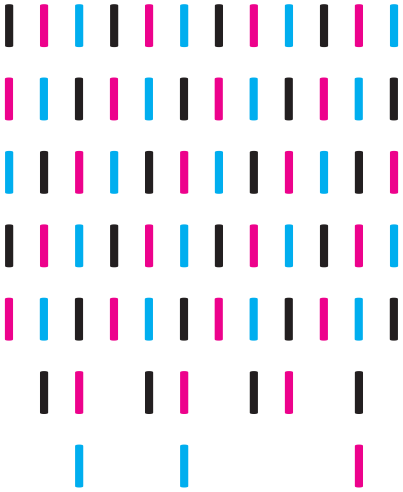

NEPS SURVEY PAPERS

Martin Senkbeil and Jan Marten Ihme

NEPS TECHNICAL REPORT FOR COMPUTER LITERACY: SCALING RESULTS OF STARTING COHORT 3 FOR GRADE 9

NEPS Survey Paper No. 29
Bamberg, November 2017

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

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

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

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

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

The NEPS Survey Papers are available at https://www.neps-data.de (see section “Publications“).

Editor-in-Chief: Corinna Kleinert, LIfBi/University of Bamberg/IAB Nuremberg

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

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

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. (2017). *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. doi:10.5157/NEPS:SP29:1.0

Acknowledgements:

This report is an extension to NEPS working paper 39 (Senkbeil, Ihme, & Adrian, 2014) that presents the scaling results for computer literacy of starting cohort 3 for grade 6. Therefore, various parts of this report (e.g., regarding the instruction and the analytic strategy) are reproduced verbatim from previous working papers (Senkbeil et al., 2014) to facilitate the understanding of the presented results.

We would like to thank Luise Fischer, Theresa Rohm, and 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 9 of starting cohort 3 (fifth grade). The computer literacy test contained 60 items (distributed among 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 4,877 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 but one fitted the model in a satisfactory way. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test were 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 4
2 Testing Computer Literacy 4
3 Data 5
3.1 The Design of the Study 5
3.2 Sample 7
4 Analyses 8
4.1 Missing Responses 8
4.2 Scaling Model 9
4.3 Checking the Quality of the Scale 9
5 Results 11
5.1 Missing Responses 11
5.1.1 Missing responses per person 11
5.1.2 Missing responses per item 13
5.2 Parameter Estimates 17
5.2.1 Item parameters 17
5.2.2 Test targeting and reliability 21
5.3 Quality of the Test 21
5.3.1 Fit of the subtasks of complex multiple choice items 21
5.3.2 Distractor analyses 21
5.3.3 Item fit 21
5.3.4 Differential item functioning 22
5.3.5 Rasch homogeneity 27
5.3.6 Unidimensionality 27
6 Discussion 29
7 Data in the Scientific Use File 30
7.2.1 Samples 30
7.2.2 The design of the link study 30
7.2.3 Results 30
7.3 Computer literacy scores 32
References 33
Appendix 35

1 Introduction

Within the National Educational Panel Study (NEPS), 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 (2016).

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 9. 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.

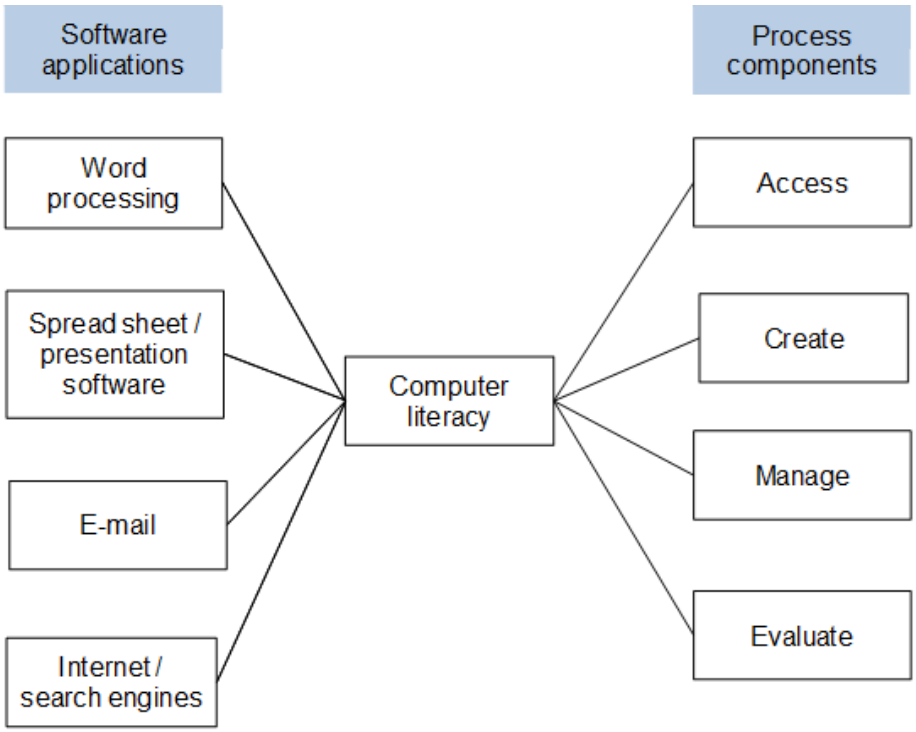

Figure 1. Assessment framework for computer literacy (process components and software applications).

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 9 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).

3 Data

3.1 The Design of the Study

The study followed a two-factorial (quasi-)experimental design. These factors referred to (a) the position of the computer literacy test within the test battery, and (b) the difficulty of the administered test.

The study assessed different competence domains including, among others, computer literacy and science. The competence tests for these two domains were always presented first within the test battery. In order to control for test position effects, the tests were administered to participants in a different sequence (see Table 4). For each participant the computer literacy test was either administered as the first or the second test (i.e., after the science test). There was no multi-matrix design regarding the order of the items within a specific test. All students received the test items in the same order. The competence test for computer literacy that was administered in the present study included 60 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.

In order to measure participants' computer literacy with great accuracy, the difficulty of the administered tests should adequately match the participants' abilities, Therefore, the study adopted the principals of longitudinal multistage testing (Pohl, 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 36 items that represented the four process components (see Table 1) and the four software applications (see Table 2). Twelve items were identical in all three test versions, twenty-four items were identical in the tests with low and medium level of difficulty, and twenty-four items were identical in the tests with medium and high level of difficulty (see Tables 1 and 2). Twelve items were unique to the test with low medium of difficulty and to the test with high level of difficulty (see Appendix B for the detailed assignment of the test items to each test version). The different response formats of the items are summarized in Table 3. Participants were assigned to the test version based on their computer literacy competence in the previous assessment (Senkbeil et al., 2014). Participants with an abilty estimate below a percentile rank of 33 received the test with a low level of difficulty, participants with an ability estimate between the percentile ranks of 34 and 66 receive the test with a medium level of difficulty, and participants with an abilty estimate above a percentile rank of 66 received the test with a high level of difficulty.

Table 1

Number of Items for the Different Process Components by Difficulty of the Test

Process components	**Low level**	**Medium level**	**High level**	**All tests**	**Low and medium level**	**Medium and high level**
Access	10	12	11	19	15	16
Create	7	6	7	13	9	10
Manage	8	11	11	14	12	13
Evaluate	11	7	7	14	12	9
Total number of items	36	36	36	60	48	48

Table 2

Number of Items for the Different Software Applications by Difficulty of the Test

Software applications	**Low level**	**Medium level**	**High level**	**All tests**	**Low and medium level**	**Medium and high level**
Word processing	10	9	9	17	13	13
Spreadsheet / presentation software	11	11	10	17	15	13
E-mail / communication tools	4	4	5	8	6	6
Internet / search engines	11	12	12	18	14	16
Total number of items	36	36	36	60	48	48

Table 3

Number of Items by Different Response Formats and Difficulty of the Test

Response format	**Low level**	**Medium level**	**High level**
Simple multiple choice items	33	29	22
Complex multiple choice items	3	7	14
Total number of items	36	36	36

3.2 Sample

A total of 4,877 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 4,876 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

	Test version:	**Low level**	**Medium level**	**High level**	**Total**
Test	**First position**	591	1222	637	2450
position	**Second position**	608	1105	713	2426
	Total	1199	2327	1350	4876

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 multi-stage design not all items were administered to all participants. For respondents receiving the test with low level of difficulty 24 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 12 items of the test with low level of difficulyt and 12 items of the test with high level of difficulty were missing by design, and for respondents receiving the test with high level of difficulty 24 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 all of the 16 CMC items categories were collapsed (see Appendix A). To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items were scored dichotomously as 0 for an incorrect and as 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats).

4.3 Checking the Quality of the 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 socioeconomic status), 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.

4.4 Software

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

5 Results

5.1 Missing Responses

5.1.1 Missing responses per person

Figure 2 shows the number of invalid responses per person by experimental condition (i.e., test difficulty). Overall, there were very few invalid responses. Between 95% and 96% of the respondents did not have any invalid response at all; overall less than two percent had more than one invalid response. There was only a slight difference in the amount of invalid responses between the different experimental conditions.

Missing responses may also occur when respondents omit items. As illustrated in Figure 3 most respondents, 61% to 69%, did not skip any item, and less than six percent omitted more than three items. There was only a slight difference in the amount of omitted items between the different experimental conditions.

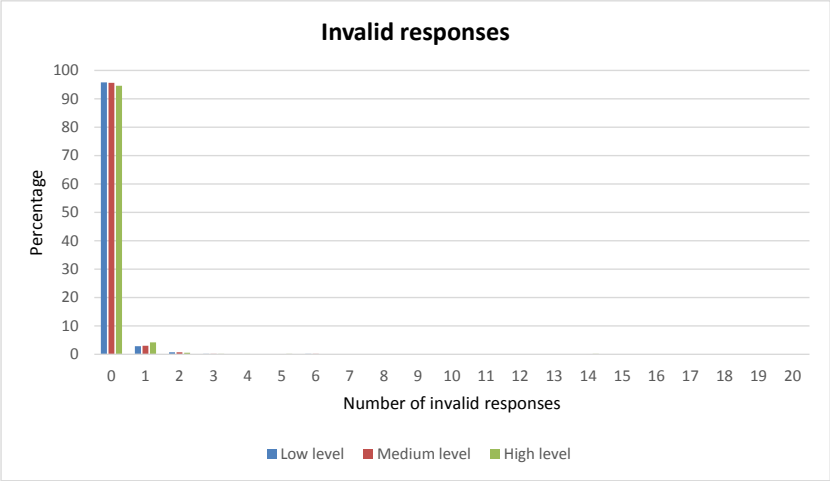

Figure 2. Number of invalid responses by test difficulty.

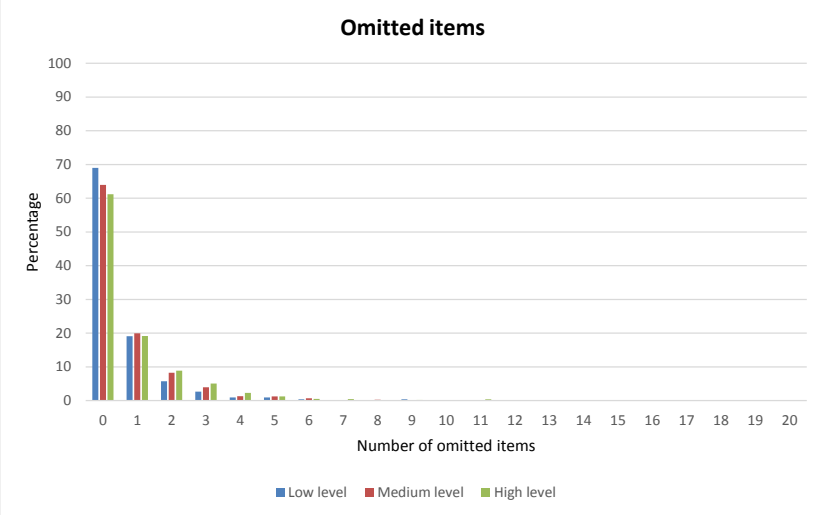

Figure 3. Number of omitted items by test difficulty.

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 85% and 94% of the respondents finished the entire test. About 3% to 5% of the participants did not reach the last five items.

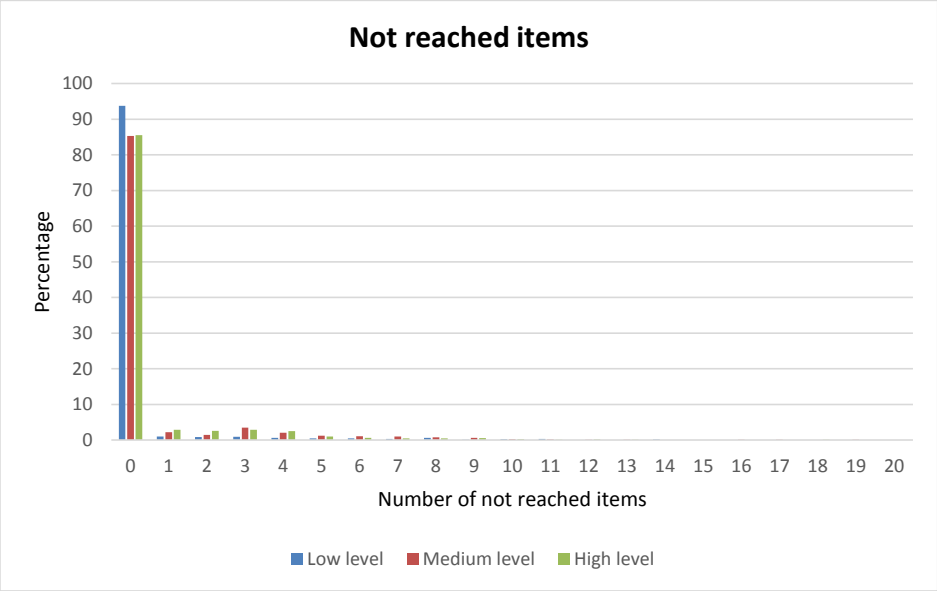

Figure 4. Number of not reached items by test difficulty.

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* = 1.01 (*SD*= 2.34; test with low level of difficulty) and *M* = 1.51 (*SD* = 2.58; test with high level of difficulty) missing responses in the different experimental conditions. About 52% to 63% of the respondents had no missing response at all and about 8% to 14% of the participants had four or more missing responses. Particularly, respondents receiving the test with medium or high level of difficulty showed more missing responses (13% and 14%) than respondents receiving the test with low level of difficulty.

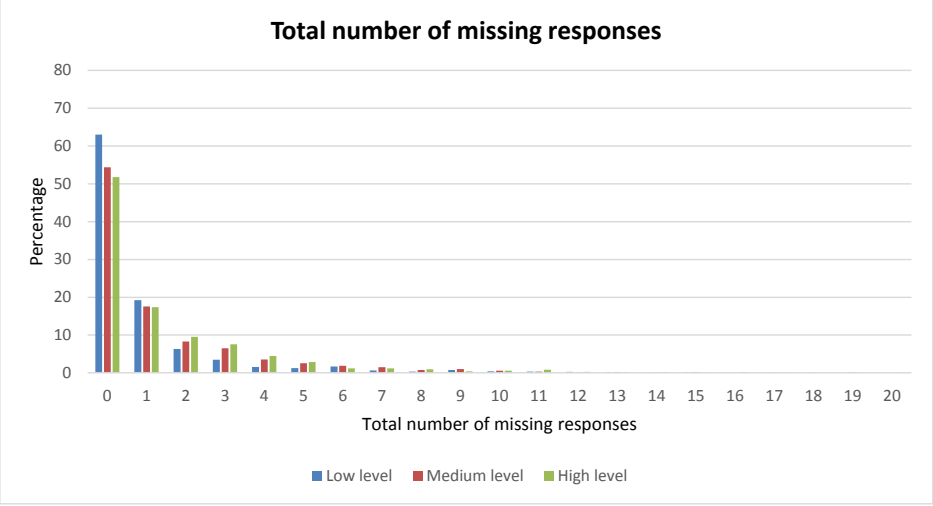

Figure 5. Total number of missing responses by test difficulty.

Overall, the amount of invalid and not-reached items is small, whereas a reasonable part of missing responses occurred due to omitted items.

5.1.2 Missing responses per item

Tables 5 to 7 provide information on the occurrence of different kinds of missing responses per item by assessment setting (at school and at home). Overall, in all of the three tests the omission rates were rather low, varying across items between 0.0 % and 8.5%. There was only one item with an omission rate exceeding 8% (icg9128_sc3g9_c in the test with medium level of difficulty). The omission rates correlated with the item difficulties at about .15 in the test with low level of difficulty, about .21 in the test with medium level of difficulty, and about .39 in the test with high level of difficulty. Generally, the percentage of invalid responses per item (column 6 Tables 5 to 7) was rather low with the maximum rate being 1.0%. With an item's progressing position in the test, the amount of persons that did not reach the item (column 4 in Tables 5 to 7) rose up to a reasonable amount of 8% to 15% for the different experimental conditions. Particularly, the last items of the tests with medium or high level of difficulty were not reached by all respondents (see Figure 6).

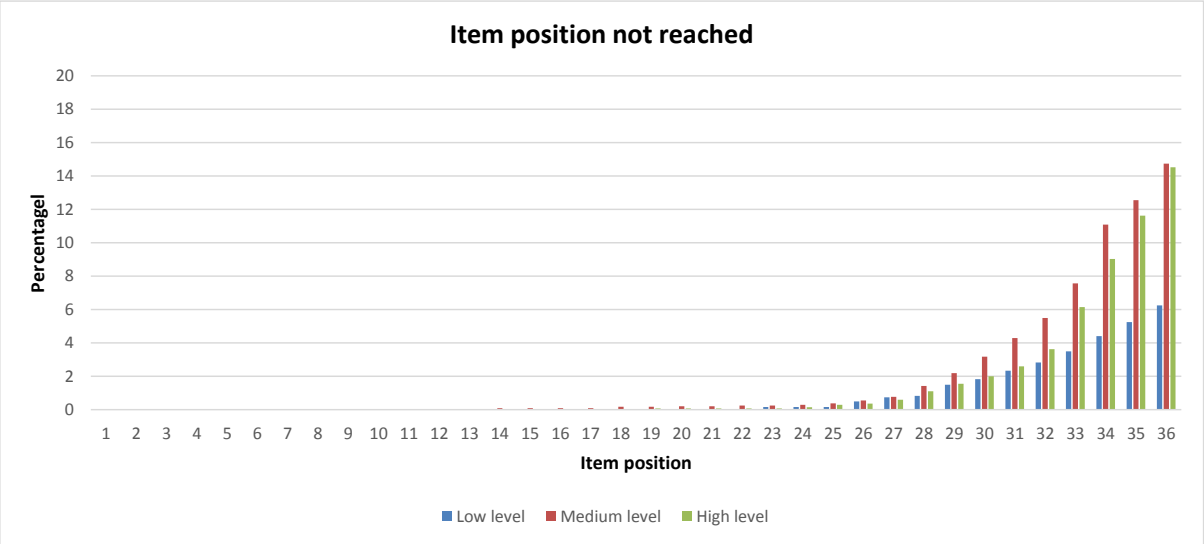

Figure 6. Item position not reached by test difficulty.

Table 5

Percentage of Missing Values by Test Difficulty (low level)

Item	Position	N	NR	OM	NV
icg5005x_sc3g9_c	1	1193	0.00	0.42	0.17
icg5034x_sc3g9_c	2	1191	0.00	0,42	0.33
icg5009x_sc3g9_c	3	1196	0.00	0,25	0.08
icg5051x_c	4	1178	0.00	1.42	0.42
icg5018x_sc3g9_c	5	1196	0.00	0.00	0.33
icg9106x_sc3g9_c	6	1190	0.00	0.50	0.33
icg5015x_sc3g9_c	7	1183	0.00	1.25	0.17
icg5046x_sc3g9_c	8	1157	0.00	3.08	0.50
icg5033x_sc3g9_c	9	1142	0.00	4.67	0.17
Icg9110x_sc3g9_c	10	1183	0.00	1.25	0.17
icg5045x_c	11	1167	0.00	2.58	0.17
icg5054x_sc3g9_c	12	1190	0.00	0.75	0.08
icg5021x_sc3g9_c	13	1192	0.00	0.58	0.08
Icg9114x_sc3g9_c	14	178	0.00	1.17	0.67
icg5059x_sc3g9_c	15	1185	0.00	1.08	0.17
icg9116x_sc3g9_c	16	1179	0.00	1.42	0.33
icg5035x_c	17	1193	0.00	0.33	0.25
icg9118x_sc3g9_c	18	1190	0.00	0.67	0.17
icg9119x_sc3g9_c	19	1142	0.00	4.58	0.25
icg5003x_sc3g9_c	20	1194	0.00	0.42	0.08
icg5029x_c	21	1173	0.00	2.00	0.25
icg9122x_sc3g9_c	22	1179	0.00	1.50	0.25
icg9123x_sc3g9_c	23	1184	0.17	1.00	0.17
icg12041x_sc3g9_c	24	1145	0.17	4.00	0.42
icg12042x_c	25	1167	0.17	2.50	0.08
icg12060s_sc3g9_c	26	1109	0.50	7.08	0.00
icg12036x_c	27	1183	0.75	0.50	0.17
icg5039x_sc3g9_c	28	1177	0.83	1.00	0.08
icg12018s_sc3g9_c	29	1112	1.50	5.50	0.33
icg5053x_sc3g9_c	30	1153	1.83	1.50	0.58
icg9131x_sc3g9_c	31	1153	2.33	1.33	0.25
icg9132x_sc3g9_c	32	1132	2.83	2.50	0.33
icg5049x_sc3g9_c	33	1132	3.50	1.75	0.42
icg12022x_c	34	1138	4.42	0.42	0.33
icg9138x_sc3g9_c	35	1121	5.25	1.00	0.33
icg9140s_sc3g9_c	36	1107	6.25	1.42	0.08

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 by Test Difficulty (medium level)

Item	Position	N	NR	OM	NV
icg5005x_sc3g9_c	1	2312	0.00	0.47	0.17
icg9102s_sc3g9_c	2	2191	0.00	5.72	0.13
icg5009x_sc3g9_c	3	2319	0.00	0.30	0.04
icg5047x_sc3g9_c	4	2296	0.00	0.30	1.03
icg12034x_sc3g9_c	5	2315	0.00	0.43	0.09
icg9106x_sc3g9_c	6	2313	0.00	0.39	0.21
icg5015x_sc3g9_c	7	2293	0.00	1.33	0.13
icg5046x_sc3g9_c	8	2253	0.00	2.49	0.69
icg5033x_sc3g9_c	9	2240	0.00	3.39	0.34
icg9110x_sc3g9_c	10	2290	0.00	1.50	0.09
icg5045x_c	11	2292	0.00	1.38	0.13
icg5054x_sc3g9_c	12	2318	0.00	0.30	0.09
icg9113x_sc3g9_c	13	2309	0.00	0.73	0.04
icg12040x_sc3g9_c	14	2304	0.09	0.43	0.47
icg5059x_sc3g9_c	15	2302	0.09	0.77	0.21
icg12043x_c	16	2284	0.09	1.59	0.17
icg9117s_sc3g9_c	17	2264	0.09	2.36	0.26
icg9118x_sc3g9_c	18	2308	0.17	0.39	0.26
ica5021s_sc3g9_c	19	2152	0.17	6.83	0.09
icg5003x_sc3g9_c	20	2302	0.21	0.69	0.17
icg5029x_c	21	2268	0.21	2.11	0.21
icg9122x_sc3g9_c	22	2280	0.26	1.50	0.26
icg9123x_sc3g9_c	23	2294	0.26	0.86	0.30
icg12041x_sc3g9_c	24	2271	0.30	2.02	0.09
icg12042x_c	25	2290	0.39	0.90	0.30
icg12060s_sc3g9_c	26	2217	0.56	4.08	0.09
icg12036x_c	27	2287	0.77	0.73	0.21
icg9128x_sc3g9_c	28	2094	1.42	8.51	0.09
icg12018s_sc3g9_c	29	2200	2.19	3.09	0.17
icg5053x_sc3g9_c	30	2217	3.18	1.29	0.26
icg9131x_sc3g9_c	31	2205	4.30	0.89	0.09
icg9132x_sc3g9_c	32	2169	5.50	0.95	0.34
icg9133s_sc3g9_c	33	1983	7.56	6.83	0.39
icg9136s_sc3g9_c	34	1947	11.09	4.86	0.39
icg9138x_sc3g9_c	35	2026	12.55	0.13	0.26
icg12027x_c	36	1982	14.74	0.00	0.09

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 7

Percentage of Missing Values by Test Difficulty (high level)

Item	Position	N	NR	OM	NV
icg9101x_sc3g9_c	1	1343	0.00	0.44	0.07
icg9102s_sc3g9_c	2	1293	0.00	4.15	0.07
icg9103x_sc3g9_c	3	1317	0.00	2.22	0.22
icg5047x_sc3g9_c	4	1341	0.00	0.07	0.59
icg12034x_sc3g9_c	5	1349	0.00	0.00	0.07
icg9106x_sc3g9_c	6	1341	0.00	0.22	0.44
icg9107s_sc3g9_c	7	1301	0.00	3.41	0.22
icg12138s_sc3g9_c	8	1317	0.00	2.22	0.22
icg12016s_sc3g9_c	9	1321	0.00	1.93	0.22
icg9110x_sc3g9_c	10	1326	0.00	1.04	0.74
icg9111x_sc3g9_c	11	1261	0.00	6.44	0.15
icg12047s_sc3g9_c	12	1272	0.00	5.26	0.52
icg9113x_sc3g9_c	13	1342	0.00	0.52	0.07
icg12040x_sc3g9_c	14	1330	0.00	1.11	0.37
icg12046s_sc3g9_c	15	1276	0.00	5.26	0.22
icg12043x_c	16	1331	0.00	1.26	0.15
icg9117s_sc3g9_c	17	1310	0.00	2.81	0.15
icg9118x_sc3g9_c	18	1344	0.00	0.37	0.07
ica5021s_sc3g9_c	19	1313	0.07	2.52	0.15
icg5003x_sc3g9_c	20	1336	0.07	0.44	0.52
icg5029x_c	21	1316	0.07	2.37	0.07
icg9122x_sc3g9_c	22	1310	0.07	2.44	0.44
icg5025s_sc3g9_c	23	1295	0.07	3.70	0.30
icg12041x_sc3g9_c	24	1334	0.15	0.74	0.30
icg9125s_sc3g9_c	25	1330	0.30	1.04	0.15
icg12060s_sc3g9_c	26	1293	0.37	3.85	0.00
icg12036x_c	27	1326	0.59	0.67	0.52
icg9128x_sc3g9_c	28	1230	1.11	7.56	0.22
icg9129x_sc3g9_c	29	1289	1.56	2.44	0.52
icg5053x_sc3g9_c	30	1307	2.00	0.74	0.30
icg9131x_sc3g9_c	31	1295	2.59	1.26	0.22
icg9132x_sc3g9_c	32	1296	3.63	0.96	0.15
icg9133s_sc3g9_c	33	1153	6.15	7.93	0.52
icg9136s_sc3g9_c	34	1146	9.04	5.63	0.44
icg12050s_sc3g9_c	35	1134	11.63	4.07	0.30
icg12027x_c	36	1152	14.52	0.00	0.15

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.

5.2 Parameter Estimates

5.2.1 Item parameters

The second column in Table 8 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 24% and 78% with an average of 57% (*SD* = 13%) correct responses.

The estimated item difficulties (for dichotomous variables) and location parameters (for the polytomous variable) are given in Table 8. The step parameters for the polytomous variable are depicted in Table 9. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties (or location parameters for the polytomous variable) ranged from -2.26 (item icg9140s_sc3g9_c) to 1.44 (item icg12040x_sc3g9_c) with an average difficulty of -0.56. 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 ≤ 0.08).

Table 8

Item parameters

Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
icg5005x_sc3g9_c	49.39	-0.117	0.036	0.92	-7.5	0.47	1.28	.030
icg5034x_sc3g9_c	62.72	-1.020	0.063	1.04	1.9	0.30	0.62	.030
icg5009x_sc3g9_c	60.65	-0.637	0.037	1.09	6.8	0.25	0.43	.035
icg5051x_c	58.56	-0.826	0.063	0.96	-1.9	0.44	1.04	.042
icg5018x_sc3g9_c	72.16	-1.498	0.068	1.09	2.8	0.20	0.33	.032
Icg9106x_sc3g9_c	64.47	-0.672	0.032	0.90	-8.5	0.48	1.44	.028
icg5015x_sc3g9_c	68.07	-1.002	0.039	1.01	0.8	0.32	0.78	.016
icg5046x_sc3g9_c	71.00	-1.154	0.040	1.04	2.5	0.30	0.60	.025
icg5033x_sc3g9_c	60.26	-0.610	0.037	1.02	1.3	0.35	0.74	.027
icg9110x_sc3g9_c	44.20	0.273	0.031	1.01	1.4	0.32	0.73	.022
icg5045x_c	58.66	-0.543	0.037	1.02	1.4	0.35	0.76	.029
icg5054x_sc3g9_c	60.40	-0.624	0.037	1.07	5.2	0.28	0.53	.026
icg5021x_sc3g9_c	73.74	-1.590	0.069	1.02	0.5	0.32	0.76	.023
icg9114x_sc3g9_c	57.05	-0.752	0.062	1.02	1.0	0.34	0.76	.030
icg5059x_sc3g9_c	63.12	-0.753	0.037	1.03	2.2	0.33	0.69	.022
icg9116x_sc3g9_c	70.48	-1.406	0.067	0.97	-1.1	0.40	1.02	.023
icg5035x_c	77.70	-1.823	0.073	1.02	0.4	0.30	0.81	.033
icg9118x_sc3g9_c	59.60	-0.436	0.031	0.95	-4.3	0.42	1.03	.025
icg9119x_sc3g9_c	59.11	-0.847	0.064	1.03	1.2	0.34	0.75	.025
icg5003x_sc3g9_c	64.38	-0.668	0.032	0.93	-5.7	0.43	1.21	.028
icg5029x_c	66.45	-0.770	0.033	1.01	0.7	0.37	0.73	.022
icg9122x_sc3g9_c	50.91	-0.034	0.031	0.95	-4.8	0.42	1.06	.027

Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*T*	r_{it}	Dicr.	Q_3
icg9123x_sc3g9_c	65.84	-0.833	0.038	0.96	-2.6	0.42	1.04	.029
icg12041x_sc3g9_c	55.89	-0.256	0.031	1.08	7.2	0.27	0.50	.024
icg12042x_c	68.90	-1.044	0.039	1.01	0.8	0.34	0.75	.022
icg12060s_sc3g9_c	n.a.	-1.520	0.043	0.96	-2.1	0.36	1.05	.022
icg12036x_c	55.23	-0.232	0.031	1.10	9.4	0.25	0.41	.028
icg5039x_sc3g9_c	72.13	-1.505	0.068	0.94	-1.9	0.44	1.28	.032
icg12018s_sc3g9_c	n.a.	-1.494	0.038	0.87	-5.8	0.54	2.34	.035
icg5053x_sc3g9_c	66.22	-0.762	0.033	0.91	-6.7	0.46	1.33	.033
icg9131x_sc3g9_c	62.93	-0.597	0.032	0.91	-7.6	0.46	1.32	.031
icg9132x_sc3g9_c	65.77	-0.733	0.033	0.97	-2.0	0.38	0.95	.022
icg5049x_sc3g9_c	63.87	-1.087	0.065	1.03	1.2	0.31	0.67	.027
icg12022x_c	70.83	-1.441	0.069	1.05	1.5	0.26	0.54	.030
icg9138x_sc3g9_c	63.23	-0.773	0.039	1.11	6.9	0.24	0.38	.028
icg9140s_sc3g9_c	n.a.	-2.259	0.071	0.93	-1.4	0.44	1.93	.036
icg9102s_sc3g9_c	n.a.	-0.603	0.042	0.98	-1.3	0.35	0.94	.023
icg5047x_sc3g9_c	40.20	0.602	0.036	1.03	2.2	0.30	0.67	.032
icg12034x_sc3g9_c	55.87	-0.116	0.035	1.07	5.6	0.27	0.49	.025
icg9113x_sc3g9_c	35.31	0.828	0.037	1.03	2.3	0.29	0.60	.026
icg12040x_sc3g9_c	23.91	1.443	0.041	1.01	0.6	0.27	0.69	.020
icg12043x_c	50.98	0.109	0.035	1.03	3.1	0.32	0.63	.027
icg9117s_sc3g9_c	n.a.	-1.465	0.036	0.91	-3.3	0.48	2.71	.033
ica54021s_sc3g9_c	n.a.	-1.846	0.044	0.91	-3.3	0.44	2.22	.029
icg9128x_sc3g9_c	43.02	0.477	0.037	1.07	5.5	0.28	0.49	.025
icg9133s_sc3g9_c	n.a.	-0.833	0.028	1.02	0.7	0.49	1.53	.031
icg9136s_sc3g9_c	n.a.	-0.221	0.027	0.94	-2.8	0.55	2.57	.030
icg12027x_c	36.53	0.769	0.039	1.08	5.8	0.21	0.39	.026
icg9101x_sc3g9_c	40.66	0.808	0.058	1.11	5.8	0.16	0.18	.027
icg9103x_sc3g9_c	33.26	1.156	0.061	0.99	-0.5	0.34	0.81	.025
icg9107s_sc3g9_c	n.a.	-1.226	0.067	0.97	-0.7	0.36	1.48	.026
icg12138s_sc3g9_c	n.a.	-0.720	0.062	1.08	2.7	0.18	0.46	.029
icg12016s_sc3g9_c	n.a.	-0.997	0.062	1.05	1.8	0.23	0.64	.026
icg9111x_sc3g9_c	50.28	0.382	0.059	0.99	-0.6	0.36	0.80	.022
icg12047s_sc3g9_c	n.a.	0.667	0.041	1.02	0.5	0.46	1.70	.030
icg12046s_sc3g9_c	n.a.	0.311	0.040	1.12	3.6	0.36	1.02	.037
ica54052s_sc3g9_c	n.a.	-0.528	0.052	1.00	-0.1	0.41	1.65	.044
icg9125s_sc3g9_c	n.a.	-1.200	0.081	0.99	-0.3	0.27	0.83	.022
icg9129x_sc3g9_c	28.78	1.393	0.064	1.01	0.2	0.31	0.72	.028
icg12050s_sc3g9_c	n.a.	-0.467	0.064	0.98	-0.5	0.38	1.43	.031

Note. 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, 1983). Percent correct scores are not informative for polytomous CMC item scores. These are denoted by n.a.

For the dichotomous items, the item-total correlation corresponds to the point-biserial correlation between the correct response and the total score; for polytomous items it corresponds to the product-moment correlation between the corresponding categories and the total score (discrimination value as computed in ConQuest).

Table 9

Step parameters (with Standard Errors) for the Polytomous Items

Item	Step 1	Step 2	Step 3	Step 4	Step 5
icg12060s_sc3g9_c	-0.252	0.252			
	(0.032)				
icg12018s_sc3g9_c	0.079	-0.154	0.075		
	(0.036)	(0.039)			
icg9140s_sc3g9_c	0.275	-0.080	-0.195		
	(0.066)	(0.074)			
icg9102s_sc3g9_c	0.469	-0.469			
	(0.041)				
icg9117s_sc3g9_c	0.008	-0.118	-0.541		
	(0.034)	(0.034)	(0.035)		
ica5021s_sc3g9_c	-0.334	-0.359	0.693		
	(0.035)	(0.036)			
icg9133s_sc3g9_c	-0.235	0.948	0.047	-0.760	
	(0.038)	(0.043)	(0.052)		
icg9136s_sc3g9_c	-0.356	-0.362	-0.020	-0.097	0.835
	(0.042)	(0.038)	(0.037)	(0.041)	
icg9107s_sc3g9_c	-0.835	0.052	0.783		
	(0.057)	(0.057)			
icg12138s_sc3g9_c	-1.407	0.809	0.598		
	(0.059)	(0.061)			
icg12016s_sc3g9_c	-1.191	0.706	0.485		
	(0.057)	(0.061)			
icg12047s_sc3g9_c	-0.447	-0.519	-0.068	0.063	0.971
	(0.069)	(0.059)	(0.059)	(0.073)	
icg12046s_sc3g9_c	-0.275	-0.556	0.142	0.086	0.604
	(0.065)	(0.058)	(0.059)	(0.071)	
ica5052s_sc3g9_c	-0.811	-0.406	0.507	0.709	
	(0.062)	(0.056)	(0.062)		
icg9125s_sc3g9_c	-0.293	0.293			
	(0.059)				
icg12050s_sc3g9_c	-0.210	-0.339	0.549		
	(0.060)	(0.062)			

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	
	X	
	XX	
	X	41 59
	XX	
	XX	
	XXXX	50
1	XXXX	
	XXXXX	
	XXXXXXX	40 48 49
	XXXXXX	55
	XXXXXXX	38
	XXXXXXX	45
	XXXXXXXX	54
	XXXXXXXXX	10 56
	XXXXXXXXXX	42
	XXXXXXXX	
0	XXXXXXXX	22
	XXXXXXXX	1 39
	XXXXXXXX	24 27 47
	XXXXXXXXX	
	XXXXXXXXX	18 57 60
	XXXXXXX	3 9 11 12 31 37
	XXXXXXX	6 20 32 52
	XXXX	4 14 15 19 21 30 35 46
	XXXXX	23
-1	XXXXX	2 7 25 53
	XXX	8 33
	XX	51 58
	XXX	
	XX	16 29 34 43
	X	5 13 26 28
	X	
	X	
	X	17 44
-2		
		36

Figure 7. Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 28.5 cases. Item difficulty is depicted on the right side of the graph. Each number represents one item (see Table 6).

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 0.64, indicating somewhat limited variability between subjects. The reliability of the test (EAP/PV reliability = .82; WLE reliability = .81) was adequate. 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 128 items. The probability of a correct response ranged from 24% to 98% across all items (*Mdn* = 72%). Thus, the number of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.87 to 1.12, the respective *t*-value from -9.1 to 10.3, 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 four items with a point-biserial-correlation between .00 and .06 (Median = -.18). 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 8). Values of the WMNSQ ranged from 0.87 (item icg12018s_sc3g9_c) to 1.12 (icg12046s_sc3g9_c). Only three items exhibited a *t*-value of the WMNSQ greater than 6 and one exceeded a value of 8. Thus, there was no indication of severe item over- or underfit. Point-biserial correlations between the item scores and the total scores ranged from .16 (item icg9101x_sc3g9_c) to .55 (item icg9136s_sc3g9_c) and had a mean of .35. 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). In addition, the effect of the experimental factor test difficulty (booklet) was also studied. 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, Rohm, Gnambs, and Carstensen (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).

The differences between the estimated item difficulties in the various groups are summarized in Table 10. 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 12).

Table 10

Differential Item Functioning

Item	Sex	Books	Migration	Migration	Migration	School	Position
	Male vs. female	< 100 vs. ≥ 100	Without vs. with	Without vs. missing	With vs. missing	no sec. vs sec.	First vs. second
icg5005x_sc3g9_c	0.560	0.190	-0.050	-0.313	-0.263	0.388	0.070
icg5034x_sc3g9_c	-0.202	-0.266	-0.085	0.190	0.275	-0.472	-0.040
icg5009x_sc3g9_c	-0.252	-0.128	0.234	0.177	-0.057	-0.292	0.068
icg5051x_c	-0.148	0.018	-0.218	-0.142	0.076	0.072	-0.298
icg5018x_sc3g9_c	-0.322	-0.538	0.127	0.386	0.259	-0.532	-0.088
icg9106x_sc3g9_c	-0.042	0.246	0.006	-0.159	-0.165	0.508	0.016
icg5015x_sc3g9_c	-0.072	-0.022	0.189	-0.090	-0.279	-0.092	-0.132
icg5046x_sc3g9_c	-0.082	-0.050	-0.122	0.023	0.145	-0.234	-0.038
icg5033x_sc3g9_c	-0.120	0.036	0.031	0.087	0.051	0.110	-0.004
icg9110x_sc3g9_c	-0.036	-0.010	-0.098	0.017	0.115	-0.070	-0.038
icg5045x_c	0.998	-0.364	0.191	0.445	0.254	-0.348	0.120
icg5054x_sc3g9_c	-0.098	-0.046	0.139	0.128	-0.011	-0.122	0.156
icg5021x_sc3g9_c	-0.130	-0.138	-0.270	0.126	0.396	0.172	0.098
icg9114x_sc3g9_c	-0.130	0.048	-0.462	-0.426	0.036	0.256	-0.114
icg5059x_sc3g9_c	0.082	-0.120	0.146	0.142	-0.004	-0.228	0.090
icg9116x_sc3g9_c	-0.120	0.416	-0.162	-0.075	0.087	0.578	-0.110
icg5035x_c	-0.214	0.038	-0.168	-0.176	-0.008	-0.282	0.056

Item	Sex	Books	Migration	Migration	Migration	School	Position
	Male vs. female	< 100 vs. ≥ 100	Without vs. with	Without vs. missing	With vs. missing	no sec. vs sec.	First vs. second
icg9118x_sc3g9_c	0.070	0.220	-0.145	-0.062	0.083	0.170	0.050
icg9119x_sc3g9_c	0.222	-0.118	0.104	-0.032	-0.136	-0.380	-0.074
icg5003x_sc3g9_c	0.466	-0.038	-0.094	-0.110	-0.016	0.052	0.082
icg5029x_c	-0.052	-0.160	-0.046	0.043	0.089	-0.166	-0.154
icg9122x_sc3g9_c	-0.094	0.174	0.034	-0.079	-0.113	0.086	-0.170
icg9123x_sc3g9_c	-0.350	0.216	-0.324	0.060	0.384	0.248	-0.172
icg12041x_sc3g9_c	0.024	-0.224	-0.023	0.302	0.325	-0.250	-0.020
icg12042x_c	-0.516	0.260	0.125	0.049	-0.076	0.392	-0.102
icg12060s_sc3g9_c	-0.344	0.156	0.105	-0.116	-0.221	0.200	-0.162
icg12036x_c	-0.142	-0.104	0.025	0.134	0.109	-0.150	0.062
icg5039x_sc3g9_c	0.012	-0.008	-0.058	-0.011	0.047	0.318	0.140
icg12018s_sc3g9_c	-0.348	0.224	-0.239	-0.340	-0.101	0.448	-0.156
icg5053x_sc3g9_c	0.062	0.446	-0.178	-0.401	-0.223	0.802	0.150
icg9131x_sc3g9_c	-0.038	0.302	0.125	-0.095	-0.220	0.282	0.054
icg9132x_sc3g9_c	-0.102	0.166	-0.231	-0.021	0.210	0.144	-0.050
icg5049x_sc3g9_c	0.024	-0.148	0.012	-0.129	-0.141	-0.188	-0.018
icg12022x_c	-0.052	-0.264	-0.121	-0.083	0.038	-0.660	0.208
icg9138x_sc3g9_c	-0.184	-0.390	0.262	0.199	-0.063	-0.536	0.232
icg9140x_sc3g9_c	0.030	-0.062	-0.478	-0.461	0.017	0.396	0.066
icg9102s_sc3g9_c	0.160	0.052	-0.107	-0.142	-0.035	-0.110	0.120
icg5047x_sc3g9_c	0.510	0.090	-0.017	-0.025	-0.008	-0.056	0.186
icg12034x_sc3g9_c	-0.300	-0.078	-0.103	-0.170	-0.067	-0.050	-0.010
icg9113x_sc3g9_c	-0.200	-0.006	-0.093	-0.270	-0.177	-0.092	-0.170
icg12040x_sc3g9_c	0.252	-0.220	0.049	0.044	-0.005	-0.248	0.044
icg12043x_c	-0.050	-0.018	-0.171	-0.075	0.096	0.048	0.014
icg9117s_sc3g9_c	0.176	0.160	0.095	-0.023	-0.118	0.222	0.006
ica54021s_sc3g9_c	0.274	0.288	-0.048	-0.181	-0.133	0.144	0.102
icg9128x_sc3g9_c	-0.290	-0.266	0.235	0.431	0.196	-0.296	0.014
icg9133s_sc3g9_c	-0.172	0.066	-0.005	0.041	0.046	0.004	-0.086
icg9136s_sc3g9_c	0.002	-0.050	0.112	-0.013	-0.125	0.098	-0.092
icg12027x_c	0.382	-0.330	0.216	0.235	0.019	-0.470	0.084
icg9101x_sc3g9_c	-0.362	-0.276	0.084	0.182	0.098	-0.302	0.338
icg9103x_sc3g9_c	-0.066	-0.282	0.113	0.241	0.128	-0.114	-0.136
icg9107s_sc3g9_c	0.482	0.042	0.011	-0.365	-0.376	0.086	0.038
icg12138s_sc3g9_c	0.266	-0.460	0.245	0.192	-0.053	-0.270	0.120
icg12016s_sc3g9_c	-0.230	-0.094	0.222	0.105	-0.117	-0.192	0.428
icg9111x_sc3g9_c	-0.010	-0.202	0.001	-0.298	-0.299	-0.166	-0.010
icg12047s_sc3g9_c	-0.320	0.000	-0.002	0.044	0.046	-0.310	0.100
icg12046s_sc3g9_c	0.042	-0.258	0.018	0.015	-0.003	-0.190	-0.024
ica5052s_sc3g9_c	0.514	-0.062	-0.072	0.293	0.365	-0.146	-0.040
icg9125s_sc3g9_c	-0.140	0.192	0.240	-0.423	-0.663	0.110	0.370

Item	Sex	Books	Migration	Migration	Migration	School	Position
	Male vs. female	< 100 vs. ≥ 100	Without vs. with	Without vs. missing	With vs. missing	no sec. vs sec.	First vs. second
icg9129x_sc3g9_c	-0.062	0.374	-0.385	0.112	0.497	0.164	-0.312
icg12050s_sc3g9_c	0.350	-0.182	0.422	-0.131	-0.553	-0.052	-0.064
Main effect	-0.042	-0.568	0.287	0.508	0.221	-0.806	0.092

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

Sex: The sample included 2,457 (50%) males and 2,420 (50%) females. On average, male participants had a higher estimated computer literacy than females (main effect = 0.042 logits, Cohen's *d* = 0.065). Only one item (item icg5045x_c) showed DIF greater than 0.6 logits. An overall test for DIF (see Table 12) 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 socioeconomic status. There were 1,813 (37%) test takers with 0 to 100 books at home and 3,026 (62%) test takers with more than 100 books at home. 38 (1%) 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 -0.568 logits (Cohen's *d* =-0.882) lower in computer literacy than participants with more than 100 books. However, there was no considerable DIF on the item level. 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 12).

Migration background: There were 3,412 participants (70%) with no migration background, 878 subjects (18%) with a migration background, and 587 individuals (12%) that did not indicate their migration background. Participants without migration background had on average a higher computer literacy than subjects with migration background (main effect = 0.287 logits, Cohen's *d* = 0.446) and than participants that did not indicate their migration background (main effect = 0.508 logits, Cohen's *d* = 0.789). Participants with a migration background had on average a higher computer literacy then individuals that did not indicate their migration background (main effect = 0.221 logits, Cohen's *d* = 0.343). There was no considerable DIF on the item level with the exception of one item that showed greater DIF than 0.6 logits (between participants with migration background and participtants that did not indicate their migration background: item icg9125s_sc3g9_c). Model comparisons using AIC and BIC both favored the main effects model. Thus, overall, there was no pronounced DIF with regard to migration background.

School type: Overall, 2,315 subjects (47%) who took the computer literacy test attended secondary school (German: "Gymnasium"), whereas 2,562 (53%) were enrolled in other school types. Subjects in secondary schools showed a higher computer literacy on average (0.806 logits; Cohen's *d* = 1.252) than subjects in other school types. There was no

considerable DIF on the item level with the exception of two items that showed greater DIF than 0.6 logits (items icg12022x_c and ic5053x_sc3g9_c). However, the overall model test using AIC and BIC indicated a slightly better fit for the more complex DIF model, because several items showed DIF effects between 0.4 and 0.6 logits; thus, these differences were not considered severe.

Position: The computer literacy test was administered in two different positions (see section 3.1 for the design of the study). A subsample of 2,451 (50%) persons received the computer literacy first and 2,420 (50%) respondents took the computer literacy test after having completed the science 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 0.092 logits (Cohen's *d* = 0.143) 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 0.428 logits (item icg12016s_sc3g9_c). As a consequence, the overall test for DIF using the BIC favored the more parsimonious main effect model (Table 12).

Test version (booklet): The computer literacy test was administered in three different test versions (see section 3.1 for the design of the study). A subsample of 3,872 (67%) persons received the computer literacy test with low level of difficulty, a subsample of 3,872 (67%) persons received the test with medium level of difficulty whereas 1,890 (33%) 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	$\Delta\sigma$	$SE_{\Delta\sigma}$	***F***	$\Delta\sigma$	$SE_{\Delta\sigma}$	***F***
icg5005x_sc3g9_c	0.36	0.08	21.06	0.36	0.08	17.61
icg5009x_sc3g9_c	-0.32	0.08	17.40			
icg9106x_sc3g9_c	0.13	0.08	2.68			
icg5015x_sc3g9_c	0.30	0.08	13.81			
icg5046x_sc3g9_c	-0.21	0.08	6.61			
icg5033x_sc3g9_c	-0.01	0.08	0.01			
icg9110x_sc3g9_c	0.10	0.08	1.41	0.18	.07	5.68
icg5045x_c	0.08	0.08	1.02			

	Tests with low and medium level of diffiuclty			Tests with medium and high level of diffiuclty		
Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	*F*	$\Delta\sigma$	$SE_{\Delta\sigma}$	*F*
icg5054x_sc3g9_c	-0.19	0.08	6.32			
icg5059x_sc3g9_c	-0.06	0.08	0.56			
icg9118x_sc3g9_c	0.13	0.08	2.90	-0.07	0.08	0.81
icg5003x_sc3g9_c	0.36	0.08	22.34	0.15	0.08	3.17
icg5029x_c	-0.03	0.08	0.16	-0.67	0.08	72.15
icg9122x_sc3g9_c	-0.04	0.08	0.32	0.08	0.08	1.00
icg9123x_sc3g9_c	0.13	0.08	2.79			
icg12041x_sc3g9_c	-0.08	0.08	0.96	-0.43	0.07	33.11
icg12042x_c	-0.01	0.08	0.01			
icg12060s_sc3g9_c	-0,21	0,10	4,08	0.01	0.11	0.01
icg12036x_c	-0,62	0,08	66,68	-0.15	0.07	4.25
icg12018s_sc3g9_c	0,07	0,08	0,80			
icg5053x_sc3g9_c	0,22	0,08	7,59	0.30	0.09	11.19
icg9131x_sc3g9_c	0,47	0,08	36,21	0.02	0.08	0.05
icg9132x_sc3g9_c	0,07	0,08	0,72	0.17	0.09	3.71
icg9138x_sc3g9_c	-0,61	0,08	56,34			
icg9102s_sc3g9_c				0.15	0.09	3.01
icg5047x_sc3g9_c				0.24	0.07	10.12
icg12034x_sc3g9_c				-0.12	0.07	2.57
icg9113x_sc3g9_c				0.00	0.07	0.00
icg12040x_sc3g9_c				0.11	0.08	1.63
icg12043x_c				-0.22	0.07	9.23
icg9117s_sc3g9_c				0.17	0.08	4.41
ica5021s_sc3g9_c				0.53	0.10	29.64
icg9128x_sc3g9_c				-0.60	0.08	61.77
icg9133s_sc3g9_c				-0.19	0.06	10.57
icg9136s_sc3g9_c				0.04	0.06	0.58
icg12027x_c				-0.12	0.08	2.31

Note. $\Delta\sigma$ = Difference in item difficulty parameters; $SE_{\Delta\sigma}$ = 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, 3,527) = 82.49 for the tests with low and medium level of difficulty and F_{0154} (1, 3,677) = 85.35 for the test with medium and high level of difficulty. A non-significant test indicates measurement invariance.

Table 12

Differential Item Functioning

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	main effect	4,877	251,424.53	100	251,624.53	252,273.76
	DIF		250,749.46	160	251,069.46	252,108.23
Books	main effect	4,839	249,215.27	100	249,415.27	250,063.72
	DIF		248,879.81	160	249,199.81	250,237.32
Migration	main effect	4,877	251,208.63	100	251,410.63	252,066.35
	DIF		250,969.62	221	251,411.62	252,846.41
School type	main effect	4,877	250,302.11	100	250,502.11	251,151.34
	DIF		249,673.48	160	249,993.48	251,032.24
Position	main effect	4,877	251,413.84	100	251,613.84	252,263.07
	DIF		251,284.37	160	251,604.37	252,418.78

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 6), ranging from 0.18 (item icg9101x_sc3g9_c) to 2.71 (item icg9117s_sc3g9_c). The average discrimination parameter fell at 0.91. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 249,794.55, BIC =250,820.33) as compared to the PCM model (AIC = 251,625.23, BIC =252,267.97). 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 13. The correlations between the dimensions varied between .87 and .95. The smallest correlation was found between Dimension 2 ("Create") and Dimension 4 ("Evaluate"). Dimension 3 ("Manage") and Dimension 4 ("Evaluate") showed the strongest correlation. All correlations deviated from a perfect correlation (i.e., they were marginally lower than r = .95, see Carstensen, 2013).

Moreover, according to model fit indices, the four-dimensional model fitted the data slightly better (AIC = 251,297.25, BIC = 251,998.42, number of parameters = 108) than the unidimensional model (AIC = 251,625.23, BIC = 252,267.97, number of parameters = 99). These results indicate that the three cognitive requirements measure a common construct, albeit it is not completely unidimensional.

Table 13

Results of Four-Dimensional Scaling (Process Components)

	Access	**Create**	**Manage**	**Evaluate**
Access (19 Items)	(0.518)			
Create (13 Items)	.906	(0.507)		
Manage (14 Items)	.947	.897	(0.860)	
Evaluate (14 Items)	.927	.871	.953	(1.014)

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 14. The correlations among the three dimensions were rather high and fell between .91 and .95. The smallest correlation was found between Dimension 1 (“Word processing”) and Dimension 2 (“Spreadsheet / presentation software”). Dimension 2 (“Spreadsheet / presentation software”) and Dimension 3 (“E-mail / communication tools”) showed the strongest correlation. However, they deviated from a perfect correlation (i.e., they were marginally lower than r = .95, see Carstensen, 2013). Moreover, according to model fit indices, the four-dimensional model fitted the data slightly better (AIC = 251,182.11, BIC = 251,883.27, number of parameters = 108) than the unidimensional model (AIC = 251,625.23, BIC = 252,267.97, number of parameters = 99). 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 = .005) — 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 14

Results of Four-Dimensional Scaling (Software Applications).

	Dim 1	**Dim 2**	**Dim 3**	**Dim 4**
Word processing (Dim1) (17 Items)	(0.480)			
Spreadsheet / presentation software (Dim 2) (17 Items)	.909	(0.526)		
E-mail / communication tools (Dim 3) (8 Items)	.911	.947	(1.035)	
Internet / search engines (Dim 4) (18 Items)	.938	.941	.942	(0.999)

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 9 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 60 items, of which 44 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. A total of 16 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 6 (see Senkbeil, Ihme, & Adrian, 2014) 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 9 that were not part of starting cohort 3) were administered all items from the grade 6 and 24 items of the grade 9 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 3,167 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 = 510 students from grade 9 received both tests within a single measurement occasion.

7.2.2 The design of the link study

The test administered in grade 6 included 30 items (see Senkbeil et al., 2014), whereas the test administered in grade 9 included 60 items, distributed among three test versions (see above). A subtest with 24 items of the grade 9 test was administered to the participants. Moreover, the computer literacy test was administered at different positions in the test battery. A random sample of 259 students received the computer literacy test before working on a science test, whereas the remaining 251 students received the science 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 = 37,141.34, number of parameters = 83; two-dimensional model: BIC = 37,142.84, number of parameters = 85), whereas the AIC favored the two-dimensional model (AIC = 36,782.91; unidimensional model: AIC = 36,789.88). 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 6 and 9 were essentially unidimensional.

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal subsample from the starting cohort. The differences in item difficulties between the link sample and starting cohort 3 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 15.

Table 15

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

		Grade 6				Grade 9		
No.	**Item**	**$\Delta\sigma$**	***SE*$_{\Delta\sigma}$**	***F***	**Item**	**$\Delta\sigma$**	***SE*$_{\Delta\sigma}$**	***F***
1	icg6001x_c	0.19	0.11	3.27	icg9101x_sc3g9_c	0.30	0.11	7.02
2	icg6003x_c	0.20	0.11	3.51	icg9102s_sc3g9_c	-0.42	0.13	11.22
3	icg6005x_c	0.81	0.11	52.02	icg9103x_sc3g9_c	0.77	0.12	43.85
4	icg6006x_c	0.33	0.18	3.30	icg9106x_sc3g9_c	0.12	0.11	1.14
5	icg6009x_c	-0.58	0.11	37.38	icg9107s_sc3g9_c	-0.19	0.12	2.46
6	icg6011x_c	0.50	0.13	15.19	icg9110x_sc3g9_c	0.03	0.11	0.08
7	icg6012x_c	0.12	0.19	0.41	icg9111x_sc3g9_c	0.10	0.12	0.71
8	icg6013x_c	-0.39	0.12	11.24	icg9113x_sc3g9_c	0.20	0.11	3.41
9	icg6014x_c	1.02	0.24	17.63	icg9114x_sc3g9_c	0.26	0.12	4.55
10	icg6015x_c	-0.02	0.11	0.03	icg9116x_sc3g9_c	0.24	0.14	3.06
11	icg6020x_c	0.12	0.23	0.27	icg9117s_sc3g9_c	-0.32	0.10	9.35
12	icg6016x_c	0.23	0.14	2.65	icg9118x_sc3g9_c	-0.05	0.11	0.26
13	icg6018x_c	-0.34	0.13	7.12	icg9119x_sc3g9_c	0.20	0.12	2.55
14	icg6021x_c	0.39	0.13	9.65	icg9122x_sc3g9_c	0.10	0.11	0.89
15	icg6024x_c	0.16	0.17	0.86	icg9123x_sc3g9_c	0.25	0.12	4.38
16	icg6025x_c	0.16	0.12	1.68	icg9125s_sc3g9_c	-0.42	0.15	8.06
17	icg6031x_c	0.29	0.11	6.24	icg9128x_sc3g9_c	0.53	0.12	19.99
18	icg6032x_c	0.01	0.14	0.00	icg9129x_sc3g9_c	-0.11	0.13	0.65
19	icg6033x_c	-0.43	0.11	15.17	icg9131x_sc3g9_c	-0.03	0.11	0.06
20	icg6034x_c	0.17	0.13	1.56	icg9132x_sc3g9_c	0.01	0.11	0.00

21	icg6036x_c	-0.52	0.13	15.08	icg9133s_sc3g9_c	0.07	0.09	0.67
22	icg6039x_c	0.21	0.15	1.89	icg9136s_sc3g9_c	0.08	0.09	0.88
23	icg6042x_c	-0.19	0.12	2.41	icg9138x_sc3g9_c	-0.03	0.13	0.07
24	icg6047x_c	-0.41	0.10	14.97	icg9140s_sc3g9_c	-0.31	0.15	4.49
25	icg6048x_c	-0.42	0.13	11.35				
26	icg6048x_c	-0.19	0.12	2.63				
27	icg6046x_c	-0.84	0.11	53.87				
28	icg6053x_c	-0.09	0.11	0.75				
29	icg6054x_c	-0.46	0.11	17.70				
30	icg6059x_c	-0.09	0.12	0.55				

Note. $\Delta\sigma$ = Difference in item difficulty parameters between the longitudinal subsample in grade 6 or 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, 3,677) = 85.35. A non-significant test indicates measurement invariance.

Analyses of differential item functioning between the link sample and starting cohort 3 identified neither for grade 6 (difference in logits: *Min* = 0.01, *Max* = 1.02) nor for grade 9 (difference in logits: *Min* = 0.01, *Max* = 0.77) items with significant (α = .05) DIF. Therefore, the computer literacy tests administered in the two grades were linked using the "mean/mean" method for the anchor-group design (see Fischer et al., 2016). The correction term was calculated as $c = 1.042$. This correction term was subsequently added to each difficulty parameter estimated in grade 9 (see Table 8) to derive the linked item parameters.

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, "icg9_sc1" and "icg9_sc1u", including their respective standard errors "icg9_sc2" and "icg9_sc2u". Both WLE scores are linked to the underlying reference scale of Grade 6. The uncorrected score "icg9_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 corrected score "icg9_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. 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. A description of these approaches can be found in Pohl and Carstensen (2012).

References

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

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. (2016). *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., & 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.

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., & 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 9)

```
title SC4 G12 Computer Literacy partial credit model;

/* load data */
datafile  >>filename.dat;
format pid 1-7 responses 9-68;
labels <<filename_with_labels.txt;

/* collapse response categories */
codes 0,1,2,3,4,5,6,7;
recode (0,1,2,3,4)          (0,0,0,1,2)          !item(26);     /* icg12060s_sc3g9_c */
recode (0,1,2,3,4)          (0,0,0,1,2)          !item(37);     /* icg9102s_sc3g9_c */
recode (0,1,2,3,4)          (0,0,0,1,2)          !item(58);     /* icg9125s_sc3g9_c */
recode (0,1,2,3,4)          (0,0,1,2,3)          !item(29);     /* icg12018s_sc3g9_c */
recode (0,1,2,3,4)          (0,0,1,2,3)          !item(36);     /* icg9140s_sc3g9_c */
recode (0,1,2,3,4)          (0,0,1,2,3)          !item(52);     /* icg12138s_sc3g9_c */
recode (0,1,2,3,4)          (0,0,1,2,3)          !item(53);     /* icg12016s_sc3g9_c */
recode (0,1,2,3,4,5)        (0,0,0,1,2,3)        !item(44);     /* ica5021s_sc3g9_c */
recode (0,1,2,3,4,5)        (0,0,0,1,2,3)        !item(51);     /* icg9107s_sc3g9_c */
recode (0,1,2,3,4,5)        (0,0,1,2,3,4)        !item(57);     /* ica5052s_sc3g9_c */
recode (0,1,2,3,4,5,6)      (0,0,0,0,1,2,3)      !item(60);     /* icg12050s_sc3g9_c */
recode (0,1,2,3,4,5,6)      (0,0,0,1,2,3,4)      !item(43);     /* icg9117s_sc3g9_c */
recode (0,1,2,3,4,5,6)      (0,0,0,1,2,3,4)      !item(46);     /* icg9133s_sc3g9_c */
recode (0,1,2,3,4,5,6)      (0,0,1,2,3,4,5)      !item(55);     /* icg12047s_sc3g9_c */
recode (0,1,2,3,4,5,6)      (0,0,1,2,3,4,5)      !item(56);     /* icg12046s_sc3g9_c */
recode (0,1,2,3,4,5,6,7)    (0,0,0,1,2,3,4,5)    !item(47);     /* icg9136s_sc3g9_c */


/* scoring */
score (0,1)          (0,1)            !item(1-25,27,28,30-35,38-42,45,48-50,54,59);
score (0,1,2)        (0,.5,1)         !item(26,37,58);
score (0,1,2,3)      (0,.5,1,1.5)     !item29,36,44,51,52,53,60);
score (0,1,2,3,4)    (0,.5,1,1.5,2)   !item(43,46,57);
score (0,1,2,3,4,5)  (0,.5,1,1.5,2,2.5) !item(47,55,56);

/* 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: Assignment of test items to the Process Components and Software Applications

No.	Item	Bookl. 1 Position	Bookl. 2 Position	Bookl. 3 Position	Process Component	Software Application
1	icg5005x_sc3g9_c	1	1		Access	Spreadsheet / presention
2	icg5034x_sc3g9_c	2			Access	Word processing
3	icg5009x_sc3g9_c	3	3		Create	Spreadsheet / presention
4	icg5051x_c	4			Create	Spreadsheet / presention
5	icg5018x_sc3g9_c	5			Manage	Word processing
6	Icg9106x_sc3g9_c	6	6	6	Manage	Internet / search engines
7	icg5015x_sc3g9_c	7	7		Create	Word processing
8	icg5046x_sc3g9_c	8	8		Manage	Internet / search engines
9	icg5033x_sc3g9_c	9	9		Create	Spreadsheet / presention
10	icg9110x_sc3g9_c	10	10	10	Access	Word processing
11	icg5045x_c	11	11		Access	Word processing
12	icg5054x_sc3g9_c	12	12		Access	Word processing
13	icg5021x_sc3g9_c	13			Access	Internet / search engines
14	icg9114x_sc3g9_c	14			Evaluate	E-mail / communication
15	icg5059x_sc3g9_c	15	15		Access	Word processing
16	icg9116x_sc3g9_c	16			Evaluate	Internet / search engines
17	icg5035x_c	17			Create	Word processing
18	icg9118x_sc3g9_c	18	18	18	Manage	Spreadsheet / presention
19	icg9119x_sc3g9_c	19			Create	Spreadsheet / presention
20	icg5003x_sc3g9_c	20	20	20	Manage	E-mail / communication
21	icg5029x_c	21	21	21	Create	Spreadsheet / presention
22	icg9122x_sc3g9_c	22	22	22	Manage	Internet / search engines
23	icg9123x_sc3g9_c	23	23		Manage	Internet / search engines
24	icg12041x_sc3g9_c	24	24	24	Manage	Word processing
25	icg12042x_c	25	25		Evaluate	Internet / search engines
26	icg12060s_sc3g9_c	26	26	26	Evaluate	Spreadsheet / presention
27	icg12036x_c	27	27	27	Access	Spreadsheet / presention
28	icg5039x_sc3g9_c	28			Access	Word processing
29	icg12018s_sc3g9_c	29	29		Access	E-mail / communication
30	icg5053x_sc3g9_c	30	30	30	Evaluate	Internet / search engines
31	icg9131x_sc3g9_c	31	31	31	Evaluate	Internet / search engines
32	icg9132x_sc3g9_c	32	32	32	Evaluate	Internet / search engines
33	icg5049x_sc3g9_c	33			Evaluate	Spreadsheet / presention
34	icg12022x_c	34			Evaluate	Spreadsheet / presention
35	icg9138x_sc3g9_c	35	35		Evaluate	Internet / search engines
36	icg9140s_sc3g9_c	36			Evaluate	E-mail / communication
37	icg9102s_sc3g9_c		2	2	Manage	E-mail / communication
38	icg5047x_sc3g9_c		4	4	Create	Word processing
39	icg12034x_sc3g9_c		5	5	Access	Spreadsheet / presention
40	icg9113x_sc3g9_c		13	13	Access	E-mail / communication
41	icg12040x_sc3g9_c		14	14	Manage	Internet / search engines

No.	Item	Bookl. 1 Position	Bookl. 2 Position	Bookl. 3 Position	Process Component	Software Application
42	icg12043x_c		16	16	Access	Spreadsheet / presention
43	icg9117s_sc3g9_c		17	17	Evaluate	Internet / search engines
44	ica5021s_sc3g9_c		19	19	Access	Word processing
45	icg9128x_sc3g9_c		28	28	Create	Spreadsheet / presention
46	icg9133s_sc3g9_c		33	33	Manage	Spreadsheet / presention
47	icg9136s_sc3g9_c		34	34	Manage	Internet / search engines
48	icg12027x_c		36	36	Access	Word processing
49	icg9101x_sc3g9_c			1	Create	Word processing
50	icg9103x_sc3g9_c			3	Manage	Internet / search engines
51	icg9107s_sc3g9_c			7	Manage	Internet / search engines
52	icg12138s_sc3g9_c			8	Access	E-mail / communication
53	icg12016s_sc3g9_c			9	Access	Word processing
54	icg9111x_sc3g9_c			11	Create	Spreadsheet / presention
55	icg12047s_sc3g9_c			12	Access	Word processing
56	icg12046s_sc3g9_c			15	Access	Spreadsheet / presention
57	ica5052s_sc3g9_c			23	Create	Word processing
58	ica9125s_sc3g9_c			25	Evaluate	Internet / search engines
59	icg9129x_sc3g9_c			29	Create	E-mail / communication
60	icg12050s_sc3g9_c			35	Evaluate	Internet / search engines

Note. Bookl. 1 = test with low level of difficulty; Bookl. 2 = test with medium level of difficulty; Bookl. 3 = test with high level of difficulty;