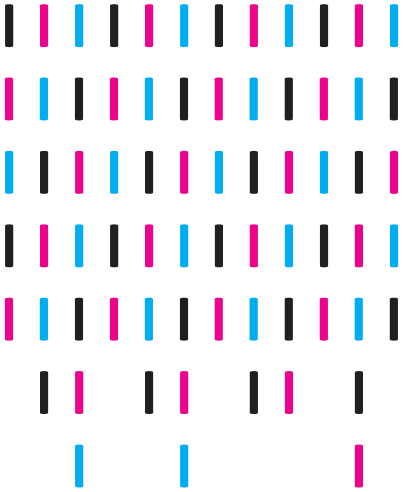

NEPS SURVEY PAPERS

Timo Gnambs, Luise Fischer, and Theresa Rohm

NEPS TECHNICAL REPORT FOR READING: SCALING RESULTS OF STARTING COHORT 4 FOR GRADE 12

NEPS Survey Paper No. 13
Bamberg, January 2017

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

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

The NEPS Survey Paper Series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

The NEPS Survey Papers are edited by a review board consisting of the scientific management of LIfBi and NEPS.

They are of particular relevance for the analysis of NEPS data as they describe data editing and data collection procedures as well as instruments or tests used in the NEPS survey. Papers that appear in this series fall into the category of 'grey literature' and may also appear elsewhere.

The NEPS Survey Papers are available at https://www.neps-data.de (see section “Publications“).

Editor-in-Chief: Corinna Kleinert, LIfBi/University of Bamberg/IAB Nuremberg

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 4 for Grade 12

Timo Gnambs, Luise Fischer, and Theresa Rohm

Leibniz Institute for Educational Trajectories, Bamberg, Germany

E-mail address of lead author:

timo.gnambs@lifbi.de

Bibliographic data:

Gnambs, T., Fischer, L., & Rohm, T. (2017). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 4 for Grade 12* (NEPS Survey Paper No. 13). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP13:1.0

Acknowledgements:

This report is an extension to NEPS working paper 16 (Haberkorn, Pohl, Hardt, & Wiegand, 2012) that presents the scaling results for reading competence of starting cohort 4 for grade 9. Therefore, various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (Haberkorn et al., 2012; Hardt, K., Pohl, S., Haberkorn, K., & Wiegand, E., 2013) to facilitate the understanding of the presented results.

We thank Anna Scharl and Micha Freund for their assistance in scaling the data.

NEPS Technical Report for Reading: Scaling Results of Starting Cohort 4 for Grade 12

Abstract

The National Educational Panel Study (NEPS) investigates the development of competencies across the life span and develops tests for the assessment of different competence domains. In order to evaluate the quality of the competence tests, a range of analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedures for the reading competence test in grade 12 of starting cohort 4 (ninth grade). The reading competence test contained 29 items (distributed among an easy and a difficult booklets) with different response formats representing different cognitive requirements and text functions. The test was administered to 5,805 students. Their responses were scaled using the partial credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, the test's dimensionality, and local item independence were evaluated to ensure the quality of the test. These analyses showed that the test exhibited an acceptable reliability and that the items fitted the model in a satisfactory way. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test were 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 related to the dimensionality analyses based on both text functions and cognitive requirements. Overall, the reading test had acceptable psychometric properties that allowed for an estimation of reliable reading competence scores. Besides the scaling results, this paper also 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

Content

1. Introduction

Within the National Educational Panel Study (NEPS) different competences are measured coherently across the life span. These include, among others, reading competence, mathematical competence, scientific literacy, information and communication technologies literacy, metacognition, vocabulary, and domain general cognitive functioning. An overview of the competences measured in the NEPS is given by Weinert and colleagues (2011) as well as Fuß, Gnambs, Lockl, and Attig (2016).

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 the NEPS, several analyses were conducted to evaluate the quality of the tests. The IRT models chosen for scaling the competence data and the analyses performed for checking the quality of the scale are described in Pohl and Carstensen (2012).

In this paper the results of these analyses are presented for reading competence in grade 12 of starting cohort 4 (ninth grade). First, the main concepts of the reading competence test are introduced. Then, the reading competence data of starting cohort 4 and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the scientific use file is presented.

Please note that the analyses in this report are based on the data available at some time before public data release. Due to ongoing data protection and data cleansing issues, the data in the scientific use file (SUF) may differ slightly from the data used for the analyses in this paper. However, we do not expect fundamental changes in the presented results.

2. Testing Reading Competence

The framework and test development for the reading competence test are described by Weinert and colleagues (2011) and Gehrer, Zimmermann, Artelt, and Weinert (2013). In the following, we briefly describe specific aspects of the reading competence test that are necessary for understanding the scaling results presented in this paper.

The reading competence test included five texts and five item sets referring to these texts. Each of these texts represented one text type or text function, namely, a) information, b) commenting or argumenting, c) literary, d) instruction, and e) advertising (see Gehrer et al., 2013, and Weinert et al., 2011, for the description of the framework). Furthermore, the test assessed 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 and Artelt, 2013, Gehrer et al., 2013, and Weinert et al., 2011, for a detailed description of the framework).

The reading competence test included two types of response formats: simple multiple choice (MC) items and complex multiple choice (CMC) items. MC items had four response options. One response option represented a correct solution, whereas the other three were distractors (i.e., they were incorrect). In CMC items a number of subtasks with two response options were

presented. Examples of the different response formats are given in Pohl and Carstensen (2012) and Gehrer, Zimmermann, Artelt and Weinert (2012).

The competence test for reading that was administered in the present study included 42 items. In order to evaluate the quality of these items extensive preliminary analyses were conducted. These preliminary analyses identified a poor fit for one item (reg12041s_c). Therefore, this item was removed from the final scaling procedure. Thus, the analyses presented in the following sections and the competence scores derived for the respondents are based on the remaining 41 items.

3. Data

3.1 The Design of the Study

The study followed a three-factorial (quasi-)experimental design. These factors referred to (a) the position of the reading test within the test battery, (b) the difficulty of the administered test, and (c) the assessment setting (i.e., the context of test administration).

The study assessed different competence domains including, among others, reading competence, computer literacy, and mathematics. The competence tests for these three domains were always presented first within the test battery. In order to control for test position effects, the tests were administered to participants in different sequence. For each participant the reading test was either administered as the first or the second test (i.e., after the computer literacy or the mathematics test). There was no multi-matrix design regarding the order of the items *within* a specific test. All students received the test items in the same order.

Table 1

Number of Items for the Different Text Types by Difficulty of the Test

Text types	**Easy test**	**Both tests**	**Difficult test**
Information text	6		5
Instruction text		6	
Advertising text		4	
Commenting text		5	1
Literary text	7		7
Total number of items	13	15	13

In order to measure participants' reading competence with great accuracy, the difficulty of the administered items should adequately match the participants' abilities. Therefore, the study adopted the principles of longitudinal multistage testing (Pohl, 2013). Based on preliminary studies two different versions of the reading competence test were developed

that differed in their average difficulty (i.e., an easy and a difficult test). Both tests included five texts and 28 items that represented the five text functions (see Table 1) and three cognitive requirements (see Table 2) as described above. Three texts with 15 items were identical in both test versions (see Table 1), whereas 13 items were unique to the easy and the difficult test. The different response formats of the items are summarized in Table 3. The number of subtasks within CMC items varied between three and six. Participants were assigned either to the easy or the difficult test based on their estimated reading competence in the previous assessment (Haberkorn et al., 2012). Participants with an ability estimate below the sample's mean ability received the easy test, whereas participants with a reading competence above the sample's mean received the difficult test.

Table 2

Number of Items by Cognitive Requirements and Difficulty of the Test

Cognitive requirements	**Easy test**	**Difficult test**
Finding information	6	6
Drawing text-related conclusions	7	7
Reflecting and assessing	15	15
Total number of items	28	28

The panel study aimed at retesting all students that were initially included in the starting cohort 4 for ninth grade (see Haberkorn et al., 2012). Because some students left their original schools during the course of the longitudinal study, the participants of the starting cohort were divided into two subsamples that exhibited different assessment settings: Students that remained at the same school as in the first assessment were tested at school in a group setting; in contrast, students that left their original school were tracked and, subsequently, individually tested at home (for details regarding the data collection process see the respective field report for wave 7). Thus, the context of test administration differed between the two groups.

Table 3

Number of Items by Different Response Formats and Difficulty of the Test

Response format	**Easy test**	**Difficult test**
Simple multiple choice items	20	20
Complex multiple choice items	8	8
Total number of items	28	28

3.2 Sample

A total of 5,805[1] individuals received the reading competence test. For four respondents less than three valid item responses were available. Because no reliable ability scores can be estimated based on such few valid responses, these cases were exclude from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 5,801 individuals. The number of participants within each (quasi-)experimental condition is given in Table 4. A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

Table 4

Number of Participants by the (Quasi-)Experimental Conditions

Assessment setting:		**At school**		**At home**		**Total**
	Test position:	**first position**	**second position**	**first position**	**second position**	
Test difficulty	**Easy test**	414	437	622	632	2,105
	Difficult test	1,499	1,531	313	353	3,696
	Total	1,913	1,968	935	985	5,801

4. Analyses

4.1 Missing Responses

Competence data include different 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, finally, e) multiple kinds of missing responses within CMC items that are not determined.

Invalid responses occurred, for example, when two response options were selected in simple MC items where only one was required, or when numbers or letters that were not within the range of valid responses were given as a response. Omitted items occurred when test takers skipped some items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response given were coded as not-reached. Because of the multi-stage testing design 26 items were not administered to all participants. For respondents receiving the easy test 13 difficult items were missing by design, whereas 13 easy items were missing by design for respondents answering the difficult test (see Table 1). As CMC items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A CMC item was coded

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

as missing if at least one subtask contained a missing response. When one subtask contained a missing response, the CMC item was coded as missing. If just one kind of missing response occurred, the item was coded according to the corresponding missing response. If the subtasks contained different kinds of missing responses, the item was labeled as a not-determinable missing response.

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

4.2 Scaling Model

Item and person parameters were estimated using a partial credit model (PCM; Masters, 1982). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

CMC items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly responded subtasks within that item. If at least one of the subtasks contained a missing response, the CMC item was scored as missing. Categories of polytomous variables with less than $N = 200$ responses were collapsed in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items; in these cases, the lower categories were collapsed into one category. For 8 of the 11 CMC items categories were collapsed (see Appendix A).

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

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

4.3 Checking the Quality of the Test

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

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

The MC items consisted of one correct response option and one or more distractors (i.e., incorrect response options). The quality of the distractors within MC items was examined using the point-biserial correlation between selecting an incorrect response option and the total correct score. Negative correlations indicate good distractors, whereas correlations between .00 and .05 are considered acceptable and correlations above .05 are viewed as problematic distractors (Pohl & Carstensen, 2012).

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC and polytomous CMC items to the partial credit model (Masters, 1982) was evaluated using three indices (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (t-value > |6|) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 (t-value > |8|) were judged as having a considerable item misfit and their performance was further investigated. Correlations of the item score with the corrected total score (equal to the corrected discrimination as computed in ConQuest) greater than .30 were considered as good, greater than .20 as acceptable, and below .20 as problematic. Overall judgment of the fit of an item was based on all fit indicators.

The reading competence test should measure the same construct for all students. If some items favored certain subgroups (e.g., they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., males and females) would be biased and, thus, unfair. For the present study, test fairness was investigated for the variables test position, gender, school degree, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012a, for a description of these variables). Moreover, in light of the quasi-experimental design measurement invariance analyses were also conducted for the test difficulty and administration setting. Differential item functioning (DIF) was examined 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. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties between the subgroups that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as considerable and noteworthy of further investigation, differences between 0.4 and 0.6 as small but not severe, and differences smaller than 0.4 as negligible DIF. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

The reading competence test was scaled using the PCM (Masters, 1982), which assumes Rasch-homogeneity. The PCM was chosen because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that might not hold for empirical data. To test the assumption of equal item discrimination parameters, a generalized partial credit model (GPCM; Muraki, 1992) was also fitted to the data and compared to the PCM.

The dimensionality of the test was evaluated by two different 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 among the dimensions as well as differences in model fit between the unidimensional model and the respective

multidimensional models were used to evaluate the unidimensionality of the test. Moreover, we examined whether the residuals of the one-dimensional model exhibited approximately zero-order correlations as indicated by Yen's (1984) Q_3. Because in case of locally independent items, the Q_3 statistic tends to be slightly negative, we report the corrected Q_3 that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) values of Q_3 falling below .20 indicate essential unidimensionality.

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

4.4 Software

The IRT models were estimated in ConQuest version 4.2.5 (Adams, Wu, & Wilson, 2015).

5. Results

5.1 Missing Responses

5.1.1 Missing responses per person

Figure 1 shows the number of invalid responses per person by experimental condition (i.e., test difficulty and administration setting). Overall, there were very few invalid responses. Between 92% and 96% of the respondents did not have any invalid response at all; less than three percent had more than one invalid response. There was no difference in the amount of invalid responses between the different experimental conditions.

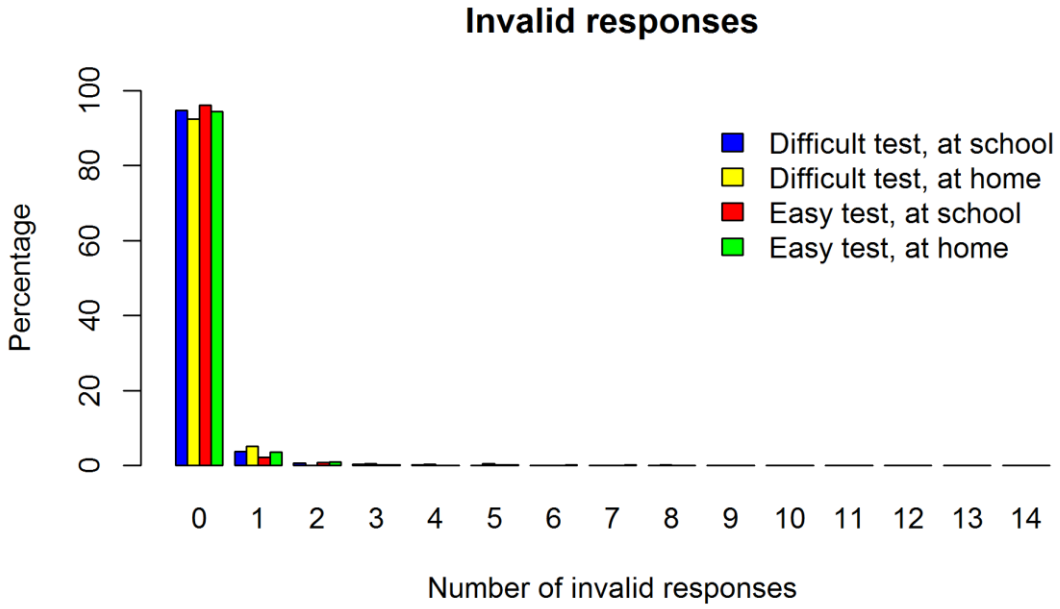

Figure 1. Number of invalid responses by experimental condition

Missing responses may also occur when respondents omit items. As illustrated in Figure 2 most respondents, 77% to 83%, did not skip any item and less than five percent omitted more

than three items. There was no difference in the amount of omitted items between the different experimental conditions.

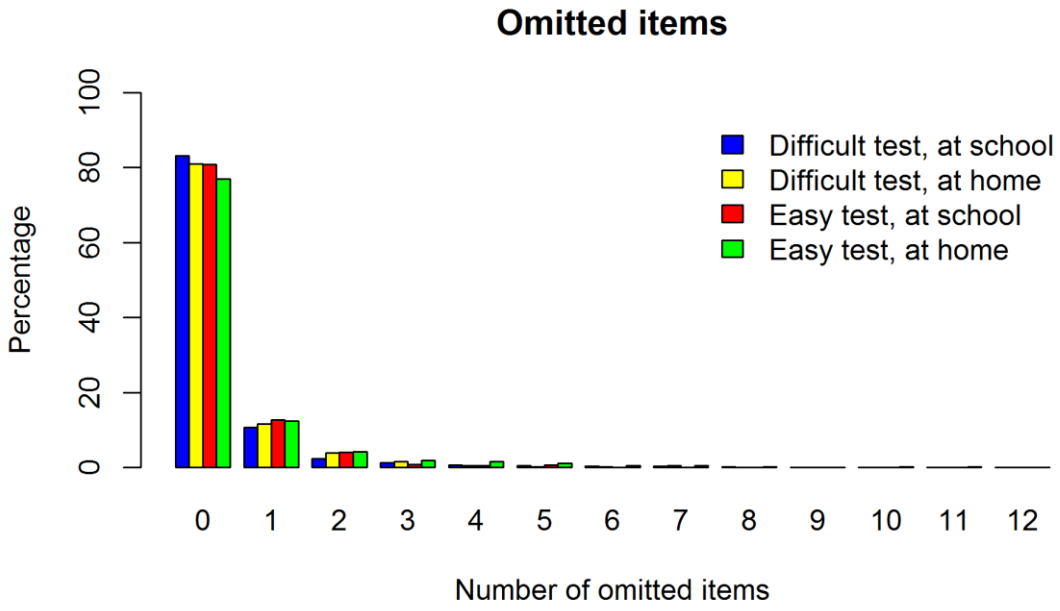

Figure 2. Number of omitted items by experimental condition

Another source of missing responses is items that were not reached by the respondents; these are all missing responses after the last valid response. The number of not-reached items was rather high because many respondents were unable to finish the test within the allocated time limit (Figure 3). Between 47% and 71% of the respondents finished the entire test. About ten percent did not reach the last of the five texts; in particular, respondents receiving the difficult test at home did not reach the last text.

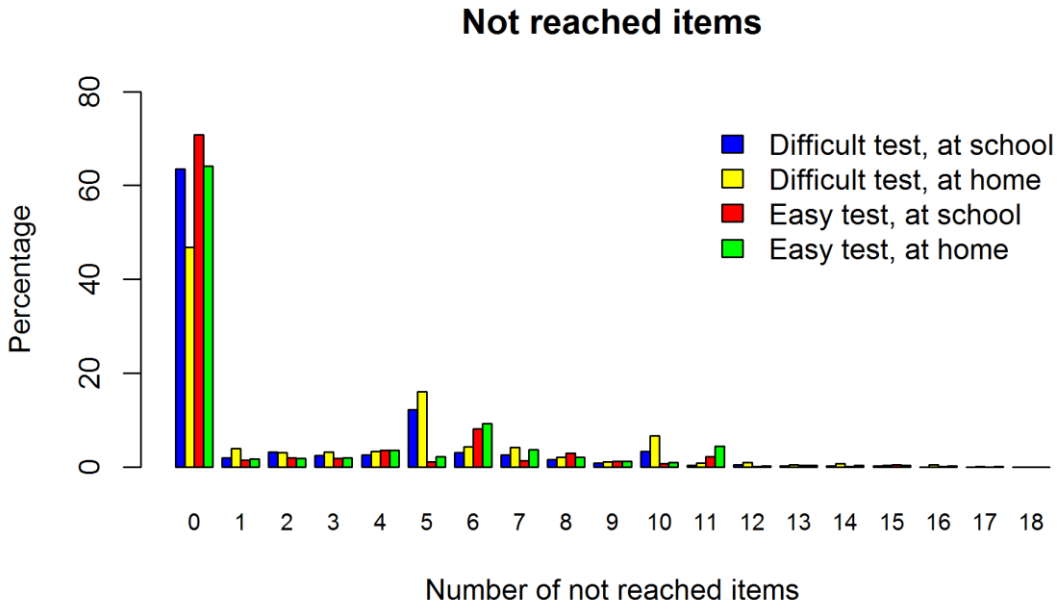

Figure 3. Number of not-reached items by experimental condition

The aggregated polytomous variables were coded as not-determinable missing response when the subtasks of CMC items contained different kinds of missing responses. Because not-determinable missing responses only occur in CMC items, the maximum number of not-determinable missing responses was eight (i.e., the number of CMC items). However, only a rather small number of not-determinable missing responses occurred. Most respondents, 99.31% to 99.60%, did not have any not-determinable missing response. There was no difference in the amount of not-determinable items between the experimental conditions.

The total number of missing responses, aggregated over invalid, omitted, not-reached, and not-determinable missing responses per person, is illustrated in Figure 4. On average, the respondents showed between $M = 2.28$ ($SD = 3.67$) and $M = 3.81$ ($SD = 4.21$) missing responses in the different experimental conditions. About 36% to 58% of the respondents had no missing response at all and about 27% to 46% of the participants had four or more missing responses. Particularly, respondents receiving the difficult test at home showed more missing responses because they did not reach the last of the five texts. Thus, it might be speculated that the test was somewhat too difficult for these respondents.

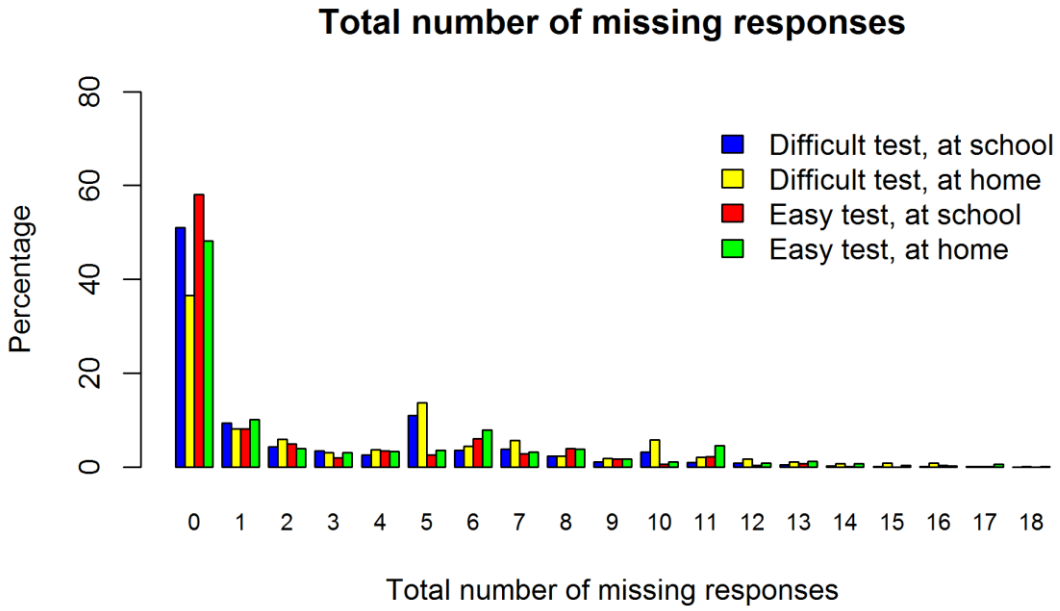

Figure 4. Total number of missing responses by experimental condition

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

Table 5

Percentage of Missing Values for the Difficult Test by Assessment Setting

		At school				*At home*			
Item	**Position**	N	**NR**	**OM**	**NV**	N	**NR**	**OM**	**NV**
reg120610_c	1	2,969	0.00	0.23	1.78	648	0.00	0.00	2.70
reg120620_c	2	2,987	0.00	0.73	0.69	654	0.00	0.30	1.50
reg120630_c	3	3,001	0.00	0.83	0.13	657	0.00	0.90	0.45
reg120640_c	4	2,982	0.00	0.92	0.66	652	0.00	0.90	1.20
reg12065s_c	5	2,969	0.00	2.01	0.00	656	0.00	1.50	0.00
reg120660_c	6	2,983	0.00	1.22	0.33	651	0.00	1.20	1.05
reg120670_c	7	3,014	0.00	0.50	0.03	662	0.00	0.30	0.30
reg12021s_c	8	3,002	0.00	0.89	0.03	656	0.15	1.35	0.00
reg120220_c	9	2,965	0.00	1.62	0.53	649	0.15	1.65	0.75
reg120230_c	10	2,994	0.03	0.43	0.73	654	0.15	0.30	1.35
reg12024s_c	11	2,951	0.03	2.57	0.00	641	0.15	3.60	0.00
reg120250_c	12	3,004	0.03	0.66	0.17	652	0.15	1.05	0.90
reg12026s_c	13	2,929	0.10	2.54	0.63	629	0.15	3.75	1.65
reg120310_c	14	2,982	0.40	0.99	0.20	651	0.75	1.05	0.45
reg120320_c	15	2,996	0.40	0.46	0.26	647	1.20	0.90	0.75
reg120330_c	16	2,925	0.76	2.64	0.07	633	1.95	2.85	0.15
reg120340_c	17	2,932	1.09	2.05	0.10	633	2.55	1.95	0.45
reg120350_c	18	2,919	1.65	1.88	0.13	625	3.60	1.80	0.75
reg120360_c	19	2,914	2.05	1.65	0.13	623	4.50	1.65	0.30
reg12042s_c	21	2,824	6.30	0.50	0.00	575	12.46	0.90	0.30
reg120430_c	22	2,739	7.92	0.40	1.29	548	14.56	1.05	2.10
reg12044s_c	23	2,633	10.59	1.32	0.89	519	18.77	2.25	0.90
reg120450_c	24	2,576	13.73	1.06	0.20	505	23.12	0.75	0.30
reg12071s_c	25	2,200	26.07	1.06	0.00	400	39.19	0.45	0.00
reg120720_c	26	2,129	28.68	0.92	0.13	377	42.64	0.15	0.60
reg12075s_c	27	2,056	31.22	0.76	0.17	355	45.95	0.00	0.75
reg120740_c	28	1,942	34.49	1.39	0.03	333	49.10	0.45	0.45
reg120730_c	29	1,911	36.50	0.33	0.03	309	53.15	0.30	0.00

Note. Position = Item position within test, N = Number of valid responses, NR = Percentage of respondents that did not reach item, OM = Percentage of respondents that omitted the item, NV = Percentage of respondents with an invalid response.
The item on position 20 was excluded from the analyses due to an unsatisfactory item fit (see section 2).

Table 6

Percentage of Missing Values for the Easy Test by Assessment Setting

		At school				*At home*			
Item	**Position**	N	**NR**	**OM**	**NV**	N	**NR**	**OM**	**NV**
reg120110_c	1	847	0.00	0.12	0.35	1,243	0.00	0.40	0.48
reg120120_c	2	848	0.00	0.35	0.00	1,241	0.00	1.04	0.00
reg120130_c	3	847	0.00	0.47	0.00	1,239	0.00	0.96	0.24
reg12014s_c	4	837	0.00	1.65	0.00	1,220	0.00	2.71	0.00
reg120150_c	5	843	0.00	0.94	0.00	1,231	0.00	1.83	0.00
reg120160_c	6	843	0.00	0.82	0.12	1,235	0.00	1.44	0.08
reg120170_c	7	847	0.00	0.12	0.35	1,237	0.00	0.80	0.56
reg12021s_c	8	843	0.00	0.94	0.00	1,226	0.08	2.15	0.00
reg120220_c	9	829	0.00	2.35	0.24	1,220	0.08	1.83	0.80
reg120230_c	10	841	0.00	0.47	0.71	1,227	0.08	0.96	1.12
reg12024s_c	11	814	0.12	4.23	0.00	1,192	0.32	4.63	0.00
reg120250_c	12	843	0.12	0.35	0.47	1,224	0.32	1.04	1.04
reg12026s_c	13	806	0.35	3.76	0.82	1,190	0.64	3.19	1.28
reg120310_c	14	834	0.94	0.82	0.24	1,179	1.12	4.55	0.32
reg120320_c	15	835	0.94	0.71	0.24	1,210	1.12	1.44	0.96
reg120330_c	16	819	1.18	2.59	0.00	1,159	1.59	5.82	0.16
reg120350_c	17	823	1.65	1.29	0.35	1,187	1.99	2.79	0.56
reg120360_c	18	825	1.88	1.18	0.00	1,186	2.31	2.95	0.16
reg12042s_c	20	801	4.94	0.94	0.00	1,143	7.89	0.96	0.00
reg120430_c	21	779	6.23	0.59	1.65	1,111	9.17	0.48	1.75
reg12044s_c	22	750	9.28	2.00	0.47	1,063	11.32	2.87	0.80
reg120450_c	23	745	10.69	1.53	0.24	1,040	15.07	1.36	0.64
reg120510_c	24	687	18.92	0.24	0.12	935	24.40	0.80	0.24
reg12052s_c	25	670	20.09	1.18	0.00	907	26.63	1.04	0.00
reg120530_c	26	646	23.74	0.12	0.24	862	30.22	0.48	0.56
reg120540_c	27	621	25.62	0.47	0.94	837	32.22	0.64	0.40
reg12055s_c	28	609	27.61	0.82	0.00	809	34.13	1.28	0.00
reg120560_c	29	601	29.14	0.00	0.24	797	35.89	0.00	0.56

Note. Position = Item position within test, N = Number of valid responses, NR = Percentage of respondents that did not reach item, OM = Percentage of respondents that omitted the item, NV = Percentage of respondents with an invalid response.
The item on position 19 was excluded from the analyses due to an unsatisfactory item fit (see section 2).

5.1.2 Missing responses per item

Tables 5 and 6 provide information on the occurrence of different kinds of missing responses per item for the easy and difficult test version. Overall, in both tests the omission rates were rather low, varying across items between 0.00% and 5.82%. There was only one item with an omission rate exceeding 5% (reg120330_c in the easy test administered at home). For the difficult test omission rates correlated with the item difficulties at about .12 in the school context and a about .15 at home; for the easy test the respective correlations were distinctly larger with .39 in the school setting and .31 in the home setting. Generally, participants were inclined to omit more difficult items. In contrast, the percentage of invalid responses per item (columns 6 and 10 in Tables 5 and 6) was rather low with the maximum rate being 2.70%.

With an item's progressing position in the test, the amount of persons that did not reach the item (columns 4 and 8 in Tables 5 and 6) rose up to a considerable amount of 29% to 53% for the different experimental conditions. Particularly, at home the last items of the difficult test were not reached by many respondents (see Figure 5).

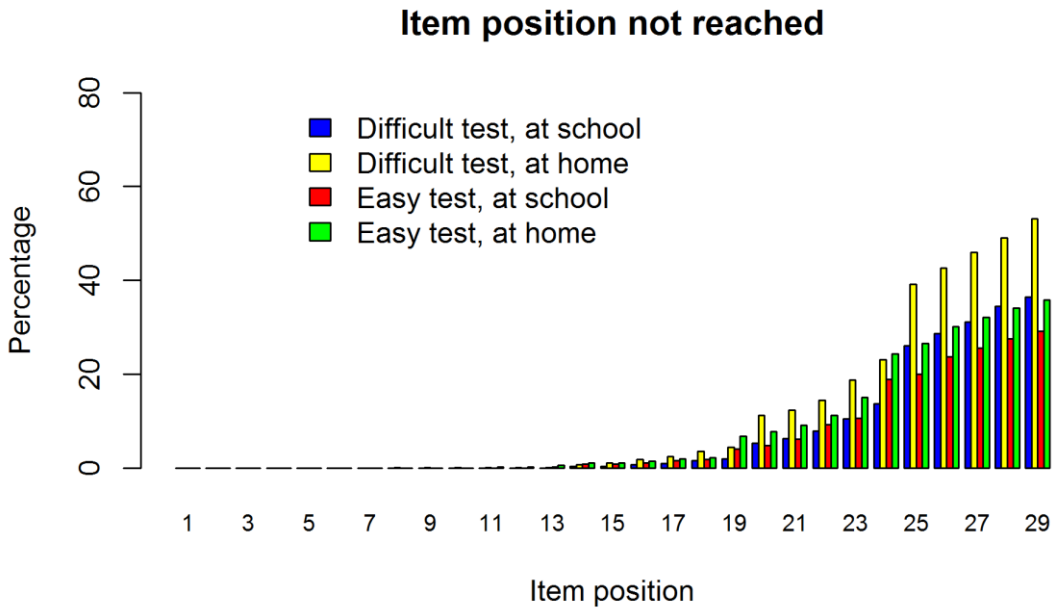

Figure 5. Item position not reached by experimental conditions

5.2 Parameter Estimates

5.2.1 Item parameters

The second column in Table 7 presents the percentage of correct responses in relation to all valid responses for each item. Because there is a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index for item difficulty. The percentage of correct responses within dichotomous items varied between 32% and 87% with an average of 67% (*SD* = 13%) correct responses.

Table 7

Item Parameters

	Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
1.	reg120110_c	0.66	-1.289	0.050	1.01	0.5	0.32	0.94	.03
2.	reg120120_c	0.57	-0.821	0.048	1.05	3.1	0.27	0.75	.03
3.	reg120130_c	0.83	-2.313	0.061	0.95	-1.5	0.35	1.45	.04
4.	reg12014s_c	0.87	-2.679	0.069	0.94	-1.2	0.32	1.53	.03
5.	reg120150_c	0.59	-0.907	0.048	1.00	0.1	0.34	0.98	.03
6.	reg120160_c	0.63	-1.141	0.049	1.02	1.2	0.30	0.88	.04
7.	reg120170_c	0.81	-2.146	0.059	1.00	0.1	0.26	0.95	.02
8.	reg12021s_c	0.67	-0.807	0.030	0.98	-1.8	0.36	0.97	.02
9.	reg120220_c	0.60	-0.488	0.029	1.07	6.1	0.24	1.04	.02
10.	reg120230_c	0.81	-1.691	0.036	1.01	0.5	0.27	0.88	.02
11.	reg12024s_c	n.a.	-1.142	0.028	0.92	-5.6	0.45	1.41	.03
12.	reg120250_c	0.83	-1.831	0.037	0.94	-2.6	0.35	1.29	.03
13.	reg12026s_c	n.a.	-0.223	0.019	1.09	5.7	0.38	0.67	.05
14.	reg120310_c	0.53	-0.145	0.029	1.04	4.3	0.28	0.74	.03
15.	reg120320_c	0.80	-1.627	0.036	0.95	-2.3	0.35	1.20	.02
16.	reg120330_c	0.69	-0.940	0.031	0.96	-2.7	0.37	1.16	.03
17.	reg120340_c	0.47	0.412	0.036	1.00	0.3	0.28	0.83	.03
18.	reg120350_c	0.59	-0.397	0.030	0.99	-1.3	0.36	0.97	.03
19.	reg120360_c	0.67	-0.809	0.031	0.98	-1.6	0.34	1.04	.02
21.	reg12042s_c	0.47	0.167	0.030	0.97	-2.7	0.36	1.04	.02
22.	reg120430_c	0.51	-0.031	0.030	1.08	7.6	0.23	0.58	.03
23.	reg12044s_c	n.a.	-1.285	0.025	0.87	-7.1	0.53	1.45	.03
24.	reg120450_c	0.65	-0.710	0.032	1.01	0.4	0.33	0.92	.02
25.	reg120510_c	0.81	-2.223	0.067	0.89	-2.9	0.42	1.85	.04
26.	reg12052s_c	n.a.	-1.178	0.063	1.04	1.9	0.24	0.74	.03
27.	reg120530_c	0.78	-2.016	0.066	0.95	-1.5	0.36	1.33	.03
28.	reg120540_c	0.69	-1.499	0.061	0.98	-0.7	0.34	1.06	.02

	Item	Percentage correct	Item difficulty	SE	WMNSQ	t	r_{it}	Discr.	Q_3
29.	reg12055s_c	n.a.	-0.423	0.067	0.98	-1.0	0.35	1.16	.03
30.	reg120560_c	0.63	-1.176	0.060	1.01	0.3	0.33	0.97	.02
31.	reg120610_c	0.65	-0.422	0.037	1.08	5.4	0.15	0.42	.03
32.	reg120620_c	0.74	-0.881	0.040	1.01	0.3	0.26	0.82	.03
33.	reg120630_c	0.83	-1.517	0.046	0.98	-0.8	0.29	1.10	.02
34.	reg120640_c	0.66	-0.475	0.037	1.02	1.2	0.26	0.83	.03
35.	reg12065s_c	n.a.	-1.460	0.039	0.98	-1.0	0.33	1.11	.03
36.	reg120660_c	0.32	1.132	0.038	1.06	3.6	0.17	0.52	.02
37.	reg120670_c	0.83	-1.500	0.046	1.00	0.1	0.22	0.79	.02
38.	reg12071s_c	0.75	-0.942	0.048	0.99	-0.6	0.30	0.98	.02
39.	reg120720_c	0.85	-1.672	0.058	1.06	1.5	0.13	0.44	.03
40.	reg120730_c	0.57	-0.073	0.044	1.05	3.1	0.22	0.62	.02
41.	reg120740_c	0.54	0.066	0.045	1.05	3.5	0.22	0.62	.02
42.	reg12075s_c	n.a.	-1.288	0.060	0.99	-0.5	0.27	0.99	.02

Note. Difficulty = Item difficulty / location parameter, SE = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, t = t-value for WMNSQ, r_{it} = Corrected item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model, Q_3 =Average absolute residual correlation for item (Yen, 1983).

Item 20 was excluded from the analyses due to an unsatisfactory item fit (see section 2).

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

For the dichotomous items, the item-total correlation corresponds to the point-biserial correlation between the correct response and the total score; for polytomous items it corresponds to the product-moment correlation between the corresponding categories and the total score (discrimination value as computed in ConQuest).

The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 7. The step parameters for polytomous variables are depicted in Table 8. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties (or location parameters for polytomous variables) ranged from -2.7 (item reg12014s_c) to 1.1 (item reg120660_c) with an average difficulty of -1.0. Overall, the item difficulties were rather low; there were no items with a high difficulty. Due to the large sample size the standard errors (SE) of the estimated item difficulties (column 4 in Table 7) were rather small (all SEs ≤ 0.07).

Table 8

Step Parameters (with Standard Errors) for Polytomous Items

Item	Step 1	Step 2	Step 3	Step 4	Step 5
reg12024s_c	-0.507 (0.040)	0.453 (0.042)	0.054		
reg12026s_c	-0.349 (0.046)	-0.267 (0.047)	0.083 (0.044)	0.805 (0.051)	-0.272
reg12044s_c	-0.224 (0.059)	0.123 (0.058)	0.354 (0.052)	-0.253	
reg12052s_c	0.295 (0.059)	-0.295			
reg12055s_c	0.176 (0.60)	-0.176			
reg12065s_c	-0.215 (0.066)	0.022 (0.061)	0.193		
reg12075s_c	0.025 (0.052)	-0.025			

5.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. In Figure 6, the 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. The variance was estimated to be 0.829, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = .795) was good. The mean of the item distribution was about 0.99 logits below the mean person ability distribution. Thus, 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.

Scale (in logits)	**Person ability**	**Item difficulty**
3		
2	X	
	X	
	X	
	XX	
	XX	
	XXXX	
	XXXX	
	XXXX	35
1	XXXXX	
	XXXXXXX	
	XXXXXXX	
	XXXXXXXX	
	XXXXXXXXX	
	XXXXXXX	17
	XXXXXXXXX	
	XXXXXXXXX	20
	XXXXXXX	40
0	XXXXXXXXXX	21 39
	XXXXXXXXXX	13 14
	XXXXXXXX	
	XXXXXXXX	18 28 30 33
	XXXXXXX	8
	XXXXXX	23
	XXXXXXX	2 9 19
	XXXXX	5 16 31 37
-1	XXXX	
	XXXX	6 11 25 29
	XXXX	1 22 41
	XXX	
	XX	27 32 34 36
	XX	10 15 38
	XX	
	X	12
-2	X	26
	X	7
	X	24
		3
		4
-3		

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 34 cases. The difficulty of the items is depicted on the right-hand side of the graph, with each number representing one item (corresponding to Table 7).

5.3 Quality of the test

5.3.1 Fit of the subtasks of complex multiple choice items

Before the subtasks of CMC items were aggregated and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the single subtasks together with the MC items in a Rasch model. Counting the subtasks of CMC items separately, there were 64 items. The probability of a correct response ranged from 32% to 95% across all items (*Mdn* = 71%). Thus, the number of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.80 to 1.07, the respective *t*-value from -11.1 to 6.7, and there were no noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. Due to the good model fit of the subtasks, their aggregation to polytomous variables seemed justified.

5.3.2 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the partial credit model, using the MC and polytomous CMC items. Altogether, item fit can be considered to be very good (see Table 7). Values of the WMNSQ ranged from 0.87 (item reg12044s_c) to 1.09 (reg12026s_c). Only three items exhibited a *t*-value of the WMNSQ greater than 6 and none exceeded a value of 8. Thus, there was no indication of severe item over- or underfit. Point-biserial correlations between the item scores and the total scores ranged from .13 (item reg120720_c) to .53 (item reg12044s_c) and had a mean of .30. All item characteristic curves showed a good fit of the items.

5.3.3 Distractor analyses

In addition to the overall item fit, we specifically investigated how well the distractors performed in the test by evaluating the point-biserial correlation between each incorrect response (distractor) and the students' total correct score. The point-biserial correlations for the distractors ranged from -.41 to .03 with a mean of -.17. These results indicate that the distractors functioned well.

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 sex, the number of books at home (as a proxy for socioeconomic status), migration background, school type, and test position (see Pohl & Carstensen, 2012, for a description of these variables). In addition, was also studied the effect of the two experimental factors. Thus, we compared the two assessment settings (at school or at home) and for the common items that were administered to all participants we examined measurement invariance for the easy and difficult test. The differences between the estimated item difficulties in the various groups are summarized in Table 9. For example, the column "Male vs. female" reports the differences in item difficulties between men and women; a positive value would indicate that the test was more difficult for males, whereas a negative value would highlight a lower difficulty for males as opposed to females. Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF to those that only estimate main effects (see Table 10).

Table 9

Differential Item Functioning

Item	Sex	Books	Migration	School	Position	Setting	Booklet
	male vs. female	< 100 vs. ≥ 100	without vs. with	no sec. vs. sec.	first vs. second	school vs. home	easy vs. difficult
reg120110_c	0.232 (0.256)	-0.006 (-0.007)	-0.076 (-0.085)	-0.134 (-0.175)	0.000 (0.000)	0.012 (0.016)	
reg120120_c	0.178 (0.197)	0.024 (0.028)	0.136 (0.153)	-0.176 (-0.230)	0.034 (0.037)	0.110 (0.144)	
reg120130_c	0.446 (0.493)	-0.040 (-0.047)	0.124 (0.139)	-0.060 (-0.078)	-0.004 (-0.004)	0.088 (0.115)	
reg12014s_c	0.676 (0.747)	-0.230 (-0.271)	0.334 (0.375)	-0.132 (-0.172)	-0.040 (-0.044)	0.254 (0.332)	
reg120150_c	0.026 (0.029)	-0.052 (-0.061)	0.096 (0.108)	-0.052 (-0.068)	0.056 (0.062)	0.106 (0.139)	
reg120160_c	0.384 (0.424)	-0.268 (-0.316)	0.154 (0.173)	-0.492 (-0.642)	-0.024 (-0.026)	0.410 (0.537)	
reg120170_c	0.142 (0.157)	0.040 (0.047)	0.228 (0.256)	-0.032 (-0.042)	0.142 (0.157)	0.090 (0.118)	
reg12021s_c	-0.190 (-0.210)	0.166 (0.196)	-0.142 (-0.159)	0.290 (0.379)	-0.002 (-0.002)	-0.232 (-0.304)	0.002 (0.002)
reg120220_c	0.008 (0.009)	-0.154 (-0.182)	0.008 (0.009)	-0.222 (-0.290)	-0.048 (-0.053)	0.242 (0.317)	-0.350 (-0.435*)
reg120230_c	-0.548 (-0.605*)	0.084 (0.099)	-0.316 (-0.354)	0.218 (0.285)	0.140 (0.154)	-0.040 (-0.052)	0.002 (0.002)
reg12024s_c	0.176 (0.194)	0.206 (0.243)	-0.020 (-0.022)	0.296 (0.386)	-0.198 (-0.218)	-0.388 (-0.508)	0.208 (0.259)
reg120250_c	0.008 (0.009)	0.342 (0.403)	-0.148 (-0.166)	0.250 (0.326)	-0.012 (-0.013)	-0.270 (-0.353)	0.170 (0.211)
reg12026s_c	-0.118 (-0.130)	-0.088 (-0.104)	-0.100 (-0.112)	-0.270 (-0.352)	-0.118 (-0.130)	0.226 (0.296)	-0.258 (-0.321*)
reg120310_c	-0.366 (-0.404)	-0.086 (-0.101)	0.188 (0.211)	-0.062 (-0.081)	0.074 (0.082)	0.066 (0.086)	-0.262 (-0.326)
reg120320_c	0.084 (0.093)	0.028 (0.033)	0.012 (0.013)	0.276 (0.360)	0.090 (0.099)	-0.298 (-0.390)	0.194 (0.241)
reg120330_c	0.132 (0.146)	0.182 (0.215)	-0.076 (-0.085)	0.296 (0.386)	0.032 (0.035)	-0.314 (-0.411)	0.098 (0.122)

Item	Sex	Books	Migration	School	Position	Setting	Booklet
reg120340_c	-0.144 (-0.159)	-0.078 (-0.092)	0.096 (0.108)	0.076 (0.099)	0.128 (0.141)	-0.148 (-0.194)	
reg120350_c	0.168 (0.186)	0.242 (0.285)	-0.050 (-0.056)	0.300 (0.392)	0.074 (0.082)	-0.288 (-0.377)	-0.094 (-0.117)
reg120360_c	0.058 (0.064)	0.090 (0.106)	-0.058 (-0.065)	0.422 (0.551)	0.036 (0.040)	-0.466 (-0.610*)	0.434 (0.540*)
reg12042s_c	-0.274 (-0.303)	0.114 (0.134)	-0.004 (-0.004)	0.348 (0.454)	-0.120 (-0.132)	-0.350 (-0.458)	-0.024 (-0.030)
reg120430_c	-0.040 (-0.044)	-0.072 (-0.085)	-0.078 (-0.087)	-0.018 (-0.023)	0.172 (0.190)	0.012 (0.016)	-0.198 (-0.246)
reg12044s_c	0.268 (0.296)	0.264 (0.311)	-0.198 (-0.222)	0.498* (0.650)	0.008 (0.009)	-0.584 (-0.764*)	0.282 (0.351)
reg120450_c	-0.148 (-0.163)	0.102 (0.12)	-0.012 (-0.013)	0.036 (0.047)	-0.010 (-0.011)	-0.042 (-0.055)	-0.202 (-0.251)
reg120510_c	0.254 (0.280)	0.078 (0.092)	-0.536 (-0.601)	0.234 (0.305)	0.018 (0.020)	-0.158 (-0.207)	
reg12052s_c	-0.098 (-0.108)	0.060 (0.071)	-0.258 (-0.289)	-0.060 (-0.078)	0.004 (0.004)	0.098 (0.128)	
reg120530_c	-0.300 (-0.331)	0.310 (0.366)	-0.206 (-0.231)	0.246 (0.321)	0.106 (0.117)	-0.328 (-0.429)	
reg120540_c	-0.008 (-0.009)	0.066 (0.078)	-0.016 (-0.018)	0.018 (0.023)	0.260 (0.287)	0.042 (0.055)	
reg12055s_c	0.420 (0.464)	0.198 (0.234)	-0.368 (-0.413)	0.050 (0.065)	-0.326 (-0.359)	-0.098 (-0.128)	
reg120560_c	-0.200 (-0.221)	0.186 (0.219)	-0.216 (-0.242)	0.156 (0.204)	0.030 (0.033)	-0.224 (-0.293)	
reg120610_c	0.188 (0.208)	-0.296 (-0.349)	0.078 (0.087)	-0.486 (-0.634)	-0.440 (-0.485)	0.436 (0.571)	
reg120620_c	-0.008 (-0.009)	-0.024 (-0.028)	0.088 (0.099)	-0.162 (-0.211)	-0.074 (-0.082)	0.048 (0.063)	
reg120630_c	-0.194 (-0.214)	-0.048 (-0.057)	-0.052 (-0.058)	-0.012 (-0.016)	-0.228 (-0.251)	-0.046 (-0.060)	
reg120640_c	-0.184 (-0.203)	-0.190 (-0.224)	0.154 (0.173)	-0.318 (-0.415)	-0.154 (-0.17)	0.410 (0.537)	
reg12065s_c	-0.332 (-0.367)	-0.032 (-0.038)	0.566 (0.635*)	-0.148 (-0.193)	-0.248 (-0.273)	0.160 (0.209)	
reg120660_c	-0.262 (-0.289)	0.052 (0.061)	-0.104 (-0.117)	-0.244 (-0.318)	0.112 (0.123)	0.392 (0.513)	

Item	Sex	Books	Migration	School	Position	Setting	Booklet
reg120670_c	-0.026 (-0.029)	-0.262 (-0.309)	0.196 (0.220)	-0.164 (-0.214)	-0.066 (-0.073)	0.142 (0.186)	
reg12071s_c	0.124 (0.137)	0.010 (0.012)	0.164 (0.184)	-0.032 (-0.042)	0.104 (0.115)	0.224 (0.293)	
reg120720_c	-0.154 (-0.170)	-0.446 (-0.526)	0.158 (0.177)	-0.482 (-0.629)	0.060 (0.066)	0.398 (0.521)	
reg12075s_c	-0.048 (-0.053)	-0.190 (-0.224)	0.204 (0.229)	0.146 (0.191)	0.086 (0.095)	-0.094 (-0.123)	
reg120740_c	-0.218 (-0.241)	-0.194 (-0.229)	0.080 (0.090)	0.034 (0.044)	0.122 (0.134)	0.036 (0.047)	
reg120730_c	-0.108 (-0.119)	-0.080 (-0.094)	-0.032 (-0.036)	-0.444 (-0.580)	0.228 (0.251)	0.368 (0.482)	
Main effect	-0.270 (-0.298)	-0.648 (-0.764)	0.370 (0.415)	-0.958 (-1.250)	0.164 (0.181)	0.982 (1.285)	-1.112 (-1.382)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. Sec. = Secondary school (German: „Gymnasium").
* Absolute standardized difference is significantly, $p < .05$, greater than 0.25 (see Fischer et al., 2016).

Sex: The sample included 2,701 (47%) males and 3,083 (53%) females. Seventeen respondents that did not indicate their sex were excluded from the analysis. On average, male participants had a lower estimated reading ability than females (main effect = -0.270 logits, Cohen's d = -0.298). Only one item (item reg12014s_c) showed DIF greater than 0.6 logits. An overall test for DIF (see Table 10) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). A model comparison using Akaike's (1974) information criterion (AIC) favored the model estimating DIF, whereas the Bayesian information criterion (BIC; Schwarz, 1978) that takes the number of estimated parameters into account and, thus, guards against overparameterization of models, indicated a better fit for the more parsimonious model including only the main effect. Thus, overall, there was no pronounced DIF with regard to sex.

Books: The number of books at home was used as a proxy for socioeconomic status. There were 1,651 (28%) test takers with 0 to 100 books at home, 3,877 (67%) test takers with more than 100 books at home, and 273 (4%) test takers without a valid response. There were considerable average differences between the two groups. Participants with 100 or less books at home performed on average 0.648 logits (Cohen's d = 0.764) lower in reading than participants with more than 100 books. There was no considerable DIF comparing participants with many or fewer books (highest DIF = 0.446 for item reg120720_c). As a consequence, also the overall test for DIF favored the main effects model (Table 10).

Table 10

Comparisons of Models with and without DIF

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	main effect	5,784	199,895.83	57	200,009.83	200,389.60
	DIF	5,784	199,581.54	97	199,775.54	200,421.80
Books	main effect	5,528	190,415.87	57	190,529.87	190,907.10
	DIF	5,528	190,270.22	97	190,464.22	191,106.10
Migration	main effect	5,693	196,551.20	57	196,665.20	197,044.10
	DIF	5,693	196,416.23	97	196,610.23	197,255.00
School	main effect	5,801	199,060.48	57	199,174.48	199,554.40
	DIF	5,801	198,610.58	97	198,804.58	199,451.20
Position	main effect	5,801	200,516.24	57	200,630.24	201,010.20
	DIF	5,801	200,391.05	97	200,585.05	201,231.60
Setting	main effect	5,801	199,022.07	57	199,136.07	199,516.00
	DIF	5,801	198,563.73	97	198,757.73	199,404.30
Difficulty	main effect	5,801	120,170.43	26	120,222.43	120,395.70
	DIF	5,801	119,959.30	40	120,039.30	120,305.90

Migration background: There were 4,285 participants (74%) with no migration background, 1,408 subjects (24%) with a migration background, and 108 individuals (2%) that did not indicate their migration background. In comparison to subjects with migration background, participants without migration background had, on average, a slightly higher reading ability (main effect = 0.370 logits, Cohen's d = 0.415). There was no noteworthy item DIF due to migration background; differences in estimated difficulties did not exceed 0.6 logits. Moreover, the overall test for DIF using the BIC also favored the main effects model that did not include item-level DIF.

School type: Overall, 3,656 subjects (63%) who took the reading test attended secondary school (German: "Gymnasium") whereas 2,145 (37%) were enrolled in other school types. Subjects in secondary schools showed a higher reading ability on average (0.958 logits, Cohen's d = 1.250) than subjects in other school types. There was no noteworthy item DIF; no item exhibited DIF greater than 0.6 logits. However, the overall model test indicated a slightly better fit for the more complex DIF model, because several items showed DIF effects between 0.4 and 0.6; however, these differences were not considered severe.

Position: The reading competence test was administered in two different positions (see section 3.1 for the design of the study). A subsample of 2,848 (49%) persons received the reading test first and 2,953 (51%) respondents took the reading test after having completed

either the computer literacy or 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. The results show minor average effect of item position[2]. Subjects who received the reading test first performed on average 0.164 logits (Cohen's d = 0.181) better than subjects who received the reading test second. There was no DIF due to the position of the test in the booklet. The largest difference in difficulty between the two design groups was 0.490 logits (item reg120610_c). As a consequence, the overall test for DIF using the BIC favored the more parsimonious main effect model.

Setting: The reading competence test was administered in two different settings (see section 3.1 for the design of the study). A subsample of 3,881 (67%) persons received the reading test in small groups at school, whereas 1,920 (33%) participants finished the test individually at their private homes. Subjects who finished the reading test at school were on average 0.982 logits (Cohen's d = 1.250) better than those working at their private homes. However, this difference must not be interpreted as a causal effect of the administration setting because respondents were not randomly assigned to the different settings. Rather, it is likely that self-selection processes occurred, for example, because less proficient students were more likely to leave school and, consequently, were tested at home. More importantly, there was no noteworthy DIF due to the administration setting; all differences in item difficulties were smaller than 0.6 logits. Again, the overall model test (see Table 10) indicated a slightly better fit for the more complex DIF model, because several items showed DIF effects between 0.4 and 0.6; however, these differences were not considered severe.

Booklet: To estimate the participants' proficiency with great accuracy the participants received different tests that either included a larger number of easy or a larger number of difficult items (see section 3.1 for the design of the study). Only a subset of 15 items that were included in both tests was administered to all participants. For these common items we examined potential DIF across the two test versions (easy versus difficult). A subsample of 2,105 (36%) persons received the easy test and 3,696 (64%) persons received the difficult test. As expected, subjects who were administered the easy test scored on average -1.112 logits (Cohen's d = -1.382) lower than subjects who received the difficult test. There was no DIF for the common items with regard to the test version. The largest difference in difficulties between the two groups was 0.434 logits (item reg120360_c).

5.3.5 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is 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 7), ranging from 0.42 (item reg120320_c) to 1.85 (item reg12071s_c). The average discrimination parameter fell at 0.97. Model fit indices suggested a slightly better model fit of the 2PL model (AIC = 199,730.80, BIC = 200,370.70) as compared to the 1PL model (AIC = 200672.80, BIC = 201046.10). 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

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

discussion of this issue). For this reason, the partial credit model (1PL) was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

5.3.6 Unidimensionality

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 models was carried out in ConQuest using Gauss-Hermite quadrature method.

The estimated variances and correlations between the three dimensions representing the different cognitive requirements are reported in Table 11. The correlations among the three dimensions were rather high and fell between .93 and .96. However, they deviated from a perfect correlation (i.e., they were marginally lower than r = .95, see Carstensen, 2013). Moreover, according to model fit indices, the three-dimensional model fitted the data slightly better (AIC = 200,536.34, BIC = 200,971.00, number of parameters = 61) than the unidimensional model (AIC = 200,672.80, BIC = 201,046.10, number of parameters = 56). These results indicate that the three cognitive requirements measure a common construct, albeit it is not completely unidimensional.

Table 11

Results of Three-Dimensional Scaling

	Dim 1	**Dim 2**	**Dim 3**
Finding information in the text (Dim 1)	(1.29)		
(8 items)			
Drawing text-related conclusions (Dim 2)	.95	(0.78)	
(13 items)			
Reflecting and assessing (Dim 3)	.96	.93	(0.77)
(20 items)			

Note. Variances of the dimensions are given in the diagonal and correlations are presented in the off-diagonal.

The estimated variances and correlations of the five-dimensional model based on the five text functions are given in Table 12. The correlations between the dimensions varied between r = .72 and r = .93. The smallest correlation was found between Dimension 1 (“literary function”) and Dimension 5 (“information”). Dimension 2 (“instruction text”) and Dimension 4 (“commenting function”) 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 = 200,301.14, BIC = 200,827.70, number of parameters = 70) fitted the data better than the unidimensional model (AIC = 200,672.80, BIC = 201,046.10, number of parameters = 56). As each text function corresponded to one of the five texts, local item dependence (LID) and the text functions were confounded. As a

consequence, the deviation of the correlations from a perfect correlation shown in Table 12, may result from multidimensionality as well as from local item dependence. Given the testing design in the main studies, it is not possible to disentangle the two sources. In pilot studies (Gehrer et al., 2013), a larger number of texts were presented to test takers, so that the impact of text functions could be investigated independently of LID. The correlations estimated in the pilot study ranged from .78 to .91. As the correlations found in Gehrer and colleagues (2013) differ from a perfect correlation, it is concluded that text functions form subdimensions of reading competence. Comparing the correlations found in Gehrer et al. (2013), which are due to text functions, to those found in the main study (Table 12), which are due to both text functions and LID, allows us to evaluate the impact of LID. The correlations found in the present study of starting cohort 4 were slightly lower (between 0.72 and 0.93) than those found in Gehrer et al. (between 0.78 and 0.91), indicating that there is some amount of local item dependence. However, according to the test developers a balanced assessment of reading competence can only be achieved by a heterogeneity of text functions (Gehrer et al., 2013).

However, for the unidimensional model the average absolute residual correlations as indicated by the Q_3 statistic (see Table 7) were quite low (M = .03, SD = .01)—the largest individual residual correlation was .14—and thus indicated an essentially unidimensional test. Because the reading test is constructed to measure a single dimension, a unidimensional reading competence score was estimated.

Table 12

Results of Five-Dimensional Scaling

	Dim 1	**Dim 2**	**Dim 3**	**Dim 4**	**Dim 5**
Literary (Dim 1)	(0.92)				
(14 items)					
Instruction (Dim 2)	.77	(0.96)			
(6 items)					
Commenting (Dim 3)	.81	.83	(1.19)		
(6 items)					
Advertising (Dim 4)	.76	.93	.87	(1.07)	
(4 items)					
Information (Dim 5)	.72	.81	.73	.87	(1.04)
(15 items)					

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

6. Discussion

The analyses in the previous sections aimed at providing detailed information on the quality of the reading test in starting cohort 4 for grade 12 and at describing how the reading competence score was estimated.

We investigated different kinds of missing responses and examined the item and test parameters. We thoroughly checked item fit statistics for simple MC items, subtasks of CMC items, as well as the aggregated polytomous CMC items, and examined the correlations between correct and incorrect responses and the total score. Further quality inspections were conducted by examining differential item functioning, testing Rasch-homogeneity, investigating the tests' dimensionality as well as local item dependence.

Various criteria indicated a good fit of the items and measurement invariance across various subgroups. However, the amount of not-reached items was rather high, indicating that the test was too long for the allocated testing time. Other types of missing responses were reasonably small.

The test had a high reliability and distinguished well between test takers. However, the test is mainly targeted at low-performing students and did not accurately measure reading competence of high-performing students. As a consequence, ability estimates will be precise for low-performing students but less precise for high performing students.

Some degree of multidimensionality is present for different text functions. In combination with the high amount of missing responses at the end of the test (i.e., there are students with no valid responses to some of the text functions), the estimation of a single reading competence score is challenged. This should be addressed in further studies. Nevertheless, Gehrer et al. (2013) argue that a balanced assessment of reading competence can only be achieved by heterogeneity of text functions and they provide theoretical arguments for a unidimensional measure of reading competence.

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

7. Data in the Scientific Use File

7.1 Naming conventions

The data in the Scientific Use File contain 42 items, of which 31 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. A total of 11 items were scored as polytomous variables (CMC items). MC items are marked with a '0_c' at the end of the variable name, whereas the variable names of CMC items end in 's_c'. For further details on the naming conventions of the variables see Fuß and colleagues (2016). In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category.

7.2 Linking of competence scores

In starting cohort 4, the reading competence tests administered in grades 9 (see Haberkorn et al., 2012) and 12 include different items that were constructed in such a way as to allow for

an accurate measurement of reading competence within each age group. As a consequence, the competence scores derived in the different grades cannot be directly compared; differences in observed scores would reflect differences in competences as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competences across grades, we adopted the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016). Following an anchor-group design, an independent link sample including students from grade 11 that were not part of starting cohort 4 were administered all items from the grade 9 and the grade 12 reading competence tests within a single measurement occasion. These responses were used to link the two tests administered in starting cohort 4 across the two grades.

7.2.1 Samples

In starting cohort 4, a subsample of 5,488 students participated at both measurement occasions, in grade 9 and also in grade 12. Consequently, these respondents were used to link the two tests across both grades (see Fischer et al., 2016.). Moreover, an independent link sample of N = 935 students (448 women) from grade 11 received both tests within a single measurement occasion.

7.2.2 The design of the link study

The test administered in grade 9 included 31 items (see Haberkorn et al., 2012), whereas the test administered in grade 12 included 28 items (see above). Again, two versions of the grade 12 test were used in the link study (easy and difficult). Because preliminary analyses identified severe differential item functioning for two items of the grade 12 test (reg12014s_c and reg12075s_c) between the link sample and the longitudinal main sample, these items were removed from the final linking procedure. A random sample of 464 students received the easy test version and 471 students were administered the difficult version. Moreover, the reading test was administered at different positions in the test battery. A random sample of 476 students received the reading test before working on a mathematics test, whereas the remaining 459 students received the mathematics test before the reading test. No multi-matrix design regarding the selection and order of the items within a test was established. Thus, all test takers were given the reading items in the same order.

7.2.3 Results

To examine whether the two tests administered in the link sample measured a common scale, we compared a one-dimensional model that specified a single latent factor for all items to a two-dimensional model that specified separate latent factors for the two tests. The information criteria slightly favored the two-dimensional model, AIC = 47,058 BIC = 47,508, over the one-dimensional model, AIC = 47,148, BIC = 47,588. However, an examination of the residual correlations for the one-dimensional model using the corrected Q_3 statistic (Yen, 1984) indicated a largely unidimensional scale—the average absolute residual correlation was $M = .04$ ($SD = .03$, $Max = .03$). This indicates that the reading competence tests administered in grades 9 and 12 were essentially unidimensional.

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal subsample from the starting cohort. The differences in item difficulties between the link sample and starting cohort 4 and the respective tests for

measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 13.

Table 13

Differential Item Functioning Analyses between the Starting Cohort and the Link Sample.

	Grade 9				Grade 12			
	Item	**Δσ**	$SE_{\Delta\sigma}$	F	**Item**	**Δσ**	$SE_{\Delta\sigma}$	F
1.	reg90110_c	0.69	0.16	19.76	reg120110_c	-0.21	0.11	3.44
2.	reg90120_c	0.63	0.19	11.63	reg120120_c	-0.43	0.11	16.18
3.	reg90150_c	0.36	0.09	14.93	reg120130_c	-0.45	0.14	11.20
4.	reg9016s_c	0.17	0.08	5.07	reg120150_c	-0.13	0.11	1.40
5.	reg9017s_c	-0.03	0.15	0.03	reg120160_c	-0.34	0.11	9.43
6.	reg90210_c	0.34	0.11	9.75	reg120170_c	0.09	0.12	0.51
7.	reg90220_c	-0.01	0.08	0.03	reg12042s_c	0.40	0.08	25.56
8.	reg90230_c	-0.15	0.12	1.79	reg120430_c	0.11	0.08	1.79
9.	reg90240_c	0.31	0.12	6.56	reg12044s_c	-0.07	0.07	1.08
10.	reg90250_c	-0.11	0.07	2.31	reg120450_c	0.14	0.09	2.57
11.	reg90310_c	-0.23	0.12	3.31	reg120510_c	-0.49	0.15	10.98
12.	reg90320_c	-0.27	0.15	3.07	reg12052s_c	-0.10	0.13	0.56
13.	reg9033s_c	-0.07	0.12	0.34	reg120530_c	-0.64	0.15	18.27
14.	reg90340_c	-0.34	0.13	6.52	reg120540_c	-0.29	0.14	4.57
15.	reg90350_c	-0.11	0.14	0.68	reg12055s_c	-0.29	0.14	4.47
16.	reg90360_c	0.01	0.09	0.03	reg120560_c	-0.24	0.13	3.24
17.	reg90370_c	-0.16	0.08	3.56	reg12021s_c	-0.07	0.08	0.85
18.	reg90410_c	-0.14	0.12	1.36	reg120220_c	0.00	0.07	0.00
19.	reg90420_c	-0.26	0.10	6.80	reg120230_c	-0.15	0.09	2.92
21.	reg90430_c	0.13	0.09	2.06	reg12024s_c	-0.15	0.07	4.32
22.	reg90440_c	0.10	0.10	1.00	reg120250_c	0.10	0.09	1.28
23.	reg90450_c	0.38	0.11	13.17	reg12026s_c	-0.36	0.05	50.63
24.	reg90460_c	0.37	0.08	20.35	reg12071s_c	0.62	0.12	26.80
25.	reg9047s_c	0.12	0.11	1.31	reg120720_c	0.04	0.13	0.09
26.	reg90510_c	-0.21	0.09	6.26	reg120740_c	0.16	0.79	0.04
27.	reg90520_c	-0.17	0.09	3.60	reg120730_c	0.78	0.12	40.25
28.	reg90530_c	-0.19	0.09	4.35	reg120310_c	0.21	0.07	8.05

	Grade 9				Grade 12			
	Item	**Δσ**	$SE_{\Delta\sigma}$	F	**Item**	**Δσ**	$SE_{\Delta\sigma}$	F
29.	reg90540_c	-0.26	0.09	9.06	reg120320_c	0.25	0.08	8.60
30.	reg90550_c	-0.66	0.09	54.58	reg120330_c	0.28	0.08	12.77
31.	reg90560_c	-0.39	0.10	14.83	reg120340_c	0.52	0.10	27.01
32.	reg90570_c	0.13	0.12	1.22	reg120350_c	0.12	0.08	2.68
33.					reg120360_c	0.22	0.08	8.22
34.					reg120610_c	-0.18	0.10	3.63
35.					reg120620_c	-0.08	0.10	0.68
36.					reg120630_c	0.15	0.11	1.99
37.					reg120640_c	0.04	0.09	0.19
38.					reg12065s_c	0.36	0.09	14.92
39.					reg120660_c	0.00	0.10	0.00
40.					reg120670_c	0.11	0.11	1.11

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in grade 9 or 12 and the link sample (positive values indicate easier items in the link sample); $SE_{\Delta\sigma}$ = Pooled standard error; F = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is F_{0154} (1, 6,423) = 68.72. A non-significant test indicates measurement invariance.

Analyses of differential item functioning between the link sample and starting cohort 4 identified neither for grade 9 (difference in logits: *Min* = 0.01, *Max* = 0.69) nor for grade 12 (difference in logits: *Min* = 0.00, *Max* = 0.78) items with significant (α = .05) DIF. Therefore, the reading competence tests administered in the two grades were linked using the "mean/mean" method for the anchor-group design (see Fischer et al., 2016).

The correction term was calculated as $c = 0.045$. This correction term was subsequently added to each difficulty parameter estimated in grade 12 (see Table 7) to derive the linked item parameters. The link error reflecting the uncertainty in the linking process was calculated according to equation 4 in Fischer et al. (2016) as 0.07 and has to be included into the *SE* when statistical tests are used to compare groups concerning their mean change of ability between two linked measurements.

7.3 Reading competence scores

In the SUF manifest reading competence scores are provided in the form of two different WLEs, "reg12_sc1" and "reg12_sc1u", including their respective standard error, "reg12_sc2" and "reg12_sc2u". For "reg12_sc1u", person abilities were estimated using the linked item difficulty parameters. Subsequently, the estimated WLE scores were corrected for differences in the test position. In grade 9, the reading test was always presented first within the test battery, whereas in grade 12 the reading test was either presented as the first or the second test within the test battery (see page 5). To correct for differences in the test position, we

added the main effect related to the test position (see Table 9) to the WLE scores of respondents that received the reading test after working on another test. As a result the WLE scores provided in “reg12_sc1u” can be used for longitudinal comparisons between grades 9 and 12. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in “reg12_sc1” are not linked to the underlying reference scale of grade 9. However, they are corrected for the position of the reading test within the booklet. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions. The ConQuest Syntax for estimating the WLE is provided in Appendix A. For persons who either did not take part in the reading test 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.

Plausible values that allow for an investigation of 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

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ConQuest 4*. Camberwell, Australia: Acer.

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* (pp. 199-214). New York, NY: Springer.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. (2016). *Linking the Data of the Competence Tests* (NEPS Survey Paper No. 1). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2016). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Gehrer, K., & Artelt, C. (2013). Literalität und Bildungslaufbahn: Das Bildungspanel NEPS. In A. Bertschi-Kaufmann, & C. Rosebrock (Eds.), *Literalität erfassen: bildungspolitisch, kulturell, individuell* (pp. 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. *Journal for Educational Research Online, 5*, 50-79.

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

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.

Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement, 50*, 447-468. doi:10.1111/jedm.12028

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 of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5*, 189-216.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft, 14*, 67-86. doi:10.1007/s11618-011-0182-7

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145. doi:10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

Appendix

Appendix A: ConQuest-Syntax for estimating linked WLEs in starting cohort 4

```
Title SC4 G12 READING: Partial Credit Model;

/* load data */
datafile [FILENAME].sav ! filetype=spss,
         responses = reg120110_c reg120120_c reg120130_c reg12014s_c
                     reg120150_c reg120160_c reg120170_c reg12021s_c
                     reg120220_c reg120230_c reg12024s_c reg120250_c
                     reg12026s_c reg120310_c reg120320_c reg120330_c
                     reg120340_c reg120350_c reg120360_c reg12042s_c
                     reg120430_c reg12044s_c reg120450_c reg120510_c
                     reg12052s_c reg120530_c reg120540_c reg12055s_c
                     reg120560_c reg120610_c reg120620_c reg120630_c
                     reg120640_c reg12065s_c reg120660_c reg120670_c
                     reg12071s_c reg120720_c reg120730_c reg120740_c
                     reg12075s_c,
         pid=ID_t >> daten.dat;

/* collapse response categories with less than 200 responses */
recode (0,1,2)         (0,1,1)         ! item (4);  /* reg12014s_c */
recode (0,1,2)         (0,0,1)         ! item (8);  /* reg12021s_c */
recode (0,1,2,3,4)     (0,0,1,2,3)     ! item (11); /* reg12024s_c */
recode (0,1,2,3,4,5,6) (0,0,1,2,3,4,5) ! item (13); /* reg12026s_c */
recode (0,1,2)         (0,0,1)         ! item (20); /* reg12042s_c */
recode (0,1,2,3)       (0,0,1,2)       ! item (28); /* reg12055s_c */
recode (0,1,2,3)       (0,0,0,1)       ! item (37); /* reg12071s_c */
recode (0,1,2,3)       (0,0,1,2)       ! item (41); /* reg12075s_c */

/* scoring */
codes 0,1,2,3,4,5;
score (0,1)         (0,1)               ! items (1-10, 12, 14-19, 20-21,
                                                 23-24, 26-27, 29-33,
                                                 35-40);
score (0,1,2)       (0,0.5,1)           ! items (25, 28, 41);
score (0,1,2,3)     (0,0.5,1,1.5)       ! items (11, 34);
score (0,1,2,3,4)   (0,0.5,1,1.5,2)     ! items (22);
score (0,1,2,3,4,5) (0,0.5,1,1.5,2,2.5) ! items (13);


/* load linked item parameters */
import anchor_parameters << anchor_parameters.txt;

/* model specification */
set constraint=none;
model item + item*step;

/* estimate model */
estimate ! method=gauss, nodes=15, iterations=1000, convergence=0.0001,
stderr=empirical;

/* save results to file */
show ! estimate=latent    >> show.txt;
show cases ! estimate=wle >> wle.txt;
```