NEPS
National Educational Panel Study

NEPS Working Papers

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

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

NEPS Working Paper No. 16

Bamberg, October 2012

SPONSORED BY THE
Federal Ministry
of Education
and Research

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

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

[1]*Otto-Friedrich-Universität Bamberg, National Educational Panel Study*
[2]*University of Mannheim*

Email address of the lead author:

kerstin.haberkorn@uni-bamberg.de

Bibliographic data:

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

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

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

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competences across the whole life span and tests for assessing the different competence domains are developed. In order to evaluate the quality of the competence tests, a wide range of analyses has been performed based on Item Response Theory (IRT). This paper describes the reading competence data of starting cohort 4 in ninth grade. Next to descriptive statistics of the data, the scaling model applied to estimate competence scores, analyses performed to investigate the quality of the scale, as well as the results of these analyses are presented. The reading test in ninth grade consisted of 33 items, which represented different cognitive requirements and text functions and used different response formats. 13,933 subjects participated in the reading test. For scaling the competence test, a partial credit model was applied to the data. Item fit statistics, differential item functioning, Rasch-homogeneity, the tests' dimensionality, and local item independence were evaluated to ensure the quality of the test. The results show that the items fitted well to the model and that test fairness could be confirmed. The test's high reliability guarantees precise and differentiating ability estimates for the students. However, many items are targeted towards a lower reading ability. While the different comprehension requirements seem to form a unidimensional structure, the findings point at some multidimensionality based on text functions. Altogether, the reading test exhibited good psychometric properties and, therefore, the estimation of a reliable reading competence score is supported. The data available in the Scientific Use File are described and ConQuest-Syntax for scaling the data is provided.

Keywords

Item Response Theory, Scaling, Reading Competence, Scientific Use File

Content

1. Introduction

Within the National Educational Panel Study (NEPS) different competences are measured coherently across the life span. Tests have been developed for different competence domains. These include, among others, reading competence, mathematical competence, scientific literacy, information and communication technologies literacy, metacognition, vocabulary, and domain general cognitive functioning. Weinert et al. (2011) give an overview of the competences 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 scale are described in Pohl and Carstensen (2012a). In this paper the results of these analyses are presented for reading competence in starting cohort 4 – ninth grade. We will first introduce the main concepts of the reading competence test. Then, we will describe the reading competence data of starting cohort 4 and the analyses performed on the data 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 give on overview of 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 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 test that are necessary for understanding the scaling results presented in this paper.

The reading test consists of five texts and five item sets referring to these 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 (see Gehrer et al., 2012, and Weinert et al., 2011, for the description of the framework). 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.

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 function as distractors (i.e., they are incorrect). In CMC items a number of subtasks with two response options are presented. MA items require the subject 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 given in Pohl and Carstensen (2012a).

3. Data

3.1 The design of the study

In the present study, all tests were administered in the same order. Each student received the reading test in first position followed by other competence tests. Furthermore, no multi-matrix design regarding the selection and order of the items within a test was established. Thus, all test takers were given the same reading items in the same order.

Table 1: Cognitive requirements of the items in the reading test grade 9

Cognitive requirement	Frequency
Finding information in text	12
Drawing text-related conclusions	11
Reflecting and assessing	8
Total number of items	31

Table 2: Number of items for the different text types in the reading test grade 9

Text types/functions	Frequency
Information texts	5
Instruction texts	5
Advertising texts	7
Commenting or argumenting texts	7
Literary texts	7
Total number of items	31

The reading test in grade nine consisted of 33 items which represented different cognitive requirements and were assigned to specific text functions. Different response formats were implemented in the test. Extensive analyses were undertaken to detect items with insufficient characteristics. Based on these results, two of the items were excluded from final analyses due to unsatisfactory item fit. The characteristics of the remaining 31 items concerning cognitive requirements are depicted in Table 1, concerning text functions in Table 2, and concerning the response formats in Table 3. The CMC and MA items contained between two and four subtasks.

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

Response format	Frequency
Simple multiple choice	27
Complex multiple choice	3
Matching	1
Total number of items	31

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].

Overall, 13,933 subjects participated in the reading test[2]. 36 of them had less than three valid responses to the test items. Since no reliable reading competence score may be estimated for them, they were excluded from further analyses. The results from the remaining 13,897 persons are presented in the following sections.

4. Analyses

4.1 Missing responses

There are different kinds of missing responses. These are a) invalid responses, b) missing responses due to omitted items, c) items that have not been reached by the test takers, d) items that have not been administered, and finally, e) multiple kinds of missing responses within an item that are not determined. In this study, all persons received the same set of items, and, thus, there are no items that were not administered to a person.

Invalid responses occurred, for example, when two response options were selected in simple MC items where only one was required, or when numbers or letters that were not within the range of valid responses were given as a response. Omitted items occurred when test persons skipped items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response given were coded as not-reached. As complex multiple choice and matching items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A CMC or MA item was coded as missing if at least one subtask contained a missing response. 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 was labeled as a not-determinable missing response.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions, handling of different response formats). They also need to be accounted for in the estimation of item and person parameters. We therefore thorougly investigated the occurrence of missing responses in the test. We evaluated the amount of different types of missing responses per person to get an impression of how well the persons were coping with the test. We also examined the missing responses per item in order to evaluate how well each of the items functioned.

[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.

4.2 Scaling model

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

For the final scaling model the subtasks of the CMC and MA items 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 value, the whole CMC or MA item was scored as a missing response. When categories of the polytomous variables had less than N = 200, the categories were collapsed in order to avoid possible estimation problems. For one of the four CMC and MA items the first and second category were collapsed into one category because of the small number of persons in these categories. Note that, as a consequence, the values of the polytomously scored CMC and MA items in the Scientific Use File do not necessarily contain 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 1 for the correct response (see Haberkorn, Pohl, Carstensen, & Wiegand, 2012, and Pohl & Carstensen, 2012b, 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 (2012a), while the data available in the SUF are explicated in section 7.

4.3 Checking the quality of the test

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

Before the responses to the subtasks of CMC and MA items were aggregated and analyzed via a partial credit model, the psychometric properties of the subtasks were evaluated separately. For this purpose, they were analyzed together with the simple multiple choice items in a Rasch model (Rasch, 1960). The weighted mean square (WMNSQ), the respective t-value, point biserial correlations of the correct responses with the total score, and the item characteristic curves were checked to investigate whether the subtasks functioned appropriately. Only when the characteristics of the items were satisfactory, they were aggregated and included in the final scaling model.

To get a detailed view of the items' quality, we specifically evaluated the performance of the distractors (the incorrect response options) within the items. It was investigated whether the distractors were predominantly chosen by students with a lower ability rather than by those who answered correctly. Therefore, the point biserial correlations between the incorrect responses and the total score were regarded. We judged correlations below zero as very good, correlations below 0.05 as acceptable and correlations above 0.05 as problematic.

Next, the partial credit model was applied to the data, and the item fit of dichotomous MC items and the polytomous MA and CMC items was evaluated based on different fit indicators: the weighted mean square error (WMNSQ), the respective t-value, the correlations of the item score with the total score (equal to the discrimination value as computed in ConQuest), and the item characteristic curves. Because of the large sample size, rather tight fit criteria were used. Items with a WMNSQ > 1.10 (t-value > |5|) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.15 (t-value > |7|) were judged as having a considerable item misfit and their performance was further investigated. Correlations of the item score with the total score greater than 0.3 were considered as good, greater than 0.2 as acceptable, and below 0.2 as problematic. Overall, judgment of the fit of an item was based on all fit indicators.

We aim at constructing a reading competence test that measures the same construct for all students. 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. For the present study, test fairness was investigated for the variables gender, the number of books at home (as a proxy for socioeconomic status), the type of school, and migration background (see Pohl & Carstensen, 2012a, for a description of these variables). To detect test unfairness, differential item functioning (DIF), analyses were undertaken using a multigroup IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were estimated. 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 sincerely, and differences smaller than 0.4 as no considerable 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.

Based on theoretical considerations (see Pohl & Carstensen, 2012a), the partial credit model (1PL), in which Rasch-homogeneity is assumed, was applied to the competence data in NEPS. 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, 2012a). Nevertheless, Rasch's assumption of equal item discrimination was tested. Thus, the data were analyzed with a generalized partial credit model (2PL) (Muraki, 1992) using the software mdltm (von Davier, 2005), and the deviations of the estimated discrimination parameters from a uniform discrimination were evaluated. Moreover, the model fit indices of the 2PL model were compared to those of the partial credit model.

To examine the dimensionality of the test, two different multidimensional analyses based on the construction criteria for the reading test were conducted. In the first model, three dimensions representing the three cognitive requirements were modeled. In the second model, five dimensions reflecting the five text functions were specified. The correlations among the dimensions as well as differences in model fit between the unidimensional model and the multidimensional models were used to evaluate the tests' dimensionality.

Finally, local item dependence (LID) was appraised for the five item sets that referred each to one of the five texts. As each text function corresponded to one of the five texts, local item dependence and the text function were confounded. To disentangle the amount of multidimensionality and local item dependence, preliminary studies on dimensionality were used as reference.

5. Results

5.1 Missing responses

5.1.1 Missing responses per person

Figure 1 shows the number of invalid responses per person in the test. Overall, there were very few invalid responses. 96.42% of the subjects did not have any invalid responses at all and only 1.15% of the persons gave more than one invalid response. There was no difference in the amount of invalid responses between different response formats.

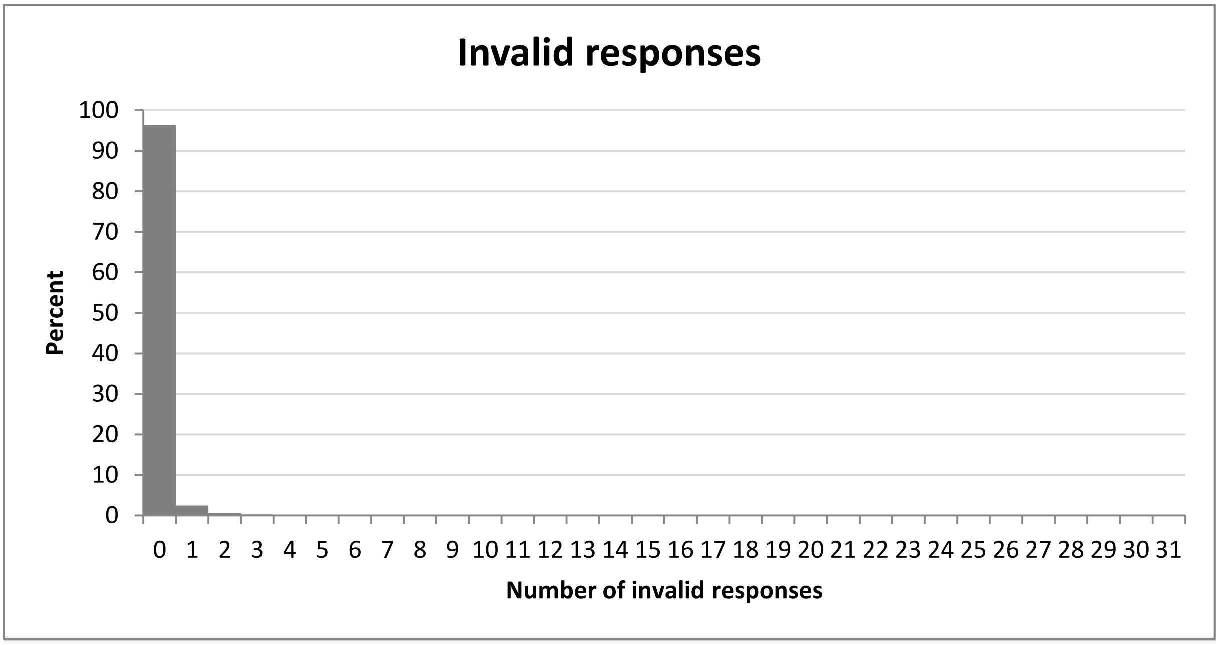

Figure 1: Number of invalid responses

Another kind of missing responses are omitted items. The frequency of participants who omitted items is depicted in Figure 2. There were, on average, 0.50% omitted items per person. 79.12% of the subjects omitted no item at all. A still rather small amount of 1.96% of the participants omitted more than five items. Considering the response format, more omitted items were found in CMC and MA items than in simple MC items.

When test persons cannot finish the test within the given time, not-reached items occur. In Figure 3, the number of not-reached items per person is presented. In comparison with other kinds of missing responses, most of the missing responses in this test arise from items that were not answered due to time limits (on average, 1.72% per person). About three quarters of test takers reached the end of the test, about 92% finished four of the five texts and 99% completed three of the five texts and responded to the corresponding items. Overall, the amount of not-reached items is still acceptable.

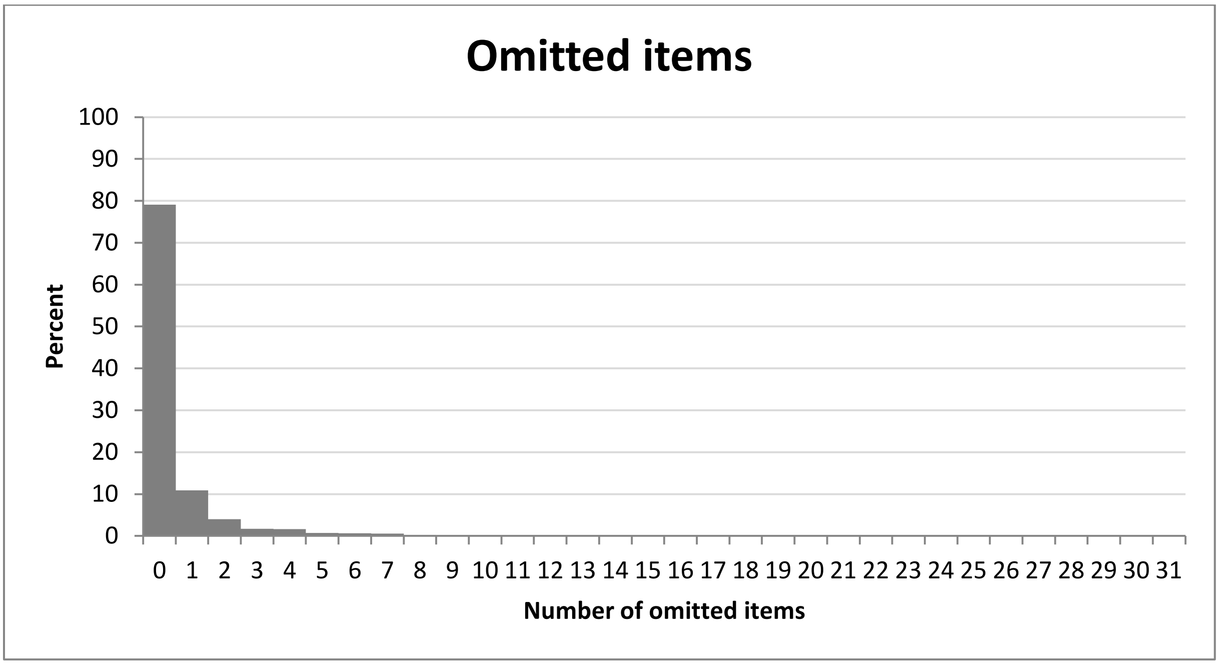

Figure 2: Number of omitted items

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. Figure 4 shows the amount of not-determinable missing responses for the four CMC and MA items in the test. As can be seen in the figure, there is only a very small number of not-determinable missing responses.

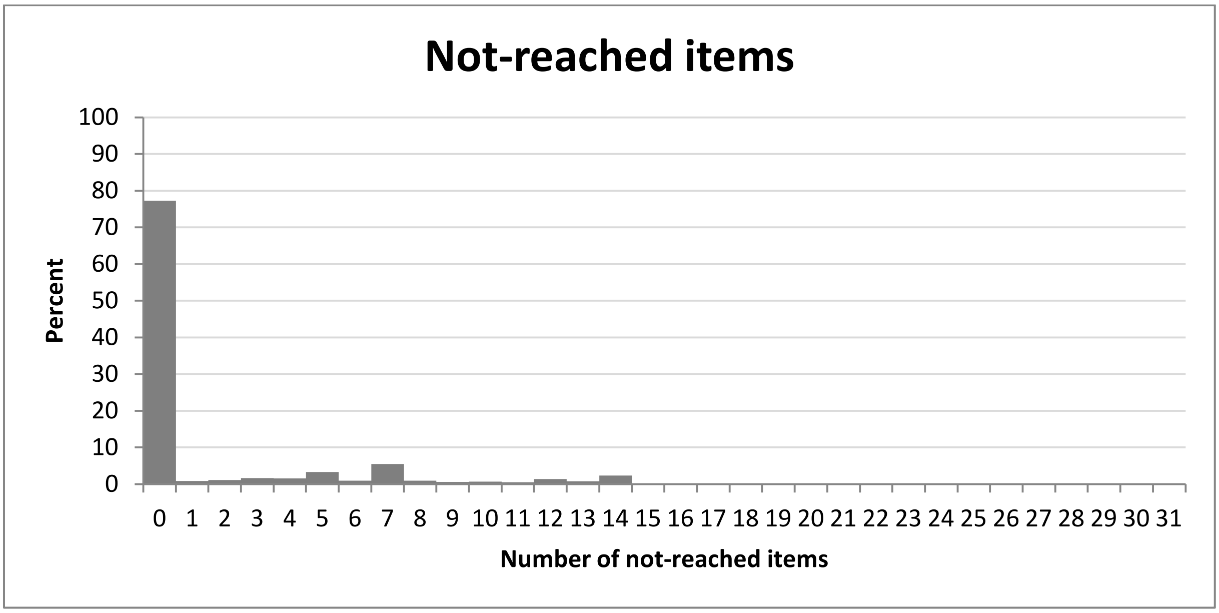

Figure 3: Number of not-reached items

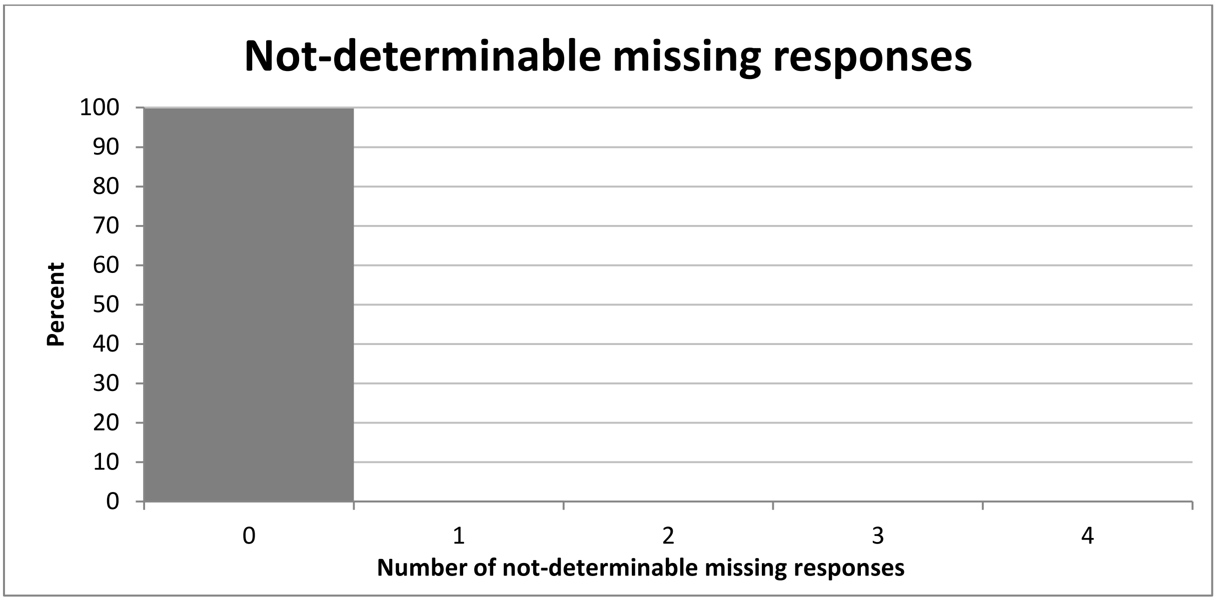

Figure 4: Number of not-determinable missing responses

Figure 5 gives the total number of missing responses per person consisting of invalid, omitted, not-reached, and not-determinable missing responses. Regarding all kinds of missing responses, 59.45% of test persons showed no missing response at all. 10.65% of the students had more than one quarter of missing responses and only about 0.94% completed less than half of the test. Altogether, the subjects had, on average, 2.29% missing responses in the test.

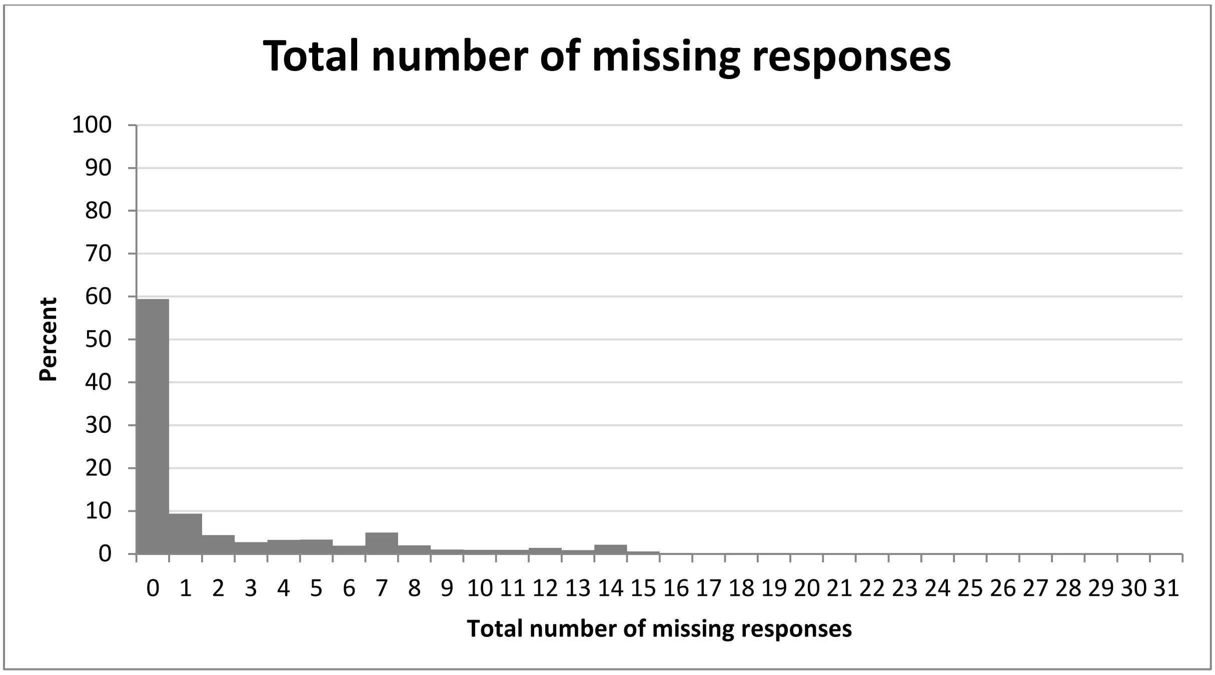

Figure 5: Total number of missing responses

In sum, the amount of invalid and not-determinable missing responses is small, whereas a reasonable part of missing responses occurs due to omitted items. The major part of missing responses results from items that were not reached. Overall, the number of missing responses is acceptable.

5.1.2 Missing responses per item

While the frequency of missing responses per person has already been presented in section 5.1.1, in Table 4 the percentage of the different types of missing responses per item is depicted. Only a small number of invalid responses or of not-determinable missing responses occurred. Regarding the amount of omissions, four items had an omission rate above 4%. A positive correlation (r = .551) between the amount of omitted responses and item difficulty was found. The more difficult the items are, the more likely they are omitted. The percentage of not-reached items was rather small in the first three texts. The percentage of missing responses within the items due to time limits increased gradually for items of the last three texts. The total number of missing responses per item ranged from 0.09% (reg90120_c) to 26.34% (reg90560_c).

5.2 Parameter estimates

5.2.1 Item parameters

Column 2 in Table 5 shows the percentage of correct responses in relation to all valid responses for each item. Note that since there is a nonnegligible amount of missing responses, this probability cannot be interpreted as an index for item difficulty. The percentage of correct responses within items varied between 39.17% and 98.16% with an average of 74.78% correct responses.

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 were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties (or location parameters for polytomous variables) ranged from -4.613 (item reg90120_c) to 0.572 (item reg90250_c). In total, the estimated item difficulties had a mean of -1.76. There are many items with a low item difficulty and only a limited number of items with a high difficulty. Due to the large sample size no standard error of the estimated item difficulties exceeded 0.07.

5.2.2 Person parameters

Person parameters are estimated as WLEs and plausible values (Pohl & Carstensen, 2012a). In the first release of the SUF, WLEs will be provided, whereas plausible values will be given in later analyses. 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 (2012a).

5.2.3 Test targeting and reliability

Test targeting was investigated in order to evaluate the measurement precision of the estimated ability scores and to judge the appropriateness of the test for the specific target population. In the partial credit model, the mean of the latent ability distribution was set to be zero. The variance of the ability distribution was 1.373, indicating that the test differentiates well between subjects. The test also exhibited a high reliability (EAP/PV reliability = .805 and WLE reliability = .749).

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 %
reg90110_c	1	13,744	0.00	0.44	0.66
reg90120_c	2	13,885	0.00	0.04	0.05
reg90150_c	5	13,673	0.00	1.36	0.25
reg9016s_c	6	13,495	0.00	2.04	0.76
reg9017s_c	7	13,584	0.00	2.23	0.02
reg90210_c	8	13,718	0.04	0.12	1.13
reg90220_c	9	13,720	0.04	0.47	0.76
reg90230_c	10	13,788	0.05	0.39	0.35
reg90240_c	11	13,823	0.06	0.33	0.14
reg90250_c	12	13,711	0.06	0.99	0.28
reg90310_c	13	13,794	0.14	0.42	0.18
reg90320_c	14	13,764	0.20	0.55	0.21
reg9033s_c	15	13,665	0.30	1.17	0.09
reg90340_c	16	13,718	0.40	0.66	0.22
reg90350_c	17	13,721	0.51	0.43	0.32
reg90360_c	18	13,698	0.56	0.83	0.04
reg90370_c	19	13,650	0.73	0.81	0.24
reg90410_c	20	13,303	3.04	1.12	0.12
reg90420_c	21	13,051	3.83	2.12	0.14
reg90430_c	22	12,567	5.20	4.25	0.13
reg90440_c	23	12,798	5.67	2.11	0.13
reg90450_c	24	12,764	6.33	1.68	0.15
reg90460_c	25	12,536	6.94	2.76	0.09
reg9047s_c	26	12,489	7.86	2.18	0.09
reg90510_c	27	11,760	13.36	1.85	0.17

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 %
reg90520_c	28	11,758	14.27	0.99	0.14
reg90530_c	29	10,840	17.61	4.25	0.14
reg90540_c	30	10,692	19.13	3.84	0.09
reg90550_c	31	10,297	20.78	4.97	0.16
reg90560_c	32	10,237	21.88	4.35	0.10
reg90570_c	33	10,728	22.70	0.00	0.11

Remarks.
The items on position 3 and 4 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	Discrimination – 2 PL
reg90110_c	96.15	-3.801	0.046	0.98	-0.5	0.28	1.09
reg90120_c	98.16	-4.613	0.064	0.97	-0.5	0.24	1.37
reg90150_c	79.99	-1.726	0.024	1.00	0.3	0.45	0.92
reg9016s_c	n.a.	-1.627	0.018	0.96	-2.8	0.59	0.96
reg9017s_c	n.a.	-3.753	0.041	0.99	-0.3	0.29	0.64
reg90210_c	87.36	-2.375	0.028	0.98	-1.1	0.41	0.61
reg90220_c	63.62	-0.696	0.020	1.08	9.4	0.42	1.62
reg90230_c	88.11	-2.456	0.028	1.05	2.8	0.31	0.38
reg90240_c	88.14	-2.460	0.028	0.88	-6.5	0.51	1.20
reg90250_c	39.17	0.572	0.020	1.18	20.9	0.30	1.83
reg90310_c	87.05	-2.345	0.027	0.95	-3.2	0.45	1.20
reg90320_c	93.32	-3.176	0.036	0.88	-4.5	0.46	1.40
reg9033s_c	n.a.	-3.125	0.029	0.90	-4.5	0.49	0.44
reg90340_c	89.35	-2.597	0.030	0.94	-3.3	0.45	0.63
reg90350_c	91.06	-2.818	0.032	0.92	-3.7	0.31	0.88
reg90360_c	77.35	-1.531	0.023	1.16	12.6	0.40	1.04
reg90370_c	66.47	-0.852	0.020	1.09	9.8	0.38	1.08
reg90410_c	87.29	-2.362	0.028	1.01	0.8	0.51	1.13
reg90420_c	75.82	-1.421	0.023	0.97	-2.8	0.53	1.29
reg90430_c	68.79	-0.983	0.022	0.95	-4.6	0.51	0.76
reg90440_c	80.26	-1.734	0.024	0.95	-3.7	0.51	1.09
reg90450_c	84.68	-2.105	0.027	0.92	-5.0	0.46	0.87
reg90460_c	65.06	-0.766	0.021	1.04	4.4	0.46	0.67
reg9047s_c	n.a.	-1.968	0.027	0.97	-2.2	0.55	1.17

Item	Percentage correct	Difficulty/location parameter	SE (difficulty/location parameter)	WMNSQ	t-value of WMNSQ	Correlation of item score with total score	Discrimination – 2 PL
reg90510_c	53.96	-0.171	0.021	0.93	-8.0	0.49	0.49
reg90520_c	63.27	-0.661	0.022	1.01	1.0	0.40	1.05
reg90530_c	60.25	-0.498	0.022	1.10	10.3	0.56	0.98
reg90540_c	48.29	0.129	0.022	0.90	-11.2	0.36	0.98
reg90550_c	51.70	-0.053	0.023	1.16	16.5	0.54	0.99
reg90560_c	64.75	-0.754	0.024	0.95	-4.4	0.45	1.36
reg90570_c	82.87	-1.966	0.028	1.00	0.0	0.28	1.09

Remarks.

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).

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

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

Item	Step 1 (SE)	Step 2 (SE)	Step 3 (SE)	Step 4 (SE)
reg9016s_c	0.01 (0.018)	-0.03 (0.018)	-0.42 (0.019)	0.44
reg9017s_c	1.07 (0.038)	-1.07		
reg9033s_c	0.19 (0.026)	0.66 (0.032)	-0.86	
reg9047s_c	1.22 (0.032)	-1.22		

Figure 6 depicts the item difficulty distribution and the reading ability distribution on the same scale. In the left panel, the persons' ability estimates are shown. In the right panel, the estimated item difficulties are given. Subjects with an ability corresponding to the difficulty of an item have a probability of 50% of correctly responding to this item. As a consequence, the item information is highest for subjects with an ability that corresponds to the difficulty of the item. Figure 6 shows that the items covered a wide range of the persons' ability distribution. However, most of the items showed a low difficulty, and rather few items were located at the upper ability distribution. This leads to precise estimates for subjects with a low and medium ability and higher standard errors for the ability estimates of subjects with a high reading ability.

5.3 Quality of the test

5.3.1 Fit of the subtasks of complex multiple choice and matching items

Before the subtasks of CMC and MA items were aggregated for the partial credit analysis, the fit of the subtasks was checked by analyzing the single subtasks together with the simple MC items in a Rasch model. The Rasch analysis was undertaken with 39 items: 27 MC items and the 12 subtasks of the CMC and MA items. There were no matching items with perfect stochastic dependence (see Pohl & Carstensen, 2012b, for a description of the problem), so that no subtasks had to be excluded from the analysis.

For all subtasks, a satisfactory item fit was obtained. 10 of the 12 items had a WMNSQ between 0.90 and 1.10 with a t-value ranging from -6.9 to 12.0. There were two items with a WMNSQ below 0.9 indicating a somewhat higher discrimination in comparison with the other items. Note that, due to the large sample size, t-values were rather high. Thus, the sample size was taken into consideration in the interpretation of the t-values. The empirical item characteristic curves were similar to the model-implied characteristic curves for all items. Since all of the subtasks of CMC and MA items showed a good item fit, all of them were used to construct aggregated polytomous scores for CMC and MA items. The polytomous CMC and MA items are marked with an 's_c' at the end of the variable name (whereas the variable name of MC items ends with a '0_c').

5.3.2 Distractor analyses

To get a detailed view of how the items performed, the quality of the items' distractors was evaluated additionally to the overall fit indices. For this purpose, the point biserial correlations (pt.bis) between the incorrect responses and the distractors based on the simple Rasch analysis, where the single subtasks of CMC and MA items were scaled together with the simple MC items (see section 5.2), were examined. Overall, the distractors correlated highly negative with the total score (on average pt.bis = -.24). For all items negative correlations were found with the pt.bis correlations ranging from -.03 to -.47. The results provide evidence for a proper functioning of the distractors.

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

Figure 6: Test targeting. Distribution of person ability (left side of the graph) and item difficulties (right side of the graph). Each ‘X’ represents 92.8 cases. Each number represents an item (which corresponds to the item position given in Table 4).

5.3.3 Item fit

The evaluation of the item fit was performed based on the results of the partial credit model. When examining the various fit indices (see Table 5), the partial credit model shows a good fit to the data. The WMNSQ for the different items varied from 0.88 (item reg90240_c) to 1.18 (item reg90250_c) with correspondent t-values that ranged from -11.2 to 20.9 (see Pohl & Carstensen, 2012b, for rules of thumb concerning the item fit). Overall, three items had a WMNSQ above 1.15 and the t-values of six items exceeded 6. The item characteristic curves of these items were still acceptable. For all items the empirical approximations of the item characteristic curves did not deviate noticeably from the model-implied curves.

The correlations between the item score and the total score varied between 0.24 (item reg90120_c) and 0.59 (item reg9016s_c). No items had correlation smaller than 0.20 and the average correlation (0.44) was good. Altogether, the different fit statistics indicate a good fit of the items to the scaling model.

5.3.4 Differential item functioning

To examine for test fairness (i.e., measurement invariance), differential item functioning (DIF) was investigated using the variables gender, the number of books at home (a proxy for socioeconomic status), migration background, and the type of school (see Pohl & Carstensen, 2012a, for a description of these variables). In this cohort, no DIF for a position effect was estimated, because there was no rotation of domains for reading. The test fairness was investigated as an overall measure for the whole test as well as for each item.

In Table 7, the absolute differences between the estimated item difficulties in the different subgroups are depicted in logits. For example, gender DIF “male vs. female” in Table 7 gives the differences between the item difficulties of males and females. A positive value indicates a higher difficulty for males, a negative value indicates a lower difficulty for males as compared to females. For example, the item reg90110_c is 0.472 logits more difficult for male students than for female students, leading to an item difficulty of -3.565 for male students on reg90110_c and an item difficulty of -4.037 for female students.

The sample of persons who took the reading competence test consisted of 6,909 (49.8%) female and 6,975 (50.2%) male students. 13 persons did not specify their gender and thus were excluded from the analysis on gender DIF. Overall, female students performed better than male students (main effect = 0.334 logits, Cohen’s d = 0.288)[3]. No considerable gender DIF was found for the items - except for item reg9033s_c which showed an absolute difference in difficulty of 0.610 logits. However, test developers found no evidence of a violation for test fairness when checking the content of this item.

DIF was also tested for the number of books at home, a proxy for socioeconomic status (SES). Students with 0 to 100 books at home were compared to students with more than 100 books at home. 5,516 (39.7%) students possessed 0 to 100 books and 7,451 (53.6%) students had more than 100 books at home. For 930 (6.7%) subjects, there was a missing value on

[3] Note that this main effect does not indicate a threat to measurement invariance. Instead, it indicates overall differences in ability between groups.

this variable. Because of the large amount of missing responses to this variable, the persons with missing responses were included in the DIF analysis as a separate group. An examination of the main effects showed a considerable difference in ability (-0.869 logits, Cohen's d = -0.797) between students with 0 to 100 books and those with more than 100 books. Students with a missing response on the DIF variable performed in a similar way as those with 0 to 100 books and 0.775 logits (Cohen's d = -0.775) worse than persons with more than 100 books at home. One item (reg90360_c) had DIF greater than 0.4 logits. There was no item with DIF that exceeded 0.6.

There were 9,802 (70.5%) test takers with no migration background, 3,472 (25.0%) persons with migration background and 623 (4.9%) persons for whom migration background could not be determined. The DIF analysis was performed comparing these three groups. The analysis exhibited a higher reading ability for students without migration background in comparison to students with migration background (main effect = 0.622 logits, Cohen's d = 0.548). Differences in reading ability were also found between students without migration background and students without information on migration background (main effect = 0.689 logits, Cohen's d = 0.608). Regarding item DIF no critical differences occurred since all items showed absolute differences below 0.6.

Finally, DIF was investigated for school type. 4887 subjects (35.2%) who took the reading test attended "Gymnasium" (type of school leading to upper secondary education and Abitur) and 9,010 (64.8%) did not. There were no missing values for this variable. The main effect of school type is quite large. Students enrolled in Gymnasium exhibited, on average, a better reading ability than students from other schools (main effect = 1.324 logits, Cohen's d = 1.338). The items show some amount of DIF. Item reg90240_c exhibited the highest DIF with an absolute difference in difficulty of 1.07 logits. For students not attending Gymnasium, the item was 1.07 logits more difficult than for students of other school types. Overall, two items had a DIF greater than 0.8 and the DIF of four items was above 0.6. With regard to their contents, no evidence for unfairness was found. Therefore, the items were not deleted for ability estimation.

In addition to examining DIF on item level, models including main effects only and models that additionally estimated DIF were compared. In Table 8 fit indices of the models including only main effects and those additionally including DIF are given for all four considered DIF variables. As can be seen in the table, Akaike's (1974) information criterion (AIC) always favored the models estimating DIF for all four DIF variables. The Bayesian information criterion (BIC; Schwarz, 1978) takes the number of estimated parameters into account and, thus, accounts for overparametrization of models. The values of the BIC indicate that the model additionally estimating DIF was preferred for the variables school, gender, and books. Regarding migration background, the more parsimonious model including only the main effect was preferred over the more complex DIF model.

Although the models additionally estimating DIF result in an overall better fit, for most of the variables the size of DIF is negligible. Those few items exhibiting a larger DIF show no substantive indication of test unfairness.

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

Item	Gender	Books			Migration status			School
	Male vs. female	<100 vs. >100	<100 vs. missing	>100 vs. missing	Without vs. with	Without vs. missing	With vs. missing	Gymnasium vs. non-Gymnasium
reg90110_c	0.472	0.389	0.196	-0.193	-0.018	0.094	0.112	-0.048
reg90120_c	0.104	0.028	-0.156	-0.184	-0.239	-0.711	-0.472	0.188
reg90150_c	-0.352	-0.010	0.187	0.197	0.043	0.035	-0.008	-0.126
reg9016s_c	0.146	0.140	0.213	0.073	-0.123	-0.090	0.033	0.232
reg9017s_c	-0.180	0.095	-0.049	-0.144	-0.377	-0.231	0.146	-0.094
reg90210_c	0.024	0.249	-0.003	-0.252	-0.295	-0.308	-0.013	-0.166
reg90220_c	-0.114	-0.035	-0.064	-0.029	0.036	0.041	0.005	-0.306
reg90230_c	0.676	-0.166	0.061	0.227	0.437	0.064	-0.373	-0.564
reg90240_c	0.336	0.626	0.082	-0.544	-0.403	-0.266	0.137	1.070
reg90250_c	-0.246	-0.359	0.053	0.412	0.292	0.395	0.103	-0.538
reg90310_c	0.238	0.567	0.132	-0.435	-0.395	-0.373	0.022	0.864
reg90320_c	0.480	0.383	0.034	-0.349	-0.227	-0.334	-0.107	0.636
reg9033s_c	0.610	0.088	-0.152	0.064	-0.143	-0.034	0.109	0.178
reg90340_c	0.456	0.276	0.135	-0.141	-0.149	-0.167	-0.018	0.378
reg90350_c	0.232	0.234	0.198	-0.036	-0.268	-0.437	-0.169	0.532
reg90360_c	0.004	-0.322	0.169	0.491	0.176	0.178	0.002	-0.786
reg90370_c	-0.194	-0.150	0.241	0.391	0.071	0.088	0.017	-0.376
reg90410_c	0.222	-0.033	0.119	0.152	0.004	0.008	0.004	-0.100
reg90420_c	-0.116	-0.002	0.038	0.040	-0.018	0.038	0.056	-0.090
reg90430_c	-0.076	0.154	0.002	-0.152	-0.110	-0.016	0.094	0.040
reg90440_c	0.258	0.137	0.079	-0.058	-0.031	-0.011	0.020	-0.016
reg90450_c	-0.124	0.164	0.016	-0.148	-0.055	-0.110	-0.055	-0.020

Item	Gender	Books			Migration status			School
	Male vs. female	<100 vs. >100	<100 vs. missing	>100 vs. missing	Without vs. with	Without vs. missing	With vs. missing	Gymnasium vs. non-Gymnasium
reg90460_c	-0.264	-0.186	0.027	0.213	0.206	0.172	-0.034	-0.506
reg9047s_c	-0.104	0.061	-0.073	-0.134	0.007	0.039	0.032	0.167
reg90510_c	-0.192	0.329	0.164	-0.165	-0.227	-0.145	0.082	0.402
reg90520_c	-0.082	0.134	0.314	0.180	-0.003	-0.150	-0.147	0.124
reg90530_c	-0.092	-0.003	0.237	0.240	0.234	0.192	-0.042	0.004
reg90540_c	-0.108	0.324	0.321	-0.003	-0.177	-0.009	0.168	0.442
reg90550_c	0.008	-0.100	0.097	0.197	0.244	0.254	0.010	-0.356
reg90560_c	-0.086	0.211	0.161	-0.050	-0.046	-0.068	-0.022	0.398
reg90570_c	0.136	0.248	0.112	-0.136	-0.216	-0.192	0.024	0.480
Main effect	-0.334	-0.227	-0.869	-0.094	0.622	0.689	0.067	-1.324

Table 8: Comparison of models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
School	main effect	366419.554	40	366499.554	366801.131
	DIF	364982.036	71	365124.036	365659.335
Gender	main effect	369474.038	40	369554.037	369855.577
	DIF	368788.582	71	368930.582	369465.814
Books	main effect	368478.890	41	368560.890	368870.007
	DIF	367836.473	103	368042.473	368819.034
Migration	main effect	369371.184	41	369453.184	369762.300
	DIF	368945.201	103	369151.201	369927.762

5.3.5 Rasch-homogeneity

To test the assumption of Rasch-homogeneity, a generalized partial credit model (2PL) was applied to the data and the estimated discrimination parameters were evaluated (see Table 5). 17 of the 31 items showed a discrimination between 0.80 and 1.20. However, some of the items exhibited a rather high or low discrimination ranging from 0.39 to 1.83. Comparing the AIC and the BIC of the partial credit model (1PL) (AIC = 370125.95, BIC = 370548.16, number of parameters = 55) and the 2PL model (AIC = 366121.39, BIC = 366769.78, number of parameters = 86), the 2PL model fitted the data better than the partial credit model. From a theoretical point of view, the partial credit model was chosen as a scaling model to preserve the weighting of items as intended in the theoretical framework.

5.3.6 Unidimensionality and local item independence

Based on the construction criteria for the reading test, two different multidimensional models were fitted to the data to evaluate the dimensionality of the test. In the first model, three dimensions representing the three cognitive requirements were specified in the second model, five dimensions based on the five text functions were applied to the data.

The first multidimensional model with three dimensions was estimated using the Gauss-Hermite quadrature method in ConQuest. In Table 9, the variances and correlations of the three cognitive requirements are presented. The variances of all dimensions were high, indicating that the participants were well discriminated on all subdimensions. The overall model fit of the three-dimensional model (AIC = 370119.98, BIC = 370451.72, number of parameters = 44) was slightly better than the fit of the unidimensional model (AIC = 370167.78, BIC = 370461.82, number of parameters = 39). This may, however, also be a result of the large sample size. As the subdimensions had very high correlations with each other (>0.95, see Carstensen, in press), clear evidence is provided that the different cognitive requirements form a unidimensional construct.

Table 9: Results of the three-dimensional model, based on the three cognitive requirements. The variance of the dimensions is given in the diagonal, correlations are depicted in the off-diagonal.

	Dim 1	Dim 2	Dim 3
Finding information in text (Dim 1) (Nitems = 11)	1.384		
Drawing text-related conclusions (Dim 2) (Nitems = 10)	0.973	1.281	
Reflection and evaluation (Dim 3) (Nitems = 6)	0.976	0.978	1.629

The five-dimensional model based on the text functions was estimated using the Monte Carlo estimation in ConQuest. The variances and correlations of the five-dimensional model are depicted in Table 10. All of the five text dimensions showed a high variance. As each text function corresponded to one of the five texts, local item dependence (LID) and the text functions were confounded. As a consequence, the deviation of the correlations from a perfect correlation shown in Table 10, may result from multidimensionality as well as from local item dependence. To disentangle these two sources, a pilot study by Gehrer et al. (2012) was used for comparison. In the study by Gehrer et al. (2012) a large number of texts were administered to the subjects and, thus, the impact of text functions could be examined independently of LID. Gehrer et al. (2012) found that the correlations between different text functions differed considerably from a perfect correlation (*r* = .78 to *r* = .91). Especially the literary text exhibited weaker relations than the other text functions. These findings point at multidimensionality due to text functions.

In the present study, the five-dimensional model exhibits a better model fit (AIC = 358104.73, BIC = 358504.32, number of parameters = 53) than the unidimensional model (AIC = 370167.78, BIC = 370461.82, number of parameters = 39). Comparing the correlations between the texts with the results of Gehrer et al. (2012), similar patterns were found. As suggested in the pilot study by Gehrer et al. (2012), the information, instruction and advertising texts showed higher correlations with each other, while the lowest correlations occurred between the literary texts and the other text types. Overall, the correlations (varying from 0.763 to 0.908) found in the present study (see Table 10) are similar in size as found in the pilot study, yielding to a negligible amount of LID. However, Gehrer et al. (2012) used a different scaling model than the present study, resulting in limitations for comparing the results of the two studies.

According to the test developers (Gehrer et al., 2012), a balanced assessment of reading competence can only be achieved by a heterogeneity of text functions. They emphasize that the reading test is constructed to measure a unidimensional reading competence score (Gehrer et al., 2012) and, hence, a unidimensional reading competence score was estimated and provided in the Scientific Use File.

Table 10: Results of the five-dimensional model based on the five text functions. The variances of the dimensions are given in the diagonal, correlations are shown in the off-diagonal.

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Information (Dim 1) (Nitems = 5)	2.045				
Instruction texts (Dim 2) (Nitems = 5)	0.908	1.095			
Advertising texts (Dim 3) (Nitems = 7)	0.869	0.902	1.693		
Literary function (Dim 4) (Nitems = 7)	0.801	0.763	0.794	2.319	
Commenting function (Dim 5) (Nitems = 7)	0.821	0.779	0.795	0.813	1.717

6. Discussion

In the previous sections, the quality of the reading test for the ninth grade was evaluated by carrying out several analyses. Furthermore, the estimation of the reading competence score was described.

We investigated the occurrence of different kinds of missing responses and examined the item and test parameters. We thoroughly checked item fit statistics for simple MC items, subtasks of CMC and MA items, as well as the aggregated polytomous CMC and MA items, and examined the appropriateness of distractors. Further quality inspections include testing for differential item functioning, testing for Rasch-homogeneity, investigating the tests' dimensionality, as well as local item dependence.

The different item fit statistics give evidence for a good item fit for all items and test fairness could be confirmed. The high reliability of the test and the large variance of latent abilities ensure precise and differentiating measures for the students. The amount of missing responses is still tolerable. The amount of items that were not reached indicates that the testing time is too short for the number of items presented. Whether the reading ability score is, as a consequence, confounded by speediness needs to be investigated in further studies.

As many items are targeted towards a lower reading ability, estimates for these students are very precise, whereas ability of high-performing students are assessed less precise. Another challenge is the test's dimensionality, since the heterogeneity of the text functions denotes a multidimensional construct. However, according to Gehrer et al. (2012) a balanced assessment of reading competence can only be achieved by a certain heterogeneity of text functions and, therefore, based on theoretical arguments, a unidimensional competence score is estimated and provided in the Scientific Use File.

In summary, the reading test shows good psychometric properties that support the estimation of a reliable reading competence score for the Scientific Use File.

7. Data in the Scientific Use File

The data set in the Scientific Use File consists of 31 items. 27 items are simple MC items and they are scored dichotomously with 0 indicating an incorrect response and 1 indicating a correct response. Four of the items are CMC or MA items, that is, polytomous items that were aggregated from the respective subtasks (see section 4.2). The values of the CMC and MA items indicate the (partial) credit that a person received for the item. The polytomous CMC and MA items are marked with a 's_c' at the end of the variable name, whereas the variable name of MC items ends with a '0_c'. Note that the values of the polytomous variables do not necessarily indicate the number of correctly responded subtasks (see section 4.2) since categories may have been collapsed due to small category frequencies. In the scaling model, each category of the polytomous CMC and MA items is scored with 0.5 points. Besides the item data, manifest scale scores for reading competence are provided in the form of WLE estimates (reg9_sc1) including the corresponding standard error (reg9_sc2). The ConQuest Syntax for estimating the WLE scores based on the 31 items is given in Appendix A. Students that did not participate in the testing or that did not have enough valid responses for estimating a scale score, got a non-determinable missing value on the WLE score for reading competence.

Note that plausible values, which allow for an investigation of latent relationships, will be provided in later releases. An overview of how to work with competence data in NEPS is given in Pohl and Carstensen (2012a).

References

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

Carstensen, C. H. (in press). 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. H., & Wiegand, E. (accepted). Incorporating different response formats in the IRT-scaling model for competence data. In H.-P. Blossfeld, J. Skopek, & J. Maurice (Eds.). *Methodological issues of longitudinal surveys: The example of the national educational panel study*.

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

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

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)* (pp. 67-86). *(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 IV

```
Title Starting Cohort IV, READING: Partial credit model;

data filename.dat;
format pid 4-10 responses 13-43; /* insert number of columns with data*/

labels << filename_with_labels.txt;

codes 0,1,2,3,4;
score (0,1) (0,1)                       !items (1, 2, 5, 8-14, 16-25, 27-33);
score (0,1,2,3,4) (0,0.5,1,1.5,2)       !item (6);
score (0,1,2) (0,0.5,1)                 !item (7);
score (0,1,2,3) (0,0.5,1,1.5)           !item (15);
score (0,1,2) (0,0.5,1)                 !item (26);

set constraint=cases;

model item + item*step;
estimate;

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