NEPS SURVEY PAPERS

Jan Marten Ihme and Martin Senkbeil

NEPS TECHNICAL REPORT FOR COMPUTER LITERACY: SCALING RESULTS OF STARTING COHORT 2 FOR GRADE 3

NEPS Survey Paper No. 42
Bamberg, June 2018

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

NEPS Technical Report for Computer Literacy: Scaling Results of Starting Cohort 2 for Grade 3

Jan Marten Ihme & Martin Senkbeil

Leibniz Institute for Science and Mathematics Education at the University of Kiel

E-mail address of lead author:

ihme@ipn.uni-kiel.de

Bibliographic data:
Ihme, J. M. & Senkbeil, M. (2018). *NEPS Technical Report for Computer Literacy: Scaling results of Starting Cohort 2 for Grade 3* (NEPS Survey Paper No. 42). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP42:1.0

NEPS Technical Report for Computer Literacy: Scaling Results of Starting Cohort 2 for Grade 3

Abstract

The National Educational Panel Study (NEPS) investigates the development of competencies across the life span and develops tests for the assessment of different competence domains. In order to evaluate the quality of the competence tests, a range of analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedures for the computer literacy test in grade 3 of starting cohort 2 (kindergarten). The computer literacy test contained 30 items with MC response format representing different cognitive requirements and different content areas. The test was administered to 5,620 students. Their responses were scaled using the Rasch model. Item fit statistics, differential item functioning, Rasch-homogeneity, the test's dimensionality, and local item independence were evaluated to ensure the quality of the test. These analyses showed that the test exhibited an acceptable reliability and that all items fitted the model in a satisfactory way. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test were the low variance and the large percentage of items at the end of the test that were not reached due to time limits. Further challenges related to the dimensionality analyses based on both software applications and cognitive requirements. Overall, the computer literacy test had acceptable psychometric properties that allowed for a reliable estimation of computer competence scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the ConQuest-syntax for scaling the data.

Keywords

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

Content

1 Introduction 4

2 Testing Computer Literacy 4

3 Data 6

3.1 The Design of the Study 6

3.2 Sample 6

4 Analyses 6

4.1 Missing Responses 6

4.2 Scaling Model 7

4.3 Checking the Quality of the Scale 7

5 Results 8

5.1 Missing Responses 8

5.1.1 Missing responses per person 8

5.1.2 Missing responses per item 10

5.2 Parameter Estimates 12

5.2.1 Item parameters 12

5.2.2 Test targeting and reliability 14

5.3 Quality of the Test 15

5.3.1 Distractor analyses 15

5.3.2 Item fit 15

5.3.3 Differential item functioning 15

5.3.5 Rasch homogeneity 18

5.3.6 Unidimensionality 19

6 Discussion 21

7 Data in the Scientific Use File 21

7.2 Computer literacy scores 22

References 23

Appendix 25

1 Introduction

Within the National Educational Panel Study (NEPS), different competencies are measured coherently across the life span. Tests have been developed for different competence domains. These include, among other things, reading competence, mathematical competence, scientific literacy, information and communication literacy (computer literacy), metacognition, vocabulary, and domain-general cognitive functioning. An overview of the competences measured in the NEPS is given by Weinert et al. (2011) as well as Fuß, Gnambs, Lockl, and Attig (2016).

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

In this paper the results of these analyses are presented for computer literacy in starting cohort 2 (kindergarten) in grade 3. First, the main concepts of the computer literacy test are introduced. Then, the computer literacy data of starting cohort 2 and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the scientific use file is presented.

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

2 Testing Computer Literacy

The framework and test development for the computer literacy test is described in Weinert et al. (2011) and in Senkbeil, Ihme, and Wittwer (2013). In the following, we point out specific aspects of the computer literacy test that are necessary for understanding the scaling results presented in this paper.

Computer literacy is conceptualized as a unidimensional construct comprising the different facets of technological and information literacy. In line with the literacy concepts of international large-scale assessments, we define computer literacy from a functional perspective. That is, functional literacy is understood to include the knowledge and skills that people need to live satisfying lives in terms of personal and economic satisfaction in modern-day societies. This leads to an assessment framework that relies heavily on everyday problems, which are more or less distant to school curricula. As a basis for the construction of the instrument assessing computer literacy in NEPS, we use a framework that identifies four process components (*access*, *create*, *manage*, and *evaluate*) of computer literacy representing the knowledge and skills needed for a problem-oriented use of modern information and communication technology (see Figure 1). Apart from the process components, the test construction of TILT (Test of Technological and Information Literacy) is guided by a categorization of software applications (*operating system/word processing/presentation* *software/graphics*, *spread sheet*, *internet / search engines*, and *e-mail*) that are used to locate, process, present, and communicate information.

The framework for this test is adapted for third-graders. The process components are equal to the main framework as presented in figure 1, but the software applications differ. Communication tools like e-mails are removed from the framework for this age group. Instead, there are more items regarding the basic skills in operating systems and word processing. Hence, the software application are classified in (1) operating system, (2) word processing and spread sheet, (3) graphics, and (4) internet/search engines.

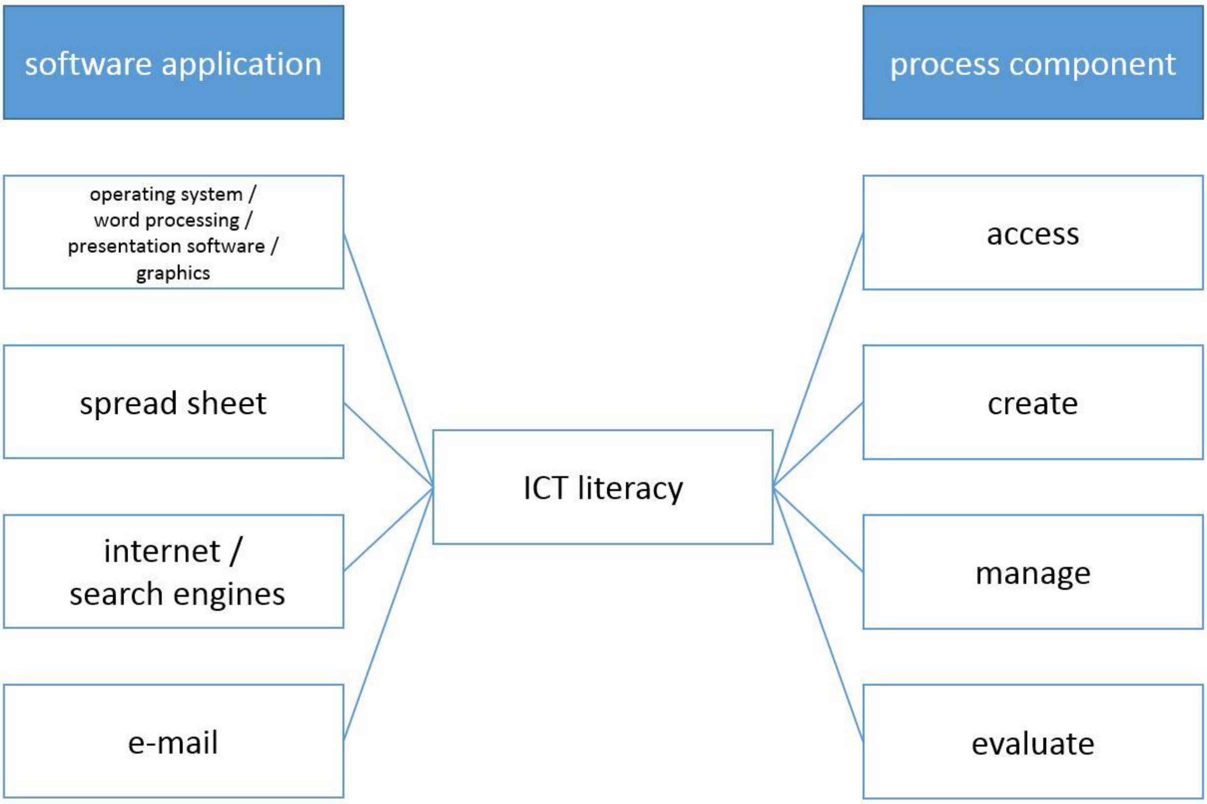

software application
process component
operating system / word processing / presentation software / graphics
spread sheet
internet / search engines
e-mail
ICT literacy
access
create
manage
evaluate

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

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

In the computer literacy test of starting cohort 2 (kindergarten) in grade 3 there is only one type of response format, that is simple multiple choice (MC). 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). Examples of the different response formats are given in Pohl and Carstensen (2012).

3 Data

3.1 The Design of the Study

The study assessed different competence domains including, among others, computer literacy. The competence tests for these two domains were always presented first within the test battery. The computer literacy test was in all cases administered on the first of two testing days as second test after the science literacy test. All students received the test items in the same order. The competence test for computer literacy that was administered in the present study included 30 items (see Table 1) which represented all four process components of the computer literacy framework. In order to evaluate the quality of these items extensive preliminary analyses were conducted. These preliminary analyses revealed that none of the items had a poor fit.

Table 1

Number of Items for the Different Process Components

Process components \ Software application	Access	Create	Manage	Evaluate	**Sum**
Operating system	3	0	0	1	4
Word processing/ spread sheet	3	5	2	2	12
Graphics	1	3	0	0	4
Internet/search engines	4	0	3	3	10
Total number of items	11	8	5	6	30

3.2 Sample

A total of 5,620 individuals received the computer literacy test. One participant had missing values on all items and was excluded from further analyses. Thus, the analyses presented in this paper are based on a sample of 5,619 individuals. A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

4 Analyses

4.1 Missing Responses

In this test, there are different kinds of missing responses. These are a) invalid responses, b) omitted items, and c) items that test takers did not reach.

Invalid responses occurred, for example, when two response options were selected in simple MC items where only one was required, or when numbers or letters that were not within the range of valid responses were given as a response. Omitted items occurred when test takers skipped some items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response given were coded as not-reached.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions) and need to be accounted for in the estimation of item and person parameters. We, therefore, thoroughly investigated the occurrence of missing responses in the test. First, we looked at the occurrence of the different types of missing responses per person. This gave an indication of how well the persons were coping with the test. We then looked at the occurrence of missing responses per item in order to obtain some information on how well the items worked.

4.2 Scaling Model

To estimate item and person parameters for computer literacy competence, a Rasch model was used. Ability estimates for computer literacy were estimated as weighted maximum likelihood estimates (WLEs). Item and person parameter estimation in NEPS is described in Pohl and Carstensen (2012), whereas the data available in the SUF are described in Section 7.

4.3 Checking the Quality of the Scale

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

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

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

The computer literacy test should measure the same construct for all students. If any items favored certain subgroups (e.g., if they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between the subgroups (e.g., males and females) would be biased and thus unfair. For the present study, test fairness was investigated for the variables gender, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Differential item functioning (DIF) analyses were estimated using a multigroup IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties between the subgroups that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as noteworthy of further investigation, differences between 0.4 and 0.6 as considerable but not severe, and differences smaller than 0.4 as negligible DIF. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

The computer literacy was scaled using the Rasch model, which assumes Rasch-homogeneity. The Rasch model was chosen because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that might not hold for empirical data. To test the assumption of equal item discrimination parameters, a 2PL model was also fitted to the data and compared to the Rasch model.

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

4.4 Software

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

5 Results

5.1 Missing Responses

5.1.1 Missing responses per person

Figure 2 shows the number of invalid responses per person. Overall, there were very few invalid responses. About 97% of the respondents did not have any invalid response at all, and less than one percent had more than one invalid response.

Missing responses may also occur when respondents omit items. As illustrated in Figure 3 most respondents (83.5%) did not skip any item, and less than two percent omitted more than three items.

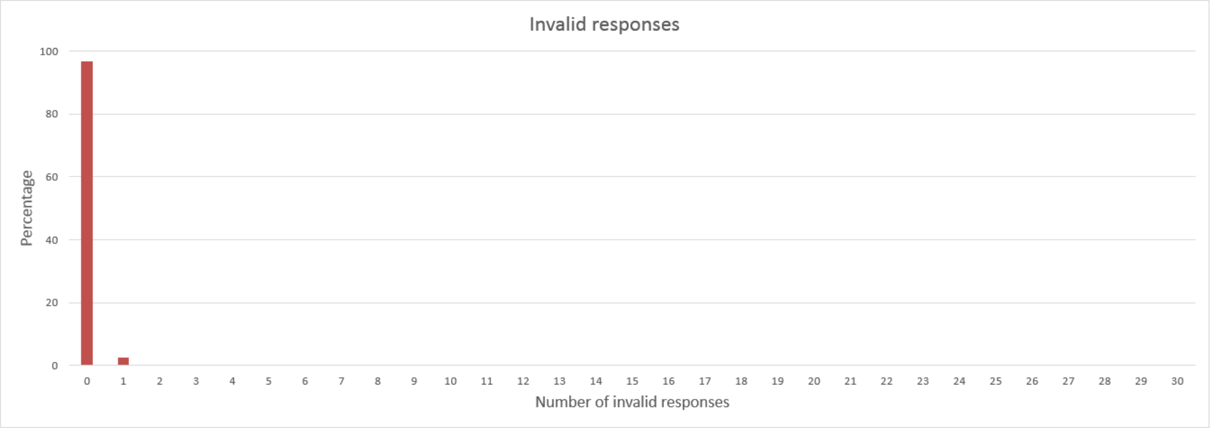


Figure 2. Number of invalid responses.

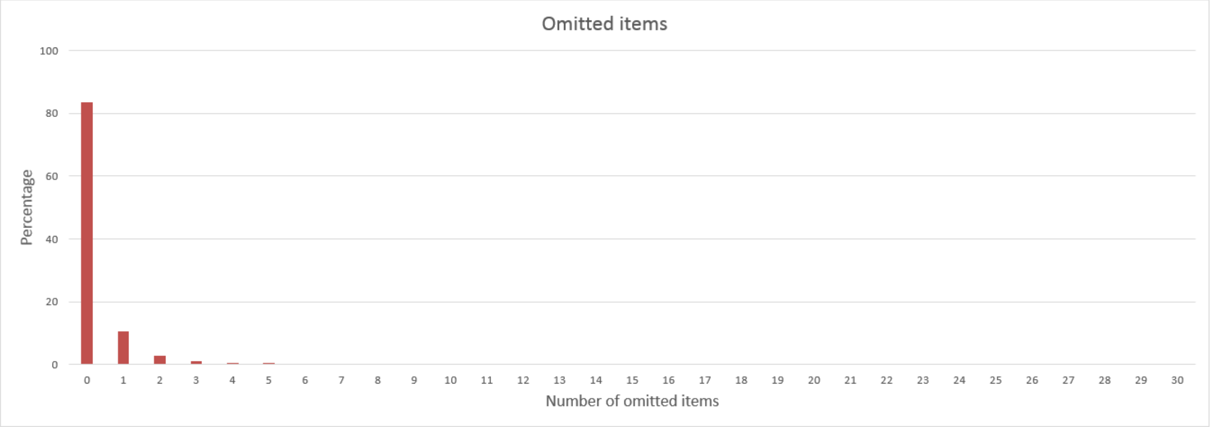


Figure 3. Number of omitted items.

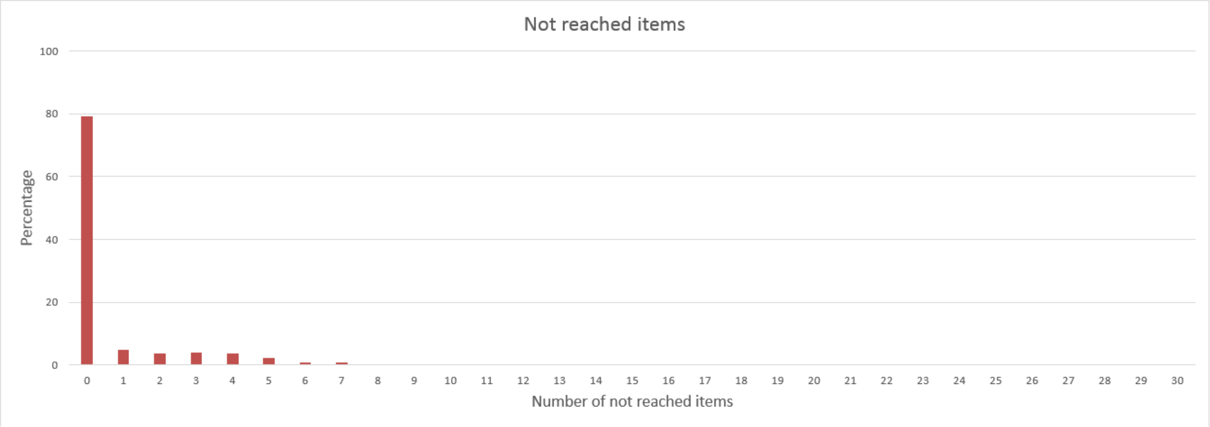


Figure 4. Number of not reached items.

Another source of missing responses are items that were not reached by the respondents; these are all missing responses after the last valid response. The number of not-reached items was rather low, most respondents were able to finish the test within the allocated time limit (Figure 4). About 79% of the respondents finished the entire test. About 2.5% of the participants did not reach the last five items or more.

The total number of missing responses, aggregated over invalid, omitted, not-reached, and not determinable per person, is illustrated in Figure 5. About 65% of the respondants had no missing responses at all. On average, the respondents showed M = 1.11 (SD = 2.34) missing responses. About 65% of the respondents had no missing response at all and about 7% of the participants had five or more missing responses.

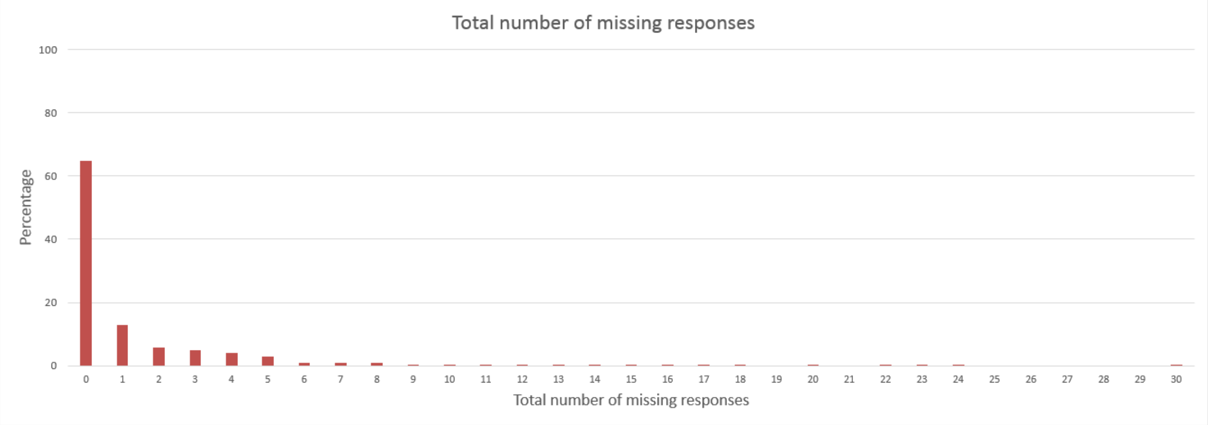


Figure 5. Total number of missing responses.

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

5.1.2 Missing responses per item

Table 2 provides information on the occurrence of different kinds of missing responses per item. Overall, the omission rates (column 5, "OM") were rather low, varying across items between 0.2 % and 3.2%. The omission rates correlated with the item difficulties at about .23. Generally, the percentage of invalid responses per item (column 6, "NV") was very low with the maximum rate being below 0.6%. With an item's progressing position in the test, the amount of persons that did not reach the item (column 4, "NR") rose up to a reasonable amount of 20.8%. Particularly, the last three items of the tests were reached by less than 90% of the respondents (see Figure 6).

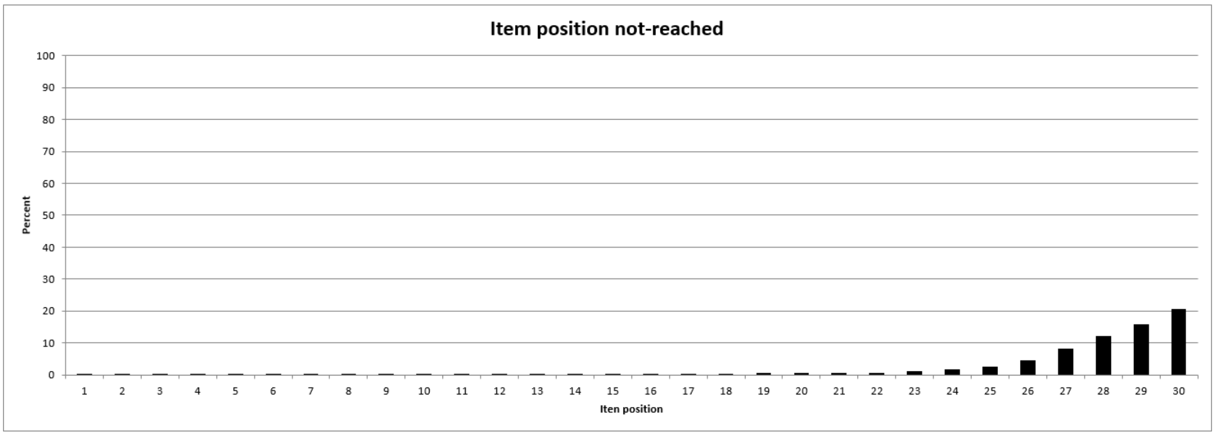


Figure 6. Item position not reached by test difficulty.

Table 2

Percentage of Missing Values

Item	Position	N	NR	OM	NV
icg3052x_c	1	5543	0.02	0.91	0.44
icg3350x_c	2	5436	0.02	3.19	0.07
icg3021x_c	3	5566	0.02	0.73	0.21
icg3610x_c	4	5508	0.02	1.73	0.25
icg3621x_c	5	5488	0.02	2.14	0.20
icg3371x_c	6	5556	0.02	1.00	0.12
icg3081x_c	7	5543	0.04	1.23	0.11
icg3102x_c	8	5542	0.07	1.19	0.12
icg3591x_c	9	5548	0.23	0.89	0.16
icg3092x_c	10	5554	0.23	0.80	0.14
icg3381x_c	11	5574	0.23	0.50	0.09
icg3400x_c	12	5522	0.23	1.46	0.05
icg3661x_c	13	5535	0.23	1.17	0.11
icg3410x_c	14	5551	0.23	0.91	0.09
icg3420x_c	15	5535	0.23	1.17	0.11
icg3432x_c	16	5527	0.23	1.25	0.18
icg3440x_c	17	5522	0.23	1.32	0.20
icg3322x_c	18	5535	0.23	1.17	0.11
icg3461x_c	19	5513	0.50	1.21	0.20
icg3211x_c	20	5532	0.53	0.48	0.55
icg3510x_c	21	5473	0.55	1.92	0.14
icg3221x_c	22	5526	0.73	0.75	0.20
icg3601x_c	23	5500	1.09	0.84	0.21
icg3260x_c	24	5426	1.85	1.32	0.28
icg3301x_c	25	5433	2.51	0.66	0.16
icg3270x_c	26	5307	4.70	0.73	0.14
icg3292x_c	27	5076	8.24	1.30	0.14
icg3481x_c	28	4879	12.26	0.64	0.28
icg3541x_c	29	4662	15.85	1.05	0.14
icg3550x_c	30	4441	20.77	0.21	0.00

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

5.2 Parameter Estimates

5.2.1 Item parameters

The second column in Table 3 presents the percentage of correct responses in relation to all valid responses for each item. Because there was a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index for item difficulty. The percentage of correct responses within dichotomous items varied between 14% and 77% with an average of 47% (*SD* = 18%) correct responses.

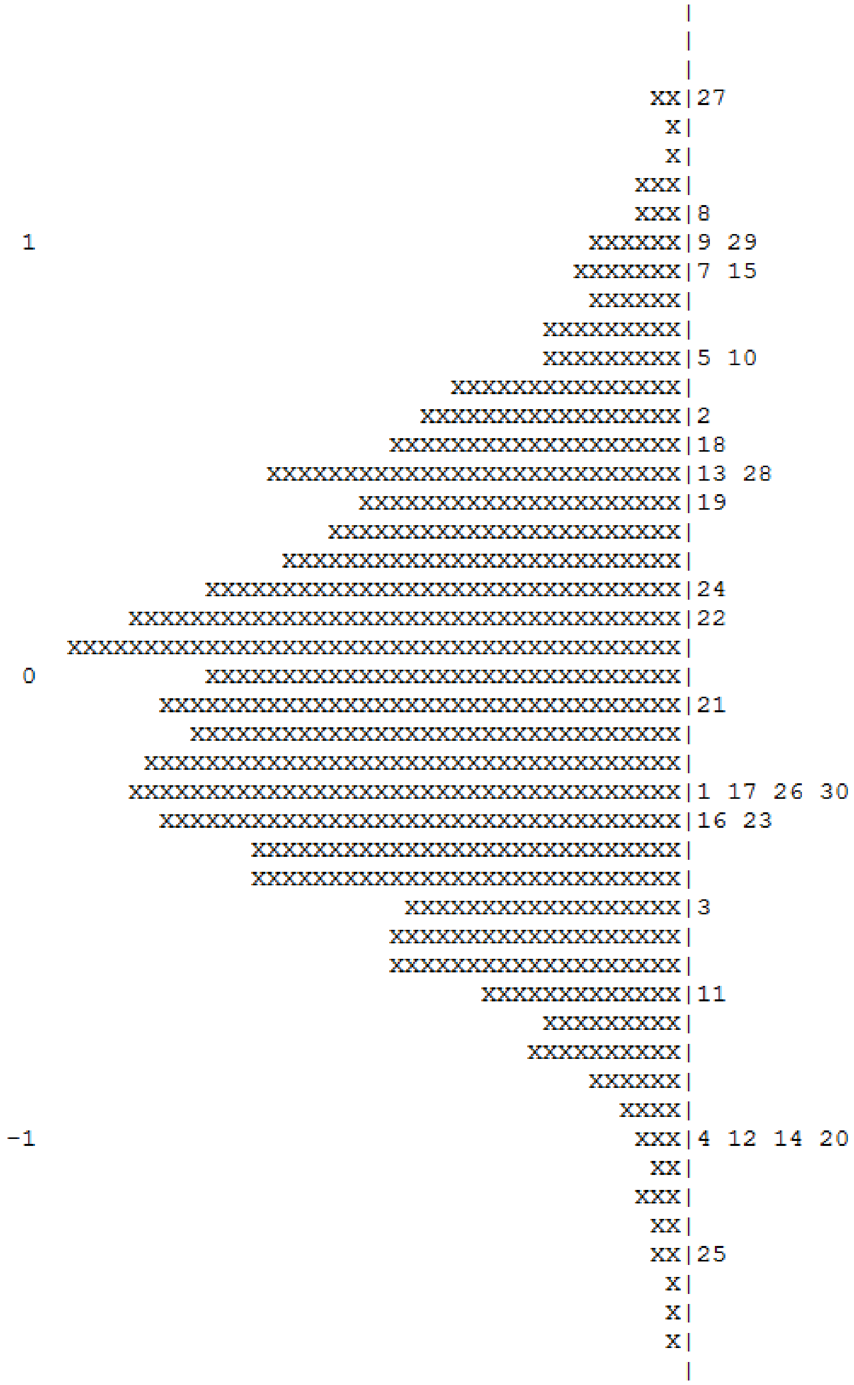
The estimated item difficulties are given in Table 3. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties ranged from -1.29 (item icg3301x_c) to 1.93 (item icg3371x_c) with an average difficulty of 0.15. Overall, the item difficulties were in a mean range with items with very high as well as items with very low difficulties.

Table 3

Item parameters

Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Q_3
icg3052x_c	56	-0.26	0.03	0.99	-2.20	.34	0.02
icg3350x_c	35	0.65	0.03	1.02	2.20	.23	0.01
icg3021x_c	62	-0.52	0.03	1.01	1.40	.27	0.02
icg3610x_c	72	-1.00	0.03	0.99	-0.70	.30	0.02
icg3621x_c	33	0.77	0.03	1.05	4.50	.13	0.02
icg3371x_c	14	1.93	0.04	1.03	1.10	.10	0.02
icg3081x_c	29	0.97	0.03	0.97	-2.60	.38	0.03
icg3102x_c	27	1.07	0.03	0.98	-1.40	.33	0.03
icg3591x_c	27	1.04	0.03	1.01	1.00	.22	0.02
icg3092x_c	32	0.77	0.03	1.01	0.90	.26	0.02
icg3381x_c	66	-0.69	0.03	1.00	-0.10	.28	0.02
icg3400x_c	73	-1.04	0.03	0.98	-1.60	.33	0.02
icg3661x_c	39	0.49	0.03	0.99	-1.50	.32	0.01
icg3410x_c	73	-1.04	0.03	0.99	-1.00	.31	0.01
icg3420x_c	29	0.92	0.03	1.01	0.40	.26	0.01
icg3432x_c	57	-0.30	0.03	0.99	-1.60	.33	0.02
icg3440x_c	56	-0.25	0.03	1.00	0.10	.30	0.02
icg3322x_c	37	0.56	0.03	1.04	4.60	.17	0.02
icg3461x_c	40	0.44	0.03	1.03	3.50	.22	0.02
icg3211x_c	73	-1.03	0.03	0.96	-3.20	.39	0.02
icg3510x_c	50	0.00	0.03	0.99	-1.60	.33	0.02
icg3221x_c	47	0.13	0.03	0.98	-2.70	.35	0.02
icg3601x_c	56	-0.27	0.03	0.97	-4.90	.39	0.02
icg3260x_c	45	0.21	0.03	0.98	-2.60	.34	0.01
icg3301x_c	77	-1.29	0.03	0.99	-0.50	.28	0.02
icg3270x_c	55	-0.21	0.03	1.02	2.80	.25	0.01
icg3292x_c	22	1.31	0.03	0.99	-0.40	.28	0.02
icg3481x_c	38	0.51	0.03	0.99	-1.40	.33	0.02
icg3541x_c	27	1.02	0.03	1.04	2.60	.16	0.02
icg3550x_c	56	-0.25	0.03	1.02	2.30	.26	0.02

Note. Difficulty = Item difficulty / location parameter, *SE* = standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, r_{it} = Corrected item-total correlation, Q_3 = Average absolute residual correlation for item (Yen, 1983).



```
                                          |
                                          |
                                          |
                                        XX|27
                                         X|
                                         X|
                                       XXX|
                                       XXX|8
  1                                 XXXXXX|9 29
                                   XXXXXXX|7 15
                                    XXXXXX|
                                 XXXXXXXXX|
                                 XXXXXXXXX|5 10
                           XXXXXXXXXXXXXXX|
                         XXXXXXXXXXXXXXXXX|2
                       XXXXXXXXXXXXXXXXXXX|18
               XXXXXXXXXXXXXXXXXXXXXXXXXXX|13 28
                     XXXXXXXXXXXXXXXXXXXXX|19
                   XXXXXXXXXXXXXXXXXXXXXXX|
                XXXXXXXXXXXXXXXXXXXXXXXXXX|
           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|24
      XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|22
  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
  0        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|21
          XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
       XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|
      XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|1 17 26 30
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX|16 23
              XXXXXXXXXXXXXXXXXXXXXXXXXXXX|
              XXXXXXXXXXXXXXXXXXXXXXXXXXXX|
                        XXXXXXXXXXXXXXXXXX|3
                       XXXXXXXXXXXXXXXXXXX|
                       XXXXXXXXXXXXXXXXXXX|
                             XXXXXXXXXXXXX|11
                                 XXXXXXXXX|
                                XXXXXXXXXX|
                                    XXXXXX|
                                      XXXX|
 -1                                    XXX|4 12 14 20
                                        XX|
                                       XXX|
                                        XX|
                                        XX|25
                                         X|
                                         X|
                                         X|
                                          |
```

Figure 7. Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 28.5 cases. Item difficulty is depicted on the right side of the graph. Each number represents one item (see Table 2).

5.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. In Figure 7, item difficulties of the computer literacy items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of item difficulties.

The mean of the ability distribution was constrained to be zero. The variance was estimated to be .24, indicating somewhat limited variability between subjects. The reliability of the test (EAP/PV reliability = .59; WLE reliability = .58) was rather low. The items covered a wide range of the ability distribution, but the mean item difficulty was *m* = .15 and slightly above the defined ability mean of 0.

5.3 Quality of the Test

5.3.1 Distractor analyses

In addition to the overall item fit, we specifically investigated how well the distractors performed in the test by evaluating the point-biserial correlation between each incorrect response (distractor) and the students' total score. All distractors had a point-biserial correlation with the total scores below zero with the exception of three items with a point-biserial-correlation between .00 and .05 (mean = -.13). The results indicate that the distractors worked well.

5.3.2 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the Rasch model. Altogether, item fit can be considered to be very good (see Table 3). Values of the WMNSQ ranged from 0.96 (item icg3211x_c _c) to 1.05 (icg3621x_c). None of the items exhibited a *t*-value of the WMNSQ greater than 6. Thus, there was no indication of any item over- or underfit. Point-biserial correlations between the item scores and the total scores ranged from .10 (item icg3371x_c) to .39 (items icg3211x_c and icg3601x_c) and had a mean of .28. All item characteristic curves showed a good fit of the items to the Rasch model.

5.3.3 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables gender, the number of books at home (as a proxy for socioeconomic status), migration background, school type, and test position (see Pohl & Carstensen, 2012, for a description of these variables).

The differences between the estimated item difficulties in the various groups are summarized in Table 4. For example, the column "Male vs. female" reports the differences in item difficulties between men and women; a positive value would indicate that the test was more difficult for males, whereas a negative value would highlight a lower difficulty for males as opposed to females. Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF to those that only estimate main effects (see Table 5).

Table 4

Differential Item Functioning

Item	**Sex**	**Books**	**Books**	**Books**	**Migration**	**Migration**	**Migration**
	Male vs. Female	<100 vs. >100	<100 vs. Missing	>100 vs. missing	without vs. with	without vs. missing	With vs. missing
icg3052x_c	-0.270	-0.004	0.064	0.068	-0.080	-0.007	0.073
icg3350x_c	-0.002	0.033	0.169	0.136	0.205	0.116	-0.089
icg3021x_c	0.130	-0.039	0.174	0.213	0.049	0.024	-0.025
icg3610x_c	0.276	0.072	0.048	-0.024	-0.125	-0.142	-0.017
icg3621x_c	0.080	-0.15	0.156	0.306	0.099	0.192	0.093
icg3371x_c	-0.032	-0.034	0.343	0.377	0.038	0.115	0.077
icg3081x_c	-0.128	-0.005	0.047	0.052	-0.006	-0.027	-0.021
icg3102x_c	0.046	0.005	0.106	0.101	0.198	0.170	-0.028
icg3591x_c	0.036	0.098	0.392	0.294	0.177	0.270	0.093
icg3092x_c	0.202	0.076	0.026	-0.050	-0.065	-0.123	-0.058
icg3381x_c	0.198	0.052	0.026	-0.026	-0.008	-0.025	-0.017
icg3400x_c	0.234	0.141	0.123	-0.018	0.032	-0.077	-0.109
icg3661x_c	-0.106	0.181	-0.052	-0.233	-0.114	-0.220	-0.106
icg3410x_c	0.018	-0.090	-0.027	0.063	-0.017	-0.004	0.013
icg3420x_c	-0.146	0.122	0.211	0.089	-0.151	-0.026	0.125
icg3432x_c	0.082	0.116	0.058	-0.058	-0.161	-0.152	0.009
icg3440x_c	-0.002	0.072	0.138	0.066	-0.023	0.080	0.103
icg3322x_c	-0.044	-0.111	0.042	0.153	0.150	0.092	-0.058
icg3461x_c	-0.136	-0.003	0.282	0.285	-0.243	0.033	0.276
icg3211x_c	0.240	0.236	-0.199	-0.435	-0.245	-0.421	-0.176
icg3510x_c	-0.534	0.166	0.020	-0.146	-0.101	-0.096	0.005
icg3221x_c	0.140	0.229	-0.031	-0.260	0.019	-0.187	-0.206
icg3601x_c	0.078	0.470	-0.014	-0.484	-0.412	-0.354	0.058
icg3260x_c	-0.138	0.275	0.034	-0.241	-0.250	-0.194	0.056
icg3301x_c	-0.262	-0.085	0.031	0.116	0.192	0.138	-0.054
icg3270x_c	0.200	-0.204	0.016	0.220	0.150	0.099	-0.051
icg3292x_c	-0.396	-0.041	0.164	0.205	0.424	0.194	-0.230
icg3481x_c	0.092	0.233	0.232	-0.001	-0.078	0.027	0.105
icg3541x_c	-0.158	-0.186	0.286	0.472	0.273	0.230	-0.043
icg3550x_c	0.244	0.151	0.212	0.061	-0.122	0.071	0.193
Main effect	0.032	-0.147	0.188	0.335	0.137	0.145	0.008

Sex: The sample included 2,751 (49%) males and 2,868 (51%) females. On average, male participants had a higher estimated computer literacy than females (main effect = 0.032 logits, Cohen's *d* = 0.07). No item showed DIF greater than 0.6 logits. An overall test for DIF (see Table 5) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). Model comparisons using Akaike's (1974) information criterion (AIC) and the Bayesian information criterion (BIC; Schwarz, 1978) both favored the model estimating DIF. The deviation was rather small in both cases. Thus, overall, there was no pronounced DIF with regard to gender.

Books: The number of books at home was used as a proxy for socioeconomic status. There were 1,676 (30%) test takers with 0 to 100 books at home and 3,116 (56%) test takers with more than 100 books at home. 827 (15%) test takers had no valid response. There was a considerable average difference between the groups. Participants with 100 or less books at home performed on average 0.147 logits (Cohen's *d* = 0.30) lower in computer literacy than participants with more than 100 books. Participants with 100 or less books at home performed on average 0.188 logits (Cohen's *d* = 0.38) better in computer literacy than participants with no valid answer on the number of books at home. Participants with more than 100 books at home performed on average 0.335 logits (Cohen's *d* = 0.68) better in computer literacy than participants with no valid answer on the number of books at home. However, there was no considerable DIF on the item level. Differences in estimated difficulties did not exceed 0.6 logits. Whereas the AIC favored the model estimating DIF, the BIC favored the main effects model (Table 5).

Migration background: There were 3.355 participants (60%) with no migration background, 1,073 subjects (19%) with a migration background, and 1,191 individuals (21%) that did not indicate their migration background. Participants without migration background had on average a higher computer literacy than subjects with migration background (main effect = 0.137 logits, Cohen's *d* = 0.28) and than participants that did not indicate their migration background (main effect = 0.145 logits, Cohen's *d* = 0.29). Participants with a migration background had on average the same computer literacy then individuals that did not indicate their migration background (main effect = 0.008 logits, Cohen's *d* = 0.02). There was no considerable DIF on the item level. Whereas the AIC favored the model estimating DIF, the BIC favored the main effects model (Table 5). Thus, overall, there was no pronounced DIF with regard to migration background.

Table 5

Differential Item Functioning

DIF variable	Model	N	Deviance	Number of Parameters	AIC	BIC
Sex	main effect	5620	200,875.52	32	200,939.52	201,151.81
	DIF		200,559.92	62	200,683.92	201,095.24
Books	main effect	5620	200,661.13	33	200,727.13	200,946.05
	DIF		200,386.68	93	200,572.68	201,189.65
Migration	main effect	5620	200,795.12	33	200,861.12	201,080.05
	DIF		200,550.86	93	200,736.87	201,353.84

5.3.5 Rasch homogeneity

An essential assumption of the Rasch (1980) model is that all item-discrimination parameters are equal. In order to test this assumption, a two-parameter logistic model (2PL) that estimates discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 6), ranging from 0.03 (item icg3621x_c) to 1.03 (item icg3211x_c). The average discrimination parameter fell at 0.50. Model fit indices suggested a better model fit of the 2PL (AIC = 200,082.35, BIC = 200,480.38, number of parameters = 60) as compared to the Rasch model (AIC = 200,940.89, BIC = 201,146.55, number of parameters = 31). Despite the empirical preference for the 2PL, the Rasch model matches the theoretical conceptions underlying the test construction more adequately (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the Rasch model was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

Table 6

Discriminations in the two-parameter logistic model (2PL)

Position	Item	Discrimination	S.E.
1	icg3052x	0.641	0.04
2	icg3350x	0.308	0.04
3	icg3021x	0.406	0.04
4	icg3610x	0.594	0.04
5	icg3621x	0.025	0.04
6	icg3371x	0.04	0.05
7	icg3081x	0.869	0.05
8	icg3102x	0.735	0.05
9	icg3591x	0.336	0.04
10	icg3092x	0.402	0.04
11	icg3381x	0.485	0.04
12	icg3400x	0.739	0.05
13	icg3661x	0.57	0.04
14	icg3410x	0.644	0.05
15	icg3420x	0.44	0.04
16	icg3432x	0.617	0.04
17	icg3440x	0.513	0.04
18	icg3322x	0.117	0.04
19	icg3461x	0.249	0.04
20	icg3211x	1.031	0.06
21	icg3510x	0.579	0.04
22	icg3221x	0.666	0.04
23	icg3601x	0.827	0.05
24	icg3260x	0.617	0.04
25	icg3301x	0.552	0.05
26	icg3270x	0.348	0.04
27	icg3292x	0.61	0.05
28	icg3481x	0.606	0.04
29	icg3541x	0.135	0.04
30	icg3550x	0.338	0.04

Note. S.E. = Standard error.

5.3.6 Unidimensionality

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

The estimated variances and correlations between the four dimensions representing the different process components are reported in Table 7. The correlations between the dimensions varied between .70 and .87. The smallest correlation was found between Dimension 2 ("Create") and Dimension 4 ("Evaluate"). Dimension 1 ("Access") and Dimension 2 ("Create") showed the strongest correlation. All correlations deviated from a perfect correlation (i.e., they were marginally lower than r = .95, see Carstensen, 2013). According to the BIC, the unidimensional model (AIC = 200,940.89, BIC = 201,146.55, number of parameters = 31) fitted the data slightly better than the four-dimensional model (AIC = 200,905.15, BIC = 201,170.52, number of parameters = 40). According to the AIC, the four-dimensional model fitted better. These results suggest that the items measure one common construct.

Table 7

Results of Four-Dimensional Scaling (Process Components)

	Access	**Create**	**Manage**	**Evaluate**
Access (11 items)	(0.218)			
Create (8 items)	.870	(0.308)		
Manage (6 items)	.864	.823	(0.317)	
Evaluate (5 items)	.745	.702	.817	(0.408)

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

Table 8

Results of Four-Dimensional Scaling (Software Applications).

	System	**Spread sheet**	**Graphics**	**Internet**
Operating System (4 items)	(0.266)			
Text processing / spread sheet (12 items)	.799	(0.289)		
Graphics (4 items)	.762	.774	(0.416)	
Internet / search engines (10 items)	.832	.798	.785	(0.256)

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

The estimated variances and correlations for another four-dimensional model based on the different types of software applications reported in Table 8. The correlations among the three dimensions fell between .76 and .83. The smallest correlation was found between Dimension 1 ("Operating system") and Dimension 3 ("Graphics"). Dimension 1 ("Operating system") and Dimension 4 ("Internet/search engines") showed the strongest correlation. However, they deviated from a perfect correlation (i.e., they were marginally lower than *r* = .95, see Carstensen, 2013). According to BIC, again the unidimensional model (AIC = 200,940.89, BIC = 201,146.55, number of parameters = 31) fitted better than the four-dimensional model (AIC = 200,901.02, BIC = 201,166.39, number of parameters = 40) indicating that all items measure one common construct.

For the unidimensional model the average absolute residual correlations as indicated by the corrected Q3 statistic (see Table 8) were quite low (M = .018, SD = .005) — the largest individual residual correlation was .034 — and thus indicated an essentially unidimensional test. Because the computer literacy test is constructed to measure a single dimension, a unidimensional computer literacy competence score was estimated.

6 Discussion

The analyses in the previous sections aimed at providing detailed information on the quality of the computer literacy test in starting cohort 2 for grade 3 and at describing how computer literacy was estimated.

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

Various criteria indicated a good fit of the items and measurement invariance across various subgroups. However, the amount of not-reached items was rather high (but still lower than in computer literacy tests for other cohorts), indicating that the test was too long for the allocated testing time for at least a part of the respondents. Other types of missing responses were reasonably small.

The test targeted at population with mainly basic competence levels. As a consequence, the test had a mediocre reliability and showed a limited variance.

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

7 Data in the Scientific Use File

7.1 Naming conventions

The data in the Scientific Use File contain data for 30 MC items, with 0 indicating an incorrect response and 1 indicating a correct response. MC items are marked as usual with a 'x_c' at the end of the variable name.

7.2 Computer literacy scores

Person abilities were estimated using the item difficulty parameters. In the SUF, manifest scale scores are provided in the form of a WLE estimates, ”icg3_sc1” and its standard errors “icg3_sc2”. The score can be used, if the research interest lies on cross-sectional issues. The ConQuest Syntax for estimating the WLE is provided in Appendix A. For persons who either did not take part in the computer literacy test or who did not give enough valid responses, no WLE is estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values.

Users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values. A description of these approaches can be found in Pohl and Carstensen (2012).

References

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

Carstensen, C. H. (2013). Linking PISA competencies over three cycles – results from Germany. In M. Prenzel, M. Kobarg, K. Schöps & S. Rönnebeck (Eds.). *Research Outcomes of the PISA Research Conference 2009*. New York, NY: Springer.

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

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2016). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

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

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

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

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

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

Senkbeil, M., Ihme, J. M., & Wittwer, J. (2013). The test of technological and information literacy (TILT) in the National Educational Panel Study: Development, empirical testing, and evidence for validity. *Journal for Educational Research Online, 5,* 139–161.

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

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

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

Appendix

Appendix A: ConQuest-Syntax for estimating WLE estimates in Starting Cohort 2 (grade 3)

```
title SC2 G3 Computer Literacy rasch model;
/* load data */
datafile  >>filename.dat;
format pid 1-7 responses 9-38;
labels <<filename_with_labels.txt;

/* scoring */
codes 0,1;
score (0,1) (0,1) ! items (1-30);

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

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

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

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

No.	Item	Process Component	Software Application
1	icg3052X	Access	Text
2	icg3350X	Access	System
3	icg3021X	Manage	System
4	icg3610X	Manage	Text
5	icg3621X	Create	Text
6	icg3371X	Create	Text
7	icg3081X	Create	Text
8	icg3102X	Create	Text
9	icg3591X	Access	Text
10	icg3092X	Evaluate	Text
11	icg3381X	Evaluate	Text
12	icg3400X	Create	Text
13	icg3661X	Access	Text
14	icg3410X	Create	Paint
15	icg3420X	Create	Paint
16	icg3432X	Create	Paint
17	icg3440X	Access	Paint
18	icg3322X	Access	Internet
19	icg3461X	Access	Internet
20	icg3211X	Access	Internet
21	icg3510X	Access	System
22	icg3221X	Manage	Internet
23	icg3601X	Evaluate	Internet
24	icg3260X	Manage	Tabellen
25	icg3301X	Access	System
26	icg3270X	Manage	Internet
27	icg3292X	Manage	Internet
28	icg3481X	Evaluate	Internet
29	icg3541X	Access	Internet
30	icg3550X	Evaluate	Internet