NEPS WORKING PAPERS

Ingrid Koller, Kerstin Haberkorn, and Theresa Rohm

# NEPS TECHNICAL REPORT FOR READING: SCALING RESULTS OF STARTING COHORT 6 FOR ADULTS IN MAIN STUDY 2012

NEPS Working Paper No. 48
Bamberg, August 2014

LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

# NEPS Technical Report for Reading: Scaling Results of Starting Cohort 6 for Adults in the Main Study 2012


*Ingrid Koller, Kerstin Haberkorn, Theresa Rohm*

*Leibniz Institute for Educational Trajectories, Bamberg, Germany*

**Email address of lead author:**

ingrid.koller@lifbi.de



**Bibliographic data:**

Koller, I., Haberkorn, K., & Rohm, T. (2014). *NEPS Technical Report for Reading: Scaling results of Starting Cohort 6 for adults in main study 2012* (NEPS Working Paper No. 48). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

We thank Michael Wenzler, Rebekka Pawlowsky, and Nadine Rita Melzner for their assistance in scaling the data and Natalie Boonyaprasop for proofreading the paper.

# NEPS Technical Report for Reading: Scaling Results of Starting Cohort 6 for Adults in the Main Study 2012


## Abstract

The National Educational Panel Study (NEPS) aims to investigate the development of competencies across the whole life span. It also develops tests for assessing the 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 results of reanalyzing the adult reading competence test. The adult reading test was first administered in the main study 2010/11. In 2012, the same test was administered to a refreshment sample, that is, it was presented to subjects who had not taken the test in the first study. As in the main study of 2010/2011, the reading competence test for the adult cohort consisted of 32 items, which represented different cognitive requirements and text functions and used different response formats. The test was administered to 3,156 persons. Because this paper describes the reanalysis of an existing reading competence test in NEPS, the detailed description of the test and the scaling procedure are given in the NEPS Working Paper No. 25 (see Hardt, Pohl, Haberkorn, & Wiegand, 2013). Thus, the description in the present paper is kept as short as possible. After reporting descriptive statistics of the data, the partial credit model was applied to investigate the quality of the scale. The results showed that the test exhibits high reliability and that the items fit the model. Moreover, measurement invariance could be confirmed for various subgroups. Dimensionality analyses showed that the different cognitive requirements foster a unidimensional construct, and there is some evidence for multidimensionality based on text functions. It should benoted that a considerable amount of items were not reached by the test takers within the given assessment time and that many items were targeted toward a lower reading ability. Altogether, as in the main study 2010/2011, the results show good psychometric properties of the reading competence test and support the estimation of a reliable reading competence score. Furthermore, measurement invariance between the two main studies could be confirmed. Therefore, the competence scores for the main study 2012 were estimated with fixed item parameters from the main study 2010/11 in order to place the subjects of the two studies on the same scale. At the end of the paper, the data available in the Scientific Use File are described and the ConQuest-Syntax for scaling the data is provided.



## Keywords

item response theory, scaling, reading competence, Scientific Use File, adults

## Content

## 1. Introduction

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

Most of the competence data are scaled using models that are based on Item Response Theory (IRT). 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).

This paper presents the results of reanalyzing the reading competence test in the third wave of Starting Cohort 6 (adults; 2012). What must be mentioned is that only short descriptions of the reading competence test and the analysis of items are given here. Please refer to Working Paper No. 25 for more detailed information on specific aspects of the reading competence test and the procedure of analysis (see Hardt, Pohl, Haberkorn, & Wiegand, 2013, for the main study 2010/2011). Moreover, the framework and test development for the reading competence test are described in Weinert et al. (2011) and in Gehrer, Zimmermann, Artelt, and Weinert (2012). Furthermore, as already mentioned, the description of the procedure for scaling the competence data and for checking the quality of the scales can be found in Pohl and Carstensen (2012). After presenting our reanalysis, we describe the data available for public use in the Scientific Use File. Please note that for reasons of maintaining comparability between abilities in Wave 1 and Wave 3, the person parameters in the Scientific Use File are estimated with constrained item parameters. Item parameters from the first wave are used (see Hardt et al., 2013) as fixed parameters during calibration to place the subjects in Wave 1 and Wave 3 on the same scale. The description of this procedure is given in Section 7.

Please also 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 any major changes in the results.

## 2. Testing Reading Competence

In the main study 2012, ICT literacy, scientific literacy, and reading competence were assessed. The test on reading competence was administered to all subjects as the last part of the assessment. Only the reading competence test is described in the following section.

The adults' reading test consists of five texts and 32 items referring to one of these five texts. Each of these texts represents one text type or text function and three cognitive requirements. The cognitive requirements do not depend on the text type, but each cognitive requirement is usually assessed within each text type (see Gehrer et al., 2012, and Weinert et al., 2011, for a detailed description of the framework). In the reading competence test there are three types of response formats: simple multiple-choice (MC) items, complex multiple-choice (CMC) items, and matching (MA) items. Examples of the different response formats are provided in Pohl and Carstensen (2012).

Because the main aim of the study was to reanalyze the reading competence test, the items rea20260 and rea20270, which had shown unsatisfactory item fit in the first study (see Hardt et al., 2013), were excluded from previous analyses.[1] Thus, the scaling results presented in the following sections are based on 30 items. The characteristics of these items are described in Table 1 to Table 3. Table 1 shows the distribution of the cognitive requirements, Table 2 the distribution of text functions, and Table 3 the response formats used. The number of subtasks within CMC and MA items varied between two and six.

Table 1

*Cognitive Requirements of Items in the Reading Test for Adults*

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

Table 2

*Number of Items for Different Text Types in the Reading Test for Adults*

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

Table 3

*Response Formats of Items in the Reading Test for Adults*

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

[1] Note that preliminary analyses were performed with all items. As in the previous study, rea20260 and rea20270 showed an unsatisfactory item fit.

## 3. Data and Sample Size

A description of the design of the study, the sample, as well as the instruments used can be found on the NEPS website[2]. In total, 3,156 subjects took the reading competence test[3]. Six of the 3,145 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. The final sample for the analyses, thus, consisted of 3,150 persons.

## 4. Analyses

### 4.1 Missing Responses

There are different types of missing responses in competence test data (see Hardt et al., 2013). 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.

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

### 4.2 Scaling Model

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

CMC and MA items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC or MA item, indicating the number of correctly responded subtasks within that item. As in Hardt et al. (2013), categories were collapsed in order to avoid possible estimation problems if the categories of the polytomous variables had less than $N = 200$ (see also Pohl & Carstensen, 2012, for an explanation of this approach). In order to ensure comparability between the main studies of 2010/11 and 2012, the same categories of the CMC and MA items were collapsed. Furthermore, it has to be mentioned that the same categories of items had cell frequencies that were too low.

Please 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, whereas simple MC items were scored dichotomously as 0 for an incorrect and as 1 for the correct response (see Haberkorn, Pohl, Carstensen, &

[2] www.neps-data.de

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

Wiegand, 2012; and Pohl & Carstensen, 2013, for studies on the scoring of different response formats).

Ability estimates for reading competence were estimated as weighted maximum likelihood estimates (WLEs; Warm, 1989). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012), and 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. The description of the procedure as well as the procedure itself is the same as in Hardt et al. (2013). For a better understanding of the results, the whole section has been adopted from Working Paper No. 25 (Hardt et al., 2013).

As in the previous study (see Hardt et al., 2013), subtasks of polytomous variables had been aggregated to polytomous variables and the item fit of dichotomous MC and polytomous CMC and MA items was examined by analyzing them via a partial credit model. The weighted mean square error (WMNSQ), the respective *t*-value, correlations of the item score with the total score, and the item characteristic curve were evaluated for each item. Items with a WMNSQ > 1.15 (*t*-value > |6|) were considered as having a noticeable item misfit and items with a WMNSQ > 1.2 (*t*-value > |8|) were judged as having a considerable item misfit, and their performance was further investigated. Correlations of the item score with the total score (equal to the discrimination as computed in ConQuest) greater than 0.3 were considered as good, greater than 0.2 as acceptable, and below 0.2 as problematic. Overall, judgment of the fit of an item was based on all fit indicators.

Our aim was to construct a reading competence test that measured 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 gender, school degree, and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Differential item functioning was estimated using a multigroup IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled. Differences in the estimated item difficulties between the subgroups were evaluated. Based on experiences with preliminary data, we judged absolute differences in estimated difficulties that were 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 significant, and differences smaller than 0.4 as no considerable DIF. In addition to DIF analyses at item level, test fairness was investigated by comparing a model including differential item functioning to a model that only estimated main effects and no DIF.

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

Additionally, we evaluated the dimensionality of the reading test by conducting several multidimensional analyses. The different subdimensions of the multidimensional models were specified based on different construction criteria. First, a model with three different subdimensions representing the three cognitive requirements, and, second, a model with five different subdimensions based on the five text functions were fitted to the data. The correlations 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 five texts, the assumption of local item dependence (LID) may not necessarily hold. However, the five texts were perfectly confounded with the five text functions. Thus, multidimensionality and local item dependence may not be evaluated separately with these data. We referred to preliminary studies on reading competence to disentangle the amount of multidimensionality and local item dependence.

# 5. Results

## 5.1 Missing Responses

### Missing responses per person

Figure 1 depicts the number of invalid responses per person. As can be seen, with 84.13%, the vast majority of the respondents did not have any invalid response at all and less than 5% had more than one invalid response.

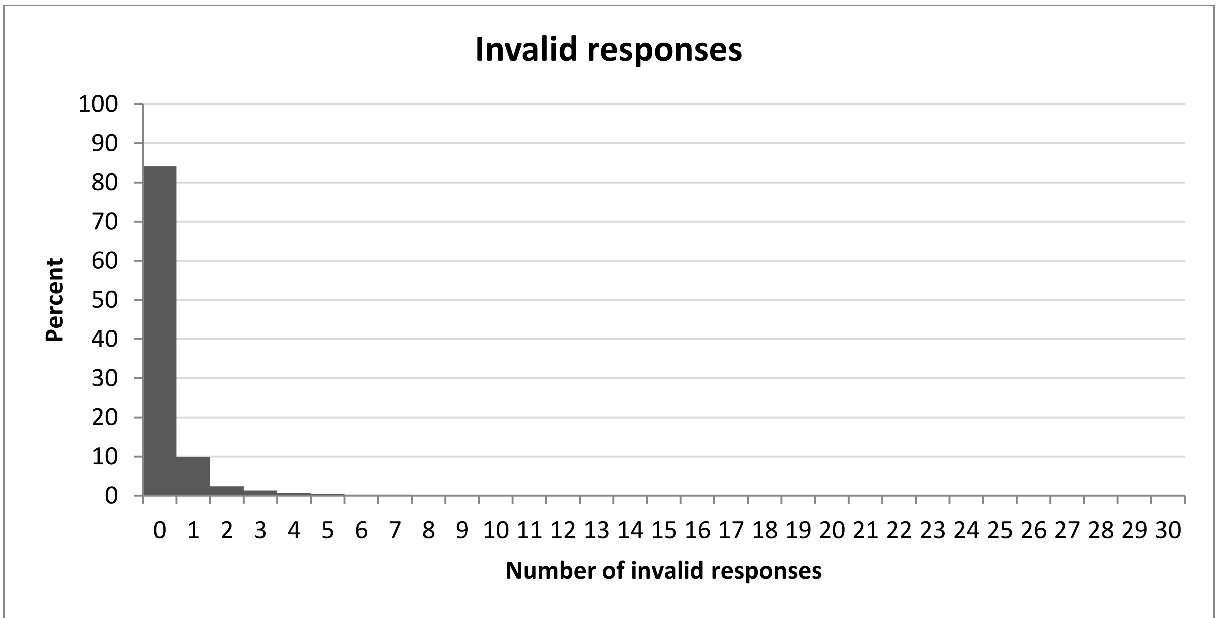


*Figure 1.* Number of invalid responses.

Missing responses may also occur when respondents omit items. As can be seen in Figure 2, the majority of subjects—almost 48%—did not skip any item at all and only about 5% omitted more than four items of the reading test.

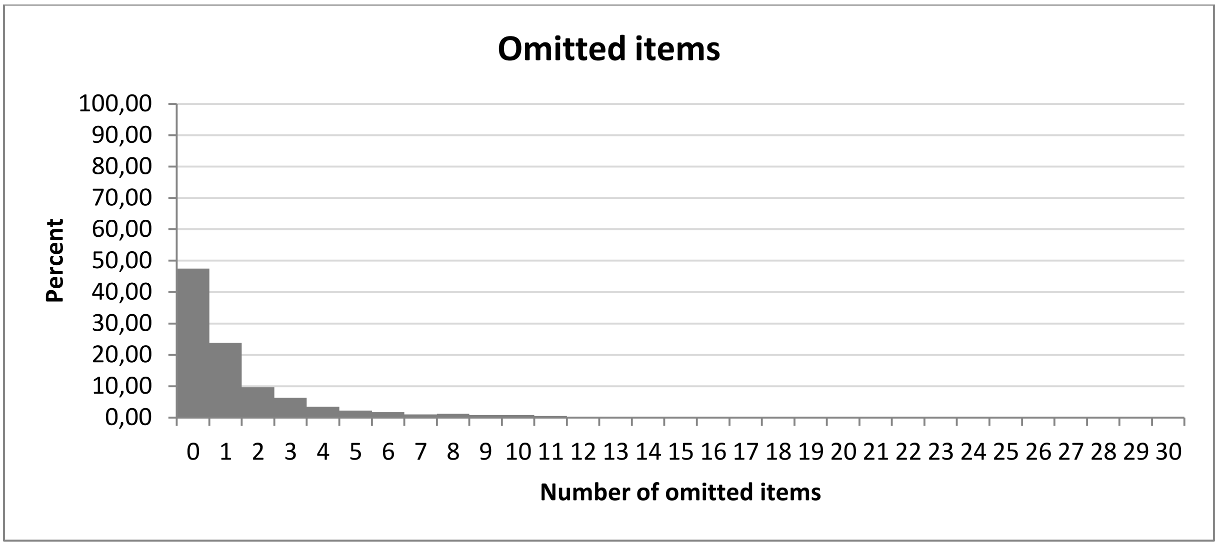


*Figure 2.* Number of omitted items.

Another source of missing responses are items that were not reached by the subjects. Per definition, these are all missing responses after the last valid response. The number of not-reached items was rather high (see Figure 3). With 38.29%, less than half of the participants were able to finish the reading competence test within the given time. Almost 42% of the subjects did not reach the last text and around 15% did not reach the items of the last two of the five texts.

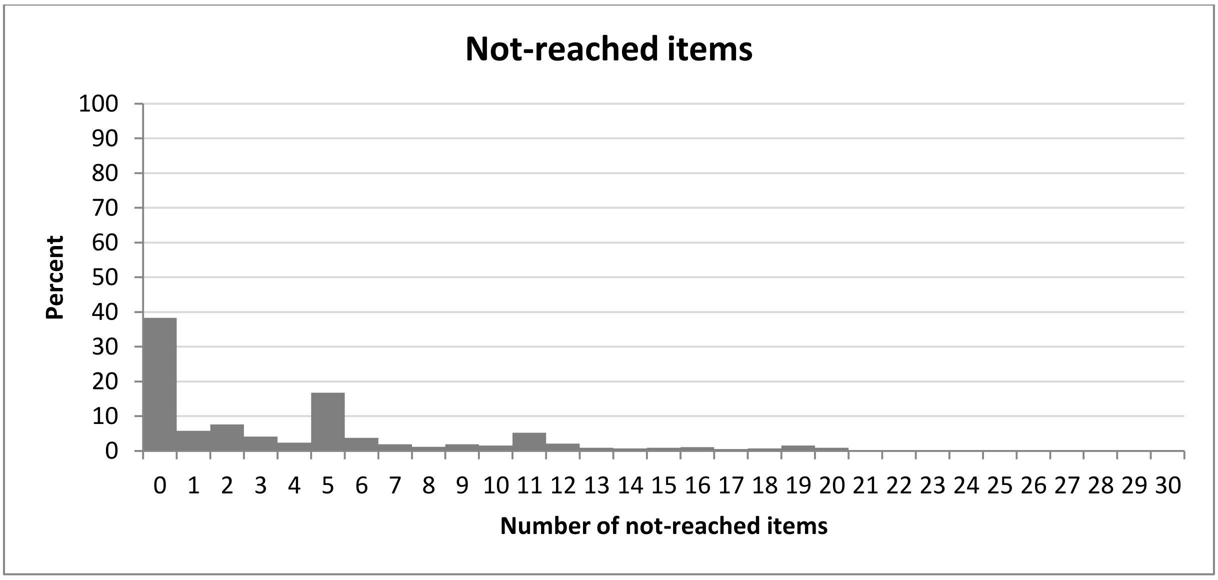


*Figure 3.* Number of not-reached items.

The aggregated polytomous variables were coded as not-determinable missing responses when the subtasks of CMC and MA items contained different kinds of missing responses. Because not-determinable missing responses might only occur in CMC and MA items, the maximum number of not-determinable missing responses was seven (i. e., the number of CMC and MA items). Only a small amount of not-determinable missing responses occurred (see Figure 4). Overall, 96.7% of the subjects had no non-determinable missing responses and only 0.28% of the persons gave a not-determinable missing response to more than one of the items.

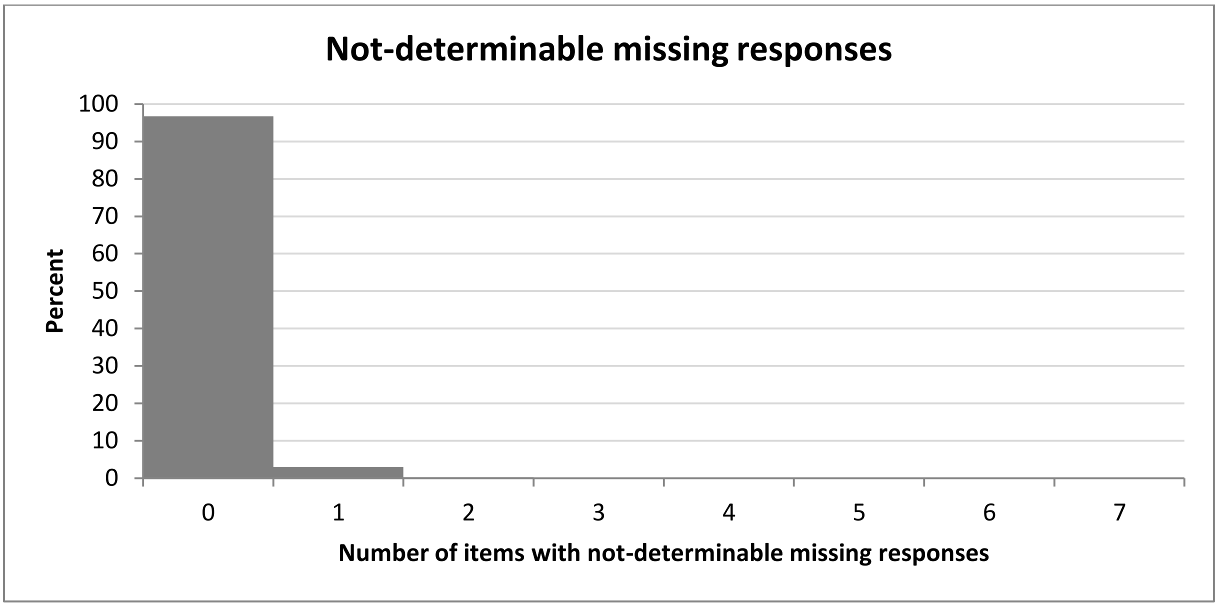


*Figure 4.* Number of not-determinable missing responses.

The total number of missing responses aggregated over invalid, omitted, not-reached, and not-determinable missing responses per person is illustrated in Figure 5. On average, the subjects showed 6.01 (*SD* = 5.81) missing responses. In total, 19.14% of the persons had no missing response at all and about 50% of the participants gave five or more missing responses.

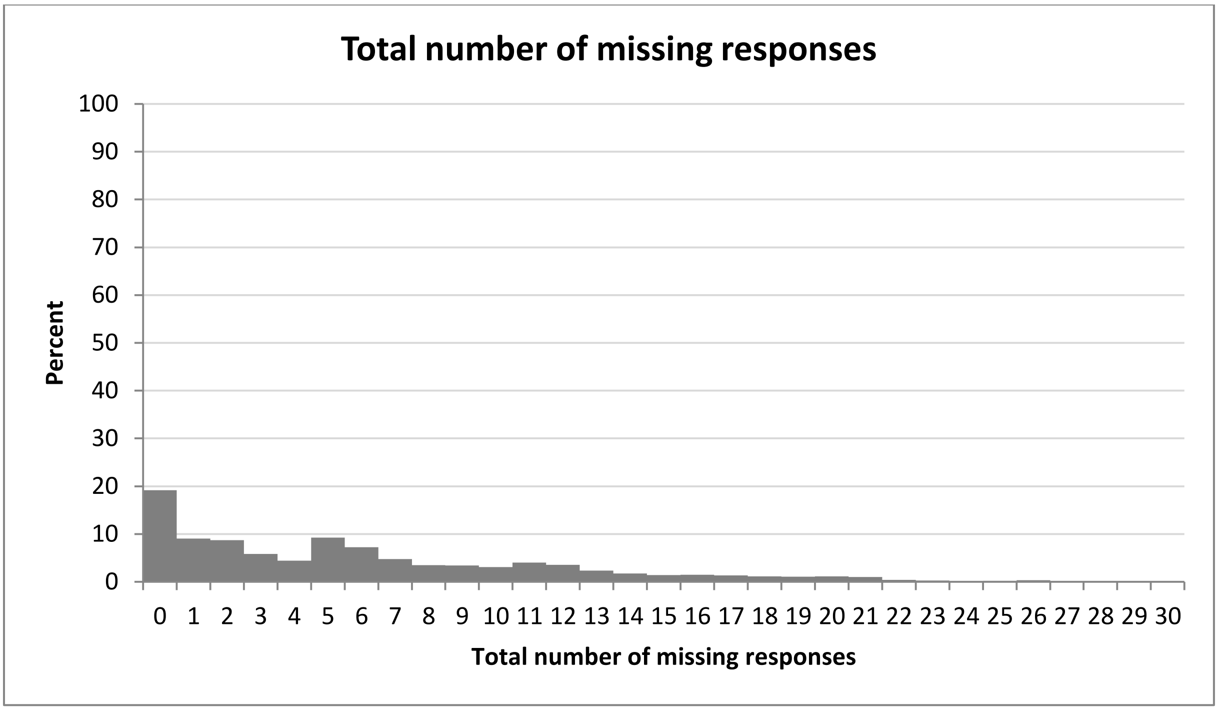


*Figure 5*. Total number of missing responses.

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

**Missing responses per item**

Table 4 provides information on the occurrence of different kinds of missing responses per item. Overall, the omission rate is acceptable, varying across items between 0% (rea20550_c) and 18.60% (rea2028s_c). There were 10 items with an omission rate exceeding 5%. On average, CMC and MA items had higher omission rates (11.94% and 11.63%, respectively) than MC items (2.91%). With an increase in the number of items being positioned toward the end of the test, the amount of persons failing to reach those items (Column 4) rose up to a considerable amount of 61.71% (for the last item rea20550_c). On the contrary, the percentage of invalid responses per item (Column 5) was rather low (maximum of 4.16% for item rea20140_c). Matching items seemed to be more prone to cause invalid responses than were multiple-choice items in both single and complex form.

## 5.2 Parameter Estimates

### 5.2.1 Item parameters

The second column in Table 5 shows the percentage of correct responses relative to all valid responses for each item. Please note that, because there is a nonnegligible amount of missing responses, this probability cannot be interpreted as an index of item difficulty. The percentage of correct responses within MC items varied between 17.05% and 91.05% with an average of 64.29% ($SD$ = 19.74%) correct responses.

For reasons of model identification, in the partial credit model, the mean of the ability distribution was constrained to be zero. The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 5. The step parameters for polytomous variables are depicted in Table 6. The item difficulties ranged from -3.299 (item rea20110_c) to 0.724 (item rea20460_c) logits with an average difficulty of -1.456 logits ($SD$ = 1.043). 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 ($SE(\beta) \leq 0.120$).

Table 4

*Missing Values*

| Item | Position in the test | Number of valid responses | Relative frequency of not-reached items in % | Relative frequency of omitted items in % | Relative frequency of invalid responses in % |
|---|---|---|---|---|---|
| **rea20110_c** | 1 | 3,053 | 0.00 | 2.29 | 0.79 |
| **rea2012s_c** | 2 | 2,573 | 0.00 | 18.16 | 0.16 |
| **rea20130_c** | 3 | 2,968 | 0.00 | 3.71 | 2.06 |
| **rea20140_c** | 4 | 2,915 | 0.00 | 3.30 | 4.16 |
| **rea2015s_c** | 5 | 2,804 | 0.16 | 10.48 | 0.35 |
| **rea20210_c** | 6 | 3,048 | 0.22 | 1.87 | 1.14 |
| **rea20220_c** | 7 | 2,970 | 0.29 | 2.32 | 3.11 |
| **rea20230_c** | 8 | 3,031 | 0.29 | 2.79 | 0.70 |
| **rea20240_c** | 9 | 3,036 | 0.29 | 2.10 | 1.24 |
| **rea20250_c** | 10 | 3,002 | 0.35 | 2.95 | 1.40 |
| **rea2028s_c** | 13 | 2,475 | 1.21 | 18.60 | 1.62 |
| **rea20310_c** | 14 | 2,867 | 2.73 | 3.30 | 2.95 |
| **rea20320_c** | 15 | 2,850 | 3.40 | 3.90 | 2.22 |
| **rea20330_c** | 16 | 2,822 | 3.94 | 5.81 | 0.67 |
| **rea20340_c** | 17 | 2,719 | 5.02 | 7.90 | 0.76 |
| **rea20350_c** | 18 | 2,718 | 5.94 | 7.08 | 0.70 |
| **rea20360_c** | 19 | 2,739 | 6.67 | 4.92 | 1.46 |
| **rea20370_c** | 20 | 2,732 | 7.56 | 4.67 | 1.05 |
| **rea2038s_c** | 21 | 2,353 | 9.65 | 15.65 | 0.00 |
| **rea20410_c** | 22 | 2,606 | 14.86 | 1.84 | 0.57 |
| **rea2042s_c** | 23 | 2,434 | 16.38 | 6.35 | 0.00 |
| **rea20430_c** | 24 | 2,519 | 18.32 | 0.76 | 0.95 |
| **rea20440_c** | 25 | 2,504 | 19.49 | 0.79 | 0.22 |
| **rea20450_c** | 26 | 2,431 | 21.43 | 0.79 | 0.60 |

| 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 % |
|---|---|---|---|---|---|
| **rea20460_c** | 27 | 2,269 | 25.14 | 2.57 | 0.25 |
| **rea20510_c** | 28 | 1,805 | 41.84 | 0.67 | 0.19 |
| **rea2052s_c** | 29 | 1,518 | 44.19 | 7.62 | 0.00 |
| **rea20530_c** | 30 | 1,609 | 48.29 | 0.48 | 0.16 |
| **rea2054s_c** | 31 | 1,198 | 55.94 | 5.81 | 0.22 |
| **rea20550_c** | 32 | 1,102 | 61.71 | 0.00 | 3.30 |

*Note*. The items in positions 11 and 12 were excluded from the analyses due to unsatisfactory item fit (see Section 2).

Table 5

*Item Parameters*

| Item | Percentage correct | Difficulty/ location parameter | *SE* (difficulty/location parameter) | WMNSQ | *t*-value of WMNSQ | Correlation of item score with total score | Discrimi-nation – 2PL |
|---|---|---|---|---|---|---|---|
| **rea20110_c** | 91.05 | -3.299 | 0.080 | 1.01 | 0.2 | 0.30 | 0.97 |
| **rea2012s_c** | n. a. | -2.569 | 0.069 | 0.94 | -1.8 | 0.46 | 1.06 |
| **rea20130_c** | 85.37 | -2.762 | 0.067 | 0.99 | -0.2 | 0.38 | 0.86 |
| **rea20140_c** | 69.65 | -1.379 | 0.048 | 1.03 | 1.4 | 0.45 | 1.10 |
| **rea2015s_c** | n. a. | -1.488 | 0.056 | 0.90 | -4.6 | 0.54 | 1.21 |
| **rea20210_c** | 90.48 | -3.196 | 0.077 | 0.98 | -0.3 | 0.33 | 1.30 |
| **rea20220_c** | 80.67 | -2.164 | 0.057 | 0.96 | -1.0 | 0.47 | 1.22 |
| **rea20230_c** | 85.68 | -2.559 | 0.063 | 0.94 | -1.5 | 0.46 | 1.09 |
| **rea20240_c** | 82.16 | -2.153 | 0.056 | 0.95 | -1.6 | 0.48 | 0.60 |
| **rea20250_c** | 79.46 | -1.976 | 0.054 | 0.97 | -0.8 | 0.47 | 1.26 |
| **rea2028s_c** | n. a. | -0.579 | 0.030 | 0.92 | -3.0 | 0.74 | 0.74 |
| **rea20310_c** | 62.06 | -0.918 | 0.045 | 1.13 | 5.8 | 0.38 | 0.87 |
| **rea20320_c** | 72.06 | -1.659 | 0.051 | 0.94 | -2.0 | 0.53 | 1.09 |

| 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 |
|---|---|---|---|---|---|---|---|
| **rea20330_c** | 69.05 | -1.484 | 0.050 | 1.07 | 2.6 | 0.41 | 1.12 |
| **rea20340_c** | 45.87 | -0.094 | 0.044 | 1.02 | 1.2 | 0.47 | 0.72 |
| **rea20350_c** | 71.59 | -1.903 | 0.056 | 0.97 | -0.8 | 0.47 | 0.51 |
| **rea20360_c** | 68.25 | -1.558 | 0.052 | 0.98 | -0.8 | 0.50 | 0.76 |
| **rea20370_c** | 55.75 | -0.678 | 0.045 | 1.08 | 3.9 | 0.44 | 1.66 |
| **rea2038s_c** | n. a. | -0.625 | 0.061 | 0.99 | -0.4 | 0.40 | 1.11 |
| **rea20410_c** | 42.48 | 0.019 | 0.045 | 1.16 | 8.6 | 0.36 | 0.60 |
| **rea2042s_c** | n. a. | -0.766 | 0.054 | 0.93 | -3.5 | 0.51 | 0.81 |
| **rea20430_c** | 60.03 | -1.307 | 0.052 | 1.08 | 2.8 | 0.42 | 0.83 |
| **rea20440_c** | 67.90 | -2.123 | 0.062 | 0.87 | -3.4 | 0.56 | 0.57 |
| **rea20450_c** | 63.05 | -1.779 | 0.058 | 0.97 | -0.9 | 0.49 | 1.32 |
| **rea20460_c** | 27.94 | 0.724 | 0.049 | 1.10 | 4.9 | 0.37 | 1.63 |
| **rea20510_c** | 51.87 | -2.658 | 0.087 | 1.05 | 0.8 | 0.34 | 1.23 |
| **rea2052s_c** | n. a. | -1.832 | 0.120 | 0.94 | -2.3 | 0.39 | 1.02 |
| **rea20530_c** | 39.11 | -1.354 | 0.066 | 1.03 | 0.8 | 0.47 | 1.28 |
| **rea2054s_c** | n. a. | 0.150 | 0.084 | 0.98 | -0.7 | 0.43 | 1.43 |
| **rea20550_c** | 17.05 | 0.293 | 0.069 | 1.13 | 4.4 | 0.38 | 1.01 |

*Note.* The percentage of correct scores gives no information on polytomous CMC and MA item scores. These are denoted by n. a.
As 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 the correlation corresponds to the product moment correlation between the corresponding categories and the total score (discrimination value as computed by ConQuest).

Table 6

*Step Parameters (and Standard Errors) of the Polytomous Items*

| Item | Step 1 (*SE*) | Step 2 (*SE*) | Step 3 (*SE*) | Step 4 (*SE*) | Step 5 (*SE*) |
|---|---|---|---|---|---|
| **rea2012s_c** | 0.164 (0.053) | -0.164 | | | |
| **rea2015s_c** | -0.194 (0.042) | 0.194 | | | |
| **rea2028s_c** | 0.372 (0.042) | -0.117 (0.042) | -0.202 (0.045) | 0.084 (0.053) | -0.137 |
| **rea2038s_c** | -0.655 (0.042) | 0.655 | | | |
| **rea2042s_c** | 0.165 (0.047) | -0.165 | | | |
| **rea2052s_c** | n. a. | | | | |
| **rea2054s_c** | -0.551 (0.059) | 0.551 | | | |

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

### 5.2.2 Test targeting and reliability

Test targeting focuses on matching item difficulties with person abilities (WLEs) 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 1.469, which implies good differentiation between the subjects. The reliability of the test (EAP/PV reliability = .797, WLE reliability = .743) was good. Although the items covered a wide range of the ability distribution, the items were slightly too easy. As a consequence, person ability in medium- and low-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors of measurement.

## 5.3 Quality of the Test

### 5.3.1 Item fit

Item fit was investigated for MC and polytomous CMC and MA items. Altogether, item fit can be considered to be very good (see Table 5). Values of the WMNSQ ranged from 0.87 (item rea20440_c) to 1.16 (rea20410_c), and only one *t*-value of the WMNSQ exceeded a *t*-value of 5. There is no indication of any severe item over- and even less item underfit. Point-biserial correlations between the item scores and the total scores ranged from .30 (item rea20110_c) to .74 (item rea2028s_c) and had a mean of .447. All item characteristic curves showed a good fit of the items.

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

*Figure 6.* Test targeting. The distribution of person ability in the sample is depicted on the left-hand side of the graph, with each ‘X’ representing 18.7 cases. The difficulty of the items is depicted on the right-hand side of the graph, with each number representing one item (corresponding to the item position indicated in Table 4).

### 5.3.2 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i. e., measurement invariance). For this purpose, DIF was examined for the variables gender, school degree, and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Table 7 provides a summary of the results of the DIF analyses.

Table 7

*Differential Item Functioning (Absolute Differences Between Difficulties)*

| **Item** | **School degree** | | | **Gender** | **Migration background** |
|---|---|---|---|---|---|
| | Lower degree vs. Abitur | Lower degree vs. missing | Abitur vs. missing | Male vs. female | Without vs. with |
| **rea20110_c** | -0.122 | 0.362 | 0.484 | 0.188 | 0.126 |
| **rea2012s_c** | 0.181 | 0.299 | 0.118 | 0.280 | -0.138 |
| **rea20130_c** | -0.102 | 0.219 | 0.321 | 0.202 | -0.120 |
| **rea20140_c** | 0.047 | 0.322 | 0.275 | 0.184 | 0.368 |
| **rea2015s_c** | 0.197 | 0.160 | -0.037 | -0.070 | -0.326 |
| **rea20210_c** | -0.089 | 0.181 | 0.270 | -0.174 | -0.446 |
| **rea20220_c** | -0.217 | 0.162 | 0.379 | 0.106 | -0.006 |
| **rea20230_c** | -0.032 | 0.267 | 0.299 | 0.280 | 0.172 |
| **rea20240_c** | -0.283 | 0.109 | 0.392 | -0.180 | -0.426 |
| **rea20250_c** | -0.586 | -0.075 | 0.511 | 0.222 | -0.036 |
| **rea2028s_c** | -0.016 | 0.180 | 0.196 | 0.108 | -0.256 |
| **rea20310_c** | -0.317 | 0.048 | 0.365 | 0.002 | -0.230 |
| **rea20320_c** | 0.146 | 0.481 | 0.335 | 0.144 | -0.042 |
| **rea20330_c** | -0.286 | 0.291 | 0.577 | 0.162 | 0.214 |
| **rea20340_c** | 0.163 | 0.305 | 0.142 | -0.224 | 0.144 |
| **rea20350_c** | 0.055 | 0.403 | 0.348 | 0.104 | 0.010 |
| **rea20360_c** | 0.050 | 0.298 | 0.248 | 0.382 | 0.446 |
| **rea20370_c** | -0.482 | -0.031 | 0.451 | -0.040 | -0.116 |
| **rea2038s_c** | 0.057 | 0.310 | 0.253 | 0.026 | 0.290 |
| **rea20410_c** | -0.296 | 0.062 | 0.358 | -0.220 | 0.078 |
| **rea2042s_c** | 0.392 | 0.448 | 0.056 | -0.382 | 0.298 |
| **rea20430_c** | -0.268 | 0.326 | 0.594 | -0.424 | 0.328 |
| **rea20440_c** | 0.169 | 0.224 | 0.055 | -0.220 | -0.006 |
| **rea20450_c** | -0.340 | -0.104 | 0.236 | -0.100 | -0.078 |
| **rea20460_c** | 0.066 | 0.138 | 0.072 | -0.206 | -0.016 |
| **rea20510_c** | 0.083 | 0.397 | 0.314 | 0.038 | -0.254 |
| **rea2052s_c** | -0.140 | 0.218 | 0.358 | 0.204 | 0.434 |
| **rea20530_c** | -0.175 | 0.094 | 0.269 | 0.272 | -0.232 |
| **rea2054s_c** | 0.192 | -0.006 | -0.198 | -0.096 | -0.316 |
| **rea20550_c** | -0.288 | 0.036 | 0.324 | -0.290 | 0.406 |
| **Main effect** | -0.935 | -0.244 | 0.691 | -0.184 | 0.408 |

Gender: The table depicts the differences in the estimated item difficulties between the respective groups. “Male vs. female”, for example, indicates the difference in difficulty $ß_{male}$ - $ß_{female}$. A positive value indicates a higher difficulty for males, whereas a negative value indicates a lower difficulty for males as opposed to females. Differential-item-functioning analysis for gender was based on 1,606 (50.98%) males and 1,544 (49.02%) females. On average, male participants had a lower estimated reading ability than females (main effect = -0.092 logits, Cohen’s *d* = 0.184). There was no considerable item DIF. Only one item (item rea20430_c) showed DIF greater than 0.4 logits.

School degree: Overall, 684 subjects (21.71%) who took the reading test had a university entrance qualification (Abitur) and 1,706 (54.16%) held a lower school degree. A total of 760 subjects gave a missing response to the question of school degree; these persons were treated as a group of their own in the DIF analysis. Subjects who had obtained a university entrance qualification showed a higher reading ability on average (0.542 logits, Cohen’s *d* = 0.809) than subjects with a lower school degree. Furthermore, subjects who were in the missing-response group had a lower reading ability on average (-0.149 logits, Cohen’s *d* = 0.598). There was no considerable item DIF. No item exhibited DIF greater than 0.6 logits. The results of the pairwise group-comparison showed DIF greater than 0.4 logits for several items (see Table 7).

Migration background: There were 2,693 participants (85.49%) with no migration background and 457 subjects (14.51%) with a migration background. In comparison to subjects with migration background, participants without migration background had, on average, a slightly higher reading ability (main effect = 0.204 logits, Cohen’s *d* = 0.338). There was no considerable DIF due to migration background. Differences in estimated difficulties did not exceed 0.6 logits. Two items exhibited a higher estimated difficulty for subjects with migration background than for subjects without, and three items exhibited a lower estimated difficulty for subjects with migration background.

The results of comparing models that include main effects only with models additionally allowing for DIF are displayed in

Table **8**. Regarding Akaike's (1974) information criterion (AIC), the more parsimonious model including only main effects is preferred over the variable school degree. The Bayesian information criterion (BIC; Schwarz, 1978) takes into account the number of estimated parameters and, thus, prevents the overparameterization of models. Using BIC, the more parsimonious model including only the main effect was preferred over the more complex DIF model for all DIF variables.

Table 8

*Comparison of Models With and Without DIF*

| DIF variable | Model | Deviance | Number of parameters | AIC | BIC |
|---|---|---|---|---|---|
| Gender | main effect | 81595.513 | 41 | 81677.513 | 81925.774 |
| | DIF | 81491.173 | 71 | 81633.173 | 82063.089 |
| School degree | main effect | 81369.967 | 42 | 81453.967 | 81708.283 |
| | DIF | 81269.824 | 102 | 81473.824 | 82091.450 |
| Migration | main effect | 81574.750 | 41 | 81656.750 | 81905.010 |
| | DIF | 81506.330 | 71 | 81648.330 | 82078.250 |

### 5.3.3 Rasch-homogeneity

One essential assumption of the Rasch (1960) model is Rasch-homogeneity. Rasch-homogeneity implies that all item-discrimination parameters are equal. In order to test this assumption, a generalized partial credit model (2PL) that estimates discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 5), ranging from 0.508 (item rea20410_c) to 1.663 (item rea20440_c). Model fit indices suggested a slightly better model fit of the 2PL model (AIC = 81096.03, BIC = 81586.50) as compared to the 1PL model (AIC = 81679.98, BIC = 82000.91). Despite the empirical preference for the 2PL model, the 1PL model more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the partial credit model (1PL) was chosen as our scaling model to preserve the weighting of items as intended in the theoretical framework.

### 5.3.4 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 five different text types constituted the second multidimensional model.

Estimation of the three-dimensional model was carried out by ConQuest using the Gauss-Hermite quadrature method. The estimated variances and correlations between the three dimensions representing the different cognitive requirements are reported in Table 9. All three dimensions had substantial variance estimates with the highest obtained for “finding information in the text” and the lowest for “reflecting and assessing”. Intercorrelations among the three dimensions were high (all > .95), supporting the unidimensionality of the test (see Carstensen, 2013). Nonetheless, according to model fit indices, the three-dimensional model fitted the data slightly better (AIC = 81679.23, BIC = 81951.71, number of parameters = 45) than the unidimensional model (AIC = 81689.61, BIC = 81931.82, number of parameters = 40). This may, however, also be a result of the large sample size. From these results we conclude that the three cognitive requirements do not measure different constructs but a unidimensional construct.

Table 9

*Results of Three-Dimensional Scaling*

| | Dim 1 | Dim 2 | Dim 3 |
|---|---|---|---|
| **Finding information in the text (Dim 1)** (Nitems = 13) | 1.595 | | |
| **Drawing text-related conclusions (Dim 2)** (Nitems = 8) | 0.971 | 1.579 | |
| **Reflecting and assessing (Dim 3)** (Nitems = 9) | 0.951 | 0.953 | 1.433 |

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

The five-dimensional model based on the five text functions was estimated using the Monte Carlo estimation algorithm implemented in ConQuest. Estimated variances and correlations are given in Table 10. The estimated variances differed between the five dimensions. Especially the texts located at the end of the booklet showed smaller variance estimates. This may be a consequence of the fact that the items constituting these dimensions were not reached by large percentages of the test takers. Correlations between the dimensions varied between $r = .783$ and $r = .936$. The lowest correlation was found between Dimension 2 (“instruction texts”) and Dimension 5 (“literary function”). Dimension 2 and Dimension 4 (“communication”) showed the strongest correlation. All correlations deviated from a perfect correlation (i. e., they were considerably lower than r = .95, see Carstensen, 2013). Moreover, the five-dimensional model (AIC = 81478.86, BIC = 81805.84, number of parameters = 54) fitted the data better than the unidimensional model (AIC = 81689.61, BIC = 81931.82, number of parameters = 40). These results are consistent with the results given in Hardt et al. (2013).

Table 10

*Results of Five-Dimensional Scaling*

| | Dim 1 | Dim 2 | Dim 3 | Dim 4 | Dim 5 |
|---|---|---|---|---|---|
| **Advertising texts (Dim 1)** (Nitems = 5) | 2.709 | | | | |
| **Instruction texts (Dim 2)** (Nitems = 6) | 0.884 | 2.416 | | | |
| **Commenting function (Dim 3)** (Nitems = 8) | 0.840 | 0.908 | 1.398 | | |
| **Communication (Dim 4)** (Nitems = 6) | 0.864 | 0.936 | 0.895 | 1.351 | |
| **Literary function (Dim 5)** (Nitems = 5) | 0.806 | 0.783 | 0.813 | 0.840 | 1.742 |

*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 to document the quality of the adults' reading competence test. The occurrence of different kinds of missing responses was evaluated and item as well as test quality was examined. Furthermore, measurement invariance, Rasch-homogeneity, and unidimensionality, as well as local item dependence were examined. Item fit statistics provided evidence of well-fitting items that are measurement invariant across various subgroups. The test is very reliable. However, because the test is mainly targeted at low- and medium-performing participants, ability estimates for those kind of participants will be very precise, but less precise for high-performing persons.

Results of the dimensionality analyses challenge the conclusion of a unidimensional test. Whereas cognitive requirements form a unidimensional construct, multidimensionality based on text functions seems to be present. In combination with the high amount of missing responses due to not-reached items at the end of the test (i. e., there are participants with no valid responses to some of the text functions), the estimation of a single reading competence score is challenged. This issue might need to be addressed in further studies. Nonetheless, Gehrer et al. (2012) argue that a balanced assessment of reading competence can only be achieved by heterogeneity of text functions, and they provide theoretical arguments for a unidimensional measure of reading competence.

In summary, the reanalysis showed equal results as given in Hardt et al. (2013). Thus, the reading test exhibits good psychometric properties that facilitate the estimation of a reliable reading competence score.

## 7. Data in the Scientific Use File

The data in the Scientific Use File contain 30 items, of which 23 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. A total of 7 items were scored as polytomous variables (CMC or MA items). MC items are marked with a '0_c' at the end of the variable name, whereas the variable names of CMC and MA items end in 's_c'. Note that the values of the polytomous variables in the Scientific Use File do not necessarily correspond to the number of correctly responded subtasks. This is due to 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.

#### The person parameters in the Scientific Use File

Manifest reading competence scores are provided in the form of WLEs (rea2_sc1) together with their corresponding standard error (rea2_sc2). 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 give enough valid responses, no WLE is estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values.

In order to place the competence scores in this study on the same scale as reading competence scores of the previous study, in which the same reading test was administered (Hardt et al., 2013), we estimated the WLEs of the current study with fixed item parameters chosen from the main study 2010/2011. Therefore, we first examined whether item difficulties were constant over time. The analyses of differential item functioning between the two waves showed that there was no considerable DIF for the items (*MIN* = 0.00 logits, *MAX* = 0.34 logits). Additionally, comparison of the approach with fixed item parameters and freely estimated parameters exhibited no differences, neither for the person parameter estimation nor for the investigation of model fit. Thus, because the main aim of the longitudinal study is to investigate the development of person ability, it is the WLEs estimated by the fixed item parameters that are only included in the Scientific Use File. The ConQuest-Syntax used to estimate WLEs and WLEs with fixed item parameters are provided in Appendix A and Appendix B, respectively.

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

## References

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

Carstensen, C. H. (2013). Linking PISA competencies over three cycles – Results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.). *Research Outcomes of the PISA Research Conference 2009,* 199-214. 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.

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

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

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

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.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report—Scaling results of Starting Cohort 3 in fifth grade.* (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

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),* Zeitschrift für Erziehungswissenschaften, Special Issue 14, 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 WLE estimates in Starting Cohort 6[4]

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

datafile filename.dat;
format pid 4-10 responses 13-42;
labels << filename_with_labels.txt;

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

score (0,1) (0,1) !items (1-23);
score (0,1,2) (0,0.5,1) !item (24-25,27-28,30);
score (0,1,2,3,4,5) (0,0.5,1,1.5,2,2.5) !item (26);
score (0,1) (0,0.5) !item (29);

set constraint=cases;

model item + item*step;
estimate;

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

[4] These estimated WLEs are used for investigating model fit and are not included in the Scientific Use File.

## Appendix B: ConQuest-Syntax for estimating fixed WLE estimates in Starting Cohort 6[5]

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

datafile filename.dat;
format pid 4-10 responses 13-42;
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 = none;
import anchor_parameters << B67_AD_RE_PCM.prm; /*insert estimated item parameters
from study B67*/
model item + item*step;
estimate;

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

[5] These estimated WLEs are based on fixed item parameters estimated in the first wave by Hardt et al. (2013). These WLEs estimators are reported in the Scientific Use File.