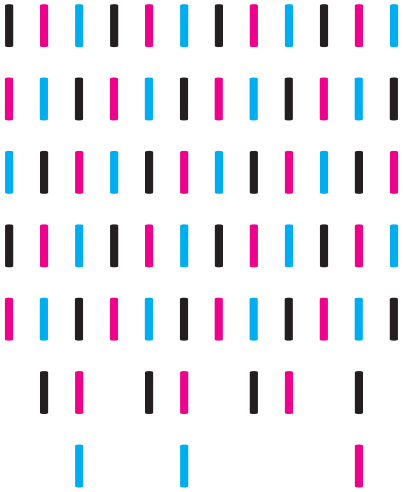

NEPS *SURVEY PAPERS*

Anna Scharl, Theresa Rohm, and Eva Zink

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

NEPS *Survey Paper* No. 85
Bamberg, April 2021

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).

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 edited by a review board consisting of the scientific management of LIfBi and NEPS.

The NEPS *Survey Papers* are available at www.neps-data.de (see section “Publications“) and at www.lifbi.de/publications.

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 in Seventh Grade

Anna Scharl[1], Theresa Rohm[1,2], Eva Zink[1,2]

[1]Leibniz Institute for Educational Trajectories, Germany
[2]University of Bamberg, Germany

E-mail address of the lead author:

anna.scharl@lifbi.de

Bibliographic data:

Scharl, A., Rohm, T., Zink, E. (2021). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 2 in Seventh Grade* (NEPS Survey Paper No. 85). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP85:1.0

Acknowledgments:

The present report has been modeled along previous reports published by LIfBi. To facilitate the understanding of the presented results many text passages (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (e.g., Krannich, Jost, Rohm, Koller, Carstensen, Fischer, & Gnambs, 2017).

NEPS Technical Report for Reading – Scaling Results of Starting Cohort 2 in Seventh Grade

Abstract

The National Educational Panel Study (NEPS) investigates the development of competences across the life span and develops tests for assessing these competence domains in different age groups. To evaluate the quality of the competence tests, a wide range of analyses based on item response theory (IRT) are performed. This paper describes the data and scaling procedure for the reading competence test administered in Grade 7 of Starting Cohort 2 (Kindergarten). The study represents a follow-up to the reading competence test administered in Grade 4 of Starting Cohort 2 and is again the same test as administered in Starting Cohort 3 in the same grade. Two different test versions were administered to the students. Based on their reading competence scores in Grade 4, students were either administered an easy or a difficult test version. The easy version of the reading competence test contained 29 items, whereas the difficult version included 30 items with different response formats representing different cognitive requirements and text functions. The two test versions were administered to 2,578 students; 1,471 of them received the easy test version, whereas 1,107 students received the difficult test version. Their responses were scaled using the partial credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, the tests' dimensionality, and local item independence were evaluated to ensure the quality of the test. These analyses demonstrated that the test exhibited an acceptable reliability and that the items showed an acceptable model fit. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test were the large percentage of items at the end of the difficult test that were not reached due to time limits and some differential item functioning between the easy and difficult test versions for some items. Overall, the reading competence 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 in the Scientific Use File and provides the R 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).

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 scales are described in Pohl and Carstensen (2012).

In this paper the results of these analyses are presented for the reading competence test administered in Grade 7 of Starting Cohort 2 (Kindergarten). The same test has been previously administered to Grade 7 of Starting Cohort 3 (5^{th} graders). This study represents a follow-up to the reading competence test administered in Grade 4 of Starting Cohort 2 (see Rohm, Krohmer, & Gnambs, 2017). First, we introduce the main concepts of the reading competence test. 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 is available for public use in the Scientific Use File (SUF) is presented.

Please note that the analyses in this report are based on the data set available at some time before the public data release. Due to ongoing data protection and data cleaning issues, the data in the Scientific Use File may differ slightly from the data used for the analyses in this paper. However, we do not expect fundamentally different results for the published data.

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, c) literary, d) instruction, or e) advertising. Furthermore, the test assessed three cognitive requirements. These were (a) finding information in the text, (b) drawing text-related conclusions, and (c) reflecting and assessing. The cognitive requirements did not depend on the text type, but each cognitive requirement was 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: 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 several subtasks with

two response options were presented. MA items required the test taker to match several responses to a given set of statements. MA items were 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).

3. Data

3.1 The Design of the Study

The study assessed different competence domains including reading competence, scientific competence, and mathematical competence. The reading competence test was always presented second within the test battery as was the case for the Grade 4 reading competence test. Thus, there was no change in the rotation design. Furthermore, there was no multi-matrix design regarding the order of the items *within* a specific test. All students received the same test items in the same 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) and administered two different difficulty-tiered reading competence tests within the sample. Based on preliminary studies, two different versions of the reading competence test were developed that differed in their average difficulty and included either more easy or more difficult items. Because reading competence was already measured in Grade 4, students were administered either test version depending on their reading competence score in Grade 4. Students with a reading competence score below the median reading competence in Grade 4 received the easy test version, whereas students with a reading competence score above the median received the difficult test version.

3.2 Instrument

In both test versions five texts, with the five text functions as described above, were presented. The second (*information*), third (*instruction*), and fourth text (*literary*) were identical in both test versions. In contrast, the first (*advertising*) and fifth text (*commenting*) differed between the easy and the difficult test version. In total, the reading competence test in Grade 7 consisted of 42 items with different response formats (see Table 1) representing different cognitive requirements and text functions (see Tables 2 and 3). The 17 common items referring to three texts (*information*, *instruction*, *literary*) were presented in both test versions. Moreover, one additional item referring to one of these texts was only included in the difficult test version. The 12 items referring to the remaining two texts (*advertising*, *commenting*) differed between the easy and the difficult test version. In the easy version, the first and the last text contained easier items whereas in the difficult test version these texts contained more difficult items. An overview of the assignment of text functions and cognitive requirements to items is depicted in Appendix A.

Preliminary analyses revealed that three items exhibited differential item functioning between the difficult and easy test version. Therefore, these items were split between the two test versions and treated as test version unique items. Thus, the descriptive characteristics of 42 items are summarized in Tables 1 to 3 and 5, while the results of the item

response theory analyses with, then, 45 items are depicted in Tables 6 to 12. The number of subtasks within CMC and MA items varied between two and five.

Table 1

Number of Items by Response Formats and Test Version

Response format	Easy test	Difficult test
Simple multiple choice	17	20
Complex multiple choice	7	5
Matching	3	4
Total number of items	27	29

Table 2

Number of Items by Cognitive Requirements and Test Version

Cognitive requirements	Easy test	Difficult test
Finding Information in text	6	7
Drawing text-related conclusions	14	13
Reflecting and assessing	7	9
Total number of items	27	29

Table 3

Number of Items by Text Types and Test Version

Text type/functions	Easy test	Difficult test
Advertising texts	6	6
Information texts	5	6
Instruction texts	6	6
Literary texts	5	6
Commenting texts	5	5
Total number of items	27	29

3.3 Sample

A total of 2,578 students received the reading competence test[1]. All students gave at least the minimum number of three valid item responses to estimate reliable competence scores for them (see Pohl & Carstensen, 2012). The number of participants receiving the easy and

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

difficult test version 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 Test Version

Easy test	Difficult test	Total
1,471	1,107	2,578

4. Analyses

Some of the following analyses are based on both test versions whereas other analyses examined the two test versions separately. Results that are based on separate analyses are explicitly indicated in the text and are reported in separate tables for the two test versions. Otherwise, the results refer to both test versions. These analyses did neither correct for the position of the reading competence test nor the difficulty of the different test versions.

4.1 Missing responses

Competence data include different kinds of missing responses. There are missing responses due to a) invalid responses, b) omitted items, c) items that test-takers did not reach, d) items that have not been administered, and finally e) multiple kinds of missing responses within CMC items (not-determinable missing values).

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 were coded as not-reached. Because of the different test versions, some items were not administered to all participants. For respondents receiving the easy test version some difficult items were missing by design, whereas some easy items were missing by design for respondents answering the difficult test version (see Table 1). Because complex multiple-choice and matching 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. For a detailed description of the different kinds of missing data see also Pohl et al. (2012).

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 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). Person abilities were derived using the weighted maximum likelihood estimator (WLE; Warm, 1989). 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 indicating the number of correctly solved subtasks within that item. Categories of polytomous variables with less than N = 200 responses were collapsed to avoid possible estimation problems (Pohl et al., 2012). An overview of items with collapsed categories for this study is given in Appendix B.

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).

4.3 Checking the quality of the test

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

To assess the quality of the distractors of the MC items and whether it was justified to aggregate the CMC and MA subtasks to polytomous items, the items were analyzed together in a Rasch model (Rasch, 1960).

The distractors within MC items were evaluated using the point-biserial correlation between selecting an incorrect response option and the total correct score treating all subtasks of CMC and MA items as single items. Negative correlations indicated good distractors, whereas correlations between .00 and .05 were considered acceptable and correlations above .05 were viewed as problematic distractors (Pohl & Carstensen, 2012). The fit of the CMC and MA subtasks was evaluated based on the weighted mean square error (WMNSQ), the respective t-value, point-biserial correlations of selecting the correct response with the total correct score, and the item characteristic curve. Only if the subtasks showed a satisfactory item fit, they were used to construct polytomous CMC and MA variables that were included in the final scaling model.

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC items and the 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 > 1.20 (t-value > |8|) were considered as having noticeable item misfit and their performance was further investigated. Correlations of the item scores with the total correct scores greater than .30 were considered as good, greater than .20 as acceptable, and below .20 as problematic. The 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), school type, age, and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Moreover, given the different test versions we also examined the measurement invariance of the 17 common items that were included in the easy and difficult test versions. Differential item functioning (DIF) was examined using a multigroup IRT model, in which the 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 greater than 1.0 logit as very strong DIF, absolute differences between 0.6 and 1.0 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, 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 may not hold for empirical data. To test the assumption of equal item discrimination parameters a generalized partial credit model (GPCM; Muraki, 1992) was 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 fit to the data. The correlations among the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the test.

Since the reading competence test consisted of item sets that referred to one of five texts, the assumption of local item independence (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 Statistical software

The item response models were estimated with the *TAM* package version 3.5-19 (Robitzsch, Kiefer, & Wu, 2020) in *R* version 3.6.1 (R Core Team, 2019).

5. Results

5.1 Missing responses

5.1.1 Missing responses per person

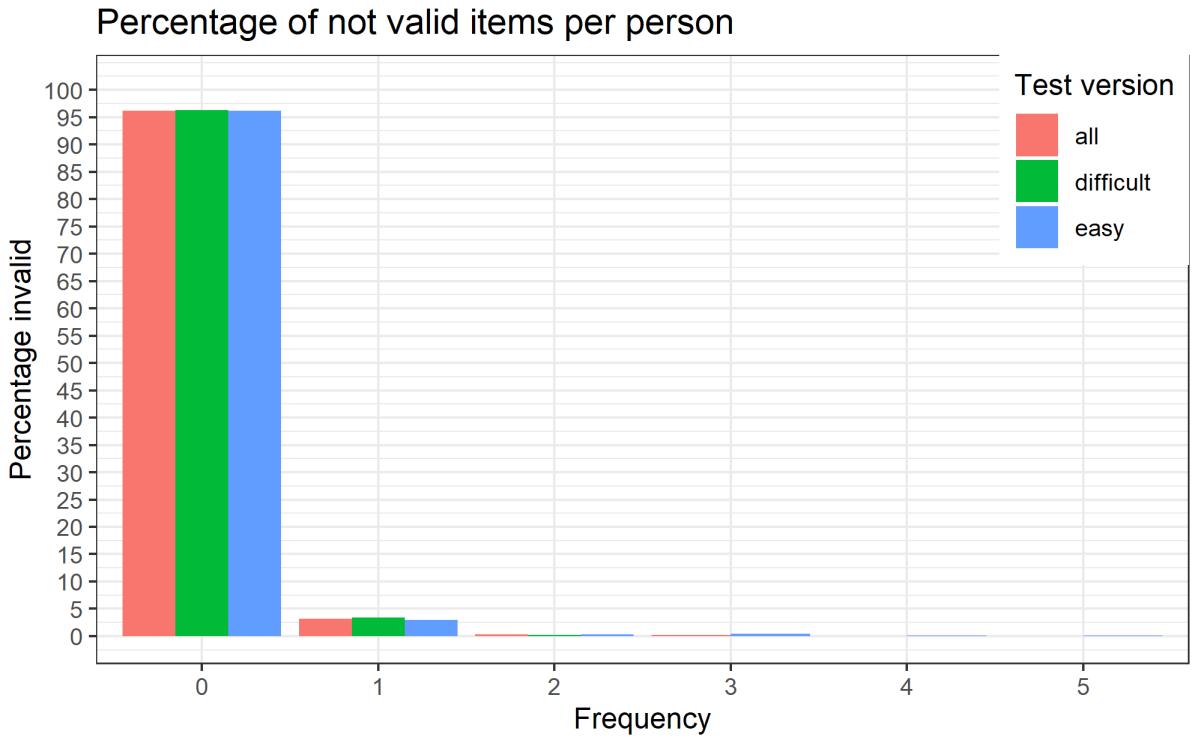

Figure 1. Number of invalid responses

The number of invalid responses per person is shown in Figure 1. The number of invalid responses was very low for both test versions. In the easy test version, 96% of the students had no invalid responses at all and only about one percent of the students had more than one invalid response. In the difficult test version, 96% of the students had no invalid responses at all and less than one percent of the students had more than one invalid response.

Missing responses may also occur when respondents omit some items. As can be seen in Figure 2, there was a non-negligible amount of omitted items even if the number of omitted items was low. In the easy test version, 75% of the students omitted no item at all, whereas only five percent of the students omitted more than three items. In the difficult test version, 76% of the students omitted no item at all and 3.4 percent of the students omitted more than three items.

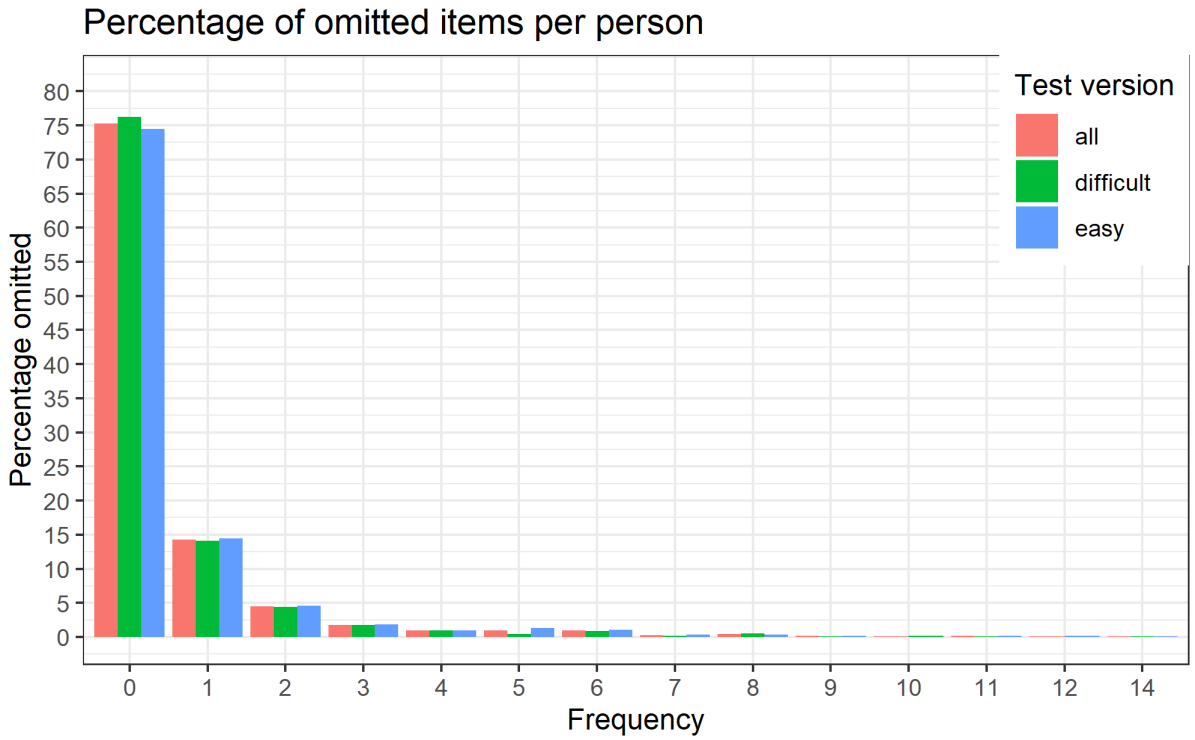

Figure 2. Number of omitted items

Per definition, all missing responses after the last valid response were not reached. The number of not-reached items for the easy test version was acceptable whereas the number of the not-reached items for the difficult test version was rather high. This is illustrated in Figure 3. About 70% of the students reached the end of the easy reading competence test; 16% of the students did not reach the items of the last text and 5% did not reach the last two of the five texts. Note that only 53% of the students reached the end of the difficult reading test. In this case, 34% of the students did not reach the items of the last text, 11% did not reach the last two of the five texts. Around 62% of the respondents reached the end of the test, with about 70% receiving the easy and 53% the difficult test version. This is similar to the numbers found in Starting Cohort 3, Grade 7 (Krannich et al., 2017). Conversely, in the previous assessment of Starting Cohort 2, Grade 4, which received a test originally designed for Starting Cohort 3, Grade 5, not even 40% were able to complete the test in the allotted time (Rohm, Krohmer, & Gnambs, 2017). This might indicate better test targeting.

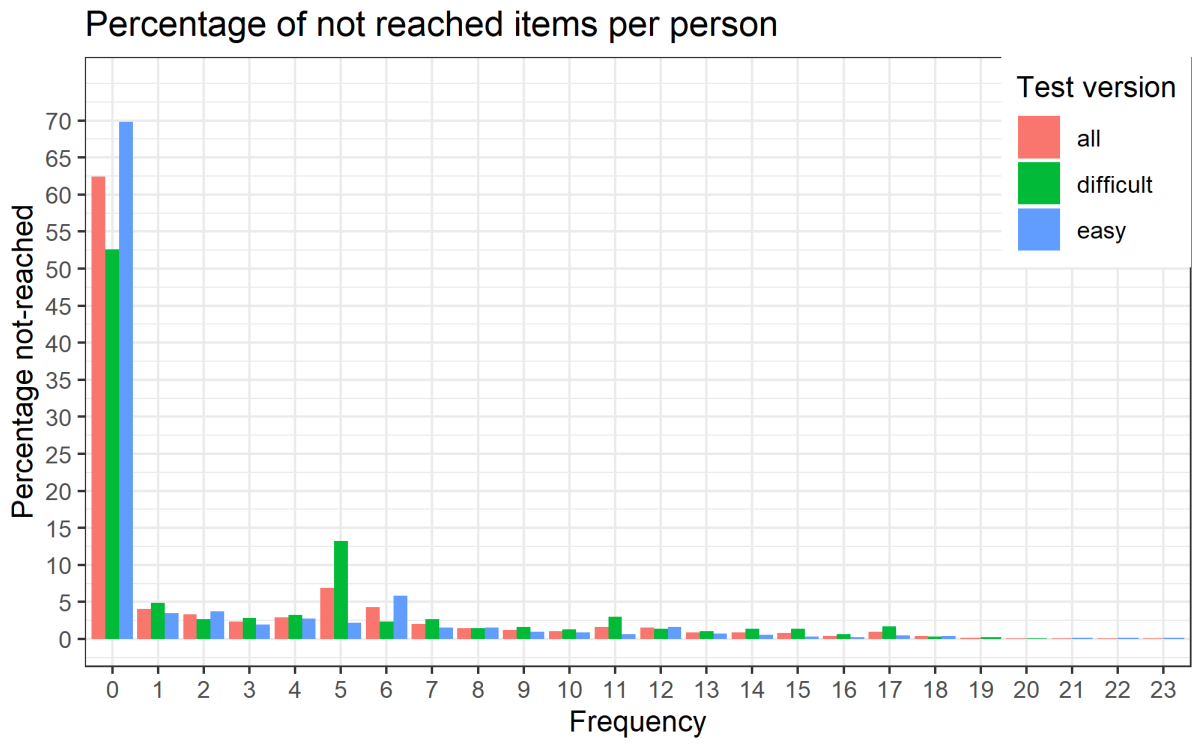

Figure 3. Number of not-reached items

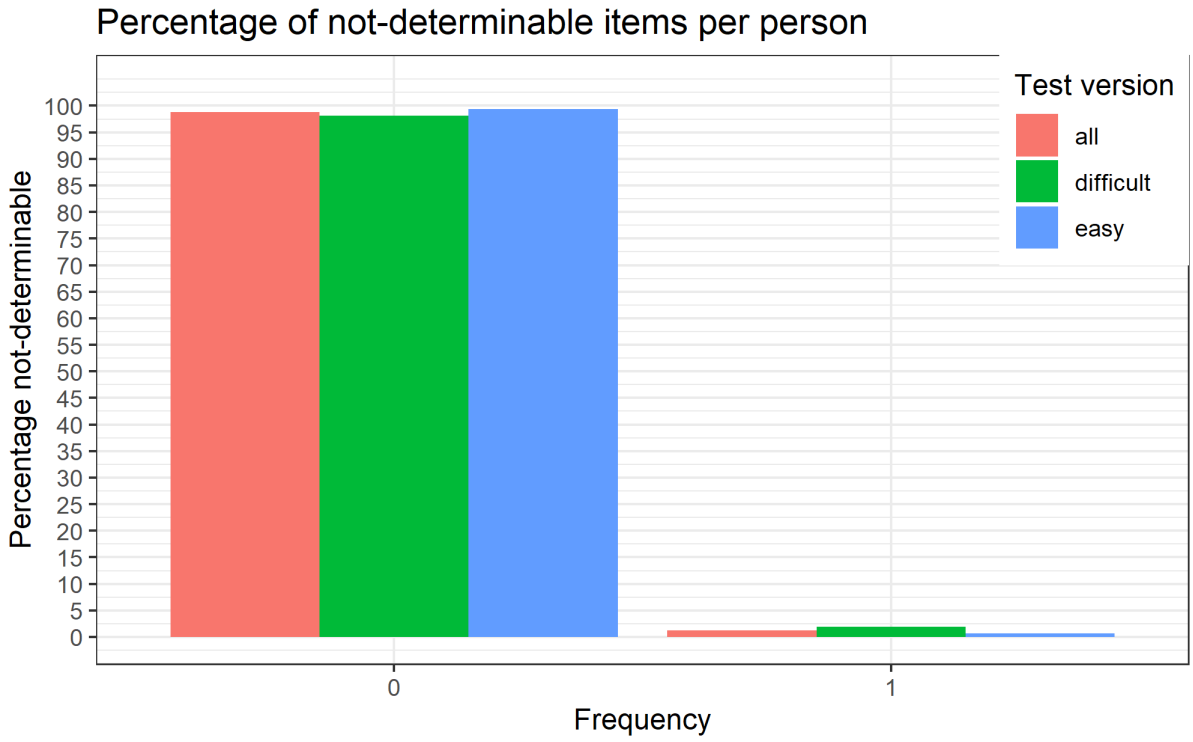

Figure 4. Number of not-determinable missing responses

The aggregated polytomous variables were coded as a not-determinable missing response when the subtasks of CMC and MA items contained different kinds of missing responses. Because not-determinable missing responses may only occur in CMC and MA items, the maximum number of not-determinable missing responses was nine (for the difficult test version) or ten (for the easy test version). There was only a very small amount of not-determinable missing responses for both test versions (see Figure 4). About 98% of the students in both test versions did not have a single not-determinable missing response.

The total number of missing responses aggregated over invalid, omitted, not-reached, and not-determinable missing responses per person is illustrated in Figure 5. It can be seen that 53% of the students that were administered the easy test version had no missing response at all. Only about 19% of these test students had more than five missing responses. In the difficult test version, there were about 41% of the students who had no missing response at all. Almost 27% of these test students had more than five missing responses.

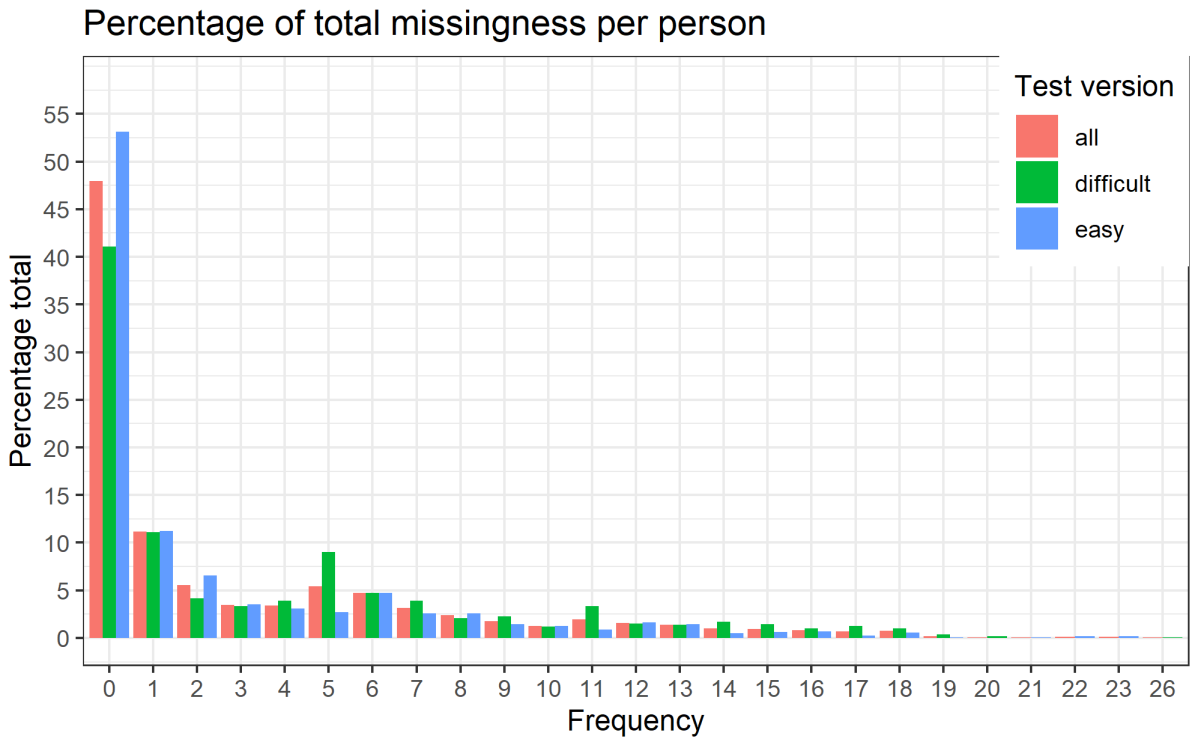

Figure 5. Total number of missing responses

Summarizing these results, there was a small amount of invalid and not-determinable missing responses for both test versions and a reasonable amount of omitted items. The number of not-reached items was – at least for the difficult test version – rather large and, therefore, the larger amount of total missingness per person in the difficult version is primarily due to the items not reached.

5.1.2 Missing responses per item

Table 5 gives information on the number of valid responses for each item, as well as the percentage of missing responses. Overall, the omission rate was quite good. In the easy test version, there was only one item with an omission rate above 6%; in the difficult test version,

there was one item with an omission rate above 6% as well. The highest omission rate occurred for item reg7024s_sc2g7_c (6.73% of the students omitted this item in the easy test version, 4.88% of the students in the difficult test version). The number of students that did not reach an item (see Figure 6) increased with the position of the item in the test to up to 30.18% (easy test version) or 47.43% (difficult test version). This is a rather large amount, especially for the difficult test version. The number of invalid responses per item was small. The highest number was for item reg7045s_sc2g7_c; 0.82% in the easy test version or 1.17% in the difficult test version. The total number of missing responses per item varied between 0.27% (reg70610_sc2g7_c) and 50.77% (item reg7075s_sc2g7_c).

Table 5

Missing values by test version

	Easy test version						Difficult test version					
Item	**Position**	***N***	**NR**	**OM**	**NV**	**ND**	**Position**	***N***	**NR**	**OM**	**NV**	**ND**
reg70110_sc2g7_c	1	1,464	0.00	0.00	0.48	0.00						
reg70120_sc2g7_c	2	1,443	0.00	1.56	0.34	0.00						
reg7013s_sc2g7_c	3	1,450	0.00	1.43	0.00	0.00						
reg70140_sc2g7_c	4	1,466	0.00	0.34	0.00	0.00						
reg7015s_sc2g7_c	5	1,431	0.00	2.72	0.00	0.00						
reg7016s_sc2g7_c	6	1,413	0.00	3.20	0.68	0.07						
reg70210_sc2g7_c	7	1,449	0.14	1.29	0.07	0.00	8	1,102	0.00	0.45	0.00	0.00
reg70220_sc2g7_c	8	1,426	0.27	2.65	0.14	0.00	9	1,095	0.00	1.08	0.00	0.00
reg7023s_sc2g7_c	9	1,400	0.41	4.42	0.00	0.00	10	1,078	0.00	2.62	0.00	0.00
reg7024s_sc2g7_c	10	1,365	0.41	6.73	0.07	0.00	11	1,052	0.09	4.88	0.00	0.00
reg70250_sc2g7_c	11	1,430	0.41	1.90	0.48	0.00	12	1,080	0.36	1.90	0.18	0.00
reg7026s_sc2g7_c							13	1,025	0.72	6.14	0.36	0.18
reg70310_sc2g7_c	12	1,438	0.82	1.36	0.07	0.00	14	1,073	2.44	0.63	0.00	0.00
reg70320_sc2g7_c	13	1,407	1.29	2.99	0.07	0.00	15	1,041	3.07	2.89	0.00	0.00
reg7033s_sc2g7_c	14	1,373	1.56	4.96	0.07	0.07	16	1,015	4.43	3.43	0.27	0.18
reg70340_sc2g7_c	15	1,403	1.90	2.31	0.41	0.00	17	1,025	5.78	1.54	0.09	0.00
reg70350_sc2g7_c	16	1,400	2.45	2.11	0.27	0.00	18	1,016	6.87	1.36	0.00	0.00
reg70360_sc2g7_c	17	1,372	3.20	3.40	0.14	0.00	19	997	8.22	1.63	0.09	0.00
reg70410_sc2g7_c	18	1,397	4.83	0.20	0.00	0.00	20	982	11.20	0.09	0.00	0.00
reg70420_sc2g7_c	19	1,382	5.44	0.54	0.07	0.00	21	966	12.47	0.27	0.00	0.00
reg70430_sc2g7_c	20	1,363	6.32	1.02	0.00	0.00	22	947	14.09	0.36	0.00	0.00
reg70440_sc2g7_c	21	1,348	7.27	1.02	0.07	0.00	23	932	15.45	0.27	0.00	0.00
reg7045s_sc2g7_c	22	1,279	8.77	3.20	0.82	0.27	24	858	18.25	2.80	1.17	0.27
reg70460_sc2g7_c	23	1,294	10.27	1.50	0.27	0.00	25	869	20.60	0.81	0.09	0.00
reg7051s_sc2g7_c	24	1,197	16.11	2.38	0.14	0.00						
reg70520_sc2g7_c	25	1,197	18.29	0.34	0.00	0.00						
reg7053s_sc2g7_c	26	1,135	21.01	1.70	0.14	0.00						
reg70540_sc2g7_c	27	1,115	22.98	1.16	0.07	0.00						
reg7055s_sc2g7_c	28	1,047	26.72	1.16	0.68	0.27						
reg70560_sc2g7_c	29	1,027	30.18	0.00	0.00	0.00						

reg70610_sc2g7_c	1	1,104	0.00	0.27	0.00	0.00
reg70620_sc2g7_c	2	1,085	0.00	1.45	0.54	0.00
reg7063s_sc2g7_c	3	1,092	0.00	1.26	0.00	0.09
reg70640_sc2g7_c	4	1,088	0.00	1.63	0.09	0.00
reg70650_sc2g7_c	5	1,085	0.00	1.99	0.00	0.00
reg7066s_sc2g7_c	6	1,068	0.00	2.80	0.54	0.18
reg70670_sc2g7_c	7	1,093	0.00	1.08	0.18	0.00
reg7071s_sc2g7_c	26	709	33.79	2.08	0.00	0.09
reg70720_sc2g7_c	27	686	37.04	0.90	0.09	0.00
reg70730_sc2g7_c	28	658	39.84	0.72	0.00	0.00
reg70740_sc2g7_c	29	628	42.55	0.72	0.00	0.00
reg7075s_sc2g7_c	30	545	47.43	2.17	0.27	0.90

Note. Position = Item position within testlet, *N* = Number of valid responses, NR = Percentage of respondents that did not reach the item, OM = Percentage of respondents that omitted the item, NV = Percentage of respondents with an invalid response, ND = Percentage of respondents with not-determinable missing values.

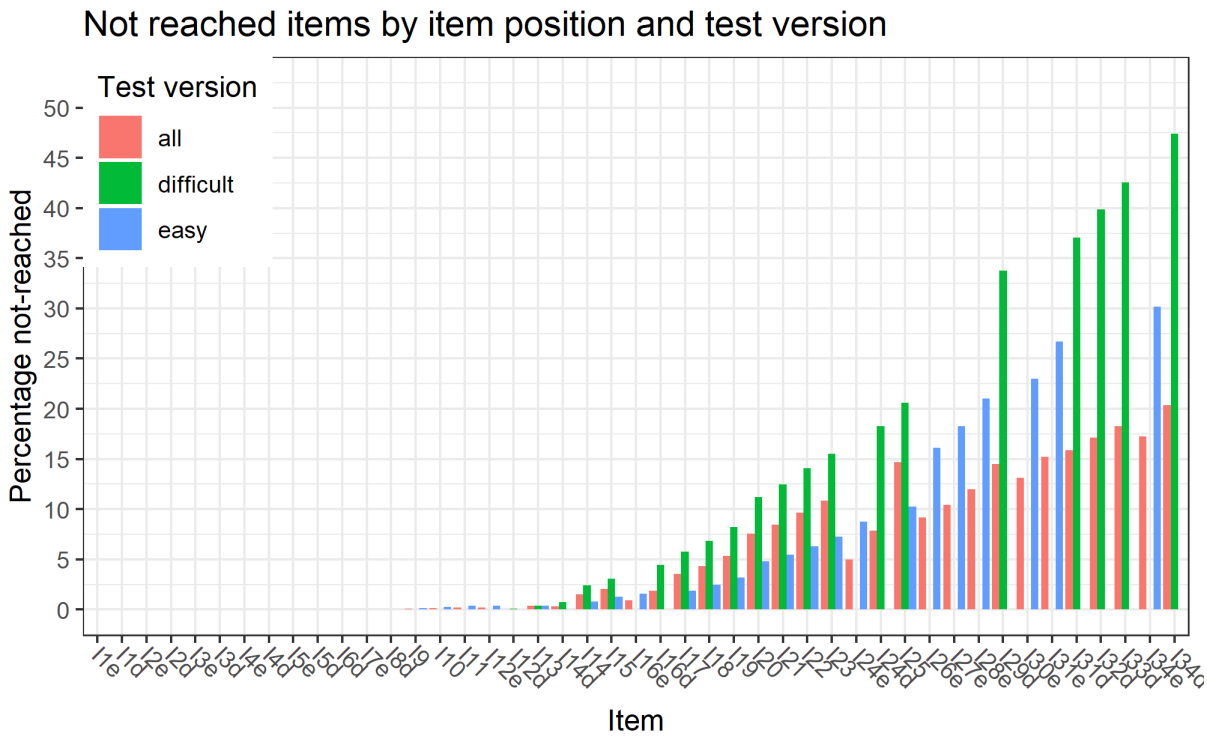

Figure 6. Not reached missing values by item position

5.2 Parameter estimates

5.2.1 Item parameters

The percentage of correct responses relative to all valid responses for each item is summarized in Table 6 (third column). Because there was a non-negligible amount of missing responses this value cannot be interpreted as an index of item difficulty. The percentage of correct responses within dichotomous items varied between 29% and 96% with an average of 70.4% correct responses.

Table 6

Item Parameters

Position	Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
1	reg70110_sc2g7_c	57	-0.68	0.06	1.17	6.52	0.20	0.62	0.04
2	reg70120_sc2g7_c	86	-2.59	0.08	1.11	1.99	0.18	0.72	0.03
3	reg7013s_sc2g7_c	n.a.	-1.46	0.06	0.94	-1.73	0.26	1.00	0.04
4	reg70140_sc2g7_c	96	-3.99	0.13	0.94	-0.50	0.27	1.78	0.03
5	reg7015s_sc2g7_c	n.a.	-1.85	0.07	1.02	0.47	0.14	0.44	0.03
6	reg7016s_sc2g7_c	n.a.	-0.09	0.04	1.00	0.05	0.30	0.59	0.03
7	reg70210_sc2g7_c	90	-2.67	0.07	0.89	-2.34	0.33	1.87	0.03
8	reg70220_sc2g7_c	85	-2.18	0.06	1.05	1.17	0.25	1.10	0.03
9	reg7023s_sc2g7_c	n.a.	-1.06	0.03	1.00	-0.14	0.27	0.63	0.03
10	reg7024s_sc2g7_c	38	0.35	0.06	1.09	3.23	0.23	0.84	0.04
11		58	0.04	0.07	1.06	2.17	0.29	0.96	0.04
12	reg70250_sc2g7_c	79	-1.69	0.06	1.06	1.92	0.29	1.02	0.03
13	reg7026s_sc2g7_c	n.a.	-0.50	0.07	0.89	-4.47	0.39	1.81	0.05
14	reg70310_sc2g7_c	91	-2.78	0.07	1.01	0.24	0.22	1.14	0.03
15	reg70320_sc2g7_c	80	-1.70	0.06	1.08	2.60	0.24	0.90	0.03
16	reg7033s_sc2g7_c	29	0.83	0.07	1.05	1.64	0.24	0.96	0.04
17		47	0.64	0.07	1.06	2.46	0.23	0.78	0.04
18	reg70340_sc2g7_c	80	-1.73	0.06	0.95	-1.72	0.37	1.53	0.04
19	reg70350_sc2g7_c	88	-2.44	0.07	0.89	-2.50	0.37	1.91	0.04
20	reg70360_sc2g7_c	78	-1.57	0.06	0.88	-4.22	0.43	1.87	0.03
21	reg70410_sc2g7_c	92	-2.96	0.08	0.96	-0.76	0.32	1.46	0.03
22	reg70420_sc2g7_c	88	-2.47	0.07	0.89	-2.43	0.42	1.86	0.04
23	reg70430_sc2g7_c	92	-2.98	0.08	0.85	-2.67	0.42	2.36	0.04
24	reg70440_sc2g7_c	89	-2.62	0.07	0.88	-2.44	0.41	1.94	0.04
25	reg7045s_sc2g7_c	54	-0.52	0.06	1.09	3.51	0.32	0.87	0.03
26		71	-0.57	0.08	1.01	0.29	0.33	1.20	0.04
27	reg70460_sc2g7_c	40	0.53	0.05	1.00	-0.13	0.34	1.12	0.03
28	reg7051s_sc2g7_c	n.a.	-1.49	0.07	0.92	-2.02	0.37	1.09	0.04
29	reg70520_sc2g7_c	81	-2.14	0.08	0.96	-0.96	0.41	1.51	0.04
30	reg7053s_sc2g7_c	n.a.	-0.78	0.06	0.96	-1.82	0.33	0.81	0.04
31	reg70540_sc2g7_c	49	-0.29	0.07	1.08	2.93	0.32	0.87	0.03
32	reg7055s_sc2g7_c	n.a.	-0.34	0.03	1.01	0.34	0.44	0.57	0.04

33	reg70560_sc2g7_c	42	0.07	0.07	1.07	2.37	0.35	0.92	0.02
34	reg70610_sc2g7_c	95	-2.99	0.14	0.94	-0.47	0.28	1.68	0.04
35	reg70620_sc2g7_c	74	-0.84	0.08	1.06	1.49	0.22	0.90	0.04
36	reg7063s_sc2g7_c	n.a.	-1.35	0.08	0.98	-0.33	0.20	0.77	0.04
37	reg70640_sc2g7_c	58	0.06	0.07	1.05	2.06	0.27	0.90	0.04
38	reg70650_sc2g7_c	62	-0.16	0.07	1.02	0.72	0.28	1.09	0.04
39	reg7066s_sc2g7_c	n.a.	-0.14	0.06	0.93	-4.26	0.34	1.23	0.04
40	reg70670_sc2g7_c	71	-0.67	0.07	1.15	3.98	0.15	0.55	0.04
41	reg7071s_sc2g7_c	n.a.	-0.36	0.08	1.00	0.12	0.20	0.55	0.03
42	reg70720_sc2g7_c	45	0.75	0.08	1.08	2.48	0.25	0.71	0.03
43	reg70730_sc2g7_c	56	0.22	0.09	1.10	3.10	0.24	0.69	0.04
44	reg70740_sc2g7_c	83	-1.40	0.12	1.06	0.87	0.30	0.93	0.04
45	reg7075s_sc2g7_c	n.a.	-0.06	0.09	0.89	-4.50	0.45	1.66	0.06

Note. 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 the item (Yen, 1993). The items reg7024s_sc2g7_c, reg7033s_sc2g7_c, and reg7045s_sc2g7_c were split by test version; the first line corresponds to the values for respondents receiving the easy test version, whereas the second line corresponds to the difficult test version.

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.

The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 6, whereas the step parameters (for polytomous variables) are summarized 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) varied between -3.99 (item reg70140_sc2g7_c) and 0.83 (item reg7033s_sc2g7_c in the easy testlet) with a mean of -1.12. Overall, the item difficulties ranged from low to medium difficulty; however, there were no highly difficult items. Due to the large sample size, the standard errors (*SE*) of the estimated item difficulties (column 4 in Table 6) were rather small, $SE(\beta) \leq 0.14$.

Table 7

Step parameters (and standard errors) of polytomous items

Item	Step 1 (*SE*)		Step 2 (*SE*)		Step 3 (*SE*)
reg7016s_sc2g7_c	-0.45	(0.05)	0.45		
reg7023s_sc2g7_c	-0.25	(0.05)	0.25		
reg7055s_sc2g7_c	-0.32	(0.06)	0.27	(0.07)	0.05

Note. For the last step parameter no standard error is reported because it represents a fixed 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. Because some items in the reading test were polytomous, we calculated Thurstonian thresholds for each response category (Wu, Adams, Wilson, & Haldane, 2007). These indicate the location at the latent dimension at which the probability of achieving a score above the respective threshold is 50%. Thus, it is similar to the item difficulties of dichotomous items. In Figure 7, the item difficulties of the reading competence 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 category thresholds. The variance was estimated to be 1.270, which indicates good differentiation between the students. The reliabilities of the test (EAP/PV reliability = .800, WLE reliability = .751) were good. The mean of the category threshold distribution was about one logit below the mean person ability distribution. Although the items covered a wide range of the ability distribution, on average, the items were slightly too easy. As a consequence, person abilities in medium- and low-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors.

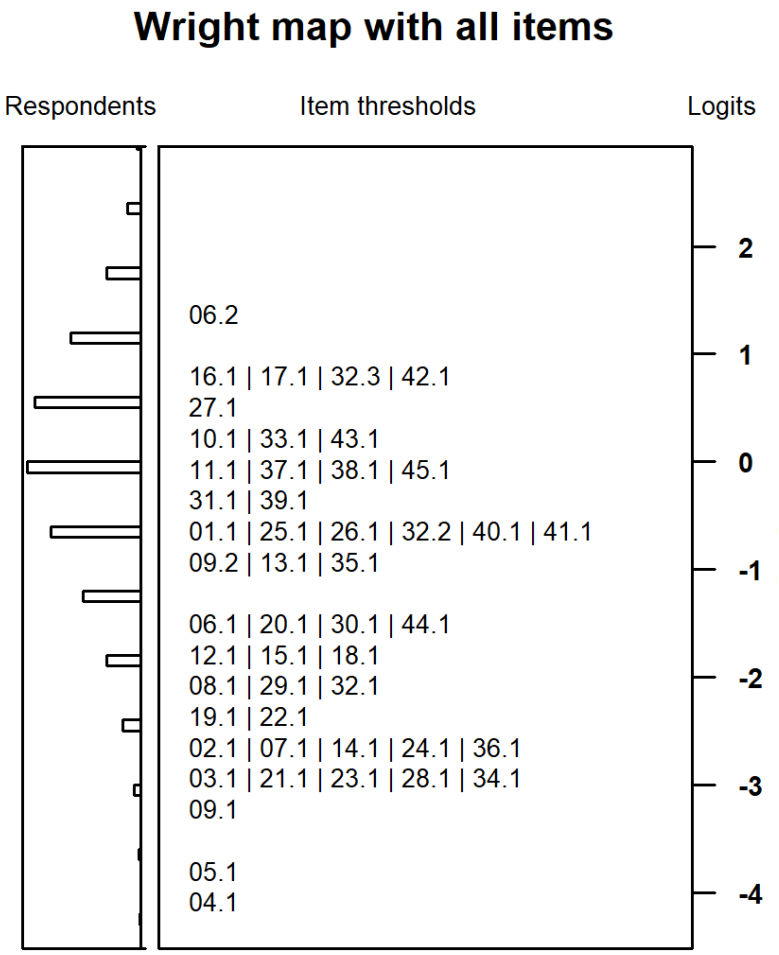

Figure 7. Test targeting. The distribution of the person ability in the sample is given on the left-hand side of the graph. The category thresholds of the items are given on the right-hand side of the graph. Each number represents one threshold with the first part (before the dot) corresponding to the item position given in Table 6 and the second part indicating the threshold.

5.3 Quality of the test

5.3.1 Fit of the subtasks of complex multiple choice and matching 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 simple MC items in a Rasch model. Counting the subtasks of CMC and MA items separately, there were 50 items in the easy and 53 items in the difficult test version. The probability of a correct response ranged from 40% to 96% except for reg7051_sc2g7_c (13%). Thus, the number of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.84 to 1.2, the respective *t*-value from -3.6 to 7.29, and there were no noticeable deviations of the empirically estimated probabilities from the model-implied item characteristic curves. Due to the satisfying model fit of the subtasks, their aggregation to polytomous variables seemed justified.

5.3.2 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 selecting an incorrect response (distractor) and the students' total correct score. The distractors consistently yielded negative point-biserial correlations ranging from -.35 to -.04 for the easy testlet and mostly negative point-biserial correlations between -.34 and .01 for the difficult test version. These results indicate that the distractors functioned well for the easy and at least sufficiently for the difficult test version.

5.3.3 Item fit

The evaluation of item fit was performed based on the final scaling model, the partial credit model, with concurrent calibration (i.e., the easy and difficult test were scaled together). Altogether, the item fit can be considered good (see Table 6). The values of the WMNSQ were reasonably close to 1 with the lowest value being .85 (item reg70430_sc2g7_c) and the highest being 1.17 (item reg70110_sc2g7_c). Two items exhibited a WMNSQ of at least 1.15 (item reg70670_sc2g7_c and reg70110_sc2g7_c) and no item a *t*-value above 8. There were no further indications of pronounced misfit of these items. Therefore, they were retained for estimating the reading competence scores. The correlations between the item scores and the total correct scores varied between .14 (item reg7015s_sc2g7_c) and .45 (item reg7075s_sc2g7_c) with an average correlation of .30. All item characteristic curves showed a good fit for the items.

5.3.4 Differential item functioning

Differential item functioning (DIF) was used to evaluate the test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables sex, the number of books at home (as a proxy for socioeconomic status), migration background, and test position (see Pohl & Carstensen, 2012, for a description of these variables). Also, for the common items that were administered to all participants, we studied them for measurement invariance between the easy and difficult test version. 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 men and women; a positive value would indicate that the item was more difficult for males, whereas a negative

value would highlight a lower difficulty for males as opposed to females. Besides investigating DIF for every single item, an overall test for DIF was performed by comparing models that allowed for DIF to those that only estimated main effects (see Table 9).

Sex: The sample included 1,263 (48.99%) boys and 1,315 (51.01%) girls. On average, male students had a higher reading ability than female students (main effect = 0.247 logits, Cohen's *d* = 0.207). Three items exhibited DIF above 0.6 (items reg7024s_sc2g7_c in both testlets, reg70250_sc2g7_c, and reg70220_sc2g7_c). Five items exhibited noticeable, but not severe DIF between 0.4 and 0.6 logits (items reg7026s_sc2g7_c, reg70520_sc2g7_c, reg70460_sc2g7_c, reg70610_sc2g7_c, and reg7063s_sc2g7_c). An overall test for DIF (see Table 9) was conducted by comparing the DIF model to a model that only estimated the main effects (but ignored potential DIF). Model comparisons using Akaike's (1974) information criterion (AIC) and also the Bayesian information criterion (BIC; Schwarz, 1978) that takes the number of estimated parameters into account and, thus, penalizes overparameterized models were conducted. The model comparison using the AIC favored the model estimating DIF, whereas the BIC indicated a better fit for the model indicating only the main effect. Thus, the DIF regarding sex did not have a large impact.

Table 8

Differential Item Functioning

Item	Sex	School	Age	Books	Difficulty	Migration
	male vs. female	no sec vs. sec	<Md vs. ≥ Md	<100 vs. ≥ 100	easy vs. difficult	without vs. with
reg70110_sc2g7_c	0.157 (0.131)	0.122 (0.102)	0.096 (0.081)	-0.026 (-0.022)		0.413 (0.347)
reg70120_sc2g7_c	-0.097 (-0.081)	0.225 (0.188)	0.285 (0.239)	0.123 (0.107)		0.099 (0.083)
reg7013s_sc2g7_c	0.241 (0.201)	0.027 (0.023)	-0.031 (-0.026)	0.195 (0.169)		0.283 (0.238)
reg70140_sc2g7_c	-0.085 (-0.071)	0.179 (0.150)	-0.049 (-0.041)	-0.386 (-0.335)		0.353 (0.297)
reg7015s_sc2g7_c	0.078 (0.065)	0.014 (0.011)	-0.133 (-0.112)	0.325 (0.282)		-0.139 (-0.117)
reg7016s_sc2g7_c	-0.068 (-0.057)	-0.025 (-0.021)	-0.046 (-0.038)	0.323 (0.288)		-0.114 (-0.096)
reg70210_sc2g7_c	0.071 (0.060)	-0.397 (-0.332)	0.165 (0.138)	0.095 (0.083)	-0.565 (-0.435)	-0.173 (-0.145)
reg70220_sc2g7_c	0.775 (0.648)	0.0672 (0.060)	-0.033 (-0.028)	-0.094 (-0.082)	-0.054 (-0.041)	-0.008 (-0.007)
reg7023s_sc2g7_c	0.165 (0.138)	0.0012 (0.010)	-0.052 (-0.044)	0.139 (0.120)	0.405 (0.311)	-0.078 (-0.066)
reg7024s_sc2g7_c	0.638 (0.533)	-0.012 (-0.010)	-0.083 (-0.070)	0.283 (0.245)		0.073 (0.061)
	0.622 (0.520)	0.124 (0.103)	-0.174 (-0.146)	-0.035 (-0.030)		0.539 (0.454)
reg70250_sc2g7_c	0.626 (0.523)	0.005 (0.004)	-0.007 (-0.006)	0.002 (0.002)	0.424 (0.326)	0.061 (0.051)
reg7026s_sc2g7_c	0.400 (0.334)	0.223 (0.187)	-0.035 (-0.029)	0.207 (0.179)		0.077 (0.065)
reg70310_sc2g7_c	-0.085 (-0.071)	-0.242 (-0.203)	0.091 (0.076)	-0.115 (-0.100)	-0.017 (-0.013)	0.067 (0.056)
reg70320_sc2g7_c	-0.163 (-0.136)	-0.036 (-0.030)	-0.052 (-0.044)	0.288 (0.249)	0.339 (0.261)	-0.548 (-0.461)
reg7033s_sc2g7_c	0.335 (0.280)	-0.234 (-0.196)	-0.131 (-0.110)	0.070 (0.060)		0.099 (0.083)
	0.059 (0.049)	-0.182 (-0.152)	0.056 (0.047)	0.162 (0.140)		0.140 (0.118)
reg70340_sc2g7_c	0.260 (0.217)	-0.396 (-0.331)	-0.078 (-0.066)	-0.378 (-0.328)	-0.157 (-0.121)	0.162 (0.137)

reg70350_sc2g7_c	-0.187	-0.208	0.185	-0.352	-0.243	-0.028
	(-0.156)	(-0.174)	(0.155)	(-0.305)	(-0.187)	(-0.024)
reg70360_sc2g7_c	0.007	-0.167	0.159	-0.465	-0.209	0.301
	(0.006)	(-0.140)	(0.133)	(-0.404)	(-0.161)	(0.254)
reg70410_sc2g7_c	-0.238	-0.189	0.360	-0.342	-0.021	-0.210
	(-0.199)	(-0.158)	(0.301)	(-0.296)	(-0.016)	(-0.177)
reg70420_sc2g7_c	0.024	0.137	0.237	-0.081	0.343	-0.083
	(0.020)	(0.114)	(0.199)	(-0.070)	(0.263)	(-0.070)
reg70430_sc2g7_c	-0.348	0.099	0.190	-0.565	0.082	0.127
	(-0.291)	(0.083)	(0.160)	(-0.490)	(0.063)	(0.107)
reg70440_sc2g7_c	-0.086	-0.229	0.133	-0.354	-0.241	0.227
	(-0.154)	(-0.192)	(0.111)	(-0.307)	(-0.185)	(0.191)
reg7045s_sc2g7_c	-0.184	-0.022	0.110	-0.038		-0.076
	(-0.154)	(-0.018)	(0.092)	(-0.033)		(-0.064)
	-0.216	-0.071	-0.229	0.051		-0.061
	(-0.181)	(-0.059)	(-0.192)	(0.044)		(-0.052)
reg70460_sc2g7_c	-0.426	0.190	0.008	-0.277	0.128	0.337
	(-0.356)	(0.159)	(0.006)	(-0.240)	(0.098)	(0.283)
reg7051s_sc2g7_c	-0.250	0.035	0.178	0.253		0.099
	(-0.209)	(0.030)	(0.149)	(0.220)		(0.083)
reg70520_sc2g7_c	-0.436	-0.262	0.136	-0.454		0.033
	(-0.364)	(-0.219)	(0.114)	(-0.394)		(0.028)
reg7053s_sc2g7_c	-0.005	0.205	-0.057	0.213		0.093
	(-0.004)	(0.171)	(-0.048)	(0.185)		(0.078)
reg70540_sc2g7_c	0.261	-0.217	-0.094	0.020		0.172
	(0.218)	(-0.181)	(-0.079)	(0.018)		(0.144)
reg7055s_sc2g7_c	-0.056	0.004	-0.067	0.163		-0.171
	(-0.047)	(0.004)	(-0.056)	(0.142)		(-0.144)
reg70560_sc2g7_c	0.207	0.231	-0.168	-0.283		0.166
	(0.173)	(0.194)	(-0.141)	(-0.245)		(0.139)
reg70610_sc2g7_c	-0.419	0.016	-0.380	0.024		0.156
	(-0.350)	(0.014)	(-0.319)	(0.021)		(0.131)
reg70620_sc2g7_c	-0.240	-0.052	0.064	0.076		-0.356
	(-0.201)	(-0.044)	(0.053)	(0.066)		(-0.300)
reg7063s_sc2g7_c	-0.487	-0.224	-0.251	0.099		-0.303
	(-0.407)	(-0.187)	(-0.211)	(0.086)		(-0.255)
reg70640_sc2g7_c	0.155	0.039	0.163	0.186		-0.468
	(0.129)	(0.032)	(0.137)	(0.161)		(-0.394)
reg70650_sc2g7_c	0.054	0.221	-0.049	0.083		-0.212
	(0.045)	(0.184)	(-0.041)	(0.072)		(-0.178)
reg7066s_sc2g7_c	-0.090	0.075	-0.058	0.216		0.175
	(-0.076)	(0.063)	(-0.048)	(0.188)		(0.147)
reg70670_sc2g7_c	0.040	0.074	-0.030	-0.259		-0.262

	(0.033)	(0.062)	(-0.025)	(-0.224)		(-0.221)
reg7071s_sc2g7_c	-0.148 (-0.124)	0.051 (0.043)	-0.271 (-0.227)	0.602 (0.521)		0.391 (0.265)
reg70720_sc2g7_c	-0.105 (-0.028)	0.246 (0.205)	0.028 (0.024)	0.086 (0.074)		-0.080 (-0.062)
reg70730_sc2g7_c	-0.107 (-0.089)	0.014 (0.012)	0.247 (0.207)	0.171 (0.148)		-0.502 (-0.422)
reg70740_sc2g7_c	-0.355 (-0.297)	0.633 (0.529)	-0.41 4 (-0.347)	0.004 (0.003)		-0.018 (-0.016)
reg7075s_sc2g7_c	0.224 (0.187)	0.143 (0.120)	0.082 (0.069)	-0.093 (-0.081)		0.673 (0.567)
Main effect	0.247 (0.207)	0.065 (0.054)	-0.109 (-0.091)	0.370 (0.320)	1.134 (0.877)	-0.177 (-0.149)

School type: Overall, 805 respondents (31.24%) who took the reading test attended grammar school (German: "Gymnasium") whereas 1,773 (68.76%) did not. Participants attending grammar school showed on average a higher reading ability (0.065 logits, Cohen's d = 0.054). One item exhibited DIF greater than 0.6 logits (reg70740_sc2g7_c). Both the AIC and the BIC comparing the models favored the main model (AIC = 66893.4, BIC = 67192, number of parameters = 51). Therefore, the DIF regarding school type did not have a large impact.

Age: The participants were on average 12.71 years old. Participants who were older than the median age of the sample were slightly less proficient than those beneath the median age (main effect = -0.109 logits, Cohen's d = -0.091). One item exhibited noteworthy DIF (item reg70740_sc2g7_c) (DIF = -0.414 logits). Furthermore, both AIC and BIC favor the main effects model without item-level DIF. Therefore, reading competencies were measured comparably in the two groups.

Number of books: The number of books at home was used as a proxy for socioeconomic status. There were 800 (31.0%) students with 0 to 100 books at home and 1,778 (69.0%) students with more than 100 books at home. There were noticeable average differences between the two groups. Participants with 100 or fewer books at home performed on average 0.370 logits (Cohen's d = 0.320) worse than participants with more than 100 books (see Table 8). Three items showed noticeable but not severe DIF between participants with many or fewer books (items reg70360_sc2g7_c, reg70430_sc2g7_c, and reg70520_sc2g7_c) and reg7071s_sc2g7_c showed considerable DIF. As a consequence, the overall test for DIF using the BIC favored the main effects model without DIF effects, whereas comparing the models using the AIC favored the model estimating DIF (Table 9).

Test difficulty: To estimate the participants' proficiency with greater accuracy, the participants received different tests that either included a larger number of easy or a larger number of difficult items (see section 3.1 for the design of the study). Only a subset of 16 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 1,471 participants (57.1%) received the easy test version and 1,107 participants (42.9%) received the difficult test version. As expected, students who were administered the difficult

test version outperformed the participants receiving the easy test version (main effect = 1.134 logits, Cohen's *d* = 0.877). Three items (reg70210_sc2g7_c, reg7023s_sc2g7_c, and reg70250_sc2g7_c) showed noticeable, but not severe DIF. The model comparison using the AIC favored the model estimating DIF (AIC = 26845.8, BIC = 27021.4, number of parameters = 30), whereas the BIC indicated a better fit for the model indicating only the main effect (AIC = 26884.9, BIC = 26984.4, number of parameters = 17). Thus, the DIF regarding difficulty did not have a large impact.

Migration: There were 2,264 (80.8%) participants without a migration background, 230 (9.2%) participants with a migration background. There was a little difference in the average performance of participants with and without migration background. Participants without a migration background had a higher reading ability than participants with a migration background (main effect = -0.177 logits, Cohen's *d* = -0.149). One item (reg7075s_sc2g7_c) exhibited considerable DIF of 0.674 logits; five items (reg70110_sc2g7_c, reg70640_sc2g7_c, reg70320_sc2g7_c, reg70730_sc2g7_c, and reg7024s_sc2g7_c for the difficult testlet respondents) exhibited noticeable, but not severe DIF between participants with and without migration background (in absolute numbers: DIF = [0.413; 0.547]). However, the overall test for DIF using the BIC and AIC favored the main effects model that did not include item-level DIF. Therefore, reading competencies were measured comparably in the two groups.

Table 9

Comparison of models with and without DIF

DIF Variable	Model	Deviance	Number of parameters	AIC	BIC
Sex	main effect	66778.199	51	66880.199	67178.792
	DIF	66565.044	95	66755.044	67311.247
School	main effect	66791.437	51	66893.437	67192.011
	DIF	66731.468	95	66921.468	67477.634
Age	main effect	66813.038	51	66915.038	67213.631
	DIF	66769.066	95	66959.066	67515.269
Books	main effect	63301.655	51	63403.655	63699.859
	DIF	63181.255	95	63371.255	63923.007
Difficulty	main effect	26850.926	17	26884.926	26984.444
	DIF	26785.803	30	26845.803	27021.423
Migration	main effect	64344.908	51	64446.908	64743.812
	DIF	64297.586	95	64487.586	65040.642

In summary, most of the differences in item difficulties across the different subgroups were (in absolute values) below 0.6. There were only two larger effects for the gender of the participants, one for their migration background and one for the school type the participants attended. Concerning AIC and BIC as overall model fit indices, at least one of the measures supported the models without DIF; thus, there was no substantial indication of test unfairness.

5.3.5 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item discrimination parameters are equal. To test this assumption, a generalized partial credit model (GPCM) that estimates

discrimination parameters was fitted to the data. The estimated discrimination parameters differed moderately among items (see Table 6), ranging from 0.44 (item reg7015s_sc2g7_c) to 2.36 (item reg70430_sc2g7_c). The average discrimination parameter fell at 1.13. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 69521, BIC = 70060, number of parameters = 92) as compared to the PCM (AIC = 70250, BIC = 70543, number of parameters = 50). Despite the empirical preference for the GPCM, the PCM 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 two test versions 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. The estimation of the models was carried out using the Gauss-Hermite quadrature method.

Table 10

Results of three-dimensional scaling

	Dim 1	**Dim 2**	**Dim 3**
Finding information in the text (Dim 1) (11 items)	1.402		
Drawing Text-related conclusions (Dim 2) (17 items)	0.946	1.350	
Reflecting and assessing (Dim 3) (15 items)	0.915	0.939	1.248

Note. The variances of the dimensions are given in the diagonal; correlations are given in the off-diagonal.

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 .92 and .95. However, they deviated from a perfect correlation (i.e., they were lower than r = .95, see Carstensen, 2013). Moreover, according to model fit indices the three-dimensional model (AIC = 66813.9, BIC = 67135.9, number of parameters = 55) fitted the data slightly better than the unidimensional model (AIC = 66829.7, BIC = 67122.4, number of parameters = 50). These results indicate that the three cognitive requirements measure a common construct, albeit it is not completely unidimensional.

Table 11

Results of five-dimensional scaling

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Information (Dim 1) (13 items)	0.9207				
Instruction (Dim 2) (6 items)	0.812	1.602			
Advertising (Dim 3) (6 items)	0.845	0.871	1.971		
Commenting (Dim 4) (7 items)	0.858	0.903	0.841	2.386	
Literary (Dim 5) (11 items)	0.884	0.864	0.830	0.856	1.118

Note. The variances of the dimensions are given in the diagonal; 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 = .81 and r = .90. All correlations deviated from a perfect correlation (i.e., they were considerably lower than r = .95, see Carstensen, 2013). For the easy test version, the five-dimensional model (AIC = 66577.4, BIC = 66952.1, number of parameters = 64 fitted the data better than the unidimensional model (AIC = 66829.7, BIC = 67122.4, number of parameters = 50).

As each text function corresponded to one of the five texts, local item dependence (LID) and the text functions were confounded. Consequently, the deviation of the correlations from a perfect correlation shown in Tables 10 and 11, may result from multidimensionality as well as from local item dependence. Given the testing design in the main studies, it was not possible to disentangle the two sources. In pilot studies (Gehrer et al., 2013), a larger number of texts were presented to test-takers, so that the impact of text functions could be investigated independently of LID. The correlations estimated in the pilot study ranged from .78 to .91. As the correlations found in Gehrer and colleagues (2013) differed 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 were due to text functions, to those found in the main study (Table 11), which were due to both text functions and LID, allowed us to evaluate the impact of LID. The correlations found in the present study of Starting Cohort 2 were similar (between 0.83 and 0.91) 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).

6. Discussion

The analyses in the previous sections provided detailed information on the quality of the reading test in Starting Cohort 2 for Grade 7. 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 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 in the difficult test version was rather high, indicating that the test was too long for the allocated testing time and the difficulty of the items. However, the amount of not-reached items was still comparable with other reading competence tests (see section 5.1.1). Other types of missing responses were reasonably small. The test had a high reliability and distinguished well between test-takers. However, the test was mainly targeted at low-performing students and did not accurately measure the reading competence of high-performing students. As a consequence, ability estimates will be precise for low-performing students but less precise for high performing students. Some degree of multidimensionality was present for different text functions. In combination with the high amount of missing responses at the end of the test (i.e., there were students with no valid responses to some of the text functions), the estimation of a single reading competence score might be challenged. This should be addressed in further studies. Nevertheless, Gehrer et al. (2013) argued that a balanced assessment of reading competence can only be achieved by a heterogeneity of text functions and they provide theoretical arguments for a unidimensional measure of reading competence.

In this study, two difficulty-tiered tests were administered. Students were assigned to one of the two test versions based on their previous performance on the reading competence test in Grade 4. Because the complex design provided additional challenges, additional analyses were conducted that showed that the common items of the two test versions measured the same latent dimension as the test unique items. Moreover, the common items were largely measurement invariant across the two test versions. Dimensionality analyses showed that the latent associations between the content dimensions (cognitive requirements and text functions) were comparatively high.

In sum, it was shown that it is feasible to implement a macro-adaptive procedure that assigned students to an easy or more difficult test version depending on their prior performance. Overall, the administered test had satisfactory 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 contain 42 items, of which 27 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. A total of 15 items were scored as polytomous variables (CMC items). MC items are marked with a '0_c' at the end of the variable name, whereas the variable names of CMC items end in 's_c'. In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category.

7.2 Linking of reading competence scores of Grade 4 and Grade 7

In Starting Cohort 2, the reading competence tests administered in Grades 4 (see Rohm, Krohmer, & Gnambs, 2017) and 7 included different items that were constructed in such a way as to allow for an accurate measurement of reading competence within each age group. As a consequence, the competence scores derived in the different grades cannot be directly compared; differences in observed scores would reflect differences in competencies as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competencies across grades, we adopted the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016). Following an anchor-group design, an independent link sample including students from Grade 7 was administered both the reading tests from Grades 4 and 7 within one measurement occasion. These responses were used to link the two tests administered in Starting Cohort 2 across the two grades.

7.2.1 Samples

In Starting Cohort 2, a subsample of 2,279 students participated at both measurement occasions, in Grade 4 and Grade 7. 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 = 1,189 students (about 50% girls) from Grade 7 received both tests within a single measurement occasion. This link sample has been used before to link the reading competence tests of Grades 5 and 7 in Starting Cohort 3 (Fischer et al., 2016; Krannich et al., 2017). The present analyses relied on the subsample of N = 555 students who received the computer-based version of the Grade 7 and the paper-based version of the Grade 5 test to mirror the assessment modes in the main study.

7.2.2 The design of the link study

The reading test in Grade 4 included 31 items, whereas the test in Grade 7 consisted of 42 items. Because retest effects are expected for the reading items, no common items could be administered in the two tests. Instead, an overlap of information was accomplished by using an independent link sample including 555 participants attending Grade 7. In the link sample, a random subset of the participants received the Grade 5 test before the Grade 7 test and vice versa. The Grade 5 test was administered paper-based, whereas the Grade 7 test was a computer-based version. Moreover, because the Grade 7 test was administered difficulty-tiered, the two test versions were randomly assigned to respondents. In the link sample, the test was scaled concurrently, whereas in the main sample the tests in Grades 4 and 7 were scaled independently.

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 = 25,997, BIC = 26,386), over the one-dimensional model, (AIC = 26,042, BIC = 26,423).

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal subsample from the starting cohort. The differences in item difficulties between the link sample and Starting Cohort 2 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 12.

Table 12

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

Item (Grade 5)	Starting Cohort 2 v. link sample	*SE*	*F*	Item (Grade 7)	Starting Cohort 2 v. link sample	*SE*	*F*
reg50110_sc2g4_c	-1.71***	-1.49	426.68	reg70110_sc2g7_c	1.08***	0.94	270.57
reg50130_sc2g4_c	-1.31***	-1.14	375.72	reg70120_sc2g7_c	0.66	0.58	65.71
reg50140_sc2g4_c	-1.16***	-1.01	339.57	reg7013s_sc2g7_c†	0.44	0.38	37.80
reg50150_sc2g4_c	-0.74*	-0.65	167.77	reg70140_sc2g7_c†	0.10	0.08	0.87
reg5016s_sc2g4_c†	-0.39***	-0.34	301.03	reg7015s_sc2g7_c†	0.43	0.38	32.62
reg50170_sc2g4_c	-1.29***	-1.12	495.72	reg7016s_sc2g7_c	0.75***	0.65	263.85
reg50210_sc2g4_c	-1.37***	-1.2	312.94	reg70210_sc2g7_c	0.56	0.49	57.84
reg50220_sc2g4_c†	-0.34	-0.3	36.99	reg70220_sc2g7_c	0.55	0.48	65.55
reg50230_sc2g4_c	-0.91*	-0.79	155.31	reg7023s_sc2g7_c†	0.45	0.39	125.62
reg50240_sc2g4_c	-0.96***	-0.84	231.13	reg7024s_sc2g7_c	0.75	0.66	122.60
reg50250_sc2g4_c	-0.80**	-0.70	188.93		-0.98*	-0.86	172.35
reg50310_sc2g4_c	-1.35***	-1.18	355.16	reg70250_sc2g7_c	0.99***	0.86	255.45
reg50320_sc2g4_c	-1.20***	-1.04	233.54	reg70310_sc2g7_c	0.66	0.58	76.52
reg50330_sc2g4_c	-1.02*	-0.89	174.40	reg70320_sc2g7_c†	0.17	0.15	7.27
reg50340_sc2g4_c	-1.13***	-0.98	306.60	reg7033s_sc2g7_c†	0.47	0.41	43.05
reg50350_sc2g4_c	-0.76*	-0.66	165.14		-1.77***	-1.54	529.57
reg50360_sc2g4_c	-1.18***	-1.03	263.25	reg70340_sc2g7_c	0.75	0.65	139.83
reg50370_sc2g4_c	-0.50	-0.43	60.87	reg70350_sc2g7_c	0.87	0.76	146.41

reg50410_sc2g4_c	-1.09***	-0.95	288.02	reg70360_sc2g7_c	0.77	0.67	148.40
reg5042s_sc2g4_c†	-0.21	-0.18	29.74	reg70410_sc2g7_c	1.18***	1.02	225.46
reg50430_sc2g4_c	-0.97***	-0.84	204.70	reg70420_sc2g7_c	0.96*	0.84	178.20
reg50440_sc2g4_c	-1.02***	-0.89	231.72	reg70430_sc2g7_c	0.73	0.64	81.11
reg50460_sc2g4_c	-1.07***	-0.93	239.59	reg70440_sc2g7_c	1.05**	0.92	194.50
reg50510_sc2g4_c	-1.30***	-1.14	241.30	reg7045s_sc2g7_c	1.28***	1.12	340.52
reg5052s_sc2g4_c†	0.21	0.18	18.84		0.67	0.58	70.83
reg50530_sc2g4_c†	-0.27	-0.24	13.11	reg70460_sc2g7_c	0.61	0.53	107.43
reg50540_sc2g4_c	-0.83	-0.73	109.25	reg7051s_sc2g7_c	0.77	0.67	104.46
reg5055s_sc2g4_c†	0.14	0.12	10.26	reg70520_sc2g7_c	0.82	0.71	102.90
reg50570_sc2g4_c	-0.52	-0.45	41.04	reg7053s_sc2g7_c†	0.36	0.32	26.68
				reg70540_sc2g7_c†	0.46	0.40	41.02
				reg7055s_sc2g7_c	0.56***	0.49	203.20
				reg70560_sc2g7_c	0.97*	0.84	167.32
				reg70610_sc2g7_c	0.95	0.83	92.18
				reg70620_sc2g7_c†	0.39	0.34	25.45
				reg7063s_sc2g7_c†	0.20	0.18	6.51
				reg70640_sc2g7_c†	0.38	0.33	27.11
				reg70650_sc2g7_c	0.91	0.79	151.13
				reg7066s_sc2g7_c	1.84***	1.61	641.94
				reg70670_sc2g7_c†	-0.39	-0.34	25.94
				reg7071s_sc2g7_c	0.72	0.63	79.40
				reg70720_sc2g7_c†	0.09	0.08	1.08
				reg70730_sc2g7_c†	0.13	0.11	2.40
				reg70740_sc2g7_c†	0.11	0.10	1.38
				reg7075s_sc2g7_c	-1.60	-1.39	7.02

Note. Differences in item difficulty parameters between the sample in Grades 4 and 7, Starting Cohort 2, and the link sample. *SE* = Standard error of the DIF estimates, *F* = *F*-value of the estimates with the critical *F*-value of 151.56. † = The item is used for calculating the link constant. The items reg5012s_sc2g4_c, reg5026s_sc2g4_c, and reg7026s_sc2g7_c had to be excluded from the link because they did not contain any correct responses either in the link sample or the longitudinal main sample.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Analyses of differential item functioning between the link sample and Starting Cohort 2 identified several items for Grade 4 (difference in absolute logits: *Min* = 0.136, *Max* = 1.712) and for Grade 7 (difference in absolute logits: *Min* = 0.086, *Max* = 1.843) with significant (α =.05) DIF and large differences ($d \geq 0.5$ logits). Therefore, the reading competence tests

administered in the two grades were linked using only the subsample of items without considerable DIF. For Grade 4, six items (about 21% of the test) and for Grade 7, 15 items (about 34% of the test) could be used for linking the tests. The "mean/mean" method for the anchor-group design was used (see Fischer et al., 2016). The correction term was calculated as $c = 0.49$. The estimated WLEs were subsequently transformed so that the mean difference of the longitudinal subsample corresponds to the correction term, thus effectively bringing the two tests onto a common scale. The link error reflecting the uncertainty in the linking process was calculated according to equation 4 in Fischer et al. (2016) as 0.18 and has to be included in 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, "reg7_sc1" and "reg7_sc1u", including their respective standard errors, "reg7_sc2" and "reg7_sc2u". For "reg7_sc1u", person abilities were estimated using the linked item difficulty parameters. As a result, the WLE scores provided in "reg7_sc1u" can be used for longitudinal comparisons between Grades 4 and 7. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in "reg7_sc1" are neither linked to the underlying reference scale of Grade 4 nor to the scores of Starting Cohort 3, Grade 7, which was administered the same test. As a consequence, they cannot be used for longitudinal purposes, but only for cross-sectional research questions. Nor can they be used to compare across starting cohorts. The R Syntax for estimating the WLE is provided in Appendix B. For persons who either did not take part in the reading test or who did not give enough valid responses, no WLE is estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values.

Plausible values allow investigating latent relationships of competence scores with other variables. For preliminary analyses, plausible values for the reading test will be provided in the SUFs ("reg7_pv1" to "reg7_pv10" for cross-sectional and "reg7_pv1u" to "reg7_pv10u" for longitudinal analyses). Because these plausible values are estimated using a minimal background model, we recommend using the *R* package *NEPSscaling*[2] (Scharl, Carstensen, & Gnambs, 2020) to estimate more precise plausible values fitting the current research question at hand.

[2] https://www.neps-data.de/Data-Center/Overview-and-Assistance/Plausible-Values

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-722. https://doi.org/10.1109/TAC.1974.1100705

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). Springer.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. (2016). *Linking the Data of the Competence Tests* (NEPS Survey Paper No. 1). 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). 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. 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.

Krannich, M., Jost, O., Rohm, T., Koller, I., Carstensen, C. H., Fischer, L., & Gnambs, T. (2017). *NEPS Technical Report for Reading –Scaling Results of Starting Cohort 3 for Grade7*(NEPS Survey Paper No. 14). Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176. https://doi.org/10.1177/014662169201600206

Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement, 50*, 447-468. https://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). Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., & Carstensen, C. (2013). Scaling the 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). Otto-Friedrich-Universität, Nationales Bildungspanel.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Robitzsch, A., Kiefer, T., & Wu, M. (2020). *TAM: Test Analysis Modules.* R package version 3.5-19. https://CRAN.R-project.org/package=TAM

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

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). Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Scharl, A., Carstensen, C. H., & Gnambs, T. (2020). *Estimating Plausible Values with NEPS Data: An Example Using Reading Competence in Starting Cohort 6* (NEPS *Survey Paper* No. 71). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP71:1.0

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464. https://doi.org/10.1214/aos/1176344136

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. In H. P. Blossfeld, H. G. Roßbach, & J. v. Maurice (Eds.), *Education as a lifelong process: The German National*

Educational Panel Study (NEPS) (pp. 67-86). (Zeitschrift für Erziehungswissenschaft, Sonderheft 14 . VS Verlag für Sozialwissenschaften.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest version 2.0: Generalised item response modelling software. *Victoria: ACER Press*.

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: Different Text Types and Cognitive Requirements

Item	Position	Text Types	Cognitive Requirements
reg70110_sc2g7_c	1	Advertising	Type 3 - Reflecting and Assessing
reg70120_sc2g7_c	2	Advertising	Type 1 - Finding Information in Text
reg7013s_sc2g7_c	3	Advertising	Type 1 - Finding Information in Text
reg70140_sc2g7_c	4	Advertising	Type 1 - Finding Information in Text
reg7015s_sc2g7_c	5	Advertising	Type 2 - Drawing text-related conclusions
reg7016s_sc2g7_c	6	Advertising	Type 3 - Reflecting and Assessing
reg70210_sc2g7_c	7	Information	Type 1 - Finding Information in Text
reg70220_sc2g7_c	8	Information	Type 2 - Drawing text-related conclusions
reg7023s_sc2g7_c	9	Information	Type 1 - Finding Information in Text
reg7024s_sc2g7_c	10	Information	Type 2 - Drawing text-related conclusions
reg70250_sc2g7_c	11	Information	Type 3 - Reflecting and Assessing
reg7026s_sc2g7_c	12	Information	Type 3 - Reflecting and Assessing
reg70310_sc2g7_c	13	Instruction	Type 1 - Finding Information in Text
reg70320_sc2g7_c	14	Instruction	Type 2 - Drawing text-related conclusions
reg7033s_sc2g7_c	15	Instruction	Type 2 - Drawing text-related conclusions
reg70340_sc2g7_c	16	Instruction	Type 2 - Drawing text-related conclusions
reg70350_sc2g7_c	17	Instruction	Type 2 - Drawing text-related conclusions
reg70360_sc2g7_c	18	Instruction	Type 3 - Reflecting and Assessing
reg70410_sc2g7_c	19	Literary	Type 2 - Drawing text-related conclusions
reg70420_sc2g7_c	20	Literary	Type 2 - Drawing text-related conclusions
reg70430_sc2g7_c	21	Literary	Type 2 - Drawing text-related conclusions
reg70440_sc2g7_c	22	Literary	Type 2 - Drawing text-related conclusions
reg7045s_sc2g7_c	23	Literary	Type 3 - Reflecting and Assessing
reg70460_sc2g7_c	24	Literary	Type 3 - Reflecting and Assessing
reg7051s_sc2g7_c	25	Commenting	Type 2 - Drawing text-related conclusions
reg70520_sc2g7_c	26	Commenting	Type 2 - Drawing text-related conclusions
reg7053s_sc2g7_c	27	Commenting	Type 2 - Drawing text-related conclusions
reg70540_sc2g7_c	28	Commenting	Type 1 - Finding Information in Text
reg7055s_sc2g7_c	29	Commenting	Type 3 - Reflecting and Assessing
reg70560_sc2g7_c	30	Commenting	Type 3 - Reflecting and Assessing
reg70610_sc2g7_c	31	Advertising	Type 1 - Finding Information in Text
reg70620_sc2g7_c	32	Advertising	Type 1 - Finding Information in Text
reg7063s_sc2g7_c	33	Advertising	Type 1 - Finding Information in Text

reg70640_sc2g7_c	34	Advertising	Type 2 - Drawing text-related conclusions
reg70650_sc2g7_c	35	Advertising	Type 1 - Finding Information in Text
reg7066s_sc2g7_c	36	Advertising	Type 3 - Reflecting and Assessing
reg70670_sc2g7_c	37	Advertising	Type 3 - Reflecting and Assessing
reg7071s_sc2g7_c	38	Commenting	Type 2 - Drawing text-related conclusions
reg70720_sc2g7_c	39	Commenting	Type 2 - Drawing text-related conclusions
reg70730_sc2g7_c	40	Commenting	Type 3 - Reflecting and Assessing
reg70740_sc2g7_c	41	Commenting	Type 3 - Reflecting and Assessing
reg7075s_sc2g7_c	42	Commenting	Type 3 - Reflecting and Assessing

Note. Position = Item position as used in the scaling of the test.

Appendix B: R Syntax for estimating WLE estimates in Starting Cohort 2

```
library(dplyr)
library(TAM)

# First the data has to be loaded from the competence file in the SUF.
# Then the items split by test versions have to be generated because
# they are not split in the SUF.

# Given the object 'dat' including the SUF competence data with the
# items split by test version, the relevant variables are selected.
dat <- dat %>%
         select(ID_t, [INSERT VARIABLE NAMES])

# Collapse response categories according to SC2 G7 scaling
dat <- dat %>%
  mutate(
    reg7013s_sc2g7_c = recode(as.numeric(reg7013s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 0, `3` = 1,
                              .default = NA_real_),
    reg7015s_sc2g7_c = recode(as.numeric(reg7015s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 1,
                              .default = NA_real_),
    reg7016s_sc2g7_c = recode(as.numeric(reg7016s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 1, `3` = 1, `4` = 2,
                              .default = NA_real_),
    reg7023s_sc2g7_c = recode(as.numeric(reg7023s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 1, `3` = 2,
                              .default = NA_real_),
    reg7026s_sc2g7_c = recode(as.numeric(reg7026s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 0, `3` = 0, `4` = 0,
                              `5` = 1L, .default = NA_real_),
    reg7051s_sc2g7_c = recode(as.numeric(reg7051s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 0, `3` = 1,
                              .default = NA_real_),
    reg7053s_sc2g7_c = recode(as.numeric(reg7053s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 0, `3` = 1,
                              .default = NA_real_),
    reg7063s_sc2g7_c = recode(as.numeric(reg7063s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 0, `3` = 1,
                              .default = NA_real_),
    reg7066s_sc2g7_c = recode(as.numeric(reg7066s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 0, `3` = 0, `4` = 1,
                              .default = NA_real_),
    reg7071s_sc2g7_c = recode(as.numeric(reg7071s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 0, `3` = 1,
                              .default = NA_real_),
    reg7075s_sc2g7_c = recode(as.numeric(reg7075s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 0, `3` = 1, `4` = 1,
                              .default = NA_real_),
    reg7045s_sc2g7_c = recode(as.numeric(reg7045s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 1, `3` = 1,
                              .default = NA_real_),
    reg7045s_sc2g7_c_d = recode(as.numeric(reg7045s_sc2g7_c_d),      # item split
                                `0` = 0, `1` = 0, `2` = 1, `3` = 1,  # by test
                                .default = NA_real_),                # version
    reg7024s_sc2g7_c = recode(as.numeric(reg7024s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 1,
                              .default = NA_real_),
    reg7024s_sc2g7_c_d = recode(as.numeric(reg7024s_sc2g7_c_d),      # item split
                                `0` = 0, `1` = 0, `2` = 1,           # by test
                                .default = NA_real_),                # version
    reg7033s_sc2g7_c = recode(as.numeric(reg7033s_sc2g7_c),
                              `0` = 0, `1` = 0, `2` = 1,
                              .default = NA_real_),
    reg7033s_sc2g7_c_d = recode(as.numeric(reg7033s_sc2g7_c_d),      # item split
                                `0` = 0, `1` = 0, `2` = 1,           # by test
                                .default = NA_real_)                 # verson
  )

# Identify polytomous items
poly <- (apply(dat[, -1], 2, max, na.rm = TRUE) > 1) + 1
poly[poly == 2] <- 0.5

# Estimate model for cross-sectional analyses
mod <- tam.mml(resp = dat, irtmodel = "PCM2",
               Q = as.matrix(poly),
```

```
                pid = dat$ID_t, verbose = FALSE)

# Estimate WLEs
wle <- tam.wle(mod, Msteps = 500)
```