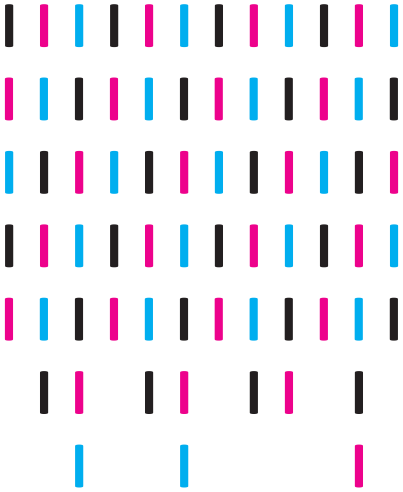

NEPS SURVEY PAPERS

Tanja Kutscher and Anna Scharl

NEPS TECHNICAL REPORT FOR READING: SCALING RESULTS OF STARTING COHORT 3 FOR GRADE 12

NEPS Survey Paper No. 67
Bamberg, January 2020

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 12

Tanja Kutscher and Anna Scharl

Leibniz Institute for Educational Trajectories, Bamberg, Germany

E-mail address of lead author:

tanja.kutscher@lifbi.de

Bibliographic data:

Kutscher, T., & Scharl, A. (2020). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 3 for Grade 12* (NEPS Survey Paper No. 67). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP67:1.0

Acknowledgements:

This report is an extension to NEPS Survey Paper 20 (Scharl, Fischer, Gnambs, & Rohm, 2017) that presents the scaling results for reading competence of starting cohort 3 for grade 9. Therefore, various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (e.g., Krannich et al., 2017; Pohl, Haberkorn, Hardt, & Wiegand, 2012) to facilitate the understanding of the presented results.

NEPS Technical Report for Reading: Scaling Results of Starting Cohort 3 for Grade 12

Abstract

The National Educational Panel Study (NEPS) investigates the development of competencies across the life span and develops tests for the assessment of different competence domains. In order to evaluate the quality of the competence tests, a range of analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedures for the reading competence test in grade 12 of starting cohort 3 (fifth grade). The reading competence test contained 29 items (distributed among an easy and a difficult booklet) with different response formats representing different cognitive requirements and text functions. The test was administered to 3,663 students. Based on the preliminary analysis, one item was excluded from the analyses due to substantial differential item functioning between the booklets and one respondent was ignored due to a high number of invalid responses. The students' responses were scaled using the partial credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, the test's dimensionality, and local item independence were evaluated to ensure the quality of the test. These analyses showed that the test exhibited an acceptable reliability and that the items fitted the model in a satisfactory way. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test were the large number of items targeted toward a lower reading ability as well as the large percentage of items at the end of the test that were not reached due to time limits. Further challenges related to the dimensionality analyses based on both text functions and cognitive requirements. Overall, the reading test had acceptable psychometric properties that allowed for an estimation of reliable reading competence scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the R syntaxes 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, Artelt, Prenzel, Senkbeil, Ehmke, and Carstensen (2011) and Fuß, Gnambs, Lockl, and Attig (2019).

Most of the competence data are scaled using models that are based on item response theory (IRT). Because most of the competence tests were developed specifically for implementation in the NEPS, several analyses were conducted to evaluate the quality of the tests. The IRT models chosen for scaling the competence data and the analyses performed for checking the quality of the scale are described in Pohl and Carstensen (2012).

In this paper, the results of these analyses are presented for reading competence in grade 12 of starting cohort 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, we briefly describe specific aspects of the reading competence test that are necessary for understanding the scaling results presented in this paper.

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

The reading competence test included three types of response formats: (i) simple multiple choice (MC) items, (ii) complex multiple choice (CMC) items, and (iii) 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 required the respondents 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 42 items. In order to evaluate the quality of these items extensive preliminary analyses were conducted. These preliminary analyses identified a significant DIF effect for the test difficulty booklets for one item (reg122307s_sc3g12_c). Therefore, this item was removed from the final scaling procedure. Thus, the analyses presented in the following sections and the competence scores derived for the respondents are based on the remaining 41 items.

3. Data

3.1 The Design of the Study

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

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

In order to measure participants' reading competence with great accuracy, the difficulty of the administered items should adequately match the participants' abilities. Therefore, the study adopted the principles of longitudinal multistage testing (Pohl, 2013). Based on preliminary studies two different versions of the reading competence test were developed that differed in their average difficulty (i.e., an easy and a difficult test). Both tests included five texts and 28 items that represented the five text functions (see Table 1) and three cognitive requirements (see Table 2) as described above. Three texts with 15 items were identical in both test versions (see Table 1), whereas 13 items were unique to the easy and the difficult test. The different response formats of the items are summarized in Table 3 (for an overview of the items in the reading test, see Appendix, part A). The number of subtasks varied within CMC items between three and six and within MA items between four and six. Participants were assigned either to the easy or the difficult test based on their estimated reading competence in the previous assessment (Scharl, Fischer, Gnambs, & Rohm, 2017). Participants with an ability estimate below the sample's mean ability received the easy test, whereas participants with a reading competence above the sample's mean received the difficult test.

Table 1

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

Text types	**Only in easy test**	**Both tests**	**Only in difficult test**
Literary	7		7
Instruction		5	
Commenting		5	1
Advertising		5	
Information	6		5
Total number of items	13	15	13

Table 2

Number of Items by Cognitive Requirements and Difficulty of the Test

Cognitive requirements	**Easy test**	**Difficult test**
Finding information	6	8
Drawing text-related conclusions	8	7
Reflecting and assessing	14	13
Total number of items	28	28

The panel study aimed at retesting all students that were initially included in the starting cohort 3 for fifth grade (see Krannich et al., 2017; Pohl, Haberkorn, Hardt, & Wiegand, 2012). Because some students left their original schools during the course of the longitudinal study or left the school context altogether, the participants of the starting cohort were divided into two subsamples that exhibited different assessment settings: Students that remained at the same school as in the previous assessment were tested at school in a group setting; in contrast, students that left their original school were tracked and, subsequently, individually tested at home (for details regarding the data collection process, see the respective field report for wave 9). Thus, the context of test administration differed between the two groups.

Table 3

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

Response format	Easy test	Difficult test
Simple multiple choice	20	20
Complex multiple choice	7	7
Matching	1	1
Total number of items	28	28

3.2 Sample

A total of 3,663[1] students (50% women) received the reading competence test. For one respondent less than three valid item responses were available. Because no reliable ability scores can be estimated based on such few valid responses, this case was exclude from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 3,662 individuals. The number of participants within each (quasi-) experimental condition is given in Table 4. A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

Table 4

Number of Participants by the (Quasi-)Experimental Conditions

Assessment setting:		**At school (*n* = 1,766)**		**At home (*n* = 1,896)**		**Total**
	Test position:	**first position**	**second position**	**first position**	**second position**	
Test difficulty	**Easy test**	214	239	493	495	1,441
	Difficult test	666	647	464	444	2,221
	Total	880	886	957	939	3,662

[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 respondents did not reach, d) items that have not been administered, and, finally, e) multiple kinds of missing responses within CMC and MA 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 participants 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 branched testing design 26 items were not administered to all participants. For respondents receiving the easy test 13 difficult items were missing by design, whereas 13 easy items were missing by design for respondents answering the difficult test (see Table 1). As CMC 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. If just one kind of missing response occurred, the item was coded according to the corresponding missing response. If the subtasks contained different kinds of missing responses, the item was labeled as a not-determinable missing response.

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

4.2 Scaling Model

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

CMC and MA items consisted of a set of subtasks. For each item, they were aggregated to a polytomous variable, 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 were scored as missing. Categories of polytomous variables with less than $N = 200$ responses were collapsed in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items; in these cases, the lower categories were collapsed into one category. For 10 of the 11 CMC and MA items categories were collapsed.

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 (for studies on the scoring of different response formats, see Pohl & Carstensen, 2013).

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 (for an R syntax for collapsing response categories, fitting the scaling model and estimating WLEs, see Appendix, part B).

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 variables that were included in the final scaling model.

For the MC items, the quality of the distractors 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 (used as an item discrimination index) 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 participants. If some items favored certain subgroups (e.g., they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., males and females) would be biased and, thus, unfair. For the present study, test fairness was investigated for the variables test position, gender, school type, 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 test difficulty and administration setting. Differential item functioning (DIF) was examined using a series of multi-group IRT models, in which main effects of the subgroups and 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 adjusted Q_3 that has an expected value of null. 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 R version 3.6.1 (R Core Team, 2019) using the TAM package version 3.3.10 (Robitzsch, Kiefer, & Wu, 2019).

5. Results

5.1 Missing Responses

5.1.1 Missing responses per person

Figure 1 shows the number of invalid responses per person by experimental condition (i.e., test difficulty and administration setting). Overall, there were very few invalid responses. Between 94.6% and 97.5% of the respondents did not have any invalid response at all; less than 2.5% had more than one invalid response. There were slightly more invalid responses for the easy test version.

Missing responses may also occur when respondents omit items. As illustrated in Figure 2, most respondents, 84.3% to 87.2%, did not skip any item and no more than five percent omitted more than one item. Slightly more items were omitted when tested at home.

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 46.4% and 80.8% of the respondents finished the entire test. Between 13.3% and 41.7% did not reach the last of the five texts. In particular, respondents receiving the difficult test at home did not reach the last text.

The aggregated polytomous variables were coded as not-determinable missing response when the subtasks of CMC or MA items contained different kinds of missing responses. Only a rather small number of not-determinable missing responses occurred. Most respondents, 98.9% to 99.5%, did not have any not-determinable missing response. There was no difference in the amount of not-determinable items between the experimental conditions.

The total number of missing responses, aggregated over invalid, omitted, not-reached, and not-determinable missing responses per person, is illustrated in Figure 4. On average, the respondents showed between *M* = 1.57 (*SD* = 3.20) and *M* = 3.91 (*SD* = 4.41) missing responses in the different experimental conditions. About 37.9% to 68.7% of the respondents had no missing response at all and about 17.9% to 46.1% of the participants had four or more missing responses.

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

5.1.2 Missing responses per item

Tables 5 and 6 provide information on the occurrence of different kinds of missing responses per item for the easy and difficult test version. Overall, in both tests the omission rates were rather low, varying across items between 0.0% and 3.1%. There was only two items with an omission rate exceeding 3.0% (reg122305s_sc3g12_c in the easy test administered at home and reg1226040_sc3g12_c in the difficult test administrated at home). For the easy test, omission rates correlated with the item difficulties at about .29 in the school setting and .31 in the home setting; for the difficult test, the respective correlations were distinctly smaller with .13 at school and .15 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 Tables 5 and 6) was rather low with the maximum rate being 3.2%.

With an item's progressing position in the test, the amount of persons that did not reach the item (columns 4 and 8 in Tables 5 and 6) rose up to a considerable amount of 19.2% to 53.6% for the different experimental conditions. Particularly, at home the last items of the difficult test were not reached by many respondents (see Figure 5).

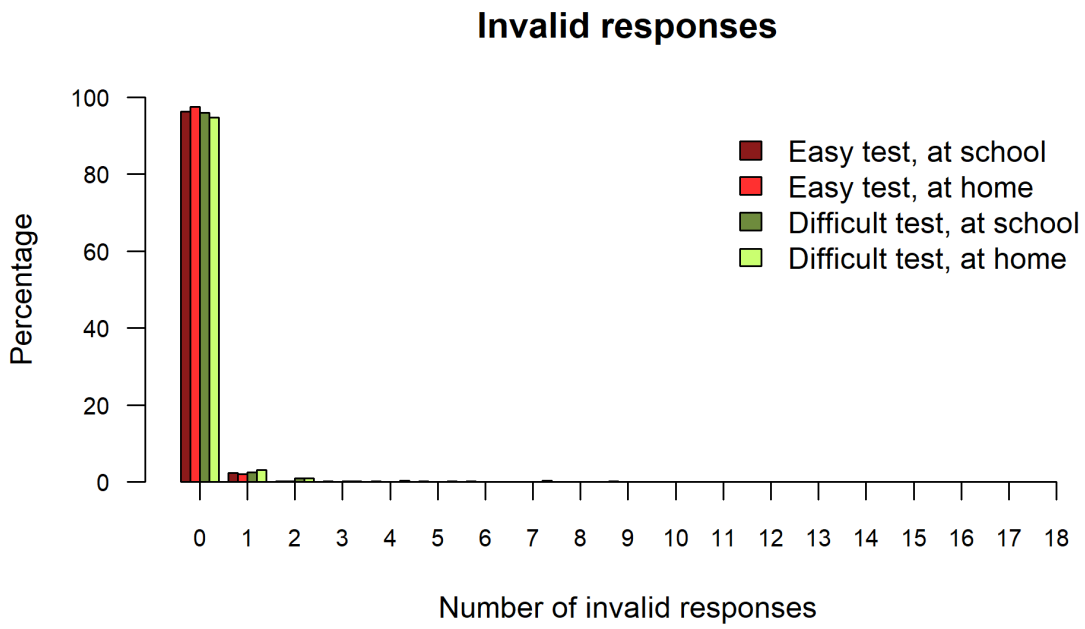

Figure 1. Number of invalid responses by experimental condition

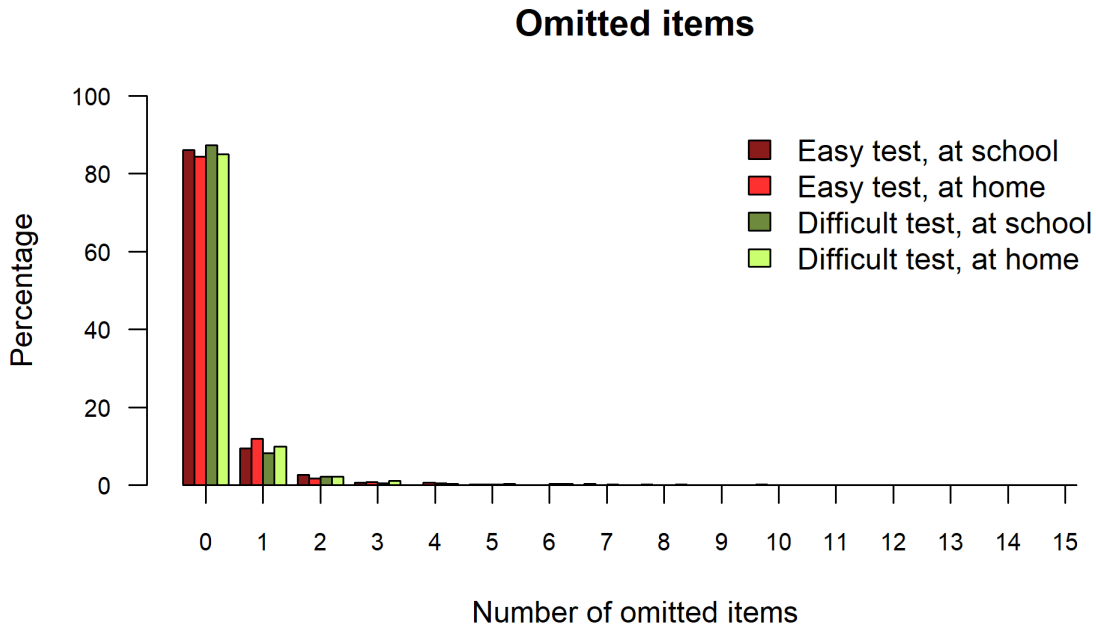

Figure 2. Number of omitted items by experimental condition

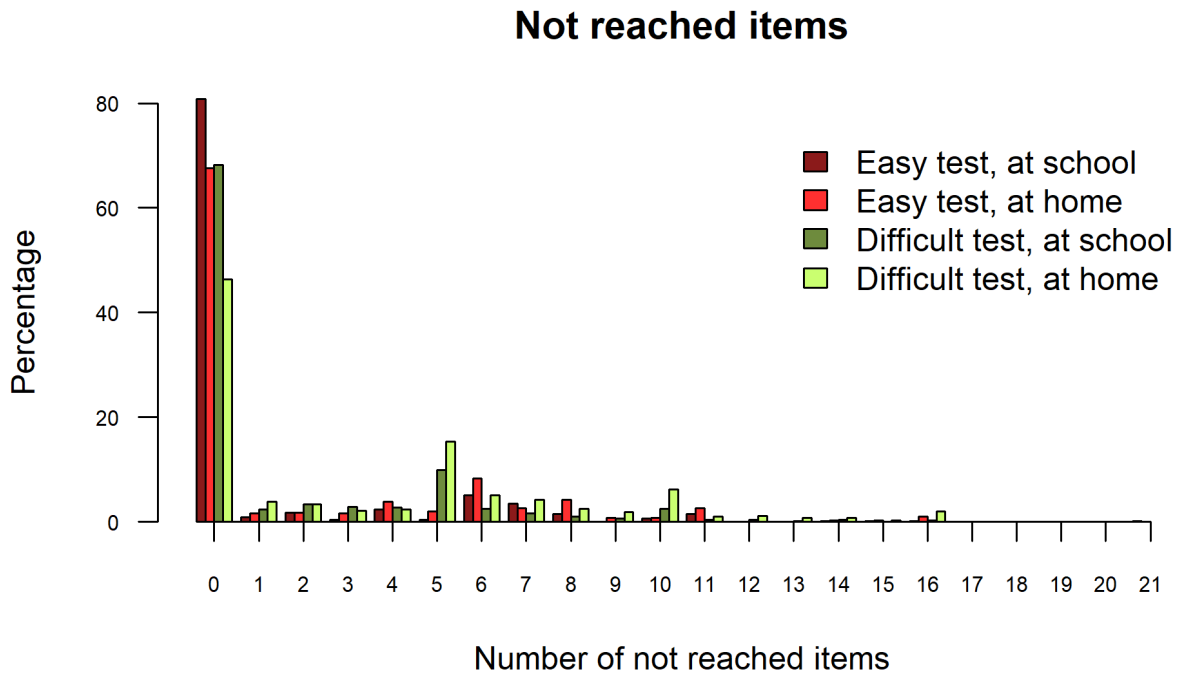

Figure 3. Number of not-reached items by experimental condition

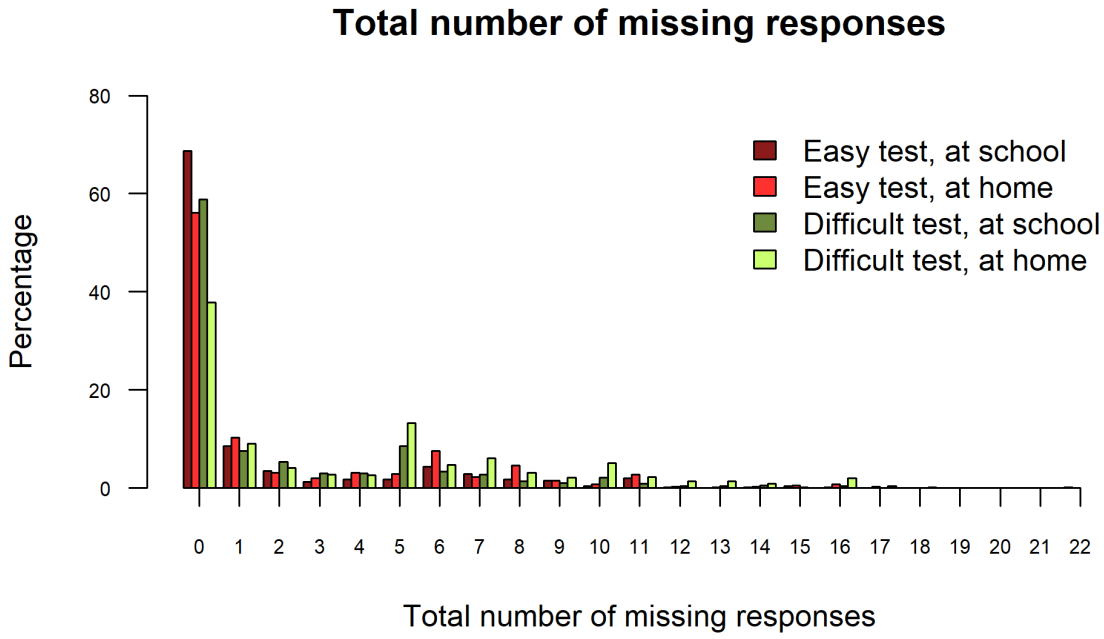

Figure 4. Total number of missing responses by experimental condition

Table 5

Percentage of Missing Values for the Easy Test by Assessment Setting

		At school				*At home*			
Item	**Pos.**	***n***	**NR**	**OM**	**NV**	***n***	**NR**	**OM**	**NV**
reg1205010_sc3g12_c	1	452	0.00	0.22	0.00	986	0.00	0.00	0.20
reg1205020_sc3g12_c	2	452	0.00	0.22	0.00	984	0.00	0.20	0.20
reg1205030_sc3g12_c	3	453	0.00	0.00	0.00	987	0.00	0.10	0.00
reg120504s_sc3g12_c	4	444	0.00	1.99	0.00	974	0.00	1.42	0.00
reg1205050_sc3g12_c	5	449	0.00	0.66	0.22	981	0.00	0.71	0.00
reg1205060_sc3g12_c	6	449	0.00	0.66	0.22	981	0.00	0.61	0.10
reg1205070_sc3g12_c	7	449	0.00	0.66	0.22	983	0.00	0.30	0.20
reg122301s_sc3g12_c	8	450	0.22	0.44	0.00	969	0.00	1.92	0.00
reg1223020_sc3g12_c	9	446	0.22	0.88	0.44	979	0.00	0.81	0.10
reg1223040_sc3g12_c	10	449	0.22	0.00	0.66	983	0.00	0.10	0.40
reg122305s_sc3g12_c	11	441	0.22	2.21	0.22	956	0.00	3.04	0.00
reg1223060_sc3g12_c	12	450	0.22	0.22	0.22	982	0.00	0.51	0.10
reg1226020_sc3g12_c	14	443	0.44	1.10	0.66	963	1.11	1.21	0.20
reg1226030_sc3g12_c	15	444	0.66	0.88	0.44	966	1.42	0.61	0.20
reg1226040_sc3g12_c	16	437	0.88	2.21	0.44	944	1.72	2.63	0.10
reg1226060_sc3g12_c	17	445	0.88	0.66	0.22	959	1.82	0.81	0.30
reg1226080_sc3g12_c	18	443	0.88	1.10	0.22	960	1.92	0.81	0.10
reg121602s_sc3g12_c	19	436	2.43	1.10	0.22	924	4.55	1.72	0.20
reg121603s_sc3g12_c	20	436	3.09	0.66	0.00	924	5.36	1.01	0.10
reg1216040_sc3g12_c	21	427	3.09	0.88	1.77	915	6.17	0.20	1.01
reg121605s_sc3g12_c	22	420	4.64	1.99	0.66	866	10.43	1.42	0.20
reg1216060_sc3g12_c	23	408	8.17	1.55	0.22	847	13.06	1.01	0.20
reg1220010_sc3g12_c	24	387	13.25	1.10	0.22	774	21.36	0.10	0.20
reg122002s_sc3g12_c	25	387	13.69	0.88	0.00	751	23.38	0.61	0.00
reg1220030_sc3g12_c	26	376	16.11	0.44	0.44	712	27.23	0.40	0.30
reg1220040_sc3g12_c	27	372	16.56	0.88	0.44	699	28.95	0.00	0.30
reg122005s_sc3g12_c	28	362	18.32	1.32	0.44	677	30.77	0.71	0.00
reg1220060_sc3g12_c	29	365	19.21	0.00	0.22	667	32.39	0.00	0.10

Note. Pos. = Item position within the easy test version, *n* = Number of valid responses, NR = Percentage of respondents that did not reach item, OM = Percentage of respondents that omitted the item, NV = Percentage of respondents with an invalid response.
The item on position 13 was excluded from the analyses due to a significant DIF effect for test difficulty (see section 2).

Table 6

Percentage of Missing Values for the Difficult Test by Assessment Setting

		At school				*At home*			
Item	**Pos.**	***n***	**NR**	**OM**	**NV**	***n***	**NR**	**OM**	**NV**
reg1229010_sc3g12_c	1	1,287	0.00	0.00	1.98	879	0.00	0.00	3.19
reg1229020_sc3g12_c	2	1,289	0.00	0.23	1.60	890	0.00	0.22	1.76
reg1229030_sc3g12_c	3	1,306	0.00	0.53	0.00	894	0.00	1.10	0.44
reg1229060_sc3g12_c	4	1,293	0.00	0.84	0.69	896	0.00	0.55	0.77
reg122907s_sc3g12_c	5	1,277	0.00	2.59	0.00	893	0.00	1.65	0.00
reg1229080_sc3g12_c	6	1,302	0.00	0.61	0.23	889	0.00	1.21	0.88
reg1229100_sc3g12_c	7	1,308	0.00	0.38	0.00	905	0.00	0.22	0.11
reg122301s_sc3g12_c	8	1,298	0.00	1.07	0.08	891	0.00	1.87	0.00
reg1223020_sc3g12_c	9	1,289	0.00	1.68	0.15	883	0.00	2.42	0.33
reg1223040_sc3g12_c	10	1,298	0.00	0.61	0.53	898	0.00	0.33	0.77
reg122305s_sc3g12_c	11	1,280	0.00	2.51	0.00	886	0.11	2.31	0.00
reg1223060_sc3g12_c	12	1,305	0.00	0.53	0.08	899	0.11	0.33	0.55
reg1226020_sc3g12_c	14	1,298	0.30	0.61	0.23	877	2.09	1.21	0.11
reg1226030_sc3g12_c	15	1,302	0.38	0.38	0.08	876	2.42	0.88	0.22
reg1226040_sc3g12_c	16	1,281	0.76	1.68	0.00	850	3.19	3.08	0.11
reg1226050_sc3g12_c	17	1,286	0.99	1.07	0.00	857	3.96	1.54	0.11
reg1226060_sc3g12_c	18	1,281	1.37	0.99	0.08	850	5.18	0.99	0.22
reg1226080_sc3g12_c	18	1,276	1.75	1.07	0.00	837	6.28	1.54	0.00
reg121602s_sc3g12_c	20	1,248	4.34	0.53	0.00	785	12.56	0.99	0.00
reg121603s_sc3g12_c	21	1,244	5.03	0.23	0.00	770	14.43	0.66	0.11
reg1216040_sc3g12_c	22	1,217	6.09	0.61	0.61	743	16.96	0.77	0.44
reg121605s_sc3g12_c	23	1,192	7.77	1.07	0.08	688	21.26	2.09	0.44
reg1216060_sc3g12_c	24	1,167	10.36	0.76	0.00	665	26.32	0.33	0.11
reg122501s_sc3g12_c	25	1,029	20.26	1.14	0.00	519	41.74	0.88	0.00
reg1225030_sc3g12_c	26	997	23.08	0.76	0.23	504	44.16	0.11	0.22
reg1225060_sc3g12_c	27	961	25.97	0.69	0.15	486	46.37	0.00	0.11
reg1225050_sc3g12_c	28	914	29.40	0.99	0.00	452	49.78	0.44	0.00
reg122504s_sc3g12_c	29	892	31.76	0.15	0.00	413	53.63	0.44	0.00

Note. Pos. = Item position within the difficult test version, n = Number of valid responses, NR = Percentage of respondents that did not reach item, OM = Percentage of respondents that omitted the item, NV = Percentage of respondents with an invalid response.
The item on position 13 was excluded from the analyses due to a significant DIF effect for test difficulty (see section 2).

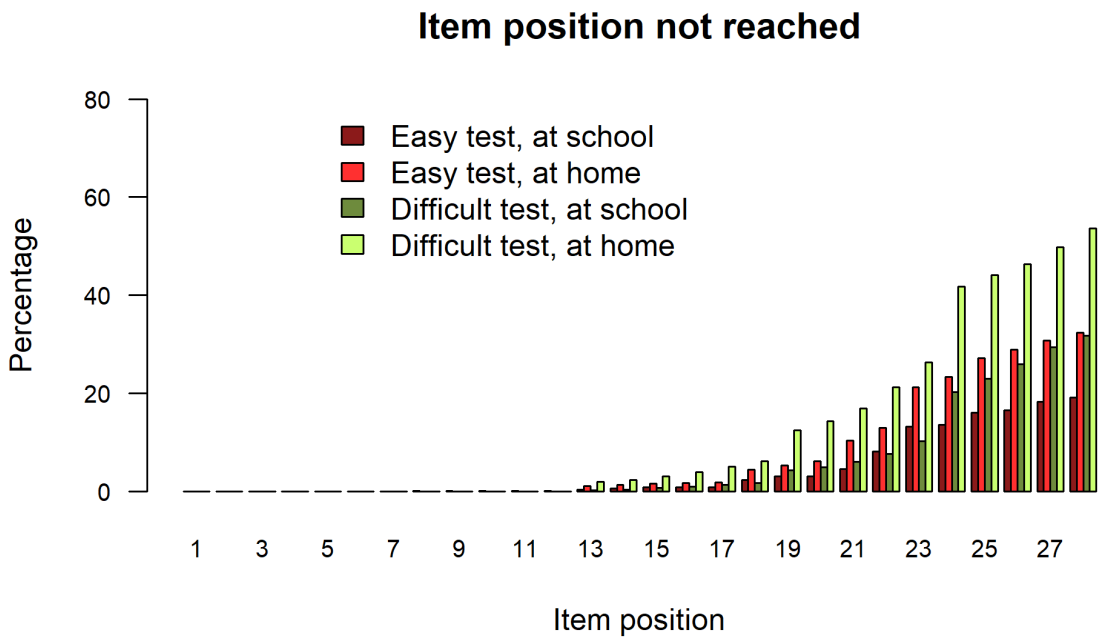

Figure 5. Item position not reached by experimental conditions

5.2 Parameter Estimates

5.2.1 Item parameters

The fourth column in Table 7 presents the percentage of correct responses in relation to all valid responses for each item. Because there is a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index for item difficulty. The percentage of correct responses within dichotomous items varied between 34.3% and 84.6% with an average of 66.4% (*SD* = 13.8%) correct responses.

The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 7. The step parameters for polytomous variables are summarized in Table 8. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties (or location parameters for polytomous variables) ranged from -2.31 (item reg120504s_sc3g12_c) to 1.09 (item reg1229080_sc3g12_c) with an average difficulty of -0.87. 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 (see column 5 in Table 7) were rather small (all *SE*s ≤ 0.08).

Table 7

Item Parameters

	Item	Pos.1	Pos.2	Percentage correct	Item difficulty (SE)	WMNSQ	*t*	r_{it}	Discr.	aQ_3
1.	reg1205010_sc3g12_c	1		65.37	-1.26 (0.06)	1.00	0.10	0.32	1.00	0.03
2.	reg1205020_sc3g12_c	2		52.72	-0.64 (0.06)	1.09	4.18	0.23	0.63	0.03
3.	reg1205030_sc3g12_c	3		81.67	-2.23 (0.07)	0.95	-1.09	0.35	1.50	0.04
4.	reg120504s_sc3g12_c	4		82.65	-2.31 (0.07)	0.95	-1.12	0.35	1.49	0.04
5.	reg1205050_sc3g12_c	5		55.87	-0.79 (0.06)	1.03	1.23	0.29	0.83	0.03
6.	reg1205060_sc3g12_c	6		65.46	-1.26 (0.06)	1.05	2.00	0.24	0.71	0.04
7.	reg1205070_sc3g12_c	7		80.94	-2.18 (0.07)	1.01	0.20	0.28	1.04	0.03
8.	reg1229010_sc3g12_c		1	65.10	-0.37 (0.05)	1.10	4.96	0.16	0.61	0.02
9.	reg1229020_sc3g12_c		2	75.40	-0.94 (0.05)	1.03	1.18	0.25	0.79	0.02
10.	reg1229030_sc3g12_c		3	84.09	-1.55 (0.06)	0.99	-0.36	0.27	0.95	0.03
11.	reg1229060_sc3g12_c		4	68.98	-0.58 (0.05)	1.02	0.95	0.26	0.36	0.03
12.	reg122907s_sc3g12_c		5	n.a.	-0.53 (0.03)	0.96	-1.82	0.33	0.51	0.02
13.	reg1229080_sc3g12_c		6	34.32	1.09 (0.05)	1.05	2.43	0.19	0.62	0.02
14.	reg1229100_sc3g12_c		7	84.59	-1.59 (0.06)	1.01	0.21	0.22	0.56	0.03
15.	reg122301s_sc3g12_c	8	8	67.05	-0.84 (0.04)	1.00	-0.15	0.35	1.05	0.04
16.	reg1223020_sc3g12_c	9	9	59.72	-0.47 (0.04)	1.05	3.53	0.28	1.08	0.03
17.	reg1223040_sc3g12_c	10	10	82.36	-1.81 (0.05)	1.03	1.05	0.26	0.56	0.02
18.	reg122305s_sc3g12_c	11	11	n.a.	-0.55 (0.02)	0.93	-3.75	0.44	0.69	0.04
19.	reg1223060_sc3g12_c	12	12	81.38	-1.73 (0.05)	0.93	-2.83	0.40	0.94	0.03
20.	reg1226020_sc3g12_c	14	14	53.56	-0.16 (0.04)	1.04	3.20	0.29	0.99	0.04

	Item	Pos.1	Pos.2	Percentage correct	Item difficulty (SE)	WMNSQ	*t*	r_{it}	Discr.	aQ_3
21.	reg1226030_sc3g12_c	15	15	77.84	-1.48 (0.04)	0.95	-2.32	0.39	0.41	0.03
22.	reg1226040_sc3g12_c	16	16	67.45	-0.86 (0.04)	1.00	-0.25	0.34	0.75	0.02
23.	reg1226050_sc3g12_c		17	43.58	0.63 (0.05)	1.01	0.82	0.27	1.06	0.02
24.	reg1226060_sc3g12_c	17	18	54.03	-0.19 (0.04)	1.02	1.60	0.32	0.81	0.02
25.	reg1226080_sc3g12_c	18	19	64.68	-0.72 (0.04)	0.99	-0.77	0.34	0.67	0.02
26.	reg121602s_sc3g12_c	19	20	n.a.	-1.08 (0.03)	0.88	-4.83	0.47	1.01	0.02
27.	reg121603s_sc3g12_c	20	21	45.53	0.21 (0.04)	0.97	-2.00	0.37	0.73	0.02
28.	reg1216040_sc3g12_c	21	22	44.73	0.24 (0.04)	1.10	7.04	0.23	0.86	0.02
29.	reg121605s_sc3g12_c	22	23	n.a.	-0.54 (0.02)	0.91	-4.24	0.52	0.68	0.03
30.	reg1216060_sc3g12_c	23	24	61.94	-0.60 (0.04)	1.01	0.72	0.35	1.48	0.03
31.	reg1220010_sc3g12_c	24		80.97	-2.24 (0.08)	0.92	-1.92	0.42	0.77	0.03
32.	reg122002s_sc3g12_c	25		n.a.	-0.64 (0.04)	1.06	2.27	0.19	1.33	0.03
33.	reg1220030_sc3g12_c	26		76.47	-1.95 (0.08)	0.95	-1.33	0.36	1.01	0.02
34.	reg1220040_sc3g12_c	27		70.03	-1.59 (0.07)	0.94	-1.80	0.40	0.85	0.03
35.	reg122005s_sc3g12_c	28		n.a.	-0.17 (0.04)	1.01	0.35	0.29	0.86	0.03
36.	reg1220060_sc3g12_c	29		66.96	-1.44 (0.07)	1.01	0.41	0.32	1.04	0.02
37.	reg122501s_sc3g12_c		25	73.97	-0.88 (0.06)	0.99	-0.18	0.32	1.87	0.06
38.	reg1225030_sc3g12_c		26	82.95	-1.48 (0.07)	1.09	2.05	0.11	0.33	0.02
39.	reg1225060_sc3g12_c		27	54.87	0.09 (0.06)	1.09	4.57	0.19	0.49	0.04
40.	reg1225050_sc3g12_c		28	49.93	0.32 (0.06)	1.06	3.15	0.24	1.43	0.04
41.	reg122504s_sc3g12_c		29	n.a.	-0.53 (0.04)	0.97	-0.86	0.31	1.35	0.03

Note. Pos.1 and Pos.2 = item position within the easy and difficult test versions, respectively. SE = standard error of item difficulty / location parameter. WMNSQ = weighted mean square. t = t-value for WMNSQ. r_{it} = crrected item-total correlation. Discr. = discrimination parameter of a generalized partial credit model. aQ_3 = adjusted average absolute residual correlation for item (Yen, 1993).

Item 13 was excluded from the analyses due to a significant DIF effect for test difficulty (see section 2).

Percent correct scores are not informative for some 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.

Table 8

Step Parameters (with Standard Errors) for Polytomous Items

	Item	Step 1	Step 2	Step 3	Step 4
12.	reg122907s_sc3g12_c	0.03 (0.05)	-0.03		
18.	reg122305s_sc3g12_c	-0.48 (0.03)	0.45 (0.04)	0.03	
26.	reg121602s_sc3g12_c	0.04 (0.04)	-0.04		
29.	reg121605s_sc3g12_c	-0.20 (0.04)	0.07 (0.04)	0.34 (0.05)	-0.21
32.	reg122002s_sc3g12_c	0.13 (0.07)	-0.13		
35.	reg122005s_sc3g12_c	0.06 (0.07)	-0.06		
41.	reg122504s_sc3g12_c	-0.0003 (0.06)	0.0003		

Note. The last step parameter for each item 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 respondents are plotted on the same scale. The distribution of the estimated respondents' ability is mapped onto the left side whereas the right side shows the distribution of item difficulties. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 0.91, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = .80; WLE reliability = .76) was good. The mean of the item distribution was about 0.87 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.

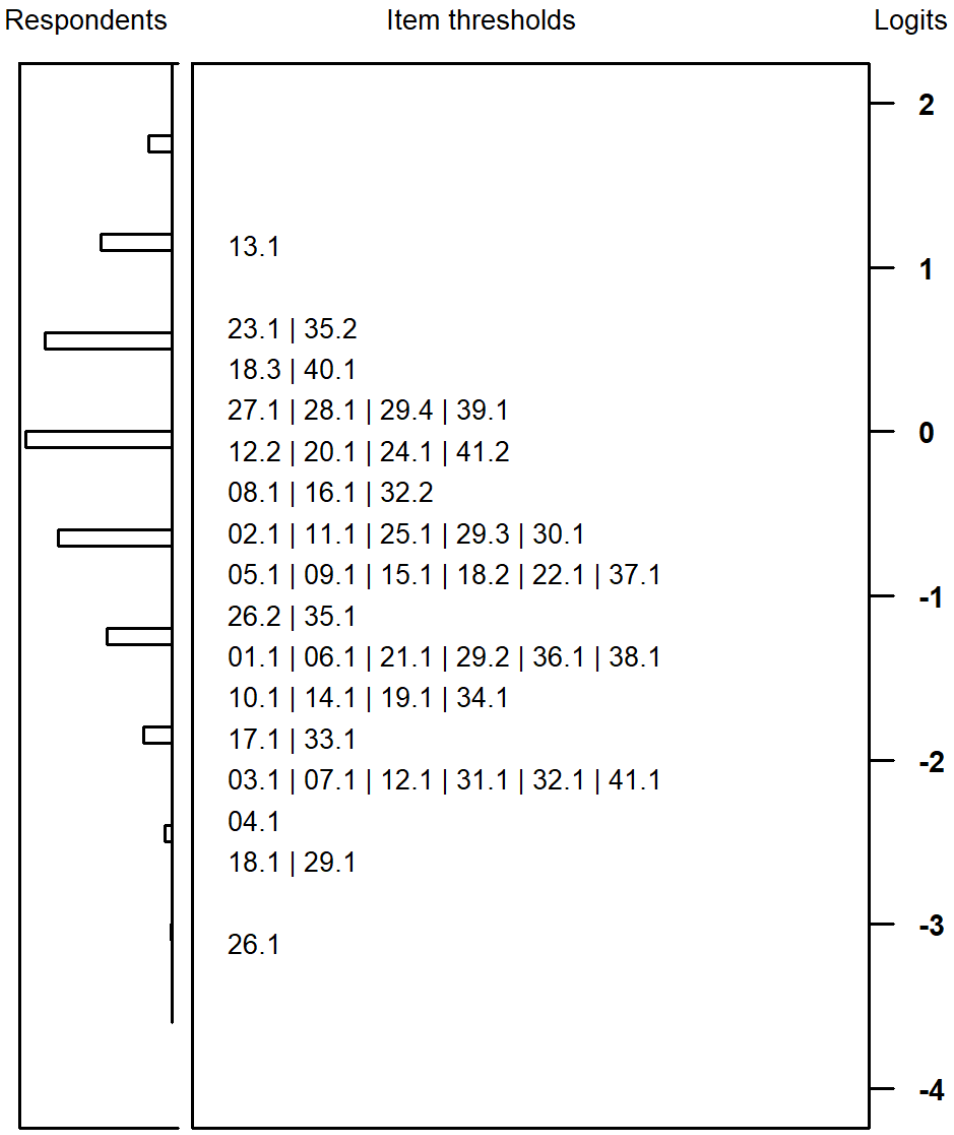

Figure 6. Test targeting. The distribution of person ability in the sample is depicted on the left-hand side of the graph. The difficulty of the items is depicted on the right-hand side of the graph, with each number representing one item (corresponding to Table 7).

5.3 Quality of the test

5.3.1 Fit of the subtasks of complex multiple choice items

Before the subtasks of CMC and MA 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 and MA items separately, there were 62 items. The probability of a correct response ranged from 34% to 94% across all items (*Md* = 75%). Thus, the number of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.88 to 1.14, the respective *t*-value from -6.65 to 6.97, 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 items and polytomous CMC and MA items. Altogether, item fit can be considered very good (see Table 7). Values of the WMNSQ ranged from 0.88 (item reg121602s_sc3g12_c) to 1.10 (item reg1229010_sc3g12_c). Only one item exhibited a *t*-value of the WMNSQ greater than 6 (*t*-value = 7.04 for item reg1216040_sc3g12_c) and none exceeded a value of 8. Thus, there was no indication of severe item over- or underfit. Point-biserial correlations between the item scores and the total scores ranged from .11 (item reg1225030_sc3g12_c) to .52 (item reg121605s_sc3g12_c) and had a mean of .30. Item characteristic curves showed a good fit of all 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 participants' total correct score. The point-biserial correlations for the distractors ranged from -.47 to -.02 with a mean of -.22. 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), migration background, school type, and test position (for a description of these variables, see Pohl & Carstensen, 2012). In addition, the effect of the two experimental factors was also studied. Thus, we compared the two assessment settings (at school or at home) and for the common items that were administered to all participants, we examined measurement invariance for the easy and difficult test. The differences between the estimated item difficulties in the various groups are summarized in Table 9. For example, the column "Male vs. female" reports the differences in item difficulties between men and women; a positive value would indicate that the test was more difficult for males, whereas a negative value would highlight a lower difficulty for males as opposed to females. Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF to those that only estimate main effects (see Table 10).

Table 9

Differential Item Functioning

Item	Gender	Books	Migration	School	Position	Setting	Difficulty
	male vs. female	≤ 100 vs. > 100	without vs. with	no sec. vs. sec	first vs. second	school vs. home	easy vs. difficult
reg1205010_sc3g12_c	0.21 (0.26)	0.20 (0.25)	-0.06 (-0.07)	-0.10 (-0.13)	-0.16 (-0.19)	0.20 (0.26)	
reg1205020_sc3g12_c	0.25 (0.31)	0.19 (0.25)	0.04 (0.05)	0.00 (0.00)	0.20 (0.25)	-0.09 (-0.12)	
reg1205030_sc3g12_c	*0.64* (0.79)	0.24 (0.31)	0.03 (0.04)	-0.24 (-0.32)	-0.14 (-0.17)	0.11 (0.15)	
reg120504s_sc3g12_c	*0.67* (0.82)	0.03 (0.03)	-0.05 (-0.07)	-0.07 (-0.09)	0.10 (0.12)	0.36 (0.48)	
reg1205050_sc3g12_c	-0.24 (-0.29)	0.03 (0.04)	-0.04 (-0.05)	0.15 (0.20)	0.19 (0.23)	0.01 (0.02)	
reg1205060_sc3g12_c	0.12 (0.15)	-0.16 (-0.21)	-0.06 (-0.07)	-0.41 (-0.56)	-0.04 (-0.05)	*0.66* (0.87)	
reg1205070_sc3g12_c	0.34 (0.42)	0.00 (0.01)	0.40 (0.49)	-0.14 (-0.19)	0.01 (0.01)	0.14 (0.18)	
reg1229010_sc3g12_c	0.17 (0.20)	0.11 (0.14)	-0.05 (-0.06)	0.38 (0.52)	-0.02 (-0.02)	-0.38 (-0.51)	
reg1229020_sc3g12_c	-0.01 (-0.01)	0.01 (0.02)	0.00 (0.00)	0.21 (0.29)	0.08 (0.10)	-0.25 (-0.33)	
reg1229030_sc3g12_c	-0.22 (-0.27)	0.04 (0.05)	-0.36 (-0.44)	0.03 (0.04)	0.05 (0.06)	-0.04 (-0.06)	
reg1229060_sc3g12_c	0.06 (0.08)	0.01 (0.01)	0.08 (0.10)	0.14 (0.20)	-0.03 (-0.04)	-0.09 (-0.12)	
reg122907s_sc3g12_c	0.02 (0.02)	0.15 (0.20)	0.08 (0.10)	0.34 (0.46)	0.08 (0.10)	-0.25 (-0.33)	
reg1229080_sc3g12_c	0.17 (0.21)	-0.34 (-0.43)	-0.38 (-0.47)	-0.38 (-0.51)	-0.12 (-0.14)	0.33 (0.44)	
reg1229100_sc3g12_c	0.02 (0.02)	0.07 (0.09)	0.26 (0.32)	-0.12 (-0.16)	0.05 (0.07)	0.16 (0.22)	
reg122301s_sc3g12_c	0.09 (0.11)	-0.05 (-0.07)	-0.17 (-0.22)	0.20 (0.28)	0.07 (0.09)	-0.18 (-0.24)	-0.01 (-0.01)
reg1223020_sc3g12_c	-0.28 (-0.34)	0.19 (0.25)	-0.14 (-0.17)	0.35 (0.47)	0.03 (0.04)	-0.30 (-0.40)	-0.13 (-0.18)
reg1223040_sc3g12_c	-0.13 (-0.16)	0.09 (0.12)	-0.27 (-0.34)	-0.09 (-0.13)	0.14 (0.17)	0.07 (0.09)	0.05 (0.08)
reg122305s_sc3g12_c	0.14 (0.18)	-0.14 (-0.18)	0.04 (0.05)	-0.06 (-0.08)	0.01 (0.01)	-0.01 (-0.01)	-0.25 (-0.34)
reg1223060_sc3g12_c	-0.01 (-0.01)	-0.01 (-0.01)	-0.11 (-0.14)	0.09 (0.12)	0.09 (0.11)	-0.06 (-0.08)	0.30 (0.41)

Item	Gender	Books	Migration	School	Position	Setting	Difficulty
	male vs. female	≤ 100 vs. > 100	without vs. with	no sec. vs. sec	first vs. second	school vs. home	easy vs. difficult
reg1226020_sc3g12_c	0.37 (0.46)	0.20 (0.25)	-0.01 (-0.01)	0.17 (0.23)	-0.13 (-0.15)	0.03 (0.04)	-0.27 (-0.37)
reg1226030_sc3g12_c	-0.18 (-0.23)	-0.06 (-0.07)	-0.07 (-0.09)	-0.33 (-0.44)	0.23 (0.28)	0.27 (0.36)	0.24 (0.33)
reg1226040_sc3g12_c	-0.05 (-0.06)	0.10 (0.13)	-0.12 (-0.15)	0.32 (0.43)	-0.07 (-0.08)	-0.54 (-0.72)	0.27 (0.36)
reg1226050_sc3g12_c	0.19 (0.23)	0.33 (0.42)	-0.04 (-0.04)	0.28 (0.38)	0.04 (0.05)	-0.33 (-0.44)	
reg1226060_sc3g12_c	-0.09 (-0.11)	-0.01 (-0.01)	-0.02 (-0.03)	-0.11 (-0.15)	0.05 (0.06)	0.05 (0.07)	-0.19 (-0.26)
reg1226080_sc3g12_c	-0.28 (-0.34)	0.23 (0.29)	-0.24 (-0.30)	0.19 (0.26)	0.14 (0.17)	-0.15 (-0.20)	0.57 (0.79)
reg121602s_sc3g12_c	-0.28 (-0.35)	0.21 (0.27)	-0.08 (-0.09)	0.25 (0.34)	-0.02 (-0.03)	-0.15 (-0.20)	-0.03 (-0.04)
reg121603s_sc3g12_c	-0.07 (-0.08)	-0.07 (-0.09)	-0.01 (-0.02)	-0.09 (-0.12)	-0.08 (-0.09)	0.11 (0.15)	0.20 (0.28)
reg1216040_sc3g12_c	*-0.61* (-0.76)	0.18 (0.24)	-0.26 (-0.33)	0.21 (0.29)	0.13 (0.16)	-0.31 (-0.41)	-0.36 (-0.50)
reg121605s_sc3g12_c	0.04 (0.04)	-0.14 (-0.18)	-0.05 (-0.06)	-0.08 (-0.11)	-0.04 (-0.05)	0.04 (0.05)	-0.26 (-0.36)
reg1216060_sc3g12_c	0.06 (0.08)	0.35 (0.45)	-0.29 (-0.35)	0.38 (0.51)	-0.11 (-0.13)	-0.52 (-0.69)	-0.15 (-0.20)
reg1220010_sc3g12_c	0.11 (0.13)	0.04 (0.05)	-0.16 (-0.19)	-0.14 (-0.19)	0.03 (0.04)	0.12 (0.16)	
reg122002s_sc3g12_c	-0.11 (-0.13)	-0.58 (-0.75)	0.19 (0.24)	-0.36 (-0.49)	-0.28 (-0.34)	0.36 (0.48)	
reg1220030_sc3g12_c	0.02 (0.03)	-0.21 (-0.27)	0.03 (0.04)	-0.03 (-0.03)	0.07 (0.08)	0.04 (0.05)	
reg1220040_sc3g12_c	-0.15 (-0.19)	-0.15 (-0.20)	*0.60* (0.74)	0.04 (0.05)	-0.04 (-0.04)	-0.05 (-0.07)	
reg122005s_sc3g12_c	0.09 (0.11)	-0.16 (-0.21)	0.28 (0.34)	-0.52 (-0.70)	0.30 (0.37)	0.40 (0.53)	
reg1220060_sc3g12_c	-0.43 (-0.53)	-0.10 (-0.12)	0.12 (0.14)	-0.05 (-0.06)	-0.07 (-0.09)	0.02 (0.03)	
reg122501s_sc3g12_c	0.01 (0.02)	-0.06 (-0.08)	-0.20 (-0.25)	0.15 (0.20)	-0.27 (-0.33)	-0.25 (-0.33)	
reg1225030_sc3g12_c	-0.28 (-0.35)	-0.18 (-0.22)	0.27 (0.33)	-0.12 (-0.17)	-0.33 (-0.40)	0.12 (0.16)	
reg1225060_sc3g12_c	-0.20 (-0.24)	-0.11 (-0.14)	0.21 (0.26)	-0.18 (-0.24)	-0.12 (-0.14)	0.15 (0.20)	
reg1225050_sc3g12_c	-0.30 (-0.36)	-0.16 (-0.21)	0.12 (0.15)	-0.17 (-0.23)	-0.01 (-0.01)	0.06 (0.09)	

Item	Gender	Books	Migration	School	Position	Setting	Difficulty
	male vs. female	≤ 100 vs. > 100	without vs. with	no sec. vs. sec	first vs. second	school vs. home	easy vs. difficult
reg122504s_sc3g12_c	0.10 (0.13)	-0.31 (-0.40)	0.50 (0.61)	-0.12 (-0.16)	-0.07 (-0.08)	0.14 (0.19)	
Main effect (with DIF)	-0.30 (-0.37)	-0.52 (-0.66)	0.33 (0.41)	-0.69 (-0.93)	0.14 (0.17)	0.61 (0.82)	-1.07 (-1.48)
Main effect (without DIF)	-0.28 (-0.34)	-0.49 (-0.62)	0.36 (0.44)	-0.70 (-0.94)	0.13 (0.16)	0.63 (0.84)	-1.00 (-1.36)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. The differences in item difficulty parameters larger than 0.60 logits are indicated in italics. Sec. = Secondary school (German: „Gymnasium").

All absolute standardized differences are not significantly greater than 0.4 (α = 5%; see Fischer et al., 2016).

Gender. The sample included 1,826 (50%) males and 1,836 (50%) females. On average, male participants had a lower estimated reading ability than females (main effect = -0.30 logits, Cohen's *d* = -0.37). However, three items showed DIF greater than 0.6 logits (DIF = 0.64 for item reg1205030_sc3g12_c, DIF = 0.67 for item reg120504s_sc3g12_c, DIF = -0.61 for item reg1216040_sc3g12_c). An overall test for DIF (see Table 10) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). A model comparison using Akaike's (1974) information criterion (AIC) favored the model estimating DIF, whereas the Bayesian information criterion (BIC; Schwarz, 1978) that takes the number of estimated parameters into account and, thus, guards against overparameterization of models, indicated a better fit for the more parsimonious model including only the main effect. Thus, overall, there was no pronounced DIF with regard to gender.

Number of books. The number of books at home was used as a proxy for socioeconomic status. There were 1,375 (38%) individuals with up to 100 books at home, 2,253 (62%) individuals with more than 100 books at home. Thirty-four individuals (0.01%) that did not report the number of books at home. were excluded from the analysis. There were considerable average differences between the two groups. Participants with up to 100 books at home performed on average -0.52 logits (Cohen's *d* = -0.66) lower in reading than participants with more than 100 books. There was no considerable DIF comparing participants with many or fewer books (the highest DIF = |0.58| for item reg122002s_sc3g12_c). As a consequence, also the overall test for DIF using the BIC favored the main effects model (see Table 10).

Migration background. There were 3,184 participants (87%) with no migration background, 469 subjects (13%) with a migration background, and 9 individuals (0.002%) that did not report their migration background. In comparison to subjects with migration background, participants without migration background had, on average, a slightly higher reading ability (main effect = 0.33 logits, Cohen's *d* = 0.41). There was no noteworthy item DIF due to migration background; differences in estimated difficulties did not exceed 0.6 logits (with an

exception of the item reg1220040_sc3g12_c which indicated DIF = 0.60 logits). Moreover, the overall test for DIF using the AIC and BIC also favored the main effects model.

Table 10

Comparisons of Models with and without DIF

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Gender	DIF	3,662	120,979	96	*121,171*	121,767
	main effect	3,662	121,222	56	121,334	*121,681*
Books	DIF	3,628	119,870	96	*120,062*	120,657
	main effect	3,628	119,995	56	120,107	*120,454*
Migration	DIF	3,653	120,917	96	121,109	121,705
	main effect	3,653	120,988	56	*121,100*	*121,447*
School	DIF	3,662	120,511	96	*120,703*	121,299
	main effect	3,662	120,732	56	120,844	*121,192*
Position	DIF	3,662	121,219	96	121,411	122,007
	main effect	3,662	121,285	56	*121,397*	*121,745*
Setting	DIF	3,662	120,618	96	*120,810*	121,405
	main effect	3,662	120,836	56	120,948	*121,296*
Difficulty	DIF	3,662	70,042	40	*70,122*	*70,371*
	main effect	3,662	70,242	26	70,294	70,456

Note. The AIC and BIC values of the best fitting model are shown in italics.

School type. Overall, 1,991 subjects (54%) who took the reading test attended secondary school (German: "Gymnasium") whereas 1,672 (46%) were enrolled in other school types. Subjects in secondary schools showed a higher reading ability on average (0.69 logits, Cohen's *d* = 0.93) than subjects in other school types. There was no noteworthy item DIF (with an exception of two items, item reg122005s_sc3g12_c with DIF = |0.52| logits and item reg1205060_sc3g12_c with DIF = |0.41| logits); no item exhibited DIF greater than 0.6 logits. The overall test for DIF using the BIC favored the main effects model.

Test position. The reading competence test was administered in two different positions (see section 3.1 for the design of the study). A subsample of 1,837 (50%) persons received the reading test first and 1,825 (50%) respondents took the reading test after having completed either the computer literacy or the mathematics test. Differential item functioning of the position of the test may, for example, occur if there are differential fatigue effects for certain items. The results show minor average effect of item position[2]. Subjects who received the reading test first performed on average 0.14 logits (Cohen's *d* = 0.17) better than subjects

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

who received the reading test second. There was no DIF due to the position of the test in the booklet. The largest difference in difficulty between the two design groups was |0.33| logits (item reg1225030_sc3g12_c). As a consequence, the overall test for DIF using the AIC and BIC favored the more parsimonious main effect model.

Assessment setting. The reading competence test was administered in two different settings (see section 3.1 for the design of the study). A subsample of 1,766 (48%) persons received the reading test in small groups at school, whereas 1,896 (52%) participants finished the test individually at their private homes. Subjects who finished the reading test at school were on average 0.61 logits (Cohen's *d* = 0.82) better than those working at their private homes. However, this difference must not be interpreted as a causal effect of the administration setting because respondents were not randomly assigned to the different settings. Rather, it is likely that self-selection processes occurred, for example, because less proficient individuals were more likely to leave school and, consequently, were tested at home. More importantly, there was no noteworthy DIF due to the administration setting; all differences in item difficulties were smaller than 0.6 logits (with an exception of item reg1205060_sc3g12_c which indicated DIF = 0.66 logits). Using the BIC, the overall model test indicated a better fit for the main effect model (see Table 10), indicating that DIF effects between 0.4 and 0.6 found for a few items were not considered severe.

Test difficulty. To estimate the participants' proficiency with great accuracy the participants received different tests that either included a larger number of easy or difficult items (see section 3.1 for the design of the study). Only a subset of 15 items that were included in both tests was administered to all participants. For these common items we examined potential DIF across the two test versions (easy versus difficult). A subsample of 1,441 (39%) persons received the easy test and 2,221 (61%) persons received the difficult test. As expected, subjects who were administered the easy test scored on average -1.07 logits (Cohen's *d* = -1.48) 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 0.57 logits (item reg1226080_sc3g12_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 (GPCM) that estimates discrimination parameters was fitted to the data. As reported in Table 7, the estimated discrimination parameters differed moderately among items, ranging from 0.33 (item reg1225030_sc3g12_c) to 1.87 (item reg122501s_sc3g12_c). The average discrimination parameter fell at 0.90. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 120,720, BIC = 121,291, number of parameters = 92) as compared to the Rasch model AIC = 121,492, BIC = 121,815, number of parameters = 52). Despite the empirical preference for the GPCM, the Rasch model more adequately matches the theoretical conceptions underlying the test construction (for a discussion of this issue, see Pohl & Carstensen, 2012; 2013). For this reason, the partial credit model (1PL) was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

5.3.6 Unidimensionality

The unidimensionality of the test was investigated by specifying two different multidimensional models and comparing them to a unidimensional model. In the first

multidimensional model, three different cognitive requirements were specified, whereas the five different text types constituted the second multidimensional model. Estimation of the models was carried out in R using Gauss-Hermite quadrature method.

The estimated variances and correlations between the three dimensions representing the different cognitive requirements are reported in Table 11. The correlations among the three dimensions were rather high and fell between .93 and .95. 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 = 121,394, BIC = 121,747, number of parameters = 57) than the unidimensional model (AIC = 121,491, BIC = 121,814, number of parameters = 52). These results indicate that the three cognitive requirements measure a common construct, albeit it is not completely unidimensional.

Table 11

Results of Three-Dimensional Scaling

	Dim. 1	**Dim. 2**	**Dim. 3**
Dim. 1: Finding information in the text (9 items)	(1.46)		
Dim. 2: Drawing text-related conclusions (13 items)	.94	(0.80)	
Dim. 3: Reflecting and assessing (19 items)	.95	.93	(0.84)

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

The estimated variances and correlations of the five-dimensional model based on the five text functions are given in Table 12. The correlations between the dimensions varied between *r* = .71 and *r* = .89. The smallest correlation was found between Dimension 3 (“Commenting”) and Dimension 5 (“information”). Dimension 2 (“Instruction”) and Dimension 4 (“Advertising”) showed the strongest correlation. All correlations deviated from a perfect correlation (i.e., *r* < .95, see Carstensen, 2013). Moreover, the five-dimensional model (AIC = 121,260, BIC = 121,670, number of parameters = 66) fitted the data better than the unidimensional model (AIC = 121,491, BIC = 121,814, number of parameters = 52). As each text function corresponded to one of the five texts, local item dependence (LID) and the text functions were confounded. As a consequence, the deviation of the correlations from a perfect correlation shown in Table 12, may result from multidimensionality and 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 respondents, so that the impact of text functions could be investigated independently of LID. The correlations estimated in the pilot study ranged from .78 to .91. As the correlations found in Gehrer and colleagues (2013) differ from a perfect correlation, it is concluded that text functions form subdimensions of reading competence. Comparing the correlations found in Gehrer et al. (2013), which are due to text functions, to those found in the main study (Table 12), which are due to both text functions and LID, allows us to

evaluate the impact of LID. As reported in Table 12, the correlations found in the present study of starting cohort 3 were slightly lower than those found in Gehrer et al. (2013), indicating that there is some amount of local item dependence. However, according to the test developers a balanced assessment of reading competence can only be achieved by a heterogeneity of text functions (Gehrer et al., 2013).

However, for the unidimensional model the average absolute residual correlations as indicated by the adjusted Q_3 statistic (see Table 7) were quite low (M = 0.03, SD = 0.01) - the largest individual residual correlation was 0.06 - and thus indicated an essentially unidimensional test. Because the reading test is constructed to measure a single dimension, a unidimensional reading competence score was estimated.

Table 12

Results of Five-Dimensional Scaling

	Dim. 1	**Dim. 2**	**Dim. 3**	**Dim. 4**	**Dim. 5**
Dim. 1: Literary (14 items)	(0.92)				
Dim. 2: Instruction (5 items)	.78	(1.29)			
Dim. 3: Commenting (6 items)	.76	.83	(1.18)		
Dim. 4: Advertising (5 items)	.81	.89	.77	(1.42)	
Dim. 5: Information (11 items)	.77	.79	.71	.75	(1.03)

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 12 and at describing how the reading competence score was estimated.

We investigated different kinds of missing responses and examined the item and test parameters. We thoroughly checked item fit statistics for simple MC items, subtasks of CMC 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 amount of not-reached items was rather high, indicating that the test was too long for the allocated testing time. Specifically, it referred to respondents who received the difficult test at home. They did not reach the last of the five texts and therefore showed more not-reached missing responses. Other types of missing responses were reasonably small.

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

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 participants with no valid responses to some of the text functions), the estimation of a single reading competence score is challenged. This should be addressed in further studies. Nevertheless, Gehrer et al. (2013) argue that a balanced assessment of reading competence can only be achieved by heterogeneity of text functions and they provide theoretical arguments for a unidimensional measure of reading competence.

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

7. Data in the Scientific Use File

7.1 Naming conventions

The data in the Scientific Use File contain 42 items, of which 30 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. A total of 12 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 items end in 's_c'. Items containing the suffix sc3_g12_ have originally been administrated in starting cohort 4, grade 12 (for details on the naming conventions of the variables, see Fuß et al., 2016). In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category.

7.2 Linking of competence scores

In starting cohort 3, the reading competence tests administered in grades 9 (see Scharl et al., 2017) and 12 include different items that were constructed in such a way as to allow for an accurate measurement of reading competence within each age group. As a consequence, the competence scores derived in the different grades cannot be directly compared because differences in observed scores would reflect both differences in competences and 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 method of "mean/mean" linking for the anchor-group design as described by Fischer, Rohm, Gnambs, & Carstensen, 2016. Following an anchor-group design, an independent link sample including students from grade 11 that were not part of starting cohort 3 were administered all items from the grade 9 and the grade 12 reading competence tests within a single measurement occasion. These responses were used to link the two tests administered in starting cohort 3 across the two grades.

7.2.1 Samples

In starting cohort 3, a subsample of 2,703 students participated at both measurement occasions, in grade 9 and also in grade 12 (so called longitudinal main subsample from the

starting cohort). Consequently, these respondents were used to link the two tests across both grades (see Fischer et al., 2016). Moreover, an independent link sample of *N* = 935 students (448 women) from grade 11 received both tests within a single measurement occasion.

7.2.2 The design of the link study

The students of the link study responded to 18 common items from the easy and difficult test versions administrated in grade 9 (see Scharl et al., 2017) and to 40 items of the grade 12 reading test (see above)[3]. Again, two versions of the grade 12 test were used in the link study (easy and difficult). A random sample of 464 students received the easy test version and 471 students were administered the difficult version. Moreover, the reading test was administered at different positions in the test battery. A random sample of 476 students received the reading test before working on a mathematics test, whereas the remaining 459 students received the mathematics test before the reading test. No multi-matrix design regarding the selection and order of the items within a test was established. Thus, all participants were given the reading items in the same order.

7.2.3 Results

To examine whether the two tests administered in the link sample measured a common scale, we compared a one-dimensional model that specified a single latent factor for all items to a two-dimensional model that specified separate latent factors for the two tests. The information criteria slightly favored the two-dimensional model (AIC = 38,159; BIC = 38,513; number of parameters = 73) over the one-dimensional model (AIC = 38,201; BIC = 38,545; number of parameters = 71). However, an examination of the residual correlations for the one-dimensional model using the adjusted Q_3 statistic (Yen, 1984) indicated a largely unidimensional scale — the average absolute residual correlation was *M* = 0.05 (*SD* = 0.01, *Max* = 0.08). This indicates that the reading competence tests administered in grades 9 and 12 were essentially unidimensional.

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal main 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 13. A positive value for the difference in item difficulty parameters indicates that the item is more difficult for the linking sample compared to the longitudinal main subsample, whereas a negative value indicates a lower difficulty for the link sample. Minimum effect hypothesis tests revealed no significant DIF for any item (α = .05). However, a couple of items exhibited considerable DIF greater that 0.4 logits and two items indicated strong DIF that was close to 1 logit or larger (*Max* = |1.42|). This concerns five items from the grade 9 test (reg90210_c, reg90230_c, reg90250_c, reg90460_c, reg90560_c) and eight items from the grade 12 test (reg1205020_c, reg1205030_c, reg120504s_c, reg121603s_c, reg1220030_c, reg122501s_c, reg1225060_c, reg122504s_c). Therefore, these items were removed from the final linking procedure.

[3] Note that due to problematic DIF the item “reg90240_sc3g9_c” and the item “reg12307s_sc3g12_c” were eliminated from the grade 9 test and the grade 12 test, respectively.

Table 13

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

	Grade 9				Grade 12			
	Item	**Δσ**	***SE*$_{Δσ}$**	***F***	**Item**	**Δσ**	***SE*$_{Δσ}$**	***F***
1	reg90210_c	*0.63*	0.14	20.23	reg1205010_c	-0.21	0.14	2.23
2	reg90220_c	0.37	0.10	14.66	reg1205020_c	*-0.58*	0.14	18.14
3	reg90230_c	*-0.54*	0.15	13.05	reg1205030_c	*-0.53*	0.20	7.45
4	reg90250_c	*0.53*	0.09	37.92	reg120504s_c	*-1.42*	0.28	25.19
5	reg90410_c	0.09	0.16	0.32	reg1205050_c	-0.21	0.13	2.47
6	reg90420_c	-0.28	0.12	5.01	reg1205060_c	-0.26	0.14	3.43
7	reg90430_c	0.10	0.11	0.82	reg1205070_c	0.21	0.16	1.65
8	reg90440_c	0.03	0.13	0.05	reg122301s_c	0.03	0.09	0.08
9	reg90450_c	0.29	0.13	4.84	reg1223020_c	0.02	0.14	2.19
10	reg90460_c	*0.54*	0.10	30.40	reg1223040_c	0.03	0.14	17.64
11	reg9047s_c	0.36	0.07	25.44	reg122305s_c	-0.03	0.20	7.33
12	reg90510_c	-0.17	0.11	2.47	reg1223060_c	0.11	0.28	24.98
13	reg90520_c	-0.09	0.11	0.65	reg1226020_c	0.27	0.13	2.40
14	reg90530_c	-0.28	0.11	6.04	reg1226030_c	0.19	0.14	3.35
15	reg90540_c	-0.40	0.11	13.34	reg1226040_c	0.23	0.17	1.61
16	reg90550_c	-0.39	0.11	11.53	reg1226060_c	-0.04	0.09	0.08
17	reg90560_c	*-0.51*	0.13	15.63	reg1226080_c	0.16	0.09	0.03
18	reg90570_c	-0.29	0.16	3.38	reg121602s_c	-0.09	0.11	0.05
19					reg121603s_c	*0.48*	0.04	0.61
20					reg1216040_c	-0.11	0.11	1.12
21					reg121605s_c	-0.05	0.09	9.78
22					reg1216060_c	0.07	0.10	3.46
23					reg1220010_c	-0.30	0.09	6.36
24					reg122002s_c	0.10	0.09	0.20
25					reg1220030_c	*-0.56*	0.09	2.94
26					reg1220040_c	-0.12	0.07	1.45
27					reg122005s_c	-0.06	0.09	26.07
28					reg1220060_c	0.18	0.10	1.43
29					reg1229010_c	-0.24	0.04	1.32
30					reg1229020_c	0.04	0.10	0.48

	Grade 9				Grade 12			
	Item	**Δσ**	***SE*$_{Δσ}$**	***F***	**Item**	**Δσ**	***SE*$_{Δσ}$**	***F***
31					reg1229030_c	0.28	0.24	1.47
32					reg1229060_c	0.16	0.09	1.03
33					reg122907s_c	0.06	0.25	5.16
34					reg1229080_c	0.03	0.20	0.34
35					reg1229100_c	0.17	0.10	0.36
36					reg122501s_c	*0.64*	0.19	0.90
37					reg1225030_c	-0.18	0.12	3.83
38					reg1225060_c	*0.58*	0.13	0.12
39					reg1225050_c	-0.04	0.14	3.89
40					reg122504s_c	*0.98*	0.12	1.67

Note. Δσ = difference in item difficulty parameters between the longitudinal main subsample in grade 9 or 12 and the link sample (positive values indicate that items were more difficult for the link sample); *SE*Δσ = pooled standard error; *F* = test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The differences in item difficulty parameters larger than 0.40 logits are indicated in italics. The critical value for the minimum effects hypothesis test using an α of 5% is F_{0154} (2; 3,641) = 42.86. A non-significant test indicates measurement invariance.

To apply the mean/mean linking method, the correction term was calculated as *c* = 0.4989. Added to the correction term for grade 5 to 9 (see Scharl et al., 2017), a total correction term of 1.7459 was derived. This correction term was subsequently added to each difficulty parameter estimated in grade 12 (see Table 7) to derive the linked item parameters (see Appendix, part C). The link error reflecting the uncertainty in the linking process was calculated according to equation 4 in Fischer et al. (2016) as 0.0816 and has to be included into the *SE* when statistical tests are used to compare groups concerning their mean change of ability between two linked measurements.

7.3 Reading competence scores

In the SUF, manifest reading competence scores are provided in the form of two different WLEs, “reg12_sc1” and “reg12_sc1u”, including their respective standard error, “reg12_sc2” and “reg12_sc2u”. The WLE scores in “reg12_sc1”are were corrected for differences in the position of reading test within the booklet. In grade 9, the reading test was always presented first within the test battery, whereas in grade 12 the reading test was either presented as the first or the second test within the test battery (see page 5). To correct for differences in the test position, we re-estimated the WLE scores by including the test position variable in the IRT scaling model. As a consequence, they can be used only for cross-sectional research questions but not if the focus of research lies on longitudinal issues. (Note that the WLE scores in “reg12_sc1” are not linked to the underlying reference scale of grade 9.) In contrast, WLE scores in “reg12_sc1u” were estimated using the linked item difficulty parameters (they are uncorrected for the position of the reading test within the booklet). As a result, these WLE scores can be used for longitudinal comparisons from grade 5 to grade 12. The resulting differences in WLE scores can be interpreted as development trajectories

across measurement points. 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.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-722. http://doi.org/10.1007/978-1-4612-1694-0_16

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. http://doi.org/10.1007/978-94-007-4458-5_12

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. Available online at: https://www.neps-data.de/Portals/0/Survey%20Papers/SP_I.pdf

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2019). *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. Available online at: https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/1-0-0/com_re_2012_en.pdf

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. Available online at: http://www.j-e-r-o.com/index.php/jero/article/viewFile/361/170

Krannich, M., Odin, J., Rohm, T., Koller, I., Pohl, S., Haberkorn, K., Carstensen, C. H., Fischer, L., & Gnambs, T. (2017). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 3 for Grade 7* (NEPS Survey Paper No. 15). Bamberg: Leibniz Institute

for Educational Trajectories, National Educational Panel Study. Available online at: https://www.neps-data.de/Portals/0/Survey%20Papers/Update_SP_XIV.pdf

Robitzsch, A., Kiefer, T., & Wu, M. (2019). *TAM: Test analysis modules. R package version 3.3-10.* Available online at: https://CRAN.R-project.org/package=TAM

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

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

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

Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement, 50*, 447-468. http://doi.org/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. Available online at: https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIV.pdf

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. Available online at: https://www.pedocs.de/volltexte/2013/8430/pdf/JERO_2013_2_Pohl_Carstensen_Scaling_of_competence_tests.pdf

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. Available online at: https://www.neps-data.de/Portals/0/Working%20Papers/WP_XV.pdf

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

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available online at: https://www.R-project.org/

Scharl, A., Fischer, L., Gnambs, T. and 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. Available online at: https://www.neps-data.de/Portals/0/Survey%20Papers/SP_XX.pdf

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450. https://doi.org/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. http://doi.org/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. http://doi.org/10.1177/014662168400800201

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

Appendix

Part A. Overview of the Items in the Reading Test

Item	Pos.1	Pos.2	Text type	Cognitive requirement	Response format
reg1205010_sc3g12_c	1		literary	Reflecting and assessing	MC
reg1205020_sc3g12_c	2		literary	Reflecting and assessing	MC
reg1205030_sc3g12_c	3		literary	Drawing conclusions	MC
reg120504s_sc3g12_c	4		literary	Drawing conclusions	CMC
reg1205050_sc3g12_c	5		literary	Drawing conclusions	MC
reg1205060_sc3g12_c	6		literary	Drawing conclusions	MC
reg1205070_sc3g12_c	7		literary	Reflecting and assessing	MC
reg1229010_sc3g12_c		1	literary	Drawing conclusions	MC
reg1229020_sc3g12_c		2	literary	Finding information	MC
reg1229030_sc3g12_c		3	literary	Drawing conclusions	MC
reg1229060_sc3g12_c		4	literary	Drawing conclusions	MC
reg122907s_sc3g12_c		5	literary	Finding information	CMC
reg1229080_sc3g12_c		6	literary	Reflecting and assessing	MC
reg1229100_sc3g12_c		7	literary	Drawing conclusions	MC
reg122301s_sc3g12_c	8	8	instruction	Finding information	CMC
reg1223020_sc3g12_c	9	9	instruction	Drawing conclusions	MC
reg1223040_sc3g12_c	10	10	instruction	Reflecting and assessing	MC
reg122305s_sc3g12_c	11	11	instruction	Finding information	CMC
reg1223060_sc3g12_c	12	12	instruction	Reflecting and assessing	MC
reg122307s_sc3g12_c[1]	13	13	instruction	Reflecting and assessing	MA
reg1226020_sc3g12_c	14	14	commenting	Reflecting and assessing	MC
reg1226030_sc3g12_c	15	15	commenting	Reflecting and assessing	MC
reg1226040_sc3g12_c	16	16	commenting	Finding information	MC
reg1226050_sc3g12_c		17	commenting	Reflecting and assessing	MC
reg1226060_sc3g12_c	17	18	commenting	Reflecting and assessing	MC
reg1226080_sc3g12_c	18	19	commenting	Reflecting and assessing	MC
reg121602s_sc3g12_c	19	20	advertising	Finding information	CMC
reg121603s_sc3g12_c	20	21	advertising	Finding information	CMC
reg1216040_sc3g12_c	21	22	advertising	Reflecting and assessing	MC
reg121605s_sc3g12_c	22	23	advertising	Reflecting and assessing	MA
reg1216060_sc3g12_c	23	24	advertising	Drawing conclusions	MC
reg1220010_sc3g12_c	24		information	Drawing conclusions	MC
reg122002s_sc3g12_c	25		information	Drawing conclusions	CMC
reg1220030_sc3g12_c	26		information	Reflecting and assessing	MC
reg1220040_sc3g12_c	27		information	Reflecting and assessing	MC
reg122005s_sc3g12_c	28		information	Finding information	CMC
reg1220060_sc3g12_c	29		information	Reflecting and assessing	MC
reg122501s_sc3g12_c		25	information	Finding information	CMC
reg1225030_sc3g12_c		26	information	Reflecting and assessing	MC
reg1225060_sc3g12_c		27	information	Reflecting and assessing	MC
reg1225050_sc3g12_c		28	information	Reflecting and assessing	MC
reg122504s_sc3g12_c		29	information	Drawing conclusions	CMC

Note. Pos.1 and Pos.2 = Item position within the easy and difficult test versions, respectively. MC = multiple-choice item; CMC = polytomous items with complex multiple-choice response format; MA = matching item.

[1] The item on the position 13 in both the easy and difficult test versions was excluded from the analyses due to a significant DIF effect for test difficulty (see section 2).

Part B. R Syntax for fitting the partial credit model as a scaling model in starting cohort 3 grade 12

```
library(haven) # contains read_sav-function
library(doBy)  # contains recodeVar-function
library(TAM)   # contains tam.mml- and tam.wle-functions
library(dplyr) # contains %>%-funcion

### load data
dat <- read_sav(file = "SC3_xTargetCompetencies_D_9-0-0.sav")
reading.items <- c( [add the items provided in Appendix, part A without
                     item "reg122307s_sc3g12_c"] )

### Collapse response categories with N < 200
dat$reg120504s_sc3g12_c <- recodeVar(dat$reg120504s_c_sc3g12,
                                     c(-97, -94, -54, 0, 1, 2),
                                     c(-97, -94, -54, 0, 0, 1))
dat$reg122301s_sc3g12_c <- recodeVar(dat$reg122301s_sc3g12_c,
                                     c(-97, -94, -54, 0, 1, 2),
                                     c(-97, -94, -54, 0, 0, 1))
dat$reg121603s_sc3g12_c <- recodeVar(dat$reg121603s_sc3g12_c,
                                     c(-97, -94, -54, 0, 1, 2),
                                     c(-97, -94, -54, 0, 0, 1))
dat$reg122305s_c_sc3g12 <- recodeVar(dat$reg122305s_c_sc3g12,
                                     c(-97, -95, -94, -55, 0, 1, 2, 3, 4),
                                     c(-97, -95, -94, -55, 0, 0, 1, 2, 3))
dat$reg122307s_c_sc3g12 <- recodeVar(dat$reg122307s_c_sc3g12,
                                     c(-97, -95, -94, -55, 0, 1, 2, 3, 4, 5, 6),
                                     c(-97, -95, -94, -55, 0, 0, 1, 2, 3, 4, 5))
dat$reg121602s_c_sc3g12 <- recodeVar(dat$reg121602s_c_sc3g12,
                                     c(-97, -95, -94, -55, 0, 1, 2, 3, 4),
                                     c(-97, -95, -94, -55, 0, 0, 0, 1, 2))
dat$reg122005s_c_sc3g12 <- recodeVar(dat$reg122005s_c_sc3g12,
                                     c(-97, -95, -94, -54, 0, 1, 2, 3),
                                     c(-97, -95, -94, -54, 0, 0, 1, 2))
dat$reg122907s_sc3g12_c <- recodeVar(dat$reg122907s_sc3g12_c,
                                     c(-97, -95, -94, -54, 0, 1, 2, 3),
                                     c(-97, -95, -94, -54, 0, 0, 1, 2))
dat$reg122504s_sc3g12_c <- recodeVar(dat$reg122504s_sc3g12_c,
                                     c(-97, -95, -94, -54, 0, 1, 2, 3),
                                     c(-97, -95, -94, -54, 0, 0, 1, 2))
dat$reg122501s_c_sc3g12 <- recodeVar(dat$reg122501s_c_sc3g12,
                                     c(-97, -94, -55, -54, 0, 1, 2, 3),
```

```
                                    c(-97, -94, -55, -54, 0, 0, 0, 1))

### Scaling of the reading test using the partial credit model (PCM)

# Identify polytomous items
poly <- apply(dat[, reading.items], 2, max, na.rm = TRUE) > 1

# Define Q-matrix for the scaling model
Q <- matrix( 1 , nrow = length(reading.items), ncol = 1)
Q[poly, 1] <- 0.5    # score of 0.5 for polyomous items

# Fit the model
pcm <- list()
pcm$model <- tam.mml(resp = dat[, reading.items], irtmodel = "PCM2", Q=Q,
                     pid = dat$ID_t)

# Estimate WLEs
pcm$wle <- tam.wle(pcm$model, Msteps = 1000)
```

Part C. Fixed item parameters used for estimating linked WLEs in the starting cohort 3 grade 12

	Item	Linked item difficulty		Item	Linked item difficulty
1	reg1205010_sc3g12_c	0.490	22	reg1226040_sc3g12_c	0.883
2	reg1205020_sc3g12_c	1.104	23	reg1226050_sc3g12_c	2.380
3	reg1205030_sc3g12_c	-0.488	24	reg1226060_sc3g12_c	1.556
4	reg120504s_sc3g12_c	-0.561	25	reg1226080_sc3g12_c	1.025
5	reg1205050_sc3g12_c	0.954	26	reg121602s_sc3g12_c	0.670
6	reg1205060_sc3g12_c	0.483	27	reg121603s_sc3g12_c	1.954
7	reg1205070_sc3g12_c	-0.435	28	reg1216040_sc3g12_c	1.989
8	reg1229010_sc3g12_c	1.373	29	reg121605s_sc3g12_c	1.202
9	reg1229020_sc3g12_c	0.804	30	reg1216060_sc3g12_c	1.147
10	reg1229030_sc3g12_c	0.195	31	reg1220010_sc3g12_c	-0.491
11	reg1229060_sc3g12_c	1.163	32	reg122002s_sc3g12_c	1.104
12	reg122907s_sc3g12_c	1.215	33	reg1220030_sc3g12_c	-0.207
13	reg1229080_sc3g12_c	2.831	34	reg1220040_sc3g12_c	0.156
14	reg1229100_sc3g12_c	0.152	35	reg122005s_sc3g12_c	1.580
15	reg122301s_sc3g12_c	0.904	36	reg1220060_sc3g12_c	0.306
16	reg1223020_sc3g12_c	1.279	37	reg122501s_sc3g12_c	0.869
17	reg1223040_sc3g12_c	-0.061	38	reg1225030_sc3g12_c	0.270
18	reg122305s_sc3g12_c	1.194	39	reg1225060_sc3g12_c	1.837
19	reg1223060_sc3g12_c	0.016	40	reg1225050_sc3g12_c	2.064
20	reg1226020_sc3g12_c	1.583	41	reg122504s_sc3g12_c	1.216
21	reg1226030_sc3g12_c	0.267			