NEPS WORKING PAPERS

Steffi Pohl, Kerstin Haberkorn, & Katinka Hardt

NEPS TECHNICAL REPORT FOR READING – SCALING RESULTS OF STARTING COHORT 5 FOR FIRST-YEAR STUDENTS IN MAIN STUDY 2010/11

NEPS Working Paper No. 34
Bamberg, March 2014

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

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

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

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

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

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

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de

NEPS Technical Report for Reading—Scaling Results of Starting Cohort 5 for First-Year Students in Main Study 2010/11

Steffi Pohl, Freie Universität Berlin
Kerstin Haberkorn, Leibniz Institute for Educational Trajectories
Katinka Hardt, University of Luxembourg

Email address of the lead author:

steffi.pohl@fu-berlin.de

Bibliographic data:

Pohl, S., Haberkorn, K., & Hardt, K. (2014). *NEPS Technical Report for Reading – Scaling results of Starting Cohort 5 for first-year students* (NEPS Working Paper No. 34). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

NEPS Technical Report for Reading—Scaling Results of Starting Cohort 5 for First-Year Students in Main Study 2010/11

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competencies across the whole life span and developing tests for assessing different competence domains. In order to evaluate the quality of the competence tests, a wide range of analyses have been performed based on item response theory (IRT). This paper describes the data and scaling procedures of the students' reading competence data in Starting Cohort 5. The reading competence test for the students contains 29 reading items with different response formats representing different cognitive requirements and text functions. The test was administered to 7,085 students and the data were scaled using the partial credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, the tests' dimensionality, and local item independence were evaluated to ensure the quality of the test. The results showed that the test exhibits an acceptable reliability and that the items fit the model in a satisfactory way. Furthermore, test fairness could be confirmed for different subgroups. Challenges of the test include the large number of items targeted toward a lower reading ability as well as the large percentage of items at the end of the test that were not reached due to time limits. Further challenges arise from dimensionality analyses based on both text functions and cognitive requirements. Overall, the reading test had acceptable psychometric properties and results of the quality of the scale support the estimation of a reliable reading competence score. Besides scaling results, this paper describes the data available in the Scientific Use File and presents the ConQuest-syntax for scaling the data.

Keywords

Item response theory, scaling, reading competence, Scientific Use File, university and college students

Contents

1. Introduction

Within the National Educational Panel Study (NEPS) different competencies are measured coherently across the life span. Tests have been developed for different competence domains. These include, among other things, reading competence, mathematical competence, scientific literacy, information and communication technologies literacy, metacognition, vocabulary, and domain-general cognitive functioning. Weinert et al. (2011) give an overview of the competencies measured in NEPS.

Most of the competence data are scaled using models that are based on item response theory (IRT). Because most of the competence tests were developed specifically for implementation in NEPS, several analyses have been conducted to evaluate the quality of the tests. The IRT models chosen for scaling the competence data and the analyses performed for checking the quality of the scales are described in Pohl and Carstensen (2012). In this paper the results of these analyses are presented for reading competence in the first wave of Starting Cohort 5 – First-Year Students at university. We will first introduce the main concepts of the reading competence test. Then, we will describe the reading competence data of Starting Cohort 5 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 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 (2013). In the following, we will point out specific aspects of the reading competence test that are necessary for understanding the scaling results presented in this paper.

The theoretical framework of the reading test comprises five text types or text functions, namely, 1. information texts, 2. commenting or argumenting texts, 3. literary texts, 4. instruction texts, and 5. advertising texts. Furthermore, 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 & Artelt, 2013; Gehrer et al., 2013;- and Weinert et al., 2011, for a detailed description of the framework).

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

assign headings to paragraphs of a text. Examples of the different response formats are provided in Pohl and Carstensen (2012) and Gehrer, Zimmermann, Artelt and Weinert (2012).

3. Data

3.1 The Design of the Study

In the main study 2010/11, reading speed, reading competence, and mathematical competence, as well as procedural metacognition were assessed. In order to control for position and order effects, the reading and the mathematical competence test were administered to participants in different order. Half of the subjects received a booklet that first contained the reading test followed by the mathematics test, whereas the other half of the sample completed the two tests in reverse order. The subjects were randomly assigned to one of the two booklets. Note that there was no multi-matrix design regarding the choice and the order of the items *within* a specific test. All subjects received the same set of reading items in the same order.

The students' reading test consisted of 36 items that represented different cognitive requirements and text functions and featured different response formats. Prior to final scaling, extensive analyses had been conducted in order to detect any items with insufficient psychometric properties. The items referring to the last text in the reading test showed a large amount of missing responses, because the participants did not reach these items within the constraints of testing time[1]. Due to the limited information available on these seven items of the last text, these items were excluded from the final scaling model. Furthermore, one item showed considerable differential item functioning and was, thus, also excluded from the final scaling.

The scaling results presented in the following sections are based on the remaining items, that are 28 items referring to one of four texts. The characteristics of these items are described in Tables 1 to 3. Table 1 shows the distribution of the cognitive requirements, Table 2 the distribution of text functions, and Table 3 the response formats used. The number of subtasks within CMC and MA items varied between three and six.

Table 1

Cognitive Requirements of Items in the Reading Test for First-Year Students

Cognitive requirement	Frequency
Finding information in text	6
Drawing text-related conclusions	14
Reflecting and assessing	8
Total number of items	28

[1] Missing responses as a result of not reaching items of the last text ranged from 65.86% to 85.86%.

Table 2

Number of Items Referring to a Specific Text Type in the Reading Test for First-Year Students

Text types/functions	Frequency
Instruction text	7
Advertising text	5
Commenting or argumenting text	8
Literary text	8
Total number of items	28

Table 3

Response Formats of Items in the Reading Test for First-Year Students

Response format	Frequency
Simple multiple choice	22
Complex multiple choice	5
Matching	1
Total number of items	28

3.2 Sample

A description of the study design, the sample, and instruments used can be found on the NEPS website.[2] In total, 7,085 subjects took the reading competence test. 3,522 of them received the mathematics test before the reading test; 3,563 subjects first completed the reading competence test and then took the mathematics competence test. Six of the 7,085 subjects gave less than three valid responses to the reading items. Because no reliable reading competence score may be estimated on the basis of such a low number of valid responses, these cases were excluded from further analyses. Thus, a sample of 7,079 persons underlies the results presented in the following sections. Note that data of about 1000 subjects are not available in the SUF due to ongoing data cleaning issues.

4. Analyses

4.1 Missing Responses

Within the reading test, there are several kinds of missing responses. These are missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach, d) items that have not been administered, and e) multiple kinds of missing responses within CMC or MA items that are not determinable.

Invalid responses occurred, for example, when two response options were selected in simple MC items where just one was required, or when numbers or letters that are not within the range of valid responses were given as a response. Items were coded as omitted when subjects skipped a particular item. Due to the limited testing time, some subjects did not complete the entire test. Items that were not completed at the end of the test were labeled

[2] www.neps-data.de

as not reached. Because CMC and MA items consist of a number of subtasks, a mixture of different types of missing responses and/or a mixture of missing and valid responses might be found. As soon as one subtask contained a missing response, the CMC or MA item was coded as missing. When just one kind of missing response occurred, the item was coded according to the corresponding missing response. When the subtasks contained different kinds of missing responses, the item value was coded as a not-determinable missing response.

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

4.2 Scaling Model

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

CMC and MA items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC or MA item, indicating the number of correctly solved 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. Low frequencies of categories also occurred for matching items with perfect local dependence. In these cases the two highest scores were collapsed into one category (see Pohl & Carstensen, 2012, for an explanation of this approach). For each 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 indicate the number of correctly solved subtasks but should rather be interpreted as (partial) credit scores.

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

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

4.3 Checking the Quality of the Test

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

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

The MC, CMC, and MA items consisted of one correct response and one or more distractors (incorrect response options). We investigated the performance of distractors, that is, whether they were predominantly chosen by subjects with a lower ability rather than by those who gave a correct response. We evaluated the point biserial correlation between the incorrect responses and the total score treating all subtasks of CMC and MA items as single items. We judged correlations below zero as very good, correlations below 0.05 as acceptable, and correlations above 0.05 as problematic.

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

The authors of the test (Gehrer et al., 2013) aim at constructing a reading competence test that measures the same construct for all participants. If there were any items that favored certain subgroups (e. g., that were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between the subgroups (e. g., males and females) would be biased and, thus, unfair. We addressed the issue of measurement invariance by investigating test fairness for the variables test position, gender, and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Differential item functioning was estimated using a multigroup IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled. Differences in the estimated item difficulties between the subgroups were evaluated. On the basis of experiences with preliminary data, we judged absolute differences in estimated difficulties that are greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as noteworthy for further investigation, differences

between 0.4 and 0.6 as considerable but not severe, and differences smaller than 0.4 as no considerable DIF. In addition to DIF analyses on item level, test fairness was investigated by comparing a model including differential item functioning to a model that only estimated main effects and no DIF.

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

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

Because the reading test consisted of item sets that referred to one of four texts, the assumption of local item dependence (LID) may not necessarily hold. Additionally, the four texts were perfectly confounded with the four 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

Figure 1 depicts the number of invalid responses per person. Overall, the number of invalid responses was very low; as can be seen, the vast majority of 88.57% of respondents did not give any invalid responses at all and less than four percent had more than one invalid response. Missing responses may also occur when persons omit particular items. As can be seen in Figure 2, the majority of the subjects — almost 63% — did not skip any item at all, only about 20% of the subjects omitted more than one item, and around 10% omitted more than three items of the reading test.

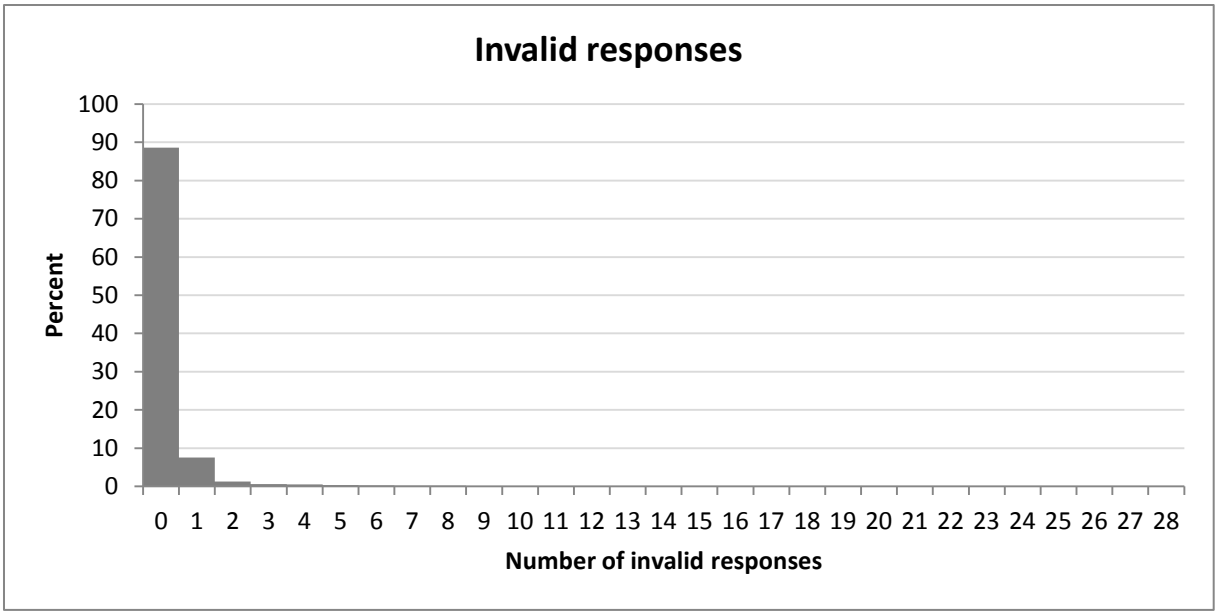

Figure 1. Number of invalid responses

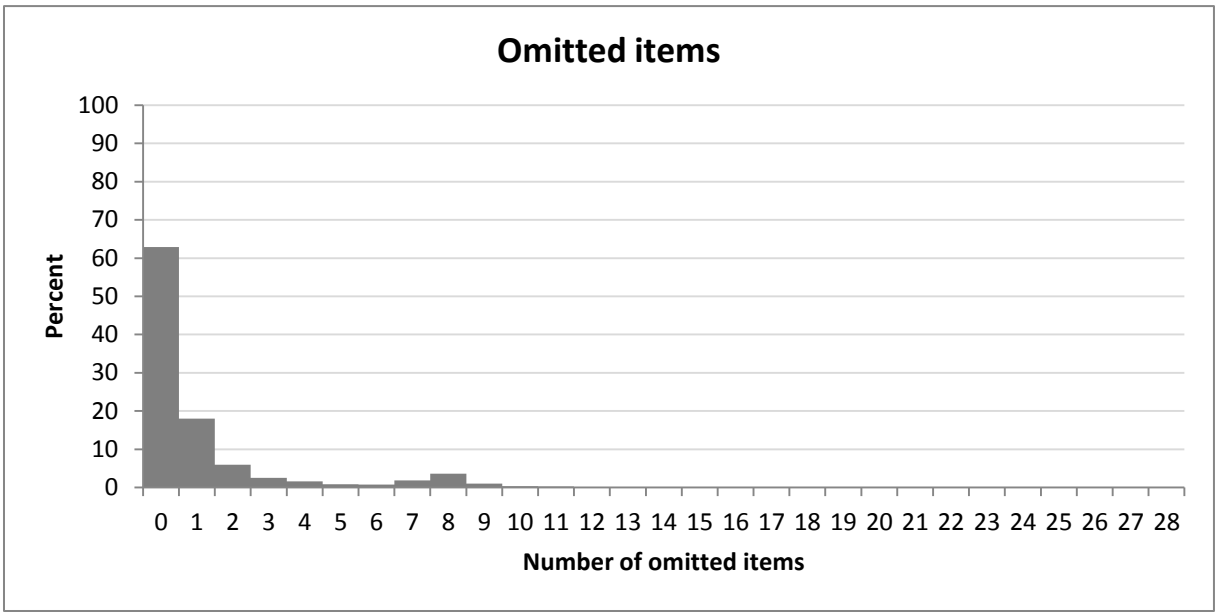

Figure 2. Number of omitted items

Another source of missing responses are items that were not reached by the subjects. Note that the items of the last text in the test were excluded from the analyses presented here. The amount of not-reached items for items of the last text was rather high (65.86% to 85.86%). Regarding the items used in the final scaling model, more than 75% of the students responded to items of the fourth text and around 60% of the subjects managed to reach the last item of this text. All respondents—with the exception of 2.73%—reached the items of the first two texts.

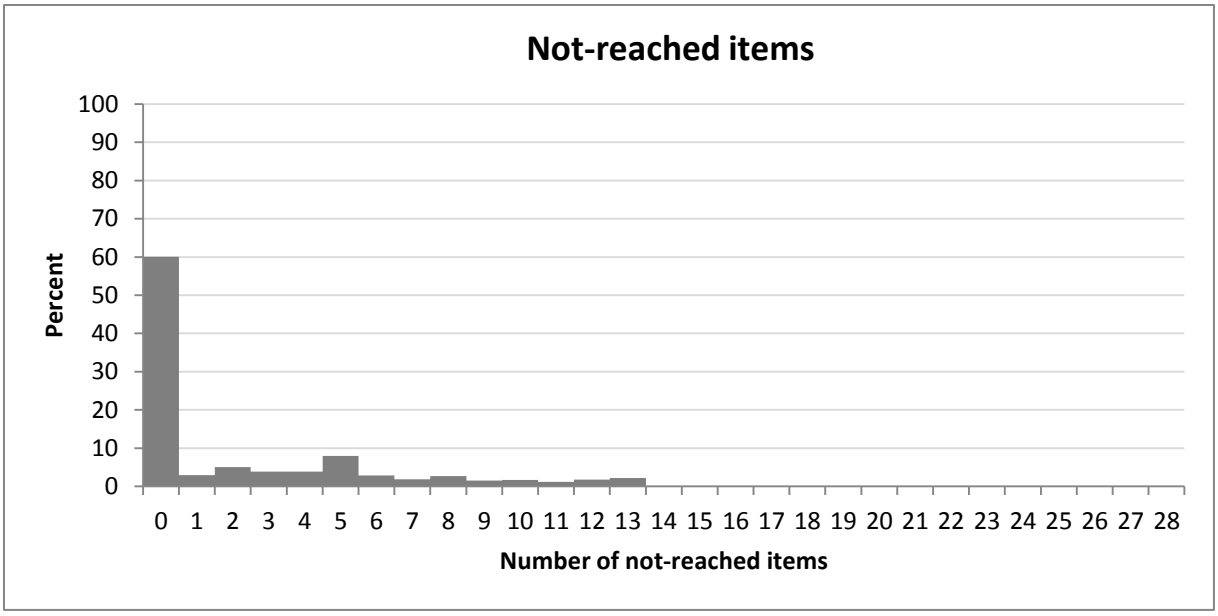

Figure 3. Number of not-reached items

When the subtasks of CMC and MA items within aggregated polytomous variables contained different kinds of missing responses, they were coded as a not-determinable missing response. Because 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). Only a small amount of not-determinable missing responses occurred (Figure 4). 98.95% of the subjects had no not-determinable missing response at all.

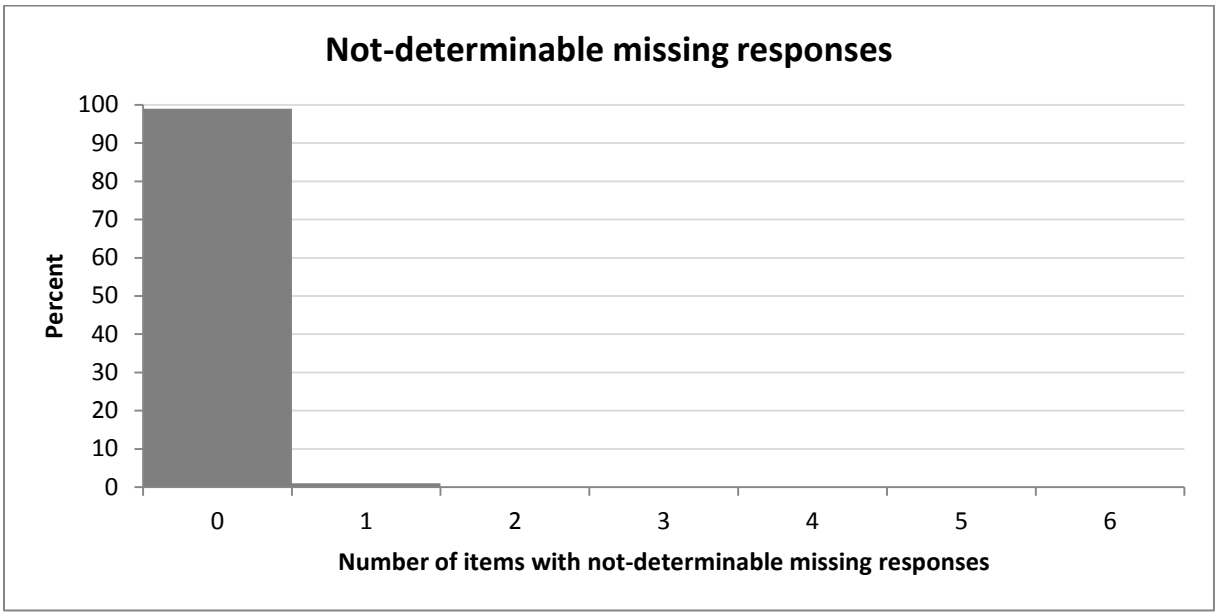

Figure 4. Number of not-determinable missing responses

The total number of missing responses, aggregated over invalid, omitted, not-reached, and not-determinable missing responses per person, is depicted in Figure 5. Nearly one third of the subjects had no missing responses at all to the 28 items in the final scaling model.

47.76% of the sample had more than two missing responses. The average number of missing responses per person was 3.75 (SD = 4.16).

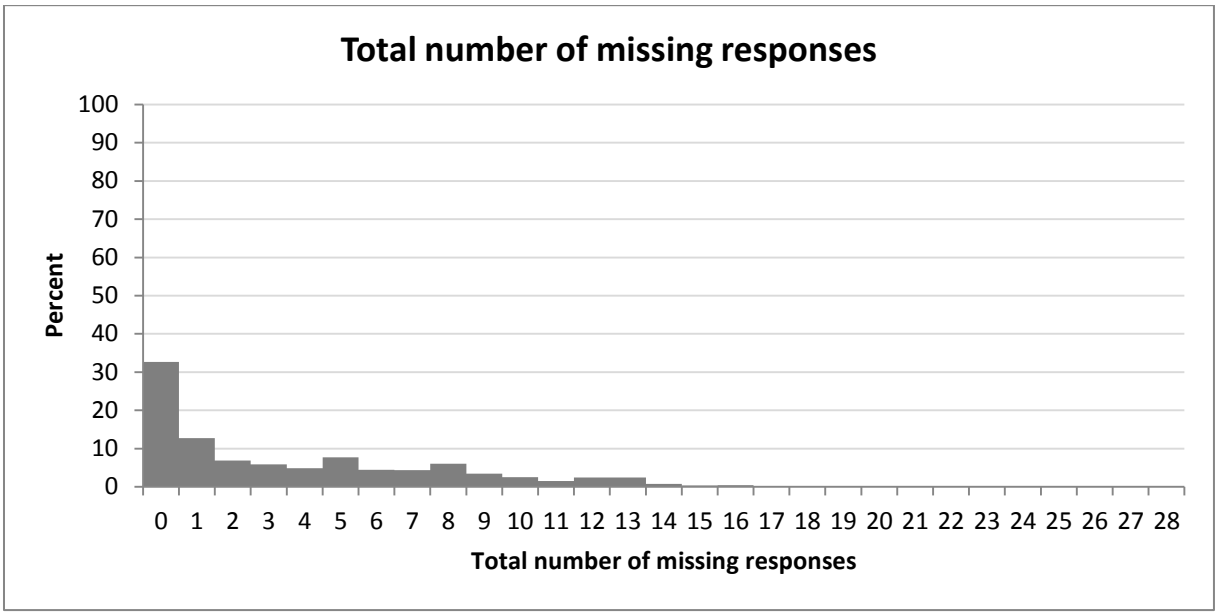

Figure 5. Total number of missing responses

Summarizing the results on the occurrence of missing responses, there is a small amount of invalid and not-determinable missing responses and a reasonable amount of omitted items. The number of not-reached items is, however, large.

5.1.2 Missing responses per item

Table 4 provides information on the occurrence of different kinds of missing responses per item. The omission rate is acceptable, varying across items between 0.25% (res10110_c) and 15.34% (res1032s_c). There were ten items with an omission rate exceeding 5%. On average, with 9.45%, CMC items had a higher omission rate than MC items (3.25%) and MA items (2.60%). Omission rate correlated with item difficulty to .12; the correlation increased to .32 when four items with bivariate extreme values were excluded. Participants are inclined to omit more difficult items. With an item's progressing position in the test, the amount of persons that did not reach the item (column 4) rose up to a considerable amount of 39.98% (for the last item res10450_c of the fourth text)[3]. On the contrary, the percentage of invalid responses per item (column 5) was rather low with the maximum rate being 4.22% (item res10110_c). Matching items seemed to be slightly more prone to invalid responses than multiple choice items in both single and complex form.

[3] Note that the items of the last text were excluded from final analyses and, thus, from the presented quality checks due a large amount of not reached items.

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 %
res10110_c	1	6,762	0.00	0.25	4.22
res1012s_c	2	6,760	0.00	4.45	0.03
res10130_c	3	6,980	0.00	0.49	0.90
res10140_c	4	7,035	0.00	0.49	0.13
res10160_c	6	6,857	0.00	0.35	2.78
res10170_c	7	6,920	0.03	0.93	1.29
res10180_c	8	7,023	0.03	0.41	0.35
res10190_c	9	7,025	0.04	0.45	0.27
res1021s_c	10	6,158	0.10	12.71	0.18
res1022s_c	11	6,700	0.11	2.02	3.15
res10230_c	12	6,941	0.18	0.86	0.90
res1024s_c	13	6,676	0.25	5.28	0.14
res10250_c	14	6,813	0.37	1.58	1.81
res10260_c	15	6,857	0.42	1.17	1.54
res10270_c	16	6,858	0.52	1.68	0.92
res10310_c	17	6,488	2.73	5.18	0.44
res1032s_c	18	5,652	4.46	15.34	0.07
res10330_c	19	6,060	5.64	7.90	0.86
res10340_c	20	5,918	7.29	8.33	0.78
res10350_c	21	5,797	8.83	8.35	0.93
res10360_c	22	5,545	11.50	9.49	0.68
res10370_c	23	5,395	13.38	9.63	0.78
res10380_c	24	5,177	16.27	10.51	0.08
res10410_c	25	5,240	24.24	0.57	1.17

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 %
res10420_c	26	4,980	28.11	1.37	0.17
res1043s_c	27	4,536	32.01	3.18	0.10
res10440_c	28	4,378	37.01	0.82	0.31
res10450_c	29	4,168	39.98	0.72	0.41

Note. The items in position 5 as well as in positions 30 to 36 were excluded from the analyses due to unsatisfactory item fit or a large amount of missing responses (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
res10110_c	74.02	-1.154	0.029	0.99	-0.9	0.39	1.01
res1012s_c	n. a.	-4.048	0.075	0.97	-1.1	0.30	1.19
res10130_c	96.15	-3.441	0.063	0.99	-0.1	0.20	0.62
res10140_c	67.18	-0.793	0.027	1.04	3.3	0.32	0.46
res10160_c	47.28	0.124	0.026	1.06	7.2	0.29	0.89
res10170_c	74.32	-1.172	0.029	1.01	0.6	0.35	1.15
res10180_c	90.45	-2.441	0.042	0.99	-0.4	0.29	0.61
res10190_c	53.65	-0.162	0.025	1.04	5.7	0.32	0.79
res1021s_c	n. a.	-2.538	0.043	1.00	0.1	0.25	0.85
res1022s_c	n. a.	-3.151	0.049	0.95	-1.7	0.37	0.55
res10230_c	66.66	-0.768	0.027	1.02	1.5	0.35	1.00
res1024s_c	n. a.	-1.384	0.037	0.99	-0.8	0.31	1.89
res10250_c	64.08	-0.639	0.027	1.01	0.7	0.37	1.34
res10260_c	84.92	-1.892	0.035	1.04	1.8	0.22	0.12
res10270_c	73.61	-1.132	0.029	0.99	-0.6	0.38	1.31
res10310_c	85.14	-1.909	0.036	0.94	-2.8	0.43	1.11
res1032s_c	n. a.	-0.533	0.029	0.97	-2.1	0.44	1.34
res10330_c	77.95	-1.384	0.032	0.97	-1.6	0.40	1.78
res10340_c	44.47	0.259	0.028	1.12	13.5	0.19	0.67
res10350_c	56.34	-0.274	0.028	0.96	-4.8	0.46	0.81
res10360_c	67.88	-0.822	0.030	0.98	-1.4	0.40	0.29
res10370_c	61.39	-0.504	0.029	0.96	-3.9	0.45	0.66
res10380_c	85.82	-1.949	0.041	0.94	-2.1	0.41	2.56

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
res10410_c	85.10	-1.896	0.040	1.02	0.8	0.26	0.83
res10420_c	91.81	-2.603	0.053	1.01	0.2	0.22	2.03
res1043s_c	n. a.	-1.197	0.025	0.94	-3.1	0.56	1.06
res10440_c	68.30	-0.838	0.034	1.09	6.4	0.22	1.28
res10450_c	77.38	-1.344	0.039	1.03	1.4	0.29	1.29

Note. Percentage of correct scores are not indicative of polytomous CMC and MA item scores. These are denoted by n. a.
For dichotomous items, the correlation with the total score corresponds to the point-biserial correlation between the correct response and the total score; for polytomous items, it corresponds to the product-moment correlation between the corresponding categories and the total score (discrimination value as computed by ConQuest).

Table 6.

Step Parameters (and Standard Errors) of the Polytomous Items

Item	Step 1 (*SE*)	Step 2 (*SE*)	Step 3 (*SE*)	Step 4 (*SE*)
res1021s_c	-0.513 (0.028)	0.513		
res1022s_c	0.157 (0.035)	-0.157		
res1024s_c	-0.735 (0.025)	0.735		
res1032s_c	-0.566 (0.028)	-0.261 (0.027)	0.827	
res1043s_c	0.036 (0.031)	0.173 (0.032)	0.029 (0.037)	-0.239

Note. Because item res1012s_c consists of only two categories, no step parameters are estimated.

5.2 Parameter Estimates

5.2.1 Item parameters

Column 2 in Table 5 shows the percentage of correct responses relative to all valid responses for each item. Note that because 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 MC items varied between 44.47% and 96.15% with an average of 72.45% (*SD* = 14.31%) correct responses.

For reasons of model identification, the mean of the ability distribution was constrained to be zero in the partial credit model. The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 5. The step parameters for polytomous variables are depicted in Table 6. The item difficulties ranged from -4.048 (item res1012s_c) to 0.259 (item res10340_c) logits with an average difficulty of -1.414 logits (*SD* = 1.068). Altogether, the item difficulties are very low. Owing to the large sample size, the corresponding standard errors of the estimated item difficulties (column 4) are small ($SE(\beta) \leq 0.076$).

5.2.2 Person parameters

Person parameters are estimated as WLEs and plausible values (Pohl & Carstensen, 2012). WLEs will be provided in the first release of the SUF, whereas plausible values will be provided in later SUF releases. A description of the data in the SUF can be found in Section 7. An overview of how to work with competence data is presented in Pohl and Carstensen (2012).

5.2.3 Test targeting and reliability

Test targeting focuses on the match of item difficulties and person abilities and was used to evaluate the appropriateness of the test for the specific target group. In Figure 6, item difficulties of the reading items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side, whereas the right side shows the distribution of item difficulties.

The mean of the ability distribution was constrained to be zero and the variance was estimated to be 0.502, which implies moderate differentiation between the subjects. The reliability of the test (EAP/PV reliability = .642, WLE reliability = .590) was acceptable.

Note that due to the reduced number of items used in the final scaling model, the reliability estimate in this test is lower than that of reading tests in other cohorts. Although the items covered some range of the ability distribution, the items were rather easy for this specific student sample. As a consequence, person ability in lower ability regions will be measured relative precisely, whereas medium and higher ability estimates will have larger standard errors of measurement.

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

Figure 6. Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph, each 'X' represents 45.7 cases. The difficulty of the items is depicted on the right side of the graph, each number represents an item (which corresponds to the item position indicated in Table 4).

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 to be analyzed via the partial credit model, the fit of the subtasks was checked by analyzing the single subtasks together with the simple MC items in a Rasch model. Counting the subtasks of CMC and MA items separately, there were 48 items. Because there was one matching task with perfect stochastic dependence (see Pohl & Carstensen, 2013, for a description of the problem), one of the subtasks of this MA item was excluded from the analyses. Consequently, 47 items entered the analysis.

Concerning relative frequencies, three subtasks of CMC items showed a probability of a correct response of greater than 95%. Overall, the fit of the subtasks was satisfactory. The WMNSQ ranged from 0.92 to 1.11, the corresponding *t*-values from -5.50 to 13.40. The good item fit of the subtasks was affirmed by the empirically estimated item characteristic curves. In conclusion, the satisfactory fit of the subtasks was considered sufficient to justify their aggregation to polytomous variables for each CMC and MA item. Note that CMC and MA items can be identified through the letters 's_c' at the end of the variable name, whereas the variable name of simple MC items ends on '0_c'.

5.3.2 Distractor analyses

In addition to the overall item fit (Section 5.3.3), we specifically investigated how well the distractors performed in the test by evaluating the point biserial correlation between each incorrect response (distractor) and the test takers' total score. The distractors consistently yielded correlations below zero with a range from -.480 to -.050 and a mean of -0.18. The results indicate that the distractors function properly.

5.3.3 Item fit

Item fit was additionally investigated for MC and polytomous CMC and MA items via the partial credit model. Altogether, item fit can be considered as good (see Table 5). Values of the WMNSQ ranged from 0.94 (item res10310_c, item res10380_c and item res1043s_c) to 1.12 (res10340_c); only two *t*-values of the WMNSQ exceeded a *t*-value of 7. Overall, there was no indication of severe item misfit. Point biserial correlations between the item scores and the total scores ranged from 0.19 (item res10340_c) to 0.56 (item res1043s_c) and had a mean of 0.34. All item characteristic curves showed an acceptable fit of the items.

5.3.4 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i. e., measurement invariance). For this purpose, DIF was examined for the variables test position, gender, as well as migration background. In contrast to other cohorts, DIF was investigated neither for the number of books at home as a proxy for socioeconomic status nor for the school type. The number of books at home is rather high for university students and might not necessarily reflect their socioeconomic status in a meaningful way. With regard to school type, the cohort of higher education students will be relatively homogeneous, because holding a higher education entrance qualification (Abitur) is the most prevailing precondition for studying, and there will be hardly any variance on this variable. Therefore, these variables were not used to explore the fairness of the test. Table 7

provides the results of the DIF-analyses for the remaining variables, depicting differences in the estimated item difficulties. “Male vs. female”, for example, indicates the difference in difficulty *ß*(male) - *ß*(female). A positive value indicates a higher difficulty for males, a negative value suggests a lower difficulty for males as opposed to females with the same level of ability.

The reading competence test was administered in two different positions (see Section 3.1 for the design of the study). 3,563 (50.29%) persons received the reading test before the mathematics test and 3,522 (49.71%) persons received the reading test after having completed the mathematics test. The subjects were randomly assigned to either of the two design groups. The results show no noticeable average effect of item position. Subjects who first took the reading test performed on average 0.040 logits (Cohen’s d = -0.057) worse than subjects who received the reading test after the mathematics test. Differential item functioning of the position of the test may, for example, occur, if there are differential fatigue effects for certain items. However, results show no difference in the estimated item difficulties between the different design groups exceeding 0.6 logits. The largest absolute difference in difficulties was -0.254 logits (item res10130_c).

Differential item functioning analysis for gender was based on 2,438 (34.72%) males and 4,584 (65.28%) females. For 57 cases, information on gender was missing; these cases were excluded from the DIF analysis. The results show no noticeable average effect of gender. Male participants performed on average slightly worse than females (main effect = -0.024 logits, Cohen’s d = -0.034). There was no considerable item DIF. Only two items (item res10410_c and item res10450_c) showed DIF greater than 0.4 logits.

Finally, test fairness was investigated for migration background. There were 5,135 participants (72.54%) with no migration background, 844 subjects (11.92%) with a migration background and 1,100 (15.54%) persons for whom migration background could not be determined. 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. Hence, the DIF analysis was performed comparing these three groups. Students without migration background had, on average, a higher reading ability than students with migration background (main effect = 0.248 logits, Cohen’s *d* = 0.356) and students without information on migration background (main effect = 0.298 logits, Cohen’s *d* = 0.427). There was no noticeable difference in reading ability between subjects without migration background and subjects without information on migration background (main effect = 0.050 logits, Cohen´s *d* = 0.072). Furthermore, no considerable DIF due to migration background occurred. Differences in estimated difficulties did not exceed 0.6 logits. The largest DIF was found on item res10310_c, which exhibited a higher estimated difficulty for subjects with migration background than for subjects without (absolute DIF = -0.592).

The results of the comparison of models including only main effects with models additionally allowing for DIF are displayed in Table 8. As can be seen in the table, for all three DIF variables, Akaike's (1974) information criterion (AIC) consistently favors the models estimating DIF. The Bayesian information criterion (BIC; Schwarz, 1978) takes into account the number of estimated parameters and, thus, accounts for overparametrization of models. According to the BIC, the model additionally estimating DIF was only preferred for the variable gender. Regarding position and migration background, the more parsimonious model including only the main effect was preferred over the more complex DIF model.

Table 7

Differential Item Functioning (Absolute Differences Between Item Difficulties)

Item	Booklet	Gender	Migration status		
	Read/math vs. math/read	Male vs. female	Without vs. with	Without vs. missing	With vs. missing
res10110_c	0.056	-0.050	-0.045	-0.117	-0.072
res1012s_c	0.114	0.394	-0.533	-0.378	0.155
res10130_c	-0.254	0.124	-0.224	-0.430	-0.206
res10140_c	0.184	0.280	-0.033	0.012	0.045
res10160_c	-0.194	-0.252	0.160	0.243	0.083
res10170_c	0.222	0.182	-0.027	0.103	0.130
res10180_c	0.002	0.184	-0.144	-0.144	0.000
res10190_c	0.092	-0.040	0.279	-0.093	-0.372
res1021s_c	0.238	0.340	0.338	-0.017	-0.355
res1022s_c	0.174	0.376	-0.378	-0.103	0.275
res10230_c	0.136	0.226	0.024	0.060	0.036
res1024s_c	0.082	-0.066	-0.264	-0.007	0.257
res10250_c	0.070	-0.352	0.027	0.111	0.084
res10260_c	-0.214	0.148	0.123	0.050	-0.073
res10270_c	0.062	-0.262	-0.057	0.036	0.093
res10310_c	-0.090	0.034	-0.592	-0.383	0.209
res1032s_c	-0.100	-0.078	-0.100	-0.161	-0.061
res10330_c	-0.068	-0.346	-0.190	-0.137	0.053
res10340_c	0.024	0.234	-0.148	0.239	0.387
res10350_c	-0.188	-0.188	-0.038	-0.035	0.003
res10360_c	0.052	-0.080	-0.041	-0.178	-0.137
res10370_c	-0.032	0.114	-0.022	-0.134	-0.112

Item	Booklet	Gender	Migration status		
	Read/math vs. math/read	Male vs. female	Without vs. with	Without vs. missing	With vs. missing
res10380_c	-0.146	-0.108	-0.038	-0.229	-0.191
res10410_c	-0.072	0.572	0.215	0.247	0.032
res10420_c	-0.172	-0.206	-0.040	-0.191	-0.151
res1043s_c	-0.052	0.066	-0.171	-0.099	0.072
res10440_c	-0.100	0.120	0.395	0.241	-0.154
res10450_c	-0.086	-0.510	0.249	0.009	-0.240
Main effect	-0.040	-0.024	0.248	0.298	-0.050

Table 8

Comparison of Models With and Without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Booklet	main effect	202502.397	38	202578.397	202839.263
	DIF	202394.080	66	202526.080	202979.163
Gender	main effect	200599.630	38	200675.630	200936.189
	DIF	200273.229	66	200405.229	200857.778
Migration	main effect	202365.298	39	202443.298	202711.028
	DIF	202182.713	95	202372.713	203024.877

Summarizing the results of DIF examination, all of the differences in item difficulties estimated via DIF-analyses are in absolute values below 0.6 and the BIC indicates a better model fit for the models without estimating DIF for almost all variables. Thus, there is no substantial indication of test unfairness.

5.3.5 Rasch-homogeneity

One essential assumption of the Rasch model is Rasch-homogeneity. Rasch-homogeneity implies that all item discrimination parameters are equal. In order to test this assumption, a generalized partial credit model (2PL) that estimates the discrimination parameter was fitted to the data. Fourteen of 28 items showed a discrimination between 0.80 and 1.20. However, some of the items exhibited a rather high or low discrimination ranging from 0.12 to 2.56 (see Table 5). Model fit indices suggested a better model fit of the 2PL model (AIC = 201125.73, BIC = 201695.51, number of parameters = 83) compared to the 1PL model (AIC = 202610.31, BIC = 202987.88, number of parameters = 55). Despite the empirical preference for the 2PL model, the 1PL model more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, for a discussion of this issue). For this reason, the partial credit model (1PL) was chosen as scaling model to preserve the weighting of items as intended in the constructional framework.

5.3.6 Unidimensionality and local item independence

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

Estimation of the three-dimensional model was carried out with ConQuest using the Gauss-Hermite quadrature method. The estimated variances and correlations between the three dimensions that represent the different cognitive requirements are reported in Table 9.

Table 9

Results of Three-Dimensional Scaling. Variances of the Dimensions are depicted in the Diagonal; Correlations are given in the Off-Diagonal.

	Dim 1	**Dim 2**	**Dim 3**
Finding information in the text (Dim 1) (Nitems = 6)	0.544		
Drawing text-related conclusions (Dim 2) (Nitems = 14)	0.892	0.453	
Reflecting and assessing (Dim 3) (Nitems = 8)	0.851	0.936	0.771

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

All three dimensions had substantial variance estimates with the highest obtained for *reflecting and* assessing and the lowest for *drawing text-related conclusions*. Intercorrelations among the three dimensions were higher than .85, but they differed from a perfect correlation (i.e., they were considerably lower than .95, see Carstensen, 2013),

indicating some degree of multidimensionality of the test. In accordance with this finding, model fit indices preferred the three dimensional model (AIC = 202504.65, BIC = 202792.98, number of parameters = 42) over the unidimensional model (AIC = 202580.15, BIC = 202834.53, number of parameters = 37). This may, however, also be a result of the large sample size. Altogether, the results suggest that the different cognitive requirements do not completely form a unidimensional construct. The four-dimensional model based on the four text functions was estimated using the Monte Carlo estimation algorithm implemented in ConQuest. Estimated variances and correlations are given in Table 10. The estimated variances differed between the four dimensions. Especially the literary and the advertising text had lower variance estimates, indicating less discrimination of persons on these subdimensions. Correlations between the dimensions varied between $r = .767$ and $r = .893$. The lowest correlation was found between Dimension 3 (*commenting text* and dimension 4 (*advertising text*). Dimension 1 (*literary text*) and Dimension 3 showed the strongest correlation. All correlations deviated from a perfect correlation (i.e., they were considerably lower than r = .95, see Carstensen, 2013). Moreover, the four-dimensional model (AIC = 202392.52, BIC = 202708.31, number of parameters = 46) fitted the data better than the unidimensional model (AIC = 202580.15, BIC = 202834.53, number of parameters = 37).

When drawing conclusions, two aspects have to be taken into account: first, missing responses increasingly occurred on items toward the end of test. As a consequence, this may result in less variation and therefore, correlations may be lower. Second, the text functions were fully determined by the texts, that is, they were perfectly confounded as one text constituted one text function. Items were organized into item sets, with each referring to one text; hence, local item dependence (LID) may be prevalent. The correlations among the texts in the four-dimensional model as shown in Table 10, thus, occur not only due to multidimensionality but also due to local item dependence.

The testing design in the main studies does not allow for disentangling these two sources. In pilot studies (Gehrer et al., 2013) a larger number of texts was presented to test takers, thus allowing us to investigate the impact of text functions independently of LID. The correlations estimated in the pilot study varied between $r = .78$ and $r = .91$. Although Gehrer et al. used a different scaling model, the results give a first idea of the impact of the text function (unconfounded with LID) on the dimensionality of the test. As the correlations found in Gehrer et al. (2013) differed from a perfect correlation, it was concluded that text functions formed subdimensions of reading competence. Comparing the correlations found in Gehrer et al. (2013), which were due to text functions, to those resulting from the main study (Table 10), which were due to both, text functions and LID, allowed us to evaluate the impact of LID. The correlations in the present study were only slightly lower (varying between $r = .767$ and $r = .893$) than those found in Gehrer et al. (ranging from $r = 0.78$ to $r = 0.91$), indicating a small amount of local item dependence. In conclusion, the text functions seem to build empirically distinguishable subdimensions of the reading competence test.

Due to theoretical considerations, Gehrer et al. (2013) argued for a unidimensional construct. Consequently, a single competence score is estimated for reading competence.

Table 10.

Results of Four-Dimensional Scaling.

	Dim 1	Dim 2	Dim 3	Dim 4
Literary function (Dim 1) (Nitems = 8)	0.425			
Instruction texts (Dim 2) (Nitems = 7)	0.861	0.616		
Commenting function (Dim 3) (Nitems = 8)	0.893	0.821	0.807	
AdvertisingTexts (Dim 4) (Nitems = 5)	0.822	0.848	0.767	0.484

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

6. Discussion

Descriptions and analyses presented in the previous sections have aimed at documenting the quality of the first-year students' reading competence test and at providing information on the estimation procedure of the reading competence score published in the Scientific Use File for first-year university students in Starting Cohort 5.

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

There is a rather small amount of missing responses due to invalid, not-determinable, and omitted items. However, in particular, the items at the end of the test showed a large amount of not reached items. Therefore, items of the last text were excluded from the analyses. Given the testing time, the test is too long. In further studies the test length should be reduced.

Overall, the items of the test show a good fit to the model and measurement invariance can be confirmed across various subgroups. The reliability of the test is acceptable. Because the test is mainly targeted at low-performing participants, ability estimates for those participants will be very precise but less precise for medium and especially for high-performing students.

Results of the dimensionality analyses challenge the conclusion of a unidimensional test. There are indications for multidimensionality referring to cognitive requirements, as well as to text functions. In combination with the exclusion of items belonging to the last text (thus, having the test rely on only four text functions), the estimation of a single reading competence score that is comparable to the reading competence score of other cohorts is challenged. The validity and comparability of this score to other cohorts might need to be investigated in further studies. Nonetheless, Gehrer et al. (2013) provide substantial theoretical arguments for a unidimensional measure of reading competence. Moreover,

they argue that a balanced assessment of reading competence can only be achieved by heterogeneity of text functions.

In summary, test quality is challenged by testing time and by indications of multidimensionality. Still, the items of the reading test exhibit good psychometric properties in terms of item fit and reliability that facilitate the estimation of a reliable reading competence score.

7. Data in the Scientific Use File

The data in the Scientific Use File contain 28 items, of which 22 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. Six items were scored as polytomous variables (CMC or MA items). MC items are marked with a '0_c' at the end of the variable name, whereas the variable names of CMC and MA items end with 's_c'. Note that the values of the polytomous variables in the Scientific Use File do not necessarily correspond to the number of correctly solved subtasks. This is due to the collapsing of categories (cf. Section 4.2 for a description of the aggregation of CMC and MA items). In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category. Manifest reading competence scores are provided in the form of WLEs (res1_sc1), together with their corresponding standard error (res1_sc2). In the estimation of WLEs the effect of the test position (first vs. second position) is controlled for. The ConQuest Syntax used to estimate WLEs is provided in Appendix A. For persons who either did not take part in the reading test, for whom no information on the sequence of tests was available, or who did not have enough valid responses, no WLE is estimated. The value of the WLE and the respective standard error for these persons are denoted as not-determinable missing value.

Plausible values, that allow us to investigate latent relationships of competence scores with other variables, will be provided in later data releases. Users interested in examining latent relationships may alternatively either include the measurement model in their own analyses or estimate plausible values themselves. A description of these approaches can be found in Pohl and Carstensen (2012).

References

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

Carstensen, C. H. (2013). Linking PISA competencies over three cycles – results from Germany. In M. Prenzel, M. Kobarg, K. Schöps & S. Rönnebeck (Eds.). *Research Outcomes of the PISA Research Conference 2009*; p.199-213. New York: Springer.

Gehrer, K., & Artelt, C. (2013). Literalität und Bildungslaufbahn: Das Bildungspanel NEPS. In A. Bertschi-Kaufmann, & C. Rosebrock (Eds.), *Literalität erfassen: bildungspolitisch, kulturell, individuell.* S. 168-187. Weinheim, Germany: Juventa.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). The assessment of reading competence (including sample items for grade 5 and 9). Scientific Use File 2012, Version 1.0.0. Bamberg: University of Bamberg, National Educational Panel Study.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. Manuscript submitted for publication.

Haberkorn, K., Pohl, S., Carstensen, C., & Wiegand, E. (2013). Incorporating different response formats in the IRT-scaling model for competence data. Manuscript submitted for publication.

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

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

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

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

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

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

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

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

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in item response theory. *Psychometrika, 54*, 427-450.

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

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

Appendix

Appendix A: ConQuest-Syntax for estimating WLEs in Starting Cohort 5

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

datafile filename.dat;

format pid 4-10 responses 13-42 position 48;

labels << filename_with_labels.txt;

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

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

set constraint=cases;

model item + item*step + position;
estimate;

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