NEPS WORKING PAPERS

Martin Senkbeil and Jan Marten Ihme

NEPS TECHNICAL REPORT FOR COMPUTER LITERACY: SCALING RESULTS OF STARTING COHORT 6–ADULTS

NEPS Working Paper No. 61
Bamberg, December 2015

LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

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

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).
The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS Consortium.

The NEPS Working Papers are available at
https://www.neps-data.de/projektübersicht/publikationen/nepsworkingpapers

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 6–Adults

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. (2015). NEPS Technical Report for Computer Literacy: Scaling results of Starting Cohort 6–Adults (NEPS Working Paper No. 61). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

We would like to thank Steffi Pohl and Kerstin Haberkorn for developing and providing standards for the technical reports and Luise Fischer for giving valuable feedback on previous drafts of this manuscript. We also thank Natalie Boonyaprasop for English proofreading.

NEPS Technical Report for Computer Literacy: Scaling Results of Starting Cohort 6 in Adults

Abstract

The National Educational Panel Study (NEPS) aims to investigate the development of competencies across the whole life span. Furthermore, NEPS develops tests for assessing the different competence domains. 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 6–Adults (Wave 4). Apart from descriptive statistics of the data, the scaling model applied to estimate competence scores, and the analyses performed to investigate the quality of the scale, as well as the results of these analyses are presented here. The computer literacy test in Starting Cohort 6 (Adults) consisted of 29 items representing different cognitive requirements and software applications. A multiple choice format was used. The test was administered to 6,923 adults. Of these, 6,138 adults completed the assessment in this domain. A Rasch 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 good reliability and the different comprehension requirements foster a unidimensional construct. In summary, the scaling procedures show that the test is a reliable instrument with satisfying psychometric properties for assessing computer literacy. In this 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 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. 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). Because 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 Starting Cohort 6 (Adults, Wave 4). We first introduce the main concepts of the computer literacy test. Then, we describe the computer literacy data of Starting Cohort 6 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 use in the Scientific Use File.

The present report has been modeled on the technical report by Senkbeil and Ihme (2012). Please 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 thus slightly differ from the data set used for the analyses in this paper. We do not, however, expect any major 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 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 (access, 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.

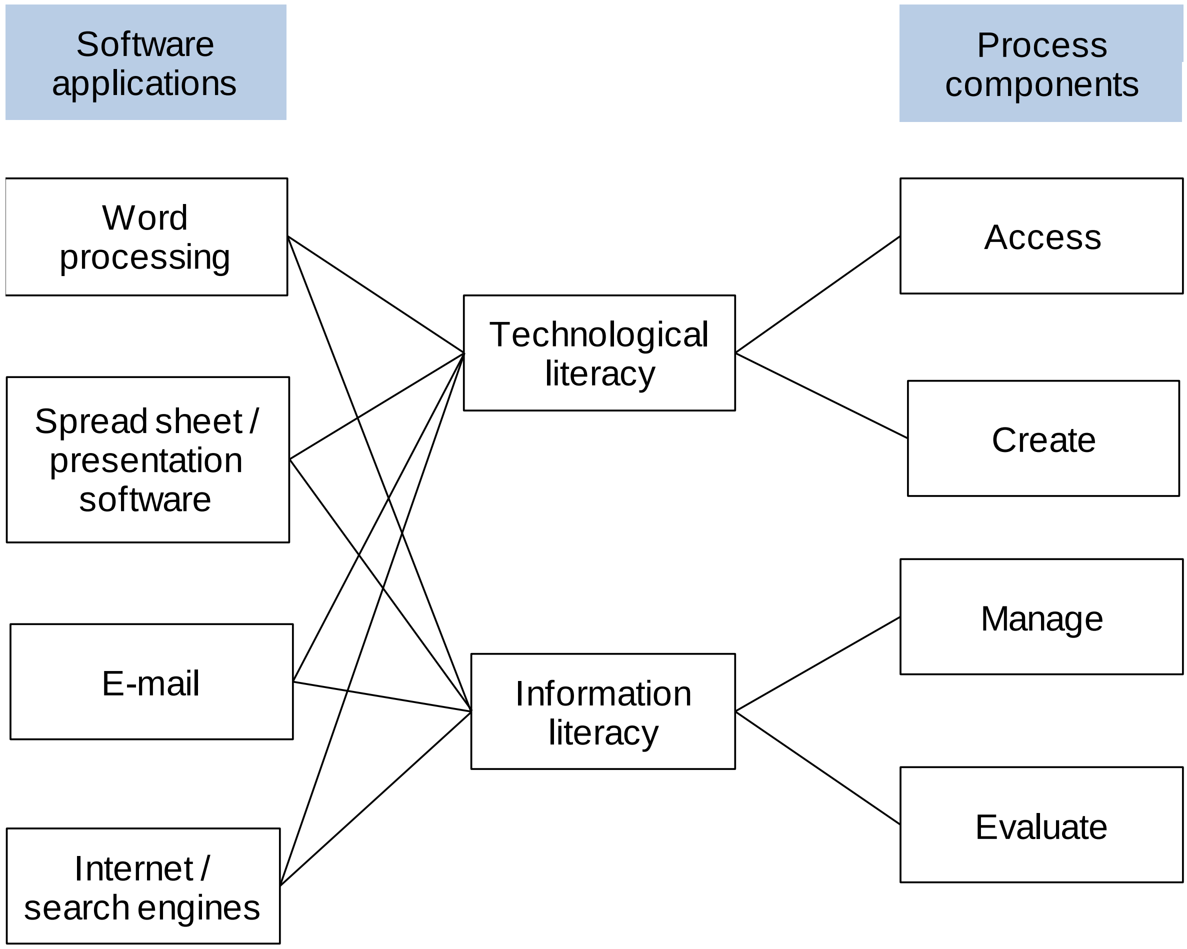

Figure 1. Assessment framework for computer literacy (process components and software applications).

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, an internet browser, an electronic database, or a spreadsheet as prompts (see Senkbeil et al., 2013).

In the computer literacy test of Starting Cohort 6 (Adults) there are two types of response formats. These are simple multiple choice (MC) and complex multiple choice (CMC) items. In MC items the test taker has to find the correct answer out of four to six response options with one option being correct and three to five response items functioning as distractors (i.e., they are incorrect). In CMC items a number of subtasks with two response options each (true / false) are presented. The number of subtasks of CMC items varies between four and eight. Examples of the different response formats are given in Pohl and Carstensen (2012a).

3. Data

3.1 The Design of the Study

Overall, the test was administered to 6,923 adults. Of these, 6,167 adults (89.1%) completed the assessment. Five hundred and twelve (7.4%) adults did not take part because they had no prior computer experience, and 244 (3.5%) adults refused to take the computer literacy test. There were two testing groups that differed in terms of the test order received. Overall, 3,100 subjects received the computer literacy test first, then the science test; while 3,038 subjects received the computer literacy test after completing the science test. The test time for the computer literacy test was 29 min, with one additional minute for the procedural metacognition item. There was no multimatrix design regarding the choice and order of the items within a test. All adults got the same test items in the same order.

The computer literacy test for adults consists of 30 items representing the knowledge and skills needed for a problem-oriented use of modern information and communication technology (for more information please refer to the NEPS website).1 One MC item (ica4048x_c) was eliminated from the analysis because of an insufficient item fit (weighted mean square > 1.20; see 4.3 for further explanation). For one CMC item (ica4015s_c), two partial items out of five were excluded because of a negative point-biserial correlation with the correct answer (see 4.3 for further explanation). The characteristics of the final set of 29 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 Adults

Process components	**Frequency**
Access	8
Create	7
Manage	9
Evaluate	5
Total number of items	29

Table 2

Software Applications of Items in the Computer Literacy Test Adults

Software applications	**Frequency**
Word processing	6
Spreadsheet / presentation software	10
E-mail / communication tools	4
Internet / search engines	9
Total number of items	29

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

Table 3

Response Formats of Items in the Computer Literacy Test Adults

Software applications	**Frequency**
Simple multiple choice	19
Complex multiple choice	10
Total number of items	29

3.2 Sample

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

In total, 6,167 persons took the computer literacy test. Twenty-nine of the cases provided less than three valid responses to the test items. Because no reliable computer literacy score may be estimated on the basis of such few responses, these cases were excluded from further analyses. The results of the remaining 6,138 test takers 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) 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 are, for example, ticking two response options in simple MC items where just one is required. Missing responses due to omitted items occur when a person skips some items. Due to time limits, it may occur that not every person will complete the test in time. As a consequence, missing responses due to items that are not reached may result from this.

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.

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

4.2 Scaling Model

To estimate item and person parameters for computer literacy competence, a Rasch 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 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 three CMC items (ica4015s_c, ica4020s_c, ica4050s_c) the lowest two categories were collapsed, and for six CMC items (ica4004s_c, ica4017s_c, ica4016s_c, ica4047s_c, ica4021s_c, ica4052s_c) the lowest three categories were collapsed and scored with 0 points. In the following analyses, a scoring of 0.5 points for each category of the polytomous items was applied; whereas simple MC items were scored dichotomously as 0 for an incorrect and as 1 for the correct response (see Haberkorn, Pohl, Carstensen, & Wiegand, 2012; and Pohl & Carstensen, 2012b, for studies on the scoring of different response formats).

A special case is item ica4018s_c. This item consists of eight subtasks. Whereas the lowest category (none of the subtasks answered correctly, scored as 0 points) had more than N = 200, some of the intermediate categories had less than N = 200. To avoid estimation problems, the intermediate categories were collapsed as follows: The categories "1, 2, or 3 correct answers" were collapsed (0.5 points), and the categories "4, 5, 6, or 7 correct answers" (1 point) were collapsed. The category "8 correct answers" (all subtasks answered correctly) was scored with 1.5 points.

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 (2012a), 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 reviewed by several analyses.

The responses to the subtasks of CMC items were aggregated to a polytomous variable for each CMC. In order to justify such an aggregation, the fit of individual subtasks was checked through several analyses. For this, individual subtasks were included separately in a Rasch model together with the MC items. The fit of the subtasks was evaluated based on the weighted mean square value (WMNSQ), the corresponding t-value, the point-biserial correlations of the responses with total correct score, and the item characteristic curve. Only if the subtasks had a satisfactory item fit were they 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 were chosen more often by 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 rated 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 the results of a partial credit model. Again, the weighted mean square value (WMNSQ), the corresponding *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 as 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 to construct a computer literacy test that measures the same construct in all adults. 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. Test fairness was investigated for the variables test position, gender, age, the number of books at home (as a proxy for socioeconomic status), duration of education ($\leq$ 12 years vs. > 12 years), 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) analysis is done using a multigroup 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 as noteworthy for further investigation, differences between 0.4 and 0.6 as considerable but not severe, and differences smaller than 0.4 as no considerable DIF. Additionally, model fit was investigated by comparing a model including DIF to a model that includes only main effects and no DIF.

The competence data in NEPS were scaled using the Rasch model (1PL). This 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). The deviations of the estimated discrimination parameters from a uniform discrimination were evaluated. 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 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

5.1 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. 94.3% of persons did not give any invalid response. Only 2.1% of subjects have more than one invalid response.

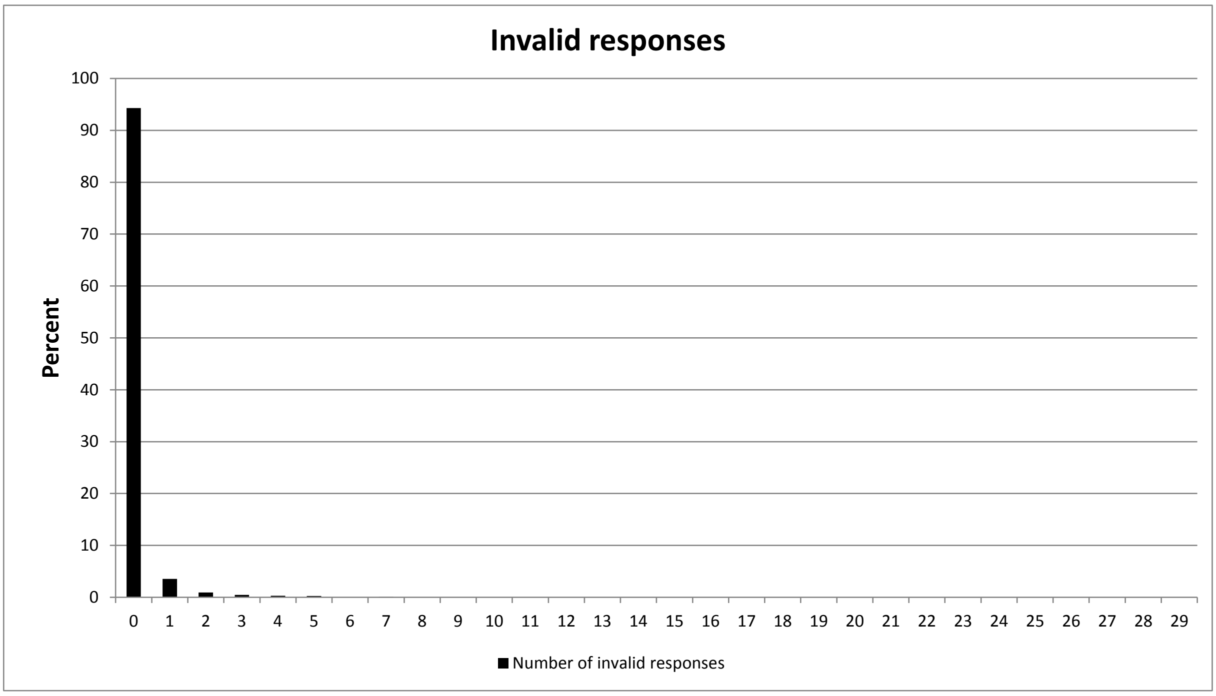

Figure 2. Number of invalid responses.

Missing responses may occur when people skip (omit) some items. The number of omitted responses per person is depicted in Figure 3. This figure shows that there is some tendency to omit items. Only 40% of the subjects omitted no item at all, and 21% of the subjects omitted more than 3 items.

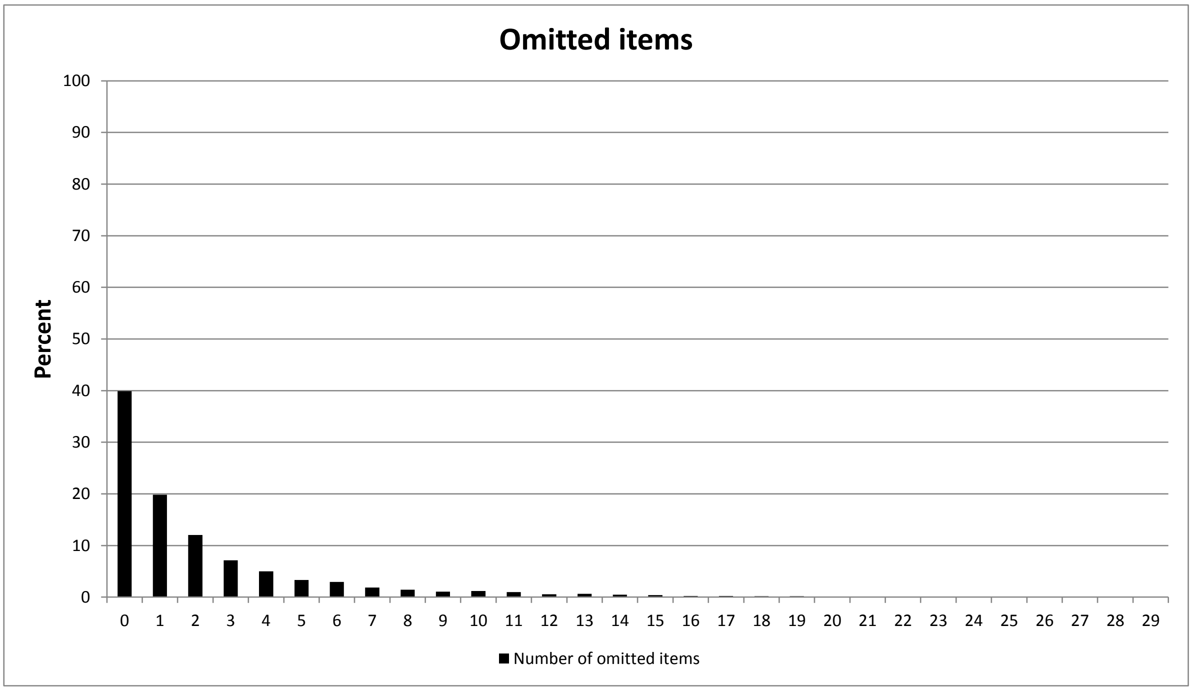

Figure 3. Number of omitted items.

Due to time limits, not all subjects reached the end of the test within the given time. Items are considered as not reached when they are omitted, standing in the position 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 large. Only 45% of the subjects reached the end of the test, and 43% of the subjects did not reach the last three items.

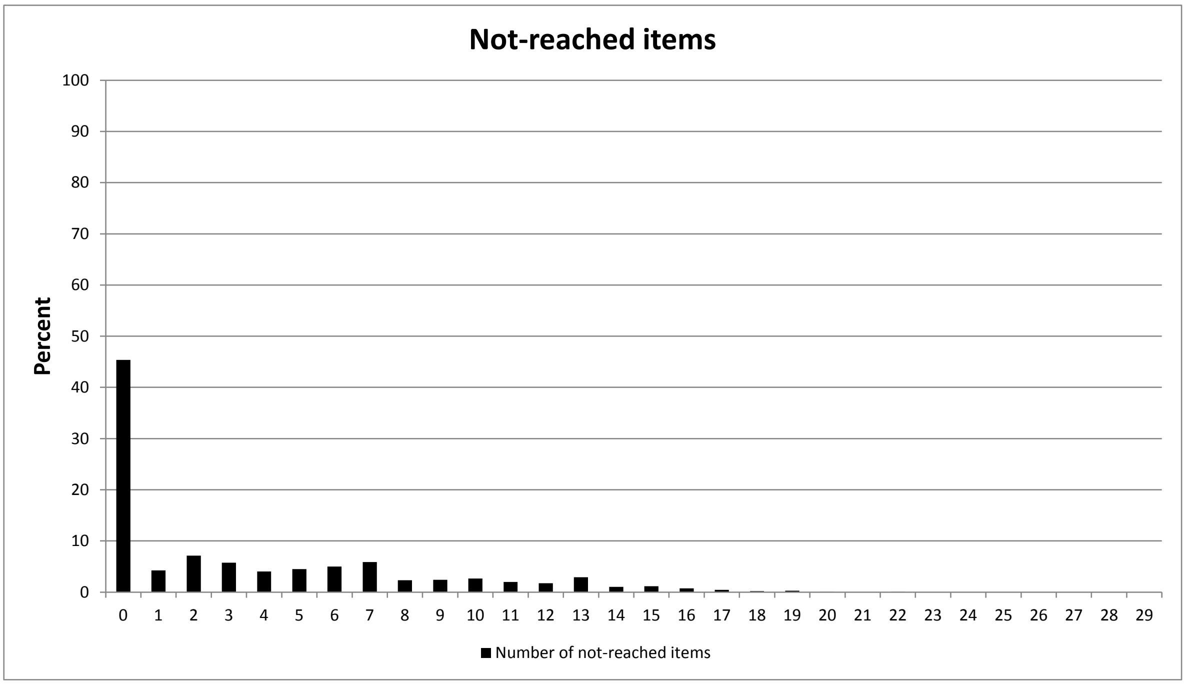

Figure 4. Number of not reached items.

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 fifth of the subjects (19.7%) showed no missing response at all, and 57.1% of the adults had more than three missing values or more. On the other hand, only 8.6% of the subjects had missing responses for more than half of the items.

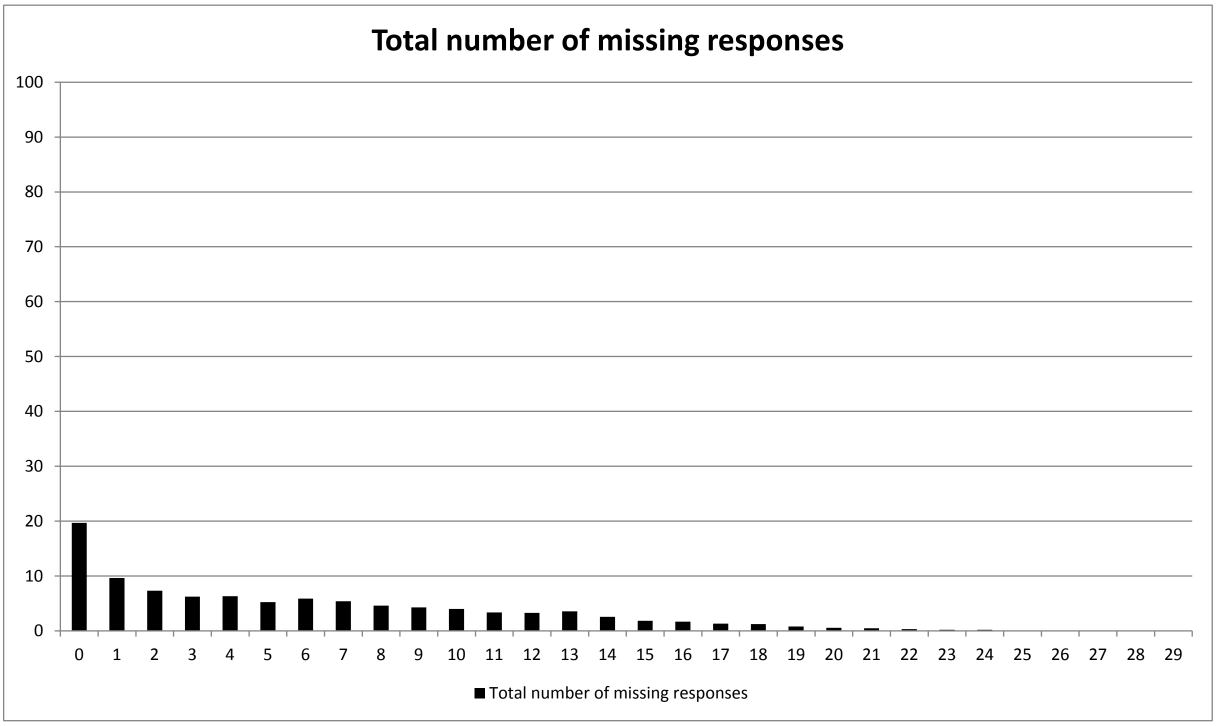

Figure 5. Total number of missing responses.

Overall, there is a small amount of invalid responses and a reasonable amount of omitted items. The number of not reached items is rather large and, therefore, so is 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 2.44% for item ica4019x_c. Overall, the number of persons omitting an item is acceptable. There are eight items with an omission rate above 10%. The highest omission rate occurs for item ica4010x_c (35.2% of the persons omitted this item). The number of omitted responses is correlated to .28 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 twice as often (11.1%) than simple MC items (6.1%). The number of persons that did not reach an item increases with the position of the item in the test to up to 54.6%. This is a rather large amount. The total number of missing responses (sum of invalid, omitted, and not-reached responses) per item varies between 2.2% (item ica4005x_c) and 58.2% (item ica4057x_c).

5.2 Parameter Estimates

5.2.1 Item parameters

The estimated item difficulties are depicted in Table 5. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties vary between -2.47 (item ica4050s_c) and 1.47 (item ica4005x_c) with a mean of -0.95. The mean probability for solving an item was .52, 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) ≤ 0.05). The step parameters for CMC items are depicted in Table 6.

5.2.2 Person parameters

Person parameters are estimated as WLEs (Pohl & Carstensen, 2012a). WLEs are provided in the first release of the SUF. 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 1.26, indicating that the test differentiates well between subjects. The reliability of the test (EAP/PV reliability = .81, WLE reliability = .79) is sufficient.

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 persons' ability distribution. However, only few items cover a very high degree of ability. There is a large number of items with a medium or low difficulty. As a consequence, subjects with a medium and low ability are measured relatively precisely, whereas subjects with a high ability have a larger standard error.

Table 4

Missing Values

Item	Position in the test	Number of categories	Number of valid responses	Relative frequency of not-reached items in %	Relative frequency of omitted items in %	Relative frequency of invalid responses in %
ica4001x_c	1	2	5964	0.00	2.82	0.02
ica4003x_c	2	2	6002	0.00	0.80	1.42
ica4005x_c	3	2	6004	0.00	1.61	0.57
ica4004s_c	4	6	5656	0.00	7.82	0.03
ica4006x_c	5	2	5969	0.00	2.31	0.44
ica4007x_c	6	2	5964	0.02	2.43	0.39
ica4008x_c	7	2	5671	0.02	6.17	1.42
ica4010x_c	8	2	3967	0.10	35.24	0.03
ica4017s_c	9	6	5054	0.15	17.46	0.05
ica4018s_c	10	2	5487	0.23	10.04	0.34
ica4015s_c	11	4	5253	0.49	13.83	0.10
ica4019x_c	12	2	5876	0.67	1.16	2.44
ica4016s_c	13	6	5520	1.11	8.93	0.03
ica4020s_c	14	5	5407	1.86	10.02	0.03
ica4023x_c	15	2	5141	3.01	12.66	0.57
ica4027x_c	16	2	5516	4.02	5.59	0.52
ica4026x_c	17	2	4752	6.92	15.51	0.15
ica4029x_c	18	2	5264	8.68	5.26	0.29
ica4028x_c	19	2	5340	10.67	2.31	0.02
ica4030x_c	20	2	5199	13.34	1.73	0.23
icg9119x_c	21	2	4808	15.75	5.59	0.33
ica4050s_c	22	5	4109	18.07	14.84	0.15
ica9122x_c	23	2	4214	23.97	6.94	0.44
ica4047s_c	24	6	3764	28.98	9.63	0.07

ica4046x_c	25	2	4025	33.48	0.70	0.24
ica4021s_c	26	6	3232	37.50	9.78	0.07
ica4052s_c	27	6	2982	43.24	8.11	0.07
ica4054x_c	28	2	2837	50.39	3.21	0.18
ica4057x_c	29	2	2565	54.63	3.55	0.03

Note. * For the CMC items, the maximum credit results from the discrimination times (number of categories - 1).

Table 5

Item Parameters

Item	**Percentage correct**	**Difficulty/ location parameter**	**SE (difficulty/location parameter)**	**WMNSQ**	***t*-value of WMNSQ**	**Correlation of item score with total score**	**Discrimination- 2 PL**
ica4001x_c	81.29	-1.78	0.04	1.06	2.80	0.35	0.85
ica4003x_c	30.39	1.04	0.03	1.04	3.10	0.41	0.89
ica4005x_c	23.60	1.47	0.03	0.94	-3.40	0.47	1.41
ica4004s_c	n.a.	-1.82	0.03	0.91	-4.90	0.55	*0.87
ica4006x_c	67.87	-0.93	0.03	1.06	3.90	0.43	0.88
ica4007x_c	63.43	-0.68	0.03	0.92	-6.10	0.57	1.57
ica4008x_c	75.14	-1.32	0.03	1.15	8.20	0.30	0.56
ica4010x_c	54.40	-0.01	0.04	1.00	0.00	0.49	1.10
ica4017s_c	n.a.	-1.74	0.04	0.97	-1.30	0.46	*0.59
ica4018s_c	n.a.	-0.64	0.03	1.11	7.20	0.40	*0.40
ica4015s_c	n.a.	-2.23	0.05	1.01	0.50	0.35	*0.56
ica4019x_c	44.25	0.31	0.03	1.16	12.80	0.34	0.57
ica4016s_c	n.a.	-2.16	0.04	0.88	-5.70	0.57	*1.06
ica4020s_c	n.a.	-2.22	0.04	1.05	3.00	0.36	*0.44
ica4023x_c	44.39	0.39	0.03	1.07	5.30	0.41	0.86
ica4027x_c	44.43	0.35	0.03	1.09	7.60	0.39	0.77
ica4026x_c	41.29	0.58	0.03	1.05	3.70	0.42	0.88
ica4029x_c	81.17	-1.72	0.04	1.02	1.10	0.39	1.07

ica4028x_c	74.04	-1.25	0.03	1.05	2.70	0.41	0.91
ica4030x_c	69.30	-0.96	0.03	0.83	-11.50	0.64	2.29
icg9119x_c	75.75	-1.32	0.04	1.00	0.20	0.45	1.15
ica4050s_c	n.a.	-2.47	0.04	0.87	-6.80	0.57	*1.04
ica9122x_c	56.83	-0.21	0.04	0.97	-1.80	0.52	1.24
ica4047s_c	n.a.	-2.10	0.04	0.91	-2.90	0.53	*0.86
ica4046x_c	69.74	-0.96	0.04	0.89	-6.40	0.59	1.74
ica4021s_c	n.a.	-1.80	0.05	0.88	-4.60	0.56	*1.04
ica4052s_c	n.a.	-0.99	0.04	0.88	-5.20	0.61	*0.89
ica4054x_c	71.41	-1.00	0.05	0.98	-0.80	0.50	1.23
ica4057x_c	78.56	-1.45	0.05	0.97	-1.20	0.49	1.33

Notes. For the dichotomous items, the correlation with the total score 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).
Percentage of correct scores is not informative of polytomous CMC and MA item scores. These are denoted by n.a.
* For the CMC items, the maximum credit results from the discrimination times (number of categories - 1).

Table 6

Step Parameters for CMC Items

Item	**Step 1**	**SE (Step 1)**	**Step 2**	**SE (Step 2)**	**Step 3**
Ica4004s_c	-0.524	0.027	-0.283	0.028	0.807
Ica4017s_c	-0.466	0.029	-0.398	0.029	0.844
Ica4018s_c	-0.543	0.027	0.188	0.030	0.355
Ica4015s_c	0.206	0.036	-0.206		
Ica4016s_c	-0.052	0.029	-0.150	0.032	0.202
Ica4020s_c	-1.262	0.028	0.273	0.029	0.989
Ica4050s_c	-1.692	0.033	0.335	0.032	1.356
Ica4047s_c	0.485	0.037	-0.486	0.040	0.001
Ica4021s_c	-0.393	0.036	-0.754	0.036	1.147
Ica4052s_c	-0.387	0.037	-0.094	0.040	0.481

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

Figure 6. 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 3).

5.3 Quality of the Test

5.3.1 Fit of the subtasks of complex multiple choice 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 68 items. One fourth of the subtasks (18 out of 68 subtasks) 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 0.81 to 1.21, the corresponding t-value ranged from -21.2 to 16.6, and there were no noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. Due to the sufficient 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 point-biserial correlation between each incorrect response (distractor) and the students' total score. All distractors had a point-biserial correlation with ability below zero (Median = -.19). The results indicate a good model fit.

5.3.3 Item fit

The item fit is mostly good. WMNSQ is close to 1 with the lowest value being 0.83 (item ica4030x_c) and the highest being 1.16 (item ica4019x_c). There are only four items with a WMNSQ above 1.07 and a corresponding t-value above 7. The correlation of the item score with the total score varies between .30 (for item ica4008x_c) and .64 (for item ica4030x_c) with an average correlation of .47. Many items (19 out of 29 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 .52, indicating a good targeting of item difficulties and person abilities.

5.3.4 Differential item functioning

The test fairness for different groups (i.e., measurement invariance) was investigated 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 socioeconomic status), duration of education (≤ 12 years vs. > 12 years), age (≤ 50 years vs. > 50 years), and migration background (see Pohl & Carstensen, 2012a, for a description of these variables). Table 8 shows the difference between the estimated item difficulties in different groups. Female versus male, for example, indicates the difference in difficulty ß(female) – ß(male). A positive value indicates a higher difficulty for females; a negative value indicates a lower difficulty for females as opposed to males.

The computer literacy test was administered in two different positions (see Section 3.1 for the design of the study). In total, 3,100 (50.5%) persons received the computer literacy test before the science test (Position 1), and 3,038 (49.5%) of the persons received the computer literacy test after having completed the science test (Position 2). The subjects were randomly assigned to either one 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 perform on average 0.11 logits (Cohen's d = -0.08) worse 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 0.23 logits.

The investigation of DIF for gender showed that 3,070 (50.0%) of the test takers were female and 3,068 (50.0%) were male. On average, male adults have a higher computer literacy than female adults (main effect = 0.41 logits, Cohen's d = 0.32). There is no item with a considerable gender DIF. The highest difference in difficulties between the two groups is -0.55 logits.

The number of books at home was used as a proxy for socioeconomic status. There were 3,145 (51.2%) test takers with 0 to 200 books at home, 2,589 (42.2%) test takers with more than 200 books at home, and 404 (6.6%) test takers without a valid response. DIF was investigated using these three groups. There are considerable average differences between the three groups. Participants with 200 or less books at home perform on average 0.44 logits (Cohen's d = 0.35) lower in computer literacy than participants with more than 200 books. Participants without a valid response on the variable 'books at home' performed 0.13 logits (Cohen's d = 0.10) better than participants with up to 200 books and 0.32 logits (Cohen's d = 0.25) worse than paticipants with more than 200 books. There is no considerable DIF comparing participants with many or fewer books (highest DIF = 0.43 logits). Comparing the group without valid responses to the two groups with valid responses, DIF occurs up to 0.43 logits. This is a rather small difference, so that there is no considerable socioeconomic DIF.

Regarding the duration of education, 3,010 (49.0%) of the adults had received 12 years of education or less. The other group of 2,949 (48.0%) adults had undergone an education period of more than 12 years. A proportion of 3.0% (179 adults) provided no valid response. On average, adults who spent more time in education performed much better in the computer literacy test than adults who spent less time in education (main effect = 1.072 logits, Cohen's d = 0.85). The adults without valid responses performed better than the group with 12 years of education or less (main effect = 0.545 logits, Cohen's d = 0.43) and worse than the group with more than 12 years of education (main effect = -0.527 logits, Cohen's d = -0.42). There are only two items (ica4004s_c, ica4050s_c) with considerable DIF (0.64 logits and 0.63 logits, respectively) between the groups with more or less than 12 years of education. Comparing the group without valid responses to the groups with valid responses, there is only one item (ica4021s_c) with considerable DIF (0.81 logits and 0.62 logits, respectively). The other items show no considerable DIF in terms of duration of education.

DIF was also investigated for age. Overall, 2,960 (48.2%) of the test takers were 50 years old or less, and 3,178 (51.8%) of the test takers were older than 50 years. On average, the younger adults performed much better in the computer literacy test than the older adults (main effect = 0.854 logits, Cohen's d = 0.68). There is only one item (ica4046x_c) with considerable DIF (0.86 logits).

[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.

There were 5,293 (86.2%) participants without a migration background, and 845 (13.8%) participants with a migration background. On average, adults without a migration background performed slightly better in the computer literacy test than adults with a migration background (main effect = 0.14 logits, Cohen’s d = 0.11). There is no item with a considerable DIF between adults without and with a migration background. The highest difference in difficulties between the two groups is -0.33 logits.

Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models allowing for DIF with those allowing for main effects only. In Table 7, the models including only main effects are compared with those that estimate DIF additionally. The Akaike's (1974) information criterion (AIC) favors the models estimating DIF for all DIF variables except migration background. The Bayesian information criterion (BIC, Schwarz, 1978) takes into account the number of estimated parameters and sample size 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 half of the DIF variables (position, books, migration). Only for the DIF variables gender, duration of education, and age, the more complex DIF model have slightly better information criteria.

Table 7

Comparison of Models With and Without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Position	main effect	200444.648	50	200544.648	200880.761
	DIF	200378.363	79	200536.363	201067.421
Gender	main effect	200294.653	50	200394.653	200730.765
	DIF	199881.859	79	200039.859	200570.917
Books	main effect	200247.350	51	200349.350	200692.185
	DIF	200028.283	109	200246.283	200979.009
Duration of education	main effect	199156.738	51	199258.738	199601.573
	DIF	198621.428	109	198839.428	199572.154
Age	main effect	199706.409	50	199806.409	200142.522
	DIF	199198.989	79	199356.989	199888.047
Migration	main effect	200448.189	50	200548.189	200884.301
	DIF	200390.802	79	200548.802	201079.860

Most of the differences in item difficulties estimated via the DIF analyses are in absolute values below 0.6. Only four items showed a DIF value above the threshold of 0.6: The items are ica4004s_c (duration of education), ica4050s_c (duration of education), ica4021s_c (duration of education), and ica4046x_c (age). But most values of these items (four out of six values: 0.64, 0.63, 0.62, 0.65) are only slightly above the threshold. Overall, the results indicate that there is no considerable DIF and that the test is treating the considered groups fairly.

Table 8

Differential Item Functioning (Absolute Differences Between Difficulties)

Item	Booklet	Gender	Books			Duration of education			Age	Immigration background
	Position 1 vs. 2	Female vs. Male	(< 200) vs. (> 200)	(< 200) vs. missing	(>200) vs. missing	≤ 12 years vs. > 12 years	≤ 12 years vs. missing	> 12 years vs. missing	≤ 50 years vs. > 50 years	Without vs. with
ica4001x_c	0.182	-0.382	0.052	-0.013	-0.065	0.002	-0.305	-0.307	0.140	-0.102
ica4003x_c	0.112	-0.304	0.296	0.241	-0.055	0.211	0.194	-0.017	-0.090	-0.090
ica4005x_c	0.056	0.320	0.057	0.039	-0.018	0.315	0.441	0.126	0.044	-0.248
ica4004s_c	0.168	-0.050	0.391	0.182	-0.209	0.642	-0.003	-0.645	-0.426	-0.226
ica4006x_c	0.228	-0.504	-0.017	0.038	0.055	-0.297	-0.187	0.110	0.072	-0.058
ica4007x_c	-0.034	0.168	0.119	0.225	0.106	0.128	0.130	0.002	-0.252	0.132
ica4008x_c	-0.034	-0.030	-0.052	0.082	0.134	-0.474	-0.168	0.306	0.480	0.022
ica4010x_c	0.062	0.154	-0.019	0.291	0.310	0.138	-0.051	-0.189	0.114	-0.002
ica4017s_c	-0.064	-0.178	0.271	-0.073	-0.344	0.132	-0.285	-0.417	0.264	-0.104
ica4018s_c	0.138	0.056	-0.429	-0.237	0.192	-0.427	0.038	0.465	-0.240	0.108
ica4015s_c	-0.102	0.320	-0.002	-0.042	-0.040	-0.002	0.002	0.004	0.144	-0.300
ica4019x_c	-0.122	-0.348	0.021	-0.171	-0.192	-0.447	-0.255	0.192	0.500	-0.068
ica4016s_c	0.018	0.150	0.195	-0.238	-0.433	0.427	0.194	-0.233	-0.220	-0.120
ica4020s_c	-0.006	-0.482	-0.143	0.014	0.157	-0.182	-0.055	0.127	0.164	0.018
ica4023x_c	-0.084	0.312	-0.133	0.142	0.275	0.035	0.289	0.254	0.198	0.278
ica4027x_c	-0.028	-0.034	-0.110	-0.046	0.064	-0.273	-0.255	0.018	0.286	0.144
ica4026x_c	-0.156	-0.158	-0.008	0.230	0.238	-0.122	0.133	0.255	0.004	0.116
ica4029x_c	-0.114	0.168	-0.348	-0.042	0.306	-0.239	0.007	0.246	0.214	0.064
ica4028x_c	0.014	-0.182	0.025	-0.016	-0.041	-0.104	-0.091	0.013	0.284	0.096
ica4030x_c	0.034	0.550	0.054	0.111	0.057	0.430	0.242	-0.188	-0.506	0.040

icg9119x_c	-0.108	0.202	0.084	0.285	0.201	0.194	0.241	0.047	0.122	0.088
ica4050s_c	-0.040	0.346	0.245	-0.109	-0.354	0.626	0.397	-0.229	-0.358	0.054
icg9122x_c	-0.164	0.026	0.154	0.119	-0.035	0.174	0.459	0.285	0.084	-0.110
ica4047s_c	-0.152	-0.022	0.135	0.425	0.290	0.283	-0.190	-0.473	-0.070	-0.120
ica4046x_c	0.010	0.090	-0.022	0.061	0.083	0.116	0.337	0.221	-0.860	0.328
ica4021s_c	0.092	0.474	0.029	-0.107	-0.136	0.188	0.808	0.620	-0.502	-0.318
ica4052s_c	0.034	0.062	0.395	0.096	-0.299	0.564	0.156	-0.408	0.106	0.140
ica4054x_c	-0.076	-0.108	0.036	0.372	0.336	0.032	0.304	0.272	-0.232	0.116
ica4057x_c	-0.130	0.164	-0.011	0.296	0.307	0.048	0.162	0.114	-0.128	0.060
Main effect	-0.108	-0.406	-0.443	-0.127	0.316	-1.072	-0.545	0.527	0.854	0.136

5.3.5 Rasch homogeneity

In order to test the assumption of Rasch homogeneity, we also fitted a generalized partial credit model (2PL) to the data. The estimated discrimination parameters are depicted in Table 5. They range from 0.56 (item 7) to 2.29 (item 20) for the MC items, and from 0.40 (item 10) to 1.06 (item 13) per category for the CMC items. Because the discriminations differ considerably among the items, the 2PL model (BIC = 198998, number of parameters = 94) fits the data better than the PCM model (BIC = 200456, number of parameters = 49). Because 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), the Rasch 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.

To estimate a multidimensional (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 9. All four dimensions show a substantive variance with the highest discrimination between subjects for Access and the lowest for Evaluate. The correlations between the dimensions vary between .847 and .966. The lowest correlation is found between Dimension 1 (Access) and Dimension 4 (Evaluate). All other correlations are above .90. Thus, the results indicate some degree of multidimensionality.

Table 9

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 (8 Items)	2.171			
Create (7 Items)	.929	1.065		
Manage (9 Items)	.943	.966	1.485	
Evaluate (5 Items)	.847	.910	.916	.929

To estimate a four-dimensional model based on the different types of software applications Gauss' estimation (nodes = 15) was used (see Table 10). The results of the analyses are depicted in Table 7. All four dimensions show a substantive variation. The correlations between the four dimensions are very high (between .920 and .967). The four software applications do not measure different constructs but a unidimensional construct.

Table 10

Results of Four-Dimensional Scaling (Software Applications).

	Dim 1	Dim 2	Dim 3	Dim 4
Word processing (6 Items)	2.346			
Spreadsheet / presentation software (10 Items)	.929	1.269		
E-mail / communication tools (4 Items)	.920	.967	1.134	
Internet / search engines (9 Items)	.925	.960	.964	1.124

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

6. Discussion

The analyses in the previous sections have aimed to provide information on the quality of the computer literacy test in Starting Cohort 6 (Adults, Wave 4) and describe how the computer literacy score is estimated. The analyses we conducted and presented 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 and subtasks of CMC, 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 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 but one items is 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 appropriately, although there is some evidence for multidimensionality.

The distribution of item difficulties and the distribution of person parameters overlap to a great extent; however, with one limitation: There are only few items that 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 min. 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 independent 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 29 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 an ‘x_c’ at the end of the variable name, the CMC items are marked with an ‘s_c’ at the end of the variable name. In the scaling model, each category of the polytomous CMC items is scored with 0.5 points. Manifest scale scores are provided in the 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 indicates the number of correctly solved subtasks. Therefore, the aggregation of categories has to be carried out by the data user. It is recommended to collapse categories with less than N = 200 in order to avoid any estimation problems (see also Section 4.2). We collapsed the two lowest categories for three CMC items and the lowest three categories for six CMC items for the estimation. For one item with eight subtasks a special scoring was used (see Section 4.2 and also syntax in the Appendix). We advise data users to do so as well.

Plausible values allow users to investigate latent relationships of competence scores with other variables will be provided in forthcoming data releases. Users 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. (2015). Incorporating different response formats in the IRT-scaling model for competence data. Manuscript submitted for publication.

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

Pohl, S., & Carstensen, C. H. (2012a). 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. (2012b). 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.

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

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.

von Davier, M. (2005). Mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent trait models [Computer Software]. Princeton, NJ: ETS.

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.

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 6, Adults (Wave 4, B69)

```
title ICT HE Adults scaling 29 items included, partial credit model;

datafile >>filename.dat;
format pid 1-7 rotation 9 responses 10-38;

labels <<filename_with_labels.txt;

codes 0,1,2,3,4,5,6,7,8;

recode (0,1,2,3)            (0,0,1,2)            !item(11);
recode (0,1,2,3,4)          (0,0,1,2,3)          !item(14,22);
recode (0,1,2,3,4,5)        (0,0,0,1,2,3)        !item(4,9,13,24,26,27);
recode (0,1,2,3,4,5,6,7,8)  (0,1,1,1,2,2,2,2,3)  !item(10);

score (0,1)        (0,1)          !item(1-3,5-8,12,15-21,23,25,28-29);
score (0,1,2)      (0,.5,1)       !item(11);
score (0,1,2,3)    (0,.5,1,1.5)   !item(4,9,10,13,14,22,24,26,27);

set constraint=cases;
model item + item*step - rotation;
estimate ! method=gauss, nodes = 15;

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