constraining the mean of the ability distribution to zero. The standard errors (SE) of most estimated parameters (see Tables 6 and 7) were rather small for most items. The estimated item difficulties ranged from -1.62 (item ica14s15s_c) to 2.38 (item ica14s07s_c) with a median of -0.34 and, thus, covered a rather wide range including easy as well as difficult items.

Table 6.

Item Parameters for Combined Sample.

	Item	N	Percentage correct	Difficulty	SE	WMNSQ	t	Discr.	αQ_3
1	ica5052s_sc4a14_c	3263		-0.31	0.02	1.01	0.75	0.52	0.04
2	icg12018s_sc4a14_c	2276		-1.15	0.03	0.97	-1.05	0.70	0.08
3	ica5017s_sc4a14_c	5392		-0.70	0.02	0.98	-1.46	0.61	0.03
4	ica5001x_sc4a14_c	2387	69.96	-1.54	0.05	1.03	1.33	0.92	0.09
5	ica5023x_sc4a14_c	3615	62.79	-0.25	0.04	0.98	-1.46	1.25	0.08
6	icg12054s_sc4a14_c	4869		-0.64	0.02	0.88	-7.03	1.02	0.05
7	ica5015s_sc4a14_c	1997		-0.28	0.02	1.16	5.46	0.25	0.16
8	icg12107s_sc4a14_c	3284		0.18	0.02	1.04	2.23	0.39	0.02
9	ica5050s_sc4a14_c	5416		-0.63	0.02	0.93	-4.29	0.88	0.05
10	ica5021s_sc4a14_c	1879		-0.56	0.03	0.89	-4.79	1.33	0.11
11	ica5026x_sc4a14_c	3490	58.74	-0.01	0.04	1.06	3.97	0.82	0.08
12	icg9122x_sc4a14_c	5556	62.71	-0.58	0.03	0.99	-0.97	1.22	0.03
13	ica5003x_sc4a14_c	3202	57.75	0.08	0.04	1.12	8.05	0.54	0.07
14	icg12028s_sc4a14_c	1663		-0.87	0.04	1.03	1.23	0.46	0.09
15	icg12138s_sc4a14_c	4528		-0.32	0.02	1.01	0.78	0.54	0.03
16	ica5004s_sc4a14_c	1405		-0.34	0.04	0.91	-3.33	0.98	0.14
17	icg12047s_sc4a14_c	2198		0.12	0.02	1.02	0.90	0.53	0.08
18	icg12109s_sc4a14_c	3331		-0.44	0.02	1.05	2.46	0.44	0.04
19	ica14s05s_c	914		0.92	0.04	1.02	0.73	0.49	0.12
20	ica14s12s_c	3823		-0.33	0.01	0.99	-0.58	0.59	0.05
21	ica14s14s_c	3929		-0.19	0.03	0.99	-0.60	0.64	0.03
22	ica14s40s_c	346		-1.58	0.13	0.98	-0.29	1.27	0.22
23	ica14s07s_c	910		2.38	0.08	1.01	0.29	1.25	0.14
24	ica14s15s_c	2564		-1.62	0.06	1.09	2.56	0.31	0.04
25	ica14s21s_c	3884		-0.11	0.02	0.92	-4.11	0.94	0.04
26	ica14s63s_c	272		-0.51	0.13	0.94	-1.32	1.84	0.06
27	ica14s08s_c	898		0.18	0.08	1.04	1.19	0.61	0.24
28	ica14s16s_c	3676		-0.55	0.02	0.91	-5.27	0.94	0.04
29	ica14s60s_c	3580		0.25	0.01	1.19	8.62	0.28	0.08
30	ica14s64s_c	553		-0.89	0.09	1.02	0.57	1.03	0.06
31	ica14s03s_c	515		0.68	0.06	1.04	1.04	0.23	0.04
32	ica14s20s_c	2454		-0.37	0.03	0.92	-3.99	0.87	0.05

Note. N = Number of observed responses; Difficulty = Item difficulty / location, SE = Standard error of item difficulty / location, WMNSQ = Weighted mean square, t = t-value for WMNSQ, Discr. = Discrimination parameter of a generalized partial credit model, αQ_3 = Average absolute residual correlation for item (Yen, 1983). Percent correct scores are not informative for polytomous item scores and, therefore, are not reported. Item names refer to Starting Cohort 4; the corresponding variable names for Starting Cohorts 4 are given in Appendix A.

Table 7.

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

Item	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
ica5052s_sc4a14_c	-0.08 (0.04)	0.08				
icg12018s_sc4a14_c	-0.23 (0.05)	0.23				
ica5017s_sc4a14_c	-0.50 (0.03)	0.50				
icg12054s_sc4a14_c	-0.49 (0.03)	0.27 (0.03)	0.22			
ica5015s_sc4a14_c	-0.85 (0.05)	-0.45 (0.05)	0.03 (0.05)	1.27		
icg12107s_sc4a14_c	-0.60 (0.04)	0.60				
ica5050s_sc4a14_c	-0.70 (0.03)	0.70				
ica5021s_sc4a14_c	-0.54 (0.05)	0.54				
icg12028s_sc4a14_c	-0.46 (0.05)	0.46				
icg12138s_sc4a14_c	0.01 (0.03)	0.00				
ica5004s_sc4a14_c	-0.38 (0.06)	0.38				
icg12047s_sc4a14_c	-0.23 (0.05)	-0.56 (0.05)	-0.15 (0.04)	0.22 (0.05)	0.72	
icg12109s_sc4a14_c	0.01 (0.04)	-0.01				
ica14s05s_c	0.41 (0.08)	-0.41				
ica14s12s_c	-0.33 (0.04)	0.26 (0.03)	-0.62 (0.03)	-0.11 (0.04)	0.81	
ica14s14s_c	-1.63 (0.03)	1.63				
ica14s21s_c	-1.22 (0.04)	-0.18 (0.03)	1.40			
ica14s16s_c	0.67 (0.04)	-0.67				
ica14s60s_c	0.59 (0.0354)	-0.75 (0.04)	0.30 (0.04)	-0.01 (0.04)	0.25 (0.06)	-0.38
ica14s03s_c	-0.18 (0.09)	0.18				
ica14s20s_c	0.38 (0.05)	-0.38				

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. Item names refer to Starting Cohort 4; the corresponding variable names for Starting Cohorts 6 are given in Appendix A.

5.2.2 Test targeting and Reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. Because some items in the computer literacy test were polytomous, we calculated Thurstonian thresholds for each response category (Wu et al., 2007). These indicate the location at the latent dimension at which the probability of achieving a score above the respective threshold is 50%. Thus, it is similar to the item difficulties of dichotomous items. In Figure 6, the item difficulties of the computer literacy items and the ability of the respondents are plotted on the same scale. The distribution of the respondents' estimated abilities is mapped onto the left side whereas the right side shows the distribution of the category thresholds. The respective thresholds ranged from -3.72 (item ica14s14s_c) to 2.96 (item ica14s14s_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 1.27, which implies good differentiation between respondents. The reliability of the test (EAP/PV reliability = .75, WLE reliability = .71) was acceptable. The median of the category threshold distribution was about 0.29 logits below the mean person ability distribution. Thus, although the items covered a wide range of the ability distribution, on average, the items were slightly too easy for the participants. As a consequence, proficiency estimates in low- and medium-ability regions will be measured

relative precisely, whereas higher ability estimates will have larger standard errors of measurement.

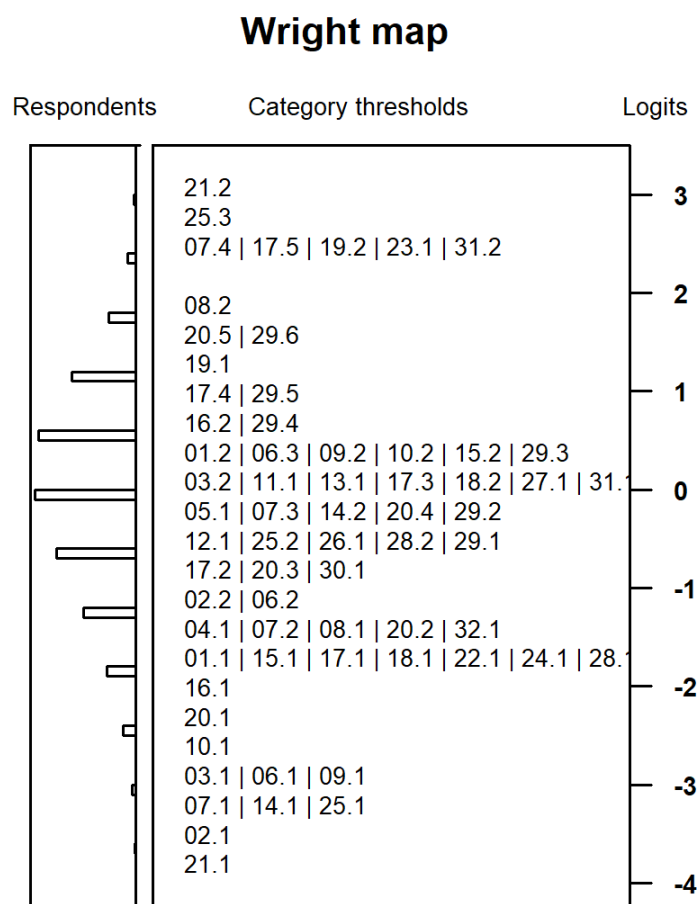


Figure 8. Test targeting.

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

5.3 Quality of the test

5.3.1 Distractor analyses

To investigate how well the distractors of the multiple-choice items performed the point-biserial correlations between each incorrect response (distractor) and the respondents' total correct scores for the remaining items were calculated. The median point-biserial correlations for the distractors fell at $-.13$ ($Min = -.28$, $Max = -.04$). In contrast, the correlations of the correct responses with the total scores varied between $.20$ and $.39$ ($Mdn = .33$). These results indicate that the distractors functioned well.

5.3.2 Item fit

The evaluation of item fit was based on the final scaling model presented above. Again, the test quality was examined for the combined sample including respondents from both starting cohorts. Altogether item fit was satisfactory (see Table 6). No item exhibited a WMSNQ greater than 1.20, while two items exhibited a WMSNQ greater than 1.15 (ica14s60s_c,

ica5015s_sc4a14_c). However, a visual inspection of the ICCs showed no pronounced deviation from the expected ICC for these items. For the remaining items, values of the WMNSQ ranged from 0.88 (item icg12054s_sc4a14_c) to 1.12 (item ica5003x_sc4a14_c), with a median of 1.01.

5.3.3 Differential item functioning

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

Sex: The sample included 3,059 men and 3,018 women. There were no substantial gender differences in computer literacy as indicated by the main effect of 0.11 logits (Cohen’s $d = 0.10$). Although one item showed small DIF greater than 0.40 logits (ica5023x_sc4a14_c), no item exhibited considerable DIF greater than 0.60 logits. However, an overall test for DIF (see Table 9) using Akaike’s (1974) information criterion (AIC) and the Bayesian information criterion (BIC; Schwarz, 1978) showed that the DIF model exhibited a better fit to the data as compared to a model that only estimated the main effect (but ignored potential DIF). But, the estimated main effects for sex were nearly identical in both models. These results indicate that there was no pronounced DIF concerning sex that might have distorted the parameter estimates.

Migration background: There were 5,593 respondents without migration background and 472 respondents with a migration background. In comparison to participants with a migration background, participants without a migration background had, on average, a slightly higher computer literacy (main effect = 0.11 logits, Cohen’s $d = 0.10$). No items exhibited a noteworthy DIF greater than 0.6 logits. Consequently, the overall test for DIF using the AIC and BIC also favored the main effect model that did not include item-level DIF (see Table 9).

Books: The number of books at home was used as an indicator of respondents’ cultural capital. There were 1,631 respondents with 0 to 100 books at home and 4,337 respondents with more than 100 books at home. On average, participants with lower cultural capital performed on average -0.26 logits (Cohen’s $d = -0.23$) worse on the computer literacy test as compared to respondents with higher cultural capital. There were no item with considerable DIF. Consequently, the overall test for DIF using the BIC also favored the main effect model that did not include item-level DIF (see Table 9).

Age group: There were 3,247 respondents younger than 50 years and 2,836 respondents that were older. In comparison to younger participants, older participants had, on average, a higher computer literacy (main effect = 0.70 logits, Cohen’s $d = 0.76$). No items exhibited a strong DIF greater than 1.0 logits, but two items (ica5021s_sc4a14_c, ica14s63s_c) had DIF exceeding 0.6. Consequently, the overall test for DIF using the AIC and BIC also favored the

DIF model (see Table 9). But, the estimated main effects for age was only hardly affected when acknowledging DIF (main effect = 0.72 logits, Cohen's $d = 0.76$).

Test position: There were 3,056 respondents received the computer literacy test first before working on the test of scientific literacy, whereas 3,027 finished the tests in the reversed order. There was no difference in computer literacy between the two groups (main effect = 0.00 logits, Cohen's $d = 0.00$). Moreover, no item exhibited a noteworthy DIF. Consequently, also the overall test for DIF using the BIC favored the main effects model that ignored potential DIF (see Table 9).

Starting cohort: Starting Cohort 4 included 3,491 respondents, whereas Starting Cohort 6 was comprised of 3,592 participants. On average, respondents in Starting Cohort 4 had a higher computer literacy (main effect = 0.59 logits, Cohen's $d = 0.60$). Two items exhibited a noteworthy DIF greater than 0.60 (ica5021s_sc4a14_c, ica5052s_sc4a14_c). Also, the overall test for DIF using the AIC and BIC favored the main effects model that ignored potential DIF (see Table 9). However, the estimated main effects for the starting cohort was only hardly affected when acknowledging DIF (main effect = 0.66 logits, Cohen's $d = 0.66$).

Table 8.

Differential Item Functioning

Item	Sex	Migration	Books	Age	Test position	Starting cohort
	men vs. women	without vs. with	<100 vs. >100	<50 vs. >50	first vs. second	4 vs. 6
ica5052s_sc4a14_c	-0.11 (-0.10)	-0.07 (-0.06)	-0.01 (-0.01)	0.60 (0.64)	0.12 (0.11)	0.67 (0.67)
icg12018s_sc4a14_c	0.20 (0.18)	-0.08 (-0.07)	0.05 (0.05)	0.11 (0.12)	0.30 (0.27)	0.01 (0.01)
ica5017s_sc4a14_c	0.13 (0.12)	-0.10 (-0.09)	0.04 (0.04)	0.29 (0.31)	0.06 (0.05)	0.32 (0.32)
ica5001x_sc4a14_c	0.00 (0.00)	0.03 (0.03)	0.28 (0.25)	-0.51 (-0.54)	0.06 (0.05)	-0.57 (-0.57)
ica5023x_sc4a14_c	-0.55 (-0.49)	0.16 (0.14)	0.19 (0.17)	-0.59 (-0.62)	-0.04 (-0.04)	-0.53 (-0.53)
icg12054s_sc4a14_c	-0.16 (-0.14)	0.09 (0.08)	-0.05 (-0.05)	-0.55 (-0.58)	-0.05 (-0.04)	-0.56 (-0.56)
ica5015s_sc4a14_c	-0.20 (-0.18)	-0.16 (-0.14)	-0.10 (-0.09)	0.39 (0.41)	-0.04 (-0.04)	0.39 (0.39)
icg12107s_sc4a14_c	0.08 (0.07)	0.03 (0.03)	-0.03 (-0.03)	0.17 (0.18)	-0.02 (-0.02)	0.10 (0.10)
ica5050s_sc4a14_c	-0.33 (-0.29)	-0.17 (-0.15)	0.04 (0.04)	-0.02 (-0.02)	-0.01 (-0.01)	0.14 (0.14)
ica5021s_sc4a14_c	-0.14 (-0.12)	0.27 (0.24)	-0.07 (-0.06)	-0.79 (-0.84)	-0.01 (-0.01)	-0.90 (-0.90)
ica5026x_sc4a14_c	-0.15 (-0.13)	-0.09 (-0.08)	0.08 (0.07)	-0.02 (-0.02)	-0.07 (-0.06)	0.04 (0.04)
icg9122x_sc4a14_c	0.20 (0.18)	0.16 (0.14)	-0.19 (-0.17)	0.46 (0.49)	0.06 (0.05)	0.39 (0.39)
ica5003x_sc4a14_c	0.27 (0.24)	-0.14 (-0.12)	0.20 (0.18)	-0.02 (-0.02)	-0.01 (-0.01)	0.11 (0.11)
icg12028s_sc4a14_c	0.01 (0.01)	0.13 (0.12)	-0.23 (-0.21)	0.25 (0.26)	0.14 (0.12)	0.26 (0.26)
icg12138s_sc4a14_c	-0.13 (-0.12)	-0.02 (-0.02)	-0.08 (-0.07)	0.28 (0.30)	-0.02 (-0.02)	0.43 (0.43)
ica5004s_sc4a14_c	-0.10 (-0.09)	-0.01 (-0.01)	0.12 (0.11)	-0.12 (-0.13)	0.11 (0.10)	-0.02 (-0.02)
icg12047s_sc4a14_c	0.22 (0.20)	-0.12 (-0.11)	-0.09 (-0.08)	-0.17 (-0.18)	-0.06 (-0.05)	-0.06 (-0.06)
icg12109s_sc4a14_c	0.07 (0.06)	-0.24 (-0.21)	-0.03 (-0.03)	0.23 (0.24)	0.02 (0.02)	0.19 (0.19)
ica14s05s_c	-0.04 (-0.04)	0.03 (0.03)	-0.09 (-0.08)	0.45 (0.48)	-0.04 (-0.04)	0.42 (0.42)
ica14s12s_c	0.12 (0.11)	-0.03 (-0.03)	-0.12 (-0.11)	-0.01 (-0.01)	0.12 (0.11)	0.04 (0.04)
ica14s14s_c	-0.04 (-0.04)	-0.22 (-0.20)	-0.12 (-0.11)	0.12 (0.13)	-0.20 (-0.18)	0.18 (0.18)
ica14s40s_c	0.18 (0.16)	0.34 (0.30)	0.02 (0.02)	0.17 (0.18)	-0.38 (-0.34)	-0.21 (-0.21)
ica14s07s_c	0.15 (0.13)	0.36 (0.32)	0.07 (0.06)	0.02 (0.02)	-0.02 (-0.02)	-0.42 (-0.42)
ica14s15s_c	-0.04 (-0.04)	-0.41 (-0.37)	-0.01 (-0.01)	0.24 (0.25)	0.29 (0.26)	0.28 (0.28)
ica14s21s_c	0.26 (0.23)	-0.03 (-0.03)	0.01 (0.01)	-0.22 (-0.23)	-0.04 (-0.04)	-0.08 (-0.08)
ica14s63s_c	0.37 (0.33)	0.08 (0.07)	0.02 (0.02)	0.73 (0.77)	-0.12 (-0.11)	0.52 (0.52)
ica14s08s_c	0.05 (0.04)	0.44 (0.39)	-0.32 (-0.29)	0.11 (0.12)	0.09 (0.08)	0.14 (0.14)
ica14s16s_c	-0.02 (-0.02)	0.05 (0.04)	-0.12 (-0.11)	-0.53 (-0.56)	-0.01 (-0.01)	-0.47 (-0.47)
ica14s60s_c	0.08 (0.07)	-0.13 (-0.12)	-0.18 (-0.16)	0.06 (0.06)	-0.04 (-0.04)	0.06 (0.06)
ica14s64s_c	-0.14 (-0.12)	-0.09 (-0.08)	0.44 (0.40)	-0.34 (-0.36)	-0.07 (-0.06)	-0.33 (-0.33)
ica14s03s_c	0.18 (0.16)	0.30 (0.27)	-0.06 (-0.05)	0.49 (0.52)	0.00 (0.00)	0.50 (0.50)
ica14s20s_c	0.00 (0.00)	-0.03 (-0.03)	-0.04 (-0.04)	-0.34 (-0.36)	0.02 (0.02)	-0.28 (-0.28)
Main effects:						
DIF model	0.14 (0.13)	0.05 (0.04)	-0.36 (-0.32)	0.72 (0.76)	0.01 (0.01)	0.66 (0.66)
Main effect model	0.11 (0.10)	0.11 (0.10)	-0.26 (-0.23)	0.70 (0.76)	0.00 (0.00)	0.59 (0.60)

Note. Raw differences between item difficulties (in logits) with standardized differences (Cohen's *d*) in parentheses.

Table 9.

Comparisons of Models with and without DIF

DIF variable	Model	N	Deviance	Number of parameters	AIC	BIC
Sex	Main effect	6077	172750	74	172898	173394
	DIF model	6077	172402	105	<u>172612</u>	<u>173317</u>
Migration	Main effect	6065	172433	74	<u>172581</u>	<u>173077</u>
	DIF model	6065	172390	105	172600	173305
Books	Main effect	5968	169142	74	169290	<u>169786</u>
	DIF model	5968	169038	105	<u>169248</u>	169951
Age group	Main effect	6083	171104	74	171252	171749
	DIF model	6083	170032	105	<u>170242</u>	<u>170946</u>
Test position	Main effect	6083	172969	74	173117	<u>173614</u>
	DIF model	6083	172876	105	<u>173086</u>	173791
Starting cohort	Main effect	6083	171673	74	171821	172318
	DIF model	6083	170550	105	<u>170760</u>	<u>171465</u>

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

5.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item-discrimination parameters are equal. To test this assumption, a generalized partial credit model (GPCM; Muraki, 1992) that estimates discrimination parameters was fitted to the data. The estimated discrimination parameters differed moderately between items (see Table 6). The median discrimination parameter fell at 0.67 (*Min* = 0.23, *Max* = 1.86). Particularly, the discrimination parameters for item ica14s03s_c was rather low. However, an inspection of the respective item characteristic curve indicated an adequate fit of the PCM. Model fit indices suggested a better model fit of the GPCM (AIC = 171921, BIC = 172586, number of parameters = 99) as compared to the PCM model (AIC = 173105, BIC = 173562, number of parameters = 68). Despite the empirical preference for the GPMC, the PCM more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the PCM was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

5.3.5 Unidimensionality

The dimensionality of the test was investigated by evaluating the correlations between the residuals of the PCM. The adjusted Q3 statistics (see last column in Table 6) were quite low (*Mdn* = 0.06)-the largest individual residual correlation was 0.22 (item ica14s40s_c). Overall, these results indicate an essentially unidimensional test.

We also examined the dimensionality of the test by specifying two multi-dimensional model based on the three process components (*access/create, manage, evaluate*) or four software applications (*word processing, spreadsheet/presentation software, e-mail/communication tools, and internet/search engines*; see Section 2.1). Each item was assigned to one of three

or four latent factors (between-item multidimensionality). The multidimensional models were estimated using Quasi Monte Carlo method with 5,000 nodes. The variances and correlations of the three dimensions reflecting the process components are summarized in Table 10. All dimensions exhibited substantial variances. As expected, the correlations between the three process components were rather high, falling between .90 and .95. But, they deviated from a perfect correlation (i.e., $r = .95$, see Carstensen, 2013). Accordingly, the three-dimensional model did fit the data better (AIC = 173022, BIC = 173512, number of parameters = 73) than the unidimensional model (AIC = 173105, BIC = 173562, number of parameters = 68). The variances and correlations of the four dimensions reflecting the software applications are summarized in Table 11. All dimensions exhibited large correlations, falling between .91 and .97. However, items referring to *internet/search engines* exhibited somewhat smaller correlations as compared to the other software applications. Therefore, the four-dimensional model exhibited a slightly better fit to the data (AIC = 172581, BIC = 173098, number of parameters = 77) than the unidimensional model. These results indicate that the four software applications measure a common construct, although they are not completely unidimensional. Because the computer literacy test was constructed to measure a single dimension, a unidimensional competence score was estimated.

Table 10.

Results of Three-Dimensional Scaling for Process Components

	Access/create	Manage	Evaluate
Access/create	1.42		
Manage	.92	0.95	
Evaluate	.95	.90	1.57

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

Table 11.

Results of Four-Dimensional Scaling for Software Applications

	Word	Spreadsheet	E-mail	Internet
Word	1.51			
Spreadsheet	.96	1.56		
E-mail	.97	.96	2.68	
Internet	.93	.92	.91	0.61

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

6. Discussion

The analyses in the previous sections reported information on the quality of computer literacy test that was administered in Starting Cohorts 4 and 6 of the NEPS. Different kinds of missing responses were examined, item fit statistics and item characteristic curves were evaluated, and item discriminations were investigated. Further quality inspections were conducted by examining differential item functioning and testing Rasch-homogeneity. Various criteria indicated a good fit of the items and measurement invariance across various subgroups. However, the number of missing values was large because many respondents did not reach the last item and, particularly in Starting Cohort 6, respondents did not finish interactive items. The test was slightly better targeted at low- to medium-performing respondents and better covered the lower ability spectrum. As a consequence, ability estimates will be more precise for low-performing respondents as compared to high-performing respondents. However, overall, the test had a good reliability and distinguished well between test-takers. In summary, the test had satisfactory psychometric properties that allowed the estimation of a unidimensional computer literacy score for the participants in both cohorts.

7. Data in the Scientific Use Files

7.1 Naming Conventions

The SUFs for the two starting cohorts contain 32 items, of which 5 were scored dichotomously (static multiple-choice items) with 0 indicating an incorrect response and 1 indicating a correct response. The remaining items were scored as polytomous variables (static CMC and interactive items) that are marked with a 's_c' at the end of the variable names. For further details on the naming conventions of the variables see Fuß and colleagues (2021). Two items are not included in the SUFs because technical problems with the testing system prevented proper recording of the responses.

7.2 Linking of Competence Scores

In both starting cohorts, computer literacy was measured in the current wave and also in a previous wave. The tests in the different waves included different items and were constructed in such a way as to allow for an accurate measurement of computer literacy within the respective age group. As a consequence, the competence scores derived in the different waves cannot be directly compared; differences in observed scores would reflect differences in competences as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competences across waves, the linking procedure described in Fischer et al. (2016) was adopted. Following an anchor-group design, an independent link sample including respondents that were not part of the present cohorts were administered all items from the previous and the current test of computer literacy within a single measurement occasion. These responses were used to link the current test to the test that was administered in the previous wave of each starting cohort.

7.2.1 Samples

In Starting Cohort 4, a subsample of 1,415 respondents (53% women) participated at both measurement occasions, in Wave 7 (i.e., Grade 12; see Senkbeil & Ihme, 2017) and also in Wave 14 (see above), while in Starting Cohort 6 the respective sample size that participated

in Wave 5 (see Senkbeil & Ihme, 2015) and Wave 14 (see above) was $N = 2,266$ (50% women). Consequently, these respondents that participated at both measurement occasions were used to link the two tests in each cohort across waves. Moreover, an independent link sample of $N = 566$ students (55% women) received all three tests within a single measurement occasion.

7.2.1.1 Design of the Link Study

The paper-based tests administered in Wave 7 of Starting Cohort 4 (Senkbeil & Ihme, 2017) and Wave 5 of Starting Cohort 6 (Senkbeil & Ihme, 2015) included 32 and 29 items, respectively, while the computerized MST in Wave 14 of both cohorts comprised 32 items. In the link sample, the three tests were administered to all participants. But, because the MST included several items that were also included in the paper-based tests, these items were replaced by item twins with similar item content and item difficulty as the original items. Thus, in the link sample only 7 items from the test originally administered in Wave 7 of Starting Cohort 4 were presented. Similarly, the link sample included only 6 items from the test originally administered in Wave 5 of Starting Cohort 6. In contrast, the same MST as originally administered in Wave 14 of both cohorts was used. Because no information on the respondents' competence level was available from previous assessments, the easy and difficult level in the first stage of the MST was randomly selected for each respondent.

In the link sample, the three computer literacy tests were administered at different positions in the test battery. A random sample of 274 respondents received the MST before working on the paper-based tests, whereas 292 respondents worked on the paper-based tests first. Moreover, a random sample of 310 respondents received the paper-based test items from Starting Cohort 4 before the paper-based test items from Starting Cohort 6, while 256 respondents were administered the paper-based tests in reversed sequence. No multi-matrix design regarding the selection and order of the items within a test was established. Thus, all test takers were given the computer literacy items in the same order.

7.2.1.2 Results

To examine whether the two tests administered in the link sample measured a common scale, we compared a one-dimensional model that specified a single latent factor for all items to a two-dimensional model that specified separate latent factors for the computerized MST and each paper-based tests. In Starting Cohort 4, the information criteria favored the unidimensional model (AIC = 22567, BIC = 22914, number of parameters = 80) as compared to the two-dimensional model (AIC = 22557, BIC = 22913, number of parameters = 82). Similarly, in Starting Cohort 6, the unidimensional model (AIC = 21590, BIC = 21928, number of parameters = 80) had a comparable fit as the two-dimensional model (AIC = 21580, BIC = 21928, number of parameters = 82). Because the differences in the information criteria between the unidimensional model and the two-dimensional model were small and, therefore, negligible, the results indicate that the computer literacy tests administered in the two waves were essentially unidimensional.

Table 12.

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

Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	F	$RMSD_{MS}$	$RMSD_{LS}$
Paper-based test from Wave 7					
icg12004s_c	0.52	0.05	104.82*	0.04	0.08
icg12010x_c	-0.36	0.12	9.59	0.03	0.09
icg12011x_c [#]	0.25	0.12	4.36	0.04	0.02
icg12060s_c [#]	-0.14	0.08	3.67	0.03	0.05
icg12016s_c [#]	-0.04	0.07	0.27	0.02	0.03
icg12121x_c	0.49	0.13	15.03	0.03	0.07
ica5027x_c	-0.73	0.15	24.92*	0.03	0.14
Computerized MST from Wave 14					
ica5052s_sc4a14_c [#]	0.27	0.10	6.84	0.02	0.04
icg12018s_sc4a14_c [#]	-0.24	0.14	3.10	0.03	0.05
ica5017s_sc4a14_c [#]	0.14	0.08	3.24	0.02	0.01
ica5001x_sc4a14_c [#]	0.21	0.20	1.05	0.03	0.04
ica5023x_sc4a14_c	-0.69	0.17	16.92	0.01	0.04
icg12054s_sc4a14_c	-0.34	0.07	26.19	0.04	0.07
ica5015s_sc4a14_c	0.75	0.08	82.67*	0.04	0.07
icg12107s_sc4a14_c [#]	0.14	0.11	1.61	-0.01	0.06
ica5050s_sc4a14_c [#]	0.15	0.08	3.19	0.03	0.06
ica5021s_sc4a14_c [#]	-0.32	0.13	6.15	0.05	0.08
ica5026x_sc4a14_c	-0.44	0.17	7.18	0.01	0.04
icg9122x_sc4a14_c [#]	-0.27	0.13	4.52	0.03	0.03
ica5003x_sc4a14_c [#]	-0.36	0.17	4.52	0.02	0.05
icg12028s_sc4a14_c	0.42	0.12	11.46	0.06	0.06
icg12138s_sc4a14_c [#]	0.30	0.07	16.74	0.02	0.04
ica5004s_sc4a14_c [#]	0.15	0.12	1.59	0.04	0.08
icg12047s_sc4a14_c [#]	0.04	0.07	0.45	0.03	0.06
icg12109s_sc4a14_c [#]	0.10	0.09	1.25	0.02	0.05

Note. $\Delta\sigma$ = Difference in item difficulty parameters between waves (negative values indicate easier items in wave 7); $SE_{\Delta\sigma}$ = Pooled standard error; F = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is $F_{0.054}(1, 1979) = 52.13$. A non-significant test indicates measurement invariance. $RMSD$ = Root mean squared deviation for main sample in Starting Cohort 4 (ms) or link sample (ls).

* $p < .05$

Table 13.

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

Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	F	$RMSD_{MS}$	$RMSD_{LS}$
Paper-based test from Wave 5					
ica5005x_c [#]	-0.29	0.12	5.64	0.02	0.02
ica5006x_c [#]	0.16	0.13	1.70	0.03	0.03
ica5007x_c	0.79	0.13	35.35	0.05	0.10
ica5008x_c [#]	-0.23	0.13	3.33	0.07	0.05
ica5016s_c [#]	-0.22	0.07	10.54	0.03	0.05
ica5020s_c [#]	-0.21	0.07	10.07	0.06	0.05
Computerized MST from Wave 14					
ica5052s_sc4a14_c [#]	-0.34	0.10	11.00	0.02	0.05
icg12018s_sc4a14_c [#]	-0.02	0.10	0.04	0.01	0.04
ica5017s_sc4a14_c [#]	-0.22	0.07	8.55	0.03	0.05
ica5001x_sc4a14_c [#]	0.20	0.16	1.49	0.02	0.01
ica5023x_sc4a14_c [#]	-0.07	0.16	0.19	0.02	0.04
icg12054s_sc4a14_c [#]	0.25	0.06	17.63	0.03	0.05
ica5015s_sc4a14_c [#]	0.01	0.08	0.04	0.04	0.07
icg12107s_sc4a14_c [#]	0.27	0.11	6.62	0.04	0.08
ica5050s_sc4a14_c [#]	-0.04	0.08	0.30	0.04	0.07
ica5021s_sc4a14_c	0.48	0.11	20.85	0.05	0.13
ica5026x_sc4a14_c [#]	-0.25	0.16	2.40	0.03	0.06
icg9122x_sc4a14_c [#]	0.07	0.12	0.33	0.01	0.03
ica5003x_sc4a14_c [#]	-0.30	0.17	3.13	0.09	0.06
icg12028s_sc4a14_c [#]	-0.04	0.11	0.11	0.01	0.05
icg12138s_sc4a14_c [#]	-0.23	0.07	9.77	0.03	0.06
ica5004s_sc4a14_c [#]	0.19	0.12	2.60	0.02	0.09
icg12047s_sc4a14_c [#]	0.17	0.07	6.56	0.03	0.06
icg12109s_sc4a14_c [#]	-0.15	0.09	3.15	0.03	0.06

Note. $\Delta\sigma$ = Difference in item difficulty parameters between waves (negative values indicate easier items in wave 7); $SE_{\Delta\sigma}$ = Pooled standard error; F = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is $F_{0.054}(1, 2830) = 69.03$. A non-significant test indicates measurement invariance. $RMSD$ = Root mean squared deviation for main sample in Starting Cohort 4 (ms) or link sample (ls).

* $p < .05$

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal subsample from the starting cohorts. Only static items were

considered as potential common items; therefore, the interactive items from the MST were excluded from these analyses. The differences in item difficulties between the longitudinal sample in Starting Cohort 4 and the link sample for the two tests from Waves 7 and 14 and the tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 12. The minimum effects hypothesis test identified significant ($\alpha = .05$) DIF for three items. Moreover, the root means squared deviation (RMSD) showed non-negligible DIF with $\text{RMSD} \geq .08$ for three additional items. Also, some items exhibited noticeable differences in item difficulties greater than 0.40 logits. Therefore, we selected 3 items from the test administered in Wave 7 with DIF effects that did not exceed 0.40 on the logit scale and 13 items from the MST administered in Wave 14.

The respective results of the DIF analyses for Starting Cohort 6 are presented in Table 13. None of the item exhibited significant DIF according to the minimum effect hypothesis test. However, two items showed non-negligible differences in item difficulties exceeding 0.40 on the logit scale. Therefore, these items were not considered for the linking. As a result, 5 items from the paper-based test administered in Wave 5 and 17 items from the MST administered in Wave 14 were used as common items to link the two tests.

The computer literacy tests administered in the two waves of each starting cohort were linked using the “mean/mean” method for the anchor-group design using these common items (see Fischer et al., 2016). For Starting Cohort 4, the correction term that was used to link the two waves was calculated as 0.601. The link error reflecting the uncertainty in the linking process was calculated according to Equation 2 in Fischer et al. (2016) as 0.134. The respective calculations for Starting Cohort 6 gave a correction term of -0.228 and a link error of 0.094.

7.3 Computer Literacy Scores

In the SUF, manifest computer literacy scores are provided in the form of WLEs (ica14_sc1) including their respective standard error (ica14_sc2). Because the responses for the two starting cohorts were concurrently scaled, these WLEs can be used to compare computer literacy across starting cohorts in Wave 14. For ica14_sc1u person abilities were estimated using the linked item difficulty parameters. As a result, these WLE scores can be used for longitudinal comparisons between different waves within a starting cohort. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in ica14_sc1 are not linked to the underlying reference scale of the previous wave. However, they are corrected for the position of the computer literacy test within the test battery. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions.

The R Syntax for estimating the WLEs is provided in Appendix B. In the IRT scaling model, all polytomous variables were scored as 0.5 for each category. For respondents who did not take part in the computer literacy test or who did not give enough valid responses no WLEs were estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values. Alternatively, users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values. Plausible values for the computer literacy test can be estimated using the R package *NEPSscaling* (Scharl et al., 2020; Scharl & Zink, 2022).

8. References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–722. <https://doi.org/10.1109/TAC.1974.1100705>
- Carstensen, C. H. (2013). Linking PISA competencies over three cycles – Results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (eds.), *Research Outcomes of the PISA Research Conference 2009* (pp. 199–214). Springer.
- Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (2016). *Linking the data of the competence tests* (NEPS Survey Paper No. 1). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP01:1.0>
- Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2021). *Competence data in NEPS: Overview of measures and variable naming conventions* (Starting Cohorts 1 to 6). Leibniz Institute for Educational Trajectories, National Educational Panel Study.
- 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>
- Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests* (NEPS Working Paper No. 14). 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. <https://doi.org/10.25656/01:8430>
- R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute of Education Research.

- Robitzsch, A., Kiefer, T., & Wu, M. (2021). *TAM: Test analysis modules*. R package version 2.12-18. URL: <https://CRAN.R-project.org/package=TAM>
- Scharl, A., Carstensen, C. H., & Gnambs, T. (2020). *Estimating plausible values with NEPS data: An example using reading competence in Starting Cohort 6* (NEPS Survey Paper No. 71). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP71:1.0>
- Scharl, A., & Zink, E. (2022). NEPSscaling: plausible value estimation for competence tests administered in the German National Educational Panel Study. *Large-scale Assessments in Education*, 10:28. <https://doi.org/10.1186/s40536-022-00145-5>
- 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., & Ihme, J. M. (2015). *NEPS Technical Report for Computer Literacy: Scaling results of Starting Cohort 6–Adults* (NEPS Working Paper No. 61). Leibniz Institute for Educational Trajectories, National Educational Panel Study.
- Senkbeil, M., & Ihme, J. M. (2017). *NEPS Technical Report for Computer Literacy: Scaling results of Starting Cohort 4 for Grade 12* (NEPS Survey Paper No. 25). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP25:1.0>
- Senkbeil, M., Ihme, J. M., & Wittwer, J. (2013). The test of technological and information literacy (TILT) in the National Educational Panel Study: Development, empirical testing, and evidence for validity. *Journal for Educational Research Online*, 5, 139–161. <https://doi.org/10.25656/01:8428>
- 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. *Zeitschrift für Erziehungswissenschaft*, 14, 67–86. <https://doi.org/10.1007/s11618-011-0182-7>
- Wu, M., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0*. Acer Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125–145. <https://doi.org/10.1177/01466216840080020>

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213. <https://doi.org/10.1111/j.1745-3984.1993.tb00423.x>

Appendix A: Process Components and Software Applications for the Items

Position	Stage	Starting Cohort 4	Starting Cohort 6	Process component	Software application
1	1	ica5052s_sc4a14_c	ica5052s_sc6a14_c	create	word processing
1	1	icg12018s_sc4a14_c	icg12018s_sc6a14_c	manage	e-mail / communication tools
2	1	ica5017s_sc4a14_c	ica5017s_sc6a14_c	manage	internet / search engines
3	1	ica5001x_sc4a14_c	ica5001x_sc6a14_c	create	spreadsheet / presentation software
3	1	ica5023x_sc4a14_c	ica5023x_sc6a14_c	create	spreadsheet / presentation software
4	1	icg12054s_sc4a14_c	icg12054s_sc6a14_c	create	word processing
5	1	ica5015s_sc4a14_c	ica5015s_sc6a14_c	evaluate	spreadsheet / presentation software
5	1	icg12107s_sc4a14_c	icg12107s_sc6a14_c	evaluate	spreadsheet / presentation software
6	1	ica5050s_sc4a14_c	ica5050s_sc6a14_c	manage	internet / search engines
7	1	ica5021s_sc4a14_c	ica5021s_sc6a14_c	access	word processing
7	1	ica5026x_sc4a14_c	ica5026x_sc6a14_c	manage	spreadsheet / presentation software
8	1	icg9122x_sc4a14_c	icg9122x_sc6a14_c	manage	internet / search engines
9	1	ica5003x_sc4a14_c	ica5003x_sc6a14_c	evaluate	internet / search engines
9	1	icg12028s_sc4a14_c	icg12028s_sc6a14_c	access	e-mail / communication tools
10	1	icg12138s_sc4a14_c	icg12138s_sc6a14_c	access	e-mail / communication tools
11	1	ica5004s_sc4a14_c	ica5004s_sc6a14_c	evaluate	internet / search engines
11	1	icg12047s_sc4a14_c	icg12047s_sc6a14_c	create	word processing
12	1	icg12109s_sc4a14_c	icg12109s_sc6a14_c	evaluate	internet / search engines
14	2	ica14s05s_c	ica14s05s_sc6a14_c	manage	internet / search engines
14	2	ica14s12s_c	ica14s12s_sc6a14_c	manage	e-mail / communication tools
15	2	ica14s14s_c	ica14s14s_sc6a14_c	manage	internet / search engines
15	2	ica14s40s_c	ica14s40s_sc6a14_c	manage	internet / search engines
16	2	ica14s07s_c	ica14s07s_sc6a14_c	manage	internet / search engines
16	2	ica14s15s_c	ica14s15s_sc6a14_c	access / create	spreadsheet / presentation software
17	2	ica14s21s_c	ica14s21s_sc6a14_c	access / create	word processing
17	2	ica14s63s_c	ica14s63s_sc6a14_c	evaluate	e-mail / communication tools
18	2	ica14s08s_c	ica14s08s_sc6a14_c	manage	internet / search engines
18	2	ica14s16s_c	ica14s16s_sc6a14_c	evaluate	internet / search engines
19	2	ica14s60s_c	ica14s60s_sc6a14_c	access / create	spreadsheet / presentation software
19	2	ica14s64s_c	ica14s64s_sc6a14_c	evaluate	e-mail / communication tools
20	2	ica14s03s_c	ica14s03s_sc6a14_c	access / create	spreadsheet / presentation software
20	2	ica14s20s_c	ica14s20s_sc6a14_c	access / create	word processing

Note. The items on position 13 are not included in the SUF because a technical error in the testing system prevented proper recording of the responses. So far, the internal validity of the individual dimensions of computer literacy has not yet been confirmed. Therefore, analyses on the facet level should not be conducted.

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

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

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

# 32 items of the computer literacy test
items <- c("ica5052s_sc4a14_c", "icg12018s_sc4a14_c",
          "ica5017s_sc4a14_c", "ica5001x_sc4a14_c",
          ...)

# polytomous items
poly <- c("ica5052s_sc4a14_c", "icg12018s_sc4a14_c",
          "ica5017s_sc4a14_c", "icg12054s_sc4a14_c",
          "ica5015s_sc4a14_c", "icg12107s_sc4a14_c",
          "ica5050s_sc4a14_c", "ica5021s_sc4a14_c",
          "icg12028s_sc4a14_c", "icg12138s_sc4a14_c",
          "ica5004s_sc4a14_c", "icg12047s_sc4a14_c",
          "icg12109s_sc4a14_c", "ica14s05s_c",
          "ica14s12s_c", "ica14s14s_c", "ica14s21s_c",
          "ica14s16s_c", "ica14s60s_c", "ica14s03s_c",
          "ica14s20s_c")

# collapse response categories
dat$icg12018s_sc4a14_c <- doBy::recodeVar(
  dat$icg12018s_sc4a14_c,
  c(0, 1, 2, 3, 4),
  c(0, 0, 0, 1, 2)
)
dat$ica5017s_sc4a14_c <- doBy::recodeVar(
  dat$ica5017s_sc4a14_c,
  c(0, 1, 2, 3, 4, 5),
  c(0, 0, 0, 0, 1, 2)
)
dat$icg12054s_sc4a14_c <- doBy::recodeVar(
  dat$icg12054s_sc4a14_c,
  c(0, 1, 2, 3, 4),
  c(0, 0, 1, 2, 3)
)
dat$ica5015s_sc4a14_c <- doBy::recodeVar(
  dat$ica5015s_sc4a14_c,
```

```
      c(0, 1, 2, 3, 4, 5),
      c(0, 0, 1, 2, 3, 4)
    )
    dat$ica5050s_sc4a14_c<- doBy::recodeVar(
      dat$ica5050s_sc4a14_c,
      c(0, 1, 2, 3, 4),
      c(0, 0, 0, 1, 2)
    )
    dat$ica5021s_sc4a14_c <- doBy::recodeVar(
      dat$ica5021s_sc4a14_c,
      c(0, 1, 2, 3, 4, 5),
      c(0, 0, 0, 0, 1, 2)
    )
    dat$icg12028s_sc4a14_c <- doBy::recodeVar(
      dat$icg12028s_sc4a14_c,
      c(0, 1, 2, 3, 4, 5),
      c(0, 0, 0, 0, 1, 2)
    )
    dat$icg12138s_sc4a14_c<- doBy::recodeVar(
      dat$icg12138s_sc4a14_c,
      c(0, 1, 2, 3, 4),
      c(0, 0, 0, 1, 2)
    )
    dat$ica5004s_sc4a14_c <- doBy::recodeVar(
      dat$ica5004s_sc4a14_c,
      c(0, 1, 2, 3, 4, 5),
      c(0, 0, 0, 0, 1, 2)
    )
    dat$icg12109s_sc4a14_c <- doBy::recodeVar(
      dat$icg12109s_sc4a14_c,
      c(0, 1, 2, 3),
      c(0, 0, 1, 2)
    )
    dat$ica5052s_sc4a14_c <- doBy::recodeVar(
      dat$ica5052s_sc4a14_c,
      c(0, 1, 2, 3, 4, 5),
      c(0, 0, 0, 0, 1, 2)
    )
    dat$icg12107s_sc4a14_c <- doBy::recodeVar(
      dat$icg12107s_sc4a14_c,
```

```
      c(0, 1, 2, 3, 4, 5),
      c(0, 0, 0, 0, 1, 2)
    )
dat$icg12047s_sc4a14_c <- doBy::recodeVar(
  dat$icg12047s_sc4a14_c,
  c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10),
  c(0, 0, 0, 0, 0, 0, 1, 2, 3, 4, 5)
)
dat$ica14s63s_c <- doBy::recodeVar(
  dat$ica14s63s_c,
  c(0, 1, 2, 3, 4, 5, 6, 7, 8),
  c(0, 0, 0, 0, 0, 0, 0, 1, 2)
)
dat$ica14s64s_c <- doBy::recodeVar(
  dat$ica14s64s_c,
  c(0, 1, 2, 3, 4, 5, 6, 7),
  c(0, 0, 0, 0, 0, 0, 1, 1)
)
dat$ica14s12s_c <- doBy::recodeVar(
  dat$ica14s12s_c,
  c(0, 1, 2, 3, 4, 5, 6, 7),
  c(0, 0, 0, 0, 1, 2)
)
dat$icg12107s_sc4a14_c <- doBy::recodeVar(
  dat$icg12107s_sc4a14_c,
  c(0, 1, 2, 3, 4, 5, 6, 7),
  c(0, 1, 1, 1, 2, 3, 4, 5)
)
dat$ica14s15s_c <- doBy::recodeVar(
  dat$ica14s15s_c,
  c(0, 1, 2, 3),
  c(0, 0, 0, 1)
)
dat$ica14s20s_c <- doBy::recodeVar(
  dat$ica14s20s_c,
  c(0, 1, 2, 3, 4, 5, 6),
  c(0, 0, 0, 0, 0, 1, 2)
)
dat$ica14s21s_c <- doBy::recodeVar(
  dat$ica14s21s_c,
```

```
      c(0, 1, 2, 3, 4),
      c(0, 0, 0, 1, 2)
    )
dat$ica14s07s_c <- doBy::recodeVar(
  dat$ica14s07s_c,
  c(0, 1, 2),
  c(0, 0, 1)
)
dat$ica14s08s_c <- doBy::recodeVar(
  dat$ica14s08s_c,
  c(0, 1, 2, 3),
  c(0, 0, 0, 1)
)
dat$ica14s40s_c <- doBy::recodeVar(
  dat$ica14s40s_c,
  c(0, 1, 2, 3, 4, 5),
  c(0, 0, 0, 0, 0, 1)
)

# define Q-matrix for 0.5 scoring of PCM
Q <- matrix(1, nrow = length(items), ncol = 1)
Q[items %in% poly, 1] <- 0.5 # score of 0.5

# estimate partial credit model
mod <- TAM::tam.mml(resp = dat[, items], Q = Q, irtmodel = "PCM2",
  pid = dat$ID_t)

summary(mod)

# item fit
TAM::tam.fit(mod)

# WLE
TAM::tam.wle(mod)
```