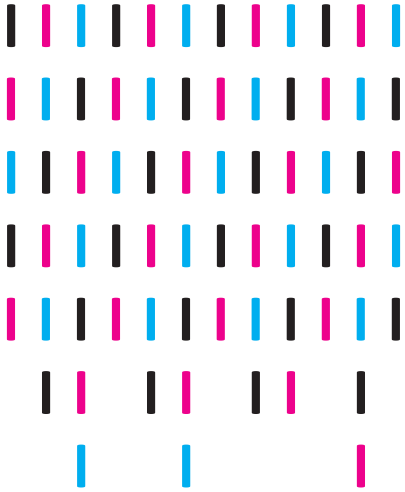

NEPS SURVEY PAPERS

Theresa Rohm, Katharina Krohmer and Timo Gnambs

NEPS TECHNICAL REPORT FOR READING: SCALING RESULTS OF STARTING COHORT 2 FOR GRADE 4

NEPS Survey Paper No. 30
Bamberg, November 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 2 for Grade 4

Theresa Rohm, Katharina Krohmer, & Timo Gnambs

Leibniz Institute for Educational Trajectories, Bamberg

E-mail address of lead author:

theresa.rohm@lifbi.de

Bibliographic data:
Rohm, T., Krohmer, K., & Gnambs, T. (2017). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 2 for Grade 4* (NEPS Survey Paper No. 30). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP30:1.0

Acknowledgements:
This report is an extension to NEPS working paper 15 (Pohl, Haberkorn, Hardt, & Wiegand, 2012) that presents the scaling results for reading competence of starting cohort 3 for grade 5. Therefore, various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from this as well as previous working papers (Pohl et al., 2012; Hardt, K., Pohl, S., Haberkorn, K., & Wiegand, E., 2013) to facilitate the understanding of the presented results.

NEPS Technical Report for Reading: Scaling Results of Starting Cohort 2 for Grade 4

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 4 of starting cohort 2 (Kindergarten). The reading competence test contained 33 items with different response formats representing different cognitive requirements and text functions. The test was administered to 6,710 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 most 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. There was also some evidence of multidimensionality related to different 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 4 of starting cohort 2 (Kindergarten). First, the main concepts of the reading competence test are introduced. Then, the reading competence data of starting cohort 2 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 assignment of the items to the different text types and cognitive requirements can be found in Appendix B.

The reading competence test included three types of response formats: simple multiple choice (MC) items, complex multiple choice (CMC) items, and matching (MA) 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. MA items require the test taker to match a number of responses to a given set of statements. MA items are usually used to assign headings to paragraphs of a text. Examples of the different response formats are given in Pohl and Carstensen (2012) and Gehrer, Zimmermann, Artelt and Weinert (2012).

The competence test for reading that was administered in the present study included 33 items. In order to evaluate the quality of these items extensive preliminary analyses were conducted. Our analyses identified a poor fit for two items (reg5045s_sc2g4_c and reg50560_sc2g4_c). 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 31 items.

3. Data

3.1 The Design of the Study

Two domains were assessed in this study – namely, reading and mathematical competence. All students received a booklet that first contained the mathematics test followed by the reading test.

The panel study aimed at retesting all students that were initially included in the starting cohort 2 for Kindergarten (see Haberkorn, Pohl, Hardt, & Wiegand, 2012). After Kindergarten, the participants of the starting cohort spread out to different elementary schools. Therefore, the participants of the starting cohort were divided into two subsamples that exhibited different assessment settings: Students who attended an elementary school together with other participants from starting cohort 2 were tested together at school in a group setting. In contrast, students who attended an elementary school without other participants were tracked and, subsequently, individually tested at home (for details regarding the data collection process see the respective field report for wave 6). Thus, the context of test administration differed between the two groups.

Table 1

Number of Items for the Different Text Types in reading test grade 4

Text types	**Frequency**
Information text	7
Instruction text	6
Advertising text	7
Commenting text	5
Literary text	6
Total number of items	31

The characteristics of the analyzed 31 items are depicted in Tables 1 to 3. Table 1 reflects the distribution of the text functions, Table 2 contains the distribution of cognitive requirements, and Table 3 informs about the distribution of the response format. The number of subtasks within CMC and MA items varied between two and seven.

Table 2

Number of Items by Cognitive Requirements

Cognitive requirements	**Frequency**
Finding information	9
Drawing text-related conclusions	12
Reflecting and assessing	10
Total number of items	31

Table 3

Number of Items by Different Response Formats

Response format	**Frequency**
Simple multiple choice items	25
Complex multiple choice items	3
Matching	3
Total number of items	31

3.2 Sample

A total of 6,710[1] individuals received the reading competence test. For nine 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 6,701 individuals. The number of participants within each assessment setting

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.

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 Assessment Setting

Assessment setting	Frequency
At school	5,272
At home	1,429
Total	6,701

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, and finally, d) 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. As CMC and MA items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A CMC or MA item was coded as missing if at least one subtask contained a missing response. When just one kind of missing response occurred, the item was coded according to the corresponding missing response. When the subtasks contained different kinds of missing responses, the item 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 and MA items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC or MA item, indicating the number of correctly responded subtasks within that item. If at least one of the subtasks contained a missing response, the CMC or MA item was scored as missing. Categories of polytomous variables with less than $N = 200$ responses were collapsed 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 five of the six CMC and MA items categories were collapsed (see Appendix A). As a consequence, the values of the polytomously scored CMC and MA items in the SUF do not necessarily contain the number of correctly solved subtasks but should rather be interpreted as (partial) credit scores.

To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items were scored dichotomously as 0 for an incorrect and 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats). A special case is item reg5026s_sc2g4_c. The item consisted of seven subtasks that showed extreme local stochastic dependences. In accordance with theoretical considerations, the item was scored as 1 only if all subtasks were solved correctly; otherwise it was scored as zero.

Reading competences were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989). 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 and MA 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 and MA 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 and MA 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, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Moreover, in light of the quasi-experimental design measurement invariance analyses were also conducted for the administration setting (school or home environment). As students attending elementary school regularly obtain a recommendation in grade four for secondary school (grammar school), test fairness was furthermore investigated for the variable school recommendation. Due to a high amount of missing values for this variable when students were assessed within the home environment (83.3 percent missing), the analyses for school recommendation refer only to students tested within the school context and that have valid information for this variable. 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. administration setting). Overall, there were very few invalid responses. About 92% of the respondents did not have any invalid response at all; less than one percent had more than one invalid response. There was no difference in the amount of invalid responses between the assessment settings.

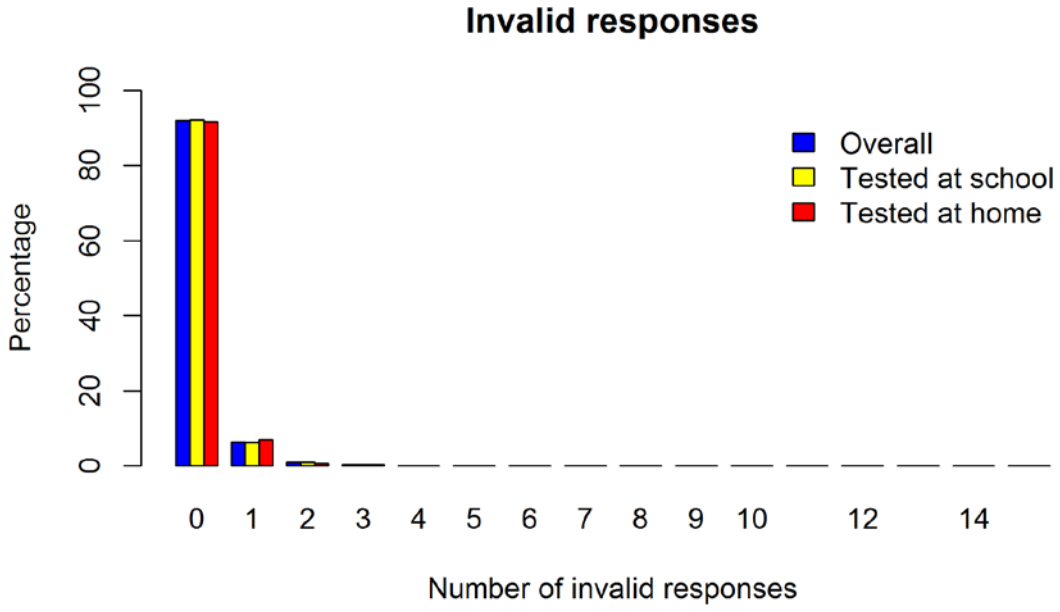

Figure 1. Number of invalid responses by assessment setting

Missing responses may also occur when respondents omit items. As illustrated in Figure 2 almost half of the respondents, 49% to 55%, did not skip any item and less than four percent omitted more than three items. There were no pronounced differences in the amount of omitted items between the assessment settings.

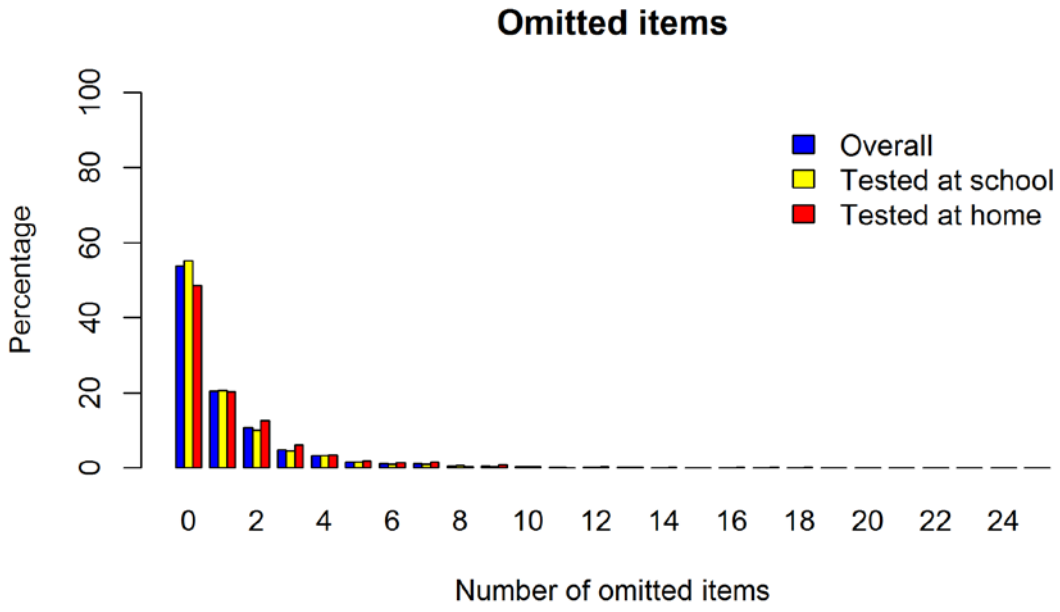

Figure 2. Number of omitted items by assessment setting

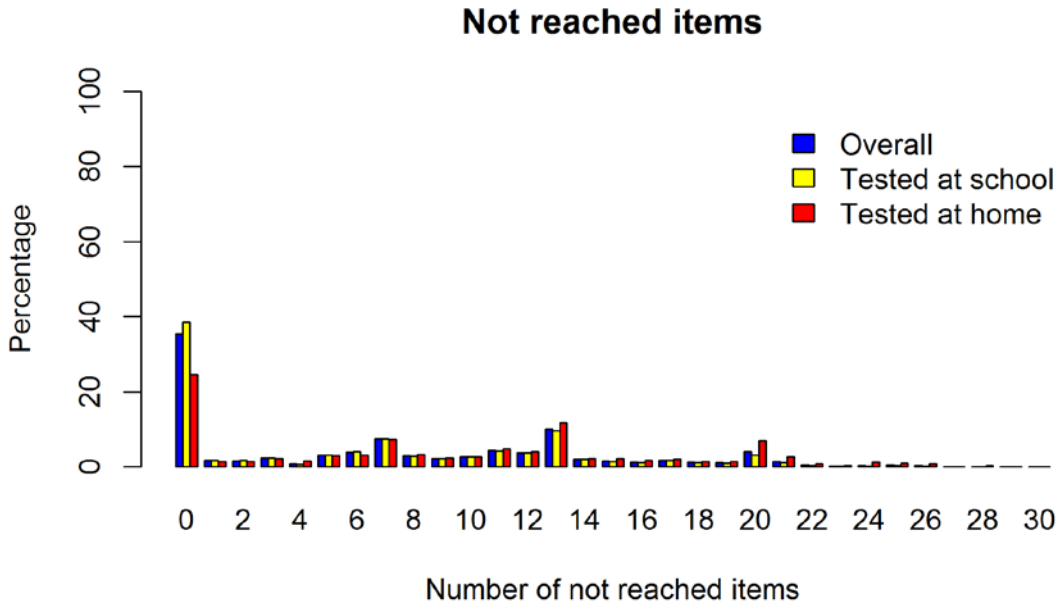

Figure 3. Number of not-reached items by assessment setting

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 25% (tested at home) and 38% (tested at school) of the respondents finished the entire test. Overall, 51% did not reach the last of the five texts; in particular, respondents tested at home did not reach the last text (62%).

The aggregated polytomous variables were coded as not-determinable missing response when the subtasks of CMC and MA items contained different kinds of missing responses. Because not-determinable missing responses only occur in CMC and MA items, the maximum number of not-determinable missing responses was three. However, only a rather small number of not-determinable missing responses occurred. Most respondents (96%) did not have any not-determinable missing response. There was no difference in the amount of not-determinable items between the assessment settings.

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* = 7.96 (*SD* = 6.95) and *M* = 11.17 (*SD* = 7.82) missing responses in the different assessment settings. About 12% to 19% of the respondents had no missing response at all and about 63% to 77% of the participants had four or more missing responses. Particularly, respondents receiving the test at home showed more missing responses because they did not reach the last of the five texts.

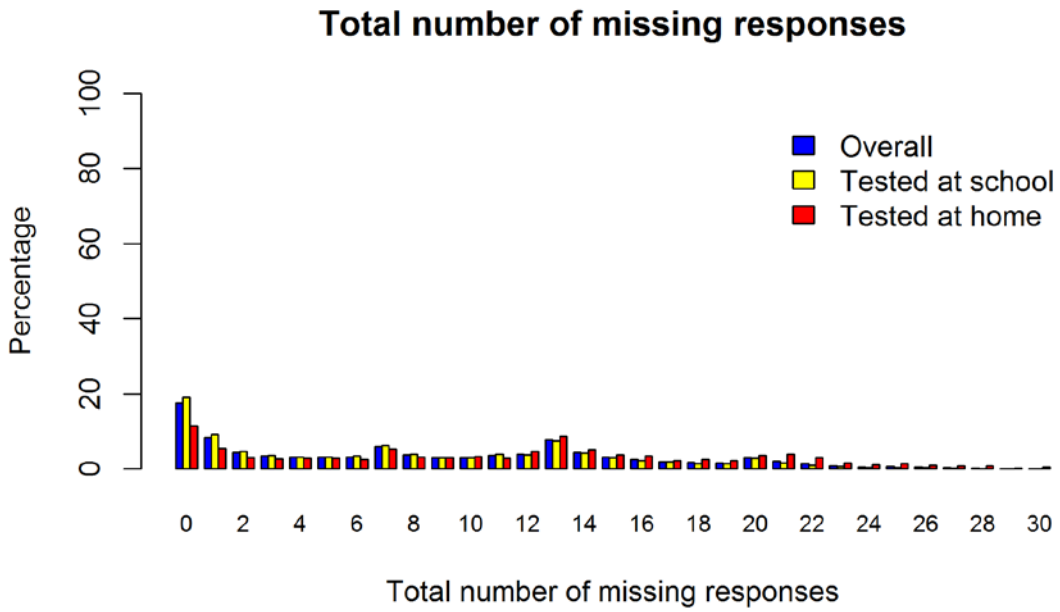

Figure 4. Total number of missing responses by assessment setting

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 rather large and had the greatest impact on the total number of missing responses.

Table 5

Percentage of Missing Values by Assessment Setting

		At school				At home			
Item	**Position**	***N***	**NR**	**OM**	**NV**	***N***	**NR**	**OM**	**NV**
reg50110_sc2g4_c	1	5,217	0.00	0.93	0.11	1,391	0.00	2.38	0.28
reg5012s_sc2g4_c	2	4,772	0.00	8.59	0.74	1,224	0.00	13.86	0.42
reg50130_sc2g4_c	3	5,198	0.00	1.33	0.08	1,392	0.00	2.52	0.07
reg50140_sc2g4_c	4	5,028	0.00	4.48	0.15	1,310	0.07	7.98	0.28
reg50150_sc2g4_c	5	5,096	0.00	2.88	0.46	1,343	0.14	5.53	0.35
reg5016s_sc2g4_c	6	4,541	0.02	10.26	2.50	1,125	0.49	16.38	2.66
reg50170_sc2g4_c	7	5,123	0.13	2.09	0.61	1,354	0.63	3.92	0.70
reg50210_sc2g4_c	8	5,176	0.38	1.31	0.13	1,374	1.54	2.10	0.21
reg50220_sc2g4_c	9	4,921	0.72	5.86	0.08	1,244	2.52	10.22	0.21
reg50230_sc2g4_c	10	5,088	0.95	2.48	0.06	1,316	3.78	3.92	0.21
reg50240_sc2g4_c	11	5,039	1.18	2.75	0.49	1,311	4.20	3.29	0.77
reg50250_sc2g4_c	12	4,964	1.54	4.17	0.13	1,287	5.11	4.76	0.07
reg5026s_sc2g4_c	13	4,381	2.69	10.36	2.05	1,061	7.77	13.72	2.45
reg50310_sc2g4_c	14	4,776	5.92	3.41	0.08	1,151	14.77	4.69	0.00
reg50320_sc2g4_c	15	4,746	6.96	2.92	0.09	1,134	16.31	4.34	0.00
reg50330_sc2g4_c	16	4,745	8.19	1.65	0.15	1,136	17.84	2.59	0.07
reg50340_sc2g4_c	17	4,540	10.00	3.77	0.11	1,076	19.87	4.76	0.07
reg50350_sc2g4_c	18	4,579	11.12	1.84	0.19	1,083	21.69	2.38	0.14
reg50360_sc2g4_c	19	4,424	12.58	3.43	0.08	1,011	23.93	5.25	0.07
reg50370_sc2g4_c	20	4,264	14.64	4.42	0.06	987	26.24	4.48	0.21
reg50410_sc2g4_c	21	3,760	24.28	4.27	0.13	799	38.00	5.88	0.21
reg5042s_sc2g4_c	22	3,548	28.07	4.42	0.17	730	41.99	6.44	0.21
reg50430_sc2g4_c	23	3,315	32.32	4.51	0.28	668	46.82	6.30	0.14
reg50440_sc2g4_c	24	3,170	34.98	4.78	0.11	636	49.48	5.88	0.14
reg50460_sc2g4_c	26	2,903	40.08	4.55	0.30	571	55.14	4.76	0.14
reg50510_sc2g4_c	27	2,684	47.63	1.40	0.06	532	62.49	0.14	0.14
reg5052s_sc2g4_c	28	2,375	51.69	2.96	0.13	451	65.64	2.24	0.28
reg50530_sc2g4_c	29	2,266	54.89	2.05	0.08	426	68.58	1.54	0.07
reg50540_sc2g4_c	30	2,255	55.61	1.56	0.06	414	70.19	0.77	0.07
reg5055s_sc2g4_c	31	1,976	58.06	3.24	0.80	345	72.50	2.59	0.49
reg50570_sc2g4_c	33	2,022	61.55	0.00	0.09	349	75.44	0.00	0.14

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 positions 25 and 32 were excluded from the analyses due to an unsatisfactory item fit (see section 2).

5.1.2 Missing responses per item

Table 5 provides information on the occurrence of different kinds of missing responses per item. Overall, the omission rates varied across items between 0.00% and 16.38 % (item reg5016s_sc2g4_c in the home environment test setting). Omission rates correlated with the item difficulties at about .29 in the school context and about .19 at home. Generally, participants were inclined to omit more difficult items. In contrast, the percentage of invalid responses per item (columns 6 and 10 in Table 5) was rather low with the maximum rate being 2.66 % (item reg5016s_sc2g4_c in the home environment test setting).

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) rose up to a considerable amount of 62% to 75% for the two assessment settings. Particularly, at home the last items were not reached by many respondents (see Figure 5).

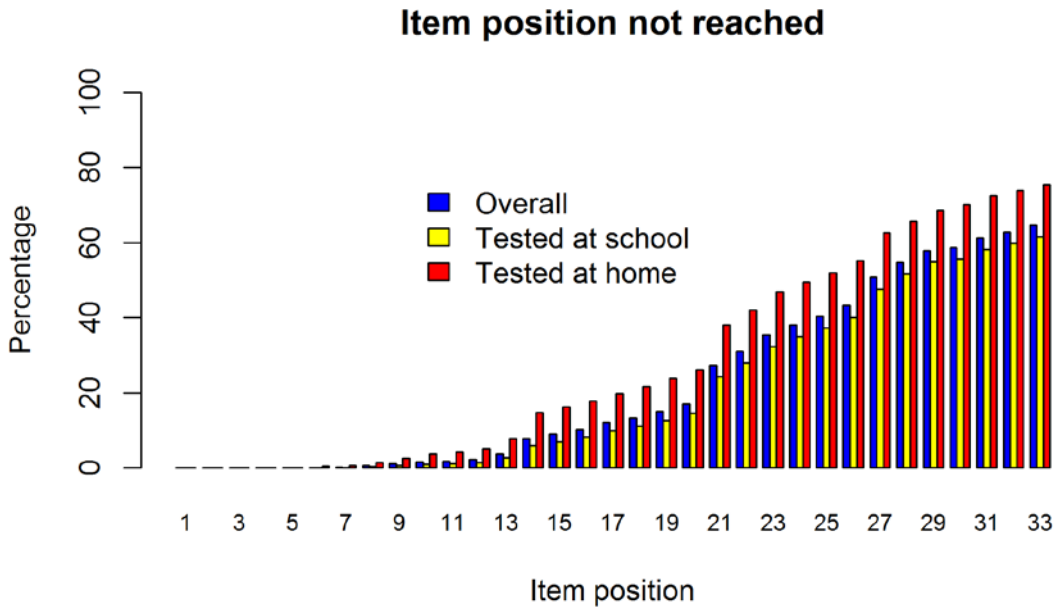

Figure 5. Item position not reached by assessment setting

5.2 Parameter Estimates

5.2.1 Item parameters

The second column in Table 6 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 24% and 90% with an average of 62% (*SD* = 19%) correct responses.

Table 6

Item Parameters

	Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
1.	reg50110_sc2g4_c	89.62	-2.688	0.047	0.920	-2.8	0.40	1.91	0.03
2.	reg5012s_sc2g4_c	n.a.	-1.976	0.043	0.800	-10.4	0.57	2.74	0.03
3.	reg50130_sc2g4_c	76.56	-1.526	0.036	0.930	-4.3	0.48	1.61	0.00
4.	reg50140_sc2g4_c	70.24	-1.113	0.035	1.000	0.2	0.43	1.25	0.00
5.	reg50150_sc2g4_c	58.21	-0.429	0.033	1.000	0.0	0.44	1.21	0.00
6.	reg5016s_sc2g4_c	n.a.	-1.052	0.029	0.960	-1.8	0.64	1.40	0.00
7.	reg50170_sc2g4_c	24.90	1.426	0.036	1.070	4.0	0.26	0.78	0.00
8.	reg50210_sc2g4_c	85.86	-2.278	0.042	0.920	-3.5	0.42	1.74	0.00
9.	reg50220_sc2g4_c	48.18	0.083	0.033	1.210	17.2	0.23	0.53	0.00
10.	reg50230_sc2g4_c	81.82	-1.918	0.039	0.890	-5.8	0.50	1.94	0.00
11.	reg50240_sc2g4_c	69.40	-1.052	0.035	0.950	-3.7	0.48	1.46	0.00
12.	reg50250_sc2g4_c	59.69	-0.502	0.034	1.050	3.6	0.40	1.03	0.00
13.	reg5026s_sc2g4_c	n.a.	1.447	0.039	0.980	-1.3	0.41	1.43	0.00
14.	reg50310_sc2g4_c	78.81	-1.685	0.039	0.920	-4.1	0.48	1.60	0.00
15.	reg50320_sc2g4_c	83.32	-2.031	0.042	0.870	-6.1	0.50	2.03	0.00
16.	reg50330_sc2g4_c	83.40	-2.050	0.042	0.890	-5.1	0.48	1.83	0.00
17.	reg50340_sc2g4_c	68.09	-0.981	0.037	0.950	-3.1	0.49	1.41	0.00
18.	reg50350_sc2g4_c	54.29	-0.223	0.035	1.050	4.0	0.40	1.00	0.00
19.	reg50360_sc2g4_c	78.58	-1.681	0.041	0.970	-1.4	0.44	1.40	0.00
20.	reg50370_sc2g4_c	63.99	-0.757	0.037	1.000	0.3	0.45	1.14	0.00
21.	reg50410_sc2g4_c	50.71	-0.105	0.038	1.210	14.1	0.27	0.61	0.00
22.	reg5042s_sc2g4_c	n.a.	-1.645	0.043	1.250	10.8	0.27	0.60	0.00

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, 1993).
Items 25 and 32 were 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).

Table 6 (continued)

	Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
23.	reg50430_sc2g4_c	23.60	1.409	0.046	1.010	0.4	0.34	1.00	0.00
24.	reg50440_sc2g4_c	35.79	0.624	0.043	1.190	10.7	0.25	0.58	0.00
26.	reg50460_sc2g4_c	41.51	0.258	0.044	1.140	7.9	0.33	0.76	0.00
27.	reg50510_sc2g4_c	74.10	-1.625	0.049	0.930	-3.1	0.49	1.54	0.00
28.	reg5052s_sc2g4_c	n.a.	-0.821	0.043	0.890	-4.9	0.61	1.71	0.00
29.	reg50530_sc2g4_c	35.07	0.501	0.051	1.110	5.0	0.36	0.83	0.00
30.	reg50540_sc2g4_c	59.72	-0.878	0.049	0.980	-0.7	0.49	1.30	0.00
31.	reg5055s_sc2g4_c	n.a.	-0.773	0.050	0.950	-2.1	0.54	1.41	0.00
33.	reg50570_sc2g4_c	50.86	-0.498	0.051	1.030	1.6	0.47	1.10	0.00

The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 6. The step parameters for polytomous variables are depicted in Table 7. 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 reg50110_sc2g4_c) to 1.4 (item reg5026s_sc2g4_c) with an average difficulty of -0.79. 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 6) were rather small (all *SE*s ≤ 0.06).

Table 7

Step Parameters (with Standard Errors) for Polytomous Items

Item	Step 1	Step 2	Step 3	Step 4	Step 5
reg5012s_sc2g4_c	1.237	-1.237			
	(0.048)				
reg5016s_sc2g4_c	-0.267	0.264	0.780	-0.193	-0.585
	(0.052)	(0.058)	(0.069)	(0.066)	
reg5042s_sc2g4_c	0.064	-0.151	0.087		
	(0.055)	(0.056)			
reg5052s_sc2g4_c	0.783	-0.842	0.059		
	(0.067)	(0.073)			
reg5055s_sc2g4_c	-0.435	0.078	0.356		
	(0.054)	(0.062)			

Note. The last step parameter is not estimated and has, thus, no standard error because it is a constrained parameter for model identification.

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 1.536, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = .824) was good. The mean of the item distribution was about 0.79 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**
4		
3	x	
	x	
	x	
	x	
	xx	
	x	
2	xxx	
	xx	
	xxxx	
	xxxx	
	xxxx	7 13 23
	xxxxx	
1	xxxxxx	
	xxxxxxx	
	xxxxxx	
	xxxxxxxx	24 28
	xxxxxxxx	
	xxxxxxxxxx	25
0	xxxxxxxx	9
	xxxxxxxx	21
	xxxxxxx	18
	xxxxxxxx	5 12 31
	xxxxxx	
	xxxxxxx	20 27 30
-1	xxxxxx	17 29
	xxxxxx	4 6 11
	xxxxx	
	xxxxx	
	xxxxx	3 22 26
	xxxx	14 19
	xxxx	2 10
-2	xxx	15 16
	xx	8
	x	
	x	
	x	1
-3		
-4		

Figure 6. Test targeting. The distribution of person ability in the sample is depicted on the left-hand side of the graph, with each 'X' representing 42.3 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 6).

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 59 items. The probability of a correct response ranged from 24% to 94% across all items (*Mdn* = 65%). Thus, the number of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.81 to 1.35, the respective *t*-value from -13.5 to 17.9, 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 acceptable (see Table 6). Values of the WMNSQ ranged from 0.80 (item reg5012s_sc2g4_c) to 1.25 (reg5042s_sc2g4_c). Five items exhibited a *t*-value of the WMNSQ greater than 6, four of these items exhibited a *t*-value of the WMNSQ greater than 10, indicating existence of item over- or underfit. Point-biserial correlations between the item scores and the total scores ranged from .23 (item reg50220_sc2g4_c) to .64 (reg5016s_sc2g4_c) and had a mean of .40. All item characteristic curves showed acceptable 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 -.49 to .06 with a mean of -.20. 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 gender, the number of books at home (as a proxy for socioeconomic status) and migration background (see Pohl & Carstensen, 2012, for a description of these variables). In addition, we estimated DIF for school recommendation and compared the two assessment settings (school or home environment). The differences between the estimated item difficulties in the various groups are summarized in Table 8. For example, the column "male vs. female" reports the differences in item difficulties between male and female students; a positive value would indicate that the test was more difficult for male students, whereas a negative value would highlight a lower difficulty for male students as opposed to female students. 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 9).

Table 8

Differential Item Functioning

Item	Gender	Books	Migration	Recommendation	Setting
	male vs. female	< 100 vs. ≥ 100	without vs. with	no sec. vs. sec.	school vs. home
reg50110_sc2g4_c	0.122 (0.099)	-0.018 (-0.015)	-0.046 (-0.037)	0.68 (0.657)	-0.162 (-0.131)
reg5012s_sc2g4_c	0.06 (0.049)	0.252 (0.215)	0.052 (0.042)	0.796 (0.769)	0.056 (0.045)
reg50130_sc2g4_c	-0.016 (-0.013)	0.012 (0.01)	0.03 (0.024)	0.186 (0.18)	-0.072 (-0.058)
reg50140_sc2g4_c	-0.046 (-0.037)	0.008 (0.007)	0.048 (0.039)	-0.002 (-0.002)	-0.048 (-0.039)
reg50150_sc2g4_c	0.024 (0.019)	0.206 (0.176)	-0.1 (-0.081)	-0.076 (-0.073)	-0.24 (-0.194)
reg5016s_sc2g4_c	-0.064 (-0.052)	0.112 (0.095)	0.024 (0.019)	0.244 (0.236)	-0.04 (-0.032)
reg50170_sc2g4_c	-0.226 (-0.183)	0.076 (0.065)	-0.096 (-0.078)	-0.466 (-0.45)	-0.178 (-0.144)
reg50210_sc2g4_c	0.202 (0.164)	0.096 (0.082)	0.04 (0.032)	0.208 (0.201)	-0.01 (-0.008)
reg50220_sc2g4_c	0.112 (0.091)	-0.312 (-0.266)	0.192 (0.156)	-0.59 (-0.57)	-0.09 (-0.073)
reg50230_sc2g4_c	0.062 (0.05)	0.01 (0.009)	0.144 (0.117)	0.506 (0.489)	0.158 (0.128)
reg50240_sc2g4_c	0.306 (0.248)	0.1 (0.085)	-0.052 (-0.042)	0.154 (0.149)	-0.046 (-0.037)
reg50250_sc2g4_c	0.376 (0.305)	-0.06 (-0.051)	0.114 (0.092)	-0.19 (-0.183)	-0.092 (-0.074)
reg5026s_sc2g4_c	0.124 (0.101)	0.24 (0.204)	0.1 (0.081)	0.12 (0.116)	0.03 (0.024)
reg50310_sc2g4_c	-0.112 (-0.091)	0.184 (0.157)	-0.266 (-0.216)	0.326 (0.315)	-0.094 (-0.076)
reg50320_sc2g4_c	-0.094 (-0.076)	0.176 (0.15)	0.026 (0.021)	0.646 (0.624)	0.136 (0.11)
reg50330_sc2g4_c	-0.018 (-0.015)	0.238 (0.203)	-0.176 (-0.143)	0.526 (0.508)	-0.19 (-0.153)
reg50340_sc2g4_c	0.256 (0.208)	0.008 (0.007)	0.026 (0.021)	0.206 (0.199)	0.006 (0.005)
reg50350_sc2g4_c	0.106 (0.086)	-0.206 (-0.176)	0.1 (0.081)	-0.232 (-0.224)	-0.094 (-0.076)

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

Table 8 (continued).

Item	Gender	Books	Migration	Recommendation	Setting
	male vs. female	< 100 vs. ≥ 100	without vs. with	no sec. vs. sec.	school vs. home
reg50360_sc2g4_c	0.242 (0.196)	-0.108 (-0.092)	-0.054 (-0.044)	0.192 (0.185)	0.152 (0.123)
reg50370_sc2g4_c	0.152 (0.123)	-0.284 (-0.242)	0.194 (0.157)	-0.066 (-0.064)	0.078 (0.063)
reg50410_sc2g4_c	-0.382 (-0.31)	-0.38 (-0.324)	0.01 (0.008)	-0.58 (-0.56)	0.228 (0.184)
reg5042s_sc2g4_c	-0.372 (-0.302)	-0.39 (-0.332)	0.184 (0.149)	-0.456 (-0.44)	-0.164 (-0.132)
reg50430_sc2g4_c	-0.282 (-0.229)	0.146 (0.124)	-0.016 (-0.013)	-0.328 (-0.317)	-0.104 (-0.084)
reg50440_sc2g4_c	-0.122 (-0.099)	-0.204 (-0.174)	-0.12 (-0.097)	-0.79 (-0.763)	-0.026 (-0.021)
reg50460_sc2g4_c	-0.188 (-0.153)	-0.196 (-0.167)	-0.016 (-0.013)	-0.544 (-0.525)	0.19 (0.153)
reg50510_sc2g4_c	-0.044 (-0.036)	0.104 (0.089)	0.174 (0.141)	0.12 (0.116)	0.22 (0.178)
reg5052s_sc2g4_c	-0.062 (-0.05)	0.222 (0.189)	-0.15 (-0.122)	0.346 (0.334)	0.248 (0.2)
reg50530_sc2g4_c	-0.074 (-0.06)	0.006 (0.005)	-0.23 (-0.187)	-0.574 (-0.554)	-0.116 (-0.094)
reg50540_sc2g4_c	-0.212 (-0.172)	0.08 (0.068)	-0.334 (-0.271)	-0.13 (-0.126)	0.012 (0.01)
reg5055s_sc2g4_c	-0.128 (-0.104)	0.152 (0.13)	-0.176 (-0.143)	0.264 (0.255)	0.364 (0.294)
reg50570_sc2g4_c	-0.164 (-0.133)	-0.004 (-0.003)	-0.146 (-0.118)	-0.144 (-0.139)	-0.116 (-0.094)
Main effect (with DIF)	-0.212 (-0.172)	-0.760 (-0.648)	0.166 (0.135)	-1.296 (-1.251)	-0.152 (-0.123)
Main effect (without DIF)	-0.220 (-0.179)	-0.764 (-0.653)	0.160 (0.130)	-1.248 (-1.218)	-0.140 (-0.113)

Gender: The sample included 2,974 (44%) males, 3,120 (47%) females, and 607 (9%) without a valid response for their gender. On average, male participants had a lower estimated reading ability than females (main effect = -0.212 logits, Cohen's d = -0.172). There was no considerable DIF greater than 0.6 logits (highest DIF = 0.38 for item reg50410_sc2g4_c). An overall test for DIF (see Table 9) 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 regarding gender.

Books: The number of books at home was used as a proxy for socioeconomic status. There were 2,070 (31%) test takers with 0 to 100 books at home, 3,730 (56%) test takers with more than 100 books at home, and 901 (13%) 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 0.760 logits (Cohen's d = -0.648) lower in reading than participants with more than 100 books. There was no considerable DIF comparing participants with many or fewer books (highest DIF = 0.39 for itemreg5042s_sc2g4_c). The overall test for DIF using the BIC favored the main effects model (Table 9).

Table 9

Comparisons of Models with and without DIF

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Gender	main effect	6,094	162,614.76	44	162,702.76	162,998.22
	DIF	6,094	162,418.58	75	162,568.58	163,072.22
Books	main effect	5,800	153,712.75	44	153,800.75	154,094.04
	DIF	5,800	153,538.39	75	153,688.39	154,188.31
Migration	main effect	6,108	163,044.57	44	163,132.57	163,428.65
	DIF	6,108	162,964.06	75	163,114.06	163,617.86
Recommendation	main effect	4,743	131,702.48	44	131,790.48	132,074.92
	DIF	4,743	131,129.63	75	131,279.63	131,764.46
Setting	main effect	6,701	181,535.97	44	181,623.97	181,923.61
	DIF	6,701	181,470.35	74	181,618.35	182,122.29

Migration background: There were 4,052 participants (60%) with no migration background, 2,061 subjects (31%) with a migration background, and 593 individuals (9%) 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 = 0.166 logits, Cohen's d = 0.135). 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 recommendation: Due to a high amount of missing values for this variable when students were assessed within the home environment (83% missing), the analyses for school recommendation refer only to students tested within the school context. Within the school context, 1,956 subjects (37%) obtained a recommendation for secondary school (grammar

school; German: “Gymnasium”) whereas 2,787 (53%) did not. Furthermore, 531 (10%) of the participants within the school context had no valid response about their school recommendation. Respondents recommended for grammar schools showed a higher reading ability on average (-1.296 logits, Cohen’s d = -1.251) than subjects in with other recommendations. There were four items with DIF greater than 0.6 logits (highest DIF = 0.796 for item reg5012s_sc2g4_c). The overall model test indicated a 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.

Setting: The reading competence test was administered in two different settings (see section 3.1 for the design of the study). A subsample of 5,274 (79%) persons received the reading test in small groups at school, whereas 1,436 (21%) participants finished the test individually at their private homes. Subjects who finished the reading test at school were on average 0.152 logits (Cohen’s d = -0.123) better than those working at their private homes. However, there was no noteworthy DIF due to the administration setting; all differences in item difficulties were smaller than 0.6 logits. Regarding the overall model test (see Table 9), the AIC indicated a slightly better fit for the more complex DIF model while the BIC indicated a better fit for the more parsimonious model including only the main effect.

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 (GPCM) that estimates discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 6), ranging from 0.53 (item reg50220_sc2g4_c) to 2.74(item reg5012s_sc2g4_c). The average discrimination parameter fell at 1.34. Model fit indices suggested a slightly better model fit of the GPCM model (AIC = 178,936.35, BIC = 179,433.48) as compared to the PCM model (AIC = 181,621.80, BIC = 181,914.63). Despite the empirical preference for the GPCM model, the PCM model more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the partial credit model 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 10. 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 = 181,253.19, BIC = 181,580.07, number of parameters = 48) than the unidimensional model (AIC = 181,621.80, BIC = 181,914.63, number of parameters = 43).

These results indicate that the three cognitive requirements measure a common construct, albeit it is not completely unidimensional.

Table 10

Results of Three-Dimensional Scaling

	Dim 1	**Dim 2**	**Dim 3**
Finding information in the text (Dim 1)	(2.48)		
(9 items)			
Drawing text-related conclusions (Dim 2)	.96	(1.61)	
(12 items)			
Reflecting and assessing (Dim 3)	.93	.94	(1.20)
(10 items)			

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

Table 11

Results of Five-Dimensional Scaling

	Dim 1	**Dim 2**	**Dim 3**	**Dim 4**	**Dim 5**
Information (Dim 1)	(2.34)				
(7 items)					
Instruction (Dim 2)	.89	(1.76)			
(6 items)					
Advertising (Dim 3)	.88	.91	(2.39)		
(7 items)					
Commenting (Dim 4)	.76	.69	.71	(0.90)	
(5 items)					
Literacy (Dim 5)	.87	.85	.88	.78	(2.28)
(6 items)					

Note. Variances of the dimensions are given in the diagonal and correlations are given 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 = .69 and r = .91. The smallest correlation was found between Dimension 2 ("instruction text") and Dimension 4 ("commenting text"). Dimension 2 and Dimension 3 ("advertising text") 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 = 180,300.88, BIC = 180,689.05, number of parameters = 57) fitted the data better than the unidimensional model (AIC = 181,621.80, BIC = 181,914.63, number of parameters = 43). As each text function corresponded to one of the five texts, local item dependence and the text functions were confounded. Consequently, the deviation of the correlations from a perfect correlation shown in Table 11 may result from multidimensionality as well as from local item dependence.

Nevertheless, for the unidimensional model the average absolute residual correlations as indicated by the Q_3 statistic (see Table 6) were quite low (M = .00, SD = .01)—the largest individual residual correlation was .0.03—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.

6. Discussion

The analyses in the previous sections aimed at providing detailed information on the quality of the reading test in starting cohort 2 for grade 4 and at describing how the reading competence score was estimated.

We investigated different types of missing responses and examined the item and test parameters. We thoroughly checked item fit statistics for simple MC items, subtasks of CMC and MA items, as well as the aggregated polytomous CMC and MA items, and examined the correlations between correct and incorrect responses and the total score. Further quality inspections were conducted by examining differential item functioning, testing Rasch-homogeneity, investigating the tests' dimensionality as well as local item dependence. Various criteria indicated a good fit of the items and measurement invariance across various subgroups. However, the number 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 result, 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 had good psychometric properties that facilitated 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 contains 33 items, of which 26 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. A total of 7 items were scored as polytomous variables (CMC and MA items). MC items are marked with a '0_c' at the end of the variable name, whereas the variable names of CMC and MA items end in 's_c'. Furthermore, CMC and MA (polytomous) items with categories of less than *N* = 200 responses were collapsed in the analyses to avoid possible estimation problems (see chapter 4.2 and Appendix A of this report). Because in grade 5 of starting cohort 3 the same test was administered, all categories were collapsed in the same way (i.e., the PCM variables in the SUF for both starting cohorts contain identical categories). In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category. Two items (reg5045s_sc2g4_c and reg50560_sc2g4_c) were removed from the final scaling procedure due to poor item fit in preliminary analyses. Nevertheless, these items are included in the Scientific Use File.

7.2 Linking of competence scores across starting cohorts

The reading competence test administered in grade 4 of starting cohort 2 is identical to the test administered in grade 5 of starting cohort 3. To place the different measurements onto a common scale and, thus, allow for the comparison of competences across starting cohorts we adopted the "mean/mean" linking approach using an anchor-items design as described in Fischer, Rohm, Gnambs, & Carstensen (2016).

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 different starting cohorts showed a non-negligible shift in item difficulties. Differential item functioning was evaluated using a multi-group IRT approach (see section 4.3). The differences between the estimated item difficulties in grade 4 (starting cohort 2) and grade 5 (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. A positive value in the second column indicates that the item was more difficult for grade 4 students compared to grade 5 students, whereas a negative value indicates a lower difficulty for grade 4 students. On average, participants from grade 4 had a lower estimated reading ability than grade 5 students (main effect = -0.52 logits, Cohen's d = -0.42). Only one item (reg5012s_c) exhibited considerable DIF greater than 0.4 logits with a DIF effect of 0.602 logits (Cohen's d = 0.494). However, minimum effects hypotheses tests revealed no significant (α = .05) DIF for any item. An overall test for DIF was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). Both AIC and BIC favored the model estimating DIF (AIC = 319,069.39, BIC = 319,623.18, number of parameters = 75) over the more parsimonious model including only the main effect (AIC = 319,438.78, BIC = 319,763.67, number of parameters = 44). Hence, when both grades are conjointly analyzed, there is difference in overall test difficulty as well as minor DIF between both starting cohorts. However, the respective effects were rather small.

Table 12

Differential Item Functioning Analyses

Item	grade 4 vs. grade 5	*F*	Item	grade 4 vs. grade 5	*F*
reg50110_c	0.214 (0.176)	2.65	reg50340_c	0.156 (0.128)	2.49
reg5012s_c	0.602 (0.494)	20.32	reg50350_c	-0.088 (-0.072)	9.09
reg50130_c	0.264 (0.217)	10.81	reg50360_c	0.192 (0.158)	3.33
reg50140_c	0.012 (0.01)	1.22	reg50370_c	0.078 (0.064)	0.06
reg50150_c	-0.328 (-0.269)	68.80	reg50410_c	-0.016 (-0.013)	2.00
reg5016s_c	-0.004 (-0.003)	10.70	reg5042s_c	-0.238 (-0.195)	5.56
reg50170_c	0.056 (0.046)	0.12	reg50430_c	0.004 (0.003)	0.36
reg50210_c	0.124 (0.102)	0.37	reg50440_c	-0.334 (-0.274)	42.65
reg50220_c	-0.378 (-0.310)	85.33	reg50460_c	-0.132 (-0.108)	9.80
reg50230_c	0.256 (0.210	7.78	reg50510_c	0.272 (0.223)	6.58
reg50240_c	-0.122 (-0.100)	14.08	reg5052s_c	0.194 (0.159)	5.81
reg50250_c	-0.192 (-0.158)	27.71	reg50530_c	-0.016 (-0.013)	0.93
reg5026s_c	-0.096 (-0.079)	6.20	reg50540_c	0.044 (0.036)	0.16
reg50310_c	0.166 (0.136)	2.17	reg5055s_c	0.150 (0.123)	0.05
reg50320_c	0.382 (0.313)	19.03	reg50570_c	-0.030 (-0.025)	1.92
reg50330_c	0.122 (0.100)	0.42			

Note. Differences in item difficulty parameters between the sample in grade 4, starting cohort 2, and the sample in grade 5, starting cohort 3 (with Cohen's *d* in parantheses). Positive values indicate easier items in the grade 4 sample; 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_{0077} (1, 11,891) = 126.69. A non-significant test indicates measurement invariance.

To place the different measurements onto a common scale and, thus, allow for the comparison of competences between starting cohorts, we applied the "mean/mean" linking

procedure as described in Fischer et al. (2016). All items of the test (except items reg5045s_c and reg50560_c that were excluded due to unsatisfactory item fit in grade 4) were used as anchor items for the "mean/mean" linking. In grade 4, the mean difficulty of item parameters between grade 4 and grade 5 is used as linking constant.

Because grade 5 adopted a rotation design and presented the reading test either as first or as second test within a test battery, the linking constant was corrected for the position effect identified in grade 5. The thus derived correction term was c = -0.567. This correction term was subsequently added to each difficulty parameter estimated in grade 4 to derive the linked item parameters.

7.3 Reading competence scores

In the SUF, manifest reading competence scores are provided in the form of one WLE per respondent, "reg4_sc1", including the respective standard error, "reg4_sc2". These WLEs are linked to the scale of the test administered in grade 5 of starting cohort 3. As a result, the WLE scores provided in "reg4_sc1" can be used for the comparison of reading competence between the two starting cohorts.

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.

Users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values. 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.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174. doi:10.1007/BF02296272

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196. doi:10.1007/BF02294457

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176. doi:10.1002/j.2333-8504.1992.tb01436.x

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.

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*, 461–464. doi 10.1214/aos/1176344136

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 2

```
Title SC2 G4 READING: Partial Credit Model;

/* load data */
datafile [FILENAME].sav ! filetype=spss,
      responses = reg50110_sc2g4_c reg5012s_sc2g4_c reg50130_sc2g4_c
                  reg50140_sc2g4_c reg50150_sc2g4_c reg5016s_sc2g4_c
                  reg50170_sc2g4_c reg50210_sc2g4_c reg50220_sc2g4_c
                  reg50230_sc2g4_c reg50240_sc2g4_c reg50250_sc2g4_c
                  reg5026s_sc2g4_c reg50310_sc2g4_c reg50320_sc2g4_c
                  reg50330_sc2g4_c reg50340_sc2g4_c reg50350_sc2g4_c
                  reg50360_sc2g4_c reg50370_sc2g4_c reg50410_sc2g4_c
                  reg5042s_sc2g4_c reg50430_sc2g4_c reg50440_sc2g4_c
                  reg50460_sc2g4_c reg50510_sc2g4_c reg5052s_sc2g4_c
                  reg50530_sc2g4_c reg50540_sc2g4_c reg5055s_sc2g4_c
                  reg50570_sc2g4_c,
         pid=ID_t >> daten.dat;


/* scoring */
codes 0,1,2,3,4,5;
score (0,1)         (0,1)                ! items (1,3-5,7-21,23-26,28-
                                                  29,31);
score (0,1,2)       (0,0.5,1)            ! items (2);
score (0,1,2,3)     (0,0.5,1,1.5)        ! items (22,27,30);
score (0,1,2,3,4,5) (0,0.5,1,1.5,2,2.5)  ! items (6);


/* 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;

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

Appendix B: Different Text Types and Cognitive Requirements

Item	Position	Text Types	Cognitive Requirements
reg50110_sc2g4_c	1	Information text	Drawing text-related conclusions
reg5012s_sc2g4_c	2	Information text	Finding information
reg50130_sc2g4_c	3	Information text	Finding information
reg50140_sc2g4_c	4	Information text	Drawing text-related conclusions
reg50150_sc2g4_c	5	Information text	Drawing text-related conclusions
reg5016s_sc2g4_c	6	Information text	Reflecting and assessing
reg50170_sc2g4_c	7	Information text	Reflecting and assessing
reg50210_sc2g4_c	8	Instruction text	Finding information
reg50220_sc2g4_c	9	Instruction text	Reflecting and assessing
reg50230_sc2g4_c	10	Instruction text	Drawing text-related conclusions
reg50240_sc2g4_c	11	Instruction text	Drawing text-related conclusions
reg50250_sc2g4_c	12	Instruction text	Drawing text-related conclusions
reg5026s_sc2g4_c	13	Instruction text	Reflecting and assessing
reg50310_sc2g4_c	14	Advertising text	Finding information
reg50320_sc2g4_c	15	Advertising text	Finding information
reg50330_sc2g4_c	16	Advertising text	Drawing text-related conclusions
reg50340_sc2g4_c	17	Advertising text	Finding information
reg50350_sc2g4_c	18	Advertising text	Drawing text-related conclusions
reg50360_sc2g4_c	19	Advertising text	Finding information
reg50370_sc2g4_c	20	Advertising text	Drawing text-related conclusions
reg50410_sc2g4_c	21	Commenting text	Finding information
reg5042s_sc2g4_c	22	Commenting text	Drawing text-related conclusions
reg50430_sc2g4_c	23	Commenting text	Reflecting and assessing
reg50440_sc2g4_c	24	Commenting text	Reflecting and assessing
reg5045s_sc2g4_c	25	Commenting text	-
reg50460_sc2g4_c	26	Commenting text	Drawing text-related conclusions
reg50510_sc2g4_c	27	Literary text	Finding information
reg5052s_sc2g4_c	28	Literary text	Reflecting and assessing
reg50530_sc2g4_c	29	Literary text	Reflecting and assessing
reg50540_sc2g4_c	30	Literary text	Drawing text-related conclusions
reg5055s_sc2g4_c	31	Literary text	Reflecting and assessing
reg50560_sc2g4_c	32	Literary text	Reflecting and assessing
reg50570_sc2g4_c	33	Literary text	Reflecting and assessing

Note. Position = Item position within test. The items on positions 25 and 32 were excluded from the analyses due to an unsatisfactory item fit (see section 2).