NEPS
National Educational Panel Study

NEPS Working Papers

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

NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grade

NEPS Working Paper No. 15

Bamberg, October 2012

SPONSORED BY THE
Federal Ministry of Education and Research

Working Papers of the German National Educational Panel Study (NEPS)
at the University of Bamberg

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

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

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

The NEPS Working Papers are available at
http://www.uni-bamberg.de/neps/publikationen/neps-working-papers/

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 3 in Fifth Grade

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

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

Email address of the lead author:

steffi.pohl@uni-bamberg.de

Bibliographic data:

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

We thank Theresa Rohm, Rebecca Stammberger, Michael Wenzler, 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 3 in Fifth Grade

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competences across the whole life span and tests for assessing the different competence domains are developed. In order to evaluate the quality of the competence tests, a wide range of analyses have been performed based on Item Response Theory (IRT). This paper describes the reading competence data of starting cohort 3 in fifth grade. Next to descriptive statistics of the data, the scaling model applied to estimate competence scores, analyses performed to investigate the quality of the scale, as well as the results of these analyses are presented. The reading test in fifth grade consisted of 33 items, which represented different cognitive requirements and text functions and used different response formats. The test was administered to 5,208 students. A partial credit model was used for scaling the data. Item fit statistics, differential item functioning, Rasch-homogeneity, the tests' dimensionality, and local item independence were evaluated to ensure the quality of the test. The results show that the items exhibited good item fit and measurement invariance across various subgroups. Moreover, the test showed a high reliability and the different comprehension requirements foster a unidimensional construct. Challenges of the test are the considerable amount of items that have not been reached by test takers due to time limits, the many items that are targeted towards a lower reading ability, and some evidence for multidimensionality of the test based on text functions. Overall, the results revealed good psychometric properties of the reading test and support the estimation of a reliable reading competence score. In the paper, the data available in the Scientific Use File are described and ConQuest-Syntax for scaling the data is provided.

Keywords

Item Response Theory, Scaling, 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, amongst 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 competence domains measured in NEPS.

Most of the competence data are scaled using models that are based on Item Response Theory (IRT). Since most of the competence tests were developed specifically for implementation in NEPS, several analyses have been 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 (2012a). In this paper the results of these analyses are presented for reading competence in starting cohort 3 – fifth grade. We will first introduce the main concepts of the reading competence test. Then, we will describe the reading competence data of starting cohort 3 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 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 the 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 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 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 the description of the framework).

In the reading competence test there are three types of response formats. These are simple multiple choice (MC), complex multiple choice (CMC), and matching (MA) items. MC items consist of 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 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 given in Pohl and Carstensen (2012a).

3. Data

3.1 The design of the study

Among others, two "lifespan" domains were assessed in this study – namely, reading and mathematical competence. In order to control for effects of position and order, the two tests were assigned to test takers in different order. Half of the subjects received a booklet that first contained the reading test followed by the mathematics test, while the other half of the sample received the two tests in the opposite order. The subjects were randomly assigned to one of the two booklets. Note that there was no multi-matrix design regarding the choice and order of the items *within* a test. All students received the same reading items in the same order.

The reading test in grade five consisted of 33 items which represented different cognitive requirements and text functions. Furthermore, different response formats were used. Extensive analyses were performed to detect items with insufficient characteristics. Due to unsatisfactory item fit in these analyses, one item was subsequently excluded from the analyses. The characteristics of the remaining 32 items are depicted in Tables 1 to 3. Table 1 contains the distribution of cognitive requirements, Table 2 reflects the distribution of the text functions and Table 3 informs about the distribution of the response format. The number of subtasks within CMC and MA items varied between four and eight.

Table 1: Comprehension requirements of the items in the reading test grade 5

Cognitive requirement	Frequency
Finding information in text	9
Drawing text-related conclusions	12
Reflecting and assessing	11
Total number of items	32

Table 2: Number of items for the different text types in the reading test grade 5

Text types/functions	Frequency
Information texts	7
Instruction texts	6
Advertising texts	7
Commenting or argumenting texts	5
Literary texts	7
Total number of items	32

Table 3: Response formats of the items in the reading test grade 5

Response format	Frequency
Simple multiple choice	26
Complex multiple choice	3
Matching	3
Total number of items	32

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].

5,208 persons took the reading test[2]. 2,613 subjects first received the reading test followed by the mathematics test, while 2,595 subjects received the reading test after having completed the mathematics test. 15 of the test takers had less than three valid responses to the test items. Since no reliable reading competence score may be estimated based on such a low number of responses, these cases were excluded from further analyses. The results of the remaining 5,193 test takers are presented in the following sections.

4. Analyses

4.1 Missing responses

There are different kinds of missing responses. These are a) invalid responses, b) missing responses due to omitted items, c) missing responses due to items that have not been reached, d) missing responses due to items that have not been administered and e) multiple kinds of missing responses that occur in an item and are not determined. In this study, all subjects received the same set of items. As a consequence, there were 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 just one was required, or when numbers or letters that were not within the range of valid responses were given as a response. Missing responses due to omitted items occurred when test persons skipped items. Due to time limits, it might happen that not every person finished the test within the given time. As a consequence, missing responses due to items that were not reached resulted. 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 value was coded as a not-determinable missing response.

[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.

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, therefore, thoroughly investigated the occurrence of missing responses in the test. First, we looked at the occurrence of the different types of missing responses per person. This gave an indication of how well the persons were coping with the test. We then examined the occurrence of missing responses per item, in order to get some information on how well the items worked.

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 (2012a).

The 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, 2012a for the explanation of this approach). For five of the six 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 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). A special case is item reg5026s_c. The item consisted of eight subtasks which showed an extreme local stochastic dependence. In accordance with substantial considerations, the item was scored as 1 only if all subtasks were solved correctly; otherwise it was scored as zero.

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 described 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.

The responses to the subtasks of CMC and MA items were aggregated to a polytomous variable for each CMC and MA item. In order to justify such an aggregation, the fit of single subtasks was checked in analyses. For this purpose, the single 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 showed a satisfactory item fit, they were used to construct polytomous CMC and MA item variables.

The MC, CMC, and MA items consisted of one correct response and a number of distractors (incorrect response options). We investigated whether the distractors worked well, that is, whether they were predominantly chosen by students 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.

Item fit was then evaluated for the MC items and the polytomous CMC and MA items based on results of a partial credit model. Again, the weighted mean square error (WMNSQ), the respective t-value, correlations of the item score with the total score, and the item characteristic curves 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 a considerable item misfit and their performance was further investigated. Correlations of the item score with the total score (equal to the discrimination value 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 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. Test fairness was investigated for the variables test position, gender, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012a, for a description of these variables). In order to test for measurement invariance, differential item functioning (DIF) 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 estimated. Differences in the estimated item difficulties between the subgroups were evaluated. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties that are 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, model fit was investigated by comparing a model, including differential item functioning, to a model that only included main effects and no DIF.

The reading competence data in NEPS were scaled using the partial credit model (1PL), in which Rasch-homogeneity is assumed. The partial credit model was chosen because it

preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012a). Nevertheless, Rasch-homogeneity is an assumption that may not hold for empirical data. We therefore checked for deviations from a uniform discrimination by estimating item discrimination with the generalized partial credit model (2PL) (Muraki, 1992) using the software mdltm (von Davier, 2005), and by comparing 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 performing 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 comprehension 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 competence test consisted of item sets that referred to one of five texts, the assumption of local item independence 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

5.1.1 Missing responses per person

The number of invalid responses per person is shown in Figure 1. The number of invalid responses was very low. 88 % of the test persons had no invalid responses. Only about three percent of the subjects had more than one invalid response. Considering the response format, missing responses occurred particularly in CMC items.

Missing responses may also occur when persons skip (omit) some items. Figure 2 presents the number of omitted responses per person. As can be seen in the figure, there was a nonnegligible amount of omitted items. Only 46 percent of the subjects omitted no item at all. However, only five percent of the subjects omitted more than five items.

All missing responses after the last valid response were defined as not reached. Figure 3 illustrates the number of items that were not reached by the persons. The number of not-reached items was rather high. Only 48.2% of the subjects reached the end of the test. 35% of the subjects did not reach the items of the last text, 14% did not reach the last two of the five texts, and 3 % of the persons only reached the first two texts.

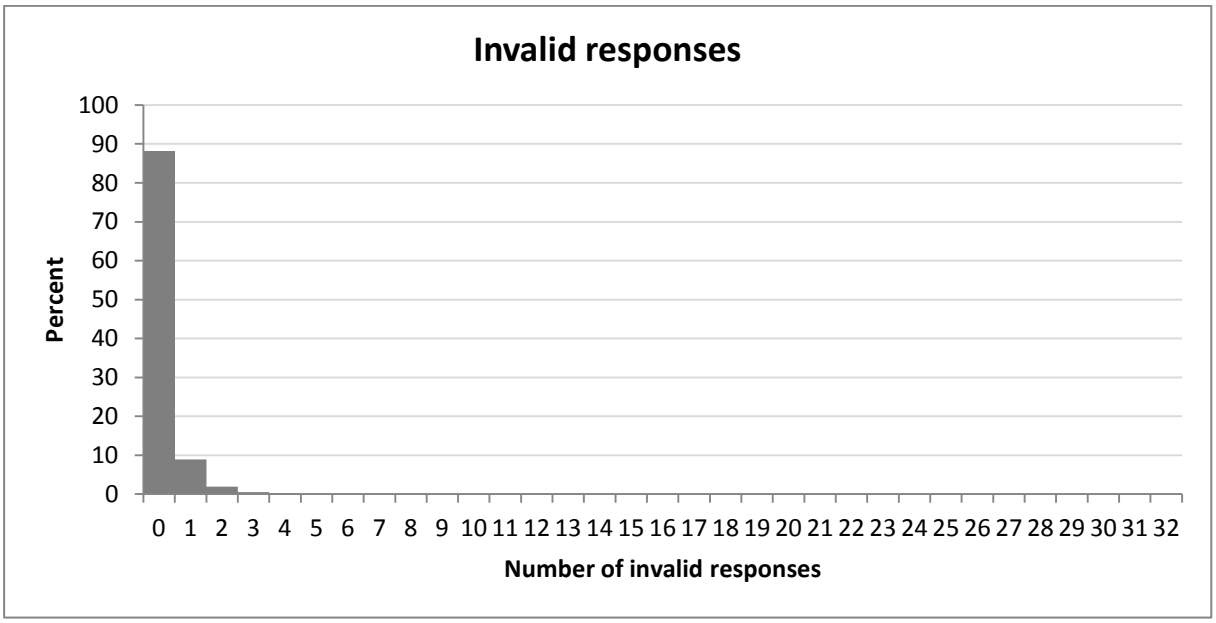

Figure 1: Number of invalid responses

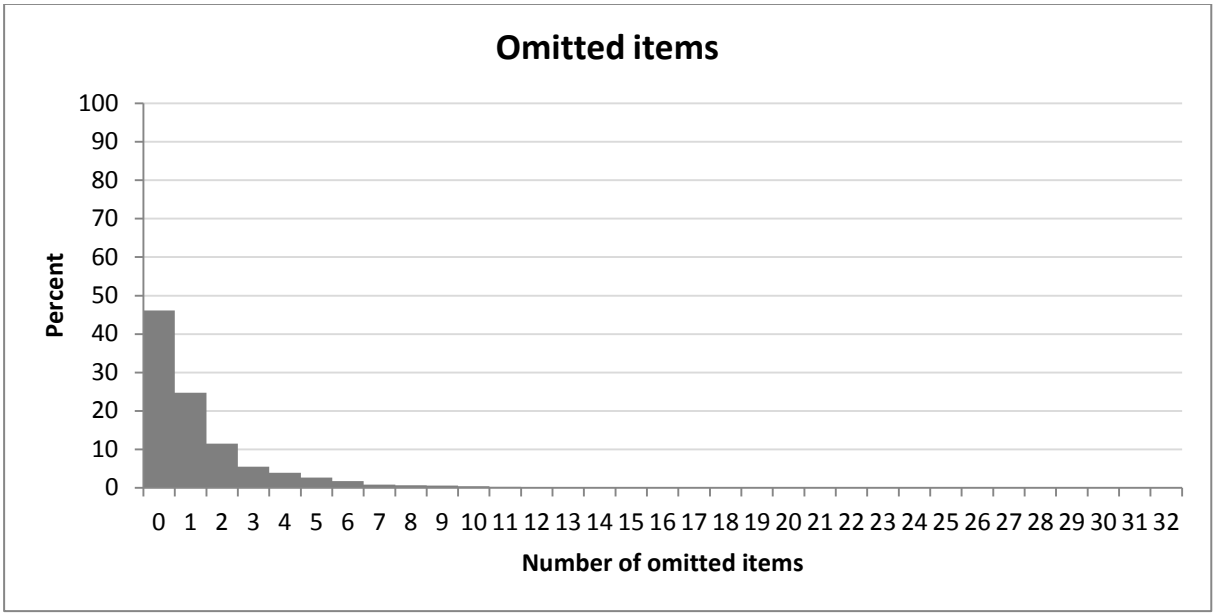

Figure 2: Number of omitted items

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 number of not-determinable missing responses in the test. Since not-determinable missing responses may only occur in CMC and MA items, the maximum number of not-determinable missing responses was six (i.e., the number of CMC and MA items). As can be seen in the figure, there was only a very small amount of not-determinable missing responses. About 97% of the persons didn't have a not-determinable missing response.

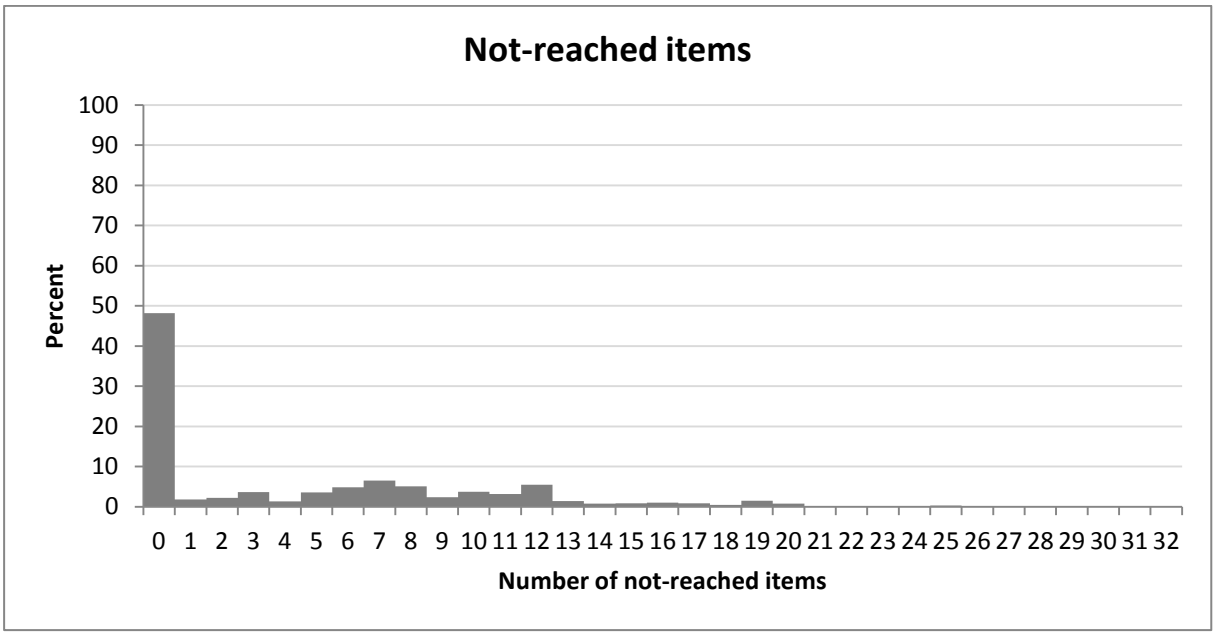

Figure 3: Number of not-reached items

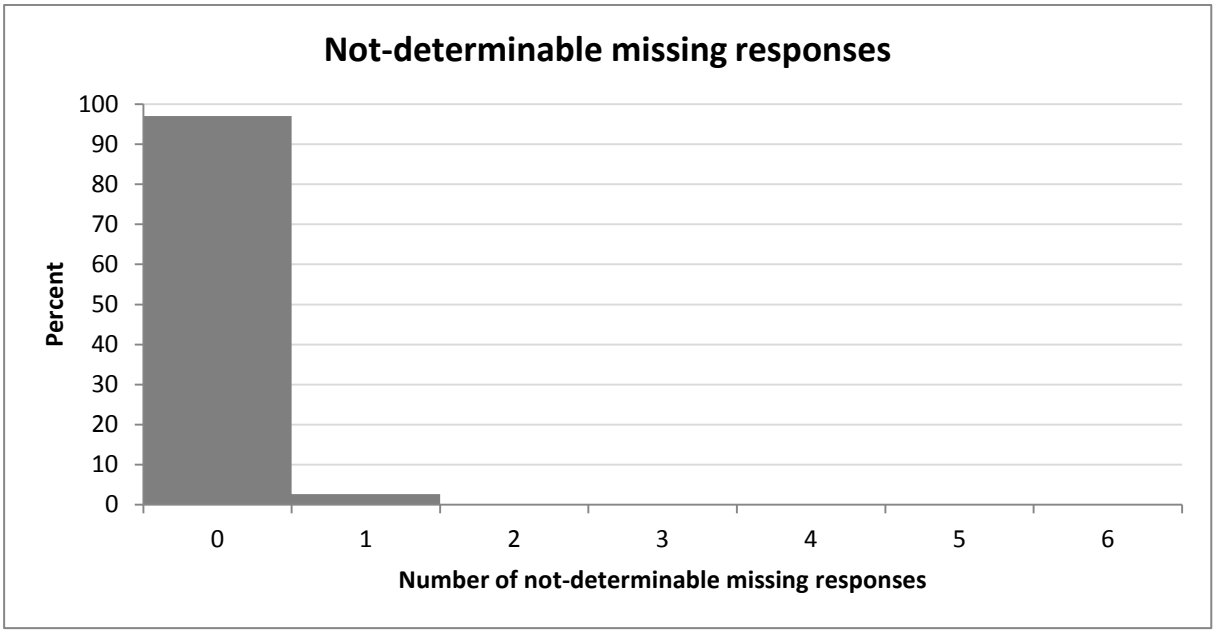

Figure 4: Number of not-determinable missing responses

In Figure 5 the total number of missing responses per person is depicted. The total number of missing responses is the sum of invalid, omitted, not-reached, and not-determinable missing responses. Figure 5 shows that only 20.3% of the subjects had no missing response at all. Almost 50% of the test persons had more than five missing responses. 6.5% of the subjects had missing responses to more than 16 (i.e., 50% of the) items.

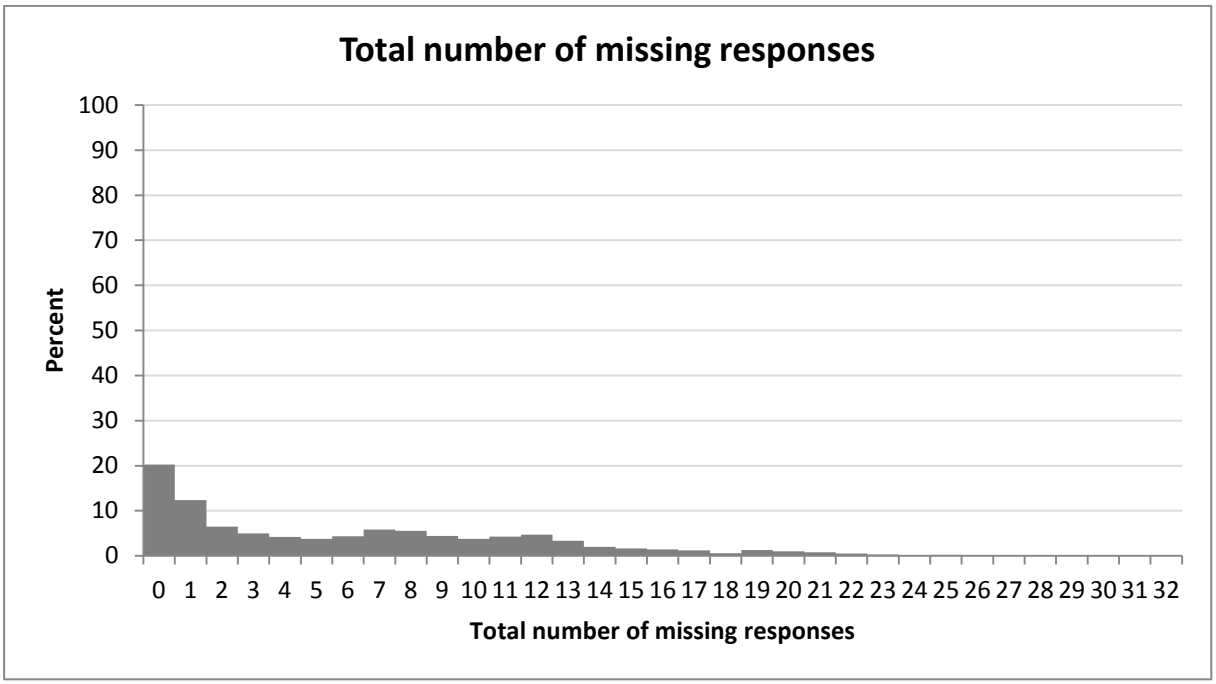

Figure 5: Total number of missing responses

Summarizing these results, 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 rather large and, therefore, it represents an essential part of the total number of missing responses.

5.1.2 Missing responses per item

Table 4 shows the number of valid responses for each item, as well as the percentage of missing responses. Overall, the number of persons that omitted an item is acceptable. There were nine items with an omission rate above 5%. The highest omission rate occurred for item reg5012s_c (17.49% of the persons omitted this item). This is probably due to the fact that the subjects had not (yet) totally understood the instructions on how to respond to such a response format. The number of missing responses was correlated to .230 with the difficulty of the item. There was one item with a bivariate extreme value, that is, a high difficulty and a very low missing rate. Excluding this item, the correlation between item difficulty and the number of omitted responses increased to .518. This result indicates that the test takers tend to omit items that are more difficult. The number of persons that did not reach an item increased with the position of the item in the test to up to 51.84%. This is a rather large amount. The number of invalid responses per item was small. The highest number was 2.75% for item reg5016s_c. The reason for invalid responses to this item is probably again due to a misunderstanding of the instruction for matching tasks. The total number of missing responses per item varied between 1.04% (item reg50110_c) and 53.3% (item reg5055s_c).

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 %
reg50110_c	1	5139	0.00	0.75	0.29
reg5012s_c	2	4248	0.00	17.48	0.54
reg50130_c	3	5109	0.00	1.31	0.31
reg50140_c	4	4914	0.00	4.97	0.40
reg50150_c	5	4928	0.00	3.43	1.67
reg5016s_c	6	4504	0.00	9.49	2.75
reg50170_c	7	4992	0.10	1.93	1.85
reg50210_c	8	5096	0.35	1.25	0.27
reg50220_c	9	4751	0.42	7.76	0.33
reg50230_c	10	5023	0.54	2.50	0.23
reg50240_c	11	5009	0.67	1.67	1.19
reg50250_c	12	4939	0.83	3.85	0.21
reg5026s_c	13	4367	1.56	11.11	2.16
reg50310_c	14	4847	3.00	3.58	0.08
reg50320_c	15	4871	3.45	2.54	0.21
reg50330_c	16	4877	4.29	1.60	0.19
reg50340_c	17	4731	5.30	3.45	0.15
reg50350_c	18	4780	6.14	1.39	0.42
reg50360_c	19	4703	6.91	2.50	0.02
reg50370_c	20	4530	8.34	4.24	0.19
reg50410_c	21	4207	13.77	4.91	0.31
reg5042s_c	22	3909	16.93	7.53	0.17
reg50430_c	23	3693	20.66	7.82	0.40

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 %
reg50440_c	24	3567	23.01	7.95	0.35
reg50460_c	26	3304	28.04	7.68	0.65
reg50510_c	27	3279	34.53	2.12	0.21
reg5052s_c	28	2791	39.32	6.08	0.46
reg50530_c	29	2792	42.85	3.22	0.17
reg50540_c	30	2796	44.16	1.85	0.15
reg5055s_c	31	2425	47.79	4.00	0.94
reg50560_c	32	2441	50.01	2.75	0.23
reg50570_c	33	2478	51.84	0.00	0.44

Remarks.
The item on position 25 was 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
reg50110_c	94.40	-3.383	0.063	0.94	-1.2	0.37	1.47
reg5012s_c	n.a.	-2.865	0.057	0.87	-3.7	0.50	2.00
reg50130_c	86.44	-2.280	0.044	0.99	-0.4	0.41	0.96
reg50140_c	78.49	-1.623	0.038	0.97	-1.5	0.49	1.09
reg50150_c	61.95	-0.609	0.033	1.01	0.4	0.50	0.87
reg5016s_c	n.a.	-1.496	0.024	1.00	-0.1	0.69	1.06
reg50170_c	34.01	0.842	0.033	0.99	-0.7	0.45	0.86
reg50210_c	91.50	-2.886	0.053	0.96	-1.1	0.39	1.24
reg50220_c	50.79	-0.053	0.033	1.21	14.8	0.31	0.40
reg50230_c	89.81	-2.662	0.050	0.92	-2.3	0.45	1.37
reg50240_c	75.86	-1.428	0.037	0.94	-3.1	0.53	1.16
reg50250_c	65.82	-0.816	0.034	1.05	3.5	0.45	0.74
reg5026s_c	n.a.	1.012	0.036	1.00	0.0	0.46	1.74
reg50310_c	86.96	-2.341	0.046	0.94	-1.9	0.46	1.21
reg50320_c	91.62	-2.900	0.055	0.91	-2.3	0.44	1.55
reg50330_c	89.81	-2.660	0.051	0.94	-1.6	0.42	1.23
reg50340_c	78.69	-1.633	0.039	0.97	-1.5	0.49	1.04
reg50350_c	62.59	-0.643	0.034	1.04	2.9	0.47	0.78
reg50360_c	87.16	-2.362	0.047	0.97	-1.1	0.43	1.09
reg50370_c	74.46	-1.337	0.038	1.02	1.0	0.46	0.81
reg50410_c	86.96	-0.598	0.036	1.11	6.8	0.40	0.64
reg5042s_c	n.a.	-2.072	0.041	1.10	4.0	0.38	0.34
reg50430_c	61.78	0.878	0.039	0.99	-0.6	0.45	0.84

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
reg50440_c	33.36	0.436	0.039	1.05	3.1	0.42	0.70
reg50460_c	41.24	-0.123	0.040	1.07	4.0	0.44	0.72
reg50510_c	51.42	-2.383	0.056	0.92	-2.2	0.49	1.37
reg5052s_c	86.46	-1.533	0.042	0.85	-5.8	0.65	1.51
reg50530_c	26.00	-0.001	0.044	1.05	2.5	0.47	0.77
reg50540_c	73.46	-1.418	0.048	0.93	-3.1	0.56	1.20
reg5055s_c	n.a.	-1.355	0.045	0.95	-2.0	0.58	1.08
reg50560_c	54.65	-0.362	0.047	1.22	10.2	0.35	0.48
reg50570_c	65.13	-0.968	0.048	1.01	0.4	0.51	0.88

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 in ConQuest).

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

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 26.00% and 94.40% with an average of 69.00% 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) varied between -3.383 (item reg50110_c) and 1.012 (item reg5026s_c) with a mean of -1.3. Overall, the item difficulties were very low. There were no items with a high difficulty. Due to the large sample size, the standard error (SE) of the estimated item difficulties was very small (SE ≤ 0.06).

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)
reg5012s_c	1.197 (0.062)	-1.197			
reg5016s_c	-0.190 (0.034)	0.494 (0.036)	0.582 (0.040)	-0.114 (0.046)	-0.773
reg5042s_c	-0.268 (0.034)	-0.110 (0.036)	0.378		
reg5052s_c	0.579 (0.042)	-0.351 (0.470)	-0.228		
reg5055s_c	-0.159 (0.042)	0.093 (0.048)	0.066		

5.2.2 Person parameters

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

5.2.3 Test targeting and reliability

Test targeting was investigated in order to evaluate the measurement precision of the estimated ability scores and to judge the appropriateness of the test for the specific target population. In the analyses, the mean of ability was constrained to be zero. The variance was estimated to be 1.361, indicating that the test differentiates well between subjects. The reliability of the test (EAP/PV reliability = .811, WLE reliability = .767) was good.

The extent to which item difficulties and location parameters are targeted towards the persons' abilities is shown in Figure 6. The figure shows that the items covered a great range of the ability distribution of the test persons. However, only few items covered a medium person ability, and no item grasped a very high degree of ability. Instead, there were a large number of easy items. As a consequence, subjects with a medium and low ability will be

measured relatively precisely, while subjects with a high reading ability will have a larger standard error of measurement.

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 and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the single subtasks together with the simple MC items in a Rasch model. Counting the subtasks of CMC and MA items separately, there were 57 items. Since there were three matching tasks with perfect stochastic dependence (see Pohl & Carstensen, 2012b, for a description of the problem), one of the subtasks of each of these MA items was excluded from the analyses. Consequently, 54 items were considered in the analysis.

Despite the very high probability of a correct response of greater than 95% for six subtasks of CMC items, no estimation problems occurred. For the remaining items the number of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.84 to 1.23, the respective t-value from -11.6 to 18.7, and there were no noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. Due to the good model fit of the subtasks, their aggregation to polytomous variables seems to be justified. Note that in the SUF, 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

In addition to the overall item fit, we specifically investigated how well the distractors performed in the test by evaluating the point biserial correlation between each incorrect response (distractor) and the students’ total score. All distractors had a point biserial correlation with the total score below zero. The point biserial correlations ranged from -.5 to -.02 with a mean of -.224. The results indicate that the distractors work well.

5.3.3 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the partial credit model. Altogether, the item fit is very good (see Table 5). WMNSQ was close to 1 with the lowest value being 0.85 (item reg5052s_c) and the highest being 1.22 (item reg50560_c). Overall, there were only two items with a WMNSQ above 1.2 and a respective t-value above 8. There were no further indications for a heavy misfit of the items. They were therefore kept in the analysis for estimating reading competence scores. The correlations of the item score with the total score varied between .31 (item reg50220_c) and .69 (item reg5016s_c) with an average correlation of .46. All item characteristic curves showed a good fit of the items.

Scale in logits	Person ability	Item difficulty
3		
	XX	
	XX	
	XXXX	
	XX	
2	XXXX	
	XXXXXX	
	XXXXXX	
	XXXXXXXX	
	XXXXXXXX	
	XXXXXXXXXXXX	
1	XXXXXXXXXXXXX	13
	XXXXXXXXXX	7 23
	XXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXX	24
	XXXXXXXXXXXXXXXXXXXX	
0	XXXXXXXXXXX	29
	XXXXXXXXXXXXXXX	9 26
	XXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXX	32
	XXXXXXXXXXXXXX	5 21
	XXXXXXXXXXXXXX	18
	XXXXXXXXXXX	12
-1	XXXXXXXXXX	33
	XXXXXXXX	
	XXXXXXXX	20 31
	XXXXXXXX	6 11 28 30
	XXXXXX	4 17
	XXXXXX	
-2	XXXX	
	XXXX	22
	XX	3 14
	XX	19 27
	XX	
	XX	10 16
	XX	2 8 15
-3		
		1
-4		

Figure 6: Test targeting. Distribution of person ability (left side of the graph) and item difficulties (right side of the graph). Each 'X' represents 17 cases. Each number represents an item (which corresponds to the item position depicted in Table 4).

5.3.4 Differential item functioning

We checked for test fairness for different groups (i.e., measurement invariance) by estimating the amount of differential item functioning (DIF). DIF was investigated for the variables test position, gender, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012a, for a description of these variables). In contrast to other cohorts, no DIF for school type was estimated. This is due to the fact that the school tracking system does not always start before grade 5 in all Federal States. Table 8 shows the difference between the estimated item difficulties in different groups. "Male vs. female", for example, indicates the difference in difficulty β(male) – β(female). A positive value indicates a higher difficulty for males, a negative value a lower difficulty for males as opposed to females.

The reading competence test was administered in two different positions (see section 3.1 for the design of the study). 2,604 (50.1%) persons received the reading test before the mathematics test (position 1), and 2,589 (49.9%) persons received the reading test after having completed the mathematics test. The subjects were randomly assigned to either of the two design groups. Differential item functioning of the position of the test may, for example, occur if there are differential fatigue effects for certain items. The results show a small average effect of item position. Subjects who received the reading test before the mathematics test performed on average 0.232 logits (Cohen's d = 0.199) better than subjects who received the reading test after the mathematics test[3]. There was no DIF due to the position of the test in the booklet. The highest difference in difficulty between the two design groups was 0.418 logits.

DIF was also investigated for gender. 2,512 (48.4%) test takers were female and 2,679 (51.6%) were male. There were two missing responses on the variable gender. These cases were excluded from the analysis. On average, male students had a lower reading ability than female students (main effect = -0.172 logits, Cohen's d = 0.147). There was no item with a considerable gender DIF. The highest difference in difficulties between the two groups was .386 logits.

The number of books at home was used as a proxy for socioeconomic status. There were 2,143 (41.3%) test takers with 0 to 100 books at home, 2,722 (52.4%) test takers with more than 100 books at home, and 328 (6.3%) test takers without a valid response. DIF was investigated using these three groups. There were considerable average differences between the three groups. Participants with 100 or less books at home performed on average 0.639 logits (Cohen's d = 0.578) lower in reading than participants with more than 100 books. Participants without a valid response on the variable 'books at home' performed 0.39 logits (Cohen's d = 0.353) or 1.029 logits (Cohen's d = 0.931) worse than participants with up to 100 and more than 100 books, respectively. There is no considerable DIF comparing participants with many or fewer books (highest DIF = 0.483). Comparing the group without valid responses to the two groups with valid responses, DIF occurred up to 0.895 logits. This is a rather large difference, which may, however, also be the result of the uncertainty in estimation due to the small number of persons with missing responses.

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

There were 3,532 (68%) participants without a migration background, 1,305 (25.11%) participants with a migration background, and 356 (6.9%) participants without a valid response. All three groups were used for investigating DIF of migration. There was a considerable difference in the average performance of participants with and those without migration background. Participants without a migration background had a higher reading ability than participants with a migration background (main effect = 0.580 logits, Cohen's d = 0.508). Also, subjects with missing values on migration differed from those without a migration background (main effect = 0.475 logits, Cohen's d = 0.417); they did not differ much from subjects with a migration background (main effect = -0.105 logits, Cohen's d = -0.092). There is no considerable DIF. The highest difference in difficulties between groups was 0.484 logits.

Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF to those that only estimate main effects. In Table 7, the models including only main effects are compared with those that additionally estimate DIF. Akaike's (1974) information criterion (AIC) 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, 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 four DIF variables.

Table 7: Comparison of models with and without DIF

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

Most of the differences in item difficulties estimated via the DIF-analyses are in absolute values below 0.5. Higher values only occur for participants with a missing value on the number of books compared to the other two groups. The comparison of a model that allows for DIF to a model without modeling DIF supports these findings. Overall, the results indicate that there is no considerable DIF and the test is fair towards the considered groups.

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

Item	Booklet	Gender	Books			Migration status		
	Position 1 vs. 2	Male vs. female	<100 vs. >100	<100 vs. missing	>100 vs. missing	Without vs. with	Without vs. missing	With vs. missing
reg50110_c	-0.108	0.242	0.024	-0.102	-0.126	0.011	-0.096	-0.107
reg5012s_c	-0.138	0.296	0.093	-0.332	-0.425	-0.059	-0.220	-0.161
reg50130_c	-0.418	0.062	0.042	0.063	0.021	0.054	-0.063	-0.117
reg50140_c	-0.098	-0.086	0.207	0.162	-0.045	0.050	-0.154	-0.204
reg50150_c	0.078	-0.090	-0.004	0.025	0.029	0.019	0.140	0.121
reg5016s_c	0.030	-0.032	0.211	0.323	0.112	0.015	-0.135	-0.150
reg50170_c	0.218	-0.078	0.199	0.010	-0.189	-0.182	-0.070	0.112
reg50210_c	0.174	0.200	0.275	-0.620	-0.895	-0.132	-0.327	-0.195
reg50220_c	0.238	0.056	-0.393	0.102	0.495	0.204	0.123	-0.081
reg50230_c	-0.396	0.248	0.100	-0.100	-0.200	0.048	-0.162	-0.210
reg50240_c	-0.074	0.272	0.101	0.136	0.035	-0.216	-0.360	-0.144
reg50250_c	-0.058	0.188	-0.095	0.399	0.494	0.026	0.004	-0.022
reg5026s_c	-0.062	0.082	0.155	-0.125	-0.280	0.187	-0.085	-0.272
reg50310_c	-0.138	-0.126	0.209	-0.089	-0.298	-0.292	-0.119	0.173
reg50320_c	-0.404	-0.284	0.143	-0.074	-0.217	-0.305	-0.143	0.162
reg50330_c	0.060	-0.124	0.301	-0.007	-0.308	-0.484	-0.466	0.018
reg50340_c	-0.162	0.226	0.165	0.045	-0.120	0.205	-0.100	-0.305
reg50350_c	-0.036	0.234	-0.059	0.368	0.427	0.075	0.033	-0.042
reg50360_c	-0.312	0.292	-0.008	-0.366	-0.358	-0.056	-0.349	-0.293
reg50370_c	-0.050	0.116	-0.146	-0.079	0.067	0.080	0.247	0.167

Item	Booklet	Gender	Books			Migration status		
	Position 1 vs. 2	Male vs. female	<100 vs. >100	<100 vs. missing	>100 vs. missing	Without vs. with	Without vs. missing	With vs. missing
reg50410_c	0.256	-0.360	0.046	0.638	0.592	-0.022	0.243	0.265
reg5042s_c	-0.014	-0.200	-0.089	0.296	0.385	-0.030	0.061	0.091
reg50430_c	0.134	-0.096	0.217	0.143	-0.074	-0.205	-0.020	0.185
reg50440_c	0.168	-0.016	0.077	0.321	0.244	0.143	0.412	0.269
reg50460_c	0.110	0.012	-0.075	0.294	0.369	0.157	0.356	0.199
reg50510_c	-0.228	0.076	0.198	-0.207	-0.405	-0.145	-0.297	-0.152
reg5052s_c	-0.096	0.130	0.381	-0.065	-0.446	-0.449	-0.272	0.177
reg50530_c	0.132	-0.284	0.209	0.612	0.403	-0.187	-0.002	0.185
reg50540_c	0.014	-0.132	0.483	0.402	-0.081	-0.247	-0.326	-0.079
reg5055s_c	-0.032	-0.064	0.332	0.403	0.071	-0.258	-0.048	0.210
reg50560_c	0.146	-0.386	-0.443	-0.298	0.145	0.437	0.295	-0.142
reg50570_c	0.058	-0.150	0.173	-0.008	-0.181	-0.303	-0.117	0.186
Main effect	0.232	-0.172	-0.639	0.390	1.029	0.580	0.475	-0.105

5.3.5 Rasch-homogeneity

In order to test the assumption of Rasch-homogeneity, we also fitted a generalized partial credit model (2PL) to the data. The estimated discrimination parameters are depicted in Table 5. They ranged from 0.33 (item reg5042s_c) to 2.00 (item reg5012s_c). The discriminations differed considerably among the items, and the 2PL model (AIC = 139382.98, BIC = 139992.60, number of parameters = 93) fitted the data better than the partial credit model (1PL) (AIC = 141297.52, BIC = 141697.38, number of parameters = 61). Nevertheless, the theoretical aim was to construct a test that equally represents the different aspects of the framework (see Pohl & Carstensen, 2012a, 2012b, for a discussion of this issue), and thus the partial credit model was used to preserve the item weightings intended in the constructional framework.

5.3.6 Unidimensionality and local item independence

The unidimensionality of the test was investigated by specifying two different multidimensional models. The first model was based on the three different cognitive requirements, and the second model was based on the five different text types.

To estimate a multidimensional (MD) model based on the three cognitive requirements, the Gauss-Hermite quadrature estimation implemented in ConQuest was used. The variances and correlations of the three dimensions are given in Table 9. All three dimensions showed a substantial variance with the highest discrimination between test takers for “Finding information in the text” and the lowest for “Reflecting and assessing”. The correlations between the three dimensions were very high (>0.95), indicating the unidimensionality of the construct. Although the correlations were very high, the three-dimensional model (AIC = 140458.87, BIC = 140780.07, number of parameters = 49) fitted the data better than the unidimensional model (AIC = 140607.24, BIC = 140895.67, number of parameters = 44). 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.

Table 9: Results of the 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 = 9)	1.904		
Drawing text-related conclusions (Dim 2) (Nitems = 12)	0.977	1.471	
Reflecting and assessing (Dim 3) (Nitems = 11)	0.952	0.952	1.178

To estimate a five dimensional model based on text functions the Monte Carlo estimation in ConQuest was used. The results of the analyses are depicted in Table 10. All five dimensions showed a substantial variation. The correlations between the dimensions varied between 0.721 and 0.875. The lowest correlations were found between dimension 4 (commenting function) and dimensions 2 (instruction texts) and 3 (advertising texts). The correlations do differ from a perfect correlation (i.e., they are considerably lower than .95, see Carstensen,

in press), indicating that there are subdimensions measured with the test. The five-dimensional model (AIC = 139992.23, BIC = 140372.43, number of parameters = 58) fitted the data better than the unidimensional model (AIC = 140607.24, BIC = 140895.67, number of parameters = 44). Note that the amount of missing responses becomes rather high in the last texts. This may result in a reduced variation and, thus, in a decreased correlation. Also note that the text types are perfectly confounded with the texts. There is one text for each text function, and, since a set of items refers to the same text, local item dependence (LID) occurs. The correlations depicted in Table 10 are, thus, due to multidimensionality based on text functions as well as due to local item dependence. Given the testing design in the main studies, it is not possible to disentangle the two sources. In pilot studies (Gehrer et al., 2012), a larger number of texts were presented to test takers, so that the impact of text functions could be investigated independently of LID. The correlations estimated in the pilot study ranged from .78 to .91. Although a different scaling model has been used in this paper, the results give a first idea about the impact of the text function (unconfounded with LID) on the dimensionality of the test. As the correlations found in Gehrer et al. (2012) differ from a perfect correlation, it is concluded that text functions form subdimensions of reading competence. Comparing the correlations found in Gehrer et al. (2012), which are due to text functions, to those found in the main study (Table 10), which are due to both text functions and LID, allows to evaluate the impact of LID. The correlations found in the present study of starting cohort 3 were lower (between 0.72 and 0.88) than those found in Gehrer et al. (between 0.78 and 0.91), indicating that there is some amount of local item dependence. Due to substantial considerations, Gehrer et al. argue for a unidimensional construct. Consequently, a single competence score is estimated for reading competence.

Table 10: Results of five-dimensional scaling. Variance 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**
Information (Dim 1) (Nitems = 7)	2.091				
Instruction texts (Dim 2) (Nitems = 6)	0.875	1.403			
AdvertisingTexts (Dim 3) (Nitems = 7)	0.849	0.881	2.008		
Commenting function (Dim 4) (Nitems = 5)	0.819	0.730	0.721	1.362	
Literary function (Dim 5) (Nitems = 7)	0.864	0.841	0.867	0.809	1.846

6. Discussion

The analyses in the previous sections aimed at providing detailed information on the quality of the reading test in starting cohort 3 and at describing how the reading competence score is estimated.

We investigated different kinds of missing responses and examined the item and test parameters. We thoroughly checked item fit statistics for simple MC items, subtasks of CMC and MA items, as well as the aggregated polytomous CMC and MA items, and examined the

correlations between correct and incorrect responses and the total score. Further quality inspections were conducted by examining differential item functioning, testing Rasch-homogeneity, investigating the tests' dimensionality as well as local item dependence.

Indicated by various fit indices the items show a good item fit and measurement invariance across various subgroups. The amount of not-reached items is higher than expected. This indicates that the test is too long for the given testing time. Besides the not-reached items, the number of other kinds of missing responses is reasonably small.

The test has a high reliability. The items distinguish well between the test takers. However, the test is mainly targeted at low-performing students and does not sufficiently measure reading competence of high-performing students. As a consequence, ability estimates will be very precise for low-performing students but less precise for high performing students.

Unidimensionality of the test could be confirmed for the different comprehension requirements. Multidimensionality, however, is present for different text functions. In combination with the high amount of missing responses at the end of the test (i.e., there are students 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. Nevertheless, 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.

Summarizing these results, the test has good psychometric properties that facilitate the estimation of a unidimensional reading competence score.

7. Data in the Scientific Use File

There are 32 items in the data set that are either scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response, or scored as polytomous variables (CMC and MA items) indicating the (partial) credit. 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'. Note that the values of the polytomous variables in the Scientific Use File do not necessarily indicate the number of correctly responded subtasks, as categories may have been collapsed within CMC and MA items (see section 4.2 for a description of the aggregation of CMC and MA items). A scoring of 0.5 points per category has been used in the scaling model for the aggregated CMC and MA items (besides for one item, see section 4.2 and the ConQuest-Syntax in Appendix A). Manifest scale scores are provided in the form of WLE estimates (reg5_sc1) including the corresponding standard error (reg5_sc2). Note that for the estimation of the WLE scores, the effect of test position in the booklet is controlled for. The ConQuest Syntax for estimating the WLE scores from the items is provided in Appendix A. Students that did not take part in the test or those that do not have enough valid responses to estimate a scale score have a not-determinable missing value on the WLE score for reading competence.

Plausible values that allow for an investigation of 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 (2012a).

References

Akaike, H. (1974). A new look at the statistical model indentification. *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. (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. (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 3

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

data filename.dat;
format pid 4-10 responses 13-44 position 46; /* insert number of columns with data*/

labels << filename_with_labels.txt;

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

score (0,1) (0,1)                       !items (1,3-5,7-12,14-21,23-26,28-29,31-32);
score (0,1,2) (0,0.5,1)                 !item (2);
score (0,1,2,3,4,5) (0,0.5,1,1.5,2,2.5) !item (6);
score (0,1) (0,1)                       !item (13);
score (0,1,2,3) (0,0.5,1,1.5)           !item (22);
score (0,1,2,3) (0,0.5,1,1.5)           !item (27,30);

set constraint=cases;

model item + item*step + position;
estimate;

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