NEPS
National Educational Panel Study

NEPS Working Papers

Martin Senkbeil & Jan Marten Ihme

NEPS Technical Report for Computer Literacy – Scaling Results of Starting Cohort 4 in Ninth Grade

NEPS Working Paper No. 17

Bamberg, November 2012

SPONSORED BY THE
Federal Ministry of Education and Research

Contact: German National Educational Panel Study (NEPS) − University of Bamberg − 96045 Bamberg − Germany − contact.neps@uni-bamberg.de

NEPS Technical Report for Computer Literacy – Scaling Results of Starting Cohort 4 in Ninth Grade

Martin Senkbeil & Jan Marten Ihme

Leibniz Institute for Science and Mathematics Education at University of Kiel

E-mail address of the lead author:

senkbeil@ipn.uni-kiel.de

Bibliographic data:

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. 17). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

We thank Steffi Pohl and Kerstin Haberkorn 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 in Ninth Grade

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competences across the whole life span and tests for assessing the different competence domains are developed. In order to evaluate the quality of the competence tests, a wide range of analyses have been performed based on Item Response Theory (IRT). This paper describes the computer literacy data of starting cohort 4 in ninth grade. Next to descriptive statistics of the data, the scaling model applied to estimate competence scores, analyses performed to investigate the quality of the scale, as well as the results of these analyses are presented. The reading test in fifth grade consisted of 36 items, which represented different cognitive requirements and text functions and used different response formats. The test was administered to 14,486 students. A partial credit model was used for scaling the data. Item fit statistics, differential item functioning, Rasch homogeneity, the tests' dimensionality, and local item independence were evaluated to ensure the quality of the test. The results show that the items exhibited good item fit and measurement invariance across various subgroups. Moreover, the test showed a high reliability and the different comprehension requirements foster a unidimensional construct. Challenges of the test are the small number of very difficult items, and the elevated number of items that have not been reached by test takers due to time limits. In summary, the scaling procedures show that the test is a reliable instrument with satisfying psychometric properties for assessing computer literacy. In the paper, the data available in the Scientific Use File are described and ConQuest-Syntax for scaling the data is provided.

Keywords

Item Response Theory, Scaling, Computer Literacy, Scientific Use File

Content

1. Introduction

Within the National Educational Panel Study (NEPS) different competences are measured coherently across the life span. Tests have been developed for different competence domains. These include, amongst others, reading competence, mathematical competence, scientific literacy, information and communication literacy, metacognition, vocabulary, and domain general cognitive functioning. Weinert et al. (2011) give an overview of the competence domains measured in NEPS.

Most of the competence data are scaled using models that are based on Item Response Theory (IRT). Since most of the competence tests were developed specifically for implementation in NEPS, several analyses have been performed 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 scales are described in Pohl and Carstensen (2012a). In this paper the results of these analyses are presented for computer literacy in the starting cohort 4. We first introduce the main concepts of the computer literacy test. Then, we describe the computer literacy data of starting cohort 4 and the analyses performed on the data for estimating competence scores and for checking the quality of the test. The results of these analyses are presented and discussed. Finally, we describe the data that are available for public in the Scientific Use File.

The present report has been modeled along the technical reports of Pohl, Haberkorn, Hardt, & Wiegand (2012) and Haberkorn, Pohl, Hardt, & Wiegand (2012). Note that the analyses of this report are based on the data set available at some time before data release. Due to data protection and data cleaning issues the data set in the Scientific Use File (SUF) may differ slightly from the data set used for the analyses in this paper. We do not, however, expect severe changes in the results.

2. Testing computer literacy

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

Computer literacy is conceptualized as a unidimensional construct comprising the 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 that assesses computer literacy in NEPS, we use a framework that identifies four process components (*access*, *create*, *manage*, and *evaluate*) of computer literacy that represent 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 (Figure 1). Apart from the process components, the test construction of TILT is guided by a categorization of software applications (*word processing*, *spreadsheet*, *presentation software*, *e-mail / communication tools* and *internet / search engines*) that are used to locate, process, present, and communicate information.

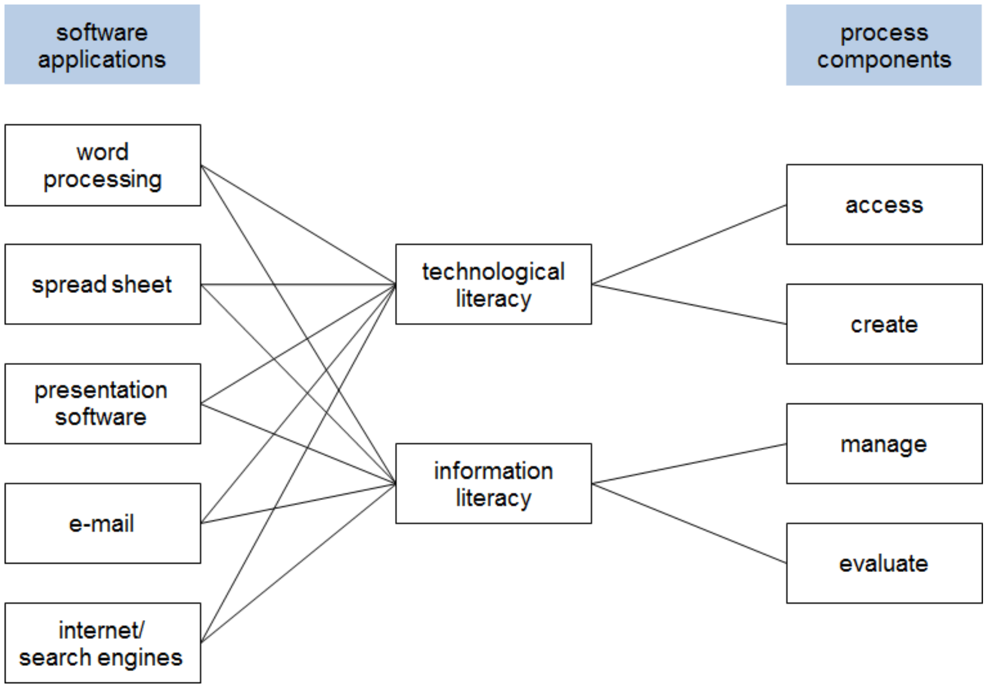

Figure 1: Assessment framework for computer literacy (process components and software applications)

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

In computer literacy 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., are incorrect). In CMC items a number of subtasks with two response options each (true / false) are presented. Examples of the different response formats are given in Pohl & Carstensen (2012a).

3. Data

3.1 The design of the study

Overall, 14,486 students in starting cohort 4 took the computer literacy test. There were two testing groups which differ in the order of the tests they received. 7,213 subjects received the mathematics test first, then the computer literacy test, and at last the science test, while

7,273 subjects received the computer literacy test after completing the mathematics and science tests. The test time for the computer literacy test was 29 minutes, with one additional minute for the procedural metacognition item. There was no multi-matrix design regarding the choice and order of the items within a test. All students got the same test items in the same order.

The computer literacy test in grade nine consists of 40 items which represent the knowledge and skills needed for a problem-oriented use of modern information and communication technology (for more information see the NEPS website)[1]. Four items were excluded from the analysis. One CMC item was excluded because a small but considerable number of subjects showed a response pattern indicating a misunderstanding of the item stimulus. Another Item had to be excluded because of a poor item fit index. Two further items had to be excluded because they were framed in a program we no longer consider essential for the construct. One of these items also showed a point-biserial correlation of the correct answer below .20. For the other item, two partial items out of six were excluded because of a negative point-biserial correlation with the correct answer.

The characteristics of the remaining 36 items are depicted in table 1, on process components, table 2, on software applications, and table 3, on response formats. The number of subtasks of CMC items varies between four and seven.

Table 1: Distribution of the number of test items by process components in the computer literacy test grade 9

Process components	**Frequency**
Access	6
Create	11
Manage	9
Evaluate	10
Total number of items	36

Table 2: Distribution of the number of the test items by software applications in the computer literacy test grade 9

Software applications	**Frequency**
Word processing	5
Spreadsheet	8

[1] https://www.neps-data.de/

Presentation software	4
E-Mail / Communication tools	6
Internet / search engines	13
Total number of items	36

Table 3: Response formats of the items in the computer literacy test grade 9

Response format	**Frequency**
Simple multiplechoice	29
Complex multiple choice	7
Total number of items	36

3.2 Sample

The description of the sample, the sampling procedure as well as information on the implementation as well as a description of the design of the study and the competence measures used can be found at the NEPS website[2].

14,486 persons took the computer literacy test. None of the cases had less than three valid responses to the test items, so that no case had to be excluded from further analyses. The results of the 14,312 subjects (loss of 174 subjects in the analyses due to delayed written consent by the parents; see also Introduction) are presented in the following sections.

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) due to items that are not reached, d) 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 are, for example, ticking two response options in simple MC items where just one is required. Missing responses due to omitted items occur when persons skip some items. Due to time limits, it may happen that not every person finishes the test within time. As a consequence, missing responses due to items that are not reached result.

[2] https://www.neps-data.de/

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 on how well the persons got along with the test. We then looked at the occurrence of missing responses per item, in order to get some information on how well the items worked.

4.2 Scaling model

For estimating item and person parameters for reading competence, a partial credit model was used and estimated in ConQuest (Wu, Adams, & Wilson, 1997). A detailed description of the scaling model can be found in Pohl and Carstensen (2012a).

CMC items consist of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly responded 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, in order to avoid possible estimation problems, the categories were collapsed. 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 to one category. For all seven CMC items the lowest two categories were collapsed.

In the following analyses, 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 1 for the correct response (see Haberkorn, Pohl, Carstensen, & Wiegand, 2012; and Pohl & Carstensen, 2012, for studies on the scoring of different response formats). A special case is item icg9140s_c. The item consists of six subtasks, but two subtasks showed only a weak point-biserial correlation with the total score and were excluded from further analysis. As a consequence, only four remaining subtasks were analyzed.

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, Warm, 1989) and later also in form of plausible values (Mislev, 1991). Person parameter estimation in NEPS is described in Pohl & Carstensen (2012a), while 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 checked in several analyses.

The responses on the subtasks of CMC items were aggregated to a polytomous variable for each CMC. In order to justify such an aggregation, the fit of the single subtasks was checked in analyses. For this, the single subtasks were separately included in a Rasch model together with the MC items and the fit of the subtasks was evaluated based on the weighted mean square error (WMNSQ), the respective t-value, point-biserial correlations of the responses with total correct score and the item characteristic curve. Only if the subtasks had a satisfactory item fit, they were used to construct polytomous CMC item variables.

In MC and CMC items there are a number of distractors (incorrect response options). We investigated if the distractors worked well, that is, if they are more often chosen by the students with a low ability than by students with a high ability. For this we evaluated the point-biserial correlation of giving a certain incorrect response and the total score in an analysis treating all subtasks of CMC items as single items. We judged correlations below zero as very good, correlations below 0.05 as acceptable and correlations above 0.05 as problematic.

Item fit was then evaluated for the MC items and the polytomous CMC items based on results of a partial credit model. Again the weighted mean square error (WMNSQ), the respective t-value, correlations of the item score with the total score (equal to the discrimination value as computed in ConQuest) and the item characteristic curve were evaluated for each item. Items with a WMNSQ > 1.15 (t-value > 6) were considered having a noticeable misfit and items with a WMNSQ > 1.2 (t-value > 8) were judged having a considerable misfit and their performance was further investigated. Correlations of the item score with the total score greater than 0.3 were considered as good, greater than 0.2 as acceptable and below 0.2 as problematic. Overall judgment of the fit of an item was based on all fit indicators.

We aim at constructing a computer literacy test that measures the same construct for all students. If there are items that favor certain subgroups (e.g., that are 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. Test fairness was investigated for the variables test position, gender, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl and Carstensen, 2012a, for a description of these variables). In order to test for measurement invariance, differential item functioning (DIF) is estimated using a multi-group IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty are estimated. Differences in the estimated item difficulties between the subgroups are evaluated. Based on experiences with preliminary data, we consider absolute differences in estimated difficulties that are greater than 1 logit as very strong DIF, absolute differences between .6 and 1 noteworthy to further investigate, and differences smaller than .6 as not considerable DIF. Additionally model fit was investigated by comparing a model including differential item functioning to a model that only includes main effects and no DIF.

The competence data in NEPS were scaled using the partial credit model (1PL), in which Rasch-homogeneity is assumed. The partial credit model was chosen because it preserves the weighting of the different aspects of the framework intended by the test developers (Pohl & Carstensen, 2012a). Nevertheless, Rasch's assumption of equal item discrimination was tested. Thus, the data were analyzed with a generalized partial credit model (2PL) (Muraki, 1992) using the software mdltm (von Davier, 2005), and the deviations of the estimated discrimination parameters from a uniform discrimination were evaluated. . Moreover, the model fit indices of the 2PL model were compared with those of the partial credit model. The computer literacy test is constructed to measure computer literacy on a unidimensional scale (Senkbeil et al., under review). The assumption of unidimensionality was, nevertheless, tested in 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 five different subdimensions based on different software applications was fitted to the data. The correlation between the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the scale.

5. Results

5.2 Missing responses

5.1.1 Missing responses per person

The number of invalid responses per person is shown in Figure 2. This number is very small. 94.5 % of persons did not give any invalid response. Only 1.3% of subjects have more than one invalid response.

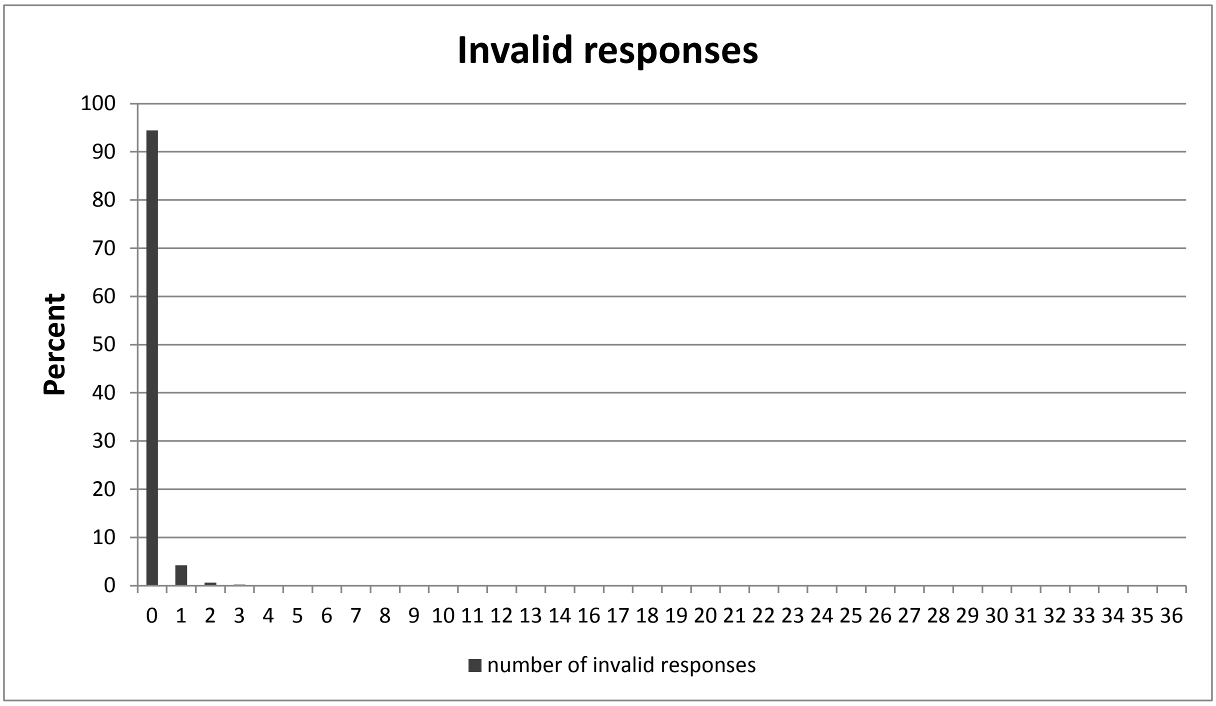

Figure 2: Number of invalid responses

Missing responses may occur when persons skip (omit) some items. The number of omitted responses per person is depicted in Figure 3. The figure shows that there is some tendency to omit items. 50 percent of the subjects omitted no item at all. Six percent of the subjects omitted more than 5 items.

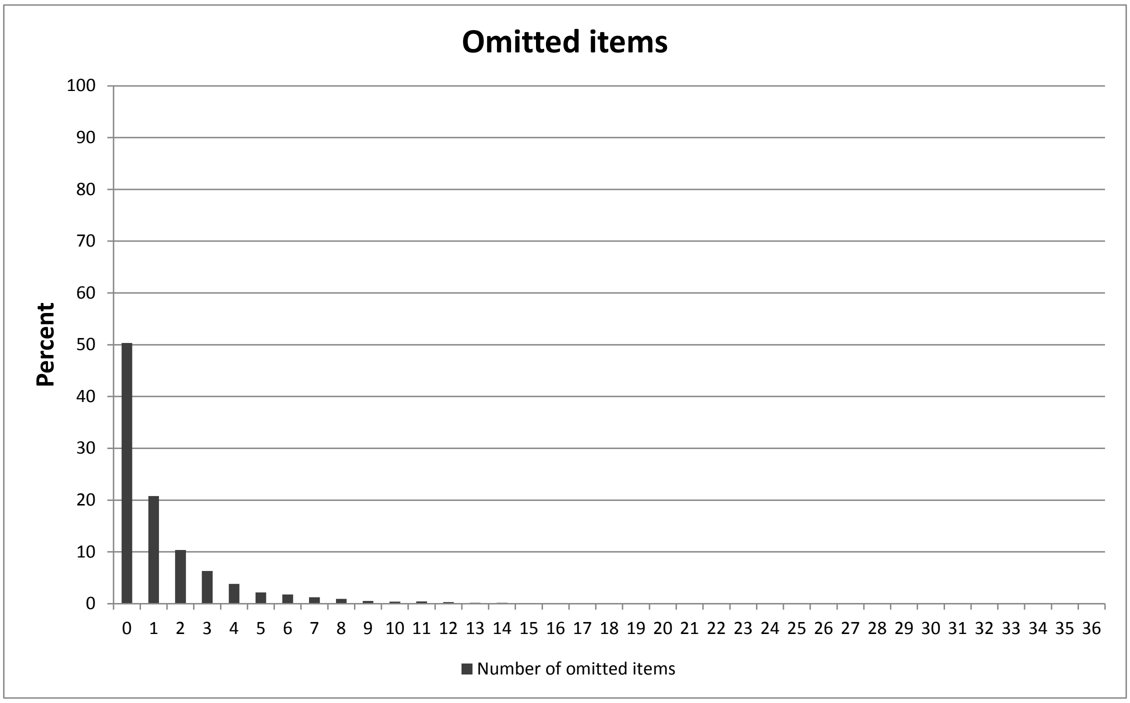

Figure 3: Number of omitted items

Due to time limits, not all persons reached the end of the test within the given time. Items are considered to be not reached if they are omitted and stand after the last response given in a test. Figure 4 shows the number of items that were not reached per person. The number of items that were not reached is rather high. Only 63.0% of the subjects reached the end of the test. 37.0% of the subjects did not reach the last item. The last five items were completed by only 80% of all subjects.

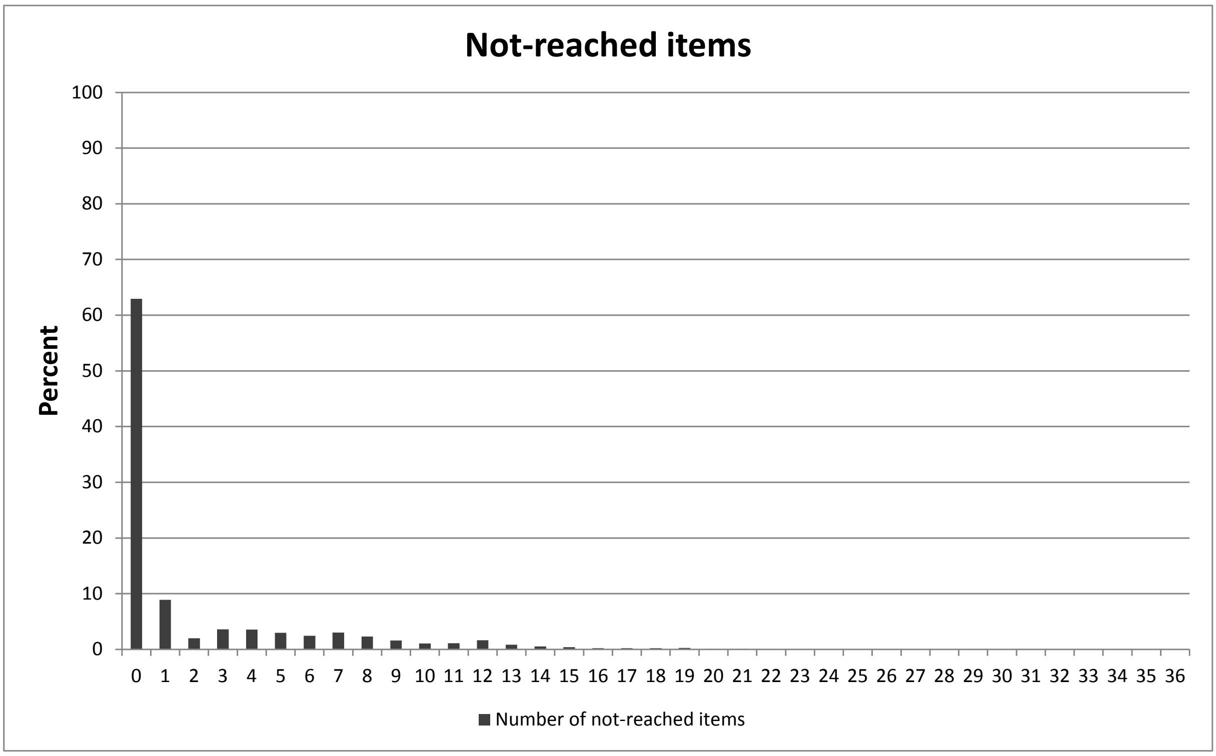

Figure 4: Number of not reached items

Figure 5 shows the total number of missing responses per person. The total number of missing responses is the sum of invalid, omitted, not reached, and not-determinable missing responses. Figure 5 shows that only one third of the subjects (32.7%) showed no missing response at all. Three quarters of the subjects (74.3%) had five missing values or less and only 1.0% of the subjects had missing responses for more than half of the items.

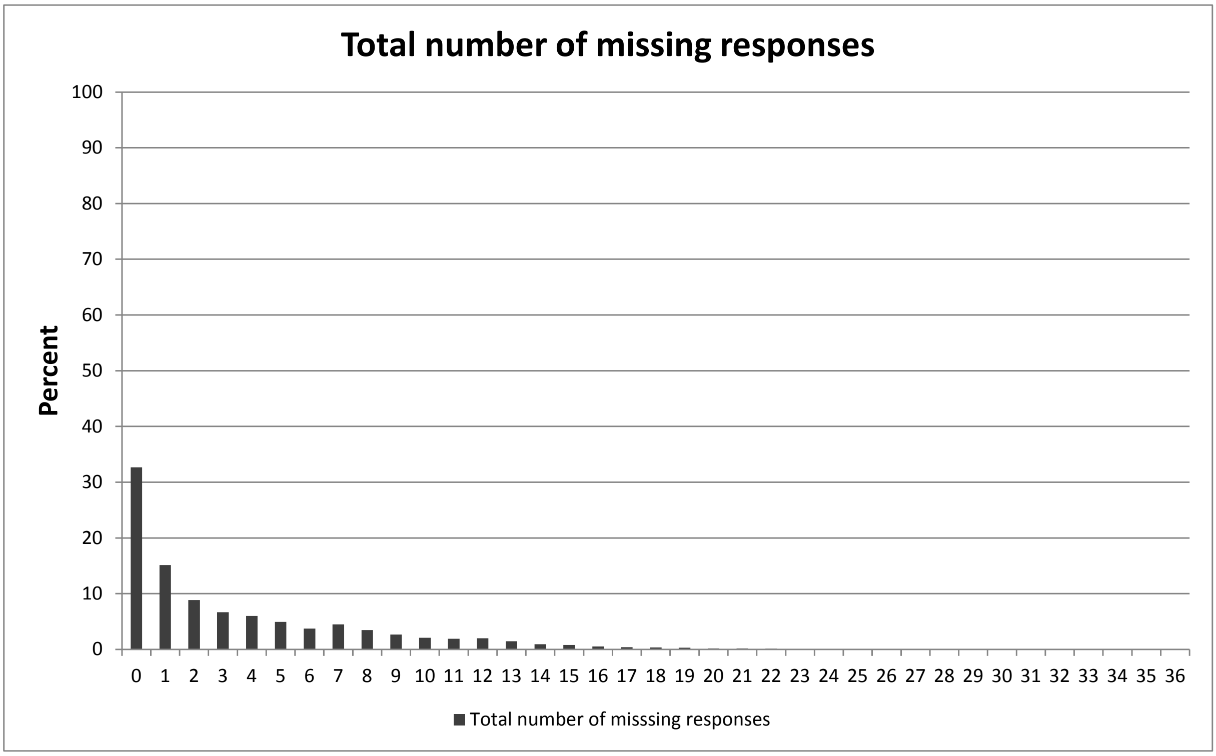

Figure 5: Total number of missing responses

Overall, there is a small amount of invalid responses and a reasonable amount of omitted items. The number of not reached items is, however, rather large and, thus, also the total number of missing responses.5.1.2 Missing responses per item

Table 4 shows the number of valid responses for each item, as well as the percentage of missing responses (total number, invalid responses, omitted responses, and not-reached responses). The number of invalid responses per item is small. The highest number is 0.55% for item icg9133s_c. The reason for invalid responses on that item is probably again due to a misunderstanding of the instruction for the CMC items. Overall, the number of persons that omit an item is acceptable. There are ten items with an omission rate above 5%. The highest omission rate occurs for item icg9128x_c (16.24% of the persons omitted this item). The number of missing responses is correlated to .32 with the difficulty of the item. This result indicates that the test takers tend to omit items that are more difficult. It is noticeable that CMC items are omitted nearly twice as often (6.1%) than MC items (3.2%). The number of persons that did not reach an item increases with the position of the item in the test to up to 37.05%. This is a rather large amount. The total number of missing responses per item varies between 0.58% (item icg9106x_c) and 39.32% (item icg9140s_c).

5.2 Parameter estimates

5.2.1 Item parameters

The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are depicted in table 4. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties (or location parameters for polytomous variables) vary between -2.05 (item icg9124x_c) and

1.56 (item icg9129x_c) with a mean of -0.61. The mean probability for solving an item was .60, indicating a good fit between item difficulties and person abilities (see Figure 6). Overall, the item difficulties are a little bit low, and there are only a few items with a high difficulty. Due to the large sample size, the standard error of the estimated item difficulties is very small (SE(ß) ≤ 0.03). The step parameters for CMC items are depicted in table 5.

5.2.2 Person parameters

Person parameters are estimated as WLEs and PVs (Pohl & Carstensen, 2012a). WLEs are provided in the first release of the SUF. PVs will be provided in later releases. A description of the data in the SUF can be found in section 7. An overview of how to work with competence data can be found in Pohl and Carstensen (2012a).

5.2.3 Test targeting and reliability

Test targeting was investigated in order to evaluate the measurement precision of the estimated ability scores and to judge the appropriateness of the test for the specific target population. In the analyses, the mean of ability is constrained to be zero. The variance was estimated to be 0.75, indicating that the test differentiates well between subjects. The reliability of the test (EAP/PV reliability =.83, WLE reliability = .81) is good.

The amount to which the item difficulties and location parameters are targeted to the ability of the persons is shown in Figure 6. The Figure shows that the items cover a great range of the ability distribution of the persons. However, only few items cover a very high degree of ability. There is a large number of easy items. As a consequence, subjects with a medium and low ability are measured relatively precisely while subjects with a high reading ability have a larger standard error.

Table 4: Item parameters

Item	Position in the test	# valid responses	Relative frequency of not-reached missings in %	Relative frequency of omitted missings in %	Relative frequency of missings due to invalid responses in %	Difficulty/ location parameter	SE (difficulty)	WMNSQ	t-value of WMNSQ	Correlation of item score with total score	Discrimination (2 PL)
icg9101x_c	1	14070	0.00	1.62	0.07	0.82	0.02	1.04	4.90	0.34	0.86
icg9102s_c	2	13372	0.00	6.44	0.13	-1.64	0.02	1.03	2.90	0.37	0.50
icg9103x_c	3	13707	0.00	4.04	0.19	0.92	0.02	0.96	-5.40	0.43	1.35
icg9104x_c	4	14119	0.00	0.93	0.42	1.12	0.02	1.10	10.80	0.22	0.49
icg9105x_c	5	14205	0.00	0.46	0.29	-1.17	0.02	0.90	-10.90	0.52	1.92
icg9106x_c	6	14230	0.00	0.29	0.29	-0.47	0.02	0.91	-14.30	0.52	1.76
icg9107s_c	7	13504	0.01	5.37	0.26	-1.86	0.02	0.99	-0.50	0.42	0.59
icg9109x_c	9	14191	0.01	0.47	0.36	-1.60	0.02	1.02	1.60	0.31	0.96
icg9110x_c	10	14039	0.01	1.57	0.07	0.33	0.02	1.02	3.00	0.38	0.97
icg9111x_c	11	12499	0.01	12.58	0.13	0.42	0.02	1.05	5.60	0.36	0.88
icg9112x_c	12	14140	0.02	0.88	0.32	-1.00	0.02	1.08	11.00	0.33	0.83
icg9113x_c	13	14215	0.03	0.55	0.08	0.68	0.02	1.01	1.40	0.28	0.64
icg9114x_c	14	14106	0.04	0.94	0.30	-0.79	0.02	0.99	-0.80	0.39	1.03
icg9116x_c	16	13918	0.07	2.21	0.10	-1.59	0.02	0.97	-2.50	0.36	1.15
icg9117s_c	17	13703	0.10	3.84	0.45	-1.55	0.02	0.98	-2.70	0.53	0.72
icg9118x_c	18	14174	0.19	0.54	0.47	-0.23	0.02	1.05	5.60	0.43	1.19
icg9119x_c	19	13974	0.27	1.83	0.31	-0.87	0.02	1.03	3.50	0.36	0.97
icg9121x_c	21	14038	0.51	1.22	0.24	-0.57	0.02	1.00	-0.10	0.41	1.11
icg9122x_c	22	13895	0.72	1.77	0.26	0.03	0.02	0.92	-13.00	0.50	1.59
icg9123x_c	23	13964	0.93	1.31	0.19	-0.85	0.02	1.01	0.80	0.39	1.05
icg9124x_c	24	13971	1.16	0.99	0.42	-2.05	0.03	0.99	-0.90	0.33	1.21
icg9125s_c	25	13392	1.54	4.67	0.19	-1.73	0.02	1.01	1.30	0.35	0.51

icg9126x_c	26	13838	2.08	1.08	0.23	-0.56	0.02	0.99	-1.10	0.42	1.14
icg9127x_c	27	13076	2.91	5.66	0.22	-0.71	0.02	0.93	-8.60	0.49	1.52
icg9128x_c	28	11320	4.54	16.24	0.13	0.07	0.02	1.09	11.90	0.31	0.68
icg9129x_c	29	12913	5.66	4.01	0.10	1.56	0.02	1.07	5.20	0.23	0.60
icg9130x_c	30	12984	6.70	2.53	0.05	-0.23	0.02	1.02	2.30	0.39	1.02
icg9131x_c	31	12857	8.31	1.72	0.13	-0.44	0.02	0.97	-4.50	0.45	1.29
icg9132x_c	32	12345	10.62	2.90	0.22	-0.77	0.02	1.00	-0.20	0.40	1.11
icg9133s_c	33	10800	13.62	10.37	0.55	-1.29	0.02	1.01	0.40	0.55	0.57
icg9134x_c	34	11441	16.07	3.81	0.17	-1.63	0.03	1.02	1.10	0.32	0.97
icg9135x_c	35	10634	19.05	6.53	0.11	-0.06	0.02	0.96	-6.20	0.47	1.32
icg9136s_c	36	9604	22.62	9.87	0.41	-0.84	0.02	1.00	0.40	0.57	0.62
icg9137x_c	37	9577	26.19	6.53	0.36	-0.62	0.02	0.92	-8.70	0.50	1.52
icg9138x_c	38	9144	28.17	7.73	0.20	-0.97	0.02	1.09	8.00	0.28	0.67
icg9140s_c	40	8685	37.05	1.97	0.30	-1.72	0.03	0.94	-3.20	0.45	0.68

Table 5: Step parameters for CMC items

Item	Step 1 (SE)	Step 2 (SE)	Step 3 (SE)	Step 4(SE)	Step 5 (SE)	Step 6 (SE)
icg9102s_c	-1.345(0.018)	1.056 (0.021)	0.289			
icg9107s_c	-1.083 (0.019)	-0.366 (0.018)	0.432 (0.018)	1.016		
icg9117s_c	-0.553 (0.018)	-0.099 (0.018)	0.191 (0.017)	-0.394 (0.018)	0.855	
icg9125s_c	-1.131 (0.018)	0.366 (0.018)	0.765			
icg9133s_c	-0.338 (0.020)	-0.694 (0.020)	0.965 (0.022)	0.385 (0.027)	-0.318	
icg9136s_c	-1.291 (0.028)	-0.486 (0.024)	0.081 (0.022)	0.269 (0.021)	0.078 (0.023)	1.350
Icg9140s_c	1.031 (0.026)	-0.594 (0.028)	-0.437			

scale in logits	Person ability	item difficulty
2	X	
	X	
	X	
	XX	
	XX	26
	XX	
	XXX	
	XXXX	
	XXXXX	4
1	XXXXX	
	XXXXXXXX	1 3
	XXXXXX	
	XXXXXXX	12
	XXXXXXX	
	XXXXXXXXX	10
	XXXXXXXXX	9
	XXXXXXXXXX	
0	XXXXXXXX	19 25
	XXXXXXXXX	32
	XXXXXXXX	27
	XXXXXXXXX	16
	XXXXXXXXX	28
	XXXXXXXXX	6 18 23
	XXXXXXXX	34
	XXXXXXX	13 24 29
	XXXX	17 20 33
-1	XXXXX	11 35
	XXXXX	
	XXXX	5
	XXX	30
	XXX	
	XX	8 14 15
	X	2 31 36
	X	22
	X	7
-2	X	21

Figure 6: Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 80.5 cases. Item difficulty is depicted on the right side of the graph. Each number represents one item (see table 4)

5.3 Quality of the test

5.3.1 Fit of the subtasks of complex multiplechoice items

Before the responses on the subtasks of CMC items were aggregated and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the single subtasks together with the simple MC items in a Rasch model. Counting the subtasks of CMC items separately, there are 75 items. Only one subtask had a probability for a correct response of higher than 90%. No estimation problems occurred. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.89 to 1.20, the respective t-value from -15.2 to 19.1, and there were no noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. Due to the good model fit of the subtasks their aggregation to polytomous variables 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 pointbiserial correlation between each incorrect response (distractor) and the students' total score. All but four distractors (from 0.00 to 0.05) had a pointbiserial correlation with ability below zero (Median = -.19). The results indicate a good model fit.

5.3.3 Item fit

The item fit is very good. WMNSQ is close to 1 with the lowest value being 0.92 (item icg9122x_c) and the highest being 1.10 (item icg9104x_c). There are only four items with a WMNSQ above 1.07 and a respective t-value above 7. The correlation of the itemscore with the total score varies between .22 (for item icg9104x_c) and .57 (for item icg9136s_c) with an average correlation of .40. Most of the items (27 out of 36 items) had a correlation with the total score between .30 und .50. All item characteristic curves showed a good fit of the items. The mean probability for solving an item was .60, indicating a good targeting of item difficulties and person abilities.

5.3.4 Differential item functioning

We checked for test fairness for different groups (i.e., measurement invariance) by estimating the amount of differential item functioning (DIF). Differential item functioning was investigated for the variables test position, gender, the number of books at home (as a proxy for socio-economic status), migration background, and school type (see Pohl & Carstensen, 2012a, for a description of these variables). Table 7 shows the difference between the estimated item difficulties in different groups. Male vs. female, for example, indicates the difference in difficulty *ß*(male) – *ß*(female). A positive value indicates a higher difficulty for males, a negative value a lower difficulty for males as opposed to females.

The computer literacy test was administered in two different positions (see section 3.1 for the design of the study). 7,126 (49.8%) persons received the computer literacy test after the mathematics test, but before the science test (position 2), and 7,186 (50.2%) of the persons received the computer literacy test after having completed the mathematics and science test (position 3). The subjects were randomly assigned to either of the two design groups. Differential item functioning of the position of the test may, for example, occur if there are differential fatigue effects for certain items. The results show a small average effect of item position. Subjects who received the computer literacy test before the science test (position 2) perform on average 0.106 logits (Cohen's d = 0.123) better than subjects who received the computer literacy test after the science test[3]. There is no DIF due to the position of the test in the booklet. The highest difference in difficulty between the two design groups is 0.312 logits.

DIF was also investigated for gender. 7,140 (49.9%) of the test takers were female and 7,166 (50.6%) were male. There were 6 missing responses on the variable gender. These cases were excluded from the analysis. On average, male students have a slightly lower computer literacy than female students (main effect = -0.022 logits, Cohen's d = 0.026). There is no item with a considerable gender DIF. The highest difference in difficulties between the two groups is 0.410 logits.

The number of books at home was used as a proxy for socio-economic status. There were 6,017 (42.0%) test takers with 0 to 100 books at home, 7,973 (55.7%) test takers with more than 100 books at home, and 322 (2.3%) test takers without a valid response. DIF was investigated using these three groups. There are considerable average differences between the three groups. Participants with 100 or less books at home perform on average 0.630 logits (Cohen's d = 0.731) lower in reading than participants with more than 100 books. Participants without a valid response on the variable 'books at home' performed 0.270 logits (Cohen's d = 0.313) or 0.895 logits (Cohen's d = 1.038) worse than participants with up to 100 and more than 100 books, respectively. There is no considerable DIF comparing participants with many or fewer books (highest DIF = 0.522). Comparing the group without valid responses to the two groups with valid responses, DIF occurs up to 0.555 logits. This is a rather large difference, which may, however, also be the result of the uncertainty in estimation due to the small number of persons with missing responses.

There were 10,021 (70.0%) participants without a migration background, 3,168 (22.1%) participants with a migration background, and 1,075 (7.5%) students could not be allocated to either group. 48 (0.4%) students were excluded from the analyses due to missing or invalid responses. The first three groups were used for investigating DIF of migration. There is a considerable difference in the average performance of participants with or without migration background (main effect = 0.395 logits, Cohen's d = 0.458). Participants without a migration background have a higher computer literacy than participants with a migration background. Also subjects with unknown background on migration differ from those without a migration background (main effect = -0.539 logits, Cohen's d = -0.625), they do not differ much from subjects with a migration background (main effect = -0.140 logits, Cohen's d = -0.162). There is no considerable DIF. The highest difference in difficulties between groups is 0.476 logits.

There were 10,021 (70.0%) students born in Germany, 2,976 (20.8%) students from other countries and 1,267 (8.8%) students who could not be allocated to either group. Again 48 (0.4%) students were excluded from the analyses due to missing or invalid responses. There is a considerable difference in the average performance of students from Germany and students from other countries (main effect = 0.387 logits, Cohen's d = 0.449). Students from

[3] Note that this main effect does not indicate a threat to measurement invariance. Instead, it may be an indication of fatigue effects that are similar for all items.

Germany have a higher computer literacy than students from other countries. Also subjects with unknown country of origin differ from students born in Germany (main effect = -0.539 logits, Cohen's d = -0.625), they do not differ much from subjects from other countries (main effect = -0.140 logits, Cohen's d = -0.162). There is no considerable DIF. The highest difference in difficulties between groups is 0.473 logits.

DIF was also investigated for school type. 5,064 (35.4%) of the test takers were high-school students and 9,248 (64.6%) were non high-school students. On average, high-school students have a higher computer literacy than non high-school students (main effect = 1.002 logits, Cohen's d = 1.162). There is no considerable DIF. The highest difference in difficulties between the two groups is 0.368 logits.

Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF with those that allow only for main effects. In Table 6, the models including only main effects are compared with those that additionally estimate DIF. The Akaike's (1974) information criterion (AIC) favors the models estimating DIF for all four DIF variables. The Bayesian information criterion (BIC, Schwarz, 1978) takes the number of estimated parameters into account and, thus, prevents from overparameterization of models. Using BIC, the more parsimonious model including only the main effect is preferred over the more complex DIF model for the most DIF variables (position, books, migration background and country of origin). Only for the DIF variables gender and school type the more complex DIF model have slightly better information criterions.

Table 6: Comparison of models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Position	main effect	140480.152	45	140570.152	140865.130
	DIF	140351.466	77	140505.466	141010.206
Gender	main effect	140411.338	45	140501.338	140796.299
	DIF	140257.848	77	140411.848	140916.559
Books	main effect	140029.608	46	140121.608	140423.141
	DIF	139769.618	110	139989.618	140710.675
Migration	main effect	140280.106	46	140372.106	140673.639
	DIF	140108.559	110	140328.559	141049.617
Country of origin	main effect	140280.106	46	140372.106	140673.639
	DIF	140108.559	110	140328.559	141049.617
School type	main effect	140280.106	46	140372.106	140673.639
	DIF	140108.559	110	140328.559	141049.617

Most of the differences in item difficulties estimated via the DIF-analyses are in absolute values below 0.5. Only four items showed DIF values above the threshold of 0.5: the items icg9104x_c (books), icg9105x_c (books, school type), icg9137x_c (school type), and icg9138x_c (school type). With one exception (icg9105x_c: -.736 in school type), these values are all only scarcely above the threshold. Overall, the results indicate that there is no considerable DIF and the test is fair for the considered groups.

Table 7: Differential item functioning (absolute differences between difficulties)

Item	Booklet	Gender	Books			Immigration background			Country of origin			School type
	Position 1 vs. 2	Male vs. Female	(< 100) vs. (> 100)	(< 100) vs. missing	(>100) vs. missing	Without vs. with	Without vs. missing	With vs. missing	Germany vs. other country	Germany vs. unknown	Other country vs. unknown	High school vs. non high school
icg9101x_c	0.312	0.180	-0.073	0.034	0.107	0.058	0.064	0.006	0.012	0.168	0.156	-0.002
icg9102s_c	0.152	-0.292	-0.040	0.208	0.248	0.053	0.076	0.023	0.044	0.088	0.044	0.162
icg9103x_c	0.152	-0.104	0.144	0.048	-0.096	-0.005	-0.025	-0.020	-0.001	-0.041	-0.040	-0.058
icg9104x_c	0.224	0.040	-0.216	0.339	0.555	0.259	0.350	0.091	0.252	0.347	0.095	0.398
icg9105x_c	0.142	0.150	0.522	-0.012	-0.534	-0.476	-0.367	0.109	-0.473	-0.396	0.077	-0.736
icg9106x_c	-0.068	-0.282	0.177	-0.126	-0.303	0.082	-0.085	-0.167	0.075	-0.054	-0.129	-0.358
icg9107s_c	0.070	-0.214	0.069	0.373	0.304	-0.001	0.028	0.029	0.010	-0.007	-0.017	-0.044
icg9109x_c	-0.128	0.182	-0.014	-0.074	-0.060	-0.039	-0.015	0.024	-0.041	-0.022	0.019	0.210
icg9110x_c	0.006	-0.210	0.040	0.044	0.004	-0.108	-0.036	0.072	-0.098	-0.076	0.022	0.138
icg9111x_c	-0.210	-0.130	-0.160	0.030	0.190	0.101	0.250	0.149	0.114	0.192	0.078	0.220
icg9112x_c	0.114	0.278	0.035	0.257	0.222	-0.055	0.136	0.191	-0.036	0.054	0.090	-0.030
icg9113x_c	-0.030	0.288	-0.198	0.163	0.361	0.068	0.163	0.095	0.090	0.087	-0.003	0.440
icg9114x_c	-0.028	0.038	0.058	0.026	-0.032	-0.242	-0.121	0.121	-0.233	-0.163	0.070	0.100
icg9116x_c	0.032	0.194	0.237	0.193	-0.044	-0.315	-0.120	0.195	-0.327	-0.126	0.201	-0.228
icg9117s_c	0.014	-0.232	0.033	0.061	0.028	0.069	-0.008	-0.077	0.065	0.004	-0.061	-0.166
icg9118x_c	-0.058	-0.148	0.153	0.100	-0.053	-0.069	-0.018	0.051	-0.057	-0.060	-0.003	-0.210
icg9119x_c	-0.042	-0.104	-0.017	0.272	0.289	0.126	0.150	0.024	0.119	0.157	0.038	0.092
icg9121x_c	-0.046	0.214	0.033	-0.114	-0.147	-0.036	0.041	0.077	-0.042	0.039	0.081	0.018
icg9122x_c	-0.120	-0.006	0.135	0.030	-0.105	0.080	-0.055	-0.135	0.077	-0.033	-0.110	-0.256
icg9123x_c	-0.158	0.058	0.029	-0.092	-0.121	-0.057	-0.162	-0.105	-0.080	-0.097	-0.017	0.048
icg9124x_c	0.020	0.064	0.055	0.030	-0.025	-0.124	-0.093	0.031	-0.154	-0.037	0.117	0.012
icg9125s_c	0.024	-0.168	0.029	0.407	0.378	0.065	0.038	-0.027	0.054	0.060	0.006	0.120
icg9126x_c	0.014	-0.240	0.032	0.069	0.037	-0.009	-0.022	-0.013	-0.029	0.020	0.049	0.128
icg9127x_c	-0.096	-0.074	0.138	0.030	-0.108	-0.203	-0.126	0.077	-0.193	-0.167	0.026	-0.342

icg9128x_c	0.016	0.130	-0.241	-0.053	0.188	0.179	0.085	-0.094	0.186	0.075	-0.111	0.394
icg9129x_c	0.016	0.064	-0.073	0.184	0.257	0.235	0.170	-0.065	0.234	0.171	-0.063	0.350
icg9130x_c	0.010	-0.054	0.028	-0.115	-0.143	-0.094	-0.132	-0.038	-0.085	-0.152	-0.067	-0.158
icg9131x_c	0.056	-0.024	0.009	-0.090	-0.099	0.117	-0.051	-0.168	0.097	0.011	-0.086	-0.060
icg9132x_c	-0.060	-0.048	0.116	-0.009	-0.125	-0.122	0.033	0.155	-0.144	0.051	0.195	0.018
icg9133s_c	-0.124	0.074	-0.040	0.112	0.152	0.042	0.003	-0.039	0.045	-0.005	-0.050	0.024
icg9134x_c	-0.032	0.212	-0.065	-0.286	-0.221	-0.082	-0.275	-0.193	-0.105	-0.207	-0.102	0.012
icg9135x_c	-0.112	0.410	0.167	-0.089	-0.256	-0.157	-0.233	-0.076	-0.119	-0.317	-0.198	-0.350
icg9136s_c	-0.032	-0.036	-0.064	-0.068	-0.004	0.131	0.031	-0.100	0.131	0.043	-0.088	0.098
icg9137x_c	-0.038	0.344	0.227	-0.008	-0.235	-0.224	-0.162	0.062	-0.226	-0.174	0.052	-0.518
icg9138x_c	0.104	0.206	-0.194	0.178	0.372	0.181	-0.046	-0.227	0.165	0.015	-0.150	0.582
icg9140s_c	0.032	0.026	0.202	-0.076	-0.278	-0.263	-0.067	0.196	-0.246	-0.140	0.106	-0.246
Main effect	0.106	-0.022	-0.630	0.270	0.895	0.395	0.535	0.140	0.387	0.539	0.140	1.002

5.3.5 Rasch homogeneity

In order to test the assumption of Rasch-homogeneity, we also fitted a generalized partial credit model (2PL) model to the data. The estimated discrimination parameters are depicted in table 4. They range from 0.49 (item icg9104x_c) to 1.92 (icg9105x_c). Although the discriminations differ considerably among the items (from .49 to 1.76 for the MC items, and from .50 to .72 for each category of the CMC items), the partial credit model (1PL) model (BIC=654816, number of parameters=59) fits the data better than the 2PL model (BIC=654875, number of parameters=112). Since, also the theoretical aim was to construct a test that equally represents the different aspects of the framework (see Pohl & Carstensen, 2012a, 2012b, for a discussion of this issue), and thus the partial credit model was used to preserve the item weightings intended in the constructional framework.

5.3.6 Unidimensionality

The unidimensionality 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 (the categories word processing and presentation software were collapsed for dimensionality analyses due to the scarce number of items in both categories).

To estimate a mulitdimensional (MD) model based on the four process components, Gauss estimation in ConQuest (nodes = 15) was used. The variances and correlations of the three dimensions are shown in table 8. All four dimensions show a substantive variance with the highest discrimination between subjects for *Manage* and the lowest for *Create*. The correlations between the four dimensions are very high (between .933 and .957). The four process components do not measure different constructs but a unidimensional construct.

Table 8: Results of four-dimensional scaling (process components). Variance of the dimensions are depicted in the diagonal; correlations are displayed in the off-diagonal

	Dim 1	**Dim 2**	**Dim 3**	**Dim 4**
Access (6 Items)	.704			
Create (11 Items)	.951	.670		
Manage (9 Items)	.949	.957	1.016	
Evaluate (10 Items)	.933	.945	.955	.769

To estimate a four-dimensional model based on the different types of software applications Gauss estimation (nodes = 15) was used (see table 9). The results of the analyses are depicted in table 9. All four dimensions show a substantive variation. The correlations between the dimensions vary between .825 and .957. The lowest correlation is found between dimension 2 (spreadsheet) and dimension 3 (e-mail / communication tools). Apart from this exception, the correlations do not differ substantially from a perfect correlation, indicating that a unidimensional construct is measured with the test. Note that the amount of missing responses is rather high for spreadsheet-items (e.g., icg9111x_c, icg9127x_c, icg9128x_c, icg9133s_c) and that the quantity of items for e-mail is rather low (6 items). This may result in a reduced variation and, thus, in a decreased correlation.

Table 9: Results of four-dimensional scaling (software applications). Variance of the dimensions are depicted in the diagonal; correlations are given in the off-diagonal

	Dim 1	Dim 2	Dim 3	Dim 4
Word processing / presentation software (9 Items)	.715			
Spreadsheet (8 Items)	.935	.823		
E-Mail / communication tools (6 Items)	.906	.825	.604	
Internet / search engines (13 Items)	.957	.931	.909	.987

6. Discussion

The analyses in the previous sections aimed at providing information on the quality of the reading test in starting cohort 4 (grade 9) and at describing how the computer literacy score is estimated. The analyses we conducted and described here indicate good measurement properties for the instrument.

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 and MA items, as well as the aggregated polytomous CMC and MA 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 and investigating the tests' dimensionality.

The results indicate a very good fit of the data to the partial credit model: The item fit (WMNSQ) of all items are within the usually accepted interval from .85 to 1.15, the comparison of the partial credit model and the 2PL model favors the partial credit model, and the dimensionality analyses indicate that the unidimensional model describes the data best.

The distribution of item difficulties and the distribution of person parameters overlap to a great extent, with one limitation: There are only few items which are very difficult, leading to an increased standard error of estimation for persons with very high ability. The distractor analysis showed a satisfying result.

The analyses of missing data revealed that only few items were omitted (skipped) by test takers, and even less of the given responses were invalid. Only the proportion of items not reached was higher than expected. This may suggest that there were too many items in the test for the given test time of 29 minutes. However, this was accounted for by regarding the missing values as missing during scaling (instead of regarding them as wrong answers), leading to an unbiased ability estimation for each subject independently from the number of processed items (see Pohl & Carstensen, 2012b).

In summary, the scaling procedures show that the test is a reliable instrument with satisfying psychometric properties for assessing computer literacy.

7. Data in the Scientific Use File

There are 36 items in the data set that are either scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response, or scored as a polytomous variable (CMC) indicating the (partial) credit. The dichotomous variables are marked with a 'x_c' at the end of the variable name, the CMC and MA items are marked with a 's_c' at the end of the variable name. In the scaling model, each category of the polytomous CMC and MA items is scored with 0.5 points. Manifest scale scores are provided in form of WLE estimates (ic_wle) including the respective standard error (ic_wle_se). The ConQuest syntax for estimating the WLE scores from the items is provided in appendix A.

Note that, different from other competence tests in the scientific use file, the value of the polytomous variables indicate the number of correctly responded subtasks. Therefore, the aggregation of categories has to be done by the user of the data. It is recommended to collapse categories with less than N=200 in order to avoid estimation problems (see also section 4.2). We collapsed the two lowest categories for all seven CMC items for the estimation (see syntax in appendix). We advise the user of the data to do so as well.

Plausible values, that allow investigating latent relationships of competence scores with other variables, will be provided in later data releases. User interested in investigating latent relationships may alternatively either include the measurement model in their analyses or estimate plausible values themselves. A description of these approaches can be found in Pohl and Carstensen (2012a).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Haberkorn, K., Pohl, S., Carstensen, C., & Wiegand, E. (2012). *Incorporating different response formats in the IRT-scaling model for competence data*. Manuscript submitted for publication.

Haberkorn, K., Pohl, S., Hardt, K., & Wiegand, E. (2012). *Technical Report of Reading– Scaling Results of Starting Cohort 4 in Ninth Grade* (NEPS Working Paper No. 16). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

Pohl, S. & Carstensen, C. H. (2012a). NEPS technical report – Scaling the data of the competence tests. (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S. & Carstensen, C. H. (2012b). Scaling the competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. Manuscript submitted for publication.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). Technical Report of Reading– Scaling Results of Starting Cohort 3 in Fifth Grade (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

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

Senkbeil, M., Ihme, J. M., & Wittwer, J. (under review). The test of technological and information literacy (TILT) in the National Educational Panel Study: Development, empirical testing, and evidence for validity. Manuscript submitted for publication.

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. v. Maurice & (Eds.). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, Sonderheft 14 .* Wiesbaden: VS Verlag für Sozialwissenschaften.

Wu, M.L., Adams, R. J., & Wilson, M.R. (1997). *ACER Conquest: Generalised item response modelling software*. Melbourne: ACER Press.

Appendix

Appendix A: ConQuest-Syntax for estimating WLE estimates in starting cohort 4, Grade 9 students (A47 / A68 / A84)

```
title ICT HE K9 scaling 36 items included, partial credit model;

datafile  >>filename.dat;
format pid 1-7 rotation 9 responses 10-45;

labels <<filename_with_labels.txt;

codes 0,1,2,3,4,5,6,7;

recode (0,1,2,3,4)       (0,0,1,2,3)            !item(2,22,36);
recode (0,1,2,3,4,5)     (0,0,1,2,3,4)          !item(7);
recode (0,1,2,3,4,5,6)   (0,0,1,2,3,4,5)        !item(15,30);
recode (0,1,2,3,4,5,6,7) (0,0,1,2,3,4,5,6)      !item(33);

score (0,1)              (0,1)                  !item(1,3-6,8-14,16-21,23-29,31-32,34-35);
score (0,1,2,3)          (0,.5,1,1.5)           !item(2,22,36);
score (0,1,2,3,4)        (0,.5,1,1.5,2)         !item(7);
score (0,1,2,3,4,5)      (0,.5,1,1.5,2,2.5)     !item(15,30);
score (0,1,2,3,4,5,6)    (0,.5,1,1.5,2,2.5,3)   !item(33);

set constraint=cases;
model item + item*step - rotation;
estimate;

show cases ! estimates=wle  >> filename.wle;
itanal  >> filename.itn;
show >> filename.shw;
```