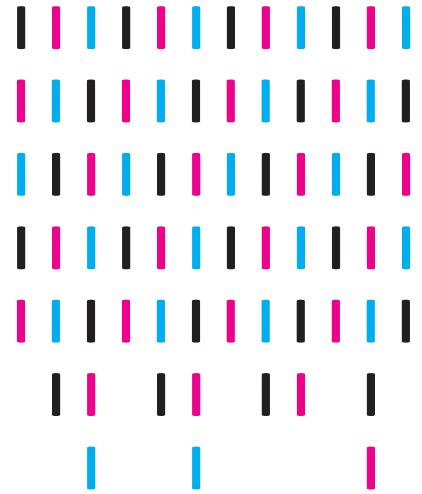




NEPS *SURVEY PAPERS*

Tanja Kutscher

NEPS TECHNICAL REPORT FOR
DIGITAL COMPETENCE:
SCALING RESULTS OF STARTING
COHORT 8 FOR GRADE 6 IN
SPECIAL SCHOOLS

NEPS *Survey Paper* No. 136
Bamberg, July 2026

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 Digital Competence: Scaling Results of Starting Cohort 8 for Grade 6 in Special Schools

Tanja Kutscher 

Leibniz Institute for Educational Trajectories

E-mail address of the lead author:

tanja.kutscher@lifbi.de

Bibliographic data:

Kutscher, T. (2026). *NEPS technical report for digital competence: scaling results of starting cohort 8 for grade 6 in special schools* (NEPS Survey Paper No. 136). Leibniz Institute for Educational Trajectories. <https://doi.org/10.5157/NEPS:SP136:1.0>

Acknowledgements. This report is an extension to *NEPS Survey Paper 137* (Kutscher, 2026) that presents the scaling results for digital competence in Grade 6 of Starting Cohort 8 for regular schools. 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 Digital Competence: Scaling Results of Starting Cohort 8 for Grade 6 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 digital competence administered in Grade 6 of Starting Cohort 8 in special schools. The digital competence test consisted of 19 items representing different facets and subfacets, using different response formats. The test was administered to 159 respondents, including 35% girls. A partial credit model was used to scale the data. Item fit statistics, Rasch-homogeneity, and test dimensionality were evaluated to ensure the quality of the test. Although the test showed a moderate reliability and a few items exhibited somewhat problematic fit, overall the test demonstrated acceptable psychometric properties that allowed for relatively reliable estimation of digital competence scores in special schools. Furthermore, analyses of differential item functioning showed partial measurement invariance for the test administered in regular and special schools. Digital 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, digital competence, scientific use file

Contents

1	Introduction	4
2	Testing Digital Competence	4
2.1	Conceptual Framework	4
2.2	Item Format	5
2.3	Study Design	6
3	Sample	6
4	Psychometric Analyses	7
4.1	Missing Responses	7
4.2	Scaling Model	7
4.3	Checking the Quality of the Test	8
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	9
5.2	Parameter Estimates	13
5.2.1	Item Parameters	13
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	17
7	Data in the Scientific Use Files	18
7.1	Naming Conventions	18
7.1.1	Linking of the Competence Tests	18
7.2	Digital Competence Scores	21
	References	22
	Appendix A: Facets	24
	Appendix B: R Syntax for Estimating WLEs	25

1. Introduction

The National Educational Panel Study (NEPS)¹ measures different competencies coherently across the lifespan. These include, among others, digital competence, reading competence, mathematical competence, and domain-general cognitive functioning. An overview of the competencies measured in the NEPS is given by Weinert et al. (2019) 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 digital competence that was administered in Grade 6 of special schools in Starting Cohort 8. The following sections first introduce the main concepts of the administered test and the test design. Then, the competence data are described. Next, the results of various psychometric analyses that were conducted to estimate the competence scores and to check the quality of the test are reported. 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, for the final data available in the SUF, no pronounced differences in the results presented are to be expected.

2. Testing Digital Competence

The framework and test development for the digital competence test are described in Lockl et al. (in prep.). In the following, several aspects of the digital competence test are described that are important for understanding the scaling results presented in this report.

2.1 Conceptual Framework

The NEPS conceptualizes digital competence as the responsible and reflective use of digital technologies, grounded in comprehensive knowledge of and critical reflection on the content, mechanisms, and effects of digital technologies (Artelt, 2025). Within this framework, digital competence is organized into four facets (see Figure 1):

1. ***Evaluating Information***

Involves assessing the credibility of sources and understanding of how algorithmic filtering shapes the access to and visibility of information. It also includes a basic understanding of how AI systems operate.

2. ***Detecting Intentions and Strategies***

Focuses on recognizing intentions that are designed to influence behavior, including commercial, social, or political intentions. It also encompasses knowledge on strategies and mechanisms of influence, including phenomena of manipulation of and influence on opinions and the restriction of diversity of opinion through filter bubbles and echo chambers.

¹<https://www.neps-data.de>

3. **Communicating and Interacting**

Focuses on effective and responsible collaboration and social engagement in digital contexts, including the awareness of potential social consequences, such as phenomena like hashtag debates, fake profiles, cyberbullying, as well as shitstorms and candystorms.

4. **Handling Data**

Addresses the responsible management of one's own data including privacy settings, questions of data collection, ways to reduce digital traces, and issues related to sharing personal information. It also refers to the consideration of other people's data and intellectual property, including ethical and legal aspects.

Figure 1

Facets of the Digital Competence Framework in the NEPS



2.2 Item Format

The digital competence test included two types of response formats, that differed in complexity, namely simple multiple choice (MC) items and complex multiple choice (CMC) items. MC items had four response options, one of which represented a correct solution and the other three were distractors (i.e., they were incorrect). In CMC items, several subtasks with two response options (each true / false) were presented. The number of subtasks within CMC items varied from four to seven. MC and CMC items typically contained textual response options or statements. However, some MC and CMC items contained visual elements in the form of images. For the sake of simplicity, the text-based as well as image-based MC and CMC items will be referred to as MC or CMC items in the remainder of this report.

The digital competence test included 19 items. Eleven of these were shared with the test version administered to Grade 6 of regular schools (Kutscher, 2026). As shown in Table 1, most of these were CMC items, while MC items were less common. Table 2 shows the distribution of the items

across the four facets (see Section 2.1). The facets were approximately equally distributed across the administered items. The allocation of each item to these facets is provided in Appendix A.

Table 1

Number of Items by Response Formats

Response format	Number of items
Complex multiple choice item: text-based	9
Complex multiple choice item: image-based	1
Simple multiple choice item: text-based	7
Simple multiple choice item: image-based	2
Total number of items	19

Table 2

Number of Items by Facets

Facet	Number of items
Evaluating information	4
Detecting intentions and strategies	6
Communicating and interacting	4
Handling data	5
Total number of items	19

2.3 Study Design

The study was conducted in small groups at the respective schools of the students. The tests were administered on computers by trained test administrators. The study assessed different competence domains, including digital competence, vocabulary, and general knowledge (Gnambs, 2026). The digital competence test was always administered 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 20 minutes. A comprehensive description of the study design is available on the NEPS website².

3. Sample

A total of 159 respondents (35% girls) were administered the digital competence test. All respondents received the digital competence test first before working on other tests. No respondents had less than three valid responses to the test items (see Pohl & Carstensen, 2012). Therefore, all cases were included in the psychometric analyses, namely 159 respondents.

²<https://www.neps-data.de>

4. Psychometric Analyses

4.1 Missing Responses

Competence data contain different types of missing responses. These are missing responses due to a) omitted items and b) items that respondents did not reach.

Omitted items occurred when respondents skipped some items. With regard to the last item, omitted means that the item was presented but some respondents did not select a response option. 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. For CMC items that had valid responses for at least half of their subitems, missing responses were imputed for the subitems in which they were missing using a Rasch (1960) model that included all MC items and CMC subitems. A CMC item was coded as missing if at least half of its subtasks had missing responses. If there was only one type of missing response, the item was coded according to the type of 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 item 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. When scoring polytomous items, missing values of subitems were imputed if at least 50% of the subitems of a polytomous item contained valid answers. The imputation was based on a Rasch model that included both multiple-choice items and subitems of polytomous items. Response categories of polytomous items with few responses were collapsed in order to avoid possible estimation problems. This was usually done for the lower categories of polytomous items. The response options were collapsed in the same way as for the scaling of digital competence in regular schools (see Kutscher, 2026). Appendix B shows how response categories were collapsed for the different items.

Digital 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 MC items were scored dichotomously as 0 for an incorrect response 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).

Ability estimates for digital competence can also be calculated as plausible values (see Mislevy, 1991) using the R package *NEPSscaling*³ (Scharl et al., 2020; Scharl & Zink, 2022; Zink et al., 2026). 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.

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

4.3 Checking the Quality of the Test

The digital competence 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 MC 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 item into a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the MC items in a Rasch (1960) model. The fit of the subtasks was evaluated using 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 judgement of item fit was based on all fit indicators.

The digital competence test should measure the same construct for all respondents. If some items favored certain subgroups, measurement invariance would be violated and a comparison of competence scores between these subgroups would be biased. In the present study, differential item functioning (DIF) was investigated for school types. DIF was examined using a multigroup item response model, in which the main effects of the subgroups and 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 to be very strong DIF, absolute differences between 0.6 and 1 to be considerable and worthy of further investigation, differences between 0.4 and 0.6 to be small but not severe, and differences less than 0.4 to be 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 digital competence 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 a multidimensional model using Quasi Monte Carlo integration (with 2,000 nodes), specifying different dimensions based on the test's construction rationale (see Section 2.1). A four-dimensional model was specified to represent the four facets of the test (see Appendix A). The correlations between the dimensions 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 using the *TAM* package (Robitzsch et al., 2025) in *R* (R Core Team, 2025).

5. Results

5.1 Missing Responses

5.1.1 Missing Responses per Person

Figure 2 shows the number of omitted items per person. A negligible number of items were omitted. Approximately 50% of respondents had no omitted items, while 32% exhibited a single omitted item. However, less than 4% of respondents had five or more omitted items. Regarding CMC items with imputed subitems, 6.3% of respondents had one item with imputed values, and 0.6% had two such 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, about 67% of respondents did not abort the test and were administered all items. However, more than 89% of the sample received at least 10 items and thus about a half of the total test.

The total number of missing responses, aggregated across omitted and not reached missing responses for each respondent, is shown in Figure 4. Because most respondents reached the end of the test and had few omitted items, the total number of missing values was relatively small. The median number of missing responses was 0; about 45% of the respondents had no missing responses at all. Only 26% of respondents had more than five missing responses, while 5% had missing responses for 11 or more items (i.e., almost half of the administered items).

Overall, there was a relatively small amount of omitted and not-reached missing responses. The total number of missing responses was also reasonable and did not indicate any substantial problems with the test.

5.1.2 Missing Responses per Item

Table 3 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

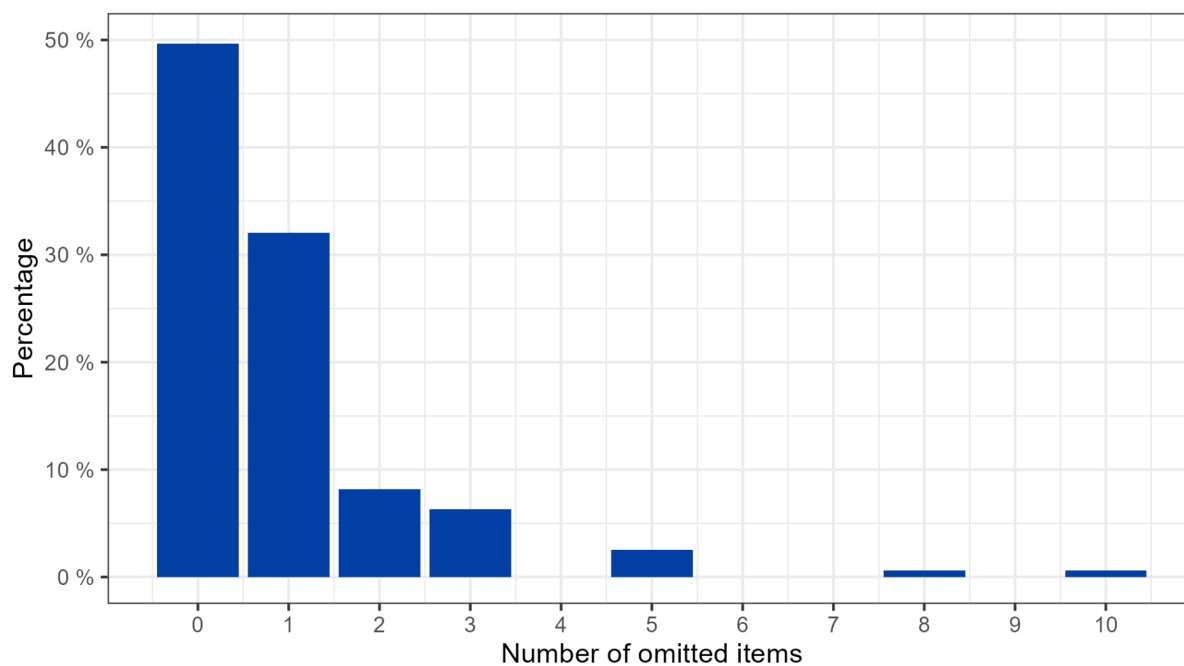
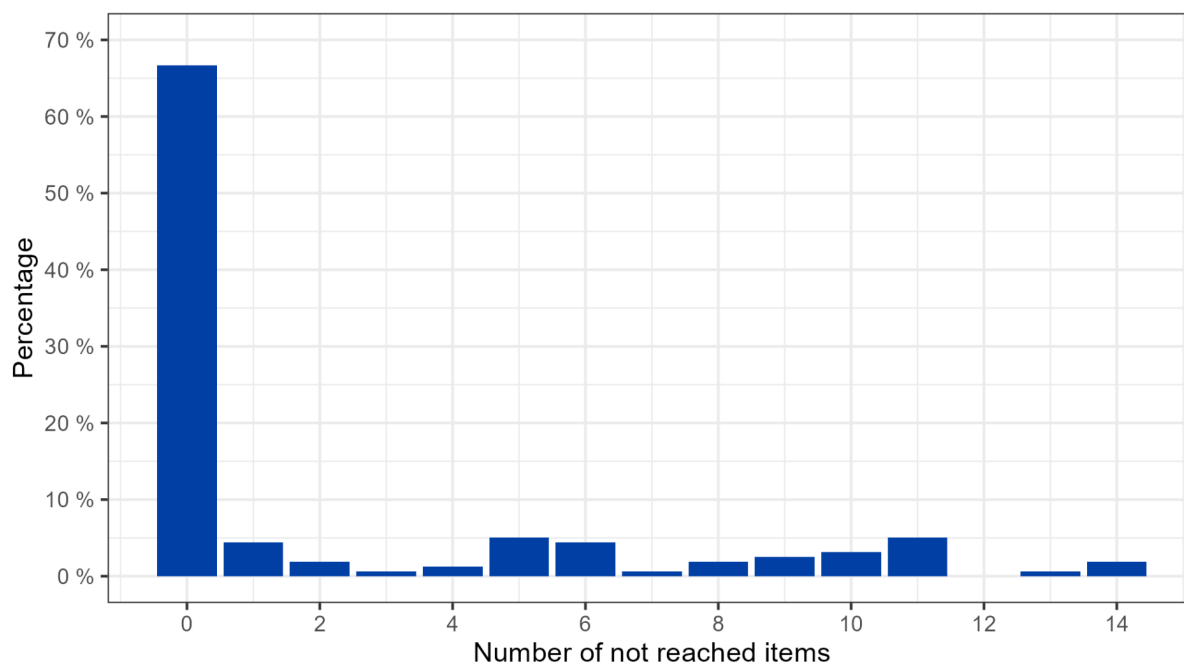
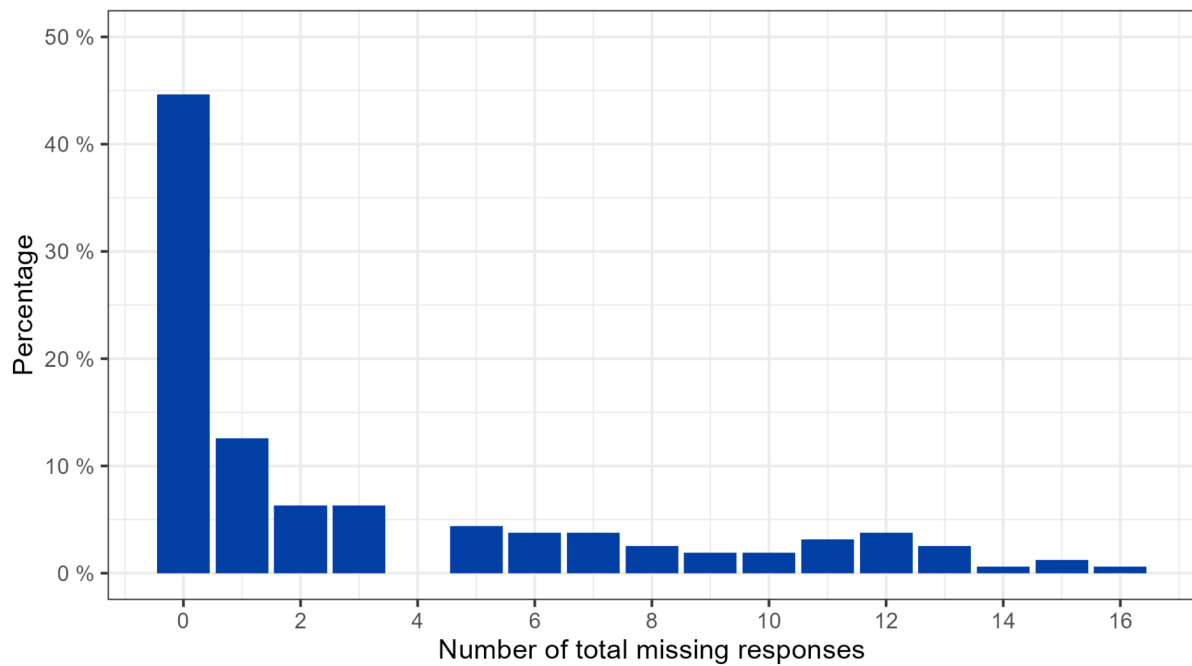
Figure 2*Number of Omitted Items***Figure 3***Number of Not-Reached Items*

Figure 4*Total Number of Missing Responses*

omitted responses per item varied between 1% and 17% ($Mdn = 4\%$). The highest omission rate was observed for the CMC item dcg63080_c. The number of imputed values per CMC item ranged from one to four ($mean = 1.7$).

Table 3*Percentage of Missing Responses by Item*

Nr.	Item	Pos.	N	OM	NR	ALL
1	dcg63050_c	1	156	1.89	0.00	1.89
2	dcg6109s_c	2	150	5.66	0.00	5.66
3	dcg6307s_c	3	154	3.14	0.00	3.14
4	dcg63080_c	4	132	16.98	0.00	16.98
5	dcg6208s_c	5	151	5.03	0.00	5.03
6	dcg6414s_c	6	150	3.77	1.89	5.66
7	dcg62040_c	7	154	0.63	2.52	3.14
8	dcg6301s_c	8	144	6.92	2.52	9.43
9	dcg61010_c	9	140	4.40	7.55	11.95
10	dcg6411s_c	10	136	3.77	10.69	14.46
11	dcg6211s_c	11	122	10.06	13.21	23.27

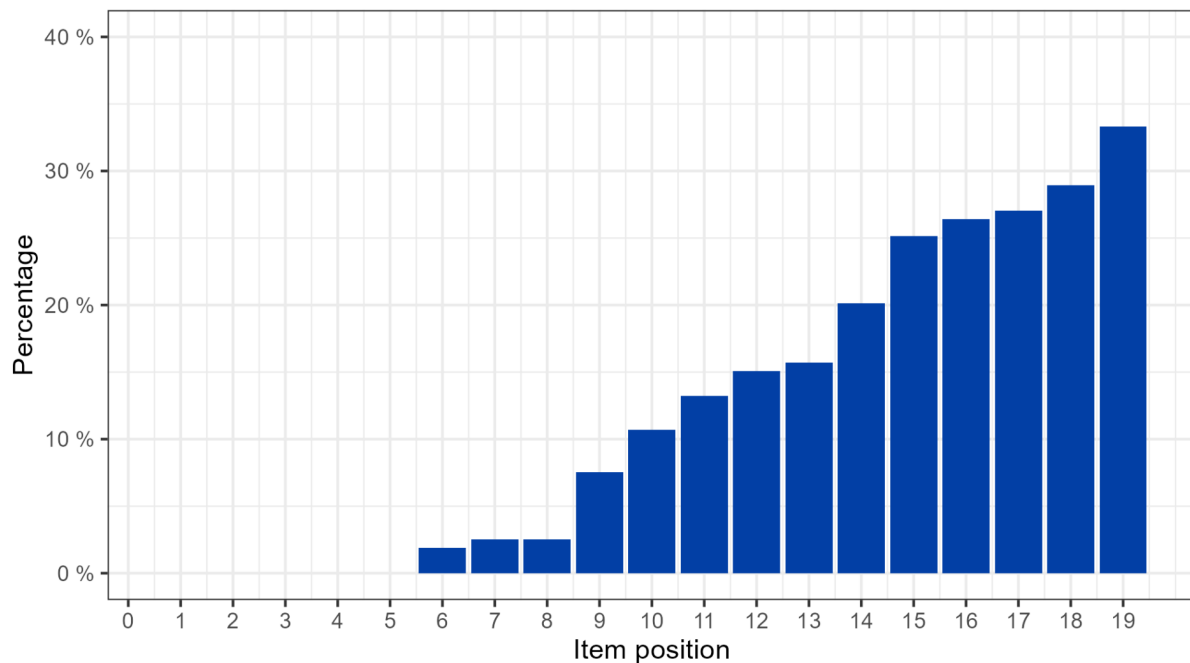
Nr.	Item	Pos.	N	OM	NR	ALL
12	dcg6214s_c	12	130	3.14	15.09	18.24
13	dcg64040_c	13	127	4.40	15.72	20.13
14	dcg6201s_c	14	121	3.77	20.13	23.90
15	dcg64150_c	15	117	1.26	25.16	26.41
16	dcg64080_c	16	105	7.55	26.41	33.96
17	dcg61040_c	17	113	1.89	27.04	28.93
18	dcg6213s_c	18	108	3.14	28.93	32.08
19	dcg61130_c	19	100	3.77	33.33	37.11

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. ALL = Percentage of all types of missing responses.

In addition, there were a moderate number of items that respondents did not reach. The percentage of missing responses because respondents did not reach an item was *Mdn* = 11%. The percentage of respondents not reaching an item increased with the position of the item in the test (see Figure 5) up to 33%. The total number of missing responses per item varied between 2% and 37% (*Mdn* = 17%).

Figure 5

Item Position not Reached



5.2 Parameter Estimates

An analysis of the subitems of the CMC items together with the MC items in a Rasch (1960) model revealed an inferior fit for some CMC subitems. These subitems had negative discrimination parameters and negative correlations with the total score. This means that the subitems did not work as expected and high ability respondents were more likely to answer them incorrectly than low ability respondents. It could be that these items were ambiguous for the present sample. For two CMC items (dcg6307s_c, dcg6411s_c), the respective subitems were excluded from further analyses. Other concerned CMC items were not modified to keep the test version for special schools consistent with the test version administered in regular schools (see Kutscher, 2026).

5.2.1 Item Parameters

The percentage of correct responses out of all valid responses for each MC item administered varied between 16.67% and 59.74%, which was considered acceptable given the specific nature of the sample (see Table 4). 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 4, while the corresponding step parameters for polytomous variables are summarized in Table 5. Item difficulties and location parameters were estimated by constraining the mean of the ability distribution to zero. The estimated item difficulties ranged from -0.46 (dcg62040_c) to 1.91 (dcg63080_c) with a median of 0.06, thus covering a range of items that varied from easy to difficult. The standard errors (*SE*) of the estimated parameters were relatively large ($SE \leq 0.25$) due to the small sample size.

Table 4

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	dcg63050_c	156	57.05	-0.35	0.17	0.97	-0.47	0.32	0.07	1.08
2	dcg6109s_c	150		-0.08	0.12	1.05	0.59	0.11	0.09	0.18
3	dcg6307s_c	154		-0.02	0.10	1.01	0.18	0.18	0.07	0.34
4	dcg63080_c	132	16.67	1.91	0.25	0.94	-0.33	0.38	0.09	1.36
5	dcg6208s_c	151		-0.15	0.08	1.07	0.86	0.20	0.10	0.32
6	dcg6414s_c	150		-0.25	0.10	0.89	-1.60	0.37	0.10	1.09
7	dcg62040_c	154	59.74	-0.46	0.18	1.05	0.88	0.23	0.08	0.54
8	dcg6301s_c	144		-0.17	0.11	0.95	-0.71	0.31	0.08	0.67
9	dcg61010_c	140	17.86	1.76	0.23	1.05	0.38	0.16	0.10	0.57
10	dcg6411s_c	136		0.32	0.11	1.01	0.11	0.29	0.07	0.44

Nr.	Item	N	Percentage correct	Difficulty	SE	WMNSQ	t	r_{it}	Q_3	Discr.
11	dcg6211s_c	122		0.86	0.21	1.13	1.44	0.18	0.07	0.29
12	dcg6214s_c	130		-0.25	0.11	0.93	-0.84	0.35	0.10	0.80
13	dcg64040_c	127	35.43	0.66	0.20	1.03	0.39	0.21	0.07	0.62
14	dcg6201s_c	121		0.42	0.20	0.93	-0.94	0.36	0.07	1.09
15	dcg64150_c	117	47.01	0.06	0.20	0.96	-0.62	0.39	0.09	1.23
16	dcg64080_c	105	33.33	0.79	0.22	1.00	0.06	0.28	0.05	0.84
17	dcg61040_c	113	31.86	0.80	0.22	1.11	1.14	0.17	0.06	0.42
18	dcg6213s_c	108		-0.02	0.21	0.88	-1.76	0.50	0.10	1.87
19	dcg61130_c	100	24.00	1.19	0.25	1.05	0.44	0.18	0.07	0.63

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). Percent correct scores are not informative for polytomous item scores and are not reported here.

Table 5

Step Parameters (with Standard Errors) for Polytomous Items

Item	Step 1	Step 2	Step 3
dcg6109s_c	-0.65 (0.16)	0.65	
dcg6307s_c	-0.11 (0.17)	0.11	
dcg6208s_c	-0.03 (0.16)	-0.33 (0.18)	0.35
dcg6414s_c	0.54 (0.20)	-0.54	
dcg6301s_c	-0.07 (0.18)	0.07	
dcg6411s_c	-0.02 (0.19)	0.02	
dcg6214s_c	-0.03 (0.19)	0.03	

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 digital 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, both the item difficulties of the digital competence items and the respondents' abilities 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 exhibited a considerable range, extending from -1.86 (dcg6109s_c) to 1.91 (dcg63080_c). The mean of the ability distribution was constrained to be zero. The variance was estimated to be 0.76, which implies a somewhat restricted differentiation between respondents. The test exhibited moderate reliability (EAP/PV reliability = 0.66, WLE reliability = 0.59), given the small sample size. The median of the category threshold distribution was about 0.42 logits above the mean person ability distribution. Consequently, although the items covered a wide range of the ability distribution, on average the items were somewhat challenging for the respondents. However, there were few items that covered the upper and lower extremes of the ability distribution. As a result, medium-level abilities were estimated 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

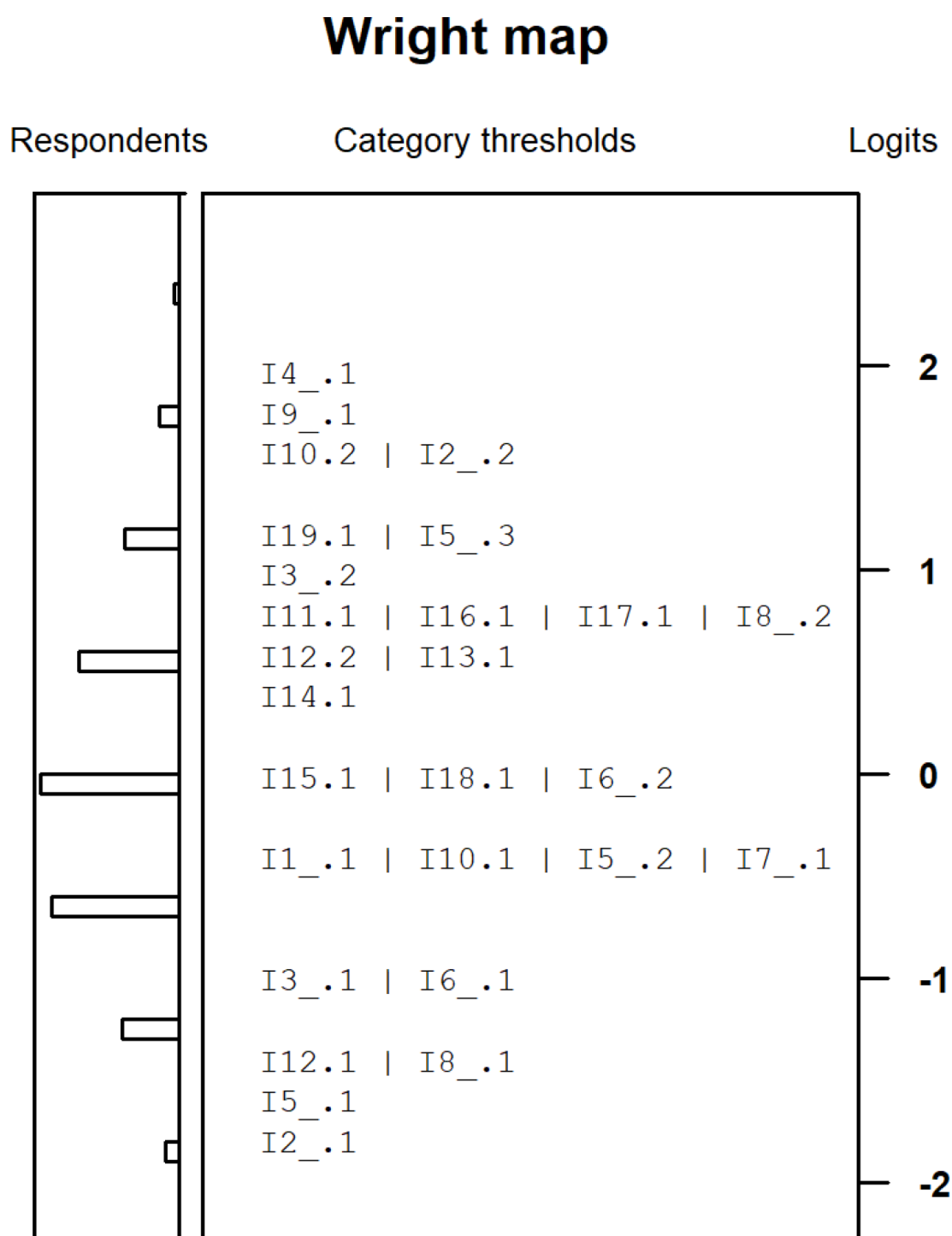
To investigate how well the distractors of the MC items performed, point-biserial correlations were calculated between each incorrect response (distractor) and the respondents' total correct scores for the remaining items. The median point-biserial correlation for the distractors was -.08 (*Min* = -.33, *Max* = .29). Only one distractor (from item dcg63080_c) out of 30 distractors across all MC items showed a correlation greater than .05. Overall, these results suggest that most 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 4). There were no items with a WMSNQ greater than 1.15. The WMNSQ values ranged from 0.88 (dcg6213s_c) to 1.13 (dcg6211s_c), with a median of 1.01. In addition, the correlations of the correct responses with the total rest scores varied between 0.11 (dcg6109s_c) and 0.50 (dcg6213s_c), with a median of 0.28.

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 4). The median discrimination parameter fell at 0.63 (*Min* = 0.18, *Max* = 1.87). The model fit indices provided inconsistent findings. BIC indicated a better model fit of the PCM (BIC = 4,111, number of parameters = 28) compared to the GPCM (BIC = 4,159, number of parameters = 46). In turn, the AIC indicated an essentially comparable fit of the GPCM (AIC = 4,017) and the PCM (AIC = 4,025). However, an inspection of the respective item characteristic curves indicated an adequate fit of the PCM. For this reason,

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 4 and the second part indicating the threshold.

the PCM was chosen as our 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 4) were quite low ($Mdn = 0.08$) - the largest absolute residual correlation was 0.10 (dcg6208s_c). Overall, these results indicate an essentially unidimensional test.

The dimensionality of the test was also examined by specifying a multi-dimensional model based on the four content facets (see Section 2.1). Each item was assigned to one of the four latent factors (between-item multidimensionality). The multidimensional model was estimated using Quasi Monte Carlo method with 5,000 nodes.

Table 6

Results of four-dimensional scaling for facets of digital competence

Dimension	Dim 1	Dim 2	Dim 3	Dim 4
Dim 1: Evaluating information	0.34			
Dim 2: Detecting intentions and strategies	.89	0.76		
Dim 3: Communicating and interacting	.90	.95	1.22	
Dim 4: Handling data	.90	.94	.95	1.08

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

The variances and correlations of the four dimensions reflecting the facets of digital competence are summarized in Table 6. All dimensions exhibited moderate variances, except for dimension 1 (evaluating information), where the variance was relatively small. As expected, the correlations between the four facets were rather high, ranging between .89 and .95. Because only the correlations of the dimension 1 with other three dimensions were smaller than $r = .95$, this facet, namely evaluating information, seems to capture some unique variance in addition to its shared variance (see Carstensen, 2013). Accordingly, the unidimensional model did fit the data better ($AIC = 4,025$, $BIC = 4,111$, number of parameters = 28) than the four-dimensional model ($AIC = 4,035$, $BIC = 4,149$, number of parameters = 37).

These results suggest that the digital competence test is essentially unidimensional, aligning with its construction to assess a single dimension. Consequently, a unidimensional competence score was estimated.

6. Discussion

The analyses in the previous sections provided information on the quality of the digital competence test that was administered in Grade 6 of NEPS Starting Cohort 8 in special schools. Different types of missing responses were examined, item fit statistics and item characteristic curves were evaluated, and item discriminations were examined. Although some subitems of

CMC items showed a problematic fit, most items had a satisfactory fit. Apart from the items that were not reached, the number of omitted items was quite small. Most distractors worked well, except for one problematic distractor.

The test was well targeted at respondents with low to high ability and covered a wide range of the ability spectrum. Few items were available that could adequately discriminate between low-ability and high-ability respondents. As a result, medium-level abilities were estimated relatively accurately, while lower or higher ability estimates had larger standard errors of measurement. Overall, however, the test had a moderate reliability, given a small sample size, and discriminated reasonably well between respondents.

The unidimensionality of the test could be confirmed for the different facets of the test. Nevertheless, Lockl et al. (in prep.) argued that a balanced assessment of digital competence can only be achieved by a heterogeneity of facets 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 digital competence score for the respondents in special school.

7. Data in the Scientific Use Files

7.1 Naming Conventions

The SUF for the starting cohort contains 19 items, of which 9 were scored dichotomously (single-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 a '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).

7.1.1 Linking of the Competence Tests

A similar digital competence test was administered in regular schools of Starting Cohort 8 in Grade 6 (Kutscher, 2026). However, this test included 23 items and shared only a subset of eleven items with the test administered in special schools. Although several items were included in both samples, only the first item was presented in the same position. For the other common items, the item position varied or was unique to the test administered in special schools. Nevertheless, we used all common items as potential anchors for linking the regular and special school samples.

To evaluate whether the digital competence test measured the same construct for all respondents, differential item functioning (DIF) was examined across school types in Starting Cohort 8 in Grade 6. The differences between the estimated item difficulties (on the logit scale) in the different subgroups are summarized in Table 7. For example, the column “special vs. lower/intermediate” reports the differences in item difficulties between special schools and lower secondary modern schools (“Hauptschule”) or intermediate secondary modern schools (“Realschule”); 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/intermediate 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 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).

The sample included 159 respondents from special schools ("Förderschule"), 843 from lower secondary modern schools ("Hauptschule") or intermediate secondary modern schools ("Realschule"), 1,057 from schools without segregating students into different educational tracks (here called integrated schools), and 2,438 from upper secondary modern schools ("Gymnasium") in Starting Cohort 8. There were substantial differences in digital competence between special schools and other school types as indicated by the main effects of -0.93 logits (Cohen's $d = -1.04$), -1.05 logits (Cohen's $d = -1.17$), and -1.49 logits (Cohen's $d = -2.16$), respectively. These reflect lower digital competencies in special schools than in other school types. Five items (d cg63050_c, d cg63080_c, d cg6211s_c, d cg64040_c, d cg64150_c, and d cg61040_c) showed DIF effects of $|0.60|$ logits or more. The largest DIF was found between special schools and upper secondary modern schools (DIF = 1.05). In addition, an overall test for DIF (see Table 8) suggested better fit for models that acknowledged DIF compared to models that ignored DIF. These results suggest that for only a half of the common items there was no pronounced DIF with respect to school type. Therefore, items with a DIF effect below 0.60 were selected for linking (d cg6109s_c, d cg6301s_c, d cg6214s_c, d cg6201s_c, and d cg6213s_c).

Table 7

Differential Item Functioning

Item	School		
	special vs. lower/intermedi- ate	special vs. integrated	special vs. upper
d cg63050_c	-0.44 (-0.49)	-0.40 (-0.45)	-0.71 (-1.03)
d cg6109s_c	-0.59 (-0.66)	-0.54 (-0.60)	-0.59 (-0.85)
d cg63080_c	0.91 (1.02)	0.81 (0.90)	0.58 (0.83)
d cg6301s_c	-0.24 (-0.27)	-0.27 (-0.30)	-0.35 (-0.50)
d cg6211s_c	-0.76 (-0.85)	-0.76 (-0.85)	-1.05 (-1.52)
d cg6214s_c	-0.03 (-0.03)	-0.04 (-0.04)	0.06 (0.09)
d cg64040_c	0.11 (0.12)	0.19 (0.21)	-0.91 (-1.31)
d cg6201s_c	-0.16 (-0.18)	-0.15 (-0.16)	0.23 (0.34)
d cg64150_c	-0.88 (-0.98)	-0.57 (-0.64)	
d cg61040_c	0.46 (0.52)	0.74 (0.82)	0.96 (1.40)
d cg6213s_c	-0.14 (-0.15)	-0.15 (-0.17)	-0.04 (-0.06)
Main effect (DIF model)	-0.93 (-1.04)	-1.05 (-1.17)	-1.49 (-2.16)

Item	School		
	special vs. lower/intermediate	special vs. integrated	special vs. upper
Main effect (Main effect model)	-0.67 (-0.75)	-0.79 (-0.88)	-1.30 (-1.89)

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

special = special schools ("Förderschule"); lower/intermediate = lower secondary modern school ("Hauptschule") or intermediate secondary modern school ("Realschule"); integrated = schools without segregating students into different educational tracks; upper = upper secondary modern schools ("Gymnasium"). Identical to Starting Cohort 8 in Grade 6 for regular schools, the item dcg64150_c was split by school type (see, Kutscher, 2026). Therefore, dcg64150_c represents the responses of respondents from other types of schools, excluding those from upper secondary schools ('Gymnasium').

Main effects refer to the main effects of respondents. A negative value indicates that the first-mentioned group has a lower ability than the second-mentioned group.

Table 8

Comparisons of Models with and without DIF

DIF variable	Model	<i>N</i>	Deviance	Number of parameters	AIC	BIC
special vs. lower/intermediate	Main effect	1,002	16,145	24	16,193	16,311
	DIF	1,002	16,062	34	<u>16,130</u>	<u>16,297</u>
special vs. integrated	Main effect	1,216	18,978	24	19,026	19,148
	DIF	1,216	18,902	34	<u>18,970</u>	<u>19,143</u>
special vs. upper	Main effect	2,597	36,370	23	36,416	36,551
	DIF	2,597	36,220	32	<u>36,284</u>	<u>36,472</u>

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

7.2 Digital Competence Scores

In the SUF, manifest digital competence scores are provided in the form of WLEs (d cg6_sc1) including their respective standard errors (d cg6_sc2). The person abilities were estimated using fixed item parameters from Starting Cohort 8 in Grade 6 for regular schools (Kutscher, 2026). These were the five items (d cg6109s_c, d cg6301s_c, d cg6214s_c, d cg6201s_c, and d cg6213s_c) that were identified as suitable for linking based on the DIF analyses. As a result, the WLE scores in special schools of Starting Cohort 8 in Grade 6 are linked to the scale of Starting Cohort 8 in Grade 6 for regular schools (see Kutscher, 2026), thus allowing comparisons between school types.

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 digital competence test or who did not provide enough valid responses. The value of the WLE and the corresponding standard error for these individuals 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>
- Artelt, C. (2025). Bildung in Zeiten von Digitalität und Künstlicher Intelligenz. In N. McElvany, C. Dignath, M. Becker, H. Gaspard, & A. Ohle-Peters (Eds.), *Welche Kompetenzen soll die Schule von heute für die Gesellschaft von morgen vermitteln?* (pp. 41–55). Waxmann. <https://doi.org/10.31244/9783818850111.04>
- Carstensen, C. H. (2013). Linking PISA competencies over three cycles – results from germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research on PISA* (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 8)*. Leibniz Institute for Educational Trajectories.
- Gnambs, T. (2026). *NEPS technical report for general knowledge: Scaling results of starting cohort 8 in grade 6 in special schools* (NEPS Survey Paper No. 135). Leibniz Institute for Educational Trajectories. <https://doi.org/10.5157/NEPS:SP135:1.0>
- 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.
- Kutscher, T. (2026). *NEPS technical report for digital competence: scaling results of starting cohort 8 for grade 6* (NEPS Survey Paper No. 137). Leibniz Institute for Educational Trajectories. <https://doi.org/10.5157/NEPS:SP137:1.0>
- Lockl, K., Tural, S., Schwaß, M., Kutscher, T., Hornberger, M., Artelt, C., & Wolter, I. (in prep.). *Measuring Digital Competence in the NEPS - Conceptual Framework and Findings From a Pilot Study*.
- 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 No. 14). University of Bamberg, National Educational Panel Study. <https://doi.org/10.5157/NEPS:WP14:1.0>
- R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.1) [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. (2025). *TAM: Test analysis modules* (Version 4.3-25) [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 No. 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, Article 28. <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., & Lockl, K. (2019). Development of competencies across the life course. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2nd ed., pp. 57–81). Springer. https://doi.org/10.1007/978-3-658-23162-0_4
- 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>
- Zink, E., Gnamb, T., Kutscher, T., & Sengewald, M.-A. (2026). *Plausible value estimates in the NEPS: An overview of the estimation routines and user examples* (NEPS Survey Paper No. 131). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP131:1.0>

Appendix A: Facets

Pos.	Item	Facet
1	dcg63050_c	Communicating and Interacting
2	dcg6109s_c	Evaluating Information
3	dcg6307s_c	Communicating and Interacting
4	dcg63080_c	Communicating and Interacting
5	dcg6208s_c	Detecting Intentions and Strategies
6	dcg6414s_c	Handling Data
7	dcg62040_c	Detecting Intentions and Strategies
8	dcg6301s_c	Communicating and Interacting
9	dcg61010_c	Evaluating Information
10	dcg6411s_c	Handling Data
11	dcg6211s_c	Detecting Intentions and Strategies
12	dcg6214s_c	Detecting Intentions and Strategies
13	dcg64040_c	Handling Data
14	dcg6201s_c	Detecting Intentions and Strategies
15	dcg64150_c	Handling Data
16	dcg64080_c	Handling Data
17	dcg61040_c	Evaluating Information
18	dcg6213s_c	Detecting Intentions and Strategies
19	dcg61130_c	Evaluating Information

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(
  "d cg61010_c", "d cg61040_c", "d cg6109s_c",
  "d cg61130_c", "d cg6201s_c", "d cg62040_c",
  "d cg6208s_c", "d cg6211s_c", "d cg6213s_c",
  "d cg6214s_c", "d cg6301s_c", "d cg63050_c",
  "d cg6307s_c", "d cg63080_c", "d cg64040_c",
  "d cg64080_c", "d cg6411s_c", "d cg6414s_c",
  "d cg64150_c"
)

# polytomous items
poly <- c(
  "d cg6109s_c", "d cg6201s_c", "d cg6208s_c",
  "d cg6211s_c", "d cg6213s_c", "d cg6214s_c",
  "d cg6301s_c", "d cg6307s_c", "d cg6411s_c",
  "d cg6414s_c"
)

# collapse response categories
dat$d cg6109s_c <-
  doBy::recodeVar(
    dat$d cg6109s_c, c(0, 1, 2, 3, 4), c(0, 0, 1, 2, 2)
  )
dat$d cg6214s_c <-
  doBy::recodeVar(
    dat$d cg6214s_c, c(0, 1, 2, 3, 4, 5), c(0, 0, 0, 1, 2, 2)
  )
dat$d cg6201s_c <-
  doBy::recodeVar(
    dat$d cg6201s_c, c(0, 1, 2, 3, 4), c(0, 0, 0, 1, 1)
  )
dat$d cg6213s_c <-
  doBy::recodeVar(
    dat$d cg6213s_c, c(0, 1, 2, 3, 4), c(0, 0, 0, 1, 1)
  )
```

```
dat$dcg6208s_c <-  
  doBy::recodeVar(  
    dat$dcg6208s_c, c(0, 1, 2, 3, 4, 5), c(0, 0, 1, 2, 3, 3)  
  )  
dat$dcg6211s_c <-  
  doBy::recodeVar(  
    dat$dcg6211s_c, c(0, 1, 2, 3, 4, 5), c(0, 0, 0, 0, 1, 1)  
  )  
dat$dcg6301s_c <-  
  doBy::recodeVar(  
    dat$dcg6301s_c, c(0, 1, 2, 3, 4, 5), c(0, 0, 0, 1, 2, 2)  
  )  
dat$dcg6307s_c <-  
  doBy::recodeVar(  
    dat$dcg6307s_c, c(0, 1, 2, 3), c(0, 0, 1, 2)  
  )  
dat$dcg6411s_c <-  
  doBy::recodeVar(  
    dat$dcg6411s_c, c(0, 1, 2, 3), c(0, 0, 1, 2)  
  )  
dat$dcg6414s_c <-  
  doBy::recodeVar(  
    dat$dcg6414s_c, c(0, 1, 2, 3, 4, 5, 6), c(0, 0, 0, 0, 0, 1, 2)  
  )  
  
# 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)
```