NEPS SURVEY PAPERS

Anna Scharl, Luise Fischer, Timo Gnambs, and Theresa Rohm

NEPS TECHNICAL REPORT FOR READING: SCALING RESULTS OF STARTING COHORT 3 FOR GRADE 9

NEPS Survey Paper No. 20
Bamberg, May 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 3 for Grade 9

Anna Scharl, Luise Fischer, Timo Gnambs, and Theresa Rohm
Leibniz Institute for Educational Trajectories, Bamberg

E-mail address of lead author:

anna.scharl@lifbi.de

Bibliographic data:

Scharl, A., Fischer, L., Gnambs, T., & Rohm, T. (2017). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 3 for Grade 9* (NEPS Survey Paper No. 20). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP20:1.0

Acknowledgements:

This report is an extension to NEPS Survey Paper No. 14 (Krannich et al., 2017) that presents the scaling results for reading competence of starting cohort 3 for grade 7. Therefore, various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (Haberkorn, Pohl, Hardt, & Wiegand, 2012; Pohl, Haberkorn, Hardt, & Wiegand, 2012) to facilitate the understanding of the presented results.

We thank Micha Freund for his assistance in scaling the data.

NEPS Technical Report for Reading: Scaling Results of Starting Cohort 3 for Grade 9

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 9 of starting cohort 3 (fifth grade). The reading competence test contained 30 and 32 items in the easy and difficult testlet, respectively. Different response formats represented different cognitive requirements and text functions. The test was administered to 4,590 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 reduced reliability for high-achieving participants as well as a 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 9 of starting cohort 3 (fifth grade). First, the main concepts of the reading competence test are introduced. Then, the reading competence data of starting cohort 3 and the analyses performed on the data to estimate competence scores and to check the quality of the test 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, specific aspects of the reading competence test will be pointed out 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 three types of response formats: simple multiple choice (MC) items, complex multiple choice (CMC) items and matching items (MA). 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. MA items required the test taker to match a number of responses to a given set of statements. 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 46 items. In order to evaluate the quality of these items, extensive preliminary analyses were conducted. These preliminary analyses identified substantial differential item functioning for the items reg90240_c (for the variables number of books at home, difficulty of the test, and school type) and reg90650_c (for the variable sex). A more detailed description of the grouping variables can be found in section 5.3.4. Therefore, these items were 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 44 items.

3. Data

3.1 The Design of the Study

The study assessed different competence domains including, among others, reading competence, declarative meta-cognition, and listening comprehension. For each participant, the reading test was administered as the first 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	7		7
Instruction text		4	
Advertising text	5		7
Commenting text		7	
Literary text		7	
Total number of items	12	18	14

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 30 or 32 items that represented the five text functions (see Table 1) and three cognitive requirements (see Table 2) as described above. Three texts with 18 items were identical in both test versions (see Table 1), whereas 12 or 14 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 two and four. Participants were assigned either to the easy or the difficult test based on their estimated reading competence in the previous assessment (Krannich et al., 2017): Participants with an ability estimate below the sample's median ability received the easy test, whereas participants with a reading competence above the sample's median received the difficult test.

Table 2. Number of Items by Cognitive Requirements and Difficulty of the Test

Cognitive requirements	**Easy test**	**Difficult test**
1 Finding information	11	10
2 Drawing text-related conclusions	11	13
3 Reflecting and assessing	8	9
Total number of items	30	32

Table 3. Number of Items by Different Response Formats and Difficulty of the Test

Response format	**Easy test**	**Difficult test**
Simple multiple choice items	27	29
Complex multiple choice items	2	2
Matching	1	1
Total number of items	30	32

3.2 Sample

The panel study aimed at retesting all students that were initially included in the starting cohort 3 for fifth grade (see Pohl et al., 2012). Thus, a total of 4,590[1] individuals received the reading competence test. For twelve 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 excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 4,578 individuals. Of these, 2,159 participants received the easy test, whereas 2,419 participants were administered the difficult test version. A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

[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.

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 14 difficult items were missing by design, whereas 12 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 as missing if at least one subtask contained a missing response. When just one kind of missing response occurred, the item was coded according to the corresponding missing response. When the subtasks contained different kinds of missing responses, the item was labeled as a not-determinable missing response.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions, handling of different response formats). They also need to be accounted for in the estimation of item and person parameters. 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. Categories of polytomous variables with less than *N* = 200 responses were collapsed in the analyses in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items. For three of the six CMC items categories were collapsed (see Appendix A). One of those (reg9083s_c) was treated as a dichotomous variable in further analyses.

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). For item reg90840_c two response categories were scored as correct after distractor analyses indicated that persons, especially with higher reading competence, perceived a second ambiguous distractor as correct. Three dichotomous items (reg90640_c, reg90840_c, reg90860_c) were scored 0.5 for the correct response due to problematic item fit.

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 three 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 rest item 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 gender, school type, the number of books at home (as a proxy for socioeconomic status), migration background (see Pohl & Carstensen, 2012, for a description of these variables), and type of testlet (easy/difficult). 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. Moreover, measurement invariance analyses were also conducted for the two test versions including either the easy or difficult items.

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 (the results are depicted in section 5.3.6). 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 (aQ_3) that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) absolute values of aQ_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) and in R (R Core Team, 2016) using the package TAM (Kiefer, Robitzsch, & Wu, 2016).

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). Overall, there were very few invalid responses. Between 96% and 98% of the respondents did not have any invalid response at all; less than two percent had more than one invalid response. There was no difference in the amount of invalid responses between the easy and difficult test.

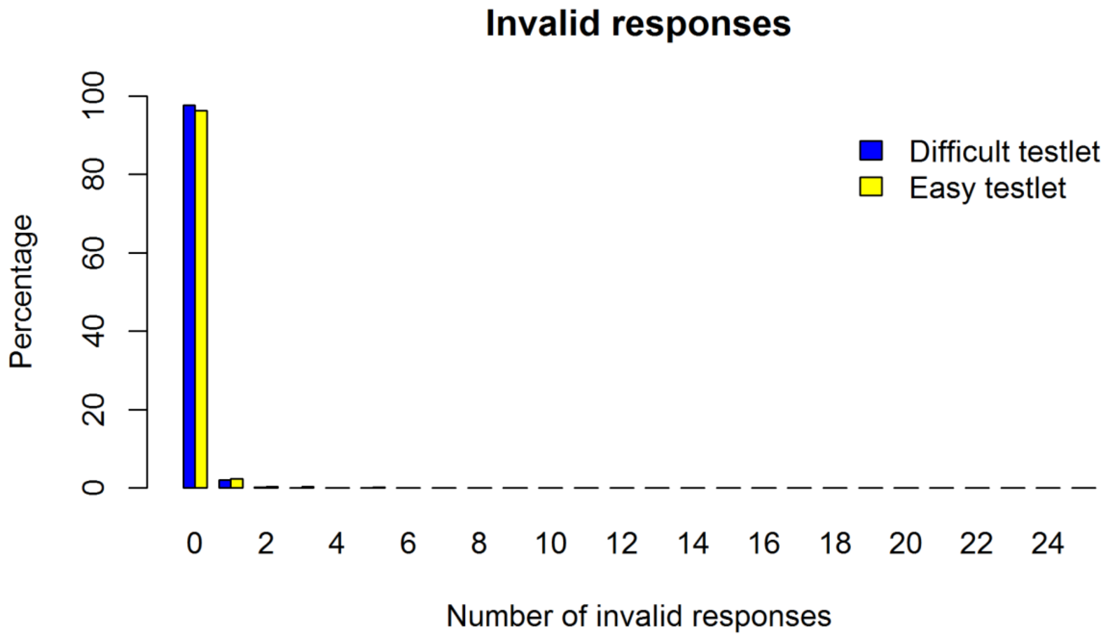

Figure 1. Number of invalid responses by test version

Missing responses also occurred when respondents omitted items. As illustrated in Figure 2 most respondents, 75% to 80%, did not skip any item and less than eight percent omitted more than three items. There was no difference in the amount of omitted items between the two test versions.

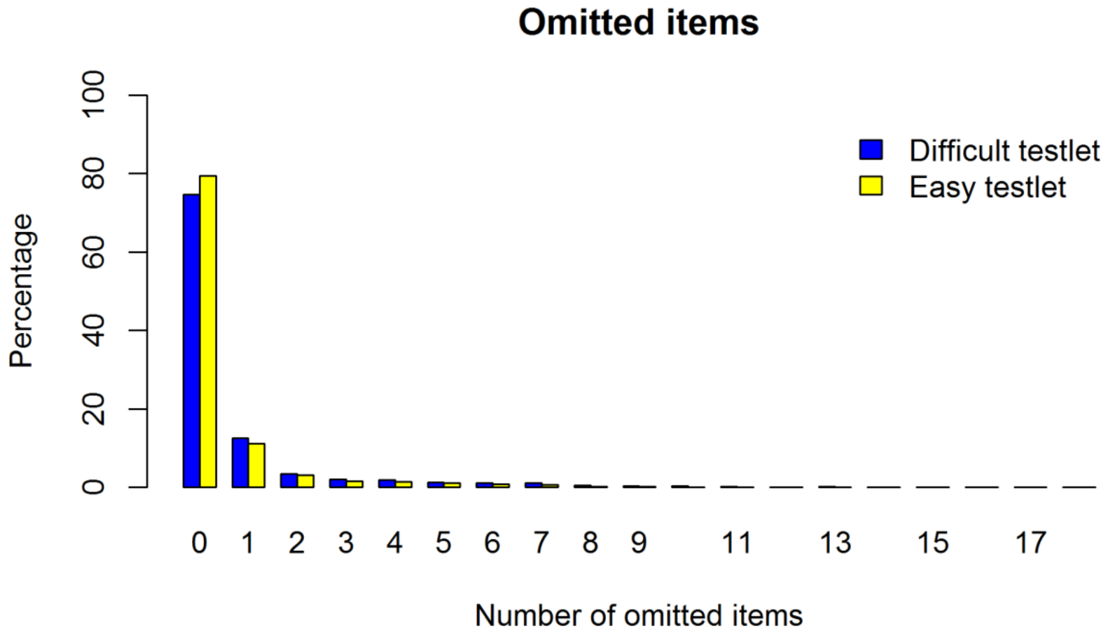

Figure 2. Number of omitted items by test version

Another source of missing responses was 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 of the easy and difficult testlet were unable to finish the test within the allocated time limit (Figure 3). Between 53% and 68% of the respondents finished the entire test. Between 22% and 37% did not reach the last of the five texts. Time constraints seemed to be a bigger issue in the difficult testlet.

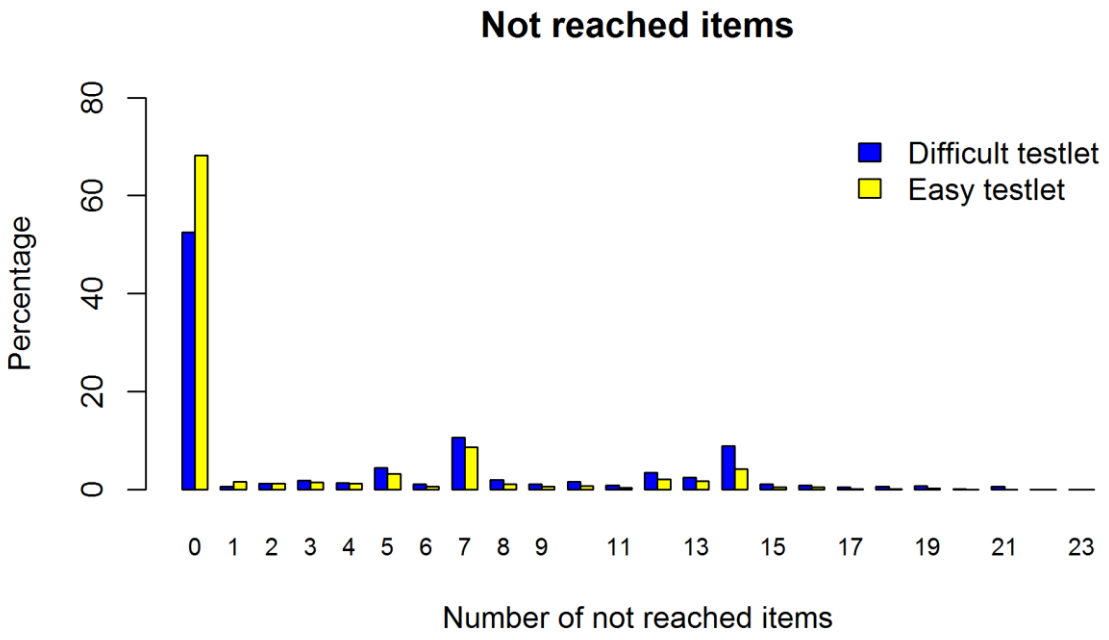

Figure 3. Number of not-reached items by test version

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 three or four (i.e., the number of CMC items according to the testlet). However, there were no 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 4. On average, the respondents showed between *M* = 3.25 (*SD* = 4.84) and *M* = 5.39(*SD* = 5.94) missing responses in the different experimental conditions (i.e., the easy and difficult testlet, respectively). About 38% to 52% of the respondents had no missing response at all and about 32% to 49% of the participants had four or more missing responses.

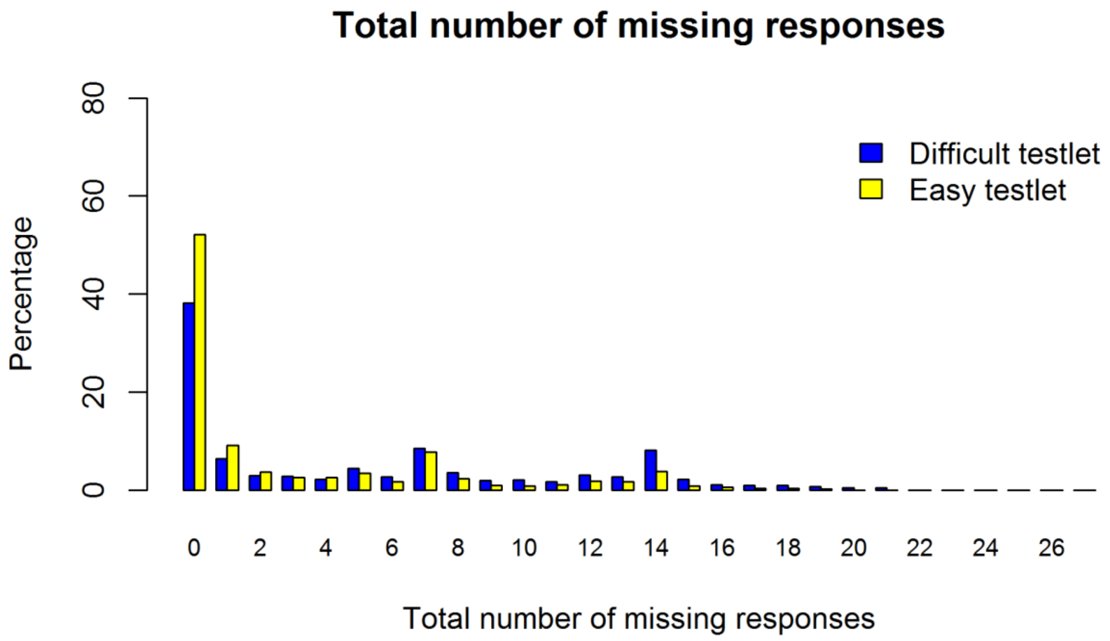

Figure 4. Total number of missing responses by test version

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

Table 4. Percentage of Missing Values by Test Difficulty

	Easy Testlet					*Difficult Testlet*				
Item	**Position**	***N***	**NR**	**OM**	**NV**	**Position**	***N***	**NR**	**OM**	**NV**
reg90610_c	1	2,134	0.00	0.51	0.65					
reg90620_c	2	2,133	0.00	1.07	0.14					
reg9063s_c	3	2,088	0.00	3.20	0.09					
reg90640_c	4	2,122	0.00	0.79	0.93					
reg90660_c	6	2,102	0.00	2.36	0.28					
reg90670_c	7	2,119	0.00	1.62	0.23					
reg90680_c	8	2,107	0.00	2.18	0.23					
reg90810_c						1	2,385	0.00	1.36	0.04
reg90820_c						2	2,393	0.00	1.07	0.00
reg9083s_c						3	2,358	0.00	2.52	0.00
reg90840_c						4	2,311	0.00	4.13	0.33
reg90850_c						5	2,365	0.00	1.98	0.25
reg90860_c						6	2,363	0.00	2.19	0.12
reg90870_c						7	2,387	0.00	1.24	0.08
reg90210_sc3g9_c	9	2,145	0.09	0.05	0.51	8	2,413	0.00	0.12	0.12
reg90220_sc3g9_c	10	2,142	0.09	0.19	0.51	9	2,413	0.00	0.04	0.21
reg90230_sc3g9_c	11	2,138	0.09	0.60	0.28	10	2,414	0.04	0.08	0.08
reg90250_sc3g9_c	13	2,139	0.09	0.42	0.42	12	2,408	0.04	0.25	0.17
reg90710_c	14	2,117	0.42	1.25	0.28					
reg90720_c	15	2,106	0.56	1.25	0.65					
reg90730_c	16	2,093	0.74	2.18	0.14					
reg9074s_c	17	2,013	1.34	3.80	1.11					
reg90750_c	18	2,081	1.95	1.53	0.14					
reg9091s_c						13	2,317	0.66	3.47	0.08
reg90920_c						14	2,338	0.87	2.32	0.17
reg90930_c						15	2,231	1.69	5.99	0.08
reg90940_c						16	2,260	2.40	3.89	0.29
reg90950_c						17	2,244	2.98	4.18	0.08
reg90960_c						18	2,246	3.93	3.18	0.04
reg9097s_c						19	2,137	5.13	5.33	0.99
reg90410_sc3g9_c	19	1,992	6.21	1.25	0.28	20	2,053	14.14	0.91	0.08
reg90420_sc3g9_c	20	1,951	7.97	1.53	0.14	21	1,984	16.62	1.32	0.04
reg90430_sc3g9_c	21	1,863	10.10	3.43	0.19	22	1,863	20.13	2.81	0.04
reg90440_sc3g9_c	22	1,876	10.56	2.32	0.23	23	1,873	21.08	1.45	0.04
reg90450_sc3g9_c	23	1,875	11.35	1.53	0.28	24	1,840	22.74	1.16	0.04
reg90460_sc3g9_c	24	1,857	12.04	1.81	0.14	25	1,800	23.89	1.65	0.04
reg9047s_sc3g9_c	25	1,829	13.20	1.85	0.23	26	1,748	25.96	1.74	0.04
reg90510_sc3g9_c	26	1,664	21.91	0.88	0.14	27	1,493	36.67	1.57	0.04
reg90520_sc3g9_c	27	1,645	22.60	1.02	0.19	28	1,474	37.78	1.12	0.17

	Easy Testlet					*Difficult Testlet*				
Item	**Position**	***N***	**NR**	**OM**	**NV**	**Position**	***N***	**NR**	**OM**	**NV**
reg90530_sc3g9_c	28	1,529	25.89	3.15	0.14	29	1,322	42.25	3.10	0.00
reg90540_sc3g9_c	29	1,510	27.23	2.69	0.14	30	1,281	43.65	3.35	0.04
reg90550_sc3g9_c	30	1,451	28.76	3.84	0.19	31	1,198	45.56	4.92	0.00
reg90560_sc3g9_c	31	1,453	30.06	2.59	0.05	32	1,178	46.80	4.46	0.04
reg90570_sc3g9_c	32	1,470	31.77	0.00	0.14	33	1,270	47.50	0.00	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 items on position 11 (difficult testlet) and 5 and 12 (easy testlet) were excluded from the analyses due to problematic DIF (see section 2).

5.1.2 Missing responses per item

Table 4 provides 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 5.99%. There were two items with an omission rate exceeding 5% (reg90930_c, reg9097s_c in the difficult testlet). For the difficult test omission rates correlated with the item difficulties at about .51; for the easy test the respective correlation was distinctly smaller with .32. Generally, participants were inclined to omit more difficult items. In contrast, the percentage of invalid responses per item (columns 6 and 11 in Table 4) was rather low with the maximum rate being 0.99%.

With an item's progressing position in the test, the amount of persons that did not reach the item (columns 4 and 9 in Table 4) rose up to a considerable amount of 32% to 48% for the two test versions (see Figure 5).

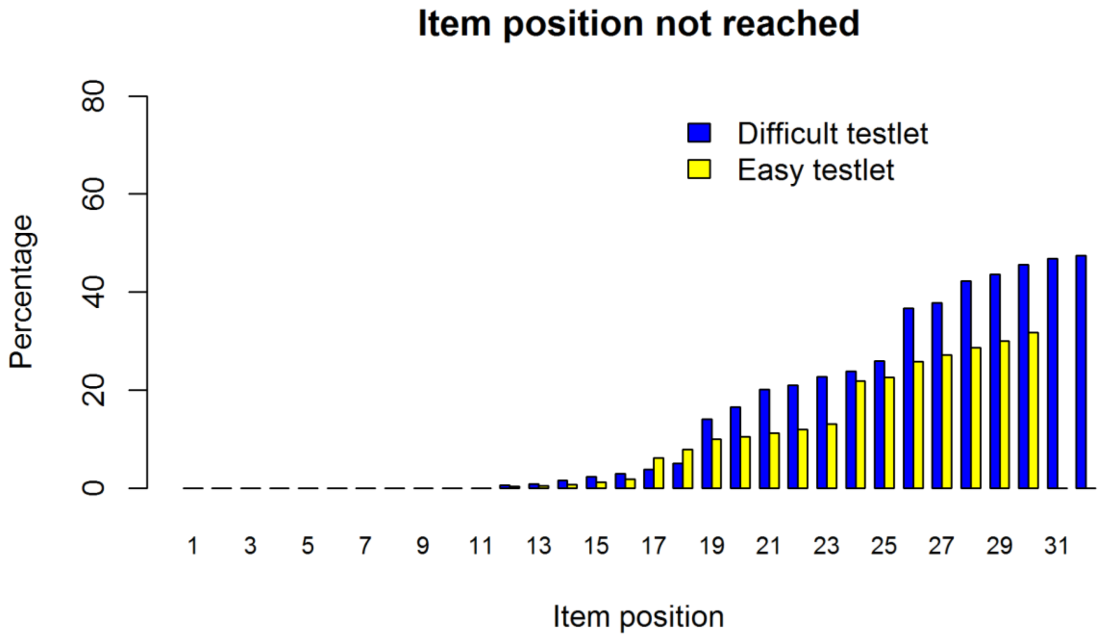

Figure 5. Item position not reached by test version

5.2 Parameter Estimates

5.2.1 Item parameters

The second column in Table 5 presents the percentage of correct responses in relation to all valid responses for each item. Because there was 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 26% and 88% with an average of 55% (*SD* = 16) correct responses.

Table 5. Item Parameters

	Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	Item-Rest correlation	Discr.	aQ₃
1.	reg90610_c	43	-0.121	0.051	1.04	2.1	0.30	0.84	0.03
2.	reg90620_c	28	0.634	0.055	1.05	2.2	0.25	0.72	0.02
3.	reg9063s_c	n.a.	0.164	0.069	1.05	1.9	0.18	0.64	0.02
4.	reg90640_c	48	-0.260	0.091	1.02	2.1	0.14	0.68	0.03
6.	reg90660_c	26	0.744	0.056	1.02	0.6	0.28	0.89	0.02
7.	reg90670_c	47	-0.319	0.051	1.10	5.9	0.22	0.59	0.03
8.	reg90680_c	31	0.475	0.054	1.10	4.4	0.18	0.53	0.03
9.	reg90810_c	57	0.105	0.048	1.06	4.1	0.23	0.63	0.02
10.	reg90820_c	50	0.429	0.048	1.08	5.2	0.21	0.59	0.03
11.	reg9083s_c	70	-0.554	0.051	1.02	0.9	0.28	0.88	0.02
12.	reg90840_c	37	1.517	0.090	1.02	1.6	0.10	0.59	0.02
13.	reg90850_c	57	0.080	0.048	0.99	-0.6	0.33	0.99	0.02
14.	reg90860_c	51	0.312	0.086	1.02	1.8	0.13	0.76	0.03
15.	reg90870_c	60	-0.057	0.048	1.06	3.6	0.23	0.66	0.03
16.	reg90210_sc3g9_c	88	-2.356	0.051	1.00	0.1	0.24	0.95	0.03
17.	reg90220_sc3g9_c	66	-0.796	0.037	1.03	2.0	0.32	0.87	0.02
18.	reg90230_sc3g9_c	82	-1.808	0.044	1.07	2.8	0.22	0.69	0.02
20.	reg90250_sc3g9_c	45	0.227	0.036	1.11	9.2	0.24	0.56	0.02
21.	reg90710_c	42	-0.114	0.051	1.03	1.7	0.30	0.86	0.02
22.	reg90720_c	56	-0.735	0.051	0.97	-1.8	0.38	1.19	0.03
23.	reg90730_c	28	0.643	0.056	0.97	-1.2	0.36	1.14	0.02
24.	reg9074s_c	n.a.	-0.280	0.048	0.96	-1.7	0.45	1.19	0.04
25.	reg90750_c	48	-0.374	0.051	0.98	-1.0	0.37	1.12	0.03
26.	reg9091s_c	n.a.	-0.186	0.058	1.01	0.5	0.26	0.91	0.03

	Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	Item-Rest correlation	Discr.	aQ$_3$
27.	reg90920_c	53	0.266	0.048	0.98	-1.4	0.34	1.04	0.03
28.	reg90930_c	51	0.341	0.049	0.99	-1.0	0.34	1.03	0.03
29.	reg90940_c	64	-0.247	0.050	1.01	0.4	0.31	0.93	0.03
30.	reg90950_c	36	1.095	0.051	0.94	-3.4	0.39	1.41	0.03
31.	reg90960_c	68	-0.434	0.052	0.99	-0.4	0.33	1.02	0.02
32.	reg9097s_c	n.a.	-0.330	0.050	0.93	-3.1	0.45	1.45	0.04
33.	reg90410_sc3g9_c	87	-2.280	0.052	0.96	-1.2	0.33	1.36	0.02
34.	reg90420_sc3g9_c	72	-1.177	0.042	0.90	-5.7	0.47	1.65	0.03
35.	reg90430_sc3g9_c	66	-0.861	0.041	0.91	-5.7	0.48	1.53	0.04
36.	reg90440_sc3g9_c	78	-1.557	0.045	0.92	-3.6	0.42	1.43	0.03
37.	reg90450_sc3g9_c	82	-1.857	0.049	0.91	-3.3	0.41	1.64	0.04
38.	reg90460_sc3g9_c	63	-0.720	0.041	1.00	0.1	0.38	1.02	0.04
39.	reg9047s_sc3g9_c	n.a.	-1.924	0.052	0.91	-3.6	0.42	1.73	0.03
40.	reg90510_sc3g9_c	49	-0.083	0.043	0.91	-6.2	0.49	1.42	0.03
41.	reg90520_sc3g9_c	59	-0.561	0.043	0.97	-2.2	0.43	1.18	0.03
42.	reg90530_sc3g9_c	54	-0.353	0.045	1.06	3.9	0.33	0.81	0.03
43.	reg90540_sc3g9_c	40	0.316	0.046	0.90	-6.1	0.50	1.47	0.03
44.	reg90550_sc3g9_c	49	-0.141	0.046	1.08	4.9	0.31	0.72	0.02
45.	reg90560_sc3g9_c	58	-0.621	0.047	0.97	-1.9	0.44	1.21	0.03
46.	reg90570_sc3g9_c	74	-1.436	0.050	1.02	1.0	0.33	0.95	0.03

Note. Difficulty = Item difficulty / location parameter, *SE* = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, Discr. = Discrimination parameter of a generalized partial credit model, aQ$_3$ =Adjusted average absolute residual correlation for item (Yen, 1984, 1993; Kiefer et al., 2016).

Items 5 and 19 were excluded from the analyses due to problematic DIF (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 item-rest correlation; 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 5. The step parameters for polytomous variables are depicted in Table 6. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties (or location parameters for polytomous variables) ranged from -2.356 (item reg90210_c) to 1.517 (item reg90840_c) with an average difficulty of -0.345. 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 5) were rather small (all *SE*s ≤ 0.09).

Table 6. Step Parameters (with Standard Errors) for Polytomous Items

Item	Step 1	Step 2	Step 3	Step 4
reg9063s_c	-0.801 (0.045)	0.801		
reg9074s_c	-0.782 (0.061)	0.021 (0.066)	0.205 (0.072)	0.566
reg9091s_c	-0.124 (0.045)	0.124		
reg9097s_c	-0.163 (0.061)	0.177 (0.068)	-0.013	
reg9047s_sc3g9_c	1.362 (0.064)	-1.362		

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 1.02, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = .809, WLE reliability = .787) was good. The mean of the item distribution was about 0.35 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 abilities in medium- and low-ability regions will be measured relatively precise, whereas higher ability estimates will have larger standard errors of measurement.

Scale (in logits)	**Person ability**	**Item difficulty**
3		
	X	
	X	
	X	
	X	
2	XX	
	XX	
	XXX	
	XXX	11
	XXXX	
	XXXXX	
	XXXXX	28
1	XXXXX	
	XXXXXXX	
	XXXXXXX	5
	XXXXXXX	2 21
	XXXXXXXXX	7 9
	XXXXXXXXX	13 26 41
	XXXXXXXXXX	3 18 25
	XXXXXXXXX	8 12
0	XXXXXXXXX	1 14 19 38
	XXXXXXXX	4 24 27 42
	XXXXXXXXXX	6 22 23 30 40
	XXXXXXX	29
	XXXXXXXX	10 39 43
	XXXXXXX	16 20 36
	XXXXXXX	33
-1	XXXXXX	
	XXXXX	32
	XXXXX	
	XXXX	44
	XXXX	34
	XXX	
	XX	17 35
-2	X	37
	X	
	X	31
	X	15
-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 25.6 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 5).

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 55 items. The probability of a correct response ranged from 11% to 88% across all items (*Mdn* = 61%). Thus, the number of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.55[2] to 1.16 (*M* = 0.98, *SD* = 0.12), the respective *t*-value from -20.1 to 9.3 (*M* = -0.91, *SD* = 5.38), 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 5). Values of the WMNSQ ranged from 0.90 (items reg90420_c and reg90540_c) to 1.11 (reg90250_c). Only two items exhibited a *t*-value of the WMNSQ smaller than -6 (reg90510_c, reg90540_c) and one exceeded a value of 8 (reg90250_c). Thus, there was no indication of severe item over- or underfit. Point-biserial correlations between the item scores and the total rest scores ranged from .10 (item reg90840_c) to .50 (item reg90540_sc3g9_c) and had a mean of .32. 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 -.43 to .17 with a mean of -.16. 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 (see Pohl & Carstensen, 2012, for a description of these variables) and the two test versions (easy vs. difficult). The differences between the estimated item difficulties in the various groups are summarized in Table 7. For example, the column "Male vs. female" reports the differences in item difficulties between men and women; a positive value would indicate that the item was more difficult for males, whereas a negative value would highlight a lower difficulty for males as opposed to females. In contrast, the main effect is to be interpreted on a group level. As such, a positive value indicates that males, on average, had a higher ability as females; whereas a negative value would highlight a lower ability, on

[2] Please note, that the extreme values of WMNSQ and the respective t-value resulted from scoring certain dichotomous items with 0.5.

average, for males 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 8).

Table 8. Differential Item Functioning

Item	Sex	Books	Migration	School	Booklet
	male vs. female	< 100 vs. ≥ 100	without vs. with	no sec. vs. sec.	easy vs. difficult
reg90610_c	-0.702 (-0.700)	-0.018 (-0.019)	-0.058 (-0.058)	0.094 (0.107)	n.a. (n.a.)
reg90620_c	-0.154 (-0.154)	0.010 (0.011)	0.228 (0.227)	-0.132 (-0.150)	n.a. (n.a.)
reg9063s_c	0.072 (0.072)	-0.022 (-0.023)	-0.088 (-0.088)	-0.158 (-0.180)	n.a. (n.a.)
reg90640_c	0.066 (0.066)	0.028 (0.029)	-0.266 (-0.265)	-0.240 (-0.273)	n.a. (n.a.)
reg90660_c	-0.342 (-0.341)	0.192 (0.202)	0.054 (0.054)	0.198 (0.225)	n.a. (n.a.)
reg90670_c	-0.386 (-0.385)	-0.030 (-0.032)	0.136 (0.135)	-0.420 (-0.478)	n.a. (n.a.)
reg90680_c	-0.646 (-0.644)	-0.282 (-0.297)	0.020 (0.020)	-0.316 (-0.360)	n.a. (n.a.)
reg90810_c	0.072 (0.072)	-0.124 (-0.131)	0.056 (0.056)	-0.156 (-0.178)	n.a. (n.a.)
reg90820_c	0.212 (0.211)	-0.208 (-0.219)	-0.008 (-0.008)	-0.012 (-0.014)	n.a. (n.a.)
reg9083s_c	-0.246 (-0.245)	0.050 (0.053)	0.032 (0.032)	-0.018 (-0.02)	n.a. (n.a.)
reg90840_c	-0.140 (-0.140)	0.114 (0.120)	0.208 (0.207)	0.004 (0.005)	n.a. (n.a.)
reg90850_c	-0.138 (-0.138)	0.032 (0.034)	0.068 (0.068)	0.100 (0.114)	n.a. (n.a.)
reg90860_c	-0.672 (-0.670)	-0.306 (-0.322)	0.102 (0.102)	-0.750 (-0.854)	n.a. (n.a.)
reg90870_c	-0.120 (-0.120)	0.146 (0.154)	0.024 (0.024)	-0.412 (-0.469)	n.a. (n.a.)
reg90210_sc3g9_c	0.144 (0.144)	0.102 (0.107)	-0.288 (-0.287)	-0.154 (-0.175)	-0.080 (-0.088)
reg90220_sc3g9_c	-0.002	-0.002	-0.134	-0.114	-0.072

Item	Sex	Books	Migration	School	Booklet
	male vs. female	< 100 vs. ≥ 100	without vs. with	no sec. vs. sec.	easy vs. difficult
	(-0.002)	(-0.002)	(-0.133)	(-0.130)	(-0.080)
reg90230_sc3g9_c	0.886	-0.074	0.212	-0.192	-0.522
	(0.884)	(-0.078)	(0.211)	(-0.219)	(-0.577)
reg90250_sc3g9_c	-0.154	-0.282	0.162	-0.316	-0.522
	(-0.154)	(-0.297)	(0.161)	(-0.360)	(-0.577)
reg90710_c	0.304	-0.198	-0.026	-0.250	n.a.
	(0.303)	(-0.209)	(-0.026)	(-0.285)	(n.a.)
reg90720_c	0.332	-0.292	0.094	-0.202	n.a.
	(0.331)	(-0.308)	(0.094)	(-0.230)	(n.a.)
reg90730_c	-0.684	0.022	0.090	0.184	n.a.
	(-0.682)	(0.023)	(0.090)	(0.210)	(n.a.)
reg9074s_c	0.416	-0.214	-0.064	-0.034	n.a.
	(0.415)	(-0.225)	(-0.064)	(-0.039)	(n.a.)
reg90750_c	0.254	-0.100	-0.022	-0.126	n.a.
	(0.253)	(-0.105)	(-0.022)	(-0.143)	(n.a.)
reg9091s_c	-0.480	-0.118	0.038	-0.188	n.a.
	(-0.479)	(-0.124)	(0.038)	(-0.214)	(n.a.)
reg90920_c	-0.120	0.000	-0.178	-0.010	n.a.
	(-0.120)	(0.000)	(-0.177)	(-0.011)	(n.a.)
reg90930_c	-0.218	-0.116	-0.172	-0.012	n.a.
	(-0.217)	(-0.122)	(-0.171)	(-0.014)	(n.a.)
reg90940_c	-0.488	0.074	-0.186	-0.096	n.a.
	(-0.487)	(0.078)	(-0.185)	(-0.109)	(n.a.)
reg90950_c	0.026	-0.044	0.068	0.152	n.a.
	(0.026)	(-0.046)	(0.068)	(0.173)	(n.a.)
reg90960_c	0.092	0.048	-0.136	-0.020	n.a.
	(0.092)	(0.051)	(-0.135)	(-0.023)	(n.a.)
reg9097s_c	-0.002	-0.030	0.108	0.308	n.a.
	(-0.002)	(-0.032)	(0.108)	(0.351)	(n.a.)
reg90410_sc3g9_c	0.180	-0.088	-0.136	0.358	0.324
	(0.180)	(-0.093)	(-0.135)	(0.408)	(0.358)
reg90420_sc3g9_c	0.108	0.088	0.140	0.232	0.270
	(0.108)	(0.093)	(0.139)	(0.264)	(0.299)
reg90430_sc3g9_c	0.204	0.082	0.048	0.178	0.320
	(0.203)	(0.086)	(0.048)	(0.203)	(0.354)

Item	Sex	Books	Migration	School	Booklet
	male vs. female	< 100 vs. ≥ 100	without vs. with	no sec. vs. sec.	easy vs. difficult
reg90440_sc3g9_c	0.498 (0.497)	-0.070 (-0.074)	-0.044 (-0.044)	0.134 (0.153)	0.182 (0.201)
reg90450_sc3g9_c	0.124 (0.124)	-0.074 (-0.078)	0.110 (0.110)	-0.108 (-0.123)	-0.074 (-0.082)
reg90460_sc3g9_c	-0.012 (-0.012)	-0.096 (-0.101)	0.150 (0.149)	-0.188 (-0.214)	-0.14 (-0.155)
reg9047s_sc3g9_c	0.168 (0.168)	0.308 (0.324)	-0.078 (-0.078)	0.350 (0.399)	0.324 (0.358)
reg90510_sc3g9_c	-0.044 (-0.044)	0.350 (0.369)	-0.200 (-0.199)	0.452 (0.515)	0.438 (0.484)
reg90520_sc3g9_c	-0.140 (-0.140)	0.232 (0.244)	-0.064 (-0.064)	0.316 (0.360)	0.222 (0.245)
reg90530_sc3g9_c	0.012 (0.012)	0.042 (0.044)	-0.064 (-0.064)	0.038 (0.043)	-0.162 (-0.179)
reg90540_sc3g9_c	0.126 (0.126)	0.380 (0.400)	-0.132 (-0.131)	0.510 (0.581)	0.424 (0.469)
reg90550_sc3g9_c	0.034 (0.034)	0.118 (0.124)	0.112 (0.112)	-0.182 (-0.207)	-0.410 (-0.453)
reg90560_sc3g9_c	-0.036 (-0.036)	0.140 (0.147)	0.158 (0.157)	0.182 (0.207)	0.052 (0.057)
reg90570_sc3g9_c	0.228 (0.227)	0.270 (0.284)	-0.162 (-0.161)	0.376 (0.428)	-0.068 (-0.075)
Main effect (DIF model)	-0.228 (-0.227)	-0.692 (-0.729)	0.220 (0.219)	-1.022 (-1.164)	-1.338 (-1.479)
Main effect (main effect model)	-0.220 (-0.220)	-0.692 (-0.727)	0.222 (0.221)	-1.024 (-1.167)	-1.322 (-1.466)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. Sec. = Secondary school (German: „Gymnasium“).
None of the absolute standardized differences was significantly, p < .05, greater than 0.25 (see Fischer, Rohm, Gnambs, & Carstensen, 2016).

Sex: The sample included 2,339 (51%) males and 2,230 (49%) females. Nine 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 = -0.228 logits, Cohen's d = -0.227). Five items (reg90610_c, reg90680_c, reg90860_c, reg90230_sc3g9_c, and reg90730_c) showed DIF greater than |0.6| logits. An overall test for DIF (see Table 8) 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, as does the Bayesian information criterion (BIC; Schwarz, 1978) that takes the number of estimated parameters into account and, thus, guards against overparameterization of models. However, the comparison of the resulting main effects of the two models suggested no pronounced DIF with regard to sex.

Books: The number of books at home was used as a proxy for socioeconomic status. There were 1,746 (38%) test takers with 0 to 100 books at home, 2,714 (59%) test takers with more than 100 books at home, and 115 (3%) test takers without a valid response. There was a considerable average difference between the two groups. Participants with 100 or less books at home performed, on average, 0.692 logits (Cohen's *d* = -0.729) lower in reading than participants with more than 100 books. There was no considerable DIF comparing participants with many or fewer books (highest DIF = 0.38 for item reg90540_c). The overall test for DIF using the AIC favored the model estimating DIF, whereas the BIC indicated a better fit for the more parsimonious model including only the main effect (Table 9). Thus, overall, there was no pronounced DIF with regard to books.

Table 9. Comparisons of Models with and without DIF

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	main effect	4,569	151,476.41	54	151,584.41	151,931.47
	DIF	4,569	151,005.33	98	151,201.33	151,831.18
Books	main effect	4,460	147,628.02	54	147,736.02	148,081.77
	DIF	4,460	147,494.90	98	147,690.90	148,318.39
Migration	main effect	4,305	143,218.10	54	143,326.10	143,669.95
	DIF	4,231	140,633.54	98	140,829.54	141,451.86
School	main effect	4,578	150,725.16	54	150,833.16	151,180.33
	DIF	4,399	144,635.50	98	144,831.50	145,457.64
Difficulty	main effect	4,576	68,617.68	21	68,659.68	68,794.68
	DIF	4,576	68,384.19	39	68,462.19	68,712.91

Migration background: There were 2,956 participants (65%) with no migration background, 1,349 subjects (29%) with a migration background, and 273 individuals (6%) 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 = 0.22 logits, Cohen's *d* = 0.219). There was no noteworthy item DIF due to migration background; differences in estimated difficulties did not exceed 0.6 logits. Nevertheless, the overall test for DIF using the AIC and BIC favored the model that included item-level DIF.

School type: Overall, 2,112 subjects (46%) who took the reading test attended secondary school (German: "Gymnasium") whereas 2,466 (54%) were enrolled in other school types. Subjects in secondary schools showed a higher reading ability on average (1.022 logits, Cohen’s *d* = 1.164) than subjects in other school types. There was one item (reg90860_c) that showed DIF greater than 0.6 logits. The overall model test indicated a slightly better fit for the more complex DIF model. Again, the comparison of the resulting main effects of the two models suggested no pronounced DIF with regard to school type.

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 18 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 2,159 (47%) persons received the easy test and 2,419 (53%) persons received the difficult test. As expected, subjects who were administered the easy test scored on average -1.338 logits (Cohen’s *d* = -1.479) 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 0.522 logits (items reg90230_c and reg90250_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 5), ranging from 0.53 (item reg90680_c) to 1.73 (item reg9047s_c). The average discrimination parameter fell at 1.01. Model fit indices suggested a slightly better model fit of the 2PL model (AIC = 150,893.18, BIC = 151,510.37) as compared to the 1PL model (AIC =151,892.89, BIC = 152,233.63). 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 and 2013, for a 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 with the R (R Core Team, 2016) package TAM (Kiefer et al., 2016) using Quasi Monte Carlo method.

The estimated variances and correlations between the three dimensions representing the different cognitive requirements are reported in Table 10. The correlations among the three dimensions were rather high and fell between .95 and .96. As such, they did not deviate significantly from a perfect correlation (see Carstensen, 2013). Moreover, according to model fit indices, the unidimensional model fitted the data slightly better (AIC = 151,890.2, BIC = 152,230.9, number of parameters = 53) than the three-dimensional model (AIC = 151,892.5, BIC = 152,265.4, number of parameters = 58). These results indicate that the three cognitive requirements measure a common construct.

Table 10. Results of Three-Dimensional Scaling

	Dim 1	Dim 2	Dim 3
Finding information in the text (Dim 1) (15 items)	(1.14)		
Drawing text-related conclusions (Dim 2) (17 items)	.95	(0.93)	
Reflecting and assessing (Dim 3) (12 items)	.96	.95	(1.20)

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 11. The correlations between the dimensions varied between *r* = .71 and *r* = .87. The smallest correlation was found between Dimension 1 ("Information") and Dimension 4 ("Literary"). Dimension 1 ("Information") and Dimension 3 ("Advertising") 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 = 150,936.6, BIC = 151,367.3, number of parameters = 67) fitted the data better than the unidimensional model (AIC = 151,890.2, BIC = 152,230.9, number of parameters = 53). 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 11, 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 11), 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 3 were slightly lower (between 0.71 and 0.87) 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 heterogeneity of text functions (Gehrer et al., 2013).

However, for the unidimensional model the average absolute residual correlations as indicated by the *aQ*$_3$ statistic (see Table 5) were quite low (*M* = .027, *SD* = .024)—the largest individual residual correlation was .21—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 11. Results of Five-Dimensional Scaling

	Dim 1	**Dim 2**	**Dim 3**	**Dim 4**	**Dim 5**
Information (Dim 1) (14 items)	(1.00)				
Instruction (Dim 2) (4 items)	.76	(0.72)			
Advertising (Dim 3) (12 items)	.87	.84	(0.91)		
Literary (Dim 4) (7 items)	.71	.82	.80	(2.67)	
Commenting (Dim 5) (7 items)	.82	.79	.82	.80	(1.52)

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 3 for grade 9 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 medium- and low-performing students and did not accurately measure reading competence of high-performing students. As a consequence, ability estimates will be precise for medium- and 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 reading 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 46 items of which 44 items were included in the previous analyses. Thirty-nine of the remaining (and the two excluded) items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. A total of five items was scored as polytomous variables (CMC and MA items). MC items are marked with a '0_c' at the end of the variable name, whereas the variable names of the CMC and MA items end in 's_c'[3]. In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category. Items containing the suffix _sc3g9_ have originally been administered in starting cohort 4, grade 9.

7.2 Linking of competence scores

In starting cohort 3, the reading competence tests administered in grades 7 (see Krannich et al., 2017) and 9 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, & Carstensen (2016). Following an anchor-group design, an independent link sample including students from grade 9 that were not part of starting cohort 3 were administered all items from the grade 7 and the grade 9 reading competence tests within a single measurement occasion. These responses were used to link the two tests administered in starting cohort 3 across the two grades.

7.2.1 Samples

In starting cohort 3, a subsample of 4,416 students participated at both measurement occasions, in grade 7 and also in grade 9. 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* = 533 students (247 women, 29 participants did not indicate their sex) from grade 9 received both tests within a single measurement occasion.

7.2.2 The design of the link study

While the grade 7 reading test (including 40 items, see Krannich et al., 2017) was administered to the link sample in its original form, the grade 9 test only included the

[3] Note that the item reg9083s_c is a CMC item, but was treated as a dichotomous item in the analyses due to collapsing of item categories.

common items between the easy and difficult test (18 items, see above). In order to avoid differences in test lengths between the link study and the main study, 14 dummy items were added. Two versions of the grade 7 test were administered in the link study (easy and difficult). A random sample of 269 students received the easy test version and 264 students were administered the difficult version. The two reading tests from grades 7 and 9 were randomly administered first or second to the participants in the link sample. 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 Correcting for a change in study design

The design of the link study was identical to the test design of the main study in grade 7. Thus, the reading test of grade 7 was either administered in first or second position. In contrast, the study design changed for the main study in grade 9; here, the reading test was always administered in the first position. In order to correct for this change in test position for some respondents (in grade 7 about half of the participants received the reading test in second position, whereas in grade 9 they received the test in first position), we substituted the position effect for the grade 9 test identified in the link sample: participants of the link sample receiving the grade 9 test in first position had, on average, a higher ability of 0.348 logits. Therefore, we added 0.174 logits (i.e., half of the position effect) to the mean item difficulty of the grade 9 test to calculate the linking correction term (see Fischer et al., 2016).

7.2.4 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. While the AIC favored the two-dimensional model (AIC = 21,318) over the one-dimensional model (AIC = 21,492), the BIC favored the more parsimonious one-dimensional model (BIC = 21,847) over the two-dimensional model, (BIC = 21,917). In addition, 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* = -.01 (*SD* = .06, *Max* = .27). This indicates that the reading competence tests administered in grades 7 and 9 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 3 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 12.

Table 12. Differential Item Functioning Analyses between the Starting Cohort and the Link Sample.

	Grade 7				Grade 9			
	Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	*F*	**Item**	$\Delta\sigma$	$SE_{\Delta\sigma}$	*F*
1.	reg70110_c	-0.518	0.173	8.9	Dummy item 1			
2.	reg70120_c	-0.443	0.278	2.5	Dummy item 2			
3.	reg7013s_c	0.806	0.447	3.3	Dummy item 3			
4.	reg70140_c	0.918	0.728	1.6	Dummy item 4			
5.	reg7015s_c	-0.431	0.318	1.8	Dummy item 5			
6.	reg7016s_c	0.514	0.286	3.2	Dummy item 6			
7.	reg70610_c	-0.903	0.267	11.4	Dummy item 7			
8.	reg70620_c	-1.225	0.164	55.6	reg90210_sc3g9_c	-0.470	0.152	9.5
9.	reg7063s_c	-0.261	0.300	0.8	reg90220_sc3g9_c	-0.418	0.122	11.7
10.	reg70640_c	-0.483	0.162	8.9	reg90230_sc3g9_c	0.403	0.160	6.3
11.	reg70650_c	-0.428	0.162	7.0	*			
12.	reg7066s_c	-0.234	0.222	1.1	reg90250_sc3g9_c	-0.271	0.121	5.0
13.	reg70210_c	0.388	0.290	1.8	Dummy item 8			
14.	reg70220_c	-0.805	0.155	26.9	Dummy item 9			
15.	reg7023s_c	0.998	0.235	18.0	Dummy item 10			
16.	reg7024s_c	1.109	0.166	44.9	Dummy item 11			
17.	reg70250_c	-0.377	0.140	7.3	Dummy item 12			
18.	reg7026s_c	-0.379	0.207	3.3	Dummy item 13			
19.	reg70310_c	0.027	0.241	0.0	Dummy item 14			
20.	reg70320_c	-0.313	0.159	3.9	reg90410_sc3g9_c	0.212	0.181	1.4
21.	reg7033s_c	0.612	0.186	10.8	reg90420_sc3g9_c	0.212	0.143	2.2
22.	reg70340_c	-0.023	0.172	0.0	reg90430_sc3g9_c	0.273	0.139	3.8
23.	reg70350_c	0.278	0.219	1.6	reg90440_sc3g9_c	0.183	0.155	1.4
24.	reg70360_c	0.154	0.167	0.9	reg90450_sc3g9_c	0.066	0.163	0.2
25.	reg70410_c	-0.139	0.215	0.4	reg90460_sc3g9_c	-0.077	0.134	0.3
26.	reg70420_c	-0.272	0.179	2.3	reg9047s_sc3g9_c	0.087	0.174	0.2
27.	reg70430_c	0.238	0.252	0.9	reg90510_sc3g9_c	-0.135	0.133	1.0
28.	reg70440_c	0.213	0.212	1.0	reg90520_sc3g9_c	0.282	0.141	4.0
29.	reg7045s_c	0.016	0.155	0.0	reg90530_sc3g9_c	-0.235	0.139	2.9
30.	reg70460_c	0.201	0.134	2.3	reg90540_sc3g9_c	0.313	0.140	5.0
31.	reg7051s_c	-0.029	0.303	0.0	reg90550_sc3g9_c	-0.364	0.142	6.6

	Grade 7				Grade 9			
	Item	$\Delta\sigma$	$SE_{\Delta\sigma}$	F	**Item**	$\Delta\sigma$	$SE_{\Delta\sigma}$	F
32.	reg70520_c	0.200	0.251	0.6	reg90560_sc3g9_c	-0.156	0.148	1.1
33.	reg7053s_c	1.088	0.285	14.6	reg90570_sc3g9_c	0.097	0.169	0.3
34.	reg7055s_c	0.614	0.211	8.5				
35.	reg70560_c	0.242	0.194	1.6				
36.	reg7071s_c	0.241	0.320	0.6				
37.	reg70720_c	-0.317	0.188	2.8				
38.	reg70730_c	-1.058	0.196	29.2				
39.	reg70740_c	-0.615	0.214	8.2				
40.	reg7075s_c	0.380	0.214	3.2				

Note. $\Delta\sigma$ = Difference in item difficulty parameters between the longitudinal subsample in grade 7/grade 9 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}(2,4947)=109,2. A non-significant test indicates measurement invariance.

*In order to scale the link study and main study identically, the item reg90240_sc3g9_c was excluded from analysis.

Analyses of differential item functioning between the link sample and starting cohort 3 did not identify items with significant (α = .05) DIF for grade 7 (difference in logits: Min = 0.02, Max = 1.23), nor for grade 9 (difference in logits: Min = 0.07, Max = 0.47). 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 for grade 7 and 9 was calculated as *c* = 0.579. Added to the correction term for grades 5 and 7 (see Krannich et al., 2017), a total correction term of 1.247 was derived. This correction term was subsequently added to each difficulty parameter estimated in grade 9 (see Table 5) 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 0.11 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, "reg9_sc1" and "reg9_sc1u", including their respective standard error, "reg9_sc2" and "reg9_sc2u". For "reg9_sc1u", person abilities were estimated using the linked item difficulty parameters. Subsequently, the estimated WLE scores were corrected for the change in test design: In grade 9, the reading test was always presented first within the test battery, whereas in grade 7 the reading test was either presented as the first or the second test within the test battery (see 7.2.2). As a result, the WLE scores provided in "reg9_sc1u" can be used for longitudinal comparisons between grades 7 and 9. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in “reg9_sc1” are not linked to the underlying reference scale of grade 7. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions. Because there was no change in test position for the reading test in grade 9, a correction was not necessary for “reg9_sc1”. 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 was 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.

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* (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.

Kiefer T., Robitzsch, A. & Wu, M. (2016). TAM: Test Analysis Modules. *[Computer software manual]. Retrieved from https://CRAN.R-project.org/package=TAM (R package version 1.995-0)*.

Krannich, M., Jost, O., Rohm, T., Koller, I., Carstensen, C. H., Fischer, L., & Gnambs, T. (2017, January). *NEPS Technical Report for reading: Scaling results of Starting Cohort 3 for grade 7* (NEPS Survey Papers No. 14). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

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

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

Pohl, S. (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.

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

R Core Team (2016). R: A language and environment for statistical computing (Version 3.2.4) [Software]. Retrieved from https://www.R-project.org/.

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.

Warm, T. A., (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450, doi:10.1007/BF02294627

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 3

```
Title A98 G9 READING: Partial Credit Model;

/* load data */
datafile [FILENAME].sav ! filetype=spss,
          responses =  reg90610_c reg90620_c reg9063s_c reg90640_c reg90660_c
                       reg90670_c reg90680_c
                       reg90810_c reg90820_c reg9083s_c reg90840_c reg90850_c
                       reg90860_c reg90870_c
                       reg90210_c reg90220_c reg90230_c reg90250_c
                       reg90710_c reg90720_c reg90730_c reg9074s_c reg90750_c
                       reg9091s_c reg90920_c reg90930_c reg90940_c reg90950_c
                       reg90960_c reg9097s_c
                       reg90410_c reg90420_c reg90430_c reg90440_c reg90450_c
                       reg90460_c reg9047s_
                       reg90510_c reg90520_c reg90530_c reg90540_c reg90550_c
                       reg90560_c reg90570_c ,
         keeps=valid,
          pid=ID_t >> daten.dat;

keepcases 1    ! valid; /* remove cases with <= 3 valid responses (1) */

/* collapse response categories with less than 200 responses */
recode (0,1,2,3)       (0,0,1,2)       ! item (3);  /* reg9063s_c */
recode (0,1,2)         (0,0,1)         ! item (10);  /* reg9083s_c */
recode (0,1,2,3)       (0,0,1,2)       ! item (24);  /* reg9091s_c */

/* scoring */
codes 0,1,2,3,4;

score (0,1)         (0,1)               ! items (1,2,5-10,12,14-21,23,25-29,31-
36,38-44);
score (0,1)         (0,0.5)             ! items (4,11,13);
score (0,1,2)       (0,0.5,1)           ! items (3,24,37);
score (0,1,2,3)     (0,0.5,1,1.5)       ! items (30);
score (0,1,2,3,4)   (0,0.5,1,1.5,2)     ! items (22);

set constraint=none, warnings=no;

/* model specification */
model item + item*step;

/* load linked item parameters */
import anchor_parameters << anker.prm;

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

/* save results to file */
show ! estimate=latent >> show.txt;
show  parameters ! tables=3, estimates=wle, filetype=excel >> reliability.xls;
export parameters      >> parameters.txt;
itanal                 >> itemanalysis.txt;
show residuals ! estimate=wle, filetype=spss >> residuals.sav;
show cases ! estimate=wle, filetype=spss >> wle.sav;

quit;
```

Appendix B: import-file of the anchor parameters for linking the grade 9 test to the grade 5/7 scale

```
1      1.12512             /* item reg90610_c  */
2      1.88076             /* item reg90620_c  */
3      1.41073             /* item reg9063s_c  */
4      0.98653             /* item reg90640_c  */
5      1.99014             /* item reg90660_c  */
6      0.92808             /* item reg90670_c  */
7      1.72162             /* item reg90680_c  */
8      1.35156             /* item reg90810_c  */
9      1.67582             /* item reg90820_c  */
10     0.69224             /* item reg9083s_c  */
11     2.76342             /* item reg90840_c  */
12     1.32638             /* item reg90850_c  */
13     1.55902             /* item reg90860_c  */
14     1.18962             /* item reg90870_c  */
15     -1.10954            /* item reg90210_sc3g9_c   */
16     0.45096             /* item reg90220_sc3g9_c   */
17     -0.56111            /* item reg90230_sc3g9_c   */
18     1.47402             /* item reg90250_sc3g9_c   */
19     1.13280             /* item reg90710_c  */
20     0.51139             /* item reg90720_c  */
21     1.88959             /* item reg90730_c  */
22     0.96640             /* item reg9074s_c  */
23     0.87295             /* item reg90750_c  */
24     1.06073             /* item reg9091s_c  */
25     1.51243             /* item reg90920_c  */
26     1.58742             /* item reg90930_c  */
27     0.99918             /* item reg90940_c  */
28     2.34168             /* item reg90950_c  */
29     0.81260             /* item reg90960_c  */
30     0.91639             /* item reg9097s_c  */
31     -1.03294            /* item reg90410_sc3g9_c   */
32     0.07003             /* item reg90420_sc3g9_c   */
33     0.38584             /* item reg90430_sc3g9_c   */
34     -0.30998            /* item reg90440_sc3g9_c   */
35     -0.61062            /* item reg90450_sc3g9_c   */
36     0.52637             /* item reg90460_sc3g9_c   */
37     -0.67786            /* item reg9047s_sc3g9_c   */
38     1.16390             /* item reg90510_sc3g9_c   */
39     0.68560             /* item reg90520_sc3g9_c   */
40     0.89386             /* item reg90530_sc3g9_c   */
41     1.56228             /* item reg90540_sc3g9_c   */
42     1.10610             /* item reg90550_sc3g9_c   */
43     0.62540             /* item reg90560_sc3g9_c   */
44     -0.18967            /* item reg90570_sc3g9_c   */
45     -0.80078            /* item reg9063s_c step 1 */
46     -0.78208            /* item reg9074s_c step 1 */
47     0.02113             /* item reg9074s_c step 2 */
48     0.20509             /* item reg9074s_c step 3 */
49     -0.12403            /* item reg9091s_c step 1 */
50     -0.16315            /* item reg9097s_c step 1 */
51     0.17665             /* item reg9097s_c step 2 */
52     1.36213             /* item reg9047s_sc3g9_c step 1 */
```