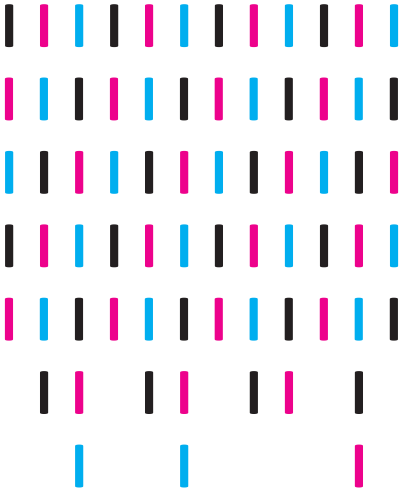

NEPS SURVEY PAPERS

Martin Senkbeil and Jan Marten Ihme

NEPS TECHNICAL REPORT FOR COMPUTER LITERACY: SCALING RESULTS OF STARTING COHORT 4 FOR GRADE 12

NEPS Survey Paper No. 25
Bamberg, June 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 4 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. (2017). *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. doi:10.5157/NEPS:SP25:1.0

Acknowledgements:

This report is an extension to NEPS working paper 17 (Senkbeil & Ihme, 2012) that presents the scaling results for computer literacy of starting cohort 4 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, 2012) 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 4 for Grade 12

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 4 (ninth grade). The computer literacy test contained 32 items with different response formats representing different cognitive requirements and different content areas. The test was administered to 5,762 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

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 and colleagues (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 4 (ninth grade) in grade 12. First, the main concepts of the computer literacy test are introduced. Then, the computer literacy data of starting cohort 4 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. The first two process components (a*ccess*, *create*) refer to the facet of technological literacy; whereas the other two process components (*manage*, *evaluate*) refer to the facet of information literacy (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.

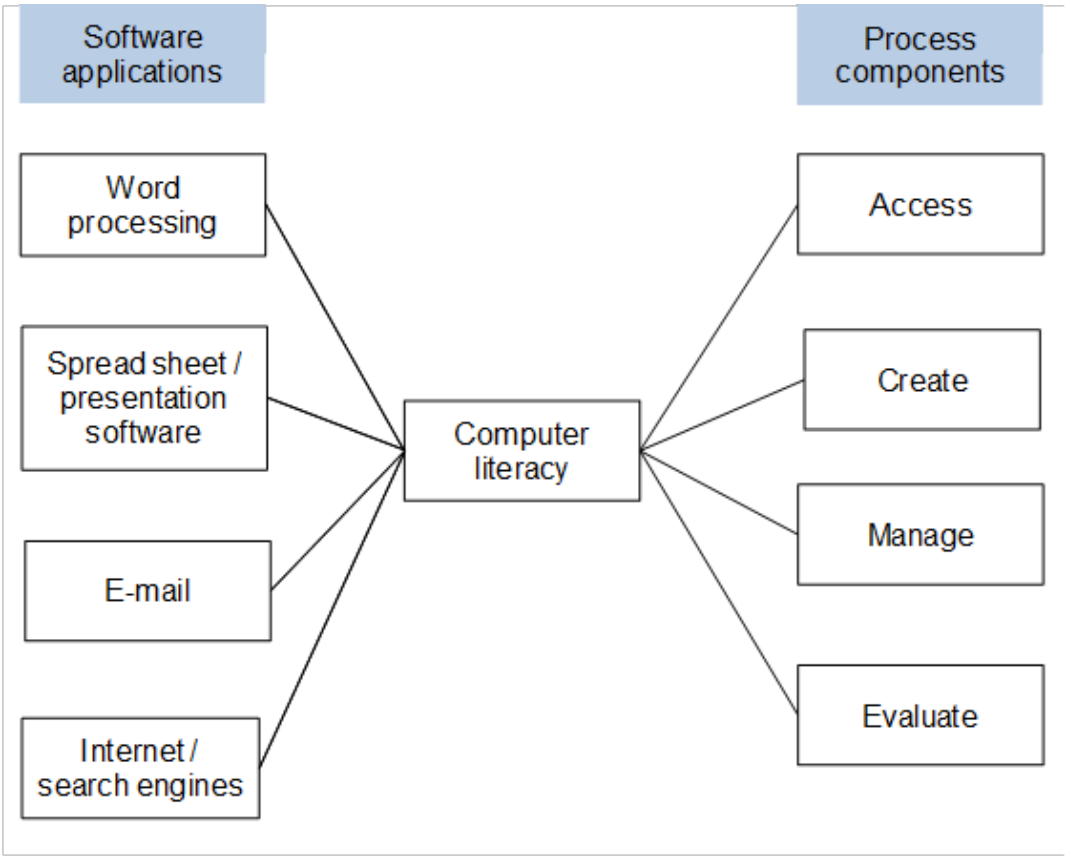

Figure 1. Assessment framework for computer literacy (process components and software applications).

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 4 (ninth 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 two 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 assessment setting (i.e., the context of test administration) and (b) the position of the computer literacy test within the test battery.

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 identified a poor fit for one CMC item (ica5018s_c; weighted mean square > 1.20; see 4.3 for further explanation). Therefore, this item was removed from the final scaling procedure. Thus, the analyses presented in the following sections and the competence scores derived for the respondents are based on the remaining 31 items. The charactersitics of the final set of 31 items are depicted in Table 1 regarding process components, in Table 2 regarding software applications, and in Table 3 regarding response formats.

Table 1. Process Components of Items in the Computer Literacy Test

Process components	**Frequency**
Access	6
Create	8
Manage	9
Evaluate	8
Total number of items	31

Table 2.Software Applications of Items in the Computer Literacy Test

Software applications	**Frequency**
Word processing	6
Spreadsheet / presentation software	11
E-mail / communication tools	5
Internet / search engines	9
Total number of items	31

The study assessed different competence domains including, among others, mathematics, computer literacy, and reading competence. The competence tests for these three domains were always presented first within the test battery. However, the tests were administered to participants in 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 reading test). There was no multi-matrix design regarding the order of the items *within* a specific test. All subjects received the test items in the same order.

Table 3. Response Formats of Items in the Computer Literacy Test

Software applications	**Frequency**
Simple multiple choice	13
Complex multiple choice	18
Total number of items	31

3.2 Sample

A total of 5,762 individuals received the computer literacy test. Since at least three valid item responses were available for all but one subjects, the analyses presented in this paper are based on the sample of 5,761 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**	**At home**	**Total**
Test	**First position**	1947	1413	3360
position	**Second position**	1925	477	2402
	Total	3872	1890	5762

4 Analyses

4.1 Missing Responses

There are different kinds of missing responses. These are a) invalid responses, b) missing responses due to omitted items, c) missing responses due to items that are not reached, d) missing responses due to items that are not administered, and e) missing responses that are not determinable. In this study, all subjects received the same set of items; thus, there are no items that were not administered to a person. 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.

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 Rasch model was used and estimated in ConQuest (Adams, Wu, & Wilson, 2015). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

CMC items consist 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 four CMC items (icg12138s_c, icg12048s_c, icg12119s_c, icg12046s_c) the lowest two categories were collapsed, for seven CMC items (icg12013s_c, icg12107s_c, icg12004s_c, icg12016s_c, icg12054s_c, icg12060s_c, ica5052s_sc4g12_c) the lowest three categories were collapsed, for four CMC items (icg12018s_c, icg12028s_c, icg12050s_c, ica5021s_scg12_c) the lowest four categories were collapsed, and for one CMC item (icg12047s_c) the lowest five categories were collapsed and scored with 0 points (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).

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). Person

parameter estimation in NEPS is described in Pohl & Carstensen (2012), whereas the data available in the SUF are described in Section 7.

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 types (high school vs. vocational school), 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. Moreover, in light of the quasi-experimental design measurement invariance analyses were also conducted for the administration setting. For this, we adopted the minimum effect null hypothesis and tested for the presence of negligible DIF. This procedure is described in Fischer, Rohm, Gnambs, and Carstensen (2016).

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 different construction criteria. First, a model with four process components representing the knowledge and skills needed for a problem-oriented use of ICT, 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., administration setting). Overall, there were very few invalid responses. Between 92% and 96% of the respondents did not have any invalid response at all; overall less than one 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, 52% (at home) and 62% (at school), did not skip any item. Less than five percent in the experimental condition "At school", and less than ten percent in the experimental

condition “At home” omitted more than three items. Thus, there was a slight difference in the amount of omitted items between the different experimental conditions.

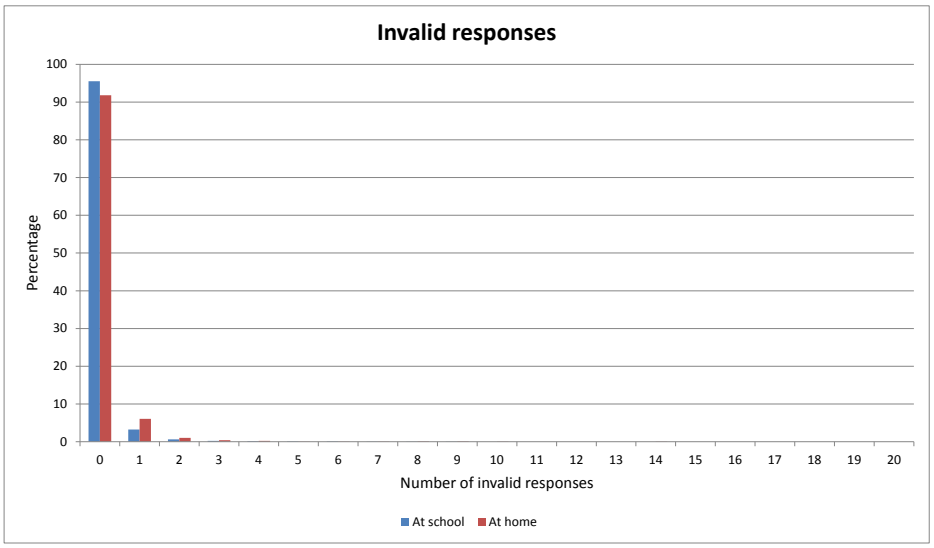

Figure 2. Number of invalid responses by administration setting.

Another source of missing responses is 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 high, because many respondents were unable to finish the test within the allocated time limit (Figure 4). Two-thirds (68%) of the respondents at school as well as at home finished the entire test. Out of those respondents who received the test at school, about 10% did not reach the last five items, whereas out of those who received the test at home, about 12% did not reach the last five items.

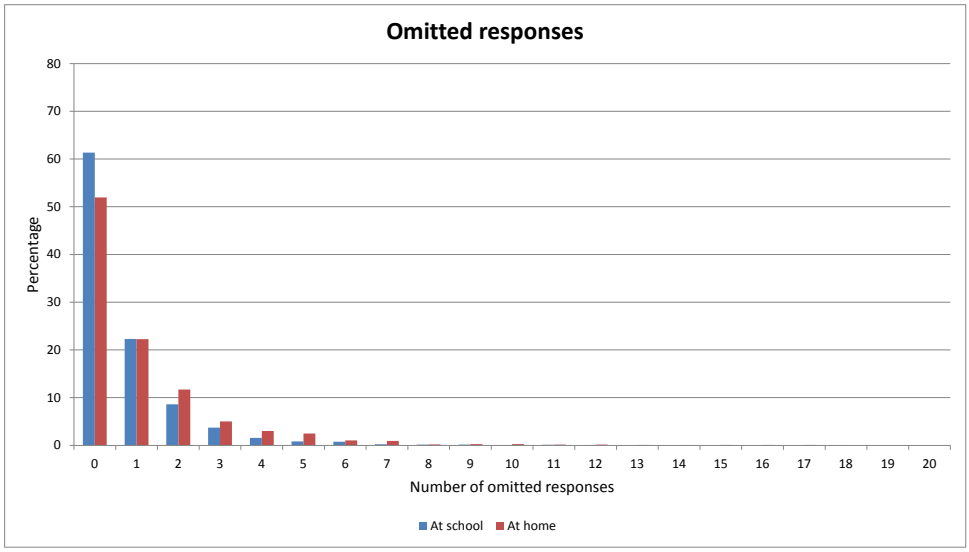

Figure 3. Number of omitted items by administration setting.

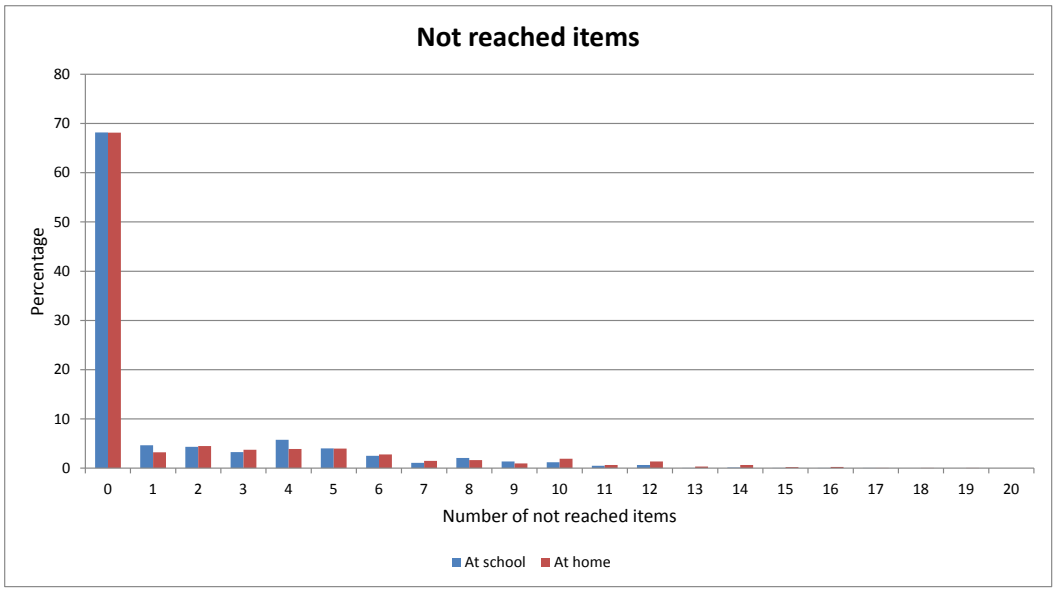

Figure 4. Number of not reached items by administration setting.

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* = 2.30 (*SD*= 3.16; at school) and *M* = 3.07 (*SD* = 3.94; at home) missing responses in the different experimental conditions. About 34% (at home) to 42% (at school) of the respondents had no missing response at all and about 20% of the participants had four or more missing responses regardless of the administration setting.

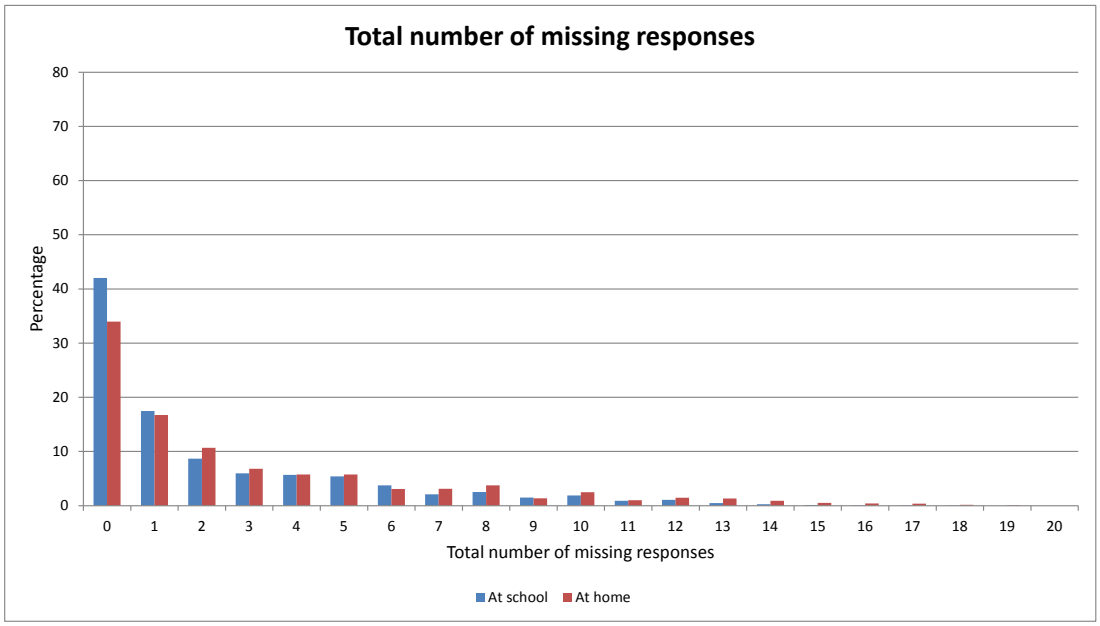

Figure 5. Total number of missing responses by administration setting.

Overall, there was a small amount of invalid responses and a reasonable amount of omitted items. The number of not reached items was rather large and, therefore, so was the total number of missing responses.

5.1.2 Missing responses per item

Tables 5 provides information on the occurrence of different kinds of missing responses per item by assessment setting (at school and at home). Overall, in both tests the omission rates were rather low, varying across items between 0.00% and 10.00%. There was only one item with an omission rate exceeding 10% (icg12016s_c, when the item was administered in the home condition). The omission rates correlated with the item difficulties at about .32 in the school context and about .16 at home. Generally, the percentage of invalid responses per item (columns 6 and 10 in Table 5) was rather low with the maximum rate being 2.28%. With an item's progressing position in the test, the amount of persons that did not reach the item (columns 4 and 8 in Table 5) rose up to a considerable amount of 32% for each experimental condition. The last items of the test were not reached by many respondents (see Figure 6). The total number of missing responses (sum of invalid, omitted, not-reached, and not determinable responses) per item varied between 1.0% (item icg12018s_c; administered at school) and 33.3% (item icg12119s_c; administered in the home condition).

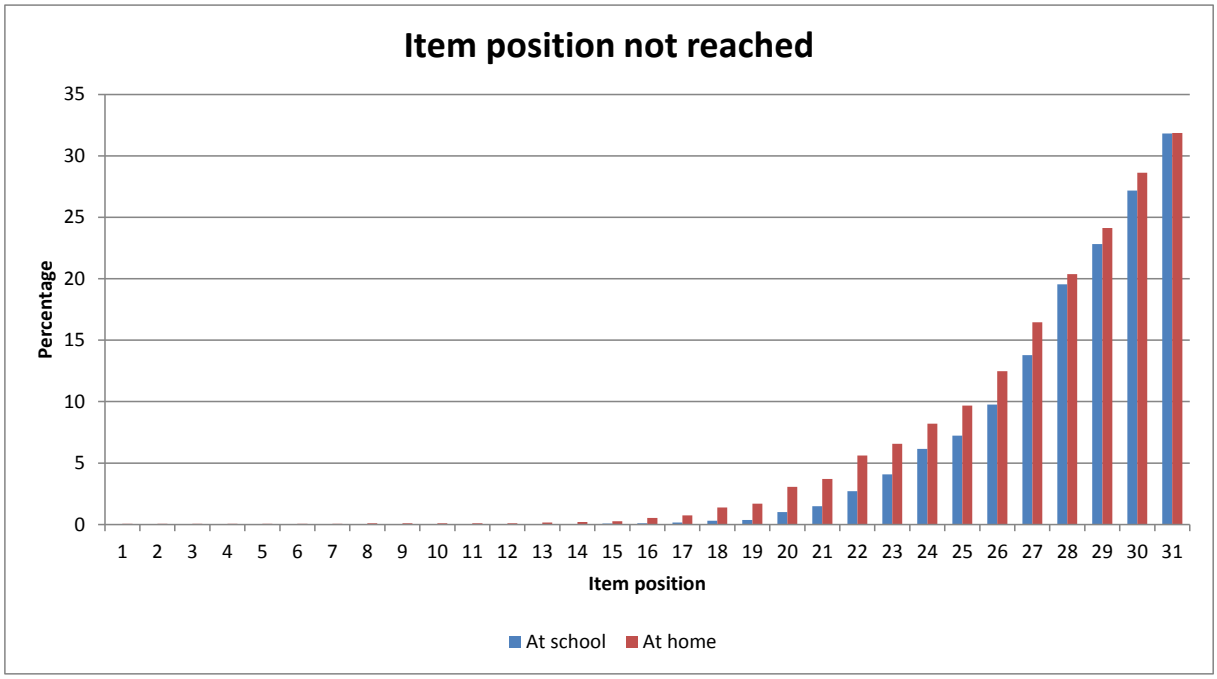

Figure 6. Item position not reached by administration setting.

Table 5. Percentage of Missing Values by Assessment Setting

		At school				*At home*			
Item	**Position**	**N**	**NR**	**OM**	**NV**	**N**	**NR**	**OM**	**NV**
icg12018s_c	1	3835	0.00	1.00	0.03	1823	0.00	3.39	0.11
ica5003x_sc4g12_c	2	3807	0.00	0,21	1.47	1848	0.00	0.21	1.96
icg12107s_c	3	3819	0.00	1,34	0.03	1823	0.00	3.49	0.00
icg12004s_c	4	3696	0.00	4,47	0.08	1752	0.00	6.88	0.37
icg12010x_c	5	3832	0.00	0,85	0.18	1847	0.00	1.90	0.32
icg12011x_c	6	3838	0.00	0,75	0.13	1856	0.00	1.32	0.42
ica5008x_sc4g12_c	7	3758	0.00	1,39	1.55	1750	0.00	5.08	2.28
icg12060s_c	8	3815	0.00	1,39	0.08	1809	0.11	4.07	0.11
icg12013s_c	9	3843	0.00	0,57	0.18	1831	0.11	2.22	0.79
icg12016s_c	11	3491	0.00	9,81	0.03	1621	0.11	14.02	0.11
ica5019x_sc4g12_c	12	3834	0.00	0,31	0.67	1815	0.11	1.90	1.96
icg12121x_c	13	3810	0.00	1,03	0.57	1827	0.11	1.75	1.48
icg12028s_c	14	3724	0.00	3,80	0.03	1751	0.16	7.20	0.00
ica5023x_sc4g12_c	15	3840	0.00	0,57	0.26	1853	0.21	1.38	0.37
ica5027x_sc4g12_c	16	3828	0.08	0.70	0.36	1838	0.26	2.01	0.48
icg12033x_c	17	3772	0.10	2,32	0.15	1786	0.53	4.66	0.32
icg12034x_c	18	3856	0.15	0,23	0.03	1860	0.74	0.63	0.21
icg12035x_c	19	3709	0.31	3,85	0.05	1766	1.38	5.03	0.16
icg12040x_c	20	3832	0.36	0,67	0.00	1839	1.69	0.74	0.26
icg12037s_c	21	3562	1.01	6,71	0.28	1692	3.07	6.98	0.42
icg12138s_c	22	3737	1.50	1,99	0.00	1731	3.70	4.66	0.05
icg12047s_c	23	3473	2.71	7,21	0.39	1610	5.61	8.84	0.37
icg12041x_c	24	3573	4.08	3,43	0.21	1712	6.56	2.17	0.69
icg12046s_c	25	3398	6.15	5,94	0.15	1595	8.20	7.20	0.21
ica5021s_sc4g12_c	26	3497	7.23	2,43	0.03	1629	9.68	3.97	0.16
ica5052s_sc4g12_c	27	3383	9.76	2,76	0.10	1590	12.49	3.33	0.05
icg12048s_c	28	3261	13.79	1,83	0.15	1528	16.46	2.54	0.16
icg12050s_c	29	3011	19.55	2,63	0.05	1428	20.37	3.76	0.32
icg12054s_c	30	2899	22.83	2,25	0.05	1394	24.13	2.12	0.00
icg12109s_c	31	2780	27.17	0,93	0.10	1330	28.62	0.95	0.05
icg12119s_c	32	2610	31.82	0,72	0.05	1260	31.85	1.43	0.05

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.

The item on position 10 was excluded from the analyses due to an unsatisfactory item fit (see section 3.1).

5.2 Parameter Estimates

5.2.1 Item parameters

The second column in Table 6 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 23% and 78% with an average of 48% (*SD* = 16%) correct responses.

The estimated item difficulties (for dichotomous variables) and location parameters (for the polytomous variable) are given in Table 6. The step parameters for the polytomous variable are depicted in Table 7. 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 -2.19 (item ica5021s_sc4g12_c) to 1.32 (item ica5003x_sc4g12_c) with an average difficulty of -0.46. 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 ≤ 0.05).

Table 6. Item parameters

Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
icg12018s_c	74.44	-1.164	0.032	1.00	0.1	0.31	0.64	.016
ica5003x_sc4g12_c	23.11	1.320	0.033	1.02	1.0	0.27	0.53	.019
icg12107s_c	n.a.	-0.587	0.032	0.96	-2.1	0.39	0.96	.020
icg12004s_c	n.a.	-0.124	0.025	0.99	-0.4	0.42	0.74	.020
icg12010x_c	49.16	0.041	0.028	1.02	2.7	0.32	0.56	.017
icg12011x_c	23.06	1.323	0.033	0.97	-1.8	0.36	0.92	.021
ica5008x_sc4g12_c	68.66	-0.850	0.030	1.08	6.0	0.20	0.26	.021
icg12060s_c	n.a.	-1.512	0.038	0.99	-0.6	0.29	0.79	.015
icg12013s_c	76.88	-1.308	0.033	1.04	2.5	0.22	0.39	.021
icg12016s_c	n.a.	-0.365	0.035	1.02	1.4	0.27	0.56	.016
ica5019x_sc4g12_c	26.57	1.124	0.031	1.05	3.1	0.22	0.35	.020
icg12121x_c	40.46	.430	0.028	1.01	1.7	0.32	0.61	.021
icg12028s_c	n.a.	-1.359	0.038	1.03	2.1	0.20	0.43	.027
ica5023x_sc4g12_c	47.62	0.109	0.028	1.02	3.1	0.31	0.55	.017
ica5027x_sc4g12_c	47.85	0.100	0.028	1.01	1.8	0.33	0.58	.021
icg12033x_c	60.15	-0.443	0.029	0.97	-2.7	0.39	0.83	.016
icg12034x_c	69.23	-0.889	0.030	1.00	-0.4	0.34	0.68	.016

icg12035x_c	52.07	-0.082	0.028	1.10	11.3	0.19	0.21	.026
icg12040x_c	44.26	0.258	0.028	0.99	-1.1	0.37	0.75	.023
icg12037s_c	n.a.	-0.342	0.037	0.93	-5.8	0.44	1.42	.027
icg12138s_c	n.a.	-1.584	0.031	1.00	0.1	0.33	0.71	.023
icg12047s_c	n.a.	-0.287	0.020	0.96	-2.2	0.54	0.82	.024
icg12041x_c	65.79	-0.719	0.030	1.01	0.8	0.33	0.64	.013
icg12046s_c	n.a.	-0.533	0.021	0.99	-0.5	0.50	0.74	.028
ica5021s_sc4g12_c	n.a.	-2.191	0.045	0.93	-3.7	0.42	1.63	.029
ica5052s_sc4g12_c	n.a.	-0.285	0.029	0.93	-4.6	0.47	1.12	.024
icg12048s_c	n.a.	-1.010	0.025	1.05	2.7	0.36	0.51	.023
icg12050s_c	n.a.	-0.928	0.031	0.93	-4.3	0.48	1.16	.021
icg12054s_c	n.a.	-0.713	0.038	0.94	-4.1	0.41	1.15	.031
icg12109s_c	n.a.	-0.826	0.028	1.11	5.1	0.28	0.35	.025
icg12119s_c	n.a.	-0.840	0.027	0.90	-5.1	0.56	1.15	.027

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, Q3 = Average absolute residual correlation for item (Yen, 1983).

Item 10 was excluded from the analyses due to an unsatisfactory item fit (see section 3.1). Percent correct scores are not informative for polytomous CMC item scores. These are denoted by na.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 7. Step parameters (with Standard Errors) for the Polytomous Items

Item	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**
icg12107s_c	-0.718 (0.029)	-0.498 (0.027)	1.215			
icg12004s_c	-0.279 (0.031)	-1.119 (0.028)	1.080 (0.037)	0.318		
icg12060s_c	-0.133 (0.029)	0.133				
icg12016s_c	0.161 (0.031)	-0.161				
icg12028s_	-0.114 (0.029)	0.114				
icg12037s_c	-0.331 (0.028)	0.331				

icg12138s_c	-1.184	0.679	0.505			
	(0.028)	(0.030)				
icg12047s_c	-0.386	-0.852	-0.290	-0.016	0.407	1.105
	(0.040)	(0.035)	(0.030)	(0.029)	(0.036)	
icg12046s_c	-0.338	-0.433	0.187	-0.043	0.627	
	(0.032)	(0.029)	(0.029)	(0.033)		
ica5021s_sc4g12_c	-0.353	0.353				
	(0.031)					
ica5052s_sc4g12_c	-0.443	0.020	0.423			
	(0.029)	(0.031)				
icg12048s_c	-0.135	-0.129	-0.002	0.266		
	(0.029)	(0.029)	(0.032)			
icg12050s_c	-0.063	-0.188	0.251			
	(0.030)	(0.033)				
icg12054s_c	0.334	-0.334				
	(0.036)					
icg12109s_c	-0.452	-0.079	-0.265	0.797		
	(0.033)	(0.031)	(0.033)			
icg12119s_c	-0.175	-0.260	0.108	0.326		
	(0.033)	(0.032)	(0.036)			

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

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 0.46, indicating somewhat limited variability between subjects. The reliability of the test (EAP/PV reliability = .74; WLE reliability = .73) 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.

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

Figure 7. Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 35.8 cases. Item difficulty is depicted on the right side of the graph. Each number represents one item (see Table 6).

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 23% to 98% across all items (*Mdn* =77%). Thus, the number of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.93 to 1.10, the respective *t*-value from -9.0 to 9.5, 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 five items with a point-biserial-correlation between .01 and .04 (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 6). Values of the WMNSQ ranged from 0.90 (item icg12119s_c) to 1.101 (icg12109s_c). Only one 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 .19 (item icg12035x_c) to .56 (item icg12119s_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 assessment setting was also studied. Thus, we examined measurement invariance for the two assessment settings (test items were administered at school vs. test items were administered at home) by adopting the minimum effect null hypothesis described in Fischer et al. (2016). The differences between the estimated item difficulties in the various groups are summarized in Table 8. 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 9).

Table 8. Differential Item Functioning

Item	Sex	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_c	0.072	0.030	-0.144	-0.060	0.134	-0.116
ica5003x_sc4g12_c	-0.126	0.030	-0.032	0.138	-0.218	0.208
icg12107s_c	0.242	0.456	-0.314	-0.492	0.018	-0.446
icg12004s_c	0.060	0.210	0.012	-0.020	-0.048	0.022
icg12010x_c	0.012	-0.060	0.012	0.070	0.016	0.134
icg12011x_c	-0.386	0.176	-0.132	0.226	-0.072	0.312
ica5008x_sc4g12_c	0.182	-0.162	0.026	0.588	-0.148	0.468
icg12060s_c	0.060	0.152	-0.068	0.120	0.010	0.084
icg12013s_c	0.552	-0.174	0.128	0.186	0.000	0.124
icg12016s_c	-0.014	-0.198	0.418	0.088	0.024	0.038
ica5019x_sc4g12_c	0.042	-0.084	0.206	0.158	0.088	0.162
icg12121x_c	-0.446	-0.234	0.368	0.246	0.016	0.338
icg12028s_c	-0.138	-0.688	0.452	0.590	-0.128	0.520
ica5023x_sc4g12_c	-0.222	-0.106	0.068	0.174	-0.018	0.148
ica5027x_sc4g12_c	0.278	-0.138	-0.064	0.094	0.054	0.048
icg12033x_c	0.196	0.070	-0.214	-0.328	0.016	-0.414
icg12034x_c	0.512	0.176	-0.096	-0.444	0.052	-0.426
icg12035x_c	0.396	-0.116	0.172	0.288	-0.134	0.232
icg12040x_c	-0.536	-0.116	0.232	0.054	0.160	0.010
icg12037s_c	-0.080	0.374	-0.282	-0.570	0.218	-0.618
icg12138s_c	-0.342	0.146	-0.034	-0.138	0.030	-0.106
icg12047s_c	0.316	-0.006	-0.058	-0.132	-0.044	-0.076
icg12041x_c	-0.070	-0.056	-0.246	0.040	0.066	-0.012
icg12046s_c	-0.024	-0.006	-0.056	-0.044	-0.048	0.044
ica5021s_sc4g12_c	-0.402	0.430	-0.156	-0.566	0.248	-0.692
ica5052s_sc4g12_c	-0.084	0.158	-0.076	-0.026	-0.008	-0.028
icg12048s_c	0.494	-0.048	-0.040	-0.010	0.102	-0.050
icg12050s_c	-0.364	0.084	0.084	-0.422	0.002	-0.412
icg12054s_c	-0.792	-0.024	0.100	-0.022	-0.090	0.028
icg12109s_c	-0.038	-0.330	0.028	0.474	-0.154	0.504
icg12119s_c	-0.410	0.222	-0.096	-0.354	0.076	-0.412

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

Sex: The sample included 2,668 (46%) males and 3,094 (53%) females. On average, male participants had a higher estimated computer literacy than females (main effect = 0.186 logits, Cohen's *d* = 0.403). Only one item (item icg12054s_c) showed DIF greater than 0.6 logits. An overall test for DIF (see Table 9) 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 1,637 (28%) test takers with 0 to 100 books at home and 3,867 (67%) test takers with more than 100 books at home. 258 (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 -0.502 logits (Cohen's *d* =-1.089) 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 (icg12028s_c). Model comparisons using AIC and BIC both favored the model estimating DIF, but the deviation was small in both cases. Thus, overall, there was no pronounced DIF with regard to books.

Migration background: There were 4,265 participants (74%) with no migration background, 1,194 subjects (21%) with a migration background, and 303 individuals (5%) 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 = 0.362 logits, Cohen's *d* = 0.785). 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 9).

School type: Overall, 3,431 subjects (60%) who took the computer literacy test attended secondary school (German: "Gymnasium"), whereas 2,331 (40%) were enrolled in other school types. Subjects in secondary schools showed a higher computer literacy on average (0.706 logits; Cohen's *d* = 1.531) than subjects in other school types. There was no noteworthy item DIF; no item exhibited DIF greater than 0.6 logits. 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 3,360 (58%) persons received the computer literacy first and 2402 (42%) respondents took the computer literacy test after having completed either the mathematics or the reading test at home. 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 0.146 logits (Cohen's *d* = 0.317) 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 0.248 logits (item ica5021s_sc4g12_c). As a consequence, the overall test for DIF using the BIC favored the more parsimonious main effect model (Table 9).

Setting: The computer literacy test was administered in two different settings (see section 3.1 for the design of the study). A subsample of 3,872 (67%) persons received the computer literacy test in small groups at school, whereas 1,890 (33%) participants finished the test individually at their private homes. Subjects who finished the computer literacy test at school were on average 0.714 logits (Cohen's *d* = 1.549) 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 two items (item icg12037s_c, ica5021s_sc4g12_c). Moreover, 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; see Appendix B). Thus, overall, there was no pronounced DIF with regard to the different settings.

Table 9. Differential Item Functioning

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	main effect	5762	300,949.19	68	301,805.19	301,538.01
	DIF		300,088.39	99	300,286.39	300,945.63
Books	main effect	5504	287,198.43	68	287,334.43	287,784.13
	DIF		286,918.51	99	287,116.51	287,771.22
Migration	main effect	5459	285,399.12	68	285,535.12	285,984.26
	DIF		285,234.48	99	285,432.48	286,086.37
School type	main effect	5762	299,799.79	68	299,935.79	300,385.81
	DIF		299,117.25	99	299,315.25	299,974.49
Position	main effect	5762	300,978.16	68	301,114.16	301,566.98
	DIF		300,901.26	99	301,099.26	301,758.51
Setting	main effect	5762	299,864.98	68	300,000.98	300,453.80
	DIF		299,257.69	99	299,455.69	300,114.93

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 (2PL) that estimates discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 6), ranging from 0.21 (item icg12035x_c) to 1.63 (item ica5021s_sc4g12_c). The average discrimination parameter fell at 0.73. Model fit indices suggested a slightly better model fit of the 2PL model (AIC = 299,626.04, BIC

=300,271.97) as compared to the 1PL model (AIC = 301,160.78, BIC =301,606.93). Despite the empirical preference for the 2PL model, the 1PL 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 C. 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 10. The correlations between the dimensions varied between .86 and .93. The smallest correlation was found between Dimension 1 ("Access") and Dimension 4 ("Evaluate"). Dimension 1 ("Access") and Dimension 2 ("Create") showed the strongest correlation. All correlations deviated from a perfect correlation (i.e., they were marginally lower than r = .95, see Carstensen, 2013). According to model fit indices, the AIC favored the four-dimensional model (AIC = 301,116.60, number of parameters = 76; unidimensional model: AIC = 301,160.78, number of parameters = 67), whereas the BIC favored the unidimensional model (BIC = 301,606.93; four-dimensional model: BIC = 301,622.69). These results indicate that the four process components measure a common construct, albeit it is not completely unidimensional.

Table 10. Results of Four-Dimensional Scaling (Process Components)

	Access	**Create**	**Manage**	**Evaluate**
Access (8 Items)	(0.519)			
Create (7 Items)	.933	(0.511)		
Manage (9 Items)	.902	.931	(0.573)	
Evaluate (5 Items)	.860	.873	.910	(0.432)

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 aoolications reported in Table 11. The correlations among the three dimensions were rather high and fell between .88 and .95. The smallest correlation was found between Dimension 2 ("Spreadsheet / presentation software") and Dimension 4 ("Internet / search engines"). Dimension 1 ("Word processing") and Dimension 2 ("Spreadsheet / presentation software"") 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 = 300,396.25, BIC = 301,442.34, number of parameters = 76) than the unidimensional model (AIC = 301,160.78, BIC = 301,606.93, number of parameters = 67).

However, for the unidimensional model the average absolute residual correlations as indicated by the corrected Q_3 statistic (see Table 6) were quite low (M = .022, SD = .004) — the largest individual residual correlation was .031 — 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 11. Results of Four-Dimensional Scaling (Software Applications).

	Dim 1	**Dim 2**	**Dim 3**	**Dim 4**
Word processing (Dim1) (6 Items)	(0.854)			
Spreadsheet / presentation software (Dim 2) (10 Items)	.952	(0.506)		
E-mail / communication tools (Dim 3) (4 Items)	.927	.914	(0.438)	
Internet / search engines (Dim 4) (9 Items)	.902	.884	.904	(0.372)

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 4 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 18 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 4, the computer literacy administered in grades 9 (see Senkbeil & Ihme, 2012) 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 4 across the two grades.

7.2.1 Samples

In starting cohort 4, a subsample of 5,559 students participated at both measurement occasions, in grade 9 and also in grade 12. 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 11 received both tests within a single measurement occasion.

7.2.2 The design of the link study

The test administered in grade 9 included 36 items (see Senkbeil & Ihme, 2012), whereas the test administered in grade 12 included 32 items (see above). Because preliminary analyses identified severe differential item functioning for one item of the grade 12 test (ica5018s_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 = 34,476.94, number of parameters = 123; two-dimensional model: BIC = 34,479.17, number of parameters = 125), whereas the AIC favored the two-dimensional model (AIC = 33,980.87; unidimensional model: AIC = 33,986.60). 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 4 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 12.

Table 12. Differential Item Functioning Analyses between the Starting Cohort and the Link Sample

	Grade 9				Grade 12			
No.	**Item**	**$\Delta\sigma$**	***SE*$_{\Delta\sigma}$**	***F***	**Item**	**$\Delta\sigma$**	***SE*$_{\Delta\sigma}$**	***F***
1	icg9101x_c	-0.09	0.11	0.70	icg12018s_c	-0.57	0.13	18.13
2	icg9102s_c	0.76	0.15	25.99	ica5003x_sc4g12_c	-1.06	0.13	62.93
3	icg9103x_c	0.73	0.11	41.10	icg12107s_c	-0.52	0.13	16.80
4	icg9104x_c	0.31	0.11	8.07	icg12004s_c	-0.50	0.09	28.18
5	icg9105x_c	2.79	0.58	23.02	icg12010x_c	-0.64	0.11	33.55
6	icg9106x_c	1.21	0.19	39.70	icg12011x_c	-0.76	0.12	37.29
7	icg9107s_c	-0.12	0.15	0.67	ica5008x_sc4g12_c	-1.08	0.12	86.65
8	icg9109x_c	0.58	0.22	7.22	icg12060s_c	-0.80	0.15	29.06
9	icg9110x_c	0.20	0.11	3.08	icg12013s_c	-0.72	0.14	28.56
10	icg9111x_c	0.52	0.12	18.40	icg12016s_c	-0.69	0.14	24.62
11	icg9112x_c	-0.07	0.14	0.21	ica5019x_sc4g12_c	-1.14	0.13	77.92
12	icg9113x_c	0.48	0.11	18.67	icg12121x_c	-0.73	0.11	44.13
13	icg9114x_c	0.42	0.15	7.66	icg12028s_c	-0.82	0.15	28.44
14	icg9116x_c	0.07	0.18	0.13	ica5023x_sc4g12_c	-0.77	0.11	49.84
15	icg9117s_c	0.79	0.18	19.49	ica5027x_sc4g12_c	-0.59	0.11	28.67
16	icg9118x_c	0.45	0.14	10.81	icg12033x_c	-0.37	0.12	9.37
17	icg9119x_c	1.06	0.20	28.94	icg12034x_c	-0.59	0.13	21.67
18	icg9121x_c	0.96	0.17	31.84	icg12035x_c	-0.76	0.11	44.99
19	icg9122x_c	0.54	0.13	17.45	icg12040x_c	-0.40	0.11	13.38

20	icg9123x_c	0.81	0.18	20.43	icg12037s_c	-0.15	0.15	0.96
21	icg9123x_c	0.29	0.24	1.46	icg12138s_c	-0.47	0.13	13.71
22	icg9125s_c	0.78	0.16	23.82	icg12047s_c	-0.61	0.08	57.01
23	icg9126x_c	0.24	0.14	3.00	icg12041x_c	-0.72	0.12	34.38
24	icg9127x_c	1.12	0.20	29.84	icg12046s_c	-0.70	0.08	67.80
25	icg9128x_c	0.10	0.12	0.63	ica5021s_sc4g12_c	0.32	0.22	2.12
26	icg9129x_c	0.38	0.11	10.79	ica5052s_sc4g12_c	-0.59	0.11	27.36
27	icg9130x_c	0.23	0.13	3.32	icg12048s_c	-0.76	0.10	59.85
28	icg9131x_c	0.87	0.16	28.28	icg12050s_c	-0.43	0.13	11.38
29	icg9132x_c	0.79	0.18	19.57	icg12054s_c	-0.54	0.15	12.37
30	icg9133s_c	0.44	0.11	16.55	icg12109s_c	-0.83	0.10	64.19
31	icg9134x_c	0.36	0.21	2.74	icg12119s_c	-0.69	0.11	39.38
32	icg9135x_c	0.76	0.15	25.31				
33	icg9136s_c	-0.83	0.08	120.18				
34	icg9137x_c	0.62	0.18	11.54				
35	icg9138x_c	0.33	0.17	3.84				
36	icg9140s_c	0.85	0.27	9.47				

Note. $\Delta\sigma$ = Difference in item difficulty parameters between the longitudinal subsample in grade 9 or 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, 5,957) = 127.84. A non-significant test indicates measurement invariance.

Analyses of differential item functioning between the link sample and starting cohort 4 identified neither for grade 9 (difference in logits: *Min* = 0.07, *Max* = 1.79) nor for grade 12 (difference in logits: *Min* = 0.15, *Max* = 1.13) 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 $c = 0.687$. This correction term was subsequently added to each difficulty parameter estimated in grade 12 (see Table 6) 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, "icg12_sc1" and "icg12_sc1u", including their respective standard errors "icg12_sc2" and "icg12_sc2u". Both WLE scores are 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 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. 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. (2012). *NEPS Technical Report for Computer Literacy – Scaling Results of Starting Cohort 4 in Ninth Grade* (NEPS Working Paper No. 14). 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 4 (grade 12)

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

/* load data */
datafile >>filename.dat;
format pid 1-7 responses 9-39;
labels <<filename_with_labels.txt;

/* collapse response categories */
codes 0,1,2,3,4,5,6,;
recode (0,1,2)          (0,1,2)          !item(20);      /* icg12037s_c */
recode (0,1,2,3)        (0,0,0,1)        !item(9);       /* icg12013s_c */
recode (0,1,2,3,4)      (0,0,0,0,1)      !item(1);       /* icg12018s_c */
recode (0,1,2,3,4)      (0,0,0,1,2)      !item(8);       /* icg12060s_c */
recode (0,1,2,3,4)      (0,0,0,1,2)      !item(10);      /* icg12016s_c */
recode (0,1,2,3,4)      (0,0,0,1,2)      !item(29);      /* icg12054s_c */
recode (0,1,2,3,4)      (0,0,1,2,3)      !item(21);      /* icg12138s_c */
recode (0,1,2,3,4)      (0,1,2,3,4)      !item(30);      /* icg12109s_c */
recode (0,1,2,3,4,5)    (0,0,0,0,1,2)    !item(13);      /* icg12028s_c */
recode (0,1,2,3,4,5)    (0,0,0,0,1,2)    !item(25);      /* ica5021s_sc4g12_c */
recode (0,1,2,3,4,5)    (0,0,0,1,2,3)    !item(3);       /* icg12107s_c */
recode (0,1,2,3,4,5)    (0,0,0,1,2,3)    !item(26);      /* ica5052s_sc4g12_c */
recode (0,1,2,3,4,5)    (0,0,1,2,3,4)    !item(27);      /* icg12048s_c */
recode (0,1,2,3,4,5)    (0,0,1,2,3,4)    !item(31);      /* icg12119s_c */
recode (0,1,2,3,4,5,6)  (0,0,0,0,1,2,3)  !item(28);      /* icg12050s_c */
recode (0,1,2,3,4,5,6)  (0,0,0,1,2,3,4)  !item(4);       /* icg12004s_c */
recode (0,1,2,3,4,5,6)  (0,0,1,2,3,4,5)  !item(24);      /* icg12046s_c */
recode (0,1,2,3,4,5,6)  (0,1,2,3,4,5,6)  !item(22);      /* icg12047s_c */

/* scoring */
score (0,1)             (0,1)            !item(1,2,5-7,9,11,12,14-19,23);
score (0,1,2)           (0,.5,1)         !item(8,10,13,20,25,29);
score (0,1,2,3)         (0,.5,1,1.5)     !item(3,21,26,28);
score (0,1,2,3,4)       (0,.5,1,1.5,2)   !item(4,27,30,31);
score (0,1,2,3,4,5)     (0,.5,1,1.5,2,2.5) !item(24);
score (0,1,2,3,4,5,6)   (0,.5,1,1.5,2,2.5,3) !item(22);

/* 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)

No.	Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	*F*
1	icg12018s_c	-0.11	0.07	2.98
2	ica5003x_sc4g12_c	0.17	0.07	5.40
3	icg12107s_c	-0.42	0.07	37.07
4	icg12004s_c	-0.06	0.05	1.07
5	icg12010x_c	0.12	0.06	3.87
6	icg12011x_c	0.28	0.07	14.86
7	ica5008x_sc4g12_c	0.47	0.06	54.24
8	icg12060s_c	0.36	0.08	19.34
9	icg12013s_c	0.13	0.07	3.82
10	icg12016s_c	0.03	0.07	0.16
11	ica5019x_sc4g12_c	0.13	0.07	3.29
12	icg12121x_c	0.32	0.06	26.58
13	icg12028s_c	0.39	0.08	24.50
14	ica5023x_sc4g12_c	0.13	0.06	4.71
15	ica5027x_sc4g12_c	0.03	0.06	0.26
16	icg12033x_c	-0.43	0.06	47.81
17	icg12034x_c	-0.43	0.06	46.71
18	icg12035x_c	0.22	0.06	13.12
19	icg12040x_c	-0.01	0.06	0.03
20	icg12037s_c	-0.63	0.08	62.59
21	icg12138s_c	-0.05	0.07	0.60
22	icg12047s_c	-0.03	0.04	0.44
23	icg12041x_c	-0.01	0.06	0.05
24	icg12046s_c	0.03	0.04	0.39
25	ica5021s_sc4g12_c	-0.56	0.09	35.22
26	ica5052s_sc4g12_c	-0.13	0.06	4.69
27	icg12048s_c	0.10	0.05	3.25
28	icg12050s_c	-0.37	0.07	30.69
29	icg12054s_c	0.05	0.08	0.40
30	icg12109s_c	0.63	0.06	107.56
31	icg12119s_c	-0.32	0.06	29.82

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, 5,762) = 124.86. A non-significant test indicates measurement invariance.

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

No.	Item	Process Component	*Software Application*
1	icg12018s_c	Manage	E-mail / communication tools
2	ica5003x_sc4g12_c	Evaluate	Internet / search engines
3	icg12107s_c	Evaluate	Spreadsheet / presentation software
4	icg12004s_c	Create	E-mail / communication tools
5	icg12010x_c	Create	Spreadsheet / presentation software
6	icg12011x_c	Manage	Spreadsheet / presentation software
7	ica5008x_sc4g12_c	Evaluate	Internet / search engines
8	icg12060s_c	Manage	Spreadsheet / presentation software
9	icg12013s_c	Manage	Internet / search engines
10	icg12016s_c	Access	Word processing
11	ica5019x_sc4g12_c	Evaluate	Internet / search engines
12	icg12121x_c	Access	Spreadsheet / presentation software
13	icg12028s_c	Access	E-mail / communication tools
14	ica5023x_sc4g12_c	Create	Spreadsheet / presentation software
15	ica5027x_sc4g12_c	Manage	E-mail / communication tools
16	icg12033x_c	Manage	Spreadsheet / presentation software
17	icg12034x_c	Access	Spreadsheet / presentation software
18	icg12035x_c	Create	Spreadsheet / presentation software
19	icg12040x_c	Manage	Internet / search engines
20	icg12037s_c	Manage	Spreadsheet / presentation software
21	icg12138s_c	Access	E-mail / communication tools
22	icg12047s_c	Create	Word processing
23	icg12041x_c	Manage	Word processing
24	icg12046s_c	Create	Spreadsheet / presentation software
25	ica5021s_sc4g12_c	Access	Word processing
26	ica5052s_sc4g12_c	Create	Word processing
27	icg12048s_c	Evaluate	Internet / search engines
28	icg12050s_c	Evaluate	Internet / search engines
29	icg12054s_c	Create	Word processing
30	icg12109s_c	Evaluate	Internet / search engines
31	icg12119s_c	Evaluate	Internet / search engines