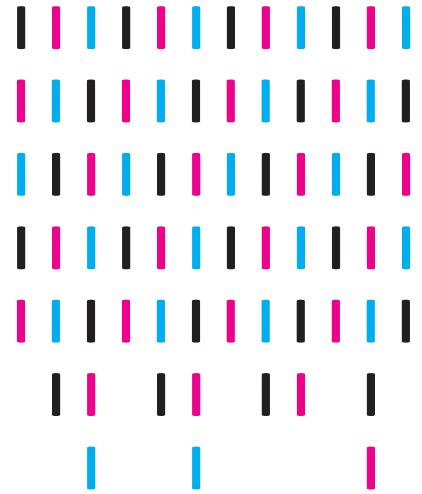




NEPS *SURVEY PAPERS*

Tanja Kutscher

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

NEPS *Survey Paper* No. 137
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

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* (NEPS Survey Paper No. 137). Leibniz Institute for Educational Trajectories. <https://doi.org/10.5157/NEPS:SP137:1.0>

Acknowledgements. This report is based on *NEPS Survey Paper No. 117* (Gnambs, 2025) that presents the scaling results for reading in Grade 5 of Starting Cohort 8. 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

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. The digital competence test consisted of 23 items representing different facets, using different response formats. The test was administered to 4,505 participants, including 51% girls. A partial credit model was used to scale the data. Item fit statistics, differential item functioning, Rasch-homogeneity, and test dimensionality were evaluated to ensure the quality of the test. The analyses showed good item fit and negligible differential item functioning (DIF) across different subgroups, except for small DIF effects by school types. In addition, the test exhibited good reliability and essential unidimensionality. Limitations of the test include a few items targeting higher digital ability, one item showing strong differential item functioning across school types and sex, and some evidence of multidimensionality based on facets. Overall, however, the test demonstrated good psychometric properties that allowed for reliable estimation of digital competence scores. 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	11
5.2	Parameter Estimates	12
5.2.1	Item Parameters	13
5.2.2	Test Targeting and Reliability	15
5.3	Quality of the test	16
5.3.1	Distractor Analyses	16
5.3.2	Item Fit	16
5.3.3	Differential Item Functioning	16
5.3.4	Rasch-homogeneity	23
5.3.5	Dimensionality	23
6	Discussion	24
7	Data in the Scientific Use Files	25
7.1	Naming Conventions	25
7.2	Digital Competence Scores	25
	References	26
	Appendix A: Facets	28
	Appendix B: R Syntax for Estimating WLEs	29

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 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 the 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 23 items. 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	12
Complex multiple choice item: image-based	2
Simple multiple choice item: text-based	5
Simple multiple choice item: image-based	4
Total number of items	23

Table 2*Number of Items by Facets*

Facet	Number of items
Evaluating information	5
Detecting intentions and strategies	7
Communicating and interacting	5
Handling data	6
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 administered on computers by trained test administrators. The study assessed different competence domains, including digital competence, information and communication technology (ICT) literacy, vocabulary, and general knowledge² (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 participants 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 website³.

3. Sample

A total of 4,505 participants (51% girls) received the digital competence test. All participants were given the digital competence test after working on the ICT literacy test and before a vocabulary test and a general knowledge test. However, 3 participants had less than three valid

²Participants whose native language is Russian or Turkish received a listening comprehension test in their native language at the end.

³<https://www.neps-data.de>

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 4,502 respondents.

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. 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*⁴ (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.

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 (e.g., they were easier for boys than for girls), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., boys and girls) would be biased. In the present study, differential item functioning (DIF) was investigated for the variables of sex, school types, the number of books at home (as a proxy for cultural status), and migrant background. 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

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

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

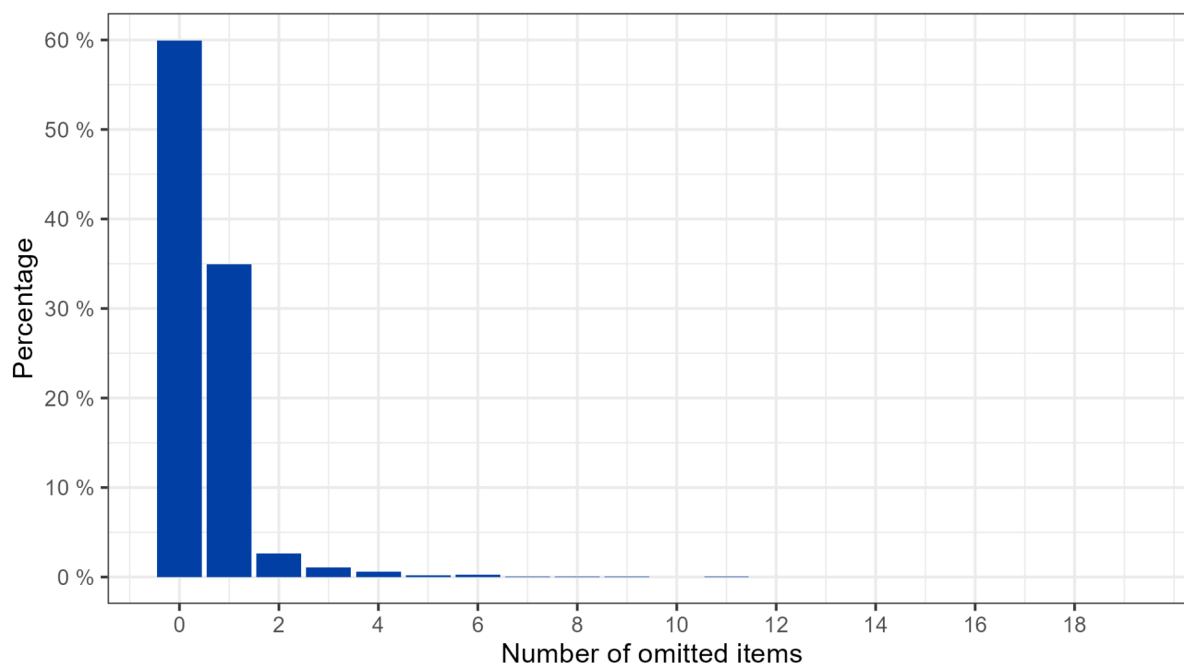
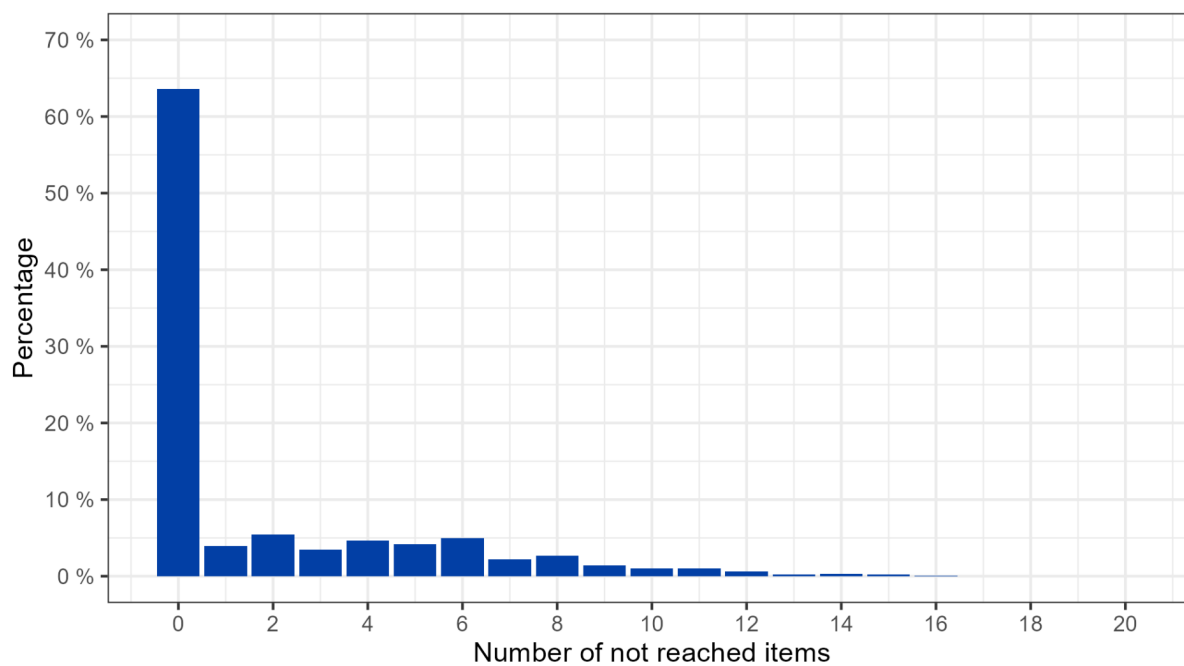
5.1 Missing Responses

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 60% of respondents had no omitted items, while 35% exhibited a single omitted item. However, less than 1% of respondents had five or more omitted items. Regarding CMC items with imputed subitems, 7.4% of respondents had one item with imputed values, and 0.4% 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 64% of respondents did not abort the test and were administered all items. However, more than 92% of the sample received at least 7 items and thus about a third of the total test.

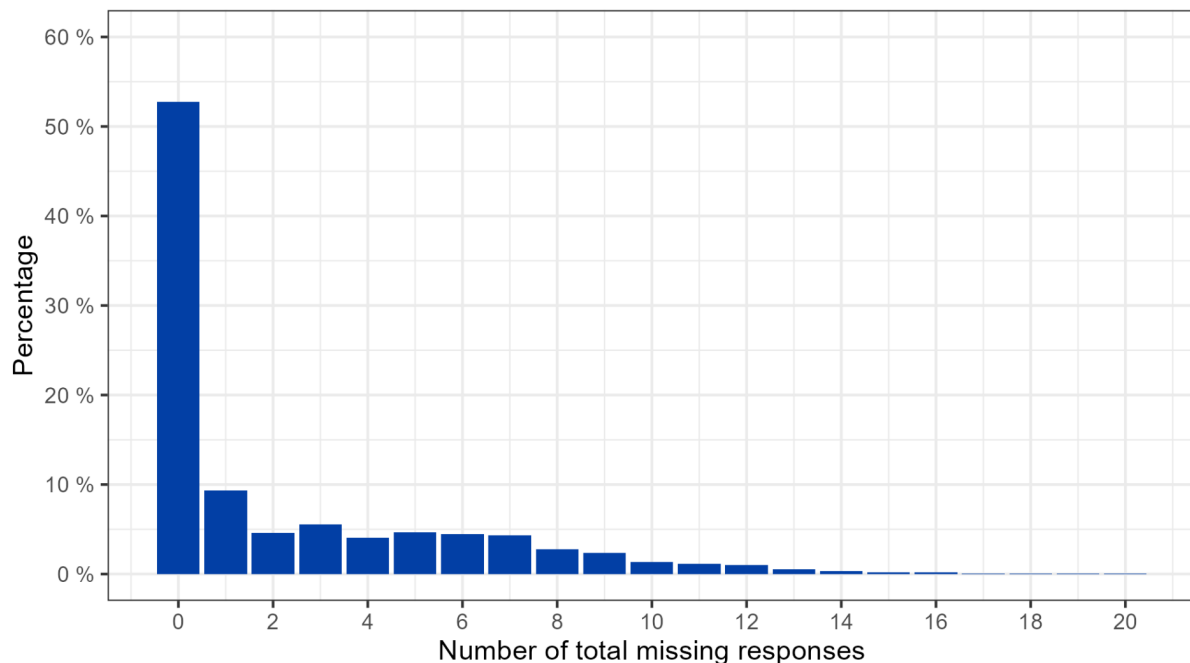
The total number of missing responses, aggregated across omitted and not reached missing responses for each respondent, is shown in Figure 4. Because most respondents reached the end of the test and had few omitted items, the total number of missing values was relatively small. The median number of missing responses was 0; about 53% of the respondents had no missing responses at all. Only 19% of respondents had more than five missing responses, while 2% had missing responses for 11 or more items (i.e., almost half of the administered items).

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

Figure 4

Total Number of Missing Responses



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 omitted responses per item varied between 0% and 7% (*Mdn* = 2%). The highest omission rate was observed for the CMC item dcg6306s_c. This may be because this was the last CMC item in the test; it was presented to the participants, but they probably did not have enough time to choose an answer. The number of imputed values per CMC item ranged from four to 106 (*mean* = 32).

Table 3

Percentage of Missing Responses by Item

Nr.	Item	Pos.	N	OM	NR	ALL
1	dcg63050_c	1	4,434	1.51	0.00	1.51
2	dcg6210s_c	2	4,480	0.49	0.00	0.49
3	dcg64170_c	3	4,473	0.64	0.00	0.64
4	dcg61040_c	4	4,485	0.35	0.02	0.38

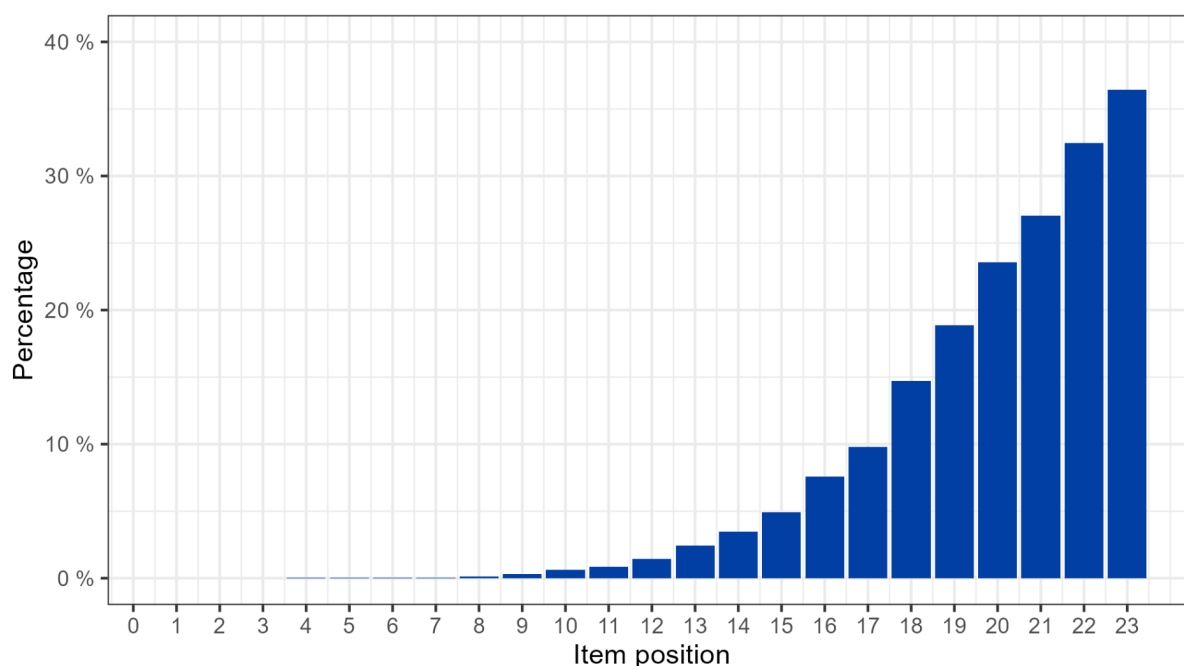
Nr.	Item	Pos.	N	OM	NR	ALL
5	dcg6209s_c	5	4,389	2.49	0.02	2.51
6	dcg62050_c	6	4,462	0.84	0.04	0.89
7	dcg6214s_c	7	4,481	0.42	0.04	0.47
8	dcg6102s_c	8	4,464	0.71	0.13	0.84
9	dcg6402s_c	9	4,463	0.53	0.33	0.87
10	dcg63080_c	10	4,421	1.18	0.62	1.80
11	dcg6114s_c	11	4,421	0.95	0.84	1.80
12	dcg6213s_c	12	4,384	1.18	1.44	2.62
13	dcg61080_c	13	4,311	1.80	2.44	4.24
14	dcg64040_c	14	4,276	1.55	3.46	5.02
15	dcg6301s_c	15	4,165	2.58	4.91	7.49
16	dcg64120_c	16	3,971	4.20	7.60	11.79
17	dcg6310s_c	17	3,845	4.82	9.77	14.59
18	dcg6201s_c	18	3,658	4.04	14.71	18.75
19	dcg6418s_c	19	3,458	4.31	18.88	23.19
20	dcg6211s_c	20	3,297	3.22	23.55	26.77
21	dcg6109s_c	21	3,066	4.87	27.03	31.90
22	dcg64150_c	22	2,898	3.15	32.47	35.63
23	dcg6306s_c	23	2,569	6.51	36.43	42.94

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* = 1%. The percentage of respondents not reaching an item increased with the position of the item in the test (see Figure 5) up to 36%. The total number of missing responses per item varied between 0% and 43% (*Mdn* = 3%).

5.2 Parameter Estimates

An analysis of subitems of the CMC items together with the MC items using a Rasch (1960) model revealed an inferior fit for two subitems. These subitems were excluded from further analyses and are not part of the generated CMC items (dcg6210s_c, dcg6310s_c). Furthermore, a multi-group analysis based on the PCM (Masters, 1982) revealed a strong DIF effect for the MC

Figure 5*Item Position not Reached*

item dcg64150_c in terms of sex and school type. Because this MC item is conceptually relevant to the test, it was split into two items based on school type (upper secondary schools versus other school types). Thus, in the remainder of this report the item dcg64150_c_g corresponds to the responses of participants attending upper secondary schools ('Gymnasium'), while item dcg64150_c corresponds to the responses of respondents attending other types of schools.

5.2.1 Item Parameters

The percentage of correct responses out of all valid responses for each MC item administered varied between 37.32% and 94.79%, thus covering a broad range (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 -2.82 (dcg64150_c_g) to 0.63 (dcg64120_c) with a median of -0.62, thus covering a range of items that varied from easy to medium. However, none of the items exhibited large difficulties. The standard errors (*SE*) of the estimated parameters were acceptable for most items ($SE \leq 0.11$).

Table 4*Item Parameters*

Nr.	Item	N	Percentage correct	Difficulty	SE	WMNSQ	t	r_{it}	Q_3	Discr.
1	dcg63050_c	4,434	71.45	-1.05	0.04	1.11	6.38	0.13	0.03	0.33
2	dcg6210s_c	4,480		-0.68	0.02	0.95	-2.90	0.37	0.03	0.60
3	dcg64170_c	4,473	45.20	0.23	0.03	1.01	0.57	0.28	0.02	0.81
4	dcg61040_c	4,485	75.18	-1.27	0.04	1.02	1.28	0.23	0.03	0.71
5	dcg6209s_c	4,389		-0.22	0.01	1.12	6.50	0.20	0.05	0.21
6	dcg62050_c	4,462	41.35	0.41	0.03	1.04	3.67	0.26	0.02	0.61
7	dcg6214s_c	4,481		-1.08	0.02	0.87	-4.46	0.40	0.06	1.05
8	dcg6102s_c	4,464		-0.15	0.01	1.05	2.45	0.20	0.04	0.32
9	dcg6402s_c	4,463		-0.57	0.02	1.04	1.71	0.24	0.03	0.34
10	dcg63080_c	4,421	52.88	-0.13	0.03	1.07	6.05	0.22	0.04	0.55
11	dcg6114s_c	4,421		-0.47	0.02	0.95	-3.05	0.36	0.03	0.57
12	dcg6213s_c	4,384		-0.85	0.02	0.91	-4.50	0.38	0.04	0.85
13	dcg61080_c	4,311	49.43	0.04	0.03	1.01	1.06	0.26	0.02	0.77
14	dcg64040_c	4,276	73.01	-1.12	0.04	0.98	-0.84	0.29	0.03	0.93
15	dcg6301s_c	4,165		-0.73	0.02	0.97	-1.32	0.31	0.03	0.49
16	dcg64120_c	3,971	37.32	0.63	0.04	1.06	4.48	0.21	0.02	0.54
17	dcg6310s_c	3,845		-0.66	0.02	1.00	0.22	0.25	0.02	0.41
18	dcg6201s_c	3,658		-0.66	0.02	0.89	-6.15	0.44	0.05	0.95
19	dcg6418s_c	3,458		-0.91	0.02	0.87	-5.20	0.45	0.07	0.99
20	dcg6211s_c	3,297		0.44	0.02	1.03	1.81	0.23	0.02	0.33
21	dcg6109s_c	3,066		-0.43	0.02	1.02	1.01	0.29	0.02	0.38
22	dcg64150_c	1,152	80.73	-1.98	0.08	0.91	-1.97	0.40	0.05	1.45
23	dcg64150_c_g	1,746	94.79	-2.82	0.11	0.95	-0.49	0.25	0.04	1.39
24	dcg6306s_c	2,569		-0.21	0.02	0.92	-3.70	0.46	0.04	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). The item dcg64150_c was split by school type. dcg64150_c_g represents the responses of respondents from upper secondary schools ('Gymnasium'), while dcg64150_c represents the responses of respondents from other types of schools.

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	Step 4
dcg6210s_c	-0.25 (0.03)	0.23 (0.04)	0.02	
dcg6209s_c	-0.59 (0.03)	-0.72 (0.03)	1.21 (0.04)	0.10
dcg6214s_c	0.41 (0.04)	0.10 (0.04)	-0.51	
dcg6102s_c	-0.14 (0.04)	-1.43 (0.03)	0.75 (0.04)	0.82
dcg6402s_c	-0.05 (0.03)	-0.59 (0.03)	-0.50 (0.03)	1.15
dcg6114s_c	-0.35 (0.03)	-0.02 (0.03)	0.36	
dcg6213s_c	0.37 (0.04)	-0.37		
dcg6301s_c	-0.25 (0.03)	0.08 (0.04)	0.17	
dcg6310s_c	-0.11 (0.04)	0.11		
dcg6201s_c	0.18 (0.04)	-0.18		
dcg6418s_c	0.54 (0.05)	-0.54		
dcg6211s_c	0.71 (0.05)	-0.71		
dcg6109s_c	-1.12 (0.04)	0.69 (0.04)	0.43	
dcg6306s_c	-0.06 (0.04)	-0.47 (0.04)	0.52	

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 ranged from -3.42 (dcg6109s_c) to 2.31 (dcg6102s_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 0.72, which implies somewhat restricted differentiation between respondents. The reliability of the test (EAP/PV reliability = 0.73, WLE reliability = 0.71) was good. The median of the category threshold distribution was about -1.16 logits below the mean person ability distribution. Thus, although the items covered a wide range of the ability distribution, on average the items were too easy for the respondents. As a result, ability estimates in the low to medium ability ranges were measured relatively accurately, while higher ability estimates had larger standard errors of measurement.

5.3 Quality of the test

5.3.1 Distractor Analyses

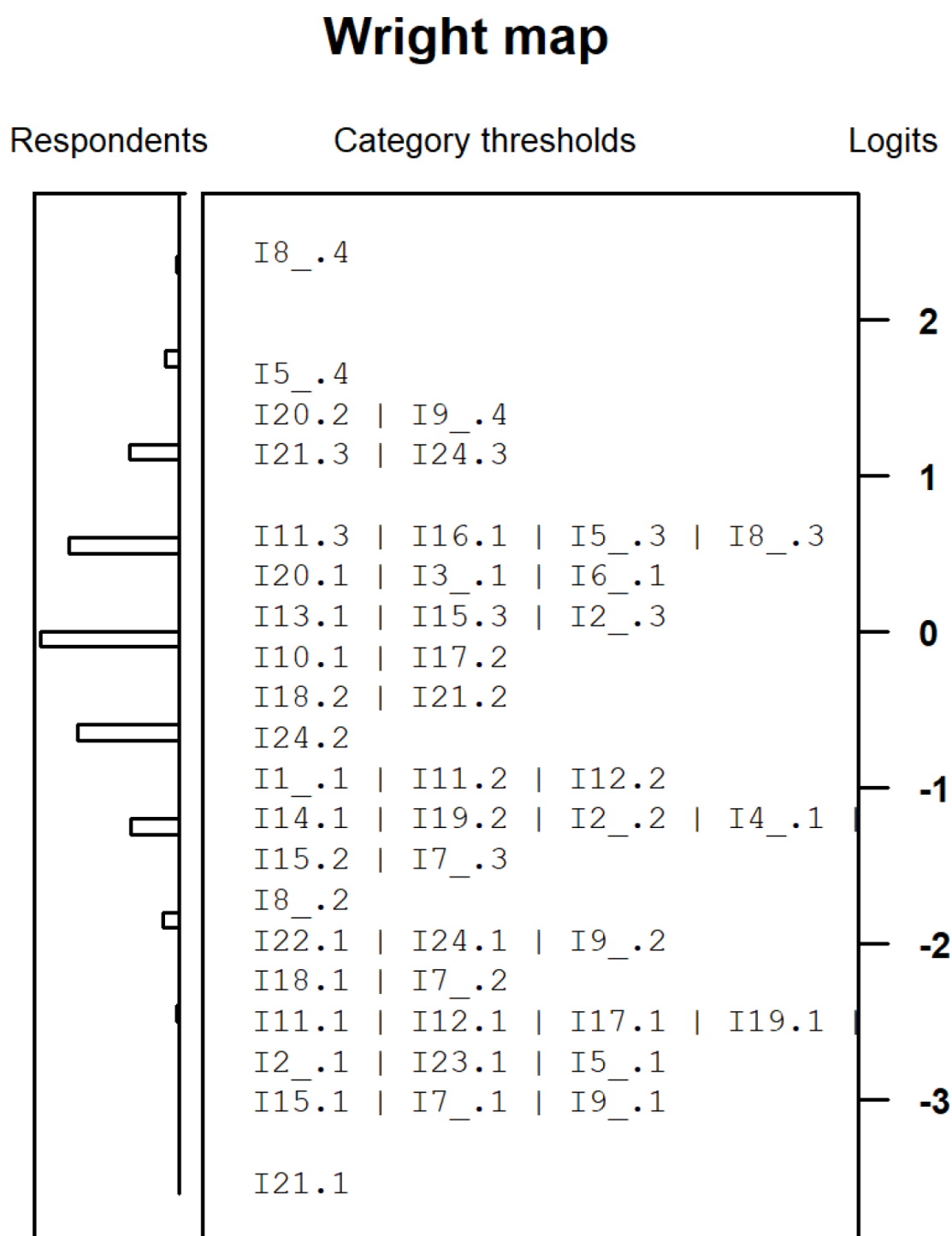
To investigate how well the distractors of the MC items performed, point-biserial correlations were calculated between each incorrect response (distractor) and the respondents' total correct scores for the remaining items. The median point-biserial correlation for the distractors was -.11 (*Min* = -.29, *Max* = .09). Only one distractor (from item dcg62050_c) out of 35 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 WMNSQ greater than 1.15. The WMNSQ values ranged from 0.87 (dcg6214s_c and dcg6418s_c) to 1.12 (dcg6209s_c), with a median of 1.00. In addition, the correlations of the correct responses with the total test scores varied between 0.13 (dcg63050_c) and 0.46 (dcg6306s_c), with a median of 0.27.

5.3.3 Differential Item Functioning

DIF was used to evaluate whether the measurement models were comparable for different subgroups. For this purpose, DIF was examined for the variables sex, school type, number of books at home (as a proxy for cultural capital), and migrant background. A description of these variables can be found in Pohl and Carstensen (2012). The differences between the estimated item difficulties (on the logit scale) in the different groups are summarized in Table 6 and Table 7. For example, the column "girls vs. boys" reports the differences in item difficulties between girls and boys; a positive value would indicate that the test was more difficult for girls, whereas a negative value would indicate that the item was less difficult for girls than for boys. In addition to examining the 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).

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.

Sex: The sample included 2,286 girls and 2,215 boys. There were small sex differences in digital competence as indicated by the main effect of 0.18 logits (Cohen's $d = 0.21$) that reflected better digital competencies for girls than boys. One item (dcg64150_c_g) showed a DIF effect of more than $|0.60|$ logits and was more difficult for boys than for girls among upper secondary school students. An overall test for DIF (see Table 8) also suggested a better fit for the model with DIF compared to a model that only estimated the main effect (but ignored potential DIF). Because an evaluation of the item content did not suggest an explanation for the observed DIF in the item dcg64150_c_g, no sex-specific item effects were acknowledged in the scaling of the test. These results suggest that for most items there was no pronounced DIF concerning sex that might have biased the parameter estimates.

School type: The sample included 843 respondents 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"). There was a negligible difference in digital competence between lower/intermediate secondary modern schools and integrated schools as indicated by the main effects of -0.11 logits (Cohen's $d = -0.15$). In contrast, there were considerable differences in digital competence between these school types and the upper secondary school as indicated by the main effects of -0.64 logits (Cohen's $d = -0.87$), and -0.53 logits (Cohen's $d = -0.72$), respectively. These reflect higher digital competencies in upper secondary schools than in other school types. One item (dcg64040_c) showed a DIF effect of $|0.60|$ logits or more. The largest DIF was found between upper secondary modern schools and lower/intermediate secondary modern schools (DIF = $|0.77|$). An overall test for DIF (see Table 8) also suggested a better fit of the model with DIF compared to a model that only estimated the main effect (but ignored potential DIF). Because an evaluation of the item content did not suggest an explanation for the observed DIF in the item dcg64040_c, no school type-specific item effects were acknowledged in the scaling of the test. These results suggest that for most items there was no pronounced DIF concerning school types that might have biased the parameter estimates.

Number of books at home: The number of books at home was used as a proxy for cultural capital. There were 1,193 participants with up to 100 books at home, 1,416 participants with more than 100 books at home, and 1,893 participants with no information on the number of books at home. All three groups were used to investigate the DIF of books at home. There were average differences between the three groups. Participants with more than 100 books at home performed on average 0.36 logits (Cohen's $d = 0.44$) higher in digital competence than participants with up to 100 books. The corresponding main effects for participants without information on this variable were 0.32 logits (Cohen's $d = 0.39$) and -0.05 logits (Cohen's $d = -0.06$), respectively. There was no considerable DIF comparing participants with many or few books at home (largest DIF = $|0.43|$ logits for the item dcg64040_c). Comparing the group without information about the number of books at home with the two groups with information, the DIF was up to $|0.37|$ logits and thus negligible. Consistent with these results, the overall test for DIF using the BIC (see Table 8) favored the model without DIF.

Migrant background: There were 2,958 participants without a migrant background, 1,179 participants with a migrant background, and 365 participants without information on their migrant background. All three groups were used to investigate the DIF of migration. There was no considerable difference between the average performance of participants with and without a migrant background. Participants without a migrant background had a higher digital

ability than participants with a migrant background (main effect = 0.21 logits, Cohen's d = 0.25). Respondents with no information on their migrant background also differed from those with no migrant background (main effect = |0.31| logits, Cohen's d = |0.38|) and those with a migrant background (main effect = |0.10| logits, Cohen's d = |0.12|). There were no items with DIF greater than |0.60| between respondents with and without a migrant background and between respondents without a migrant background and participants with no information on their migrant background. The largest difference in item difficulties between respondents with a migrant background and respondents without information on their migrant background among upper secondary school students was |0.62| logits (dgc64150_c_g). Although this appears to be a considerable difference, it may be a consequence of the uncertainty in the estimation due to the small number of respondents without information on their migrant background. Accordingly, an overall test for DIF using the BIC (see Table 8) suggested a better fit for the main effect model that did not acknowledge DIF effects.

Table 6

Differential Item Functioning (sex, school type)

Item	Sex	School		
	girls vs. boys	lower / intermediate vs. integrated	lower / intermediate vs. upper	integrated vs. upper
dgc63050_c	-0.13 (-0.15)	0.03 (0.04)	-0.33 (-0.45)	-0.36 (-0.49)
dgc6210s_c	0.09 (0.10)	0.00 (0.00)	-0.05 (-0.07)	-0.05 (-0.07)
dgc64170_c	0.24 (0.28)	0.03 (0.05)	0.28 (0.39)	0.25 (0.34)
dgc61040_c	0.10 (0.12)	0.26 (0.35)	0.48 (0.66)	0.22 (0.30)
dgc6209s_c	0.37 (0.43)	-0.11 (-0.15)	-0.35 (-0.48)	-0.24 (-0.33)
dgc62050_c	0.03 (0.04)	0.04 (0.05)	0.05 (0.07)	0.01 (0.02)
dgc6214s_c	0.07 (0.08)	-0.01 (-0.01)	0.14 (0.20)	0.15 (0.21)
dgc6102s_c	0.06 (0.07)	-0.07 (-0.09)	-0.27 (-0.36)	-0.20 (-0.27)
dgc6402s_c	-0.10 (-0.12)	-0.03 (-0.04)	-0.23 (-0.31)	-0.20 (-0.28)
dgc63080_c	-0.17 (-0.20)	-0.10 (-0.14)	-0.31 (-0.42)	-0.21 (-0.29)
dgc6114s_c	0.08 (0.10)	-0.07 (-0.10)	-0.13 (-0.18)	-0.06 (-0.09)
dgc6213s_c	0.38 (0.45)	-0.07 (-0.09)	-0.02 (-0.03)	0.04 (0.06)
dgc61080_c	-0.04 (-0.04)	0.01 (0.01)	0.24 (0.32)	0.23 (0.31)
dgc64040_c	0.05 (0.06)	0.09 (0.12)	0.77 (1.05)	0.68 (0.93)
dgc6301s_c	0.09 (0.10)	-0.03 (-0.04)	-0.20 (-0.27)	-0.16 (-0.22)
dgc64120_c	0.19 (0.22)	0.05 (0.07)	0.11 (0.15)	0.06 (0.08)

Item	Sex	School		
	girls vs. boys	lower / intermediate vs. integrated	lower / intermediate vs. upper	integrated vs. upper
dcg6310s_c	-0.16 (-0.19)	0.00 (0.01)	-0.22 (-0.30)	-0.23 (-0.31)
dcg6201s_c	0.03 (0.03)	0.02 (0.03)	0.34 (0.46)	0.32 (0.43)
dcg6418s_c	0.25 (0.29)	0.03 (0.04)	-0.18 (-0.25)	-0.21 (-0.29)
dcg6211s_c	0.25 (0.29)	0.01 (0.01)	-0.37 (-0.51)	-0.38 (-0.52)
dcg6109s_c	0.20 (0.24)	0.07 (0.10)	-0.09 (-0.13)	-0.16 (-0.22)
dcg64150_c	-0.55 (-0.65)			
dcg64150_c_g	-0.82 (-0.97)			
dcg6306s_c	-0.00 (-0.00)	-0.10 (-0.14)	-0.02 (-0.03)	0.08 (0.10)
Main effect (DIF model)	0.18 (0.21)	-0.11 (-0.15)	-0.64 (-0.87)	-0.53 (-0.72)
Main effect (Main effect model)	0.07 (0.09)	-0.07 (-0.10)	-0.50 (-0.68)	-0.43 (-0.58)

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

lower or 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 school ("Gymnasium").

na = group of respondents who did not provide information related to the group variable.

The item dcg64150_c was split by school type. dcg64150_c_g represents the responses of respondents from grammar schools, while dcg64150_c represents the responses of respondents from other types of schools.

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

Table 7*Differential Item Functioning (books at home, migrant background)*

Item	Books			Migration		
	>100 vs. <100	>100 vs. na	<100 vs. na	without vs. with	without vs. na	with vs. na
dcg63050_c	0.33 (0.40)	0.23 (0.29)	-0.09 (-0.11)	0.29 (0.34)	0.06 (0.07)	-0.23 (-0.27)
dcg6210s_c	0.07 (0.09)	0.06 (0.07)	-0.01 (-0.01)	-0.01 (-0.01)	0.07 (0.09)	0.08 (0.10)
dcg64170_c	-0.21 (-0.26)	-0.28 (-0.34)	-0.07 (-0.08)	-0.16 (-0.19)	-0.01 (-0.02)	0.14 (0.17)
dcg61040_c	-0.27 (-0.33)	-0.10 (-0.12)	0.18 (0.22)	-0.13 (-0.15)	-0.21 (-0.25)	-0.08 (-0.10)
dcg6209s_c	0.13 (0.16)	0.11 (0.14)	-0.02 (-0.02)	0.14 (0.17)	0.19 (0.23)	0.05 (0.06)
dcg62050_c	0.18 (0.21)	-0.14 (-0.17)	-0.31 (-0.38)	0.38 (0.45)	0.12 (0.14)	-0.26 (-0.31)
dcg6214s_c	0.04 (0.05)	0.04 (0.05)	-0.00 (-0.00)	-0.04 (-0.04)	0.06 (0.08)	0.10 (0.12)
dcg6102s_c	0.10 (0.12)	0.19 (0.24)	0.10 (0.12)	0.05 (0.06)	0.11 (0.13)	0.06 (0.07)
dcg6402s_c	0.06 (0.08)	0.10 (0.13)	0.04 (0.05)	-0.00 (-0.01)	0.10 (0.12)	0.11 (0.13)
dcg63080_c	0.23 (0.29)	-0.14 (-0.17)	-0.37 (-0.45)	-0.02 (-0.02)	-0.01 (-0.01)	0.01 (0.01)
dcg6114s_c	-0.02 (-0.03)	0.00 (0.00)	0.02 (0.03)	0.08 (0.10)	0.11 (0.13)	0.03 (0.04)
dcg6213s_c	-0.03 (-0.04)	-0.06 (-0.07)	-0.02 (-0.03)	0.02 (0.03)	0.02 (0.02)	-0.00 (-0.00)
dcg61080_c	-0.19 (-0.23)	-0.13 (-0.15)	0.06 (0.08)	-0.06 (-0.08)	0.02 (0.02)	0.08 (0.10)
dcg64040_c	-0.43 (-0.53)	-0.11 (-0.14)	0.32 (0.39)	-0.36 (-0.43)	-0.28 (-0.34)	0.07 (0.09)
dcg6301s_c	0.11 (0.13)	0.14 (0.17)	0.03 (0.03)	0.07 (0.08)	0.14 (0.17)	0.07 (0.08)
dcg64120_c	-0.05 (-0.06)	0.10 (0.12)	0.15 (0.18)	-0.08 (-0.10)	0.12 (0.14)	0.20 (0.24)
dcg6310s_c	0.19 (0.23)	0.21 (0.26)	0.02 (0.03)	0.29 (0.34)	0.33 (0.40)	0.04 (0.05)

Item	Books			Migration		
	>100 vs. <100	>100 vs. na	<100 vs. na	without vs. with	without vs. na	with vs. na
dcg6201s_c	-0.21 (-0.26)	-0.11 (-0.13)	0.10 (0.13)	-0.09 (-0.11)	-0.15 (-0.18)	-0.06 (-0.07)
dcg6418s_c	0.04 (0.05)	0.08 (0.10)	0.04 (0.05)	0.19 (0.23)	0.14 (0.16)	-0.06 (-0.07)
dcg6211s_c	0.29 (0.36)	0.13 (0.16)	-0.16 (-0.20)	0.14 (0.16)	0.22 (0.27)	0.09 (0.10)
dcg6109s_c	0.05 (0.06)	0.04 (0.05)	-0.01 (-0.01)	0.06 (0.08)	0.13 (0.15)	0.06 (0.08)
dcg64150_c	0.09 (0.11)	-0.07 (-0.09)	-0.16 (-0.20)	-0.44 (-0.53)	-0.39 (-0.47)	0.05 (0.06)
dcg64150_c_g	-0.41 (-0.51)	-0.18 (-0.22)	0.24 (0.29)	0.08 (0.10)	-0.54 (-0.65)	-0.62 (-0.75)
dcg6306s_c	0.01 (0.02)	0.02 (0.02)	0.01 (0.01)	-0.00 (-0.00)	-0.07 (-0.08)	-0.06 (-0.08)
Main effect (DIF model)	0.36 (0.44)	0.32 (0.39)	-0.05 (-0.06)	0.21 (0.25)	0.31 (0.38)	0.10 (0.12)
Main effect (Main effect model)	0.29 (0.36)	0.23 (0.28)	-0.06 (-0.07)	0.16 (0.19)	0.21 (0.25)	0.04 (0.05)

Note. Raw differences between item difficulties (in logits) with standardized differences (Cohen's *d*) in parentheses. The item dcg64150_c was split by school type. dcg64150_c_g corresponds to the values of respondents from the upper secondary modern schools, while dcg64150_c represents here the values of respondents from other school types.

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

Table 8*Comparisons of Models with and without DIF*

DIF variable	Model	N	Deviance	Number of parameters	AIC	BIC
Sex	Main effect	4,501	167,207	54	167,315	167,662
	DIF	4,501	166,858	77	<u>167,012</u>	<u>167,506</u>
School	Main effect	4,338	158,778	53	158,884	159,222
	DIF	4,338	158,256	95	<u>158,446</u>	<u>159,052</u>
Books	Main effect	4,502	167,021	55	167,131	<u>167,484</u>
	DIF	4,502	166,810	101	<u>167,012</u>	167,660
Migration	Main effect	4,502	167,147	55	167,257	<u>167,610</u>
	DIF	4,502	166,970	101	<u>167,172</u>	167,819

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

5.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item discrimination parameters are equal. To test this assumption, a generalized partial credit model (GPCM, Muraki, 1992) was fitted to the data to estimate the discrimination parameters. The estimated discrimination parameters differed moderately between items (see Table 4). The median discrimination parameter fell at 0.60 (*Min* = 0.21, *Max* = 1.45). Model fit indices indicated a better model fit of the GPCM (AIC = 165,911, BIC = 166,386, number of parameters = 74) compared to the PCM model (AIC = 167,358, BIC = 167,685, number of parameters = 51). However, an inspection of the respective item characteristic curves indicated an adequate fit of the PCM. Despite the empirical preference for the GPCM, the PCM is more in line with the theoretical concepts underlying the test construction (see Pohl & Carstensen, 2012). For this reason, the PCM was chosen as our scaling model in order to preserve the item weightings as intended in the theoretical framework.

5.3.5 Dimensionality

The dimensionality of the test was investigated by evaluating the correlations between the residuals of the PCM. The adjusted Q_3 statistics (see Table 4) were quite low (*Mdn* = 0.03) - the largest absolute residual correlation was 0.07 (dcg6418s_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 9*Results of four-dimensional scaling for facets of digital competence*

Dimension	Dim 1	Dim 2	Dim 3	Dim 4
Dim 1: Evaluating information	0.74			
Dim 2: Detecting intentions and strategies	.88	1.05		
Dim 3: Communicating and interacting	.87	.89	0.59	
Dim 4: Handling data	.91	.91	.87	0.82

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 9. All dimensions exhibited moderate variances. As expected, the correlations between the four facets were rather high, ranging between .87 and .91. Because they were smaller than $r = .95$, the different facets seem to capture some unique variance in addition to their shared variances (see Carstensen, 2013). Accordingly, the four-dimensional model did fit the data better (AIC = 167,252, BIC = 167,637, number of parameters = 60) than the unidimensional model (AIC = 167,358, BIC = 167,685, number of parameters = 51).

These results suggest that the digital competence test is an essentially unidimensional test. However, the four facets measure a common construct but capture some unique variance components beyond the common digital competence factor. Because the digital 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 digital competence test that was administered in Grade 6 of NEPS Starting Cohort 8. Different types of missing responses were examined, item fit statistics and item characteristic curves were evaluated, and item discriminations were examined. Apart from the items that were not reached, the number of omitted items was quite small. Further quality inspections were carried out by examining differential item functioning and testing Rasch-homogeneity. Several criteria indicated a good fit of the items and measurement invariance across different subgroups. An exception is one item (dcg64150_c) that indicated a strong DIF effect for sex and school type.

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

The different facets of the test induced some multidimensionality. 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 digital competence.

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

7. Data in the Scientific Use Files

7.1 Naming Conventions

The SUF for the starting cohort contains 23 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).

7.2 Digital Competence Scores

In the SUF, manifest digital competence scores are provided in the form of WLEs (dcg6_sc1) including their respective standard errors (dcg6_sc2).

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. Alternatively, users interested in examining latent relationships can either include the measurement model in their analyses or estimate plausible values. Plausible values for the digital competence test can be estimated using the R package *NEPSscaling*⁵ (Scharl et al., 2020; Scharl & Zink, 2022; Zink et al., 2026).

⁵<https://www.neps-data.de/Data-Center/Overview-and-Assistance/Plausible-Values>

References

- Adam, R., Cloney, D., Wu, M., Berenzer, A., Osses, A., & Schwantner, V. (2023). *ACER ConQuest Manual*. Australian Council for Educational Research. <https://conquestmanual.acer.org>
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–722. <https://doi.org/10.1109/TAC.1974.1100705>
- 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. (2025). *NEPS Technical Report for reading: Scaling results of Starting Cohort 8 in Grade 5* (NEPS Survey Paper No. 117). Leibniz Institute for Educational Trajectories. <https://doi.org/10.5157/NEPS:SP117:1.0>
- 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.
- 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	dcg6210s_c	Detecting Intentions and Strategies
3	dcg64170_c	Handling Data
4	dcg61040_c	Evaluating Information
5	dcg6209s_c	Detecting Intentions and Strategies
6	dcg62050_c	Detecting Intentions and Strategies
7	dcg6214s_c	Detecting Intentions and Strategies
8	dcg6102s_c	Evaluating Information
9	dcg6402s_c	Handling Data
10	dcg63080_c	Communicating and Interacting
11	dcg6114s_c	Evaluating Information
12	dcg6213s_c	Detecting Intentions and Strategies
13	dcg61080_c	Evaluating Information
14	dcg64040_c	Handling Data
15	dcg6301s_c	Communicating and Interacting
16	dcg64120_c	Handling Data
17	dcg6310s_c	Communicating and Interacting
18	dcg6201s_c	Detecting Intentions and Strategies
19	dcg6418s_c	Handling Data
20	dcg6211s_c	Detecting Intentions and Strategies
21	dcg6109s_c	Evaluating Information
22	dcg64150_c	Handling Data
23	dcg6306s_c	Communicating and Interacting

Appendix B: R Syntax for Estimating WLEs

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

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

# items of the competence test
items <- c(
  "dgc6102s_c", "dgc61040_c", "dgc61080_c",
  "dgc6109s_c", "dgc6114s_c", "dgc6201s_c",
  "dgc62050_c", "dgc6209s_c", "dgc6210s_c",
  "dgc6211s_c", "dgc6213s_c", "dgc6214s_c",
  "dgc6301s_c", "dgc63050_c", "dgc6306s_c",
  "dgc63080_c", "dgc6310s_c", "dgc6402s_c",
  "dgc64040_c", "dgc64120_c", "dgc64150_c",
  "dgc64170_c", "dgc6418s_c"
)

# polytomous items
poly <- c(
  "dgc6102s_c", "dgc6109s_c", "dgc6114s_c",
  "dgc6201s_c", "dgc6209s_c", "dgc6210s_c",
  "dgc6211s_c", "dgc6213s_c", "dgc6214s_c",
  "dgc6301s_c", "dgc6306s_c", "dgc6310s_c",
  "dgc6402s_c", "dgc6418s_c"
)

# collapse response categories
dat$dgc6102s_c <-
  doBy::recodeVar(
    dat$dgc6102s_c, c(0, 1, 2, 3, 4, 5), c(0, 0, 1, 2, 3, 4)
  )
dat$dgc6109s_c <-
  doBy::recodeVar(
    dat$dgc6109s_c, c(0, 1, 2, 3, 4), c(0, 0, 1, 2, 3)
  )
dat$dgc6214s_c <-
  doBy::recodeVar(
    dat$dgc6214s_c, c(0, 1, 2, 3, 4, 5), c(0, 0, 0, 1, 2, 3)
  )
dat$dgc6114s_c <-
  doBy::recodeVar(
```

```

    dat$dcg6114s_c, c(0, 1, 2, 3, 4, 5), c(0, 0, 0, 1, 2, 3)
  )
dat$dcg6201s_c <-
  doBy::recodeVar(
    dat$dcg6201s_c, c(0, 1, 2, 3, 4), c(0, 0, 0, 1, 2)
  )
dat$dcg6213s_c <-
  doBy::recodeVar(
    dat$dcg6213s_c, c(0, 1, 2, 3, 4), c(0, 0, 0, 1, 2)
  )
dat$dcg6210s_c <-
  doBy::recodeVar(
    dat$dcg6210s_c, c(0, 1, 2, 3, 4, 5), c(0, 0, 0, 1, 2, 3)
  )
dat$dcg6211s_c <-
  doBy::recodeVar(
    dat$dcg6211s_c, c(0, 1, 2, 3, 4, 5), c(0, 0, 0, 0, 1, 2)
  )
dat$dcg6301s_c <-
  doBy::recodeVar(
    dat$dcg6301s_c, c(0, 1, 2, 3, 4, 5), c(0, 0, 0, 1, 2, 3)
  )
dat$dcg6306s_c <-
  doBy::recodeVar(
    dat$dcg6306s_c, c(0, 1, 2, 3, 4, 5, 6), c(0, 0, 0, 0, 1, 2, 3)
  )
dat$dcg6310s_c <-
  doBy::recodeVar(
    dat$dcg6310s_c, c(0, 1, 2, 3), c(0, 0, 1, 2)
  )
dat$dcg6402s_c <-
  doBy::recodeVar(
    dat$dcg6402s_c, c(0, 1, 2, 3, 4, 5, 6), c(0, 0, 0, 1, 2, 3, 4)
  )
dat$dcg6418s_c <-
  doBy::recodeVar(
    dat$dcg6418s_c, c(0, 1, 2, 3, 4), c(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)
```