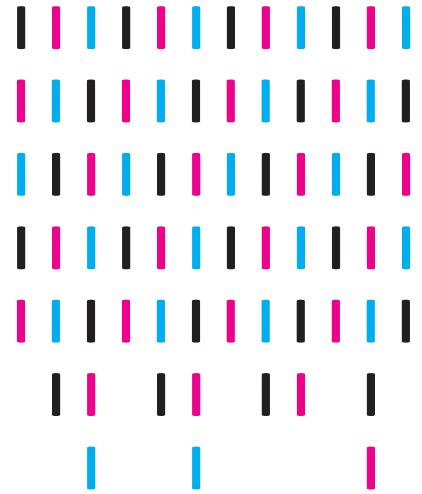




NEPS *SURVEY PAPERS*

Timo Gnambs

NEPS TECHNICAL REPORT
FOR READING:
SCALING RESULTS OF
STARTING COHORT 8
FOR GRADE 5

NEPS *Survey Paper* No. 117
Bamberg, March 2025

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.

Editor-in-Chief: Thomas Bäumer, LIfBi

Review Board: Board of Directors, Heads of LIfBi Departments, and Scientific Management of NEPS Working Units

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de



This work is licensed under CC BY 4.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

NEPS Technical Report for Reading: Scaling Results of Starting Cohort 8 for Grade 5

Timo Gnambs 

Leibniz Institute for Educational Trajectories

E-mail address of the lead author:

timo.gnambs@lifbi.de

Bibliographic data:

Gnambs, T. (2025). *NEPS technical report for reading: Scaling results of Starting Cohort 8 for grade 5* (NEPS Survey Paper No. 117). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP117:1.0>

Acknowledgements. This report is an extension to *NEPS Working Paper 15* (Pohl et al., 2012) that presents the scaling results for reading in Grade 5 of Starting Cohort 3. Therefore, various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced verbatim to facilitate the understanding of the presented results.

NEPS Technical Report for Reading: Scaling Results of Starting Cohort 8 for Grade 5

Abstract

The National Educational Panel Study (NEPS) examines the development of competencies across the lifespan. Therefore, the NEPS develops tests to assess different domains of competence in different age cohorts. To evaluate the quality of these competence tests, several analyses based on item response theory are conducted. This paper describes the data and scaling procedures for a test of reading competence that was administered in Grade 5 of Starting Cohort 8. The reading test consisted of 33 items representing different cognitive requirements and text types, and using different response formats. The test was administered to 5,424 students (50% girls) from regular schools. A partial credit model was used to scale the data. Item fit statistics, differential item functioning, Rasch-homogeneity, test dimensionality, and local item independence were evaluated to ensure the quality of the test. The analyses showed that the items had good item fit and negligible differential item functioning across different subgroups. In addition, the test showed a high reliability and the different cognitive requirements supported a unidimensional construct. Limitations of the test included many items that were not reached by test takers due to time constraints, few items targeted at a higher reading ability, and some evidence of multidimensionality based on text types. Overall, however, the test had good psychometric properties that allowed the estimation of reliable reading competence scores. Furthermore, analyses of differential item functioning showed comparable measurement models for the test administered in Starting Cohort 8 and a comparable test previously used in Starting Cohort 3. Competence scores from both tests were therefore linked to allow comparison between the two cohorts. In addition to the scaling results, this report also describes the data available in the scientific use file and presents the R syntax for scaling the data.

Keywords

item response theory, scaling, reading competence, scientific use file

Contents

1	Introduction	4
2	Testing Reading Competence	4
2.1	Conceptual Framework	4
2.2	Item Format	4
2.3	Study Design	6
3	Sample	6
4	Psychometric Analyses	6
4.1	Missing Responses	6
4.2	Scaling Model	7
4.3	Checking the Quality of the Test	7
4.4	Software	9
5	Results	9
5.1	Missing Responses	9
5.1.1	Missing Responses per Person	9
5.1.2	Missing Responses per Item	12
5.2	Parameter Estimates	14
5.2.1	Item Parameters	14
5.2.2	Test Targeting and Reliability	16
5.3	Quality of the test	17
5.3.1	Distractor Analyses	17
5.3.2	Item Fit	17
5.3.3	Differential Item Functioning	17
5.3.4	Rasch-homogeneity	24
5.3.5	Dimensionality	24
6	Discussion	26
7	Data in the Scientific Use Files	26
7.1	Naming Conventions	26
7.2	Linking of the Competence Tests	26
7.3	Reading Competence Scores	27
	References	29
	Appendix A: Cognitive Requirements and Text Types	31
	Appendix B: R Syntax for Estimating WLEs	33

1. Introduction

The National Educational Panel Study (NEPS)¹ measures different competencies coherently across the lifespan. These include, among others, reading competence, mathematical competence, and domain-general cognitive functioning. An overview of the competencies measured in the NEPS is given by Weinert et al. (2011) and Fuß et al. (2025). Most of the administered competence tests are developed specifically for use in the NEPS and are therefore routinely evaluated using psychometric models based on item response theory (IRT, see Pohl & Carstensen, 2012).

This report presents the results of these analyses for a test of reading competence that was administered in fifth grades of regular schools in Starting Cohort 8 (fifth grade). The following sections first introduce the main concepts of the administered test and the test design. Then, the competence data are described. Then, the results of various psychometric analyses are reported that were conducted to estimate competence scores and to check the quality of the test. Finally, an overview of the data that are available for public use in the scientific use file (SUF) is given.

The analyses summarized in this report are based on the data available at some point prior to the public data release. Due to ongoing data protection and data cleansing issues, the data in the SUF may differ slightly from the data used for the analyses in this report. However, pronounced differences in the presented results are not expected for the final data available in the SUF.

2. Testing Reading Competence

The framework and test development for the reading competence test are described in Gehrer and Artelt (2013) and Gehrer et al. (2013). Because the administered test is identical to the reading competence test previously administered in Grade 5 of Starting Cohort 3, additional information on the test is provided in Pohl et al. (2012). In the following, several aspects of the reading test are described that are important for understanding the scaling results presented in this report.

2.1 Conceptual Framework

The reading test administered included five texts and five sets of items referring to these texts. Each of these texts represented one text type, namely, a) information, b) commenting or argumenting, c) literary, d) instruction, and e) advertising. The reading test also 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 can be found in Gehrer and Artelt (2013) and Gehrer et al. (2013).

2.2 Item Format

The reading competence test included three types of response formats, namely, simple multiple choice (MC) items, complex multiple choice (CMC) items, and matching (MA) items. MC

¹<https://www.neps-data.de>

items had four response options, of which one option represented a correct solution, whereas the other three were distractors (i.e., they were incorrect). CMC items presented a number of subtasks with two response options that required evaluating whether they were correct or incorrect according to the information in the text. MA items required the respondents to match a number of responses to a given set of statements. MA items were usually used to assign headings to paragraphs of a text. The number of subtasks within CMC and MA items varied from four to eight. Examples of the different response formats are given in Gehrler et al. (2012) and Pohl and Carstensen (2012).

The reading test included 33 items. Because preliminary analyses revealed an unsatisfactory fit for one item (reg5045s_sc8g5_c), this item was excluded from the scaling procedure. Therefore, the results presented in this paper refer to 32 items. As shown in Table 1, most of these were MC items, while more complex item types such as CMC and MA items were less common. Table 2 and Table 3 show the distribution of the items across the two dimensions of the assessment framework, that is, the five text types and the three cognitive requirements (see Section 2.1). The text types and cognitive requirements were approximately equally distributed across the administered items. The allocation of each item to these dimensions is provided in Appendix A.

Table 1

Number of Items by Response Formats

Response format	Number of items
Simple multiple-choice items	26
Complex multiple-choice items	3
Matching items	3
Total number of items	32

Table 2

Number of Items by Text Types

Text types	Number of items
Information text	7
Instruction text	6
Advertising text	7
Commenting or argumenting text	5
Literary text	7
Total number of items	32

Table 3*Number of Items by Cognitive Requirements*

Cognitive requirement	Number of items
Finding information in the text	8
Drawing text-related conclusions	13
Reflecting and assessing	11
Total number of items	32

2.3 Study Design

The study was conducted in small groups at the respective schools of the students. The tests were presented on paper by trained test administrators. The study assessed different competence domains. These included reading and mathematical competence. In order to control for test position effects, the two tests were assigned to test takers in different order. Half of the participants were given a booklet that included the reading test followed by the mathematics test, while the other half were given the two tests in reverse order. The allocation of the two booklets was randomized. There was no multi-matrix design with respect to the order of the items within the test. All respondents received the test items in the same order. The testing time was limited to 28 minutes. A comprehensive description of the study design is available on the NEPS website².

3. Sample

A total of 5,424 participants (50% girls) were administered the reading competence test. A random subsample of 2,775 participants was given the reading test before working on the mathematics test, while 2,649 participants worked on the two tests in the reverse order. However, 25 participants had less than three valid responses to the test items. Because no reliable competence scores can be estimated from such a small number of responses, these cases were excluded from the psychometric analyses (see Pohl & Carstensen, 2012). Therefore, the analyses presented in this report are based on 5,399 respondents.

4. Psychometric Analyses

4.1 Missing Responses

Competence data contain different types of missing responses. These are missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach, and, finally, d) multiple types of missing responses within CMC and MA items that are not determined.

Invalid responses occurred, for example, when two response options were selected in MC items although 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 respondents skipped

²<https://www.neps-data.de>

some items. Due to time constraints or lack of motivation, not all participants completed the test within the allotted time. All missing responses after the last valid response were coded as not reached. Because CMC and MA items were aggregated from several subtasks, different types of missing responses or a mixture of valid and missing responses could be found. A CMC or MA item was coded as missing if at least one subtask contained a missing response. If there was only one type of missing response, the item was coded according to the type of missing response. If the subtasks contained different types of missing responses, the item was coded as a not-determinable missing response.

Missing responses provide information about how well the test worked (e.g., time limits, understanding of instructions, dealing with different response formats). Therefore, the occurrence of missing responses in the test was evaluated to get an idea of how well respondents coped 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) with Gauss-Hermite quadrature (21 nodes). A detailed description of the scaling model can be found in Pohl and Carstensen (2012). CMC items consisted of a set of subtasks, which were aggregated into a polytomous variable for each item, indicating the number of correctly solved subtasks within that item. If at least one of the subtasks contained a missing response, the partial credit item was scored as missing. Response categories of polytomous items with few responses were collapsed in order to avoid possible estimation problems. These categories were collapsed in the same way as in Grade 5 of Starting Cohort 3 (see Pohl et al., 2012). This was usually done for the lower categories of polytomous items. Appendix B shows how response categories were collapsed for the different items.

Reading competencies were estimated as weighted likelihood estimates (WLE, Warm, 1989). To estimate item and person parameters, a score of 0.5 points was applied to each category of the polytomous items, while simple multiple-choice items were scored dichotomously as 0 for an incorrect and 1 for a correct response. Studies on the scoring of different response formats are reported in Pohl and Carstensen (2012) and Haberkorn et al. (2016). An exception to these scoring rules was item `reg5026s_sc8g5_c`, which consisted of eight subtasks, but was scored as 1 if all subtasks were solved correctly and 0 otherwise to account for extreme local stochastic dependence (see Pohl et al., 2012, for a comparable procedure in Starting Cohort 3).

Ability estimates for reading competence were estimated as weighted likelihood estimates (WLEs, Warm, 1989) and can also be calculated as plausible values (see Mislevy, 1991) using the R package *NEPSscaling*³ (Scharl et al., 2020; Scharl & Zink, 2022). Person parameter estimation in the NEPS is described in more detail in Pohl and Carstensen (2012), while the data available in the SUF are described in Section 7.

4.3 Checking the Quality of the Test

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

³<https://www.neps-data.de/Data-Center/Overview-and-Assistance/Plausible-Values>

The MC items consisted of one correct response option and several distractors (i.e., incorrect response options). The quality of the distractors within the multiple-choice items was examined to evaluate whether the distractors were predominantly chosen by respondents with a lower ability rather than by those who gave a correct response. We therefore calculated the point-biserial correlation between selecting an incorrect response option for a given item and the total correct score for the remaining items. Negative correlations indicate good distractors, while correlations between .00 and .05 were considered acceptable and correlations above .05 were considered problematic distractors (Pohl & Carstensen, 2012).

Before aggregating the subtasks of a CMC and MA item into a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the multiple-choice items in a Rasch (1960) model. The fit of the subtasks was evaluated based on the weighted mean square error (WMNSQ), the respective *t*-value, and the item characteristic curves. Only when the subtasks showed a satisfactory item fit, were they used to construct polytomous variables that were included in the final scaling model.

After aggregating the subtasks into polytomous variables, the fit of the dichotomous and polytomous items to the PCM (Masters, 1982) was evaluated using the WMNSQ statistic, the respective *t*-value, and the item characteristic curves (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (*t*-value > |6|) were considered to have a noticeable item misfit, and items with a WMNSQ > 1.20 (*t*-value > |8|) were considered to have a considerable item misfit and their performance was investigated further. Correlations of the item score with the corrected total score greater than .30 were considered as good, those greater than .20 as acceptable, and those less than .20 as problematic. The overall judgment of item fit was based on all fit indicators.

The reading competence test should measure the same construct for all respondents. If some items favored certain subgroups (e.g., they were easier for boys than for girls), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., boys and girls) would be biased. In the present study, differential item functioning (DIF) was investigated for the variables sex, the number of books at home (as a proxy for cultural status), migrant background, and test position (first versus second). In addition, we compared the reading test administered in Grade 5 of Starting Cohort 3 (see Pohl et al., 2012) and Starting Cohort 8 to examine whether competence scores could be compared between the two cohorts. DIF was examined using a multigroup item response model, in which the main effects of the subgroups as well as the differential effects of the subgroups on item difficulty were modeled. Based on experience 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 worthy of further investigation, differences between 0.4 and 0.6 as small but not severe, and differences less than 0.4 as negligible DIF. In addition, measurement invariance was examined by comparing the fit of a model that acknowledged DIF to a model that included only main effects and no DIF.

The reading test was scaled using the PCM (Masters, 1982) because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nevertheless, 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 also fitted to the data and compared to the

PCM.

The dimensionality of the test was evaluated by examining the residuals of the PCM. Approximately zero-order correlations, as indicated by Yen's (1984) Q_3 , suggest unidimensionality. Because the Q_3 tends to be slightly negative in the case of locally independent items, we report the corrected Q_3 , which has an expected value of 0. According to Yen (1993), values of Q_3 falling below .20 indicate essential unidimensionality. In addition, we estimated two multidimensional models using quasi-Monte Carlo integration (with 2,000 nodes) that specified different subdimensions based on the different construction rationales for the test (see Section 2.1), that is, a model with three subdimensions representing the three cognitive requirements and a model with five subdimensions based on the different text types (see Appendix A). The correlations between the subdimensions and the differences in model fit between the unidimensional and multidimensional models were used to evaluate the unidimensionality of the test.

4.4 Software

The item response models were estimated with the *TAM* package (Robitzsch et al., 2024) in *R* (R Core Team, 2024).

5. Results

5.1 Missing Responses

5.1.1 Missing Responses per Person

The number of invalid responses per person is shown in Figure 1. The number of invalid responses was rather low with 95% of respondents having no invalid responses at all. Only about 1% of respondents had more than one invalid response.

Figure 2 illustrates the number of missing responses when respondents omitted items. The figure shows that a non-negligible number of items were omitted. Approximately 58% of respondents had no omitted items, while 22% exhibited a single omitted item. However, less than 5% of respondents had five or more omitted items.

Another source of missing responses is items that respondents did not reach because they aborted the test, for example, due to time constraints or lack of motivation. These missing values refer to items after the last valid response. As shown in Figure 3, only about 28% of respondents did not abort the test and were administered all items. However, more than 97% of the sample received at least 11 items and thus about a third of the total test.

For the aggregated polytomous variables, responses were coded as not determinable missing if the subtasks of the CMC and MA items contained different types of missing responses. Figure 4 shows the percentage of not-determinable missing responses in the test. Because not-determinable missing responses can only occur in CMC and MA items, the maximum number of not-determinable missing responses was 6 (see Table 1). There was only a rather small number of not-determinable missing responses. About 98% of the respondents did not have a single not determinable missing response.

The total number of missing responses, aggregated across invalid, omitted, not reached, and not-determinable missing responses for each respondent, is shown in Figure 5. Because most

Figure 1
Number of Invalid Responses

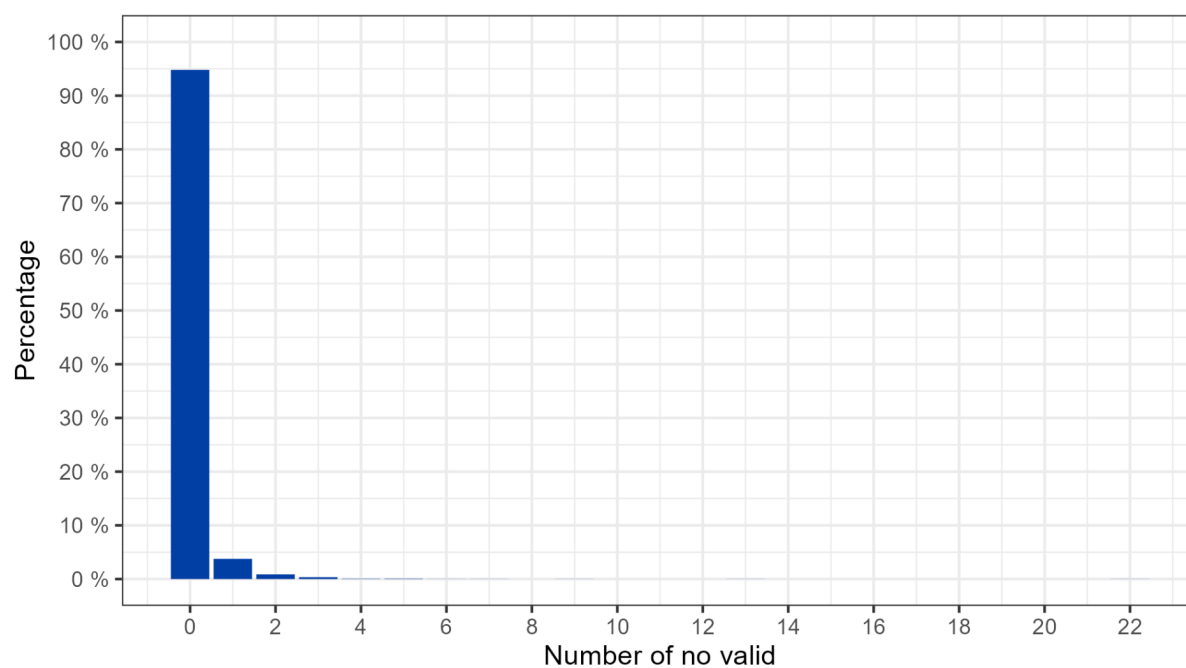


Figure 2
Number of Omitted Items

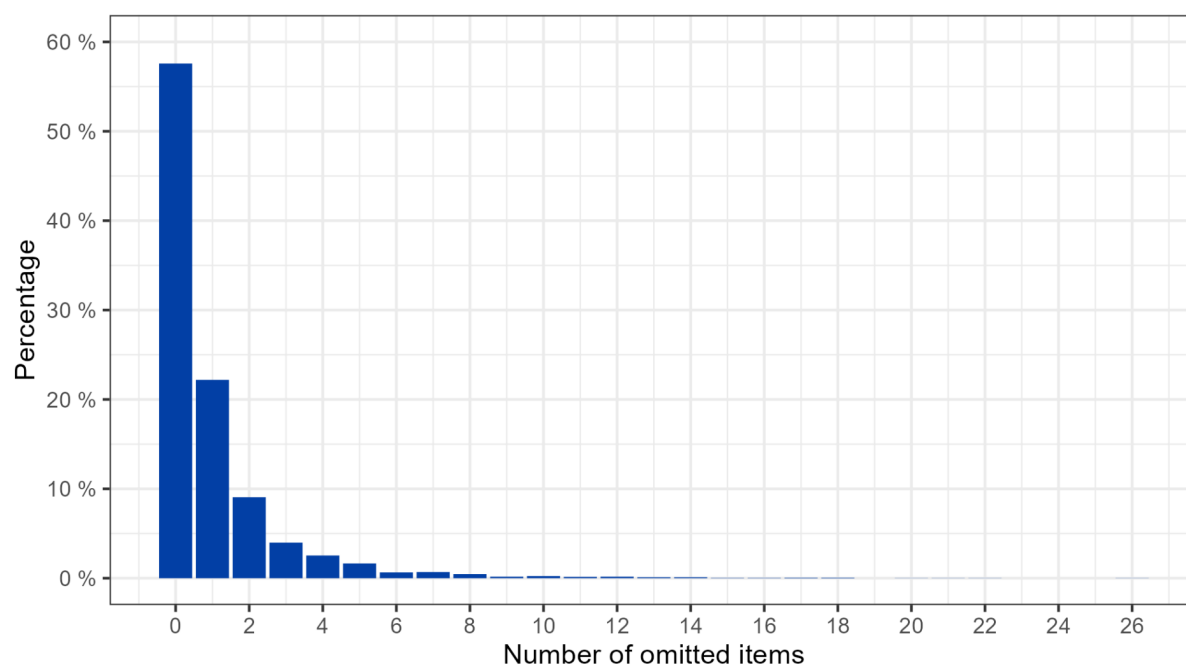
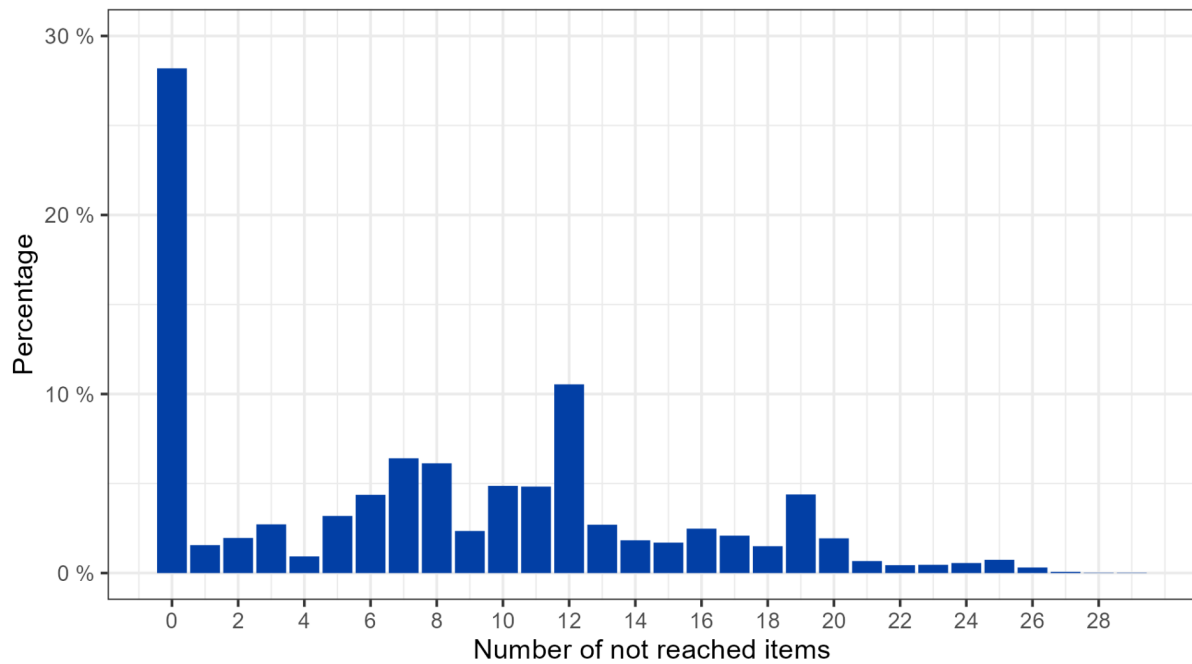
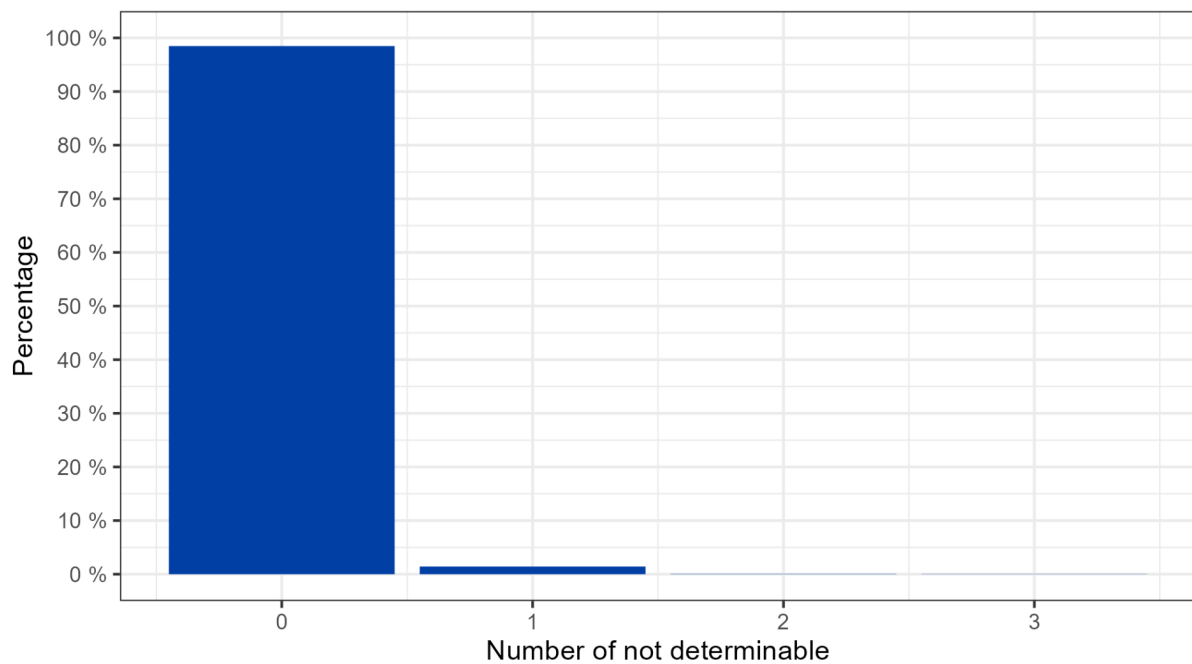
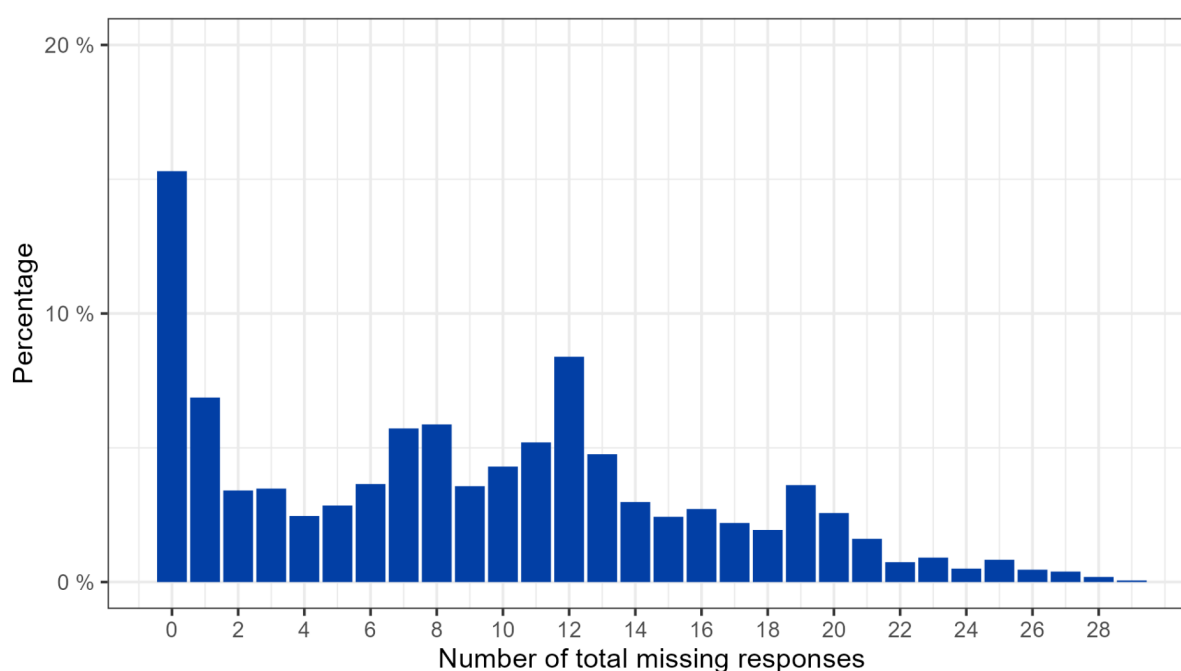


Figure 3*Number of Not-Reached Items***Figure 4***Number of Not-Determinable Missing Responses*

respondents did not reach the end of the test or omitted some items, the total number of missing values was substantial. The median number of missing responses was 9; only about 15% of the respondents had no missing response at all. Almost 66% of respondents had more than five missing responses, while 19% had missing responses for 16 or more items (i.e., almost half of the administered items).

Figure 5

Total Number of Missing Responses



Overall, there was a small amount of invalid and not-determinable missing responses, while the percentage of omitted items was reasonable. However, the total number of missing responses was rather large because many respondents did not reach the end of the test.

5.1.2 Missing Responses per Item

Table 4 provides information on the occurrence of different types of missing responses per item. Overall, the number of respondents that omitted an item was acceptable. The percentage of omitted responses per item varied between 0% and 12% (*Mdn* = 2%). The highest omission rate was observed for the CMC item reg5012s_sc8g5_c. This was probably because it was the first CMC item in the test and respondents had not (yet) fully understood the instructions on how to use this response format. In addition, most items had relatively few invalid responses. The percentage of invalid responses varied across items between 0% and 1% (*Mdn* = 0%).

Table 4*Percentage of Missing Responses by Item*

Nr.	Item	Pos.	N	OM	NR	NV
1	reg50110_sc8g5_c	1	5,350	0.63	0.00	0.28
2	reg5012s_sc8g5_c	2	4,721	11.85	0.00	0.52
3	reg50130_sc8g5_c	3	5,337	0.87	0.00	0.28
4	reg50140_sc8g5_c	4	5,143	4.39	0.02	0.33
5	reg50150_sc8g5_c	5	5,228	2.35	0.04	0.78
6	reg5016s_sc8g5_c	6	4,846	9.67	0.11	0.22
7	reg50170_sc8g5_c	7	5,205	1.76	0.43	1.41
8	reg50210_sc8g5_c	8	5,269	1.06	1.17	0.19
9	reg50220_sc8g5_c	9	4,954	6.26	1.72	0.26
10	reg50230_sc8g5_c	10	5,164	2.09	2.19	0.07
11	reg50240_sc8g5_c	11	5,098	1.89	2.63	1.06
12	reg50250_sc8g5_c	12	5,059	2.91	3.30	0.09
13	reg5026s_sc8g5_c	13	4,610	8.91	5.24	0.09
14	reg50310_sc8g5_c	14	4,704	3.17	9.63	0.07
15	reg50320_sc8g5_c	15	4,684	2.00	11.13	0.11
16	reg50330_sc8g5_c	16	4,585	1.70	13.22	0.15
17	reg50340_sc8g5_c	17	4,399	2.65	15.71	0.17
18	reg50350_sc8g5_c	18	4,380	1.33	17.41	0.13
19	reg50360_sc8g5_c	19	4,242	2.15	19.24	0.04
20	reg50370_sc8g5_c	20	4,058	2.80	21.95	0.09
21	reg50410_sc8g5_c	21	3,490	2.72	32.49	0.15
22	reg5042s_sc8g5_c	22	3,162	3.87	37.32	0.15
23	reg50430_sc8g5_c	23	2,910	3.74	42.19	0.17
24	reg50440_sc8g5_c	24	2,773	3.95	44.55	0.15
25	reg50460_sc8g5_c	26	2,449	3.56	50.68	0.41
26	reg50510_sc8g5_c	27	2,266	0.81	57.08	0.13
27	reg5052s_sc8g5_c	28	1,893	2.87	61.46	0.17
28	reg50530_sc8g5_c	29	1,832	1.30	64.64	0.13

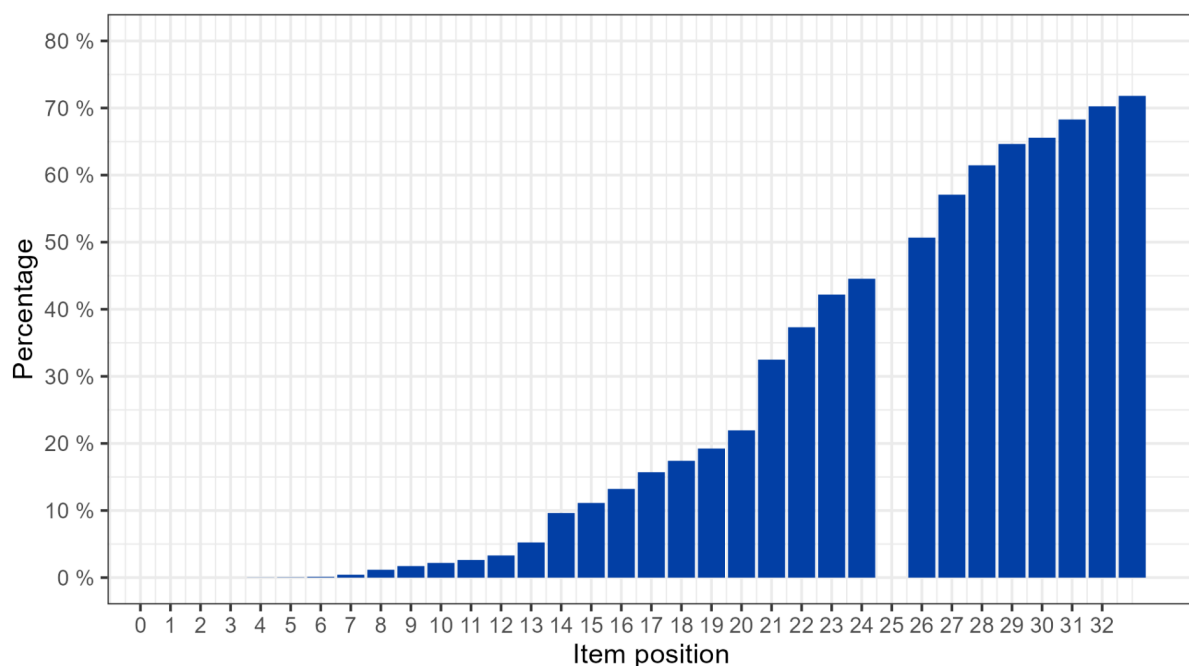
Nr.	Item	Pos.	N	OM	NR	NV
29	reg50540_sc8g5_c	30	1,803	0.94	65.57	0.09
30	reg5055s_sc8g5_c	31	1,551	2.65	68.29	0.02
31	reg50560_sc8g5_c	32	1,519	1.46	70.25	0.15
32	reg50570_sc8g5_c	33	1,513	0.00	71.81	0.17

Note. Pos. = Item position within the test. N = Number of valid responses. NR = Percentage of respondents that did not reach an item. OM = Percentage of omitted responses. NV = Percentage of invalid responses.

In contrast, there was a substantial number of items that respondents did not reach. The percentage of missing responses because participants did not reach an item was $Mdn = 14\%$. The percentage of respondents not reaching an item increased with the position of the item in the test (see Figure 6) up to 72%. Because a relatively large number of items was not reached by respondents, the total number of missing responses per item was also quite large. These varied between 1% and 72% ($Mdn = 17\%$).

Figure 6

Item Position not Reached



Note. The item on position 25 was excluded due to an unsatisfactory item fit.

5.2 Parameter Estimates

5.2.1 Item Parameters

The percentage of correct responses out of all valid responses for each multiple-choice item administered varied between 37.34% and 95.70% and was therefore quite high (see Table 5).

Because there was a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index of item difficulty.

The estimated item difficulties (for dichotomous items) and location parameters (for polytomous items) are given in Table 5, while the corresponding step parameters for polytomous variables are summarized in Table 6. Item difficulties and location parameters were estimated by constraining the mean of the ability distribution to zero. The estimated item difficulties ranged from -3.83 (reg50110_sc8g5_c) to 0.61 (reg50170_sc8g5_c) with a median of -1.19, thus covering a range of items that varied from easy to medium. However, none of the items exhibited large difficulties. The standard errors (*SE*) of the estimated parameters were rather small for most items ($SE \leq 0.07$).

Table 5

Item Parameters

Nr.	Item	<i>N</i>	Percentage correct	Difficulty	<i>SE</i>	WMNSQ	<i>t</i>	r_{it}	Q_3	Discr.
1	reg50110_sc8g5_c	5,350	95.70	-3.83	0.07	0.91	-1.56	0.21	0.05	2.30
2	reg5012s_sc8g5_c	4,721		-1.65	0.03	0.84	-4.30	0.28	0.07	1.43
3	reg50130_sc8g5_c	5,337	88.57	-2.62	0.05	0.96	-1.41	0.22	0.04	1.56
4	reg50140_sc8g5_c	5,143	83.03	-2.06	0.04	0.98	-1.00	0.28	0.03	1.50
5	reg50150_sc8g5_c	5,228	68.40	-1.01	0.03	1.00	0.18	0.29	0.03	1.32
6	reg5016s_sc8g5_c	4,846		-0.84	0.01	1.04	1.29	0.31	0.05	0.69
7	reg50170_sc8g5_c	5,205	39.10	0.61	0.03	0.99	-0.98	0.38	0.03	1.33
8	reg50210_sc8g5_c	5,269	90.78	-2.91	0.05	0.95	-1.42	0.27	0.03	1.61
9	reg50220_sc8g5_c	4,954	54.26	-0.23	0.03	1.24	15.65	0.21	0.04	0.59
10	reg50230_sc8g5_c	5,164	91.54	-3.01	0.05	0.91	-2.45	0.25	0.05	2.04
11	reg50240_sc8g5_c	5,098	80.64	-1.85	0.04	0.95	-2.35	0.30	0.03	1.65
12	reg50250_sc8g5_c	5,059	74.24	-1.38	0.04	1.02	1.00	0.29	0.02	1.22
13	reg5026s_sc8g5_c	4,610		0.40	0.03	0.90	-10.25	0.33	0.04	1.34
14	reg50310_sc8g5_c	4,704	85.99	-2.34	0.05	0.90	-3.48	0.35	0.04	1.92
15	reg50320_sc8g5_c	4,684	90.33	-2.83	0.05	0.88	-3.38	0.32	0.05	2.24
16	reg50330_sc8g5_c	4,585	87.72	-2.52	0.05	0.91	-2.83	0.34	0.03	1.82
17	reg50340_sc8g5_c	4,399	80.97	-1.88	0.04	0.94	-2.69	0.34	0.04	1.63
18	reg50350_sc8g5_c	4,380	62.08	-0.63	0.04	1.09	5.28	0.28	0.03	1.00
19	reg50360_sc8g5_c	4,242	87.79	-2.53	0.05	0.98	-0.48	0.28	0.04	1.47
20	reg50370_sc8g5_c	4,058	74.77	-1.41	0.04	1.03	1.47	0.33	0.03	1.17

Nr.	Item	N	Percentage correct	Difficulty	SE	WMNSQ	t	r_{it}	Q_3	Discr.
21	reg50410_sc8g5_c	3,490	65.87	-0.87	0.04	1.12	6.05	0.36	0.03	0.98
22	reg5042s_sc8g5_c	3,162		-1.12	0.02	1.12	4.19	0.33	0.04	0.51
23	reg50430_sc8g5_c	2,910	42.99	0.39	0.04	0.98	-1.16	0.43	0.03	1.30
24	reg50440_sc8g5_c	2,773	53.44	-0.21	0.05	1.07	3.53	0.39	0.03	1.07
25	reg50460_sc8g5_c	2,449	57.98	-0.48	0.05	1.08	3.52	0.42	0.03	1.09
26	reg50510_sc8g5_c	2,266	84.86	-2.45	0.07	0.90	-2.57	0.46	0.04	1.92
27	reg5052s_sc8g5_c	1,893		-0.70	0.03	0.84	-5.19	0.64	0.04	0.98
28	reg50530_sc8g5_c	1,832	37.34	0.58	0.06	1.20	6.97	0.35	0.03	0.77
29	reg50540_sc8g5_c	1,803	67.61	-1.26	0.06	1.00	-0.09	0.51	0.04	1.37
30	reg5055s_sc8g5_c	1,551		-0.67	0.03	0.94	-1.93	0.60	0.03	0.79
31	reg50560_sc8g5_c	1,519	53.26	-0.47	0.06	1.31	9.42	0.34	0.04	0.72
32	reg50570_sc8g5_c	1,513	57.57	-0.77	0.06	0.99	-0.30	0.55	0.04	1.35

Note. N = Number of observed responses. Difficulty = Item difficulty. SE = Standard error of item parameter. WMNSQ = Weighted mean square error. t = t-value for WMNSQ. Discr. = Discrimination parameter of a two-parametric item response model. r_{it} = Corrected item-total correlation. Q_3 = Average absolute residual correlation for item (Yen, 1983).

Table 6

Step Parameters (with Standard Errors) for Polytomous Items

Item	Step 1	Step 2	Step 3	Step 4	Step 5
reg5012s_sc8g5_c	1.43 (0.07)	-1.43			
reg5016s_sc8g5_c	0.35 (0.03)	0.17 (0.04)	0.49 (0.04)	-0.15 (0.05)	-0.86
reg5042s_sc8g5_c	-0.14 (0.04)	-0.06 (0.04)	0.20		
reg5052s_sc8g5_c	0.46 (0.05)	-0.30 (0.06)	-0.16		
reg5055s_sc8g5_c	-0.38 (0.05)	0.11 (0.06)	0.27		

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. Because some items in the reading competence test were polytomous, we calculated Thurstonian thresholds for each

response category (Adam et al., 2023). These indicate the location on the latent dimension where the probability of scoring above the threshold is 50%. This 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 respondents are plotted on the same scale. The distribution of respondents' estimated abilities is plotted on the left, while the distribution of the category thresholds is plotted on the right.

The category thresholds ranged from -3.83 (reg50110_sc8g5_c) to 0.79 (reg5026s_sc8g5_c) and thus covered a rather wide range. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 1.77, which implies a good differentiation between respondents. The reliability of the test (EAP/PV reliability = 0.80, WLE reliability = 0.71) was good. The median of the category threshold distribution was about -1.51 logits below the mean person ability distribution. Thus, although the items covered a wide range of the ability distribution, on average the items were too easy for the respondents. As a result, proficiency estimates in low to medium ability ranges were measured relatively accurately, while higher ability estimates had larger standard errors of measurement.

5.3 Quality of the test

5.3.1 Distractor Analyses

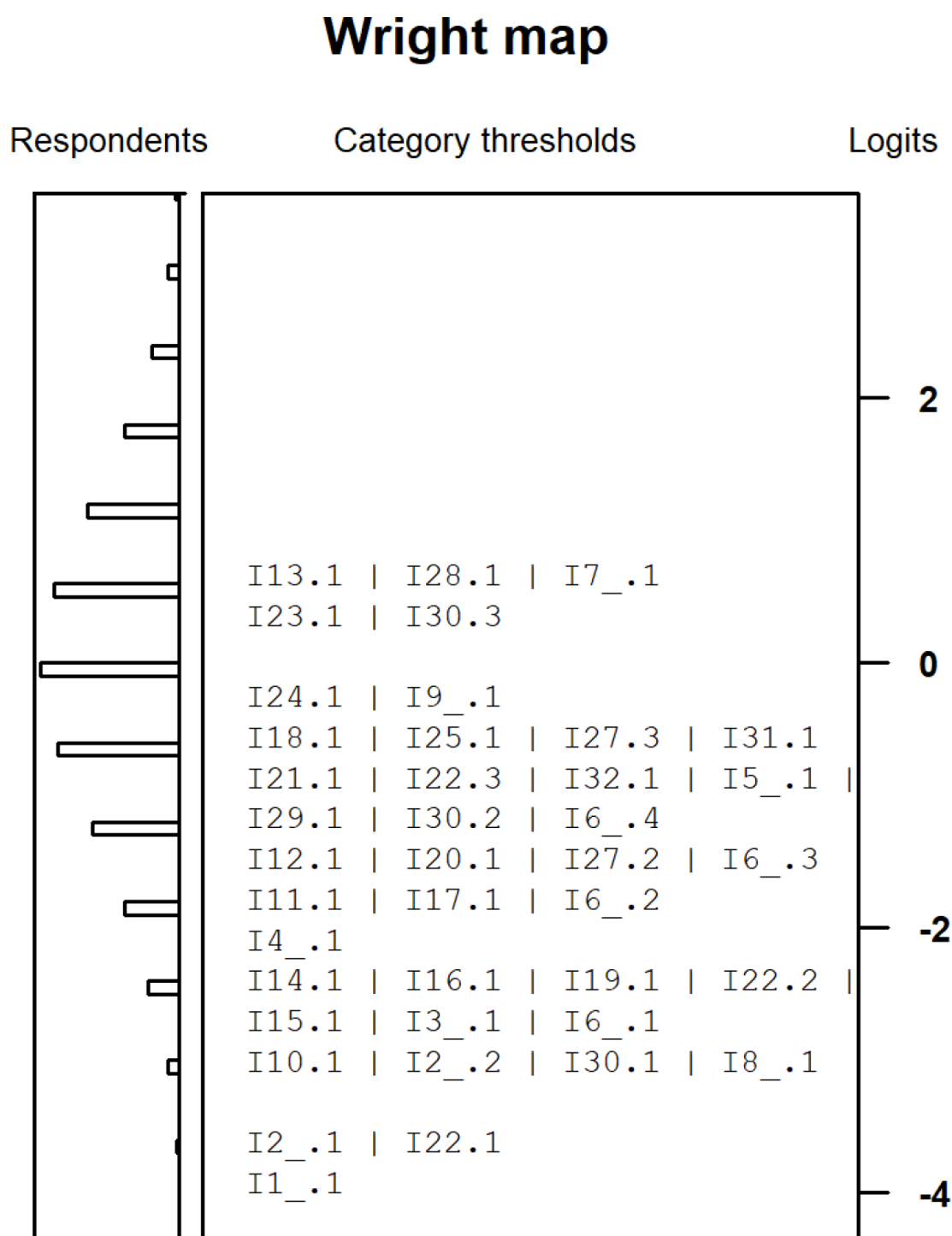
To investigate how well the distractors of the multiple-choice items performed, point-biserial correlations were calculated between each incorrect response (distractor) and the respondents' total correct scores for the remaining items. The median point-biserial correlation for the distractors fell at -.22 (*Min* = -.35, *Max* = -.02). These results indicate that the distractors worked well.

5.3.2 Item Fit

The evaluation of item fit was based on the final scaling model presented above. Overall, the item fit was satisfactory (see Table 5). There were 2 items (reg50220_sc8g5_c, reg50560_sc8g5_c) with a WMSNQ greater than 1.20, while no items had a WMSNQ between 1.15 and 1.20. However, visual inspections of the empirical ICCs for these items did not reveal any pronounced deviation from the expected ICCs. For the remaining items, the WMNSQ values ranged from 0.84 (reg5012s_sc8g5_c, reg5052s_sc8g5_c) to 1.12 (reg50410_sc8g5_c, reg5042s_sc8g5_c), with a median of 0.98. In addition, the correlations of the correct responses with the total test scores varied between 0.21 (reg50110_sc8g5_c, reg50220_sc8g5_c) and 0.64 (reg5052s_sc8g5_c), with a median of 0.33.

5.3.3 Differential Item Functioning

DIF was used to evaluate whether the measurement models were comparable for different subgroups. For this purpose, DIF was examined for the variables sex, migrant background, number of books at home (as a proxy for cultural capital), and test position. A description of these variables can be found in Pohl and Carstensen (2012). The differences between the estimated item difficulties (on the logit scale) in the different groups are summarized in Table 7. For example, the column "boys vs. girls" reports the differences in item difficulties between boys and girls; a positive value would indicate that the test was more difficult for boys, whereas

Figure 7*Test Targeting*

Note. The distribution of the person abilities 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 numbers given in Table 5 and the second part indicating the threshold.

a negative value would indicate that the item was less difficult for boys than for girls. In addition to examining DIF for each item, an overall test for DIF was performed by comparing models that allow for DIF with those that only estimate main effects. In Table 8, models including only main effects were compared with those additionally estimating DIF using Akaike's (1974) information criterion (AIC) and Bayesian information criterion (BIC, Schwarz, 1978).

Sex: The sample included 2,686 boys and 2,712 girls. There were small sex differences in reading competence as indicated by the main effect of -0.28 logits (Cohen's $d = -0.21$) that reflected better reading competencies for girls than boys. One item (reg50250_sc8g5_c) showed a DIF effect of more than 0.60 logits and was more difficult for girls than for boys. An overall test for DIF (see Table 8) also suggested a better fit for the model with DIF compared to a model that only estimated the main effect (but ignored potential DIF). Because an evaluation of the item content, did not suggest an explanation for the observed DIF in the item reg50250_sc8g5_c and the item did not show DIF in Starting Cohort 3 (see Pohl et al., 2012), no sex-specific item effects were acknowledged in the scaling of the tests. These results suggest that for most items there was no pronounced DIF concerning sex that might have biased the parameter estimates.

Migrant background: There were 4,638 participants without a migrant background, 231 participants with a migrant background, and 530 participants with no information on their migrant background. All three groups were used to investigate DIF of migration. There was a considerable difference between the average performance of participants with and without a migrant background. Participants without a migrant background had a higher reading ability than participants with a migrant background (main effect = 0.88 logits, Cohen's $d = 0.68$). Respondents without information on their migrant background also differed from those with no migrant background (main effect = 0.32 logits, Cohen's $d = 0.24$) and those with a migrant background (main effect = -0.57 logits, Cohen's $d = -0.43$). There were 6 items with DIF greater than 0.60 between respondents with and without a migrant background. The largest difference in item difficulties between these groups was 0.95 logits (reg50310_sc8g5_c). Although this appears to be a rather large difference, it may be a consequence of the uncertainty in the estimation due to the small number of respondents with a migrant background. Accordingly, an overall test for DIF using the BIC (see Table 8) suggested a better fit for the main effect model that did not acknowledge DIF effects.

Table 7*Differential Item Functioning*

Item	Sex	Migration				Books		Test position
	boys vs. girls	without vs. with	with vs. na	without vs. na	<100 vs. >100	<100 vs. na	>100 vs. na	first vs. second
reg50110_sc8g5_c	0.04 (0.03)	0.30 (0.23)	0.07 (0.05)	-0.23 (-0.18)	0.15 (0.12)	0.20 (0.16)	0.05 (0.04)	-0.17 (-0.13)
reg5012s_sc8g5_c	-0.03 (-0.02)	0.19 (0.15)	-0.17 (-0.13)	-0.36 (-0.28)	-0.28 (-0.22)	-0.17 (-0.13)	0.11 (0.09)	-0.07 (-0.05)
reg50130_sc8g5_c	0.16 (0.12)	0.14 (0.11)	-0.21 (-0.16)	-0.35 (-0.27)	-0.09 (-0.07)	0.08 (0.06)	0.17 (0.13)	-0.35 (-0.26)
reg50140_sc8g5_c	-0.28 (-0.21)	-0.18 (-0.14)	0.13 (0.10)	0.31 (0.24)	0.03 (0.02)	0.14 (0.11)	0.12 (0.09)	-0.05 (-0.04)
reg50150_sc8g5_c	0.13 (0.10)	-0.30 (-0.23)	0.19 (0.15)	0.49 (0.38)	0.00 (0.00)	-0.04 (-0.03)	-0.04 (-0.03)	-0.14 (-0.11)
reg5016s_sc8g5_c	-0.16 (-0.12)	0.40 (0.31)	0.11 (0.08)	-0.30 (-0.23)	-0.33 (-0.26)	-0.11 (-0.09)	0.22 (0.17)	0.17 (0.13)
reg50170_sc8g5_c	-0.13 (-0.10)	-0.52 (-0.40)	-0.03 (-0.02)	0.50 (0.38)	0.25 (0.19)	0.04 (0.03)	-0.21 (-0.16)	0.24 (0.18)
reg50210_sc8g5_c	0.09 (0.07)	-0.51 (-0.39)	-0.18 (-0.14)	0.33 (0.25)	0.13 (0.10)	-0.24 (-0.19)	-0.38 (-0.30)	-0.09 (-0.07)
reg50220_sc8g5_c	0.42 (0.32)	0.49 (0.38)	0.12 (0.09)	-0.37 (-0.28)	-0.16 (-0.12)	-0.07 (-0.05)	0.09 (0.07)	0.12 (0.09)
reg50230_sc8g5_c	0.17 (0.13)	-0.06 (-0.05)	-0.30 (-0.23)	-0.23 (-0.18)	-0.14 (-0.11)	-0.04 (-0.03)	0.10 (0.08)	-0.59 (-0.45)
reg50240_sc8g5_c	0.27 (0.20)	-0.32 (-0.25)	-0.11 (-0.08)	0.22 (0.17)	0.16 (0.12)	0.04 (0.03)	-0.12 (-0.09)	-0.11 (-0.08)
reg50250_sc8g5_c	0.80 (0.60)	-0.04 (-0.03)	-0.11 (-0.08)	-0.07 (-0.05)	-0.02 (-0.02)	0.00 (0.00)	0.02 (0.02)	0.05 (0.04)
reg5026s_sc8g5_c	0.13 (0.10)	0.23 (0.18)	0.08 (0.06)	-0.15 (-0.11)	-0.03 (-0.02)	0.06 (0.05)	0.10 (0.08)	-0.02 (-0.02)
reg50310_sc8g5_c	-0.21 (-0.16)	-0.95 (-0.73)	-0.13 (-0.10)	0.81 (0.62)	0.27 (0.21)	-0.02 (-0.02)	-0.30 (-0.23)	-0.19 (-0.14)

Item	Sex	Migration			Books			Test position
	boys vs. girls	without vs. with	with vs. na	without vs. na	<100 vs. >100	<100 vs. na	>100 vs. na	first vs. second
reg50320_sc8g5_c	-0.07 (-0.05)	-0.88 (-0.67)	-0.27 (-0.21)	0.60 (0.46)	0.37 (0.29)	0.01 (0.01)	-0.36 (-0.28)	-0.40 (-0.30)
reg50330_sc8g5_c	-0.02 (-0.02)	-0.79 (-0.61)	-0.33 (-0.25)	0.45 (0.34)	0.23 (0.18)	0.12 (0.09)	-0.12 (-0.09)	-0.14 (-0.11)
reg50340_sc8g5_c	0.23 (0.17)	-0.39 (-0.30)	0.07 (0.05)	0.46 (0.35)	0.30 (0.23)	0.18 (0.14)	-0.12 (-0.09)	-0.25 (-0.19)
reg50350_sc8g5_c	0.06 (0.05)	0.35 (0.27)	0.22 (0.17)	-0.14 (-0.11)	-0.14 (-0.11)	0.06 (0.05)	0.20 (0.16)	0.03 (0.02)
reg50360_sc8g5_c	0.16 (0.12)	0.02 (0.02)	0.04 (0.03)	0.02 (0.02)	-0.02 (-0.02)	-0.09 (-0.07)	-0.07 (-0.05)	-0.28 (-0.21)
reg50370_sc8g5_c	0.20 (0.15)	-0.03 (-0.02)	-0.01 (-0.01)	0.02 (0.02)	0.05 (0.04)	0.09 (0.07)	0.04 (0.03)	0.01 (0.01)
reg50410_sc8g5_c	-0.33 (-0.25)	0.15 (0.11)	0.06 (0.05)	-0.09 (-0.07)	0.04 (0.03)	0.06 (0.05)	0.01 (0.01)	0.19 (0.14)
reg5042s_sc8g5_c	-0.32 (-0.24)	0.57 (0.44)	0.15 (0.11)	-0.42 (-0.32)	-0.31 (-0.24)	-0.02 (-0.02)	0.30 (0.23)	0.17 (0.13)
reg50430_sc8g5_c	-0.13 (-0.10)	0.47 (0.36)	0.34 (0.26)	-0.13 (-0.10)	0.24 (0.19)	0.26 (0.20)	0.02 (0.02)	0.06 (0.05)
reg50440_sc8g5_c	-0.13 (-0.10)	0.25 (0.19)	0.36 (0.28)	0.11 (0.08)	-0.05 (-0.04)	-0.02 (-0.02)	0.03 (0.02)	-0.05 (-0.04)
reg50460_sc8g5_c	0.06 (0.05)	0.27 (0.21)	0.21 (0.16)	-0.06 (-0.05)	0.00 (0.00)	-0.03 (-0.02)	-0.03 (-0.02)	0.04 (0.03)
reg50510_sc8g5_c	-0.16 (-0.12)	-0.02 (-0.02)	-0.37 (-0.28)	-0.35 (-0.27)	0.20 (0.16)	-0.04 (-0.03)	-0.24 (-0.19)	0.29 (0.22)
reg5052s_sc8g5_c	-0.10 (-0.08)	0.18 (0.14)	0.18 (0.14)	0.00 (0.00)	-0.17 (-0.13)	-0.09 (-0.07)	0.08 (0.06)	0.17 (0.13)
reg50530_sc8g5_c	-0.46 (-0.35)	0.85 (0.65)	-0.09 (-0.07)	-0.94 (-0.72)	-0.08 (-0.06)	-0.16 (-0.12)	-0.09 (-0.07)	0.35 (0.26)
reg50540_sc8g5_c	-0.18 (-0.14)	-0.42 (-0.32)	-0.22 (-0.17)	0.21 (0.16)	0.03 (0.02)	-0.18 (-0.14)	-0.22 (-0.17)	0.24 (0.18)
reg5055s_sc8g5_c	-0.12 (-0.09)	0.51 (0.39)	0.27 (0.21)	-0.23 (-0.18)	-0.30 (-0.23)	-0.08 (-0.06)	0.22 (0.17)	0.31 (0.23)

Item	Sex	Migration				Books		Test position
	boys vs. girls	without vs. with	with vs. na	without vs. na	<100 vs. >100	<100 vs. na	>100 vs. na	first vs. second
reg50560_sc8g5_c	-0.14 (-0.11)	0.72 (0.55)	0.11 (0.08)	-0.62 (-0.47)	-0.49 (-0.38)	-0.08 (-0.06)	0.42 (0.33)	0.30 (0.23)
reg50570_sc8g5_c	-0.07 (-0.05)	0.68 (0.52)	0.18 (0.14)	-0.51 (-0.39)	-0.17 (-0.13)	-0.16 (-0.12)	0.00 (0.00)	-0.16 (-0.12)
Main effect (DIF model)	-0.28 (-0.21)	0.88 (0.68)	0.32 (0.24)	-0.57 (-0.43)	-0.70 (-0.54)	-0.19 (-0.15)	0.50 (0.39)	0.35 (0.27)
Main effect (Main effect model)	-0.14 (-0.10)	0.60 (0.46)	0.23 (0.18)	-0.37 (-0.28)	-0.42 (-0.33)	-0.10 (-0.08)	0.32 (0.25)	0.20 (0.15)

Note. Raw differences between item difficulties (in logits) with standardized differences (Cohen's *d*) in parentheses. na = Information not available.

Number of books: The number of books at home was used as a proxy for cultural capital. There were 1,474 participants with up to 100 books at home, 1,796 participants with more than 100 books at home, and 2,129 participants with no information on the number of books at home. There were considerable average differences between the three groups. Participants with up to 100 books at home performed on average -0.70 logits (Cohen's $d = -0.54$) lower in reading than participants with more than 100 books. The corresponding main effects for participants without information on this variable were -0.19 logits (Cohen's $d = -0.15$) and 0.39 logits (Cohen's $d = 0.39$), respectively. There was no considerable DIF comparing participants with many or few books at home (largest DIF = 0.49 logits). Comparing the group with no information about the number of books at home with the two groups with information, DIF was up to 0.42 logits and thus negligible. Consistent with these results, the overall test for DIF using the BIC (see Table 8) favored the model without DIF.

Test position: The reading competence test was administered in two different positions (see Section 2.3 for the design of the study). There were 2,764 respondents who received the reading test before the mathematics test and 2,635 respondents who received the reading test after the mathematics test. Participants were randomly assigned to one of the two groups. DIF for the position of the test can occur, for example, if there are differential fatigue effects for certain items. The results show a small average effect of item position. Respondents who received the reading test before the mathematics test performed on average 0.35 logits (Cohen's $d = 0.27$) better than respondents who received the reading test after the mathematics test. This main effect, however, does not indicate a threat to measurement invariance. Instead, it may be an indication of fatigue effects that are similar for all items. There was no noteworthy DIF due to the position of the test. The largest difference in item difficulty between the two groups was 0.59 logits (reg50230_sc8g5_c). Also, the overall test for DIF using the BIC (see Table 8) showed a better fit for the model that ignored potential DIF.

Table 8

Comparisons of Models with and without DIF

DIF variable	Model	<i>N</i>	Deviance	Number of parameters	AIC	BIC
Sex	Main effect	5,398	120,866	49	120,964	121,287
	DIF	5,398	120,549	80	<u>120,709</u>	<u>121,236</u>
Migration	Main effect	5,399	120,834	50	120,934	<u>121,263</u>
	DIF	5,399	120,595	112	<u>120,819</u>	121,557
Books	Main effect	5,399	120,739	50	120,839	<u>121,169</u>
	DIF	5,399	120,544	112	<u>120,768</u>	121,506
Test position	Main effect	5,399	120,866	49	120,964	<u>121,287</u>
	DIF	5,399	120,687	80	<u>120,847</u>	121,374

Note. The best-fitting model according to each information criterion is underlined.

5.3.4 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, Muraki, 1992) was fitted to the data to estimate the discrimination parameters. The estimated discrimination parameters differed moderately between items (see Table 5). The median discrimination parameter fell at 1.33 ($Min = 0.51$, $Max = 2.30$). Model fit indices suggested a better model fit of the GPCM (AIC = 119,519, BIC = 120,014, number of parameters = 75) compared to the PCM model (AIC = 121,011, BIC = 121,301, number of parameters = 44). However, an inspection of the respective item characteristic curves indicated an adequate fit of the PCM. Despite the empirical preference for the GPCM, the PCM is more in line with the theoretical concepts underlying the test construction (see Pohl & Carstensen, 2012, 2013). For this reason, the PCM was chosen as our scaling model in order to preserve the item weightings as intended in the theoretical framework.

5.3.5 Dimensionality

The dimensionality of the test was investigated by evaluating the correlations between the residuals of the PCM. The adjusted Q_3 statistics (see Table 5) were quite low ($Mdn = 0.04$) - the largest absolute residual correlation was 0.07 (reg5012s_sc8g5_c). Overall, these results indicate an essentially unidimensional test.

The dimensionality of the test was also examined by specifying two multi-dimensional models based on the three cognitive requirements or the five text types (see Section 2.1). Each item was assigned to one of three or five latent factors (between-item multidimensionality). The multidimensional models were estimated using Quasi Monte Carlo method with 5,000 nodes.

Table 9

Results of three-dimensional scaling for cognitive functions

Dimension	Dim 1	Dim 2	Dim 3
Dim 1: Finding information in the text	1.88		
Dim 2: Drawing text-related conclusions	.97	2.66	
Dim 3: Reflecting and assessing	.96	.94	1.46

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

The variances and correlations of the three dimensions reflecting the cognitive requirements are summarized in Table 9. All dimensions exhibited substantial variances. As expected, the correlations between the three cognitive requirements were rather high, ranging between .94 and .97. Because they were close to or above $r = .95$, these correlations suggest an essential unidimensionality of the test (see Carstensen, 2013). Although the correlations were large, the three-dimensional model did fit the data better (AIC = 120,836, BIC = 121,159, number of parameters = 49) than the unidimensional model (AIC = 121,011, BIC = 121,301, number of

parameters = 44). The results suggest that the three cognitive requirements do not measure different constructs but rather represent an essentially unidimensional construct.

The variances and correlations of the five dimensions reflecting the text types are summarized in Table 10. All dimensions exhibited substantial variances. The correlations between the five dimensions were between .80 and .90. The lowest correlations were found between Dimension 2 (instruction text) and Dimensions 1 (information text) and 4 (commenting text). However, all correlations were notably smaller than .95 (see Carstensen, 2013), indicating that the texts formed subdimensions. The five-dimensional model also exhibited a better fit to the data (AIC = 120,526, BIC = 120,908, number of parameters = 58) than the unidimensional model. Because the text types are perfectly confounded with the texts, that is, there is a single text for each text type, local item dependence (LID) occurs. The correlations reported in Table 10 are therefore due to multidimensionality based on text types as well as local item dependence. In the present study, it is not possible to disentangle these two sources. However, in pilot studies (Gehrer et al., 2013), a larger number of texts were presented to the respondents, thus, allowing the impact of text types to be investigated independently of LID. The correlations estimated in the pilot study ranged from .78 to .91. Although a different scaling model was used in this paper, the results can give an idea of the impact of the text types (unconfounded with LID) on the dimensionality of the test. As the correlations found in Gehrer et al. (2013) also differed from .95, it can be concluded that text types form subdimensions of reading competence. In addition, comparing these correlations with those found in the present study (see Table 10), which confounded the impact of text types and LID, allows for evaluating the impact of LID. The correlations found in the present study were similar (between .80 and .90) to those found in Gehrer et al. (between .78 and .91), indicating that there is little local item dependence. Based on theoretical considerations, Gehrer et al. (2013) argued for a unidimensional construct.

Table 10

Results of five-dimensional scaling for text types

Dimension	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Dim 1: Information text	2.39				
Dim 2: Instruction text	.80	1.66			
Dim 3: Advertising text	.88	.84	2.60		
Dim 4: Commenting or arguing text	.89	.81	.90	1.93	
Dim 5: Literary text	.89	.86	.85	.83	2.03

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

These results suggest that the three cognitive requirements and five text types measure a common construct. However, the test is not entirely unidimensional. Because the reading competence test was constructed to measure a single dimension, a unidimensional competence score was estimated.

6. Discussion

The analyses in the previous sections provided information on the quality of the reading competence test that was administered in Starting Cohort 8 of the NEPS. Different types of missing responses were examined, item fit statistics and item characteristic curves were evaluated, and item discriminations were investigated. Further quality inspections were conducted by examining differential item functioning and testing Rasch-homogeneity. Several criteria indicated a good fit of the items and measurement invariance across different subgroups. However, the number of missing values was high because many respondents did not reach the end of the test. This suggests that the test may have been too long for the given testing time. However, apart from the items that were not reached, other types of missing responses were quite small.

The test was better targeted at low- to medium-ability respondents and better covered the lower ability spectrum. Few items were available that could adequately discriminate between high-ability respondents. As a result, ability estimates were more accurate for low-performing respondents as compared to high-performing respondents. Overall, however, the test had a good reliability and discriminated well between respondents.

The unidimensionality of the test could be confirmed for the different cognitive requirements, while the different texts induced some multidimensionality. Nevertheless, Gehrler et al. (2013) argued that a balanced assessment of reading competence can only be achieved by a heterogeneity of texts and thus provided theoretical arguments for a unidimensional measure of reading competence.

In conclusion, the test had satisfactory psychometric properties that allowed the estimation of a unidimensional reading competence score for the participants.

7. Data in the Scientific Use Files

7.1 Naming Conventions

The SUF for the starting cohort contains 33 items, of which 27 were scored dichotomously (multiple-choice items) with 0 indicating an incorrect response and 1 indicating a correct response. The remaining items were scored as polytomous variables (complex multiple-choice and matching items) that are marked with an 's_c' at the end of the variable names. Further details on the naming conventions of the variables can be found in Fuß et al. (2025).

7.2 Linking of the Competence Tests

The same reading competence test was administered in Starting Cohort 3 (Pohl et al., 2012) and Starting Cohort 8. Therefore, $N = 5,193$ respondents from Starting Cohort 3 were used to evaluate whether the item difficulties were comparable between the two cohorts. As shown in Table 11, the median DIF effect (i.e., the difference in item difficulties) was 0.20 and thus not substantial. Only one item had a DIF effect greater than 0.60 (reg50530_sc8g5_c) and was more difficult in Starting Cohort 8. No pronounced DIF effects were observed for the remaining items. However, an overall test for DIF suggested a better fit for the model with DIF (AIC = 261,882, BIC = 261,882, number of parameters = 80) compared to a model without DIF (AIC = 262,395, BIC = 262,395, number of parameters = 49). Finally, the main effect between starting cohorts

was -0.14 logits (Cohen's $d = -0.11$), suggesting slightly better reading competencies in Starting Cohort 8 than in Starting Cohort 3.

Table 11

Differential Item Functioning for Starting Cohorts

Item	Estimate	Item	Estimate
reg50110_sc8g5_c	0.25 (0.20)	reg50340_sc8g5_c	0.07 (0.06)
reg5012s_sc8g5_c	0.00 (0.00)	reg50350_sc8g5_c	-0.18 (-0.14)
reg50130_sc8g5_c	0.15 (0.12)	reg50360_sc8g5_c	-0.03 (-0.02)
reg50140_sc8g5_c	0.26 (0.21)	reg50370_sc8g5_c	-0.10 (-0.08)
reg50150_sc8g5_c	0.23 (0.18)	reg50410_sc8g5_c	0.11 (0.09)
reg5016s_sc8g5_c	-0.02 (-0.02)	reg5042s_sc8g5_c	-0.03 (-0.02)
reg50170_sc8g5_c	0.10 (0.08)	reg50430_sc8g5_c	0.36 (0.29)
reg50210_sc8g5_c	-0.18 (-0.14)	reg50440_sc8g5_c	0.50 (0.40)
reg50220_sc8g5_c	0.02 (0.02)	reg50460_sc8g5_c	0.20 (0.16)
reg50230_sc8g5_c	0.16 (0.13)	reg50510_sc8g5_c	-0.13 (-0.10)
reg50240_sc8g5_c	0.24 (0.19)	reg5052s_sc8g5_c	-0.24 (-0.19)
reg50250_sc8g5_c	0.39 (0.31)	reg50530_sc8g5_c	-0.72 (-0.57)
reg5026s_sc8g5_c	0.31 (0.25)	reg50540_sc8g5_c	-0.34 (-0.27)
reg50310_sc8g5_c	-0.19 (-0.15)	reg5055s_sc8g5_c	-0.20 (-0.16)
reg50320_sc8g5_c	-0.26 (-0.21)	reg50560_sc8g5_c	-0.05 (-0.04)
reg50330_sc8g5_c	-0.33 (-0.26)	reg50570_sc8g5_c	0.37 (0.29)
Main effect (DIF model)	-0.14 (-0.11)	Main effect (Main effect model)	-0.11 (-0.09)

Note. Raw differences between item difficulties (in logits) with standardized differences (Cohen's d) in parentheses.

7.3 Reading Competence Scores

In the SUF, manifest reading competence scores are provided in the form of WLEs (reg5_sc1) including their respective standard errors (reg5_sc2). The person abilities were estimated using fixed item parameters from Starting Cohort 3 (Pohl et al., 2012). Because one item (reg50530_sc8g5_c) showed non-negligible DIF between cohorts (see Table 11), this item was not constrained but freely estimated. The WLE scores were also corrected for the position of the reading test within the test battery. As a result, the WLE scores in Starting Cohort 8 are linked to the scale of Starting Cohort 3, thus allowing comparisons between cohorts.

The R syntax for estimating the WLEs is provided in Appendix B. In the IRT scaling model, all polytomous variables were scored as 0.5 for each category. No WLEs were estimated for respondents who did not participate in the reading competence test or who did not provide enough valid responses. The value of the WLE and the corresponding standard error for these persons are denoted as not-determinable missing values. Alternatively, users interested in examining latent relationships can either include the measurement model in their analyses or estimate plausible values. Plausible values for the reading competence test can be estimated using the R package *NEPSscaling*⁴ (Scharl et al., 2020; Scharl & Zink, 2022).

⁴<https://www.neps-data.de/Data-Center/Overview-and-Assistance/Plausible-Values>

References

- Adam, R., Cloney, D., Wu, M., Berenzer, A., Osses, A., & Schwantner, V. (2023). *ACER ConQuest Manual*. Australian Council for Educational Research. <https://conquestmanual.acer.org>
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 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, Schöps, & S. Rönnebeck (Eds.), *Research outcomes of the PISA Research Conference 2009* (pp. 199–213). Springer. https://doi.org/10.1007/978-94-007-4458-5_12
- Fuß, D., Gnambs, T., Lockl, K., Attig, M., & Nusser, L. (2025). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6)*. Leibniz Institute for Educational Trajectories.
- 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)*. 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 of Educational Research Online*, 5(2), 50–79. <https://doi.org/10.25656/01:8424>
- Haberkorn, K., Pohl, S., & Carstensen, C. H. (2016). Incorporating different response formats of competence tests in an IRT model. *Psychological Test and Assessment Modeling*, 53(2), 223–252.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/BF02296272>
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177–196. <https://doi.org/10.1007/BF02294457>
- Muraki, E. (1992). A generalized partial credit model. Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176. <https://doi.org/10.1177/014662169201600>
- Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the data of the competence tests* (NEPS Working Paper 14). University of Bamberg, National Educational Panel Study. <https://doi.org/10.5157/NEPS:WP14:1.0>
- Pohl, S., & Carstensen, C. H. (2013). Scaling the competence tests in the National Educational Panel Study – many questions, some answers, and further challenges. *Journal of Educational Research Online*, 5(2), 189–216. <https://doi.org/10.25656/01:8430>
- 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 15). University of Bamberg, National Educational Panel Study. <https://doi.org/10.5157/NEPS:WP15:1.0>
- R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.1) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
- Robitzsch, A., Kiefert, T., & Wu, M. (2024). *TAM: Test analysis modules* (Version 4.2-21) [Computer software]. <https://doi.org/10.32614/CRAN.package.TAM>
- 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 71). Leibniz Institute for Educational Trajectories. <https://doi.org/10.5157/NEPS:SP71:1.0>
- Scharl, A., & Zink, E. (2022). NEPSscaling: Plausible value estimation for competence tests administered in the German National Educational Panel Study. *Large-Scale Assessments in Education*, 10. <https://doi.org/10.1186/s40536-022-00145-5>
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464. <https://doi.org/10.1214/aos/1176344136>
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(2), 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 (Eds.), *Zeitschrift für Erziehungswissenschaften*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67–86). VS Verlag für Sozialwissenschaften. <https://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(2), 125–145. <https://doi.org/10.1177/014662168400800201>
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213. <https://doi.org/10.1111/j.1745-3984.1993.tb00423.x>

Appendix A: Cognitive Requirements and Text Types

Pos.	Item	Cognitive requirements	Text types
1	reg50110_sc8g5_c	Drawing text-related conclusions	Information text
2	reg5012s_sc8g5_c	Finding information in the text	Information text
3	reg50130_sc8g5_c	Finding information in the text	Information text
4	reg50140_sc8g5_c	Drawing text-related conclusions	Information text
5	reg50150_sc8g5_c	Drawing text-related conclusions	Information text
6	reg5016s_sc8g5_c	Reflecting and assessing	Information text
7	reg50170_sc8g5_c	Reflecting and assessing	Information text
8	reg50210_sc8g5_c	Finding information in the text	Instruction text
9	reg50220_sc8g5_c	Reflecting and assessing	Instruction text
10	reg50230_sc8g5_c	Drawing text-related conclusions	Instruction text
11	reg50240_sc8g5_c	Drawing text-related conclusions	Instruction text
12	reg50250_sc8g5_c	Drawing text-related conclusions	Instruction text
13	reg5026s_sc8g5_c	Reflecting and assessing	Instruction text
14	reg50310_sc8g5_c	Finding information in the text	Advertising text
15	reg50320_sc8g5_c	Finding information in the text	Advertising text
16	reg50330_sc8g5_c	Drawing text-related conclusions	Advertising text
17	reg50340_sc8g5_c	Finding information in the text	Advertising text
18	reg50350_sc8g5_c	Drawing text-related conclusions	Advertising text

Pos.	Item	Cognitive requirements	Text types
19	reg50360_sc8g5_c	Finding information in the text	Advertising text
20	reg50370_sc8g5_c	Drawing text-related conclusions	Advertising text
21	reg50410_sc8g5_c	Finding information in the text	Commenting or arguing text
22	reg5042s_sc8g5_c	Drawing text-related conclusions	Commenting or arguing text
23	reg50430_sc8g5_c	Reflecting and assessing	Commenting or arguing text
24	reg50440_sc8g5_c	Reflecting and assessing	Commenting or arguing text
25	reg5045s_sc8g5_c		Commenting or arguing text
26	reg50460_sc8g5_c	Drawing text-related conclusions	Commenting or arguing text
27	reg50510_sc8g5_c	Drawing text-related conclusions	Literary text
28	reg5052s_sc8g5_c	Reflecting and assessing	Literary text
29	reg50530_sc8g5_c	Reflecting and assessing	Literary text
30	reg50540_sc8g5_c	Drawing text-related conclusions	Literary text
31	reg5055s_sc8g5_c	Reflecting and assessing	Literary text
32	reg50560_sc8g5_c	Reflecting and assessing	Literary text
33	reg50570_sc8g5_c	Reflecting and assessing	Literary text

Appendix B: R Syntax for Estimating WLEs

```
# load packages
library(rio) # to import SPSS files
library(doBy) # to recode variables
library(TAM) # for IRT analyses

# load competence data
dat <- rio::import("xTargetCompetencies.sav")

# items of the competence test
items <- c(
  "reg50110_sc8g5_c", "reg5012s_sc8g5_c", "reg50130_sc8g5_c",
  "reg50140_sc8g5_c", "reg50150_sc8g5_c", "reg5016s_sc8g5_c",
  "reg50170_sc8g5_c", "reg50210_sc8g5_c", "reg50220_sc8g5_c",
  "reg50230_sc8g5_c", "reg50240_sc8g5_c", "reg50250_sc8g5_c",
  "reg5026s_sc8g5_c", "reg50310_sc8g5_c", "reg50320_sc8g5_c",
  "reg50330_sc8g5_c", "reg50340_sc8g5_c", "reg50350_sc8g5_c",
  "reg50360_sc8g5_c", "reg50370_sc8g5_c", "reg50410_sc8g5_c",
  "reg5042s_sc8g5_c", "reg50430_sc8g5_c", "reg50440_sc8g5_c",
  "reg50460_sc8g5_c", "reg50510_sc8g5_c", "reg5052s_sc8g5_c",
  "reg50530_sc8g5_c", "reg50540_sc8g5_c", "reg5055s_sc8g5_c",
  "reg50560_sc8g5_c", "reg50570_sc8g5_c"
)

# polytomous items
poly <- c(
  "reg5012s_sc8g5_c", "reg5016s_sc8g5_c", "reg5026s_sc8g5_c",
  "reg5042s_sc8g5_c", "reg5052s_sc8g5_c", "reg5055s_sc8g5_c"
)

# collapse response categories
dat$reg5012s_sc8g5_c <-
  doBy::recodeVar(
    dat$reg5012s_sc8g5_c, c(0, 1, 2, 3, 4, 5, 6, 7), c(0, 0, 0, 0, 0, 0, 1, 2)
  )
dat$reg5016s_sc8g5_c <-
  doBy::recodeVar(
    dat$reg5016s_sc8g5_c, c(0, 1, 2, 3, 4, 5, 6), c(0, 1, 2, 3, 4, 5, 5)
  )
dat$reg5026s_sc8g5_c <-
  doBy::recodeVar(
    dat$reg5026s_sc8g5_c, c(0, 1, 2, 3, 4, 5, 6, 7), c(0, 0, 0, 0, 0, 0, 0, 1)
  )
dat$reg5052s_sc8g5_c <-
  doBy::recodeVar(
```

```
  dat$reg5052s_sc8g5_c, c(0, 1, 2, 3, 4), c(0, 0, 1, 2, 3)
)
dat$reg5f52s_c <-
  doBy::recodeVar(
    dat$reg5f52s_c, c(0, 1, 2, 3, 4), c(0, 0, 1, 2, 3)
  )
dat$reg5055s_sc8g5_c <-
  doBy::recodeVar(
    dat$reg5055s_sc8g5_c, c(0, 1, 2, 3, 4), c(0, 1, 2, 3, 3)
  )

# define Q-matrix for 0.5 scoring of PCM
Q <- matrix(1, nrow = length(items), ncol = 1)
Q[items %in% poly, 1] <- 0.5

# estimate partial credit model
mod <- TAM::tam.mml(resp = dat[, items], Q = Q, irtmodel = "PCM2", pid = dat$ID_t)
summary(mod)

# item fit
TAM::tam.fit(mod)

# WLE
TAM::tam.wle(mod)
```