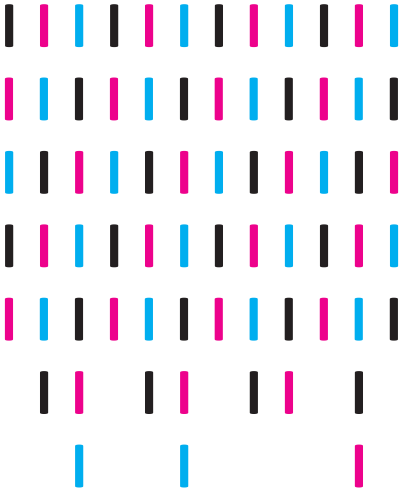


NEPS *SURVEY PAPERS*

Martin Senkbeil and Jan Marten Ihme

NEPS TECHNICAL REPORT FOR COMPUTER LITERACY: SCALING RESULTS OF STARTING COHORT 3 FOR GRADE 12

NEPS *Survey Paper* No. 90
Bamberg, October 2021

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

Survey Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

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

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.

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 Cohort 3 for Grade 12

Martin Senkbeil, Jan Marten Ihme
Leibniz Institute for Science and Mathematics Education at the University of Kiel

E-mail address of lead author:

senkbeil@ipn.uni-kiel.de

Bibliographic data:

Senkbeil, M., & Ihme, J. M. (2021). *NEPS Technical Report for Computer Literacy: Scaling results of Starting Cohort 3 for Grade 12* (NEPS Survey Paper No. 90). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP90:1.0.

Acknowledgements:

This paper uses preliminary data from the National Educational Panel Study (NEPS): Starting Cohort Grade 5, doi:10.5157/NEPS:SC3:10.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS has been carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

This report is an extension to NEPS survey paper 29 (Senkbeil & Ihme, 2017a) that presents the scaling results for computer literacy of starting cohort 3 for grade 9. Therefore, various parts of this report (e.g., regarding the instruction and the analytic strategy) are reproduced verbatim from previous working papers (Senkbeil & Ihme, 2017a; Senkbeil, Ihme, & Adrian, 2014) to facilitate the understanding of the presented results.

We would like to thank Timo Gnambs for developing and providing standards for the technical reports and for giving valuable feedback on previous drafts of this manuscript.

NEPS Technical Report for Computer Literacy: Scaling Results of Starting Cohort 3 for Grade 9

Abstract

The National Educational Panel Study (NEPS) investigates the development of competencies across the life span and develops tests for the assessment of different competence domains. In order to evaluate the quality of the competence tests, a range of analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedures for the computer literacy test in grade 12 of starting cohort 3 (fifth grade). The computer literacy test contained 32 items (distributed among one booklet with all items and three booklets with a low, medium, or high level of difficulty) with different response formats representing different cognitive requirements and different content areas. The test was administered to 3,749 students. Their responses were scaled using the partial credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, the test's dimensionality, and local item independence were evaluated to ensure the quality of the test. These analyses showed that the test exhibited an acceptable reliability and that all items fitted the model in a satisfactory way. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test was the large number of items targeted toward a lower computer literacy as well as the large percentage of items at the end of the test that were not reached due to time limits. Further challenges related to the dimensionality analyses based on both software applications and cognitive requirements. Overall, the computer literacy test had acceptable psychometric properties that allowed for a reliable estimation of computer competence scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the ConQuest-syntax for scaling the data.

Keywords

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

Content

1. Introduction

Within the National Educational Panel Study (NEPS) (Blossfeld, Roßbach, & von Maurice, 2011), different competencies are measured coherently across the life span. Tests have been developed for different competence domains. These include, among other things, reading competence, mathematical competence, scientific literacy, information and communication literacy (computer literacy), metacognition, vocabulary, and domain-general cognitive functioning. An overview of the competences measured in the NEPS is given by Weinert et al. (2011) as well as Fuß, Gnambs, Lockl, and Attig (2021).

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

In this paper the results of these analyses are presented for computer literacy in starting cohort 3 (fifth grade) in grade 12. First, the main concepts of the computer literacy test are introduced. Then, the computer literacy data of starting cohort 3 and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the scientific use file is presented.

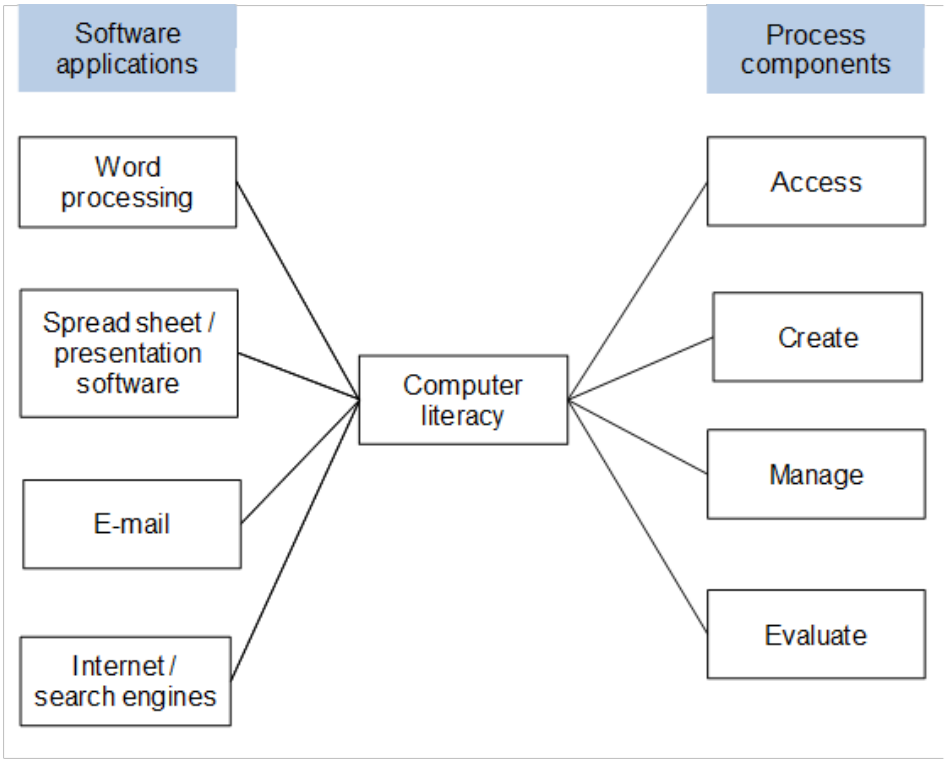
Please note that the analyses in this report are based on the data available at some time before public data release. Due to ongoing data protection and data cleansing issues, the data in the scientific use file (SUF) may differ slightly from the data used for the analyses in this paper. However, we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

2. Testing Computer Literacy

The framework and test development for the computer literacy test is 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 the different facets of technological and information literacy. In line with the literacy concepts of international large-scale assessments, we define computer literacy 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 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 TILT (Test of Technological and Information Literacy) 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.



Each item in the test refers to one process component and one software application. With the exception of a few items addressing factual knowledge (e.g., computer terminology), the items ask subjects to accomplish computer-based tasks. To do so, subjects were presented with realistic problems embedded in a range of authentic situations. Most items use screenshots, for example, of an internet browser, an electronic database, or a spreadsheet as prompts (see Senkbeil et al., 2013).

In the computer literacy test of starting cohort 3 (fifth grade) in grade 12 there are two types of response formats. These are simple multiple choice (MC) and complex multiple choice (CMC) items. In MC items the test taker has to find 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 are incorrect). In CMC items a number of subtasks with two response options each (true / false) are presented. The number of subtasks of CMC items varies between four and ten. Examples of the different response formats are given in Pohl and Carstensen (2012).

The competence test for computer literacy that was administered in the present study included 32 items. In order to evaluate the quality of these items extensive preliminary analyses were conducted. These preliminary analyses revealed that none of the items had a poor fit.

3. Data

3.1 The Design of the Study

The study followed a three-factorial (quasi-)experimental design. These factors referred to (a) the position of the computer literacy test within the test battery, (b) the difficulty of the administered test, and (c) the assessment setting (i.e., the context of test administration).

The study assessed different competence domains including, among others, computer literacy, reading competence, and mathematical competence. The competence tests for these three domains were always presented first within the test battery. 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 reading or the mathematics test).

The panel study aimed at retesting all students that were initially included in the starting cohort 3 for fifth grade (see Senkbeil & Ihme, 2017a; Senkbeil, Ihme, & Adrian, 2014). Because some students left their original schools during the course of the longitudinal study or left the school context altogether, the participants of the starting cohort were divided into two subsamples that exhibited different assessment settings: Students that remained at the same school as in the previous assessment were tested at school in a group setting; in contrast, students that left their original school were tracked and, subsequently, individually tested at home (for details regarding the data collection process, see the respective field report for wave 9). Thus, the context of test administration differed between the two groups.

Students that remained at the same school as in the previous assessment and that were tested at school in a group setting received the overall test that included 32 items. This test was identical to the computer literacy test in grade 12 of starting cohort 4 (Senkbeil & Ihme, 2017b). Students that left their original school and that were individually tested at home received a subsample of the overall test that included 19 items. Additionally, they received simulation-based test items within a computer-based test environment that are not part of the scaling in the present report. In order to measure computer literacy of the students tested individually at home with great accuracy, the difficulty of the administered tests should adequately match the participants' abilities. Therefore, the study adopted the principals of longitudinal multistage testing (Pohl, 2013). Based on preliminary studies three different versions of the computer literacy test were developed that differed in their average difficulty (i.e., a test with low level of difficulty, a test with medium level of difficulty, and a test with high level of difficulty). Each of the three tests included 11 items that represented the four process components (see Table 1) and the four software applications (see Table 2). Three items were identical in all three test versions, seven items were identical in the tests with low and medium level of difficulty, and seven items were identical in the tests with medium and high level of difficulty (see Tables 1 and 2). Four items were unique to the test with low medium of difficulty and to the test with high level of difficulty (see Appendix C for the detailed assignment of the test items to each test version). The different response formats of the items are summarized in Table 3. Participants were assigned to the test version based on their computer literacy competence in the previous assessment (Senkbeil et al., 2014).

Table 1

Number of Items for the Different Process Components by Assessment Setting and Difficulty of the Test

Process components	**Overall test, at school**	**Low level, at home**	**Medium level, at home**	**High level, at home**	**All tests, at home**	**Low and medium level, at home**	**Medium and high level, at home**
Access	7	4	1	1	4	4	1
Create	8	2	5	6	6	5	6
Manage	9	1	2	2	3	2	3
Evaluate	8	4	3	2	6	4	5
Total number of items	32	11	11	11	19	15	15

Table 2

Number of Items for the Different Software Applications by Assessment Setting and Difficulty of the Test

Software applications	**Overall test, at school**	**Low level, at home**	**Medium level, at home**	**High level, at home**	**All tests, at home**	**Low and medium level, at home**	**Medium and high level, at home**
Word processing	6	3	4	3	4	4	4
Spreadsheet / presentation software	11	2	3	5	6	4	5
E-mail / communication tools	5	2	1	1	3	3	1
Internet / search engines	10	4	3	2	6	4	5
Total number of items	32	11	11	11	19	15	15

Table 3

Number of Items by Different Response Formats, Assessment setting, and Difficulty of the Test

Response format	**Overall test, at school**	**Low level, at home**	**Medium level, at home**	**High level, at home**
Simple multiple choice items	13	3	5	7
Complex multiple choice items	19	8	6	4
Total number of items	32	11	11	11

3.2 Sample

A total of 3,749 individuals received the computer literacy test. For one participant less than three valid item responses were available. Because no reliable ability scores can be estimated based on such few valid responses, this case was excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 3,748 individuals. The number of participants within each (quasi-)experimental condition is given in Table 4. A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

Table 4

Number of Participants by the (Quasi-)Experimental Conditions

	Assessment setting:	**At school (n = 1,762)**		**At home (n = 1,986)**		**Total**
	Test position:	First position	Second position	First position	Second position	
	Overall test	887	885			1762
Test	**Low level**			126	134	260
Difficulty	**Medium level**			589	616	1205
	High level			279	242	521
	Total	887	885	994	992	3748

4. Analyses

4.1 Missing Responses

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

Invalid responses occurred, for example, when two response options were selected in simple MC items where only one was required, or when numbers or letters that were not within the range of valid responses were given as a response. Omitted items occurred when test takers skipped some items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response given were coded as not-reached. Because of the branched design (students that that left their original school and that were individually tested at home) not all items were administered to all participants. For respondents receiving the test with low level of difficulty 8 items of the tests with medium and high level of difficulty were missing by design, for respondents receiving the test with medium level of difficulty 4 items of the test with low level of difficulty and 4 items of the test with high level of difficulty were missing by design, and for respondents receiving the test with high level of difficulty 8 items of the tests with low and medium level of difficulty were missing by design (see Table 1 and Appendix B). As CMC items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A CMC item was coded as missing if at least one subtask contained a missing response. When one subtask contained a missing response, the CMC item was coded as missing. 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) and need to be accounted for in the estimation of item and person parameters. We, therefore, thoroughly investigated the occurrence of missing responses in the test. First, we looked at the occurrence of the different types of missing responses per person. This gave an indication of how well the persons were coping with the test. We then looked at the occurrence of missing responses per item in order to obtain some information on how well the items worked.

4.2 Scaling Model

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

CMC items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly solved subtasks within that item. If at least one of the subtasks contained a missing response, the whole CMC item was scored as missing. When categories of the polytomous variables had less than N = 200, the categories were

collapsed in order to avoid any possible estimation problems. This usually occurred for the lower categories of polytomous items; especially when the item consisted of many subtasks. In these cases the lower categories were collapsed into one category. For all of the 19 CMC items categories were collapsed with the exception of one item (icg12037s_sc3g12_c; see Appendix A). To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items were scored dichotomously as 0 for an incorrect and as 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats).

4.3 Checking the Quality of the Scale

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

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

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

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

The computer literacy test should measure the same construct for all students. If any items favored certain subgroups (e.g., if they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between the subgroups (e.g., males and females) would be biased and thus unfair. For the present study, test fairness was investigated for the variables test position, gender, school type (secondary school vs. other school types), the number of books at home (as a proxy for cultural capital), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Differential item functioning (DIF) analyses were estimated using a multigroup IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item

difficulty were modeled. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties between the subgroups that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as noteworthy of further investigation, differences between 0.4 and 0.6 as considerable but not severe, and differences smaller than 0.4 as negligible DIF. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

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

The test was constructed to measure a unidimensional computer literacy score. The computer literacy test is constructed to measure computer literacy on a unidimensional scale (Senkbeil et al., 2013). The assumption of unidimensionality was, nevertheless, tested on the data by specifying different multidimensional models. The different subdimensions of the multidimensional models were specified based on the construction criteria. First, a model with four process components, and second, a model with four different subdimensions based on different software applications was fitted to the data. The correlation among 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 scale. Moreover, we examined whether the residuals of the one-dimensional model exhibited approximately zero-order correlations as indicated by Yen's (1984) Q_3. Because in case of locally independent items, the Q_3 statistic tends to be slightly negative, we report the corrected Q_3 that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) values of Q_3 falling below .20 indicate essential unidimensionality.

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

5. Results

5.1 Missing Responses

5.1.1 Missing responses per person

Figure 2 shows the number of invalid responses per person for students that remained at the same school as in the previous assessment and that were tested at school in a group setting. Please note that invalid responses were not possible for students who received the items in a computer-based testing environment due to technical settings (students that left their original school and that were individually tested at home). Overall, there were very few invalid responses. More than 95% of the respondents did not have any invalid response at all; overall less than one percent had more than one invalid response.

Missing responses may also occur when respondents omit items. As illustrated in Figure 3 most respondents, 66.0% to 77.9%, did not skip any item, and less than seven percent omitted

more than one item. There was only a slight difference in the amount of omitted items between the different experimental conditions.

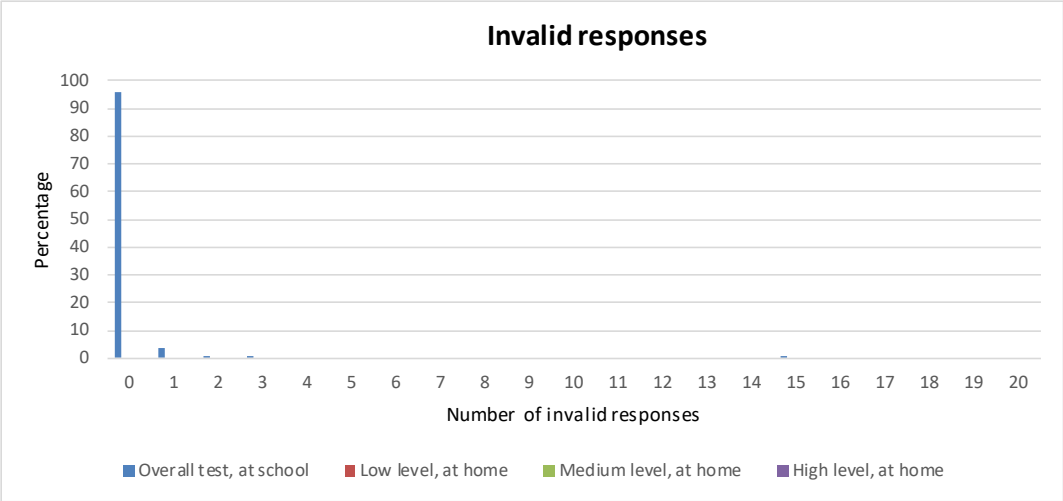


Figure 2. Number of invalid responses (students that were tested at school in a group setting).

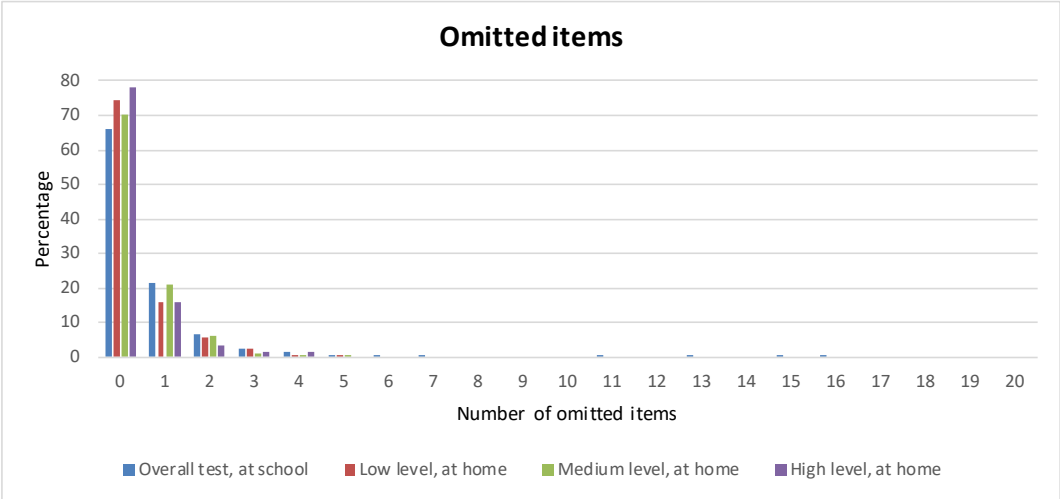


Figure 3. Number of omitted items by test difficulty.

Another source of missing responses are items that were not reached by the respondents; these are all missing responses after the last valid response. The number of not-reached items was rather low, most respondents were able to finish the test within the allocated time limit (Figure 4). Between 73.0% and 82.7% of the respondents finished the entire test. Between 7.3% and 19.1% of the respondents did not reach the last three items. In particular, this applies to the students who received the overall test at school.

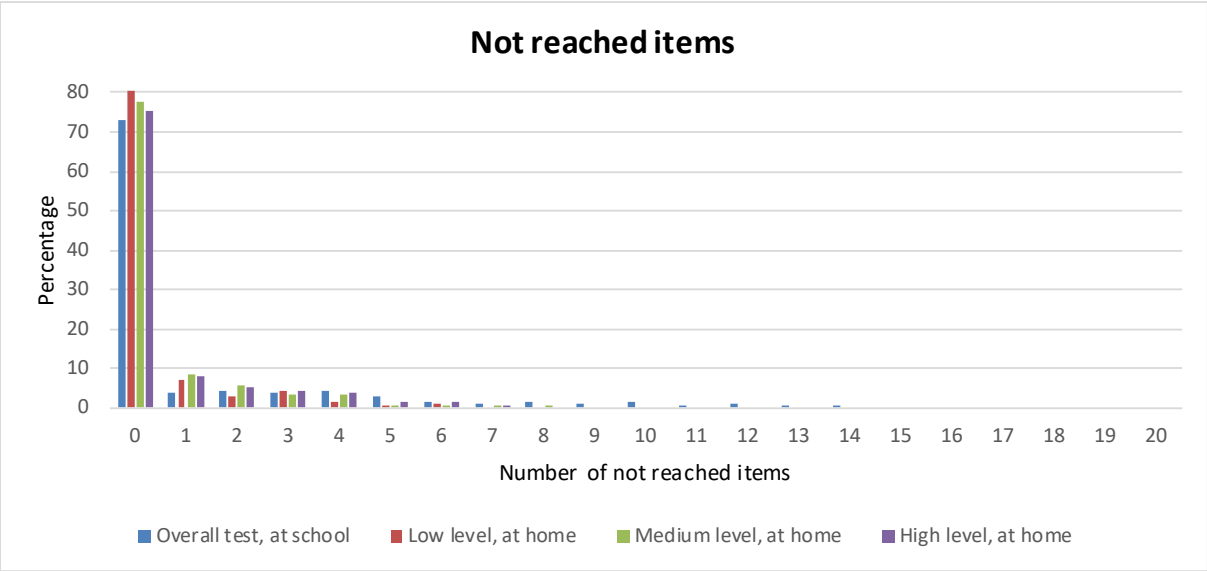


Figure 4. Number of not reached items by test difficulty.

The total number of missing responses, aggregated over invalid, omitted, not-reached, and not determinable per person, is illustrated in Figure 5. On average, the respondents showed between *M* = 0.82 (*SD*= 1.37; test with low level of difficulty) and *M* = 1.97 (*SD* = 3.10; overall test) missing responses in the different experimental conditions. About 47.6% to 60.4% of the respondents had no missing response at all and about 5.0% to 20.9% of the participants had four or more missing responses. Particularly, respondents receiving the overall test at school showed more missing responses than respondents receiving the test individually at home.

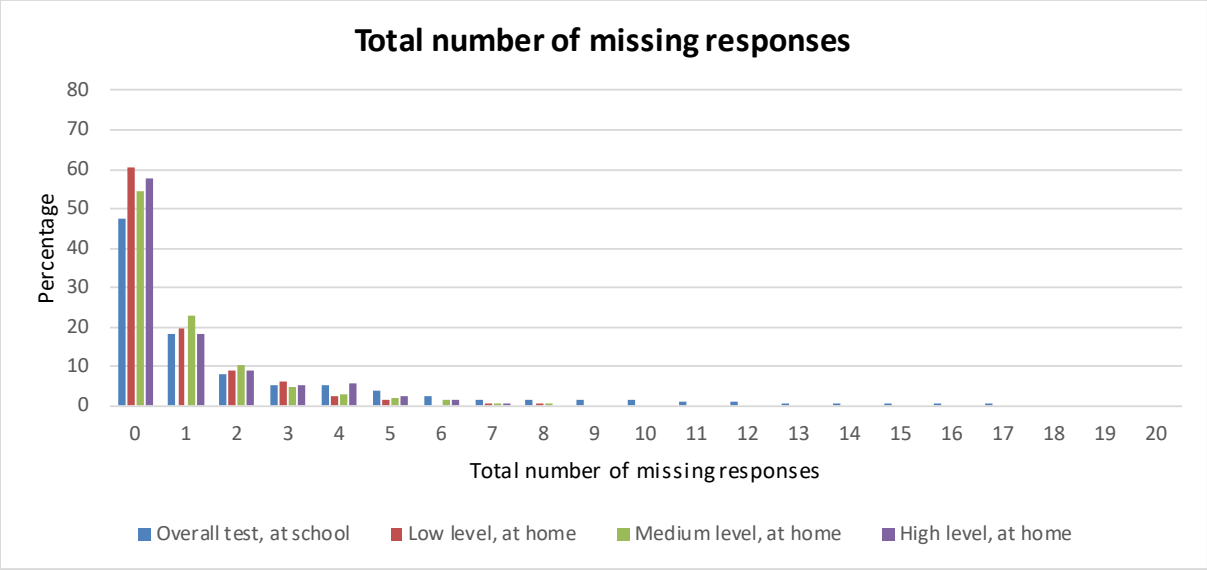


Figure 5. Total number of missing responses by test difficulty.

Overall, the amount of invalid items was small, whereas a reasonable part of missing responses occurred due to omitted items. The number of not reached items was, however, rather large and had the greatest impact on the total number of missing responses.

5.1.2 Missing responses per item

Tables 5 and 6 provide information on the occurrence of different kinds of missing responses per item by assessment setting (at school and at home) and difficulty of the administered test. Overall, in all of the three tests the omission rates were rather low. Across most items the omission rates vary between 0% and 5%. There were only six items with omission rates exceeding 5% (icg12016s_sc3g12_c in all experimental conditions, icg12047s_sc3g12_c in the overall test at school, icg12028s sc3g12_c, icg12056s sc3g12_c, and icg12050s sc3g12_c in the test with low level of difficulty, icg12046s_sc3g12_c in the tests with medium and high level of difficulty). The omission rates correlated with the item difficulties at about .10 in the overall test at school, and about .41 in the test administered individually at home. Generally, the percentage of invalid responses per item (only available for the overall test at school) per item (column 6 in Table 5) was rather low with the maximum rate being 1.2%.

With an item's progressing position in the test, the amount of persons that did not reach the item (column 4 in Table 5, columns 4, 7 and 10 in Table 6) rose up to a reasonable amount of 16.5% to 27.0% for the different experimental conditions. Particularly, the last items of the overall test at school and of the tests with medium or high level of difficulty were not reached by all respondents (see Figure 6).

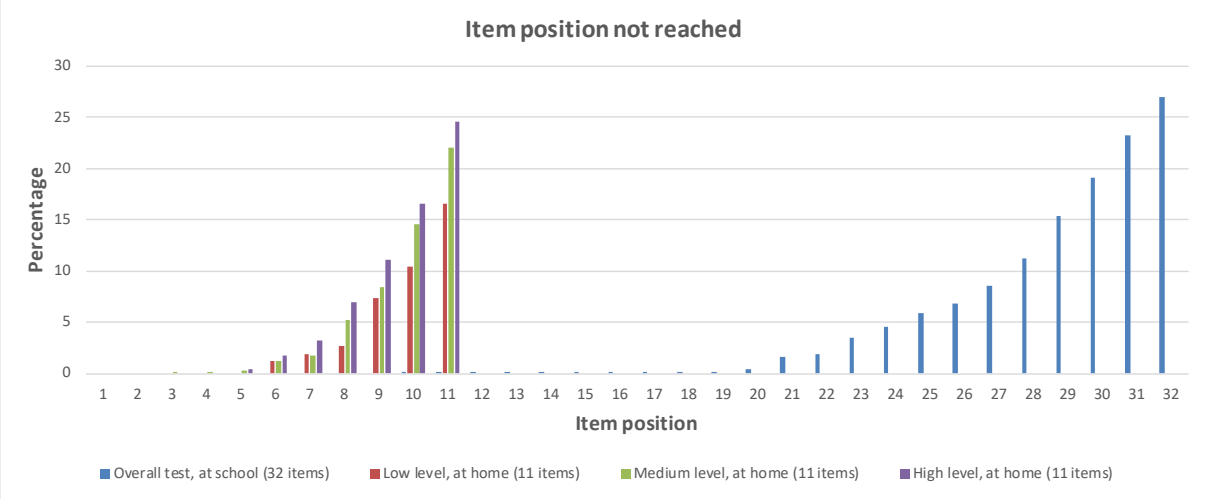


Figure 6. Item position not reached by test difficulty.

Table 5

Percentage of Missing Values for the Overall Test at School

Item	Position	N	NR	OM	NV
icg12018s_sc3g12_c	1	1749	0.00	0.62	0.11
ica4003x_sc3g12_c	2	1738	0.00	0.17	1.19
icg12107s_sc3g12_c	3	1745	0.00	0.91	0.06
icg12004s_sc3g12_c	4	1686	0.00	4.14	0.17
icg12010x_sc3g12_c	5	1752	0.00	0.40	0.17
icg12011x_sc3g12_c	6	1745	0.00	0.85	0.11
ica4008x_sc3g12_c	7	1720	0.00	1.19	1.19
icg12060s_sc3g12_c	8	1734	0.00	1.59	0.00
icg12013s_sc3g12_c	9	1748	0.00	0.45	0.34
ica4018s_sc3g12_c	10	1732	0.06	1.42	0.23
icg12016s_sc3g12_c	11	1626	0.11	7.55	0.06
ica4019x_sc3g12_c	12	1745	0.11	0.17	0.68
icg12121x_sc3g12_c	13	1730	0.11	1.36	0.34
icg12028s_sc3g12_c	14	1710	0.11	2.84	0.00
ica4023x_sc3g12_c	15	1753	0.11	0.28	0.11
ica4027x_sc3g12_c	16	1748	0.11	0.62	0.06
icg12033x_sc3g12_c	17	1726	0.11	1.80	0.20
icg12034x_sc3g12_c	18	1753	0.11	0.30	0.10
icg12035x_sc3g12_c	19	1694	0.20	3.60	0.10
icg12040x_sc3g12_c	20	1741	0.45	0.62	0.11
icg12037s_sc3g12_c	21	1624	1.59	6.19	0.06
icg12138s_sc3g12_c	22	1698	1.93	1.59	0.11
icg12047s_sc3g12_c	23	1594	3.46	6.02	0.06
icg12041x_sc3g12_c	24	1637	4.54	2.44	0.11

icg12046s_sc3g12_c	25	1598	5.90	3.35	0.06
ica4021s_sc3g12_c	26	1618	6.81	1.36	0.00
ica4052s_sc3g12_c	27	1584	8.51	1.53	0.06
icg12048s_sc3g12_c	28	1541	11.24	1.19	0.11
icg12050s_sc3g12_c	29	1447	15.38	2.44	0.06
icg12054s_sc3g12_c	30	1388	19.07	2.04	0.11
icg12109s_sc3g12_c	31	1333	23.21	0.91	0.23
icg12119s_sc3g12_c	32	1274	26.96	0.68	0.06

Note. Position = Item position within test, *N* = Number of valid responses, NR = Percentage of respondents that did not reach item, OM = Percentage of respondents that omitted the item, NV = Percentage of respondents with an invalid response.

Table 6

Percentage of Missing Values for the Test at home by Test Difficulty

		Low level			Medium level			High level		
Item	**Pos.**	**N**	**NR**	**OM**	**N**	**NR**	**OM**	**N**	**NR**	**OM**
icg12010x_sc3g12_c	1				1199	0.00	0.50	520	0.00	0.19
icg12028s_sc3g12_c	1	240	0.00	7.69						
ica4008x_sc3g12_c	2	256	0.00	1.54	1201	0.00	0.33			
ica4003x_sc3g12_c	2							521	0.00	0.00
icg12016s_sc3g12_c	3	233	0.00	10.38	1031	0.08	14.36	469	0.00	9.98
ica4027x_sc3g12_c	4				1195	0.08	0.75	520	0.00	0.19
icg12034x_sc3g12_c	4	266	0.00	0.00						
icg12041x_sc3g12_c	5	260	0.00	0.00	1201	0.25	0.08			
icg12011x_sc3g12_c	5							517	0.38	0.38
icg12046s_sc3g12_c	6	236	1.15	8.08	1024	1.24	13.78	451	1.73	11.71
icg12035x_sc3g12_c	7				1161	1.74	1.91	499	3.26	0.96
icg12138s_sc3g12_c	7	248	1.92	2.69						
icg12109s_sc3g12_c	8	249	2.69	1.54	1106	5.23	2.99			
ica4019x_sc3g12_c	8							485	6.91	0.00
icg12054s_sc3g12_c	9	231	7.31	3.85	1055	8.38	4.07	436	11.13	5.18
ica4052s_sc3g12_c	10				985	14.52	3.73	414	16.51	4.03
icg12050s_sc3g12_c	10	218	10.38	5.77						
icg12119s_sc3g12_c	11	217	16.54	0.00	840	21.99	0.00			
ica4023x_sc3g12_c	11							393	24.57	0.00

Note. Pos. = Item position within test, *N* = Number of valid responses, NR = Percentage of respondents that did not reach item, OM = Percentage of respondents that omitted the item.

5.2 Parameter Estimates

5.2.1 Item parameters

The fourth column in Table 7 presents the percentage of correct responses in relation to all valid responses for each item. Because there was a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index for item difficulty. The percentage of correct responses within dichotomous items varied between 26.1% and 79.3% with an average of 49.9% (*SD* = 16.8%) correct responses.The estimated item difficulties (for dichotomous variables) and location parameters (for the polytomous variable) are given in Table 7. The step parameters for the polytomous variable are depicted in Table 8. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. Because the students that left their original school received simulation-based test items within a computer-based test environment (students that remained at the same school received paper-pencil-based test items; see Section 3.1), their estimated item difficulties were corrected for differences in testing modes (mode effect: 0.093 logit) that was derived in an unpublished developmental study.

Table 7

Item parameters

Item	Pos. 1	Pos. 2	% correct	Item Difficulty	*SE*	WM NSQ	*t*	r_{it}	Discr.	Q_3
icg12018s_sc3g12_c	1		n.a.	-1.22	0.06	1.02	0.7	0.20	0.31	0.02
ica4003x_sc3g12_c	2	2	26.07	1.22	0.05	1.00	0.0	0.30	0.55	0.03
icg12107s_sc3g12_c	3		n.a.	-0.21	0.07	0.97	-1.3	0.33	1.00	0.03
icg12004s_sc3g12_c	4		n.a.	0.04	0.04	0.99	-0.3	0.40	1.32	0.02
icg12010x_sc3g12_c	5	1	54.22	-0.15	0.04	1.02	2.4	0.32	0.42	0.03
icg12011x_sc3g12_c	6	5	34.08	0.81	0.05	0.97	-1.8	0.37	0.84	0.02
ica4008x_sc3g12_c	7	2	65.56	-0.70	0.04	1.05	3.3	0.26	0.24	0.03
icg12060s_sc3g12_c	8		n.a.	-0.43	0.05	1.02	1.3	0.27	0.46	0.02
icg12013s_sc3g12_c	9		n.a.	-1.39	0.06	1.04	1.1	0.15	0.24	0.02
ica4018s_sc3g12_c	10		n.a.	0.60	0.04	1.07	3.3	0.26	0.36	0.03
icg12016s_sc3g12_c	11	3	n.a.	-0.42	0.04	1.03	1.9	0.24	0.34	0.03
ica4019x_sc3g12_c	12	8	27.67	1.14	0.05	1.00	-0.1	0.29	0.55	0.03
icg12121x_sc3g12_c	13		36.24	0.74	0.05	1.00	-0.3	0.31	0.59	0.03
icg12028s_sc3g12_c	14	1	n.a.	-1.90	0.07	1.00	0.0	0.22	0.52	0.03

ica4023x_sc3g12_c	15	11	56.43	-0.17	0.05	1.02	1.7	0.31	0.48	0.02
ica4027x_sc3g12_c	16	4	47.99	0.12	0.04	1.01	1.0	0.34	0.53	0.02
icg12033x_sc3g12_c	17		69.70	-0.76	0.05	1.01	0.3	0.29	0.58	0.02
icg12034x_sc3g12_c	18	4	79.28	-1.38	0.06	0.98	-0.7	0.32	0.75	0.02
icg12035x_sc3g12_c	19	7	55.16	-0.19	0.04	1.05	4.9	0.26	0.21	0.03
icg12040x_sc3g12_c	20		36.13	0.74	0.05	1.01	0.5	0.29	0.56	0.02
icg12037s_sc3g12_c	21		n.a.	-0.77	0.07	0.95	-2.2	0.39	1.30	0.03
icg12138s_sc3g12_c	22	7	n.a.	0.61	0.06	1.03	1.5	0.21	0.40	0.03
icg12047s_sc3g12_c	23		n.a.	0.15	0.04	0.96	-1.5	0.47	1.63	0.03
icg12041x_sc3g12_c	24	5	60.56	-0.47	0.04	1.02	1.6	0.33	0.48	0.03
icg12046s_sc3g12_c	25	6	n.a.	-0.59	0.03	0.97	-1.6	0.53	2.06	0.04
ica4021s_sc3g12_c	26		n.a.	-0.94	0.06	0.96	-1.4	0.36	0.90	0.02
ica4052s_sc3g12_c	27	10	n.a.	0.02	0.04	0.95	-2.8	0.45	1.63	0.04
icg12048s_sc3g12_c	28		n.a.	-0.57	0.05	0.99	-0.3	0.35	0.88	0.03
icg12050s_sc3g12_c	29	10	n.a.	-0.96	0.05	0.91	-3.0	0.50	1.99	0.03
icg12054s_sc3g12_c	30	9	n.a.	-0.44	0.04	0.96	-3.2	0.42	1.18	0.04
icg12109s_sc3g12_c	31	8	n.a.	-0.63	0.04	1.03	1.3	0.28	0.59	0.03
icg12119s_sc3g12_c	32	11	n.a.	-0.95	0.04	0.93	-2.7	0.53	2.06	0.04

Note. Pos. 1 = Item position within the overall test at school, Pos. 2 = Item position within the test versions of low, medium, and high level of difficulty; Difficulty = Item difficulty / location parameter, *SE* = standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, r_{it} = Corrected item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model, Q_3 = Average absolute residual correlation for item (Yen, 1993). Percent correct scores are not informative for polytomous CMC item scores. These are denoted by n.a. The item-total correlation corresponds to the product-moment correlation between the corresponding categories and the total score (discrimination value as computed in ConQuest).

The estimated item difficulties (or location parameters for the polytomous variable) ranged from -1.90 (item icg12028s_sc3g12_c) to 1.22 (item ica4003x_sc3g12_c) with an average difficulty of -0.28. Overall, the item difficulties were rather low, there were no items with a high difficulty. Due to the large sample size the standard errors (*SE*) of the estimated item difficulties (column 4 in Table 6) were rather small (all *SE*s ≤ 0.07).

Table 8

Step parameters (with Standard Errors) for the Polytomous Items

Item	Step 1	Step 2	Step 3	Step 4	Step 5
icg12107s_sc3g12_c	-0.643	0.643			
	(0.048)				
icg12004s_sc3g12_c	-0.087	-1.101	1.048	0.140	
	(0.054)	(0.049)	(0.067)		
Icg12013s_sc3g12_c	2.711	-2.711			
	(0.154)				
Ica4018s_sc3g12_c	1.100	-0.160	-0.940		
	(0.056)	(0.075)			
Icg12016s_sc3g12_c	0.092	-0.092			
	(0.038)				
Icg12028s_sc3g12_c	0.125	-0.125			
	(0.055)				
Icg12037s_sc3g12_c	-0.296	0.296			
	(0.051)				
Icg12138s_sc3g12_c	0.101	-0.101			
	(0.050)				
Icg12047s_sc3g12_c	-0.161	-0.401	-0.040	0.603	
	(0.054)	(0.051)	(0.058)		
icg12046s_sc3g12_c	-0.413	-0.406	0.050	0.122	0.647
	(0.040)	(0.036)	(0.035)	(0.040)	
Ica4052s_sc3g12_c	-0.391	0.052	0.340		
	(0.037)	(0.041)			
icg12048s_sc3g12_c	0.404	-0.370	-0.034		
	(0.052)	(0.057)			
icg12050s_sc3g12_c	0.035	-0.111	0.075		
	(0.050)	(0.055)			
Icg12054s_sc3g12_c	0.449	-0.449			

	(0.043)			
Icg12109s_sc3g12_c	-0.313	-0.596	0.909	
	(0.040)	(0.039)		
icg12119s_sc3g12_c	-0.262	-0.245	0.159	0.347
	(0.042)	(0.041)	(0.045)	

Note. The last step parameter is not estimated and has, thus, no standard error because it is a constrained parameter for model identification.

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

Figure 7. Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 21.2 cases. Item difficulty is depicted on the right side of the graph. Each number represents one item (see Table 7).

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. In Figure 7, item difficulties of the computer literacy items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of item difficulties.

The mean of the ability distribution was constrained to be zero. The variance was estimated to be 0.36, indicating somewhat limited variability between subjects. The reliability of the test (EAP/PV reliability = .65; WLE reliability = .65) was acceptable. Although the items covered a wide range of the ability distribution, there were no items to cover the lower and upper peripheral ability areas. As a consequence, person ability in medium ability regions will be measured relative precisely, whereas lower and higher ability estimates will have larger standard errors of measurement.

5.3 Quality of the Test

5.3.1 Fit of the subtasks of complex multiple choice items

Before the subtasks of the CMC item were aggregated and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the single subtasks together with the MC items in a Rasch model. Counting the subtasks of the CMC item separately, there were 108 items. The probability of a correct response ranged from 26.1% to 99.1% across all items (*Mdn* = 77.9%). Thus, the number of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.89 to 1.11, the respective *t*-value from -7.8 to 8.1, and there were no noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. Due to the good model fit of the subtasks, their aggregation to a polytomous variable seems to be justified.

5.3.2 Distractor analyses

In addition to the overall item fit, we specifically investigated how well the distractors performed in the test by evaluating the point-biserial correlation between each incorrect response (distractor) and the students' total score. All distractors had a point-biserial correlation with the total scores below zero with the exception of one item with a point-biserial-correlation of .00 (Median = -.16). The results indicate that the distractors worked well.

5.3.3 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the partial credit model, using the MC items and the polytomous CMC item. Altogether, item fit can be considered to be very good (see Table 7). Values of the WMNSQ ranged from 0.91 (item icg12050s_sc3g12_c) to 1.07 (item ica4018s_sc3g12_c). No item exhibited a *t*-value of the WMNSQ greater than 6. Thus, there was no indication of severe item over- or underfit. Point-biserial correlations between the item scores and the total scores ranged from .15 (item icg12013s_sc3g12_c) to .53 (items icg12046s_sc3g12_c and icg12119s_sc3g12_c) and had a mean of .33. All item characteristic curves showed a good fit of the items to the PCM.

5.3.4 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables gender, the number of books at home (as a proxy for socioeconomic status), migration background, school type, and test position (see Pohl & Carstensen, 2012, for a description of these variables). The differences between the estimated item difficulties in the various groups are summarized in Table 9. For example, the column "Male vs. female" reports the differences in item difficulties between men and women; a positive value would indicate that the test was more difficult for males, whereas a negative value would highlight a lower difficulty for males as opposed to females. Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF to those that only estimate main effects (see Table 10). Furthermore, the effect of the experimental factor assessment setting was also studied. Thus, we examined measurement invariance for the two assessment settings (for the common test items that were administered at school and at home) by adopting the minimum effect null hypothesis described in Fischer, Rohm, Gnambs, and Carstensen (2016). In addition, the effect of the experimental factor test difficulty (booklet) was also studied for the students that left their original school and that were individually tested at home. Thus, we examined measurement invariance for the three test versions (test with low, medium or high level of difficulty) by adopting the minimum effect null hypothesis described in Fischer et al. (2016). For this purpose we considered the common items of the test versions 1 and 2 (tests with low and medium level of difficulty) and of the test versions 2 and 3 (tests with medium and high level of difficulty; see Table 11).

Gender: The sample included 1,820 (48.6%) males and 1,864 (49.7%) females. Gender information was not available for 64 participants (1.7%) On average, male participants had a higher estimated computer literacy than females (main effect = 0.192 logits, Cohen's *d* = 0.538). However, three items (items icg12121x_sc3g12_c, icg12034x_sc3g12_c, and icg12054s_sc3g12_c) showed DIF greater than 0.6 logits. An overall test for DIF (see Table 10) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). Model comparisons using Akaike's (1974) information criterion (AIC) and the Bayesian information criterion (BIC; Schwarz, 1978) both favored the model estimating DIF. The deviation was rather small in both cases. Thus, overall, there was no pronounced DIF with regard to gender.

Books: The number of books at home was used as a proxy for cultural capital. There were 1,046 (27.9%) test takers with 0 to 100 books at home and 2,272 (60.6%) test takers with more than 100 books at home. 430 (11.5%) test takers had no valid response and were excluded from the analysis. There was a considerable average difference between the two groups. Participants with 100 or less books at home performed on average -0.414 logits (Cohen's *d* =-1.160) lower in computer literacy than participants with more than 100 books. However, there was no considerable DIF on the item level with the exception of one item that showed greater DIF than 0.6 logits (icg12107s_sc3g12_c). A model comparison using Akaike's (1974) information criterion (AIC) favored the model estimating DIF, whereas the Bayesian information criterion (BIC; Schwarz, 1978) that takes the number of estimated parameters into account and, thus, guards against overparameterization of models, indicated a better fit for the more parsimonious model including only the main effect. Thus, overall, there was no pronounced DIF with regard to the number of books at home.

Table 9

Differential Item Functioning

Item	Gender	Books	Migration	School	Position	Setting
	Male vs. female	< 100 vs. ≥ 100	Without vs. with	no sec. vs sec.	First vs. second	School vs. home
icg12018s_sc3g12_c	-0.044	-0.052	0.050	-0.036	-0.038	n.a.
ica4003x_sc3g12_c	-0.176	0.110	-0.370	0.170	-0.144	-0.292
icg12107s_sc3g12_c	-0.004	0.616	-0.144	0.296	0.140	n.a.
icg12004s_sc3g12_c	0.046	0.210	-0.180	0.028	-0.038	n.a.
icg12010x_sc3g12_c	0.136	-0.042	-0.138	0.060	0.062	0.006
icg12011x_sc3g12_c	-0.294	0.312	-0.170	-0.198	0.100	0.33
ica4008x_sc3g12_c	0.226	-0.338	0.074	-0.398	0.040	0.458
icg12060s_sc3g12_c	-0.074	0.006	0.262	-0.072	-0.008	n.a.
icg12013s_sc3g12_c	0.180	-0.100	0.264	-0.068	-0.074	n.a.
ica4018s_sc3g12_c	0.048	-0.364	0.072	-0.398	-0.034	n.a.
icg12016s_sc3g12_c	0.244	-0.190	0.378	-0.132	0.050	0.076
ica4019x_sc3g12_c	0.030	0.122	0.000	0.054	0.050	0.118
icg12121x_sc3g12_c	-0.648	-0.306	0.044	0.232	-0.092	n.a.
icg12028s_sc3g12_c	0.138	-0.086	-0.066	0.394	-0.124	-0.602
ica4023x_sc3g12_c	-0.250	0.056	-0.050	-0.164	0.118	0.112
ica4027x_sc3g12_c	0.250	-0.016	0.086	-0.086	-0.114	0.034
icg12033x_sc3g12_c	0.542	-0.004	-0.240	0.192	-0.180	n.a.
icg12034x_sc3g12_c	0.660	0.380	-0.244	0.658	-0.030	-0.942
icg12035x_sc3g12_c	0.416	-0.210	-0.158	-0.218	-0.070	0.2
icg12040x_sc3g12_c	-0.418	0.004	0.396	-0.040	-0.056	n.a.
icg12037s_sc3g12_c	-0.002	0.038	0.112	0.410	0.094	n.a.
icg12138s_sc3g12_c	-0.180	-0.292	0.330	-0.306	0.042	0.352

icg12047s_sc3g12_c	0.268	0.202	-0.114	0.230	-0.116	n.a.
icg12041x_sc3g12_c	-0.140	-0.260	-0.082	-0.134	0.158	0.15
icg12046s_sc3g12_c	-0.064	0.198	-0.226	0.134	-0.018	-0.206
ica4021s_sc3g12_c	-0.098	-0.188	0.516	-0.320	-0.002	n.a.
ica4052s_sc3g12_c	-0.180	0.012	0.050	-0.156	-0.016	0.264
icg12048s_sc3g12_c	0.532	0.044	0.154	-0.132	0.130	n.a.
icg12050s_sc3g12_c	-0.234	0.258	0.150	0.498	0.032	-0.684
icg12054s_sc3g12_c	-0.720	-0.032	0.204	0.140	0.044	-0.192
icg12109s_sc3g12_c	0.166	-0.354	-0.096	-0.266	0.018	0.37
icg12119s_sc3g12_c	-0.266	0.168	0.148	0.262	0.034	-0.316
Main effect	0.192	-0.414	0.202	-0.456	0.024	0.514

Note. Sec. = Secondary school (German: "Gymnasium").

Migration background: There were 2,528 participants (67.5%) with no migration background, 612 subjects (16.3%) with a migration background, and 608 individuals (16.2%) that did not indicate their migration background. In comparison to subjects with migration background, participants without migration background had on average a higher computer literacy (main effect = 0.202 logits, Cohen's *d* = 0.566). There was no noteworthy item DIF due to migration background; differences in estimated difficulties did not exceed 0.6 logits. Whereas the AIC favored the model estimating DIF, the BIC favored the main effects model (Table 10). Since the BIC takes the number of estimated parameters into account and guards against overparameterization of models, thus, overall, there was no pronounced DIF with regard to migration background.

School type: Overall, 2,009 subjects (53.6%) who took the computer literacy test attended secondary school (German: "Gymnasium"), whereas 1,739 (46.4%) were enrolled in other school types. Subjects in secondary schools showed a higher computer literacy on average (0.456 logits; Cohen's *d* = 1.277) than subjects in other school types. There was no considerable DIF on the item level with the exception of one item that showed greater DIF than 0.6 logits (item icg12034x_sc3g12_c). Whereas the AIC favored the model estimating DIF, the BIC favored the main effects model (Table 10). Since the BIC takes the number of estimated parameters into account and guards against overparameterization of models, thus, overall, there was no pronounced DIF with regard to school type.

Position: The computer literacy test was administered in two different positions (see section 3.1 for the design of the study). A subsample of 1,871 (49.9%) persons received the computer literacy first and 1,877 (50.1%) respondents took the computer literacy test after having completed either the mathematics or the reading test. Differential item functioning due to the position of the test can, for example, occur if there are differential fatigue

effects for certain items. The results showed minor average effects of the item position. Subjects who received the computer literacy test first performed on average 0.024 logits (Cohen's *d* = 0.067) better than subjects who received the computer literacy test second. There was no DIF due to the position of the test in the booklet. The largest difference in difficulty between the two test design groups was 0.180 logits (item icg12033x_sc3g12_c). As a consequence, the overall test for DIF using the BIC favored the more parsimonious main effect model (Table 10).

Setting: The computer literacy test was administered in two different settings (see section 3.1 for the design of the study). A subsample of 1,762 (67%) persons received the computer literacy test in small groups at school, whereas 1,986 (33%) participants finished the test individually at their private homes. Subjects who finished the computer literacy test at school were on average 0.514 logits (Cohen's *d* = 1.448) better than those working at their private homes. However, this difference must not be interpreted as a causal effect of the administration setting because respondents were not randomly assigned to the different settings. Rather, it is likely that self-selection processes occurred, for example, because less proficient students were more likely to leave school and, consequently, were tested at home. More importantly, there was no noteworthy DIF due to the administration setting; all differences in item difficulties were smaller than 0.6 logits with the exception of three items (items icg12028s_sc3g12_c, icg12034x_sc3g12_c, icg12050s_sc3g12_c). In addition, and of greater importance, further investigation using the procedure described in Fischer et al. (2016) identified no significant DIF (inspecting the differences in item difficulties between the two assessment settings and the respective tests for measurement invariance based on the Wald statistic: The highest empirical F value (F_{max} = 45.13) was much lower than the critical F value (F_{0154} (1, 3,748) = 86.71; see also Appendix B). Thus, overall, there was no pronounced DIF with regard to the different settings.

Table 10

Differential Item Functioning

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Gender	main effect	3,684	127,192.03	64	127,320.03	127,714.01
	DIF		126,854.71	96	127,046.71	127,637.68
Books	main effect	3,318	121,705.74	64	121,833.74	122,224.60
	DIF		121,578.14	96	121,770.14	122,356.42
Migration	main effect	3,140	114,542.06	64	114,670.06	115,057.38
	DIF		114,461.15	96	114,653.15	115,234.14
School type	main effect	3,748	132,179.96	64	132,307.96	132,706.62
	DIF		132,024.18	96	132,216.18	132,814.16
Position	main effect	3,748	132,441.59	64	132,569.59	132,698.25
	DIF		132,415.39	96	132,607.39	133,205.37
Setting	main effect	3,748	94,060.97	40	94,140.97	94,383.05
	DIF		93,848.92	59	93,966.92	94,323.98

Test version (booklet): The computer literacy test for students that left their original school and that were individually tested at home was administered in three different test versions (see section 3.1 for the design of the study). A subsample of 260 (13.1%) persons received the computer literacy test with low level of difficulty, a subsample of 1,205 (60.7%) persons received the test with medium level of difficulty whereas 521 (26.2%) participants received the test with high level of difficulty. To examine measurement invariance we considered the common items of the test versions 1 and 2 (tests with low and medium level of difficulty) and the common items of the test versions 2 and 3 (tests with medium and high level of difficulty). Adopting the minimum effect null hypothesis described in Fischer et al. (2016) the examinations identified no significant DIF (inspecting the differences in item difficulties between the test versions and the respective tests for measurement invariance based on the Wald statistic; see Table 11). Thus, overall, there was no pronounced DIF with regard to the different test versions.

Table 11

Differential Item Functioning Analyses between the Test Versions

	Tests with low and medium level of diffiuclty			Tests with medium and high level of diffiuclty		
Item	**Δσ**	***SE*$_{Δσ}$**	***F***	**Δσ**	***SE*$_{Δσ}$**	***F***
Ica4008x_sc3g12_c	-0.10	0.14	0.51			
Icg12016s_sc3g12_c	-0.01	0.18	0.00	-0.25	0.14	3.36
Icg12041x_sc3g12_c	-0.06	0.14	0.15			
Icg12046s_sc3g12_c	0.17	0.11	2.54	0.05	0.08	0.29
Icg12054s_sc3g12_c	0.21	0.18	1.29	0.17	0.13	1.70
Icg12109s_sc3g12_c	-0.08	0.15	0.27			
Icg12119s_sc3g12_c	-0.13	0.12	1.14			
Icg12010x_sc3g12_c				-0.04	0.11	0.16
Ica4027x_sc3g12_c				-0.08	0.11	0.56
Icg12035x_sc3g12_c				-0.19	0.11	2.83
Icga4052s_sc3g12_c				0.35	0.12	8.44

Note. Δσ = Difference in item difficulty parameters; $SE_{Δσ}$ = 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 using an α of .05 is F_{0154} (1, 1,465) = 41.53 for the tests with low and medium level of difficulty and F_{0154} (1, 1,726) = 46.76 for the test with medium and high level of difficulty. A non-significant test indicates measurement invariance.

5.3.5 Rasch homogeneity

An essential assumption of the Rasch (1980) model is that all item-discrimination parameters are equal. In order to test this assumption, a generalized partial credit model (GPCM) that estimates discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 7), ranging from 0.21 (item icg12035x_sc3g12_c) to 2.06 (item icg12119s_sc3g12_c). The average discrimination parameter fell at 0.81. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 131,997.27, BIC =132,582.49) as compared to the PCM model (AIC = 132,568.33, BIC =132,960.76). Despite the empirical preference for the GPCM, the PCM model matches the theoretical conceptions underlying the test construction more adequately (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the partial credit model was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

5.3.6 Unidimensionality

The dimensionality of the test was investigated by specifying two different multidimensional models. The first model is based on the four process components, and the second model is based on the four different types of software applications. To estimate a multidimensional (MD) model based on the four process components, Gauss' estimation in ConQuest (nodes = 15) was used. The assignment of the test items to the subscales (process components, software applications) is depicted in Appendix B. However, please note, that the computer literacy test is conceptualized as a unidimensional construct.

The estimated variances and correlations between the four dimensions representing the different process components are reported in Table 12. The correlations among the dimensions varied between .84 and .93. The smallest correlation was found between Dimension 1 ("Access") and Dimension 4 ("Evaluate") and Dimension 3 ("Manage") and Dimension 4 ("Evaluate"), respectively. Dimension 2 ("Create") and Dimension 3 ("Manage") showed the strongest correlation. All correlations deviated from a perfect correlation (i.e., they were marginally lower than r = .95, see Carstensen, 2013). A model comparison using Akaike's (1974) information criterion (AIC) favored the four-dimensional model (AIC = 132,541.91, number of parameters = 72 vs. AIC = 132,568.33, number of parameters = 63), whereas the Bayesian information criterion (BIC; Schwarz, 1978) indicated a better fit for the unidimensional model (BIC = 132,960.76, number of parameters = 63 vs. BIC = 132,990.39, number of parameters = 72). These results indicate that the three cognitive requirements measure a common construct, albeit it is not completely unidimensional.

Table 12

Results of Four-Dimensional Scaling (Process Components)

	Access	**Create**	**Manage**	**Evaluate**
Access (7 Items)	(0.346)			
Create (8 Items)	.896	(0.472)		
Manage (9 Items)	.866	.925	(0.372)	
Evaluate (8 Items)	.840	.855	.840	(0.490)

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

The estimated variances and correlations for the four-dimensional model based on the different types of software applications reported in Table 13. The correlations among the three dimensions varied between .81 and .90. The smallest correlation was found between Dimension 3 ("E-mail / communication tools") and Dimension 4 ("Internet / search engines"). Dimension 2 ("Spreadsheet / presentation software") and Dimension 4 ("Internet / search engines") showed the strongest correlation. However, they deviated from a perfect correlation (i.e., they were marginally lower than r = .95, see Carstensen, 2013). A model comparison using Akaike's (1974) information criterion (AIC) favored the four-dimensional model (AIC = 132,523.29, number of parameters = 72 vs. AIC = 132,568.33, number of parameters = 63), whereas the Bayesian information criterion (BIC; Schwarz, 1978) indicated

a better fit for the unidimensional model (BIC = 132,960.76, number of parameters = 63 vs. BIC = 132,971.77, number of parameters = 72). These results indicate that the three cognitive requirements measure a common construct, albeit it is not completely unidimensional.

However, for the unidimensional model the average absolute residual correlations as indicated by the corrected Q_3 statistic (see Table 8) were quite low (M = .028, SD = .007) — the largest individual residual correlation was .141 — and thus indicated an essentially unidimensional test. Because the computer literacy test is constructed to measure a single dimension, a unidimensional computer literacy competence score was estimated.

Table 13

Results of Four-Dimensional Scaling (Software Applications).

	Dim 1	**Dim 2**	**Dim 3**	**Dim 4**
Word processing (Dim1) (6 Items)	(0.653)			
Spreadsheet / presentation software (Dim 2) (11 Items)	.865	(0.391)		
E-mail / communication tools (Dim 3) (5 Items)	.882	.864	(0.315)	
Internet / search engines (Dim 4) (10 Items)	.832	.904	.810	(0.323)

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

6. Discussion

The analyses in the previous sections aimed at providing detailed information on the quality of the computer literacy test in starting cohort 3 for grade 12 and at describing how computer literacy was estimated.

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

Various criteria indicated a good fit of the items and measurement invariance across various subgroups. However, the amount of not-reached items was rather high, indicating that the test was too long for the allocated testing time. Other types of missing responses were reasonably small.

The test had a high reliability but a somewhat limited variance. However, the test was mainly targeted at low-performing students and did not accurately measure computer literacy of high-performing students. As a consequence, ability estimates will be precise for low-performing students but less precise for high performing students.

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

7. Data in the Scientific Use File

7.1 Naming conventions

The data in the Scientific Use File contain 32 items, of which 13 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. A total of 19 items were scored as polytomous variables (CMC items). MC items are marked with a 'x_c' at the end of the variable name, whereas the variable names of CMC items end in 's_c'. In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category.

7.2 Linking of competence scores

In starting cohort 3, the computer literacy administered in grades 9 (see Senkbeil & Ihme, 2017a) and 12 include different items that were constructed in such a way as to allow for an accurate measurement of computer literacy within each age group. As a consequence, the competence scores derived in the different grades cannot be directly compared; differences in observed scores would reflect differences in competences as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competences across grades, we adopted the linking procedure described in Fischer et al. (2016). Following an anchor-group design, an independent link sample including students from grade 11 that were not part of starting cohort 4 were administered all items from the grade 9 and the grade 12 computer literacy tests within a single measurement occasion. These responses were used to link the two tests administered in starting cohort 3 across the two grades.

7.2.1 Samples

In starting cohort 3, a subsample of 2,855 students participated at both measurement occasions, in grade 6 and also in grade 9. Consequently, these respondents were used to link the two tests across both grades (see Fischer et al., 2016.). Moreover, an independent link sample of *N* = 398 students from grade 9 received both tests within a single measurement occasion.

7.2.2 The design of the link study

The students of the link study responded to 24 common items from the test versions with low, medium, and high level of difficulty administered in grade 9 (see Senkbeil & Ihme, 2017a) and to 32 items of the grade 12 computer literacy test (see above). Because preliminary analyses identified severe differential item functioning for one item of the grade 12 test (ica4018s_sc4g12_c) between the link sample and the longitudinal main sample, this item was removed from the final linking procedure. Moreover, the computer literacy test was

administered at different positions in the test battery. A random sample of 204 students received the computer literacy test before working on a reading test, whereas the remaining 194 students received the reading test before the computer literacy test. No multi-matrix design regarding the selection and order of the items within a test was established. Thus, all test takers were given the computer literacy items in the same order.

7.2.3 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 two tests. According to model fit indices, the BIC favored the unidimensional model (BIC = 33,736.69, number of parameters = 112; two-dimensional model: BIC = 33,851.82, number of parameters = 114), whereas the AIC favored the two-dimensional model (AIC = 33,283.36; unidimensional model: AIC = 33,290.20). Because the differences in the information criteria between the unidimensional model and the two-dimensional model were very small and, therefore, negligible, the results indicate that the computer literacy tests administered in grades 9 and 12 were essentially unidimensional.

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal subsample from the starting cohort. The differences in item difficulties between the link sample and starting cohort 3 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Tables 14 (Grade 9) and 15 (Grade 12). A positive value for the difference in item difficulty parameters indicates that the item is easier for the linking sample compared to the longitudinal main subsample, whereas a negative value indicates higher difficulty for the linking sample. Minimum effect hypothesis test revealed significant DIF (α = .05) for only one item (icg9103x_c). However, a couple of items exhibited considerable DIF greater than 0.6 logits and five items indicated strong DIF that was larger than 1 logit (Max = $|1.26|$). This concerns thirteen items from the grade 9 test (icg9103x_c, icg9106x_c, icg9107s_c, icg9113x_c, icg9114x_c, icg9117s_c, icg9119x_c, icg9122x_c, icg9123x_c, ICG9128x_c, icg9131x_c, icg9132x_c, icg9138x_c) and thirteen items from the grade 12 test (ica4003x_c, icg12107s_c, icg12010x_c, icg12011x_c, ica4008x_c, ica4019x_c, icg12028s_c, ica4023x_c, icg12034x_c, icg12035x_c, icg12047s_c, icg12109s_c, icg12119s_c). Therefore, these items were removed from the final linking procedure.

To apply the “mean/mean” linking method, the correction term was calculated as c = 0.312. Added to the correction term for grade 6 to 9 (see Senkbeil et al., 2014), a total correction term of 1.354 was derived. This correction was subsequently added to each difficulty parameter estimated in grade 12 (see Table 7) to derive the linked item parameters. The link error reflecting the uncertainty in the linking process was calculated according to equation 4 in Fischer et al. (2016) as 0.1013 and has to be included into the SE when statistical tests are used to compare groups concerning their mean change of ability between two linked measurements.

Table 14

Differential Item Functioning Analyses between the Starting Cohort and the Link Sample (Grade 9)

Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	***F***
icg9101x_sc3g9_c	0.05	0.13	0.15
icg9102S_sc3g9_c	0.59	0.16	13.81
icg9103X_sc3g9_c*	1.22	0.13	86.98
icg9106X_sc3g9_c*	1.26	0.20	41.41
icg9107S_sc3g9_c*	0.65	0.17	15.38
icg9110X_sc3g9_c	0.33	0.12	7.98
icg9111X_sc3g9_c	0.60	0.14	19.61
icg9113X_sc3g9_c*	0.73	0.12	39.72
icg9114X_sc3g9_c*	0.65	0.17	13.99
icg9116X_sc3g9_c	0.53	0.20	6.59
icg9117S_sc3g9_c*	0.84	0.18	20.88
icg9118X_sc3g9_c	0.49	0.14	12.46
icg9119X_sc3g9_c*	1.23	0.22	32.63
icg9122X_sc3g9_c*	0.73	0.13	30.83
icg9123X_sc3g9_c*	0.88	0.18	23.04
icg9125S_sc3g9_c	0.11	0.19	0.35
icg9128X_sc3g9_c*	0.63	0.12	25.80
icg9129X_sc3g9_c	0.32	0.13	5.96
icg9131X_sc3g9_c*	0.89	0.17	28.31
icg9132X_sc3g9_c*	0.99	0.18	29.94
icg9133S_sc3g9_c	0.43	0.11	14.64
Icg9136S_sc3g9_c	-0.22	0.09	6.78
icg9138X_sc3g9_c*	0.70	0.17	17.01
icg9140S_sc3g9_c	0.23	0.29	0.62

Note. * item removed from the linking procedure due to considerable DIF; $\Delta\sigma$ = Difference in item difficulty parameters between the longitudinal subsample in grade 9 and the link sample (positive values indicate easier items in the link sample); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis using an α of .05 is F_{0154} (1, 3,153) = 77.24. A non-significant test indicates measurement invariance.

Table 15

Differential Item Functioning Analyses between the Starting Cohort and the Link Sample (Grade 12)

Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	**F**
icg12018S_sc3g12_c	-0.37	0.14	6.64
ica4003X_sc3g12_c*	-0.93	0.14	44.49
icg12107S_sc3g12_c*	-0.80	0.16	25.87
icg12004S_sc3g12_c	-0.11	0.10	1.19
icg12010X_sc3g12_c*	-0.66	0.11	33.21
icg12011X_sc3g12_c*	-1.03	0.13	63.26
ica4008X_sc3g12_c*	-0.65	0.12	28.73
icg12060S_sc3g12_c	-0.47	0.12	14.61
icg12013S_sc3g12_c	-0.30	0.15	3.87
icg12016S_sc3g12_c	-0.51	0.14	12.71
ica4019X_sc3g12_c*	-0.89	0.13	43.31
icg12121X_sc3g12_c	-0.18	0.12	2.31
icg12028S_sc3g12_c*	-1.10	0.16	44.96
ica4023X_sc3g12_c*	-0.80	0.12	47.93
ica4027X_sc3g12_c	-0.37	0.11	10.60
icg12033X_sc3g12_c	-0.43	0.13	10.86
icg12034X_sc3g12_c*	-0.84	0.14	38.85
icg12035X_sc3g12_c*	-0.64	0.12	29.08
icg12040X_sc3g12_c	0.32	0.12	7.47
icg12037S_sc3g12_c	-0.32	0.16	3.74
icg12138S_sc3g12_c	0.44	0.14	9.98
icg12047S_sc3g12_c*	-0.74	0.10	56.26
icg12041X_sc3g12_c	-0.24	0.13	3.41
icg12046S_sc3g12_c	-0.60	0.09	47.53
ica4021S_sc3g12_c	-0.57	0.14	16.84
ica4052S_sc3g12_c	-0.06	0.12	0.23

icg12048S_sc3g12_c	-0.34	0.12	8.58
icg12050S_sc3g12_c	-0.20	0.13	2.22
icg12054S_sc3g12_c	-0.04	0.16	0.08
icg12109S_sc3g12_c*	-0.86	0.12	49.80
icg12119S_sc3g12_c*	-0.63	0.11	30.14

Note. * item removed from the linking procedure due to considerable DIF; $\Delta\sigma$ = Difference in item difficulty parameters between the longitudinal subsample in grade 12 and the link sample (positive values indicate easier items in the link sample); $SE_{\Delta\sigma}$ = Pooled standard error; F = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis using an α of .05 is F_{0154} (1, 3,153) = 77.24. A non-significant test indicates measurement invariance.

7.3 Computer literacy scores

Person abilities were subsequently estimated using the linked item difficulty parameters. In the SUF, manifest scale scores are provided in the form of two different WLE estimates, ”icg12_sc1” and “icg12_sc1u”, including their respective standard errors “icg12_sc2” and “icg12_sc2u”. The corrected score “icg12_sc1” was corrected for the position of the computer literacy test within the booklet and can be used, if the research interest lies on cross-sectional issues. (Note that the WLE scores in “icg12_sc1” are not linked to the underlying reference scale of grade 9.) The uncorrected score “icg12_sc1u” (uncorrected for the position of the reading test within the booklet) can be used, if the focus of the research lies on longitudinal issues, such as competence development since differences in WLE scores can be interpreted as development trajectories across measurement points. The ConQuest Syntax for estimating the WLE is provided in Appendix A. For persons who either did not take part in the computer literacy test or who did not give enough valid responses, no WLE is estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values.

Users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values. Plausible values for competence tests administered in the NEPS can be estimated using the R package *NEPSscaling*[1] (Scharl, Carstensen, & Gnambs, 2020).

[1] https://www.neps-data.de/Data-Center/Overview-and-Assistance/Plausible-Values

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ConQuest 4*. Camberwell, Australia: Acer.

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

Blossfeld, H.-P., Roßbach, H.-G, & von Maurice, J. (Eds.) (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). [Special Issue] Zeitschrift für Erziehungswissenschaft: 14.

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*. New York, NY: Springer.

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

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)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

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

Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement, 50*, 447-468.

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg: University of Bamberg, National Educational Panel Study.

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

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

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

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

Senkbeil, M., Ihme, J. M., & Adrian, E. (2014). *NEPS Technical Report for Computer Literacy – Scaling Results of Starting Cohort 3 in Grade 6* (NEPS Working Paper No. 39). Bamberg: University of Bamberg, National Educational Panel Study.

Senkbeil, M., & Ihme, J. M. (2017a). *NEPS Technical Report for Computer Literacy: Scaling results of Starting Cohort 3 for Grade 9* (NEPS Survey Paper No. 29). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Senkbeil, M., & Ihme, J. M. (2017b). *NEPS Technical Report for Computer Literacy: Scaling results of Starting Cohort 4 for Grade 12* (NEPS Survey Paper No. 25). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011) Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaften, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS) (pp. 67–86.)* Wiesbaden: VS Verlag für Sozialwissenschaften.

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

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

Appendix

Appendix A: ConQuest-Syntax for estimating WLE estimates in Starting Cohort 3 (grade 12)

```
title SC3 G12 Computer Literacy partial credit model;

/* load data */
datafile  >>filename.dat;
format pid 1-7 responses 9-40;
labels <<filename_with_labels.txt;

/* collapse response categories */
codes 0,1,2,3,4,5,6,;
recode (0,1,2)          (0,1,2)          !item(21);     /* icg12037s_sc3g12_c */
recode (0,1,2,3)        (0,0,0,1)        !item(9);      /* icg12013s_sc3g12_c */
recode (0,1,2,3,4)      (0,0,0,0,1)      !item(1);      /* icg12018s_sc3g12_c */
recode (0,1,2,3,4)      (0,0,0,0,1)      !item(8);      /* icg12060s_sc3g12_c */
recode (0,1,2,3,4)      (0,0,0,1,2)      !item(11);     /* icg12016s_sc3g12_c */
recode (0,1,2,3,4)      (0,0,0,1,2)      !item(22);     /* icg12138s_sc3g12_c */
recode (0,1,2,3,4)      (0,0,0,1,2)      !item(30);     /* icg12054s_sc3g12_c */
recode (0,1,2,3,4)      (0,0,1,2,3)      !item(31);     /* icg12109s_sc3g12_c */
recode (0,1,2,3,4,5)    (0,0,0,0,0,1)    !item(26);     /* ica4021s_sc3g12_c */
recode (0,1,2,3,4,5)    (0,0,0,0,1,2)    !item(3);      /* icg12107s_sc3g12c */
recode (0,1,2,3,4,5)    (0,0,0,0,1,2)    !item(14);     /* icg12028s_sc3g12_c */
recode (0,1,2,3,4,5)    (0,0,0,1,2,3)    !item(27);     /* ica4052s_sc3g12_c */
recode (0,1,2,3,4,5)    (0,0,0,1,2,3)    !item(28);     /* icg12048s_sc3g12_c */
recode (0,1,2,3,4,5)    (0,0,1,2,3,4)    !item(32);     /* icg12119s_sc3g12_c */
recode (0,1,2,3,4,5,6)  (0,0,0,0,1,2,3)  !item(29);     /* icg12050s_sc3g12_c */
recode (0,1,2,3,4,5,6)  (0,0,0,1,2,3,4)  !item(4);      /* icg12004s_sc3g12_c */
recode (0,1,2,3,4,5,6)  (0,0,0,1,2,3,4)  !item(23);     /* icg12047s_sc3g12_c */
recode (0,1,2,3,4,5,6)  (0,0,1,2,3,4,5)  !item(25);     /* icg12046s_sc3g12_c */
recode (0,1,2,3,4,5,6,7,8)  (0,0,1,1,1,1,2,2,3)  !item(10);  /* ica4018s_sc3g12_c */

/* scoring */
score (0,1)             (0,1)            !item(1,2,5-7,8,12,13,15-20,24,26);
score (0,1,2)           (0,.5,1)         !item(3,9,11,14,21,22,30);
score (0,1,2,3)         (0,.5,1,1.5)     !item(10,27,28,29,31);
score (0,1,2,3,4)       (0,.5,1,1.5,2)   !item(4,23,32);
score (0,1,2,3,4,5)     (0,.5,1,1.5,2,2.5)  !item(25);

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

/* estimate model */
estimate ! method=gauss, nodes = 15; iterations = 1000; convergence = 0.0001;

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

Appendix B: Differential Item Functioning Analyses between the Assessment Settings (test administered at school vs. test administered at home) for the common test items that were administered at school and at home

Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	F
ica4003x_sc3g12_c	-0.25	0.13	3.69
icg12010x_sc3g12_c	0.06	0.07	0.74
icg12011x_sc3g12_c	0.38	0.11	12.31
ica5008x_sc4g12_c	0.52	0.08	45.13
icg12016s_sc3g12_c	0.11	0.09	1.71
ica4019x_sc3g12_c	0.16	0.12	1.75
icg12028s_sc3g12c	-0.60	0.18	11.57
ica4023x_sc3g12_c	0.17	0.12	2.08
ica4027x_sc3g12_c	0.09	0.07	1.62
icg12034x_sc3g12_c	-0.88	0.14	37.63
icg12035x_sc3g12_c	0.25	0.07	12.40
icg12138s_sc3g12_c	0.43	0.18	7.41
icg12041x_sc3g12_c	0.21	0.08	7.41
icg12046s_sc3g12_c	-0.14	0.05	6.91
ica4052s_sc3g12_c	0.21	0.07	8.46
icg12050s_sc3g12_c	-0.82	0.15	31.95
icg12054s_sc3g12_c	-0.17	0.09	3.61
icg12109s_sc3g12_c	0.50	0.08	38.35
icg12119s_sc3g12_c	-0.24	0.07	10.89

Note. $\Delta\sigma$ = Difference in item difficulty parameters between the school sample and the home sample (positive values indicate easier items in the school sample); $SE_{\Delta\sigma}$ = 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 using an α of .05 is F_{0154} (1, 3,748) = 86.71. A non-significant test indicates measurement invariance.

Appendix C: Assignment of test items to the Process Components and Software Applications

Item	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Response format	Process Component	Software Application
icg12018s_sc3g12_c	1				CMC	Manage	E-mail / communication
ica4003x_sc3g12_c	2			2	MC	Evaluate	Internet / search engines
icg12107s_sc3g12_c	3				CMC	Evaluate	Spreadsheet / presention
icg12004s_sc3g12_c	4				CMC	Create	E-mail / communication
icg12010x_sc3g12_c	5		1	1	MC	Create	Spreadsheet / presention
icg12011x_sc3g12_c	6			5	MC	Manage	Spreadsheet / presention
ica4008x_sc3g12_c	7	2	2		MC	Evaluate	Internet / search engines
icg12060s_sc3g12_c	8				CMC	Manage	Spreadsheet / presention
icg12013s_sc3g12_c	9				CMC	Manage	Internet / search engines
ica4018s_sc3g12_c	10				CMC	Manage	Internet / search engines
icg12016s_sc3g12_c	11	3	3	3	CMC	Access	Word processing
ica4019x_sc3g12_c	12			8	MC	Evaluate	Internet / search engines
icg12121x_sc3g12_c	13				MC	Access	Spreadsheet / presention
icg12028s_sc3g12_c	14	1			CMC	Access	E-mail / communication
ica4023x_sc3g12_c	15			11	MC	Create	Spreadsheet / presention
ica4027x_sc3g12_c	16		4	4	MC	Manage	E-mail / communication
icg12033x_sc3g12_c	17				MC	Manage	Spreadsheet / presention
icg12034x_sc3g12_c	18	4			MC	Access	Spreadsheet / presention
icg12035x_sc3g12_c	19		7	7	MC	Create	Spreadsheet / presention
icg12040x_sc3g12_c	20				MC	Access	Internet / search engines
icg12037s_sc3g12_c	21				CMC	Manage	Spreadsheet / presention
icg12138s_sc3g12_c	22	7			CMC	Access	E-mail / communication
icg12047s_sc3g12_c	23				CMC	Create	Word processing
icg12041x_sc3g12_c	24	5	5		MC	Manage	Word processing
icg12046s_sc3g12_c	25	6	6	6	CMC	Create	Spreadsheet / presention
ica4021s_sc3g12_c	26				CMC	Access	Word processing

ica4052s_sc3g12_c	27		10	10	CMC	Create	Word processing
icg12048s_sc3g12_c	28				CMC	Evaluate	Internet / search engines
icg12050s_sc3g12_c	29	10			CMC	Evaluate	Internet / search engines
icg12054s_sc3g12_c	30	9	9	9	CMC	Create	Word processing
icg12109s_sc3g12_c	31	8	8		CMC	Evaluate	Internet / search engines
icg12119s_sc3g12_c	32	11	11		CMC	Evaluate	Internet / search engines

Note. Pos. 1 = overall test, administered at school; Pos. 2 = booklet with low level of difficulty, administered at home; Pos 3. = booklet with medium level of difficulty, administered at home; Pos. 4 = booklet with high level of difficulty, administered at home