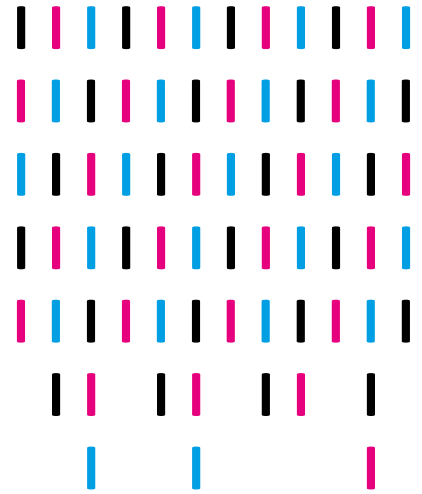




NEPS *SURVEY PAPERS*

Timo Gnambs

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

NEPS *Survey Paper* No. 118
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 in Special Schools

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 in special schools* (NEPS Survey Paper No. 118). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP118:1.0>

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 in Special Schools

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 in special schools. The reading test consisted of 25 items representing different cognitive requirements and text types, and using different response formats. The test was administered to 209 students (36% girls) from special schools. A partial credit model was used to scale the data. Item fit statistics, Rasch-homogeneity, test dimensionality, and local item independence were evaluated to ensure the quality of the test. The analyses showed that the items had satisfactory item fit and good reliability. In addition, 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 and some evidence of multidimensionality based on text types. Overall, however, the test had satisfactory 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 regular and special schools. Competence scores from both tests were therefore linked to allow comparison between the different school types. 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	5
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	8
5	Results	8
5.1	Missing Responses	8
5.1.1	Missing Responses per Person	8
5.1.2	Missing Responses per Item	10
5.2	Parameter Estimates	12
5.2.1	Item Parameters	12
5.2.2	Test Targeting and Reliability	14
5.3	Quality of the test	15
5.3.1	Distractor Analyses	15
5.3.2	Item Fit	15
5.3.3	Rasch-homogeneity	15
5.3.4	Dimensionality	17
6	Discussion	18
7	Data in the Scientific Use Files	19
7.1	Naming Conventions	19
7.2	Linking of the Competence Tests	19
7.3	Reading Competence Scores	21
	References	23
	Appendix A: Cognitive Requirements and Text Types	25
	Appendix B: R Syntax for Estimating WLEs	27

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 special 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 Artel (2013) and Gehrer et al. (2013). The administered test was similar to the reading competence test previously administered in regular schools of Grade 5 in Starting Cohort 3 (Pohl et al., 2012) and Starting Cohort 8 (Gnambs, 2025a). As compared to the test version administered in regular schools, the present test included only four texts. The first, third, and fourth texts were identical to the texts used in regular schools, while the second text was unique to the test administered in special schools. Moreover, some items referring to the texts common in both school types were exchanged with easier items to better target the expected ability level of students attending special schools. 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 four texts and four sets of items referring to these texts. Each of these texts represented one text type, namely, a) information, b) literary, c) instruction, and d) 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).

¹<https://www.neps-data.de>

2.2 Item Format

The reading competence test included three types of response formats, namely, simple multiple choice (MC) items and complex multiple choice (CMC) items. MC 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. The number of subtasks within CMC items varied from six to seven. Examples of the different response formats are given in Gehrler et al. (2012) and Pohl and Carstensen (2012).

The reading test included 25 items. Because preliminary analyses revealed an unsatisfactory fit for two items (reg50350_sc8g5_c, reg50530_sc8g5_c), these items were excluded from the scaling procedure. Therefore, the results presented in this paper refer to 23 items. As shown in Table 1, most of these were MC items, while CMC 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 four 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	21
Complex multiple-choice items	2
Total number of items	23

Table 2

Number of Items by Text Types

Text types	Number of items
Information text	6
Instruction text	5
Advertising text	6
Literary text	6
Total number of items	23

Table 3*Number of Items by Cognitive Requirements*

Cognitive requirement	Number of items
Finding information in the text	9
Drawing text-related conclusions	9
Reflecting and assessing	5
Total number of items	23

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 cognitive domains including reading competence as well as verbal and nonverbal reasoning abilities (Gnambs, 2025b). The reading competence test was always presented first before the other tests. 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 209 participants (36% girls) were administered the reading competence test. However, 12 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 197 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 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 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 items were aggregated from several subtasks, different types of missing responses or a mixture of valid and missing responses could be found. A CMC item was coded as missing if at least one subtask contained a missing response. If there was only

²<https://www.neps-data.de>

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 regular schools (see Gnambs, 2025a). Appendix B shows how response categories were collapsed for the different items.

Ability estimates for reading competence were estimated as weighted likelihood estimates (WLEs, 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). Because preliminary analyses identified low item discriminations for the items reg50180_c, reg50140_sc8g5_c, and reg50540_sc8g5_c in special schools, these items also received a score of 0.5 points for the estimation of the person parameters.

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.

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).

After aggregating the subtasks of CMC items 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 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 four 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

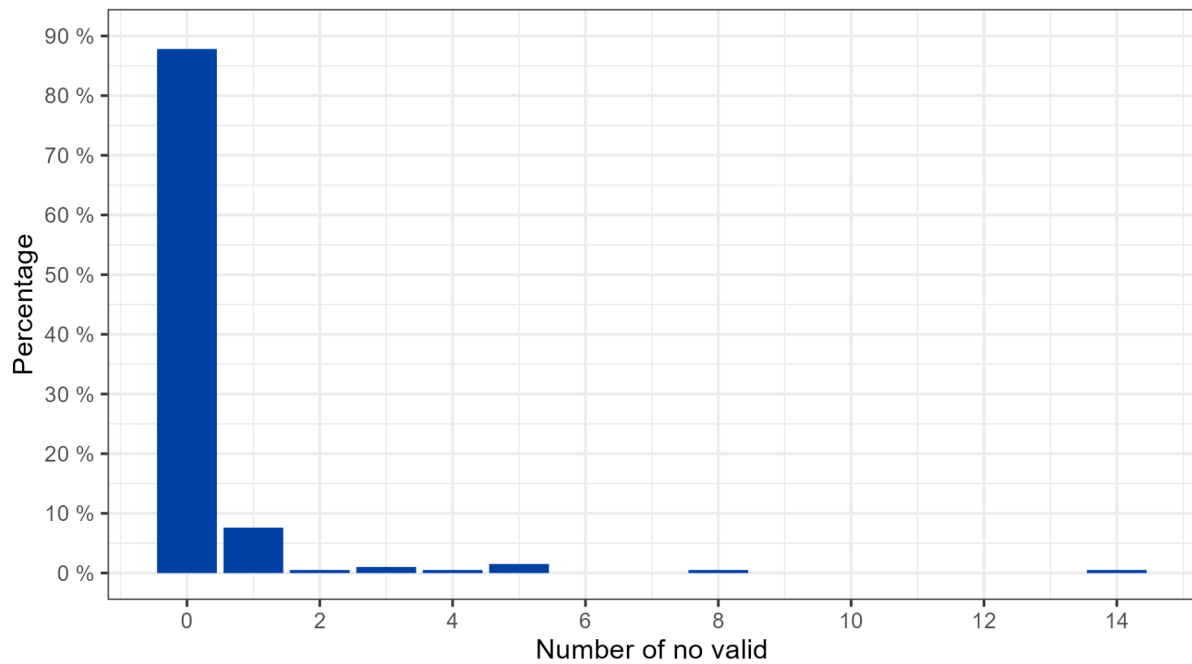
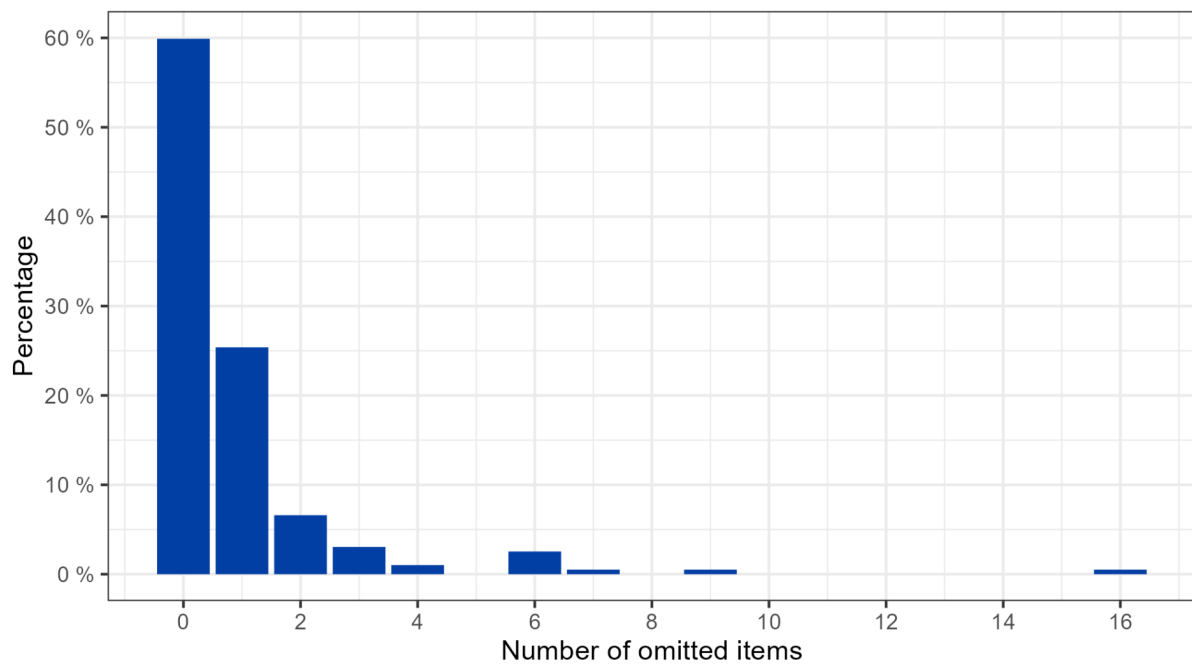
5.1 Missing Responses

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 88% of respondents having no invalid responses at all. Only about 5% 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 60% of respondents had no omitted items, while 25% exhibited a single omitted item. However, less than 4% of respondents had five or more omitted items.

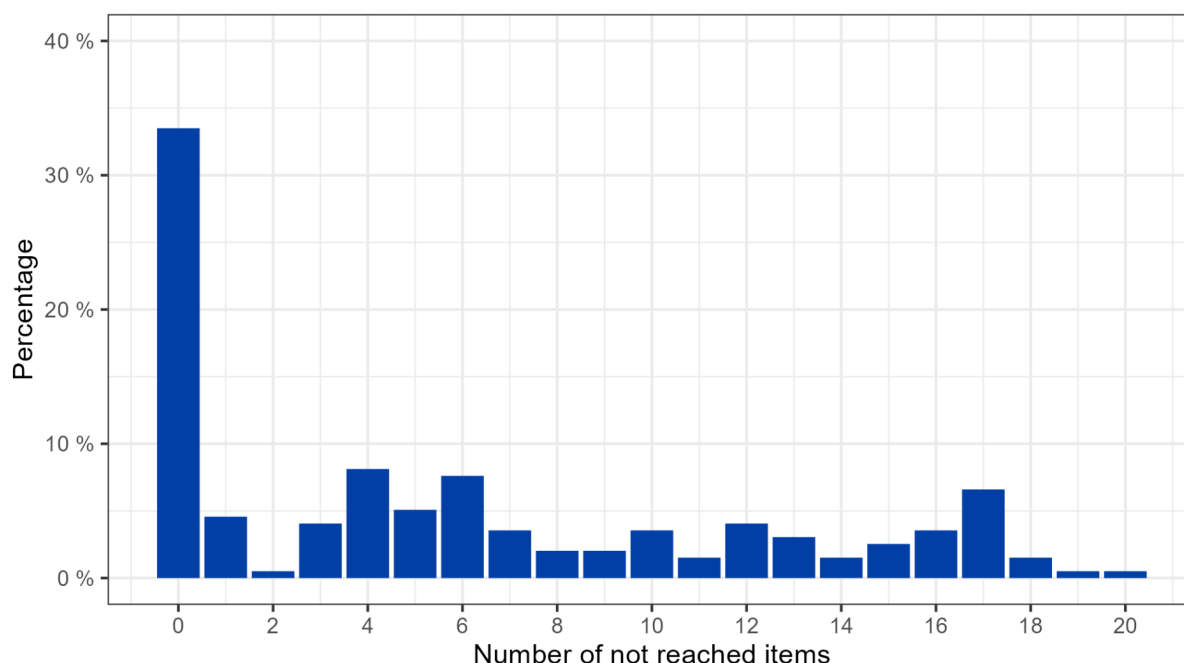
Another source of missing responses is items that respondents did not reach because they aborted the test, for example, due to time constraints or lack of motivation. These missing values refer to items after the last valid response. As shown in Figure 3, only about 34% of

Figure 1*Number of Invalid Responses***Figure 2***Number of Omitted Items*

respondents did not abort the test and were administered all items. However, more than 87% of the sample received at least 8 items and thus about a third of the total test.

Figure 3

Number of Not-Reached Items



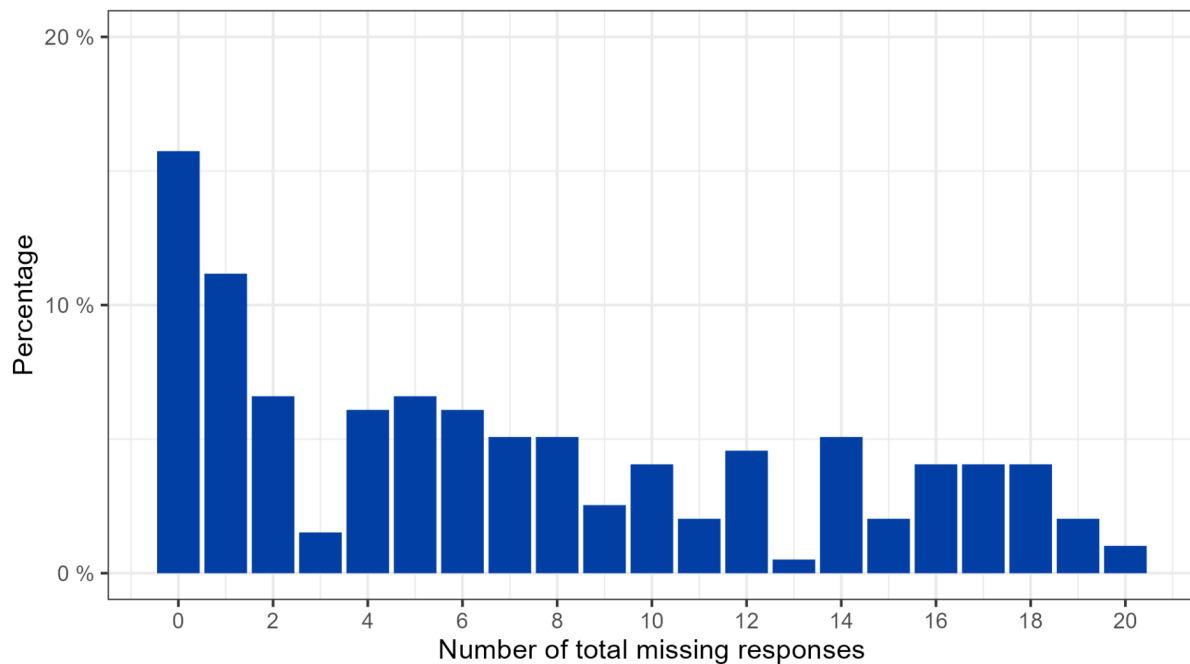
For the aggregated polytomous variables, responses were coded as not determinable missing if the subtasks of the CMC items contained different types of missing responses. Because not-determinable missing responses can only occur in CMC items, the maximum number of not-determinable missing responses was 2 (see Table 1). There was a negligible number of not-determinable missing responses. About 99% 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 4. Because most 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 16% of the respondents had no missing response at all. Almost 52% of respondents had more than five missing responses, while 27% had missing responses for 12 or more items (i.e., almost half of the administered items).

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 31% ($Mdn = 3\%$). The highest omission

Figure 4*Total Number of Missing Responses*

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 3% ($Mdn = 2\%$).

Table 4*Percentage of Missing Responses by Item*

Nr.	Item	Pos.	N	OM	NR	NV
1	reg50110_sc8g5_c	1	188	1.52	0.00	3.05
2	reg5012s_sc8g5_c	2	133	30.96	0.00	1.02
3	reg50130_sc8g5_c	3	184	4.06	0.00	2.54
4	reg50140_sc8g5_c	4	184	4.06	0.51	2.03
5	reg50150_sc8g5_c	5	185	3.05	1.02	2.03
6	reg50180_c	6	178	4.57	2.54	2.54
7	reg50910_c	7	173	0.51	9.14	2.54
8	reg50920_c	8	167	0.51	12.69	2.03
9	reg50930_c	9	156	2.54	15.23	3.05
10	reg50940_c	10	155	2.54	16.75	2.03

Nr.	Item	Pos.	N	OM	NR	NV
11	reg50950_c	11	148	3.55	19.80	1.52
12	reg50310_sc8g5_c	12	148	0.51	23.86	0.51
13	reg50320_sc8g5_c	13	142	1.52	25.38	1.02
14	reg50380_c	14	134	2.54	28.93	0.51
15	reg50340_sc8g5_c	15	125	3.55	30.96	2.03
16	reg50360_sc8g5_c	17	123	3.55	32.99	1.02
17	reg50370_sc8g5_c	18	115	3.05	36.55	2.03
18	reg50590_c	19	108	1.02	44.16	0.00
19	reg5f52s_c	20	83	7.61	49.24	0.51
20	reg50540_sc8g5_c	22	83	0.51	57.36	0.00
21	reg50580_c	23	74	0.51	61.42	0.51
22	reg50560_sc8g5_c	24	73	1.02	61.93	0.00
23	reg50570_sc8g5_c	25	66	0.00	66.50	0.00

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. The items on positions 16 and 21 were excluded due to a bad model fit.

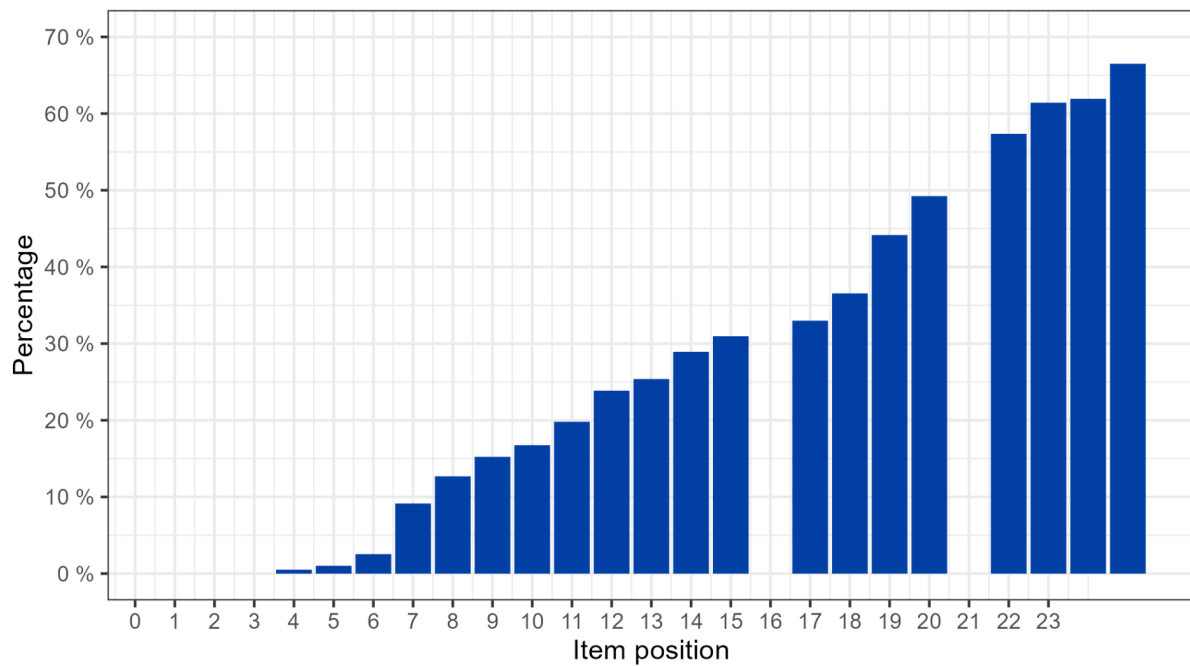
In contrast, there was a substantial number of items that respondents did not reach. The percentage of missing responses because participants did not reach an item was $Mdn = 24\%$. The percentage of respondents not reaching an item increased with the position of the item in the test (see Figure 5) up to 66%. 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 5% and 66% ($Mdn = 28\%$).

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 21.21% and 74.47% and had therefore an appropriate range for the sample (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 -1.38 (reg50110_sc8g5_c) to 1.38 (reg50150_sc8g5_c) with a median of 0.05, thus covering a fairly wide range including both easy and difficult items. Because of the small sample size,

Figure 5*Item Position not Reached*

Note. The items on positions 16 and 21 were excluded due to an unsatisfactory item fit.

the standard errors (*SE*) of the estimated item parameters were rather large with a median of 0.19 (*Min* = 0.11, *Max* = 0.34). The results reported in Table 5 and Table 6 are therefore rather imprecise.

Table 5*Item Parameters*

Nr.	Item	<i>N</i>	Percentage correct	Difficulty	<i>SE</i>	WMNSQ	<i>t</i>	<i>r_{it}</i>	<i>Q₃</i>	Discr.
1	reg50110_sc8g5_c	188	74.47	-1.38	0.19	0.95	-0.50	0.23	0.10	1.88
2	reg5012s_sc8g5_c	133		0.02	0.11	0.86	-1.74	0.34	0.11	1.22
3	reg50130_sc8g5_c	184	51.09	-0.07	0.17	1.00	0.09	0.32	0.07	1.21
4	reg50140_sc8g5_c	184	51.63	-0.08	0.15	1.01	0.22	0.15	0.06	0.56
5	reg50150_sc8g5_c	185	25.41	1.38	0.19	0.98	-0.13	0.25	0.09	1.29
6	reg50180_c	178	37.64	0.54	0.16	1.03	0.46	0.10	0.07	0.47
7	reg50910_c	173	34.10	0.82	0.18	1.08	0.91	0.22	0.08	0.90
8	reg50920_c	167	35.93	0.70	0.18	0.93	-0.81	0.43	0.07	1.50
9	reg50930_c	156	46.15	0.11	0.18	0.99	-0.07	0.44	0.08	1.30

Nr.	Item	<i>N</i>	Percentage correct	Difficulty	<i>SE</i>	WMNSQ	<i>t</i>	r_{it}	Q_3	Discr.
10	reg50940_c	155	48.39	0.01	0.18	1.08	1.03	0.35	0.09	0.82
11	reg50950_c	148	39.19	0.46	0.19	1.02	0.22	0.34	0.09	1.06
12	reg50310_sc8g5_c	148	52.70	-0.23	0.19	0.91	-1.22	0.46	0.08	1.81
13	reg50320_sc8g5_c	142	53.52	-0.26	0.19	0.92	-1.03	0.48	0.07	1.61
14	reg50380_c	134	51.49	-0.18	0.20	0.98	-0.17	0.39	0.09	1.37
15	reg50340_sc8g5_c	125	47.20	0.02	0.20	1.05	0.56	0.39	0.09	0.98
16	reg50360_sc8g5_c	123	56.91	-0.49	0.21	1.08	0.92	0.34	0.07	1.01
17	reg50370_sc8g5_c	115	36.52	0.50	0.22	0.94	-0.54	0.43	0.08	1.50
18	reg50590_c	108	44.44	0.05	0.22	1.00	-0.03	0.34	0.10	1.23
19	reg5f52s_c	83		-0.23	0.12	1.05	0.43	0.37	0.12	0.47
20	reg50540_sc8g5_c	83	33.73	0.58	0.24	1.01	0.12	0.18	0.09	0.59
21	reg50580_c	74	32.43	0.61	0.28	1.00	0.08	0.37	0.10	1.03
22	reg50560_sc8g5_c	73	42.47	0.05	0.27	1.09	0.83	0.32	0.10	0.80
23	reg50570_sc8g5_c	66	21.21	1.33	0.34	1.16	0.87	0.27	0.11	0.65

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
reg5012s_sc8g5_c	0.65 (0.23)	-0.65	
reg5f52s_c	-0.45 (0.22)	-0.15 (0.24)	0.60

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 6, 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 -2.27 (reg5f52s_c) to 1.43 (reg5f52s_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 1.54, which implies a good differentiation between respondents. The reliability of the test (EAP/PV reliability = 0.76, WLE reliability = 0.66) was satisfactory given the brevity of the test. The median of the category threshold distribution was about 0.05 log-its below the mean person ability distribution. Thus, the difficulties of the items, on average, matched the abilities of the respondents. However, there were few items covering the tails of the proficiency distribution. As a result, medium proficiencies were measured relatively accurately, while lower or higher ability estimates had larger standard errors of measurement.

5.3 Quality of the test

5.3.1 Distractor Analyses

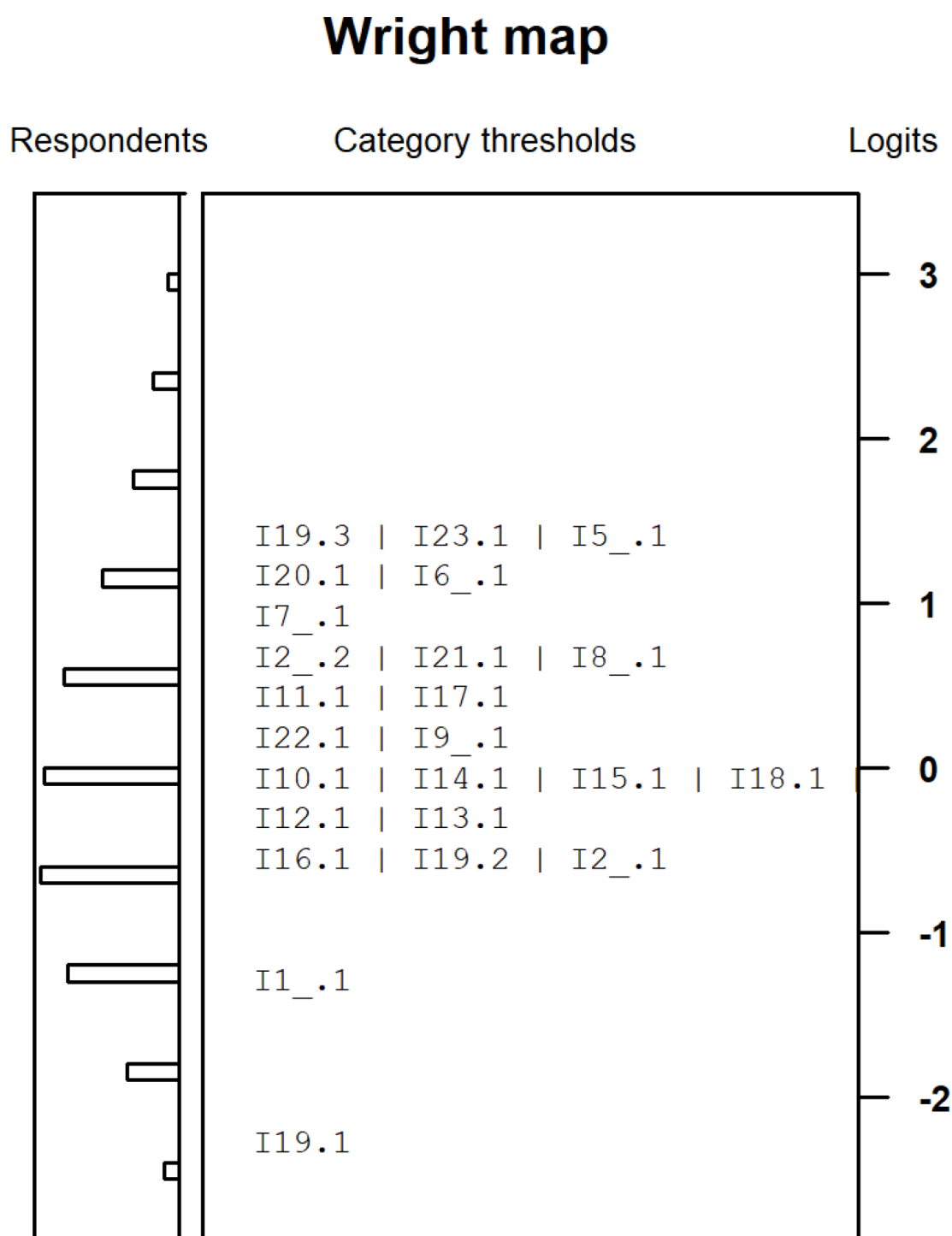
To investigate how well the distractors of the multiple-choice items performed, point-biserial correlations were calculated between each incorrect response (distractor) and the respondents' total correct scores for the remaining items. The median point-biserial correlation for the distractors fell at -.18 (*Min* = -.37, *Max* = .09). 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 was 1 item (reg50570_sc8g5_c) with a WMSNQ between 1.15 and 1.20, while no item had a WMSNQ greater than 1.20. However, visual inspections of the empirical ICC for this item did not reveal any pronounced deviation from the expected ICCs. For the remaining items, the WMNSQ values ranged from 0.86 (reg5012s_sc8g5_c) to 1.09 (reg50560_sc8g5_c), with a median of 1.00. In addition, the correlations of the correct responses with the total test scores varied between 0.10 (reg50180_c) and 0.48 (reg50320_sc8g5_c), with a median of 0.34.

5.3.3 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 1.06 (*Min* = 0.47, *Max* = 1.88). Model fit indices suggested a better model fit of the PCM (AIC = 3,998, BIC = 4,086, number of parameters = 27) compared to the GPCM (AIC = 4,017, BIC = 4,178, number of parameters = 49). An inspection of the respective item characteristic curves also indicated an adequate fit of the PCM. For this reason and in line with the theoretical concepts underlying the test construction (see Pohl & Carstensen, 2012, 2013),

Figure 6*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.

the PCM was chosen as the scaling model in order to preserve the item weightings as intended in the theoretical framework.

5.3.4 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 ($Mdn = 0.09$) - the largest absolute residual correlation was 0.12 (reg5f52s_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 four text types (see Section 2.1). Each item was assigned to one of three or four latent factors (between-item multidimensionality). The multidimensional models were estimated using Quasi Monte Carlo method with 5,000 nodes.

Table 7

Results of three-dimensional scaling for cognitive functions

Dimension	Dim 1	Dim 2	Dim 3
Dim 1: Finding information in the text	1.66		
Dim 2: Drawing text-related conclusions	.91	1.95	
Dim 3: Reflecting and assessing	.92	.90	1.15

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 7. All dimensions exhibited substantial variances. As expected, the correlations between the three cognitive requirements were rather high, ranging between .90 and .92. Because they were close to $r = .95$, these correlations suggest an essential unidimensionality of the test (see Carstensen, 2013). Also, the unidimensional model did fit the data better (AIC = 3,998, BIC = 4,086, number of parameters = 27) than the three-dimensional model (AIC = 4,001, BIC = 4,106, number of parameters = 32). 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 four dimensions reflecting the text types are summarized in Table 8. All dimensions exhibited substantial variances. The correlations between the four dimensions were between .78 and .90. The lowest correlations were found between Dimension 3 (advertising text) and Dimension 4 (literary text). However, all correlations were notably smaller than .95 (see Carstensen, 2013), indicating that the texts formed subdimensions. Nevertheless, the four-dimensional model exhibited a worse fit to the data (AIC = 4,005, BIC = 4,123, number of parameters = 36) 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 8 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 8), 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 .78 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 8

Results of four-dimensional scaling for text types

Dimension	Dim 1	Dim 2	Dim 3	Dim 4
Dim 1: Information text	2.02			
Dim 2: Instruction text	.89	2.25		
Dim 3: Advertising text	.90	.85	1.51	
Dim 4: Literary text	.87	.85	.78	1.32

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 four 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 special schools of 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 testing Rasch-homogeneity. Several criteria indicated a good fit of the items. 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 medium-ability respondents and better covered the middle ability spectrum. Few items were available that could adequately discriminate between low- or high-ability respondents. As a result, ability estimates were more accurate for

medium-performing respondents as compared to low- or high-performing respondents. Overall, however, the test had a satisfactory reliability and adequately discriminated 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 in special schools.

7. Data in the Scientific Use Files

7.1 Naming Conventions

The SUF for the starting cohort contains 25 items, of which 22 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 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

A similar reading competence test was administered in regular schools of Starting Cohort 3 (Pohl et al., 2012) and Starting Cohort 8 (Gnambs, 2025a). However, these tests included five (instead of four) texts (see Section 2) and shared only a subset of items with the test administered in special schools. Although several items were included in both samples, only the first five items of the first text and six items of the third text were presented in the same position. For the items of the fourth text, the item position varied or was unique to the test administered in special schools. Therefore, the focus is on the items contained in the first and third texts as potential anchors for linking the regular and special school samples.

To evaluate whether the reading test measured the same construct for all respondents, differential item functioning (DIF) was examined across school types in Starting Cohort 8 and across starting 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 (see Pohl & Carstensen, 2012). 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 differences between the estimated item difficulties (on the logit scale) in the different subgroups are summarized in Table 9. For example, the column “special vs. lower” reports the differences in item difficulties between special schools and lower secondary modern schools (“Hauptschule”); a positive value would indicate that the test was more difficult in special

schools, whereas a negative value would indicate that the item was less difficult in special schools than in lower secondary modern schools. 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 10, 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).

Table 9*Differential Item Functioning*

Item	School type			Starting cohort
	special vs. lower	special vs. intermediate	special vs. upper	3 vs. 8
reg50110_sc8g5_c	0.40 (0.27)	0.23 (0.16)	0.56 (0.42)	-0.02 (-0.01)
reg5012s_sc8g5_c	0.09 (0.06)	-0.28 (-0.20)	-0.53 (-0.40)	0.22 (0.15)
reg50130_sc8g5_c	0.46 (0.31)	0.79 (0.56)	0.27 (0.20)	-0.26 (-0.18)
reg50150_sc8g5_c	0.33 (0.22)	0.12 (0.08)	0.06 (0.05)	0.14 (0.10)
reg50310_sc8g5_c	-0.33 (-0.22)	-0.18 (-0.13)	0.13 (0.10)	-0.14 (-0.10)
reg50320_sc8g5_c	0.17 (0.11)	0.37 (0.26)	0.60 (0.45)	-0.67 (-0.47)
reg50340_sc8g5_c	-0.61 (-0.41)	-0.59 (-0.42)	-0.33 (-0.25)	0.39 (0.27)
reg50360_sc8g5_c	-0.32 (-0.22)	-0.06 (-0.04)	-0.28 (-0.21)	0.14 (0.10)
reg50370_sc8g5_c	0.19 (0.13)	0.42 (0.30)	0.49 (0.37)	-0.19 (-0.13)
Main effect (DIF model)	-0.95 (-0.64)	-2.00 (-1.41)	-3.02 (-2.28)	1.99 (1.38)
Main effect (Main effect model)	-1.01 (-0.68)	-1.91 (-1.36)	-2.87 (-2.18)	2.01 (1.40)

Note. Raw differences between item difficulties (in logits) with standardized differences (Cohen's d) in parentheses. special = special school ("Förderschule"); lower = lower secondary modern school ("Hauptschule"); intermediate = intermediate secondary modern school ("Realschule"); upper = upper secondary modern school ("Gymnasium").

School type: The sample included 197 respondents from special schools, 325 from lower secondary modern schools ("Hauptschule"), from intermediate secondary modern schools ("Realschule"), and from upper secondary modern schools ("Gymnasium") in Starting Cohort 8. There were substantial differences in reading competence between school types as indicated by the main effects of -0.95 logits (Cohen's $d = -0.64$), -2.00 logits (Cohen's $d = -1.41$), and -3.02 logits (Cohen's $d = -2.28$), respectively. These reflect lower reading competencies in special schools than in other school types. Three items (reg50340_sc8g5_c, reg50130_sc8g5_c, reg50320_sc8g5_c) showed DIF effects of 0.60 logits or more. The largest DIF was found between special schools and intermediate secondary modern schools (DIF = 0.79). However, an

overall test for DIF (see Table 10) suggested better fits for models without DIF compared to models that acknowledged DIF. These results suggest that for most items there was no pronounced DIF with respect to school type that could have biased the parameter estimates.

Starting Cohort: The sample included 5,193 respondents attending different school types from Starting Cohort 3 and 197 respondents attending special schools from Starting Cohort 8. There were substantial differences in reading competence between the starting cohorts as indicated by the main effects of 1.99 logits (Cohen's $d = 1.38$) that reflect lower reading competencies in special schools of Starting Cohort 8 than in Starting Cohort 3. One item (reg50320_sc8g5_c) showed a DIF effect of more than 0.60 logits. However, an overall test for DIF (see Table 10) suggested a better fit for a model that ignored DIF effects. These results suggest that for most items there was no pronounced DIF with respect to starting cohorts.

Table 10

Comparisons of Models with and without DIF

DIF variable	Model	<i>N</i>	Deviance	Number of parameters	AIC	BIC
School type: lower	Main effect	522	4,291	13	<u>4,317</u>	<u>4,372</u>
	DIF	522	4,275	21	4,317	4,406
School type: intermediate	Main effect	629	4,472	13	4,498	<u>4,556</u>
	DIF	629	4,448	21	<u>4,490</u>	4,584
School type: upper	Main effect	3,039	14,223	13	14,249	<u>14,328</u>
	DIF	3,039	14,188	21	<u>14,230</u>	14,357
Starting cohort	Main effect	5,390	35,971	13	35,997	<u>36,083</u>
	DIF	5,390	35,950	21	<u>35,992</u>	36,130

Note. The best-fitting model according to each information criterion is underlined. School type = special school vs. lower secondary modern school ("Hauptschule"), intermediate secondary modern school ("Realschule"), upper secondary modern school ("Gymnasium").

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 (reg50320_sc8g5_c) showed non-negligible DIF between starting cohorts (see Table 9), this item was not constrained but freely estimated. As a result, the WLE scores in special schools of Starting Cohort 8 are linked to the scale of Starting Cohort 3 and implicitly to that of Starting Cohort 8 (see Gnambs, 2025a), thus allowing comparisons between school types. Because a rotation design was used in regular schools of Starting Cohort 3 (Pohl et al., 2012) and Starting Cohort 8 (Gnambs, 2025a), the estimated person scores in special schools were adjusted by subtracting half the main effect of 0.20 estimated for the rotation design in regular schools (see Gnambs, 2025a).

The R syntax for estimating the WLEs is provided in Appendix B. In the IRT scaling model, all polytomous items and the MC items `reg50180_c`, `reg50140_sc8g5_c`, and `reg50540_sc8g5_c` were scored as 0.5 for each category (see Section 4.3). 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.

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>
- Gnambs, T. (2025a). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 8 in Grade 5* (NEPS Survey Paper 117). Leibniz Institute for Educational Trajectories. <https://doi.org/10.5157/NEPS:SP117:1.0>
- Gnambs, T. (2025b). *NEPS Technical Report for Verbal and Nonverbal Reasoning: Scaling results of Starting Cohort 8 in Grade 5 for Special Schools* (NEPS Survey Paper 119). Leibniz Institute for Educational Trajectories. <https://doi.org/10.5157/NEPS:SP119:1.0>
- 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>
- 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>
- 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	reg50180_c	Finding information in the text	Information text
7	reg50910_c	Drawing text-related conclusions	Instruction text
8	reg50920_c	Drawing text-related conclusions	Instruction text
9	reg50930_c	Drawing text-related conclusions	Instruction text
10	reg50940_c	Drawing text-related conclusions	Instruction text
11	reg50950_c	Reflecting and assessing	Instruction text
12	reg50310_sc8g5_c	Finding information in the text	Advertising text
13	reg50320_sc8g5_c	Finding information in the text	Advertising text
14	reg50380_c	Finding information in the text	Advertising text
15	reg50340_sc8g5_c	Finding information in the text	Advertising text
16	reg50350_sc8g5_c	Drawing text-related conclusions	Advertising text
17	reg50360_sc8g5_c	Finding information in the text	Advertising text

Pos.	Item	Cognitive requirements	Text types
18	reg50370_sc8g5_c	Drawing text-related conclusions	Advertising text
19	reg50590_c	Finding information in the text	Literary text
20	reg5f52s_c	Reflecting and assessing	Literary text
21	reg50530_sc8g5_c	Reflecting and assessing	Literary text
22	reg50540_sc8g5_c	Drawing text-related conclusions	Literary text
23	reg50580_c	Reflecting and assessing	Literary text
24	reg50560_sc8g5_c	Reflecting and assessing	Literary text
25	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("xTargetSpecialNeedsCompetencies.sav")

# items of the competence test
items <- c(
  "reg50110_sc8g5_c", "reg5012s_sc8g5_c", "reg50130_sc8g5_c",
  "reg50140_sc8g5_c", "reg50150_sc8g5_c", "reg50180_c",
  "reg50310_sc8g5_c", "reg50320_sc8g5_c", "reg50340_sc8g5_c",
  "reg50360_sc8g5_c", "reg50370_sc8g5_c", "reg50380_c",
  "reg50540_sc8g5_c", "reg50560_sc8g5_c", "reg50570_sc8g5_c",
  "reg50580_c", "reg50590_c", "reg50910_c",
  "reg50920_c", "reg50930_c", "reg50940_c",
  "reg50950_c", "reg5f52s_c"
)

# polytomous items
poly <- c(
  "reg5012s_sc8g5_c", "reg5f52s_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$reg5f52s_c <-
  doBy::recodeVar(
    dat$reg5f52s_c, c(0, 1, 2, 3, 4), c(0, 0, 1, 2, 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)