NEPS
National Educational Panel Study

NEPS Working Papers

Katinka Hardt, Steffi Pohl, Kerstin Haberkorn, & Elena Wiegand

NEPS Technical Report for Reading – Scaling Results of Starting Cohort 6 for Adults in Main Study 2010/11

NEPS Working Paper No. 25

Bamberg, April 2013

SPONSORED BY THE
Federal Ministry
of Education
and Research

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

NEPS Technical Report for Reading – Scaling Results of Starting Cohort 6 for Adults in Main Study 2010/11

Katinka Hardt[1], Steffi Pohl[1], Kerstin Haberkorn[1], & Elena Wiegand[2]

[1]University of Bamberg, National Educational Panel Study
[2]University of Mannheim

Email address of the lead author:

katinkahardt@yahoo.de

Bibliographic data:

Hardt, K., Pohl, S., Haberkorn, K., & Wiegand, E. (2013). *NEPS Technical Report for Reading – Scaling results of Starting Cohort 6 for adults in main study 2010/11* (NEPS Working Paper No. 25). Bamberg: University of Bamberg, National Educational Panel Study.

We thank Rebecca Stammberger, Theresa Rohm, Michael Wenzler, and Maike Krannich for their assistance in scaling the data.

NEPS Technical Report for Reading – Scaling Results of Starting Cohort 6 for the Adults in Main Study 2010/11

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competencies across the whole life span and tests for assessing the different competence domains are developed. In order to evaluate the quality of the competence tests, a wide range of analyses have been performed based on Item Response Theory (IRT). This paper describes the data and scaling procedures of the adult reading competence test in starting cohort 6. After reporting descriptive statistics of the data, the scaling model applied to estimate competence scores, analyses performed to investigate the quality of the scale, as well as the results of these analyses are presented. The reading competence test for the adults' cohort consisted of 32 items, which represented different cognitive requirements and text functions and used different response formats. The test was administered to 5,349 persons. A partial credit model was used for scaling the data. Item fit statistics, differential item functioning, Rasch-homogeneity, the tests' dimensionality, and local item independence were evaluated to ensure the quality of the test. The results showed that the test exhibits a high reliability and that the items fit the model. Moreover, measurement invariance could be confirmed for various subgroups. Dimensionality analyses showed that the different cognitive requirements foster a unidimensional construct, while there is some evidence for multidimensionality based on text functions. It is to note that there is a considerable amount of items that have not been reached by the test takers within assessment time and that there are many items that are targeted towards a lower reading ability. Altogether, the results show good psychometric properties of the reading competence test and support the estimation of a reliable reading competence score. In addition to scaling results, the data available in the Scientific Use File are described and the ConQuest-Syntax for scaling the data is provided.

Keywords

item response theory, scaling, reading competence, scientific use file, adults

Content

1. Introduction

Within the National Educational Panel Study (NEPS) different competencies are measured coherently across the life span and tests have been developed for different competence domains. These include, among others, reading competence, mathematical competence, scientific literacy, and information and communication technologies literacy. Weinert et al. (2011) give an overview of the competencies measured in NEPS.

Most of the competence data are scaled using models that are based on Item Response Theory (IRT). Since most of the competence tests were developed specifically for implementation in NEPS, several analyses have been 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 scales are described in Pohl and Carstensen (2012). In this paper the results of these analyses are presented for reading competence in the second wave of starting cohort 6 (adults). We will first introduce the main concepts of the reading competence test. Then, we will describe the reading competence data of starting cohort 6 and the analyses performed to estimate competence scores and to check the quality of the test. The results of these analyses will be presented and discussed. Finally, we will describe the data that are available for public use in the Scientific Use File.

Please note that the analyses in this report are based on the data set available at some time before data release. Due to data protection and data cleaning issues, the data set in the Scientific Use File (SUF) may differ slightly from the data set used for analyses in this paper. We do not, however, expect major changes in the results.

2. Testing reading competence

The framework and test development for the reading competence test are described in Weinert et al. (2011) and Gehrer, Zimmermann, Artelt, and Weinert (2012). In the following, we will point out specific aspects of the reading competence test that are necessary for understanding the scaling results presented in this paper.

The reading test consists of five texts and a number of items referring to one of the five texts. Each of these texts represents one text type or text function, namely, 1. information texts, 2. commenting or argumenting texts, 3. literary texts, 4. instruction texts, and 5. advertising texts. The test aims at assessing three cognitive requirements. These are a) finding information in the text, b) drawing text-related conclusions, and c) reflecting and assessing. The cognitive requirements do not depend on the text type but each cognitive requirement is usually assessed within each text type (see Gehrer et al., 2012, and Weinert et al., 2011, for a detailed description of the framework).

In the reading competence test there are three types of response formats: simple multiple choice (MC) items, complex multiple choice (CMC) items, and matching (MA) items. In MC items there are four response options, of which one option is correct, while the other three response options function as distractors (i. e., they are incorrect). In CMC items a number of subtasks with two response options are given. MA items require the test taker to match a number of responses to a given set of statements. MA items are usually used to assign headings to paragraphs of a text. Examples of the different response formats are provided in Pohl and Carstensen (2012).

3. Data

3.1 The design of the study

In the main study 2010/11, reading speed, reading competence, mathematical competence as well as procedural metacognition were assessed. In order to investigate the effects of test duration and to control for position and order effects, the tests were administered to participants in different selection and sequence. For this purpose, the sample was split into four groups receiving different test booklets (see Figure 1). Assignment to test booklets was random. Reading speed and procedural metacognition were assessed of all participants. In order to assess the effects of test duration, half of the sample additionally received both the reading and mathematics test, while the other half received only one of these two competence tests. The sample receiving only one of the two tests was split in two groups. In one group reading competence and in the other group mathematical competence was assessed. In order to control for position and order effects in the group receiving both tests, the two tests were assigned to the participants in different order. One testing group first completed the reading test followed by the mathematic test, while the other group completed the two tests in the opposite order. Note that there was no multi-matrix design regarding the choice and the order of the items *within* a specific test. All subjects received the same set of reading items in the same order.

Booklet 1	**Booklet 2**	**Booklet 3**	**Booklet 4**
Reading speed	Reading speed	Reading speed	Reading speed
Reading (+ meta-p)	Math (+ meta-p)	Reading (+ meta-p)	Math (+ meta-p)
Math (+ meta-p)	Reading (+ meta-p)		

Figure 1: Design of the study. Reading – reading competence, Math – mathematical competence, meta-p – procedural metacognition for the respective competence

The adults' reading test consisted of 32 items which represented different cognitive requirements and text functions and featured different response formats. Prior to the final scaling, extensive analyses were conducted in order to evaluate the quality of the items. Two items showed an unsatisfactory item fit and were excluded from the final scaling procedure. The scaling results presented in the following sections are, thus, based on 30 items. The characteristics of these items are described in Tables 1 to 3. Table 1 shows the distribution of the cognitive requirements, Table 2 the distribution of text functions, and Table 3 the response formats used. The number of subtasks within CMC and MA items varied between two and six.

Table 1: Cognitive requirements of the items in the reading test for adults

Cognitive requirement	**Frequency**
Finding information in text	13
Drawing text-related conclusions	8
Reflecting and assessing	9
Total number of items	30

Table 2: Number of items for the different text types in the reading test for adults

Text types/functions	Frequency
Information texts	6
Instruction texts	6
Advertising texts	5
Commenting or argumenting texts	8
Literary texts	5
Total number of items	30

Table 3: Response formats of the items in the reading test for adults

Response format	Frequency
Simple multiple choice	23
Complex multiple choice	4
Matching	3
Total number of items	30

3.2 Sample

A description of the design of the study, the sample, as well as the instruments used can be found on the NEPS-website[1]. In total, 5,349 subjects took the reading competence test[2]. 1,717 of them received the mathematic test followed by the reading test; 1,767 subjects first completed the reading competence test before the mathematic competence test. 1,859 persons received a test booklet including the reading test only. For 6 subjects no valid information on the booklet indicator was available.

14 of the 5,349 subjects, who took the reading test, had less than three valid responses to the reading items. Since no reliable reading competence score may be estimated based on such a low number of valid responses, these cases were excluded from further analyses. Thus, a sample of 5,335 persons underlies the results presented in the following sections.

4. Analyses

4.1 Missing responses

There are different types of missing responses in competence test data. These are missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach, d) items that have not been administered, and e) multiple kinds of missing responses within CMC or MA items that are not determinable. As described in the study design, there are 1,799 persons that did not receive a reading test. The responses of these persons to the reading items were coded as not administered. Invalid responses occurred, for example, when two response options were selected in simple MC items where just one was required, or when numbers or letters that are not within the range of valid responses were given as a response. Items were coded as omitted when subjects skipped a particular item. Due to the

[1] www.neps-data.de

[2] Note that these numbers may differ from those found in the SUF. This is due to still ongoing data protection and data cleaning issues.

limited testing time, some subjects did not complete the entire test. Items that were not completed at the end of the test were labeled as not-reached. Since CMC and MA items consist of a number of subtasks, a mixture of different types of missing responses and/or a mixture of missing and valid responses might be found. When one subtask contained a missing response, the CMC or MA item was coded as missing. When just one kind of missing response occurred, the item was coded according to the corresponding missing response. When the subtasks contained different kinds of missing responses, the item value was coded as a not-determinable missing response.

Missing responses provide information on how well the test worked (e. g., time limits, understanding of instructions, handling of different response formats), and they need to be accounted for in the estimation of item and person parameters. We thoroughly inspected 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 test persons were coping with the test. We then examined the occurrence of missing responses per item in order to obtain some information on how well the items performed.

4.2 Scaling model

In order to estimate item and person parameters, a partial credit model (Masters, 1982) was used and estimated in ConQuest (Wu, Adams, & Wilson, 1997). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

CMC and MA items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC or MA item, indicating the number of correctly responded subtasks within that item. If at least one of the subtasks contained a missing response, the whole CMC item or MA item was scored as missing. When categories of the polytomous variables had less than N = 200, the categories were collapsed in order to avoid possible estimation problems. This usually occurred in 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. Small frequencies of categories also occurred for matching tasks with perfect local dependence. In these cases the two highest scores were collapsed into one category (see Pohl & Carstensen, 2012 for the explanation of this approach). For six of the seven CMC and MA items, categories were collapsed. Note here that, as a consequence, the values of the polytomously scored CMC and MA items in the Scientific Use File do not necessarily indicate the number of correctly solved subtasks but should rather be interpreted as (partial) credit scores.

To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items were scored dichotomously as 0 for an incorrect and as 1 for the correct response (see Haberkorn, Pohl, Carstensen, & Wiegand, 2012; and Pohl & Carstensen, 2013, for studies on the scoring of different response formats).

Ability estimates for reading competence were estimated as weighted maximum likelihood estimates (WLEs; Warm, 1989) and will later also be provided in form of plausible values (Mislevy, 1991). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF are described in section 7.

4.3 Checking the quality of the test

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

Before aggregating the responses to the subtasks of CMC and MA items to a polytomous variable, the aggregation was justified by preliminary analyses. For this purpose, the subtasks were included separately in a Rasch model (Rasch, 1960) together with the MC items, and the fit of the subtasks was evaluated based on the weighted mean square error (WMNSQ), the respective t-value, point biserial correlations of the correct responses with the total score, and the item characteristic curve. Only if the subtasks exhibited a satisfactory item fit, they were used to generate polytomous variables that were then included in the final scaling model.

The MC, CMC and MA items consisted of one correct response and one or more distractors (incorrect response options). We investigated the performance of distractors, that is, whether they were predominantly chosen by subjects with a lower ability rather than by those who gave a correct response. We evaluated the point biserial correlation between the incorrect responses and the total score treating all subtasks of CMC and MA 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.

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

We aim at constructing a reading competence test that measures the same construct for all participants. If there were any items that favored certain subgroups (e. g., that 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. We addressed the issue of measurement invariance by investigating test fairness for the variables test position, gender, school degree, and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Differential item functioning was 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. Differences in the estimated item difficulties between the subgroups were evaluated. Based on experiences with preliminary data, we judged absolute differences in estimated difficulties that are greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as noteworthy for further investigation, differences between 0.4 and 0.6 as considerable but not sincerely, and differences smaller than 0.4 as no considerable DIF. In addition to DIF analyses on item level, test fairness was investigated by comparing a model including differential item functioning to a model that only estimated main effects and no DIF.

The reading competence data in NEPS were scaled using the partial credit model (1PL), which assumes Rasch-homogeneity. The partial credit model was chosen because it preserves the weighting of the different aspects of the framework as intended by test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that may not hold for empirical data. We therefore checked for deviations from a uniform discrimination. We estimated item discrimination applying the generalized partial credit model (2PL) (Muraki, 1992) using the software mdltm (von Davier, 2005) and compared model fit indices of the 2PL model to those obtained when applying the partial credit model.

Additionally, we evaluated the dimensionality of the reading test by conducting several multidimensional analyses. The different subdimensions of the multidimensional models were specified based on different construction criteria. First, a model with three different subdimensions representing the three cognitive requirements, and, second, a model with five different subdimensions based on the five text functions 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.

Since the reading test consisted of item sets that referred to one of five texts, the assumption of local item dependence (LID) may not necessarily hold. However, the five texts were perfectly confounded with the five text functions. Thus, multidimensionality and local item dependence may not be evaluated separately with these data. We referred to preliminary studies on reading competence to disentangle the amount of multidimensionality and local item dependence.

5. Results

5.1 Missing responses

Missing responses per person

Figure 2 depicts the number of invalid responses per person. As can be seen, with 85.23%, the vast majority of the respondents did not have any invalid response at all and less than five percent had more than one invalid response.

Missing responses may also occur when respondents omit items. As can be seen in Figure 3, the majority of the subjects – almost 58 percent – did not skip an item at all and only about five percent omitted more than four items of the reading test.

Another source of missing responses are items that were not reached by the subjects; per definition, these are all missing responses after the last valid response. The number of not-reached items was rather high (see Figure 4). With 41.82%, less than half of the participants were able to finish the reading competence test within time. Almost 40% of the subjects did not reach the last text and around 14% did not reach the items of the last two of the five texts.

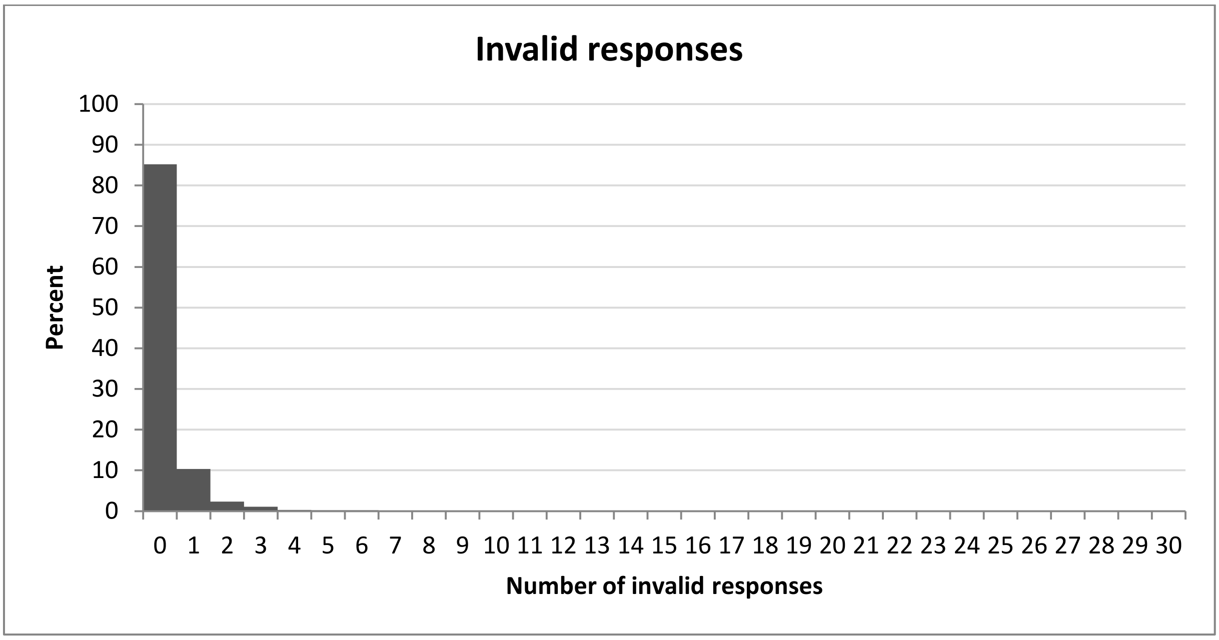

Figure 2: Number of invalid responses

The aggregated polytomous variables were coded as not-determinable missing response when the subtasks of CMC and MA items contained different kinds of missing responses. Since not-determinable missing responses might only occur in CMC and MA items, the maximum number of not-determinable missing responses was seven (i. e., the number of CMC and MA items). Only a small amount of not-determinable missing responses occurred (Figure 5). 95.5% of the subjects had no non-determinable missing responses and only 1.61% of the persons had a not-determinable missing response to more than one of the items.

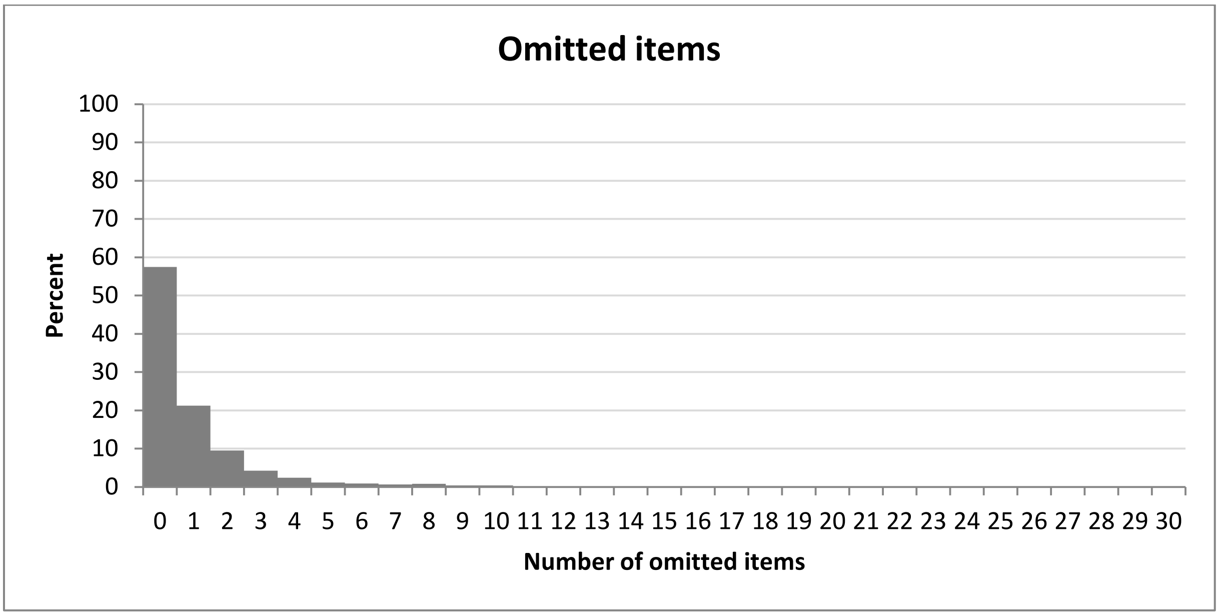

Figure 3: Number of omitted items

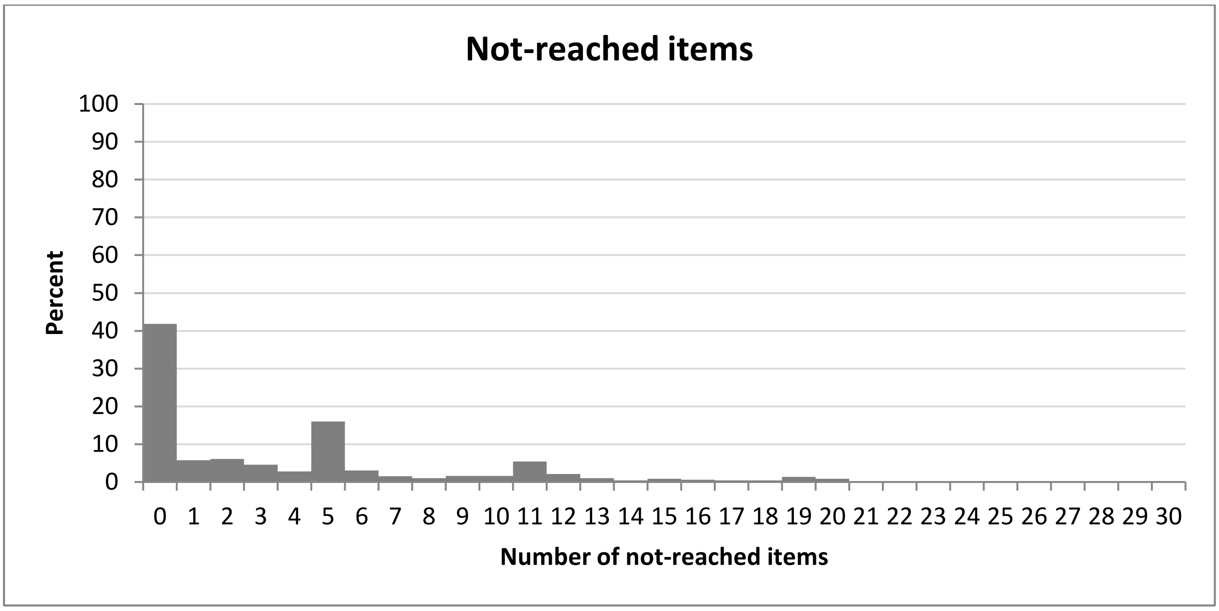

Figure 4: Number of not-reached items

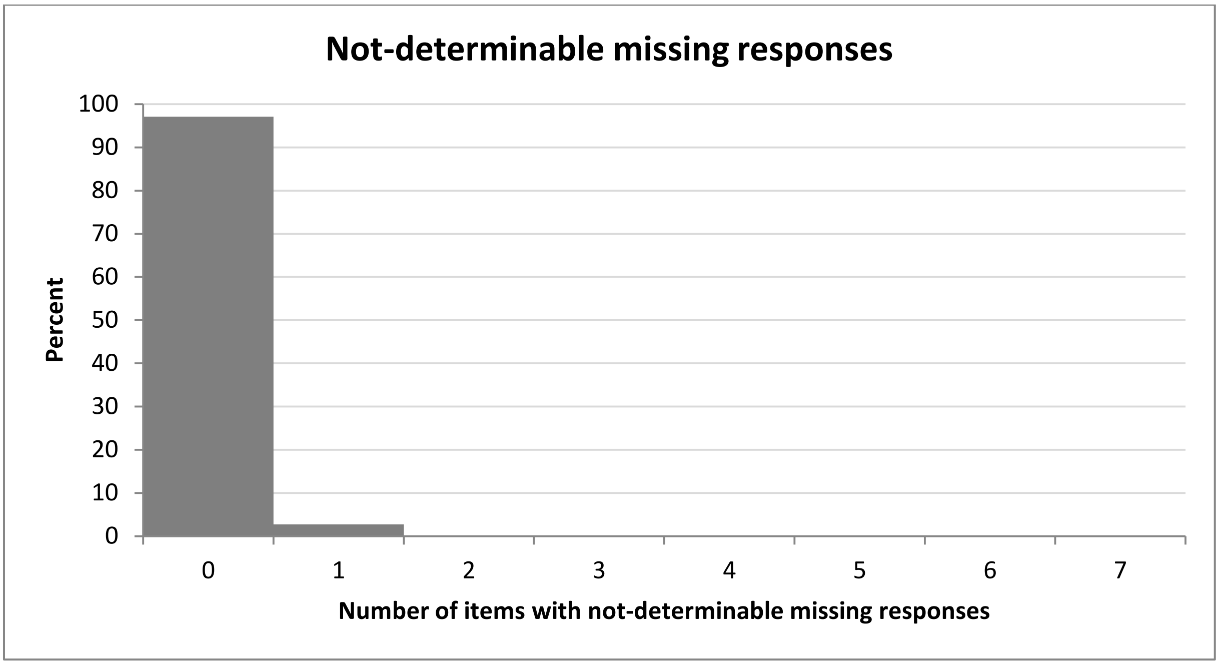

Figure 5: Number of not-determinable missing responses

The total number of missing responses, aggregated over invalid, omitted, not-reached, and not-determinable missing responses per person, is illustrated in Figure 6. On average, the subjects showed 5.15 (SD = 5.46) missing responses. 23.39% of the persons had no missing response at all and about 50% of the participants had four or more missing responses.

In sum, there is a small amount of invalid and not-determinable missing responses and a reasonable amount of omitted items. The number of not-reached items is, however, rather large and has the greatest impact to the total number of missing responses.

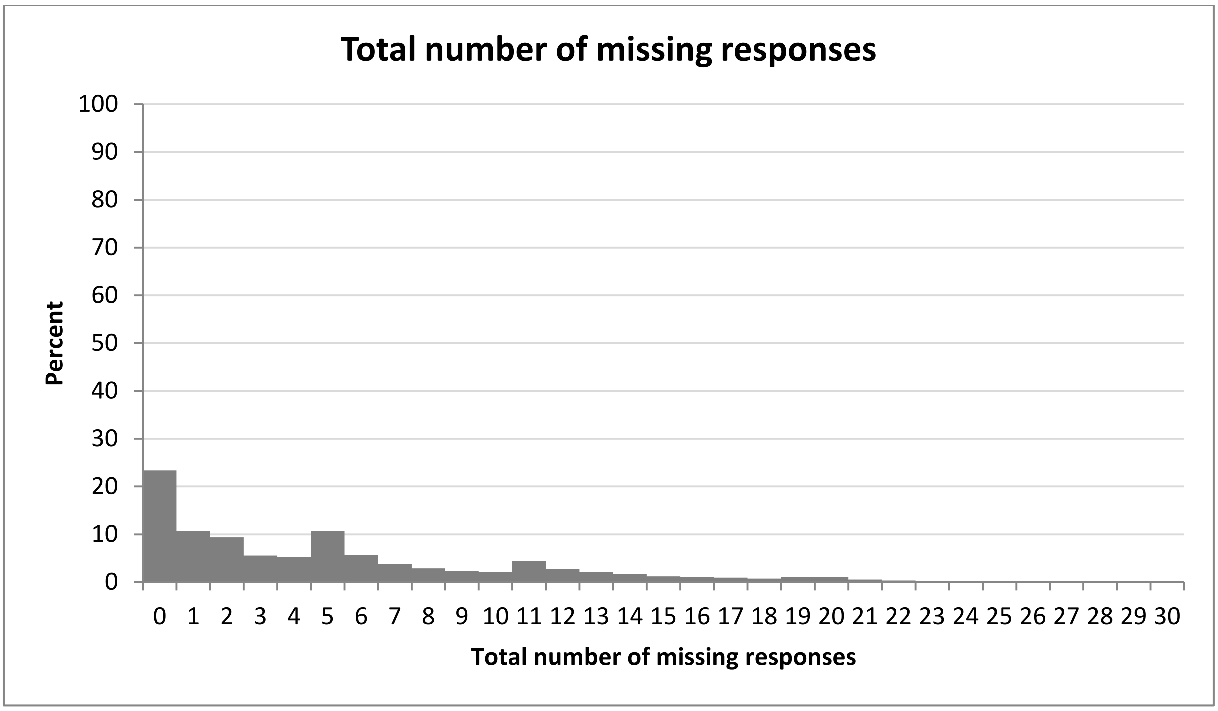

Figure 6: Total number of missing responses

Missing responses per item

Table 4 provides information on the occurrence of different kinds of missing responses per item. Overall, the omission rate is acceptable, varying across items between 0% (rea20110_c) and 13.90% (rea2028s_c). There were seven items with an omission rate exceeding 5%. On average, CMC and MA items had higher omission rates (8.93% and 8.43%, respectively) than MC items (1.84%). Omission rate correlated with item difficulty to .12; the correlation increased to .24 when three items with bivariate extreme values were excluded. Participants are inclined to omit more difficult items. With a proceeding position of an item in the test the amount of persons that did not reach the item (column 4) rose up to a considerable amount of 58.16% (for the last item rea20550_c). On the contrary, the percentage of invalid responses per item (column 5) was rather low (maximum of 3.67% for item rea20140_c). Matching items seemed to be more prone to cause invalid responses than were multiple choice items in both single and complex form.

5.2 Parameter estimates

Item parameters

Column 2 in Table 5 shows the percentage of correct responses relative to all valid responses for each item. Note that since there is a non-negligible amount of missing responses, this probability cannot be interpreted as an index for item difficulty. The percentage of correct responses within MC items varied between 42.95% and 95.37% with an average of 80.11% (SD = 14.51%) correct responses.

Table 4: Missing values

Item	Position in the test	Number of valid responses	Relative frequency of not-reached items in %	Relative frequency of omitted items in %	Relative frequency of invalid responses in %
rea20110_c	1	5,249	0.00	1.11	0.51
rea2012s_c	2	4,658	0.00	12.70	0.00
rea20130_c	3	5,147	0.00	2.38	1.14
rea20140_c	4	5,052	0.00	1.63	3.67
rea2015s_c	5	4,823	0.00	6.30	2.90
rea20210_c	6	5,230	0.13	1.22	0.62
rea20220_c	7	5,138	0.13	1.52	2.04
rea20230_c	8	5,213	0.19	1.80	0.30
rea20240_c	9	5,243	0.21	0.94	0.58
rea20250_c	10	5,175	0.34	1.78	0.88
rea2028s_c	13	4,373	1.20	13.90	1.40
rea20310_c	14	5,010	2.49	2.16	1.44
rea20320_c	15	4,989	2.96	2.16	1.37
rea20330_c	16	4,937	3.39	3.81	0.26
rea20340_c	17	4,823	4.01	5.25	0.34
rea20350_c	18	4,803	4.89	4.85	0.22
rea20360_c	19	4,857	5.32	3.00	0.64
rea20370_c	20	4,794	6.32	3.30	0.52
rea2038s_c	21	4,224	8.49	11.94	0.00
rea20410_c	22	4,523	13.89	1.01	0.32
rea2042s_c	23	4,271	15.50	4.42	0.02
rea20430_c	24	4,366	17.15	0.47	0.54
rea20440_c	25	4,335	18.16	0.52	0.06
rea20450_c	26	4,237	19.74	0.54	0.30
rea20460_c	27	4,019	22.83	1.69	0.15

Item	Position in the test	Number of valid responses	Relative frequency of not-reached items in %	Relative frequency of omitted items in %	Relative frequency of invalid responses in %
rea20510_c	28	3,223	38.86	0.69	0.04
rea2052s_c	29	2,735	41.67	6.71	0.06
rea20530_c	30	2,831	46.24	0.56	0.13
rea2054s_c	31	2,208	52.37	5.12	0.64
rea20550_c	32	2,086	58.16	0.00	2.74

Remarks.
The items on positions 11 and 12 were excluded from the analyses due to unsatisfactory item fit (see section 3.1).

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	Discrimi-nation – 2PL
rea20110_c	95.37	-3.594	0.068	1.00	0.0	0.29	1.05
rea2012s_c	n. a.	-3.169	0.058	0.98	-0.7	0.36	1.10
rea20130_c	94.09	-3.302	0.062	0.98	-0.3	0.33	1.08
rea20140_c	83.75	-2.010	0.042	1.02	0.9	0.41	0.88
rea2015s_c	n. a.	-1.913	0.045	0.95	-2.8	0.46	1.27
rea20210_c	95.07	-3.518	0.067	0.96	-0.8	0.34	1.33
rea20220_c	89.18	-2.553	0.048	0.96	-1.3	0.44	1.27
rea20230_c	92.23	-2.988	0.055	0.97	-0.8	0.39	1.19
rea20240_c	88.16	-2.451	0.046	0.99	-0.2	0.40	0.97
rea20250_c	85.41	-2.162	0.043	0.98	-0.8	0.45	1.07
rea2028s_c	n. a.	-0.801	0.023	0.93	-3.3	0.73	1.26
rea20310_c	71.84	-1.137	0.035	1.13	7.5	0.35	0.53
rea20320_c	83.92	-1.999	0.042	0.96	-1.8	0.49	1.25
rea20330_c	79.93	-1.678	0.039	1.02	1.1	0.43	0.86
rea20340_c	59.34	-0.432	0.033	0.99	-0.8	0.50	0.97
rea20350_c	86.30	-2.192	0.045	1.00	0.1	0.41	0.98
rea20360_c	83.59	-1.954	0.042	0.96	-1.7	0.48	1.25
rea20370_c	71.86	-1.117	0.036	1.07	3.9	0.42	0.70
rea2038s_c	n. a.	-1.044	0.046	0.96	-2.3	0.44	1.15
rea20410_c	56.60	-0.240	0.034	1.14	10.0	0.35	0.53
rea2042s_c	n. a.	-1.136	0.042	0.93	-4.1	0.51	1.34
rea20430_c	80.92	-1.697	0.042	1.07	2.9	0.38	0.71
rea20440_c	91.67	-2.803	0.058	0.90	-2.5	0.47	1.69
rea20450_c	85.60	-2.073	0.047	0.93	-2.4	0.49	1.29
rea20460_c	42.95	0.505	0.036	1.09	5.8	0.38	0.62

Item	Percentage correct	Difficulty/ location parameter	SE (difficulty/location parameter)	WMNSQ	t-value of WMNSQ	Correlation of item score with total score	Discrimi-nation – 2PL
rea20510_c	91.96	-2.750	0.069	1.00	0.0	0.35	0.97
rea2052s_c	n. a.	-2.207	0.092	0.98	-0.8	0.30	1.06
rea20530_c	80.15	-1.522	0.052	1.07	2.2	0.40	0.77
rea2054s_c	n. a.	-0.057	0.060	0.97	-1.2	0.44	1.05
rea20550_c	52.54	0.137	0.050	1.16	7.8	0.35	0.44

Remarks.
Percent correct scores are not informative for polytomous CMC and MA item scores. These are denoted by n. a.
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 by ConQuest).

Table 6: Step parameters (and standard errors) of the polytomous items

Item	Step 1 (SE)	Step 2 (SE)	Step 3 (SE)	Step 4 (SE)	Step 5 (SE)
rea2012s_c	0.038 (0.041)	-0.038			
rea2015s_c	-0.250 (0.032)	0.250			
rea2028s_c	0.148 (0.031)	-0.015 (0.031)	-0.245 (0.033)	0.148 (0.039)	-0.036
rea2038s_c	-0.573 (0.032)	0.573			
rea2042s_c	0.092 (0.035)	-0.092			
rea2052s_c	n. a.				
rea2054s_c	-0.473 (0.044)	0.473			

Remark.
Note that, because item rea2052s_c consists of only two categories, no step parameters are estimated.

For reasons of model identification, in the partial credit model, the mean of the ability distribution was constrained to be zero. The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 5. The step parameters for polytomous variables are depicted in Table 6. The item difficulties ranged from -3.594 (item rea20110_c) to 0.505 (item rea20460_c) logits with an average difficulty of -1.795 logits (SD = 1.094). Altogether, the item difficulties are very low. Owing to the large sample size, the corresponding standard errors of the estimated item difficulties (column 4) are small (SE(β) ≤ 0.092).

Person parameters

Person parameters in NEPS are estimated as WLEs and as plausible values (Pohl & Carstensen, 2012). WLEs will be provided in the first release of the SUF, whereas plausible values will be provided in later releases 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 is presented in Pohl and Carstensen (2012).

Test targeting and reliability

Test targeting focuses on the match of item difficulties and person abilities and was used to evaluate the appropriateness of the test for the specific target group. In Figure 7, item difficulties of the reading 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 while the right side shows the distribution of item difficulties.

The mean of the ability distribution was constrained to be zero and the variance was estimated to be 1.390, which implies good differentiation between the subjects. The reliability of the test (EAP/PV reliability = .774, WLE reliability = .717) was good. Although the items covered a wide range of the ability distribution, the items were slightly too easy. As a consequence, person ability in medium and low ability regions will be measured relative precisely, while higher ability estimates will have larger standard errors of measurement.

5.3 Quality of the test

Fit of the subtasks of complex multiple choice and matching items

Before the subtasks of CMC and MA items were aggregated to be analyzed via the 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 and MA items separately, there were 48 items. Since there were two matching tasks with perfect stochastic dependence (see Pohl & Carstensen, 2013, for a description of the problem), one of the subtasks of each of these MA items was excluded from the analyses. Consequently, 46 items entered the analysis.

Scale (in logits)	Person ability	Item difficulty
3		
	X	
	X	
	XX	
	X	
2	XX	
	XXX	
	XXX	
	XXX	
	XXXX	
	XXXXXX	
1	XXXXXXX	
	XXXX	
	XXXXXXXX	
	XXXXXXX	27
	XXXXXXXX	
	XXXXXXXXXX	
	XXXXXXX	32
0	XXXXXXXX	31
	XXXXXXXX	22
	XXXXXXXXX	17
	XXXXXXXX	
	XXXXXXX	
	XXXXXXX	13
-1	XXXXX	21
	XXXX	14 20 23
	XXXX	
	XXXXX	30
	XXX	16 24
	XX	
-2	XX	4 5 15 19
	XX	10 18 26
	X	29
	X	9
	X	7
	X	25 28
	X	
-3		8
		2
		3
		1 6
-4		

Figure 7: Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph, each 'X' represents 34.6 cases. The difficulty of the items is depicted on the right side of the graph, each number represents an item (which corresponds to the item position indicated in Table 4).

Concerning relative frequencies, only one subtask of a CMC item showed a probability of a correct response of greater than 95%. Despite the small variance for this item, no estimation problems occurred. Overall, the fit of the subtasks was satisfactory. The WMNSQ ranged from 0.84 to 1.14, the corresponding t-values from -12.70 to 7.80. The good item fit of the subtasks was affirmed by the empirically estimated item characteristic curves. In conclusion, the good fit of the subtasks was considered to justify their aggregation to polytomous variables for each CMC and MA item. Note that CMC and MA items can be identified through the letters 's_c' at the end of the variable name, whereas the variable name of simple MC items ends on '0_c'.

Distractor analyses

In addition to the overall item fit (section 5.3.3), we specifically investigated how well the distractors performed in the test by evaluating the point biserial correlation between each incorrect response (distractor) and the test takers' total score. The distractors consistently yielded correlations below zero with a range from -.480 to -.050 and a mean of -.201. The results indicate that the distractors function properly.

Item fit

Item fit was additionally investigated for MC and polytomous CMC and MA items. Altogether, item fit can be considered as very good (see Table 5). Values of the WMNSQ ranged from 0.90 (item rea20440_c) to 1.16 (rea20550_c), only three t-values of the WMNSQ exceeded a t-value of 7. There is no indication of severe item over- and even less of item underfit. Point biserial correlations between the item scores and the total scores ranged from .29 (item rea20110_c) to .52 (item rea2028s_c) and had a mean of .403. All item characteristic curves showed a good fit of the items.

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 test position, gender, school degree, the number of books at home as a proxy for the socioeconomic status, as well as migration background (see Pohl & Carstensen, 2012, for a description of these variables). Table 8 provides a summary of the results of the DIF-analyses. The table depicts the differences in the estimated item difficulties between the respective groups. "Male vs. female", for example, indicates the difference in difficulty *ß*(male) - *ß*(female). A positive value indicates a higher difficulty for males, a negative value a lower difficulty for males as opposed to females.

The reading test was randomly administered to the subjects in two positions but in three different test sequences (booklets) (see section 1.1 for the design of the study). 1,758 (33.08%) subjects first took the mathematics test before the reading test, to 1,709 (32.15%) subjects the two competence tests were presented in reverse order and 1,848 (34.77%) participants received the reading competence test only. For the remaining 20 cases[3], no information on the booklet variable was available. DIF was investigated for the test

[3] Note that this number of missing information on the booklet variable slightly differs from the amount reported in section 3.2. This is due to issues concerning data delivery and ongoing data cleaning at the time of data analysis.

sequence. Overall, there are merely small average effects of the test sequence. Subjects who first received the mathematics test and then the reading test performed on average 0.200 logits (Cohen's d = 0.171) better than subjects who took the reading competence test before the mathematics test. Note that this main effect does not indicate a threat to measurement invariance. Instead, it may be an indication of habituation effects that are similar for all items. As expected, there is on average no noticeable difference between the two design groups who both received the reading test in the first position (0.046 logits, Cohen's d = 0.039). Whether the reading test is followed by the mathematics test does not have an impact on the average reading score. Differential item functioning with regard to the position of the test may occur for instance due to item specific fatigue effects or due to item specific habituation to the testing mode. No difference in the estimated item difficulties between the different design groups exceeded 0.6 logits. The largest absolute difference in difficulties was -0.447 logits (item rea2052s_c).

Differential item functioning analysis for gender was based on 2,656 (49.86%) males and 2,671 (50.14%) females. For 8 cases information on gender was missing; these cases were excluded from the DIF analysis. On average, male participants had a lower estimated reading ability than females (main effect = -0.132 logits, Cohen's d = -0.112). There was no considerable item DIF. Only two items (item rea20510_c and item rea2052s_c) showed DIF greater than 0.4 logits.

Finally, DIF was investigated for school degree. 2,805 subjects (54.34%) who took the reading test held a high school degree (Abitur) and 2,357 (45.66%) had a lower school degree. 173 subjects had a missing response on school degree; these persons were excluded from the DIF analysis. Subjects who had obtained a high school degree had on average a higher reading ability (1.258 logits, Cohen's d = 1.271) than subjects with a lower school degree. There was no considerable item DIF. No item exhibited DIF greater than 0.6 logits. Six items showed DIF greater than 0.4 logits.

In order to examine differential item functioning for socioeconomic status, the number of books at home was dichotomized into the categories of ≤100 books at home (N = 1,960/36.81%) and >100 books at home (N = 3,365/63.19%). For 10 subjects no valid responses were available on the variable indicating the number of books at home. These cases were excluded from the DIF analysis. On average, test takers with a high socioeconomic status performed 0.838 logits (Cohen's d = 0.757) better on the reading test than subjects with a low socioeconomic status. There was no considerable item DIF, no item had DIF greater than 0.6 logits. For two items (rea20320_c and rea20550_c), differential item functioning greater than 0.4 logits was found.

Finally, test fairness was investigated for migration background. There were 4,283 participants (84.64%) with no migration background and 777 subjects (15.36%) with a migration background. 275 subjects were excluded from the DIF analysis due to missing responses to the involved variables. In comparison to subjects with migration background, participants without migration background had on average a slightly higher reading ability (main effect = 0.252 logits, Cohen's d = 0.215). There was no considerable DIF due to migration background. Differences in estimated difficulties did not exceed 0.6 logits. Only one item (item rea20530_c) exhibited a higher estimated difficulty for subjects with migration background than for subjects without (absolute DIF = -0.446).

The results of the comparison of models including only main effects with models additionally allowing for DIF are displayed in Table 7. Regarding Akaike's (1974) information criterion (AIC), the more parsimonious model including only main effects is preferred for the variables booklet and migration background. The Bayesian information criterion (BIC; Schwarz, 1978) takes the number of estimated parameters into account and, thus, prevents from overparameterization of models. Using BIC, the more parsimonious model including only the main effect was preferred over the more complex DIF model for all DIF variables, except for school degree.

Table 7: Comparison of models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Booklet	main effect	132413.520	42	132497.520	132773.808
	DIF	132308.890	102	132512.890	133183.875
Gender	main effect	132659.800	41	132741.800	133011.603
	DIF	132474.365	71	132616.365	133083.583
School degree	main effect	127192.766	41	127274.766	127543.279
	DIF	126877.747	71	127019.747	127484.732
Books	main effect	132080.708	41	132162.708	132432.495
	DIF	131932.970	71	132074.970	132542.162
Migration	main effect	125813.477	41	125895.477	126163.171
	DIF	125778.134	71	125920.134	126383.702

Summarizing the results of DIF examination, neither strong nor noteworthy DIF was found (all absolute DIF < 0.6 logits). There was no indication for test unfairness.

Rasch-homogeneity

One essential assumption of the Rasch model is Rasch-homogeneity. Rasch-homogeneity implies that all item discrimination parameters are equal. In order to test this assumption, a generalized partial credit model (2PL) that estimates discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 5), ranging from 0.439 (item rea20550_c) to 1.685 (item rea20440_c). Model fit indices suggested a better model fit of the 2PL model (AIC = 132150.72, BIC = 132683.86, number of parameters = 81) as compared to the 1PL model (AIC = 133008.06, BIC = 133343.74, number of parameters = 51). Despite the empirical preference for the 2PL model, the 1PL model more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the partial credit model (1PL) was chosen as scaling model to preserve the weighting of items as intended in the theoretical framework.

Unidimensionality and local item independence

The unidimensionality of the test was investigated by specifying two different multidimensional models and comparing them to a unidimensional model. In the first multidimensional model three different cognitive requirements were specified, while the five different text types constituted the second multidimensional model.

Table 8: Differential item functioning (absolute differences between difficulties)

Item	**Booklet**			**Gender**	**Books**	**Migration status**	**School degree**
	Math/read vs. read/math	Math/read vs. read only	Read/math vs. read only	Male vs. female	0-100 vs. >100	Without vs. with	Lower degree vs. high school degree
rea20110_c	0.045	-0.003	-0.048	0.260	0.272	0.072	0.074
rea2012s_c	0.076	0.055	-0.021	-0.052	0.076	-0.150	0.024
rea20130_c	-0.203	-0.271	-0.068	-0.158	-0.160	-0.276	-0.080
rea20140_c	-0.219	-0.126	0.093	0.264	0.082	0.196	-0.120
rea2015s_c	-0.379	-0.356	0.023	-0.232	0.036	-0.082	0.362
rea20210_c	0.200	0.277	0.077	0.346	0.108	-0.072	0.060
rea20220_c	-0.387	-0.178	0.209	0.206	-0.004	-0.170	-0.026
rea20230_c	-0.104	-0.034	0.070	0.194	0.020	0.132	-0.060
rea20240_c	0.019	-0.039	-0.058	-0.202	-0.010	-0.202	-0.202
rea20250_c	-0.154	-0.213	-0.059	0.386	-0.108	0.120	-0.256
rea2028s_c	-0.194	-0.220	-0.026	0.118	0.086	-0.008	0.178
rea20310_c	0.064	0.045	-0.019	-0.154	-0.236	0.100	-0.392
rea20320_c	-0.239	-0.210	0.029	0.174	0.422	-0.080	0.234
rea20330_c	-0.014	-0.212	-0.198	0.148	0.058	0.000	-0.134
rea20340_c	-0.030	0.067	0.097	-0.176	0.244	-0.010	0.404
rea20350_c	-0.009	-0.167	-0.158	0.080	0.300	-0.138	0.094
rea20360_c	-0.035	-0.061	-0.026	0.360	0.128	0.290	0.388
rea20370_c	0.147	0.129	-0.018	0.052	-0.252	-0.094	-0.434
rea2038s_c	-0.064	-0.059	0.005	-0.004	0.244	0.036	0.592
rea20410_c	0.150	0.182	0.032	-0.154	-0.306	0.052	-0.386
rea2042s_c	-0.169	-0.023	0.146	-0.352	0.026	0.102	0.402
rea20430_c	-0.184	-0.338	-0.154	-0.270	-0.240	0.116	-0.272
rea20440_c	0.048	0.013	-0.035	-0.278	0.202	-0.044	0.360

Item	Booklet			Gender	Books	Migration status	School degree
	Math/read vs. read/math	Math/read vs. read only	Read/math vs. read only	Male vs. female	0-100 vs. >100	Without vs. with	Lower degree vs. high school degree
rea20450_c	-0.066	0.081	0.147	-0.218	0.098	-0.040	-0.014
rea20460_c	0.059	0.149	0.090	-0.312	-0.186	0.114	0.162
rea20510_c	0.303	0.255	-0.048	0.468	-0.072	0.108	-0.076
rea2052s_c	0.349	-0.098	-0.447	0.496	-0.116	-0.116	-0.074
rea20530_c	0.069	0.113	0.044	0.096	0.044	-0.446	-0.500
rea2054s_c	0.051	0.072	0.021	-0.148	-0.016	-0.198	-0.180
rea20550_c	0.157	0.175	0.018	-0.110	-0.492	0.036	-0.434
Main effect	0.200	0.154	-0.046	-0.132	-0.838	0.254	-1.258

Table 9: Results of three-dimensional scaling. Variances of the dimensions are depicted in the diagonal, correlations are given in the off-diagonal.

	Dim 1	Dim 2	Dim 3
Finding information in the text (Dim 1) (Nitems = 13)	1.601		
Drawing text-related conclusions (Dim 2) (Nitems = 8)	0.948	1.347	
Reflecting and assessing (Dim 3) (Nitems = 9)	0.942	0.951	1.410

Table 10: Results of five-dimensional scaling. Variances of the dimensions are depicted in the diagonal, correlations are given in the off-diagonal.

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Advertising texts (Dim 1) (Nitems = 5)	2.158				
Instruction texts (Dim 2) (Nitems = 6)	0.930	2.240			
Commenting function (Dim 3) (Nitems = 8)	0.832	0.866	1.478		
Communication (Dim 4) (Nitems = 6)	0.905	0.925	0.895	1.272	
Literary function (Dim 5) (Nitems = 5)	0.780	0.830	0.838	0.833	1.367

Estimation of the three dimensional model was done in ConQuest using the Gauss-Hermite quadrature method. The estimated variances and correlations between the three dimensions that represent the different cognitive requirements are reported in Table 9. All three dimensions had substantial variance estimates with the highest obtained for “finding information in the text” and the lowest for “drawing text-related conclusions”. Intercorrelations among the three dimensions were high (all > .94), supporting the unidimensionality of the test (see Carstensen, 2013). Nonetheless, according to model fit indices, the three-dimensional model fitted the data better (AIC = 132959.79, BIC = 133255.98, number of parameters = 45) than the unidimensional model (AIC = 132998.25, BIC = 133261.53, number of parameters = 40). This may, however, also be a result of the large sample size. From the results we conclude that the three cognitive requirements do not measure different constructs but a unidimensional construct.

The five dimensional model based on the five text functions was estimated using the Monte Carlo estimation algorithm implemented in ConQuest. Estimated variances and correlations are given in Table 10. The estimated variances differed between the three dimensions. Especially the texts located at the end of the booklet had smaller variance estimates. This may be a consequence of the fact that the items constituting these dimensions were not reached by large percentages of the test takers. Correlations between the dimensions varied between r = .780 and r = .930. The lowest correlation was found between dimension 1 (“advertising texts”) and dimension 5 (“literary function”). Dimension 1 and dimension 2 (“instruction texts”) showed the strongest correlation. All correlations deviated from a perfect correlation (i. e., they were considerably lower than r = .95, see Carstensen, 2013). Moreover, the five-dimensional model (AIC = 132716.06, BIC = 133071.49, number of parameters = 54) fitted the data better than the unidimensional model (AIC = 132998.25, BIC = 133261.53, number of parameters = 40). As a conclusion, it cannot be confirmed that the test measures a unidimensional construct, instead, with the reading competence test that includes texts featuring different text functions, subdimensions seemed to be measured. When drawing conclusions, two aspects have to be taken into account: first, as already noted, missing responses occurred increasingly on items at the end of the test. As a consequence, there was less variation and therefore, correlations may be lower. Second, the text functions were fully determined by the texts, that is, they were perfectly confounded since one text constituted one text function. Items were organized into item sets each referring to one text; hence, local item dependence (LID) may be prevalent. The correlations among the texts in the five-dimensional model as shown in Table 10 are, thus, not only due to multidimensionality, but also due to local item dependence.

The testing design in the main studies does not allow to disentangle these two sources. In pilot studies (Gehrer et al., 2012) a larger number of texts was presented to test takers allowing to investigate the impact of text functions independently of LID. The correlations estimated in the pilot study varied between r = .78 and r = .91. Although Gehrer et al. used a different scaling model, the results give a first idea of the impact of the text function (unconfounded with LID) on the dimensionality of the test. As the correlations found in Gehrer et al. (2012) differed from a perfect correlation, it was concluded that text functions formed subdimensions of reading competence. Comparing the correlations found in Gehrer et al. (2012), which were due to text functions, to those resulting from the main study (Table 10), which were due to both, text functions and LID, allowed us to evaluate the impact of LID. The correlations in the present study were similar (varying between r = .78 and r = .93) to those found in Gehrer et al. (ranging from r = 0.78 to r = 0.91), indicating that there was no considerable amount of local item dependence. Due to theoretical considerations, Gehrer et al. argued for a unidimensional construct. Consequently, a single competence score is estimated for reading competence.

6. Discussion

Descriptions and analyses presented in the previous sections aimed at documenting the quality of the adults' reading competence test and at providing information on the estimation procedure of the reading competence score published in the Scientific Use File.

The occurrence of different kinds of missing responses was evaluated and item as well as test quality were examined. In detail, item fit statistics including distractor analysis were thoroughly investigated not only for the dichotomous MC and polytomous CMC and MA items belonging to the final scaling model but also for the subtasks constituting CMC and MA items. Furthermore, measurement invariance, Rasch-homogeneity, unidimensionality as well as local item dependence were examined.

Overall, there is a rather small amount of missing responses due to invalid, not determinable, and omitted items. However, in particular, items at the end of the test show large amounts of missing responses due to not reached items. Given the testing time, the test is rather too long.

Item fit statistics provide evidence of well-fitting items which are measurement invariant across various subgroups. The test is very reliable. However, since the test is mainly targeted at low- and medium-performing participants, ability estimates for those participants will be very precise but less precise for high-performing persons.

Results of the dimensionality analyses challenge the conclusion of a unidimensional test. While cognitive requirements form a unidimensional construct, multidimensionality based on text functions seems to be present. In combination with the high amount of missing responses due to not reached items at the end of the test (i. e., there are participants with no valid responses to some of the text functions), the estimation of a single reading competence score is challenged. This might need to be addressed in further studies.

Nonetheless, Gehrer et al. (2012) argue that a balanced assessment of reading competence can only be achieved by heterogeneity of text functions and they provide theoretical arguments for a unidimensional measure of reading competence.

In summary, the reading test exhibits good psychometric properties that facilitate the estimation of a reliable reading competence score.

7. Data in the Scientific Use File

The data in the Scientific Use File contain 30 items, of which 23 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. 7 items were scored as polytomous variables (CMC or MA items). MC items are marked with a ‘0_c’ at the end of the variable name, while the variable names of CMC and MA items end with ‘s_c’. Note that the values of the polytomous variables in the Scientific Use File do not necessarily correspond to the number of correctly responded subtasks. This is due to collapsing of categories (cf. section 4.2 for a description of the aggregation of CMC and MA items). In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category. Manifest reading competence scores are provided in form of WLEs (rea2_sc1) together with their corresponding standard error (rea2_sc2). In the estimation of WLEs the effect of the test position (first vs. second) is controlled for. The ConQuest-Syntax used to estimate WLEs is provided in Appendix A. For persons who either did not take part in the reading test, for whom no information on the sequence of tests was available, or who did not have 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 value.

Plausible values that allow investigating latent relationships of competence scores with other variables will be provided in later data releases. Alternatively, users interested in investigating latent relationships may 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 (2012).

References

Akaike*, H. (1974).* A new look at the statistical model identification*. IEEE Transactions on Automatic Control, 19,* 716-722*.*

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: Springer.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). *NEPS framework for assessing reading competence and results from an adult pilot study*. Manuscript submitted for publication.

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

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

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

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: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., & Carstensen, C. H. (2013). *Scaling the competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges*. Manuscript submitted for publication.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche (Expanded Edition, Chicago, University of Chicago Press, 1980).

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6* (2), 461-464.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

Warm T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika, 54*, 427-450.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of Competencies Across the Life Span. In H. P. Blossfeld, H. G. Roßbach & J. v. Maurice & (Eds.). *Education as a Lifelong Process: The German National*

Educational Panel Study (NEPS). (Zeitschrift für Erziehungswissenschaft, Sonderheft 14). Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Appendix

Appendix A: ConQuest-Syntax for estimating WLE estimates in starting cohort 6

```
title Starting Cohort VI, READING: Partial credit model;

datafile filename.dat;
format pid 4-10 responses 13-42 position 48; /* insert number of columns with data*/
labels << filename_with_labels.txt;

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

score (0,1) (0,1)                           !items (1,3,4,6-10,12-18,20,22-26,28,30);
score (0,1,2) (0,0.5,1)                     !item (2,5,19,21,29);
score (0,1,2,3,4,5) (0,0.5,1,1.5,2,2.5)     !item (11);
score (0,1) (0,0.5)                         !item (27);

set constraint=cases;

model item + item*step + position;
estimate;

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