NEPS *SURVEY PAPERS*

Anna-Lena Kock, Lara Aylin Petersen, and Kristin Litteck

NEPS TECHNICAL REPORT FOR MATHEMATICS: SCALING RESULTS OF STARTING COHORT 4 FOR GRADE 9 IN SPECIAL SCHOOLS

NEPS *Survey Paper* No. 73
Bamberg, September 2020

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

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

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 4 for Grade 9 in Special Schools

Anna-Lena Kock, Lara Aylin Petersen, and Kristin Litteck

Leibniz Institute for Science and Mathematics Education, Kiel

E-mail address of lead author:

alkock@leibniz-ipn.de

Bibliographic data:

Kock, A.-L., Petersen, L.A., & Litteck, K. (2020). *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 4 for Grade 9 in Special Schools* (NEPS *Survey Paper* No. 73). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP73:1.0

Acknowledgements:

This report is an extension to NEPS *Working Paper* 22 (Duchhardt & Gerdes, 2013) that presents the scaling results for mathematical competence of Starting Cohort 4 for Grade 9 in general schools. Therefore, various parts of this report are reproduced *verbatim* to facilitate the understanding of the presented results.

NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 4 for Grade 9 in Special Schools

Abstract

The National Educational Panel Study (NEPS) investigates the development of competencies across the life span and develops tests for the assessment of different competence domains. In order to evaluate the quality of the competence tests, various analyses based on item response theory (IRT) are performed. This paper describes the data and scaling procedures for the mathematical competence test in Grade 9 of Starting Cohort 4 (ninth grade) that was administered to students in special schools. The feasibility of including students with special educational needs in the NEPS was investigated with two different test versions. Both versions contained some items of the test from the main sample of students from general school. Additionally, version 1 contained some easier items for students of the ninth grade and some items designed for a younger age cohort. In version 2, all additional items were designed for a younger age cohort, resulting in an easier test as compared to version 1. Both tests were shorter than the test administered to the main sample. The two test versions were randomly distributed among a sample of *N* = 1,086 students (45 % girls) from special schools. The responses were scaled using the Rasch model. Item fit statistics, differential item functioning, and Rasch-homogeneity were evaluated to examine the quality of the tests. These analyses showed that the tests exhibited limited item fits, variances and reliabilities, thus, allowing only rather crude analyses of interindividual differences between students with special educational needs. For this reason, no mathematical ability score could be estimated. Importantly, there was substantial differential item functioning between special schools and lower secondary schools. Therefore, comparative analyses between the two school types are not recommended. Overall, these results highlight substantial difficulties in assessing mathematical competence among students with special educational needs at special schools in educational large-scale assessments. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the Conquest syntax for scaling the data.

Keywords

item response theory, special educational needs, mathematical competence, scientific use file

Content

1. Introduction

Within the National Educational Panel Study (NEPS) different competences are measured coherently across the life span. These include, among others, mathematical competence, reading competence, scientific literacy, and information and communication technologies literacy. An overview of the competences measured in the NEPS is given by Weinert and colleagues (2011) as well as Fuß, Gnambs, Lockl, and Attig (2019). Most of the competence data are scaled using models of item response theory (IRT). Because the tests were developed specifically for implementation in the NEPS, several analyses are conducted to evaluate their quality. The IRT model chosen for scaling the competence data and the analyses performed for checking the quality of the scale are described in Pohl and Carstensen (2012).

The main sample of the NEPS includes students from different school type across Germany (see Duchhardt & Gerdes, 2013). In Grade 9 of Starting Cohort 4 (ninth grade), a feasibility study was conducted to evaluate whether and how students from special schools (i.e. schools specializing in the education of special needs children) could be validly and meaningfully included in the NEPS. In this paper, the results of these analyses are presented for a mathematical competence test administered to students with special educational needs attending special schools. First, the main concepts of the mathematical competence tests and the test design are introduced. Then, the mathematical competence data of Starting Cohort 4 and the analyses performed to check the quality of the tests are described. Finally, an overview of the data that are available for public use in the Scientific Use File (SUF) is presented.

Please note that the analyses in this report are based on the data available at some time before 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 paper. However, we do not expect fundamental changes in the presented results.

2. Testing Mathematical Competence

The framework and test development for the mathematical competence test are described by Weinert et al. (2011), Neumann et al. (2013), and Ehmke et al. (2009). In the following, specific aspects of the mathematical competence tests will be pointed out that are necessary for understanding the scaling results presented in this paper.

In this study two different test versions were administered. The items were not arranged in units. Thus, in the tests, students faced a certain situation followed by one or two tasks related to it. Each of the items belonged to one of the following content areas:

- quantity,
- space and shape,
- change and relationships,
- data and chance.

Each item was constructed in such a way as to primarily address a specific content area (see Appendix A). The framework also describes as a second and independent dimension six cognitive components required for solving the tasks. These are distributed across the items.

In the mathematics tests there were two types of response formats. These were simple multiple-choice (MC) and short constructed responses (SCR). In MC items the test taker had to find the correct answer from either four or five response options. SCR tasks required the test taker to write down an answer into one or two empty boxes. Examples of the different response formats are given in Pohl and Carstensen (2012).

2.1 The Design of the Study

The study assessed different cognitive domains including mathematical competence, reading speed, listening comprehension at word level, and attentional capacity (see Gnambs & Freund, 2019). These four domains were randomly distributed among the participants. The study adopted an experimental design and administered two different versions of the mathematical competence test:

- Test version 1 included 20 items: eight items from the test administered to the main sample in Grade 9 of Starting Cohort 4 (see Haberkorn, Pohl, Hardt, & Wiegand, 2012), two easier items designed for ninth-grade, and ten items designed for a younger age cohort. Preliminary analyses identified severe misfit of items mag5v271_sc4g9_c, mag9r051_c, mag9d241_c, and mag9q081_c (e.g., item-total correlation). Therefore, these items were excluded from the analyses, resulting in a test with 16 items.

- Test version 2 also included 20 items: six items from the test administered to the main sample and 14 items designed for a younger age cohort. Again, some items were excluded from the analyses due to severe misfit, resulting in a test with 18 items. The following items were excluded from further analyses: mag5r101_sc4g9_c and mag9r111_c (e.g., item-total correlation).

Both test versions were also administered in reversed order. As no substantial position effects were found (see 4.4.1), tests in general and in reversed order were scaled concurrently for both test versions. Both tests were shorter than the test administered to the main sample. Table 1 shows the distribution of the four content areas (see Appendix A for the assignment of the items to the content areas), whereas Table 2 shows the distribution of the response formats.

Table 1. *Number of Items for the Different Content Areas by Test Version*

Content area	**Version 1**	**Version 2**
Quantity	6	6
Space and shape	5	6
Change and relationships	5	5
Data and chance	4	3
Total number of items	20	20

Table 2. *Number of Items for the Different Response Formats by Test Version*

Response format	Version 1	Version 2
Simple multiple choice	15	12
Short constructed response	5	8
Total number of items	20	20

2.2 Samples

Overall, a total of 1,098[1] students from special schools received the mathematical competence tests. For 12 respondents less than three valid item responses were available. Because no reliable ability scores can be estimated based on such few valid responses, these cases were excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 1,086 individuals (45 % female) from special schools. A summary of basic descriptive statistics for this sample is given in Table 3. A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

Table 3. *Sample Description by Test Version*

	Special schools (Version 1 & 2)	Special schools (Version 1)	Special schools (Version 2)
Sample size (*N*)	1086	554	532
Female (%)	45 %	45 %	45 %
Migration background (%)	27 %	28 %	26 %
100+ books at home (%)	18 %	18 %	18

3. Analyses

3.1 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach, d) items that have not been administered, and finally, e) multiple kinds of missing responses within SCR items that are not determined.

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

Invalid responses occurred, for example, when two response options were selected in simple MC items where only one was required, or when numbers or letters that were not within the range of valid responses were given as a response. Omitted items occurred when test takers skipped some items. Due to time limits, not all persons finished the test within the given time (30 minutes). All missing responses after the last valid response given were coded as not-reached. As SCR items were sometimes aggregated from several subtasks, different kinds of missing responses might be found in these items. A SCR item was coded as missing if at least one subtask contained a missing response. When just one kind of missing response occurred, the item was coded according to the corresponding missing response. When the subtasks contained different kinds of missing responses, the item was labeled as a not-determinable missing response.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions, handling of different response formats). They also need to be accounted for in the estimation of item and person parameters. Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the persons were coping with the test. Missing responses per item were examined in order to evaluate how well each of the items functioned.

3.2 Scaling Model

Item and person parameters were estimated using a Rasch model (Rasch, 1980). A detailed description of the scaling model can be found in Pohl and Carstensen (2012). All items were scored dichotomously as 0 for an incorrect and 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats).

Mathematical competencies are usually estimated as weighted maximum likelihood estimates (WLE; Warm, 1989). However, due to insufficient reliabilities and variances, no reliable mathematical ability score could be estimated in the present study. The data available in the SUF is described in section 6.

3.3 Checking the Quality of the Tests

The mathematical competence tests were specifically constructed to be implemented in the NEPS. In order to ensure appropriate psychometric properties, the quality of the tests was examined in several analyses.

The MC items consisted of one correct response option and three or four distractors (i.e., incorrect response options). The quality of the distractors within MC items was examined using the point-biserial correlation between selecting an incorrect response option and the rest item total correct score. Negative correlations indicate good distractors, whereas correlations between .00 and .05 are considered acceptable and correlations above .05 are viewed as problematic distractors (Pohl & Carstensen, 2012).

The SCR items require the test-taker to give mostly one-word answers, such as numbers. All SCR items were scored dichotomously even if there was more than one response required.

The fit of the dichotomous MC and SCR items to the Rasch model (Rasch, 1980) was evaluated using three indices (see Pohl & Carstensen, 2012). Items with a weighted mean square (WMNSQ) > 1.15 (*t*-value > |6|) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 (*t*-value > |8|) were judged as having a considerable item misfit and their performance was further investigated. Correlations of the item score with the corrected total score (equal to the discrimination value as computed in ConQuest) greater than .30 were considered as good, greater than .20 as acceptable, and below .20 as problematic. Overall, judgment of the fit of an item was based on all fit indicators. Moreover, the model-implied and empirical item characteristic curves were compared to identify a potential item misfit.

The mathematical competence tests should measure the same construct for all students. If some items favored certain subgroups (e.g., they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., school types) would be biased and, thus, unfair. For the present study, test fairness was investigated for the variables gender, the number of books at home (as a proxy for socioeconomic status), and migration background, (see Pohl & Carstensen, 2012, for a description of these variables). Moreover, test fairness was also evaluated for the different test versions administered in special schools to determine whether a common mathematical score might be derived. Differential item functioning (DIF) was examined using a multigroup IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled and compared. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties between the subgroups that were greater than 1 logit as strong DIF, absolute differences between 0.60 and 1 as considerable and noteworthy of further investigation, absolute differences between 0.40 and 0.60 as small but not severe, and differences smaller than 0.25 as negligible DIF. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

The competence data in the NEPS are usually scaled assuming Rasch-homogeneity. The Rasch (1980) model was chosen because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that might not hold for empirical data. To test the assumption of equal item discrimination parameters, a two-parametric logistic model (2PL; Birnbaum, 1968) was also fitted to the data and compared to the Rasch model.

The dimensionality of the mathematical tests was evaluated by specifying a four-dimensional model based on the four content areas. Each item was assigned to one content area (between-item-multidimensionality). To estimate this multidimensional model, Quasi Monte Carlo integration in TAM (version 3.3-10) in *R* version 3.6.2 (R Core Team, 2017) was used. To guarantee the compatibility with the multidimensional model, the unidimensional model was estimated in TAM as well. The number of nodes in the multidimensional model was chosen in such a way as to obtain stable parameter estimates (15,000 nodes). The correlations between the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the tests.

3.4 Software

The IRT models were estimated in ConQuest version 4.5.2 (Adams, Wu, & Wilson, 2015). The 2PL model was estimated in mdltm (Matthias von Davier, 2005).

4. Results

4.1 Missing Responses

4.1.1 Missing responses per person

Figure 1 shows the number of invalid responses per person by administered test version. Overall, there were very few invalid responses. 80.9 % and 66.0 % of the students in special schools gave no invalid response at all. More than one invalid response was observed for 2.7 % and 9.0 % of students in special schools respectively.

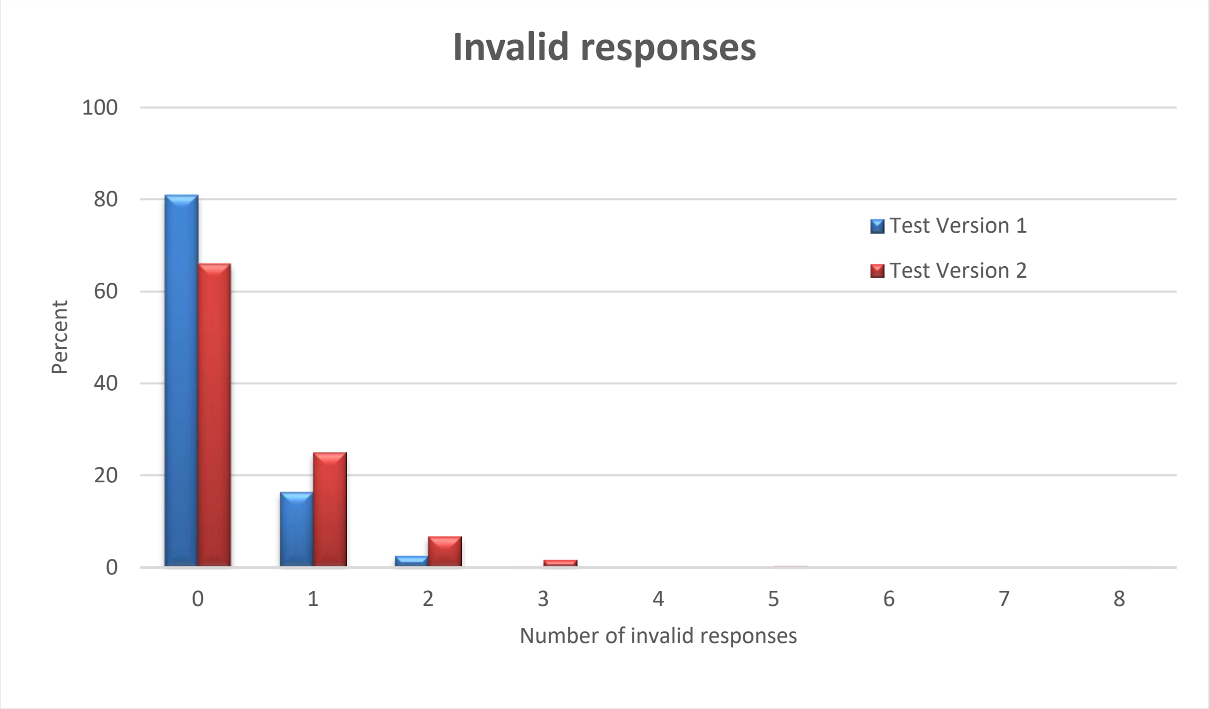

Figure 1. Number of invalid responses by test version

Missing responses also occur when respondents skip (omit) items. As illustrated in Figure 2, most respondents in special schools (47.3 % and 46.4 % respectively) did not skip any item and 19.9 % and 25.0 % omitted more than two items.

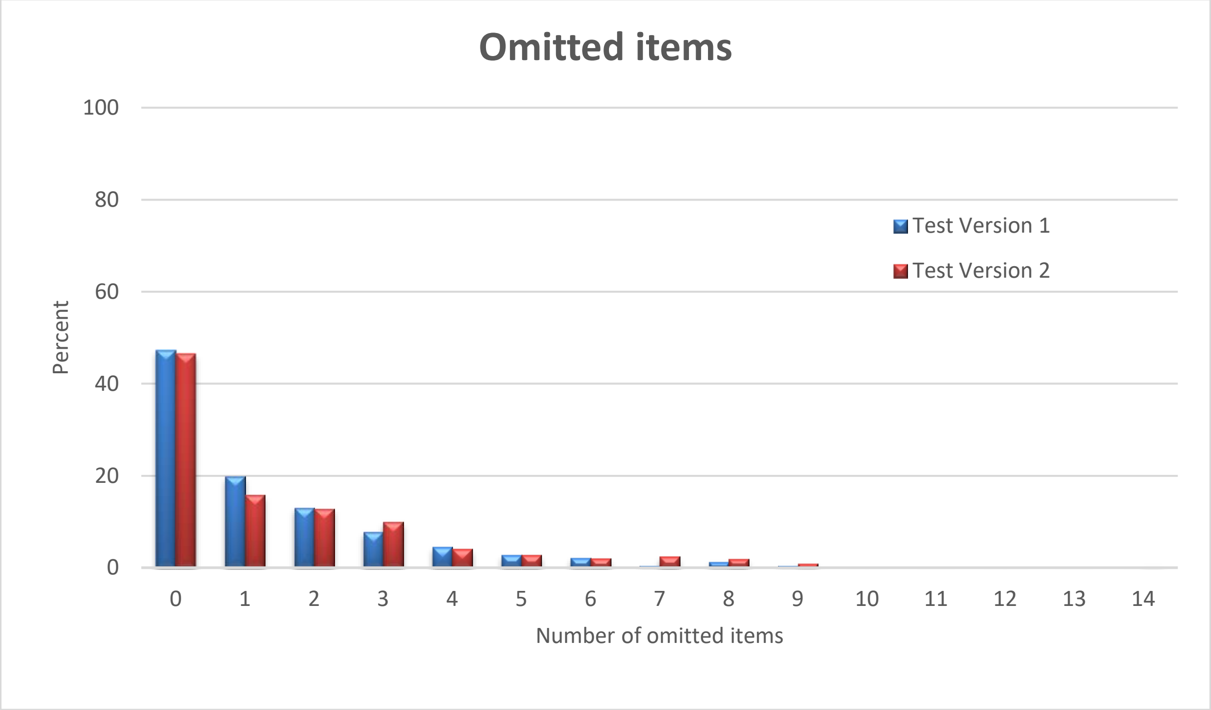

Figure 2. Number of omitted items by test version

All missing responses after the last valid response are defined as not reached. 86.6 % and 92.7 % of students in special schools finished all items, whereas 11.9 % and 5.8 % of the test takers had not reached one to five items. 1.4 % and 1.5 % of the students had not reached more than five items (Figure 3).

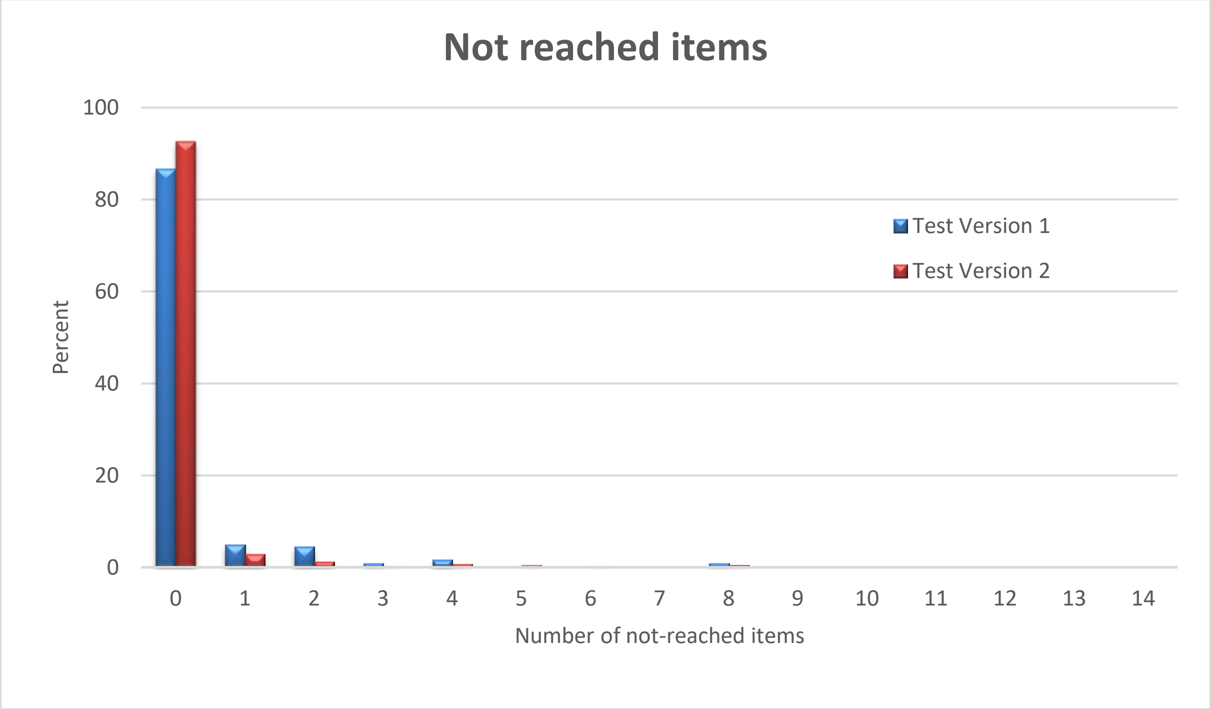

Figure 3. Number of not-reached items by test version

The SCR items were coded as not-determinable when the subtasks contained different kinds of missing responses. Because these missings only occur in SCR items, the maximum possible number of not-determinable missing responses was five in version 1 and eight in version 2 (see Table 2). There were no substantial missing responses that were not determinable (1.6 % and 1.5 %; Figure 4).

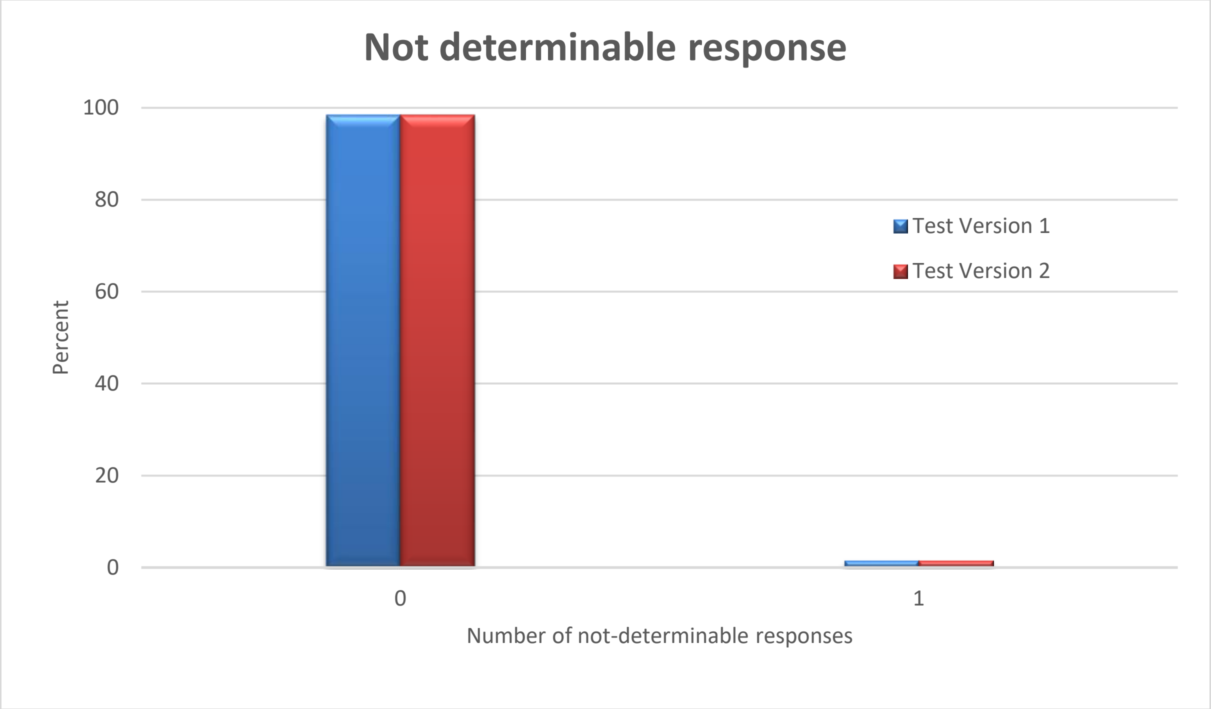

Figure 4. Number of not-determinable responses by test version

The total number of missing responses, aggregated over invalid, omitted, not-reached, and not-determinable missing responses per person, is illustrated in Figure 5. About 33.4 % and 26.3 % of the test takers had no missing response at all, whereas about 12.1 % and 16.9 % of participants had five or more missing responses.

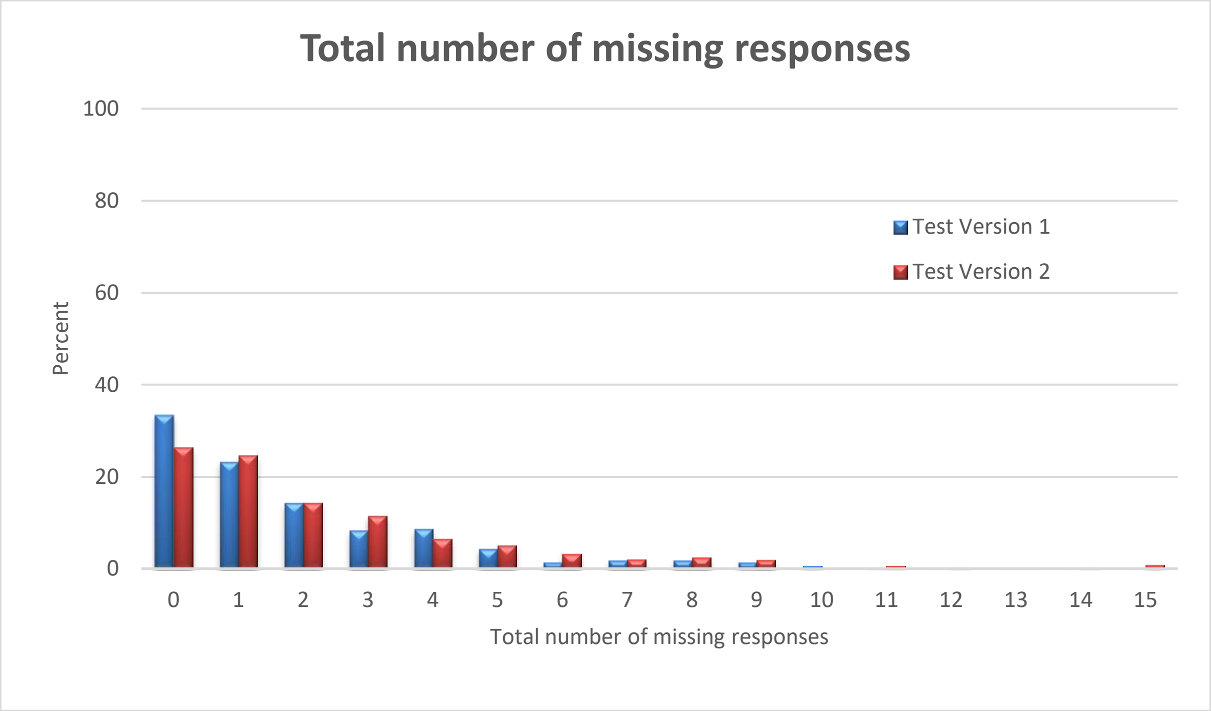

Figure 5. Total number of missing responses by test version

Overall, there is a negligible amount of invalid, not reached and not-determinable missing responses and a reasonable amount of omitted items. Furthermore, the total number of items used in this test and, thus, the time needed for the completion of the test seems appropriate for students of special schools.

4.1.2 Missing responses per item

Table 4 provides information on the occurrence of different kinds of missing responses per item for the different test versions. SCR items exhibited large omission rates (10.7 % – 22.2 % in version 1, 8.1 % – 20.1 % in version 2). The students might have rather skipped these items than guessing an answer. An allocation of items to response formats can be found in Appendix A. The items included from the test administered in the main field exhibited in this sample noticeably higher omission rates than in the main field itself (2.5 % – 11.2 % in version 1, 1.7 % – 7.7 % in version 2). This potentially indicates a higher difficulty for students with special educational needs. The rest of the items had acceptable omission rates of 0.9 % – 5.1 % for both versions. The number of persons that did not reach an item increased up to 7.8 % for version 1 and 4.1 % for version 2. As version 1 is the more difficult test version, this difference was to be expected. Both rates are rather small and acceptable. The items not reached by test position are depicted in Figure 6 for the different test booklets. The number of invalid responses per item is small for most of the items. The highest numbers are 8.3 % and 10.9 % for the related items mag5d02s_sc4g9_c and mag5v024_sc4g9_c. This might, again, be due to the open response format. The number of not-determinable items is small.

Table 4. Percentage of Missing Values by test version

		Test version 1					*Test version 2*				
	Item	***N***	**NR**	**OM**	**NV**	**ND**	***N***	**NR**	**OM**	**NV**	**ND**
1	mag5d041_sc4g9_c	503	7.8	0.9	0.5	0.0	500	4.1	1.7	0.2	0.0
2	mag5q291_sc4g9_c	392	5.1	22.2	2.0	0.0	404	2.1	20.1	1.9	0.0
3	mag5v271_sc4g9_c						498	1.3	5.1	0.0	0.0
4	mag5r171_sc4g9_c	507	2.2	5.1	1.3	0.0	496	1.3	4.0	1.5	0.0
5	mag9r111_c	481	1.8	11.2	0.2	0.0					
6	mag5q301_sc4g9_c	475	0.9	10.7	2.7	0.0	469	0.8	8.1	3.0	0.0
7	mag9d151_c	518	0.9	5.1	0.5	0.0	505	0.9	3.4	0.8	0.0
8[e]	mag9r051_c										
9	mag9v011_c	514	0.7	6.3	0.2	0.0	495	0.4	5.5	1.1	0.0
10	mag9v012_c	499	0.7	9.0	0.2	0.0	493	0.6	6.4	0.4	0.0
11	mag5q140_sc4g9_c	454	0.5	13.7	2.7	1.1	437	0.4	13.7	2.6	1.1
12[e]	mag9d241_c										
13	mag9r191_c	459	0.5	9.8	0.4	0.0	481	0.8	7.7	1.1	0.0
14	mag5r101_sc4g9_c	535	0.5	1.8	1.1	0.0					
15	mag9q181_c	535	0.7	2.5	0.2	0.0	515	0.9	1.7	0.6	0.0
16	mag9q221_c	507	1.4	3.4	3.6	0.0					
17	mag9d260_c	418	2.0	19.7	2.4	0.5					
18[e]	mag9q081_c										
19	mag5v321_sc4g9_c	442	3.6	12.6	4.0	0.0	414	2.4	17.5	2.3	0.0
20	mag5v091_sc4g9_c	508	5.6	2.5	0.2	0.0					
21	mag5q221_sc4g9_c						419	0.8	17.3	3.2	0.0
22	mag5r201_sc4g9_c						519	0.8	1.5	0.2	0.0
23	mag5q131_sc4g9_c						423	1.5	17.1	1.9	0.0
24	mag5d02s_sc4g9_c						409	1.5	13.0	8.3	0.4
25	mag5v024_sc4g9_c						363	1.9	10.9	10.9	0.0
26	mag5r191_sc4g9_c						468	3.2	1.7	7.1	0.0

Note. *N* = Number of valid responses, NR = Percentage of respondents that did not reach item, OM = Percentage of respondents that omitted the item, NV = Percentage of respondents with an invalid response, ND = Percentage of respondents with a not-determinable response, [e] Excluded from the analyses due to unsatisfactory item fit.

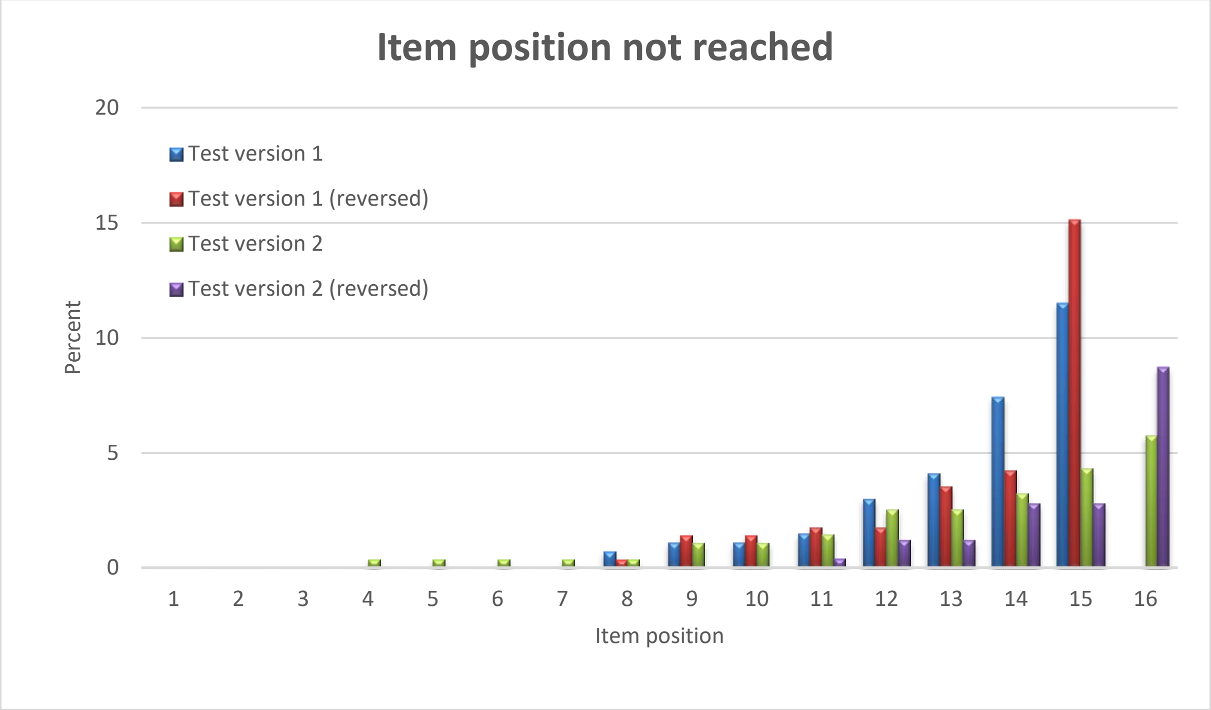

Figure 6. Item position not reached by test version

4.2 Parameter Estimates for Different Test Versions

4.2.1 Distractor analyses

To investigate how well the distractors of the MC items performed, the point-biserial correlations between each incorrect response (distractor) and the students' total correct scores were calculated (see Table 5). The mean point-biserial correlations for the distractors were -.06 for test version 1 and -.09 for test version 2. In contrast, the correlations of the correct responses with the total scores were M = .20 for version 1 and M = .23 for version 2. These results indicate that the distractors did not function well for students in special schools as compared to students in regular schools (Duchardt & Gerdes, 2013). As all items were developed for students from regular schools and have already shown good item fit in previous studies (see Duchhardt & Gerdes, 2012, 2013), this seems to be an issue that is specific for students with special educational needs.

Table 5. Distractor Analyses for Test Versions in Special Schools

	Distractors			***Correct response***		
Test version	***Mean***	***Min***	***Max***	***Mean***	***Min***	***Max***
Version 1	-.06	-.21	.08	.20	.13	.27
Version 2	-.09	-.27	-.06	.23	.12	.33

Note. Reported are point-biserial correlations between the distractor or correct response and the total score.

4.2.2 Item parameters

The item parameters for the different test versions are summarized in Table 6. Detailed results for each test version are given in Appendix B. The percentage of correct responses in relation to all valid responses for each item was higher for the more difficult test version 2. On average, the rates of correct responses were 28 % for version 1 and 32 % for version 2. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. As anticipated, the item difficulties were higher in version 1 (*M* = 1.2) than in version 2 (*M* = 1.0). The combination of low percentages of correct responses and high mean difficulties indicate that both test versions were too difficult.

Table 6. Summary of Item Parameters for Different Test Versions

Test version	**Percentage correct**	ξ	**WMNSQ**	***t***	**Item-rest correlation**	**Item-total correlation**
Version 1	28 [5, 47]	1.2 [0.2, 3.2]	1.0 [0.9, 1.0]	0.0 [-1.3, 1.3]	.2 [.1, .4]	.4 [.3, .5]
Version 2	32 [8, 61]	1.0 [-0.5, 2.8]	1.0 [0.9, 1.1]	-0.1 [-2.3, 1.4]	.3 [.1, .4]	.4 [.3, .6]

Note. Reported are mean values across all items with minimum and maximum value in parentheses. ξ = Item difficulty, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ. Percent correct scores are not informative for short constructed response items and, thus, are not acknowledged.

4.2.3 Item fit

Altogether, item fit for the different test versions administered in special schools can be considered to be satisfactory (see Table 6). The mean values of the WMNSQ for both test versions fell around 1.0. The respective *t*-values indicated no substantial misfit (|*t*| > 6). Overall, there was no indication of substantial item over- or underfit. The median correlations between the item scores and the total-rest scores were about .2 and .3 and, thus, did not indicated substantial differences between the test versions. The item-total correlations where rather low and did not show the desired item fit. According to the reported point-biserial correlations in Table 5, the items did not seem to measure the abilities of the test takers as well as they were expected to do. The item characteristic curves showed an acceptable fit of the items.

4.2.4 Test targeting and reliability

Test targeting was investigated in order to evaluate the measurement precision of the estimated ability scores and to judge the appropriateness of the tests for the specific target population. In the analyses, the mean of ability was constrained to be zero. Table 7 summarizes the estimated variances and the reliabilities of the different test versions. The easier test version 2 showed a slightly higher variance (*Var* = 0.71) as compared to test version 1 (*Var* = 0.53). The estimated reliabilities of test version 2 (EAP/PV reliability = .64, WLE reliability = .55) were also higher as compared to test version 1 (EAP/PV reliability = .54, WLE reliability = .41). Overall, neither test version showed an acceptable reliability.

Table 7. Reliabilities of the Different Test Versions

Test version	Variance	EAP/PV Rel.	WLE Rel.
Version 1	0.53	.54	.41
Version 2	0.71	.64	.55

4.3 Parameter Estimates for Concurrently Scaled Tests

4.3.1 Item parameters

Because both test versions administered in special schools presented most items at roughly the same position, they were concurrently scaled to estimate linked item parameters that can be compared across test versions in special schools. The respective item parameters are summarized in Table 8. The estimated item difficulties ranged from -0.458 (item mag5r201_sc4g9_c) to 2.939 (item mag5q301_sc4g9_c) with a mean of 0.916. The standard errors (SE) of the estimated parameters were rather large with a mean of 0.095 and a range of [0.070, 0.139]. Thus, the reported item parameters had a somewhat limited precision.

Table 8. Item Parameters for Combined Scaling of Both Test Versions in Special Schools

	Item	Pos.		*N*	Percentage correct	ξ	SE_ξ	WMNSQ	*t*	r_{it}	Discr.
1	mag5d041_sc4g9_c	1	1	1003	29.91	0.936	0.076	1.00	-0.1	0.38	0.759
2[e]	mag5q291_sc4g9_c	2	2	796	37.19	0.611	0.081	0.94	-2.3	0.52	1.625
3	mag5v271_sc4g9_c	3[e]	3	990	22.53	1.349	0.083	1.05	1.2	0.25	0.427
4	mag5r171_sc4g9_c	4	4	1003	35.00	0.691	0.073	0.98	-0.7	0.43	0.975
5	mag9r111_c	5	5[e]	963	42.99	0.305	0.072	1.08	3.6	0.27	0.250
6	mag5q301_sc4g9_c	6	6	944	6.36	2.939	0.139	0.97	-0.2	0.31	1.086
7	mag9d151_c	7	7	1023	26.98	1.105	0.077	0.98	-0.6	0.40	0.912
8[e]	mag9r051_c	8[e]									
9	mag9v011_c	9	9	1009	23.09	1.321	0.081	1.01	0.2	0.33	0.670
10	mag9v012_c	10	10	992	14.72	1.914	0.096	1.02	0.3	0.27	0.505
11	mag5q140_sc4g9_c	11	11	891	21.21	1.457	0.089	0.94	-1.2	0.47	1.411
12[e]	mag9d241_c	12[e]									
13	mag9r191_c	13	13	976	33.20	0.787	0.075	1.03	0.9	0.34	0.688
14	mag5r101_sc4g9_c	14	14[e]	1046	40.06	0.447	0.070	1.06	2.6	0.31	0.376
15	mag9q181_c	15	15	1050	39.24	0.488	0.070	0.99	-0.5	0.42	0.906
16	mag9q221_c	16		507	29.78	0.912	0.105	1.01	0.3	0.38	0.721
17	mag9d260_c	17		418	46.65	0.165	0.107	0.96	-1.3	0.47	1.148
18[e]	mag9q081_c	18[e]									
19	mag5v321_sc4g9_c	19	19	856	14.84	1.925	0.103	0.98	-0.3	0.35	0.888
20	mag5v091_sc4g9_c	20		508	15.75	1.812	0.129	1.00	0.1	0.31	0.600
21[e]	mag5q221_sc4g9_c		8	419	38.66	0.559	0.110	0.99	-0.4	0.44	0.965
22	mag5r201_sc4g9_c		12	519	60.69	-0.458	0.099	0.94	-1.9	0.47	1.262
23	mag5q131_sc4g9_c		16	423	50.35	0.060	0.107	0.92	-2.7	0.54	1.756
24	mag5d02s_sc4g9_c		17	409	60.88	-0.438	0.111	0.96	-1.2	0.45	1.151
25	mag5v024_sc4g9_c		18	363	28.65	1.104	0.126	0.96	-0.8	0.46	1.309
26	mag5r191_sc4g9_c		20	468	27.99	1.080	0.112	1.01	0.2	0.34	0.634

Note. Pos. = Item position in standard, easy, and out-of-level tests, *N* = Number of valid responses for item, ξ = Item difficulty, SE_{ξ} = Standard error of item difficulty, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, r_{it} = Item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model (2PL).
[e] Excluded from the analyses due to unsatisfactory item fit.

4.3.2 Item fit

For the concurrently scaled test versions in special schools (see Table 8) no item exhibited a noteworthy WMNSQ exceeding 1.15. The values of the WMNSQ fell between 0.92 and 1.08 (*M* = 0.99, *SD* = 0.04). All *t*-values indicated good fit (*Max* = 3.60). Although one item exhibited item-rest correlations less than .10, most items had adequate item-total correlation with a mean of .39.

4.3.3 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the tests for the specific target population.

In Figure 7, item difficulties of the mathematics items and the ability of the test takers are plotted on the same scale for the concurrently scaled test versions. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of item difficulties. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 0.473, indicating that the test did not differentiate well between students. The item difficulties ranged from -0.451 (item mag5r201_sc4g9_c) to 2.916 (item mag5q301_sc4g9_c). Thus, the items covered a smaller range than intended (see Figure 7). A larger number of easy items would have been desirable. The reliability of the test (EAP/PV reliability = 0.48, WLE reliability = 0.43) was unsatisfactory.

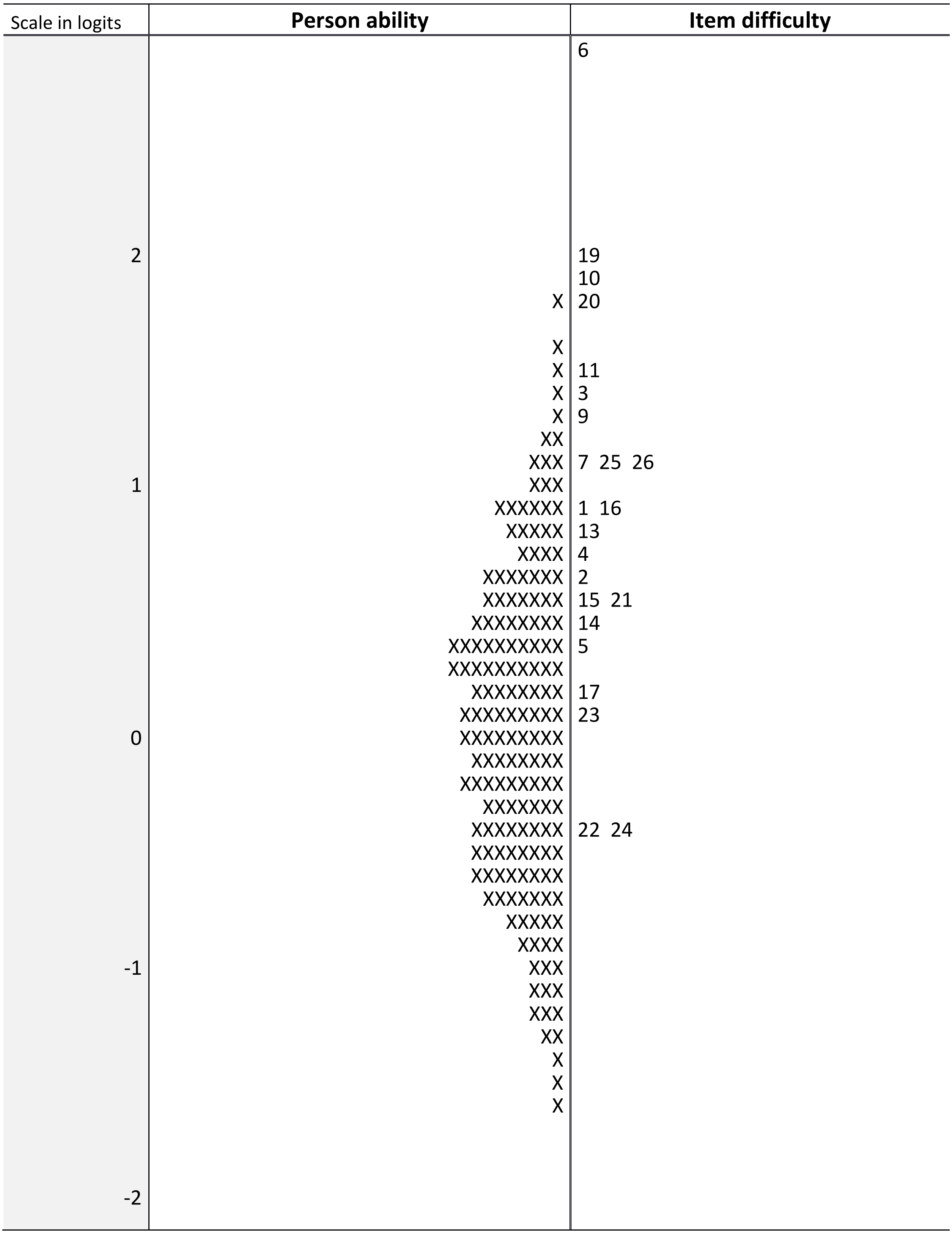

Figure 7. Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 6.2 cases. The difficulty of the items is depicted on the right-hand side of the graph. Each number represents one item (see Table 8).

Note. Items 8, 12 and 18 were excluded from the analyses due to unsatisfactory item fit.

4.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1980) model is that all item-discrimination parameters are equal. In order to test this assumption, a two-parametric model (2PL; Birnbaum, 1968) was fitted to the data. The estimated discrimination parameters for the concurrently scaled test versions are depicted in Table 8. They ranged from 0.250 (item mag9r111_c) to 1.756 (item mag5q131_sc4g9_c). The AIC suggested a slightly better model fit of the 2PL model (AIC = 19,584.43, number of parameters = 60) as compared to the 1PL Rasch model (AIC = 19,685.61, number of parameters = 37), whereas the BIC favored the Rasch model (BIC = 19,870.25) over the 2PL model (BIC = 19,883.85). The Rasch model matches the theoretical conceptions underlying the test construction more adequately (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the Rasch model (also known as one parameter logistic model, 1PL) was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework. Note that these calculations were performed in mdltm (see Davier, 2005). Therefore, slightly different results as compared to the estimations using TAM might have been observed.

4.3.5 Unidimensionality

The dimensionality of the tests was investigated by specifying a four-dimensional model based on the four different content areas. Each item was assigned to one content area (between-item-multidimensionality).

To estimate this multidimensional model, the Quasi Monte Carlo method implemented in TAM was used. The number of nodes used in TAM was set to 15,000. The standard deviations and correlations of the four dimensions are shown in Table 9. Two of the four dimensions exhibited a relatively good variance. The dimension “space and shape” had the smallest variance with 0.320. The difficulties of the six items of this dimension varied from -0.458 (item mag5r201_sc4g9_c) to 1.080 (item mag5r191_sc4g9_c), so the difficulties covered a relatively wide range and could not explain the small variance. On the other hand the difficulties of dimension “change and relationship” cover a very small range from 1.104 (item mag5v024_sc4g9_c) to 1.925 (item mag5v321_sc4g9_c) and therefore, could be a reason for the small variance of this dimension. As expected, the correlations between the four dimensions were very high, varying between 0.837 and 0.930.

Model fit between the unidimensional model and the four-dimensional model is compared in table 10. The AIC favored the four-dimensional model, whereas the BIC favored the unidimensional model. There were very few items per dimension, leading to low reliabilities and high standard errors of the WLE scores. Regarding the very high correlations between the four dimensions, it is reasonable to treat mathematical competence as a unidimensional construct.

Table 9. Results of Four-Dimensional Scaling

	Quantity	Space and shape	Change and relation-ships	Data and chance
Quantity (7 items)	1.123			
Space and shape (6 items)	0.891	0.320		
Change and relationships (6 items)	0.908	0.837	0.405	
Data and chance (4 items)	0.930	0.887	0.871	0.744

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

Table 10. Comparison of the Unidimensional and the Four-Dimensional Model

Model	Deviance	Number of parameters	AIC	BIC
Unidimensional	19,625.20	24	19,673.20	19,792.97
Four-dimensional	19,569.54	33	19,635.54	19,800.22

Note. Contrary to the calculations for the 1PL and 2PL models (see 4.3.3), the results in this table were achieved by using TAM in R.

4.4 Differential Item Functioning

Differential item functioning (DIF) was used to evaluate test fairness with regard to test order effects (general order vs. reverse order) and test version effects (common items of version 1 and version 2) of the different tests administered in special schools. DIF was also investigated for several subgroups (i.e., measurement invariance) for the variables gender, migration background and the number of books at home (see Pohl & Carstensen, 2012, for a description of these variables). Additionally, DIF was investigated for a subsample of the main field compared to the students from special schools.

The differences between the estimated item difficulties in the various groups are summarized in Tables 11, 12, and 13. For example, the column “Male vs. Female” reports the differences in item difficulties between the genders; a positive value would indicate that the item was more difficult for male students, whereas a negative value would point to a lower difficulty for male students. In contrast, the main effect is to be interpreted on a group level. As such, a positive value indicates that female students, on average, had a higher ability as compared to male students; whereas a negative value would suggest a lower ability, on average, for female students as compared to male students. Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF to those that only estimate main effects (see Table 14).

4.4.1 Test order effects

In special schools, test version 1 and test version 2 were administered in original and in reversed order. DIF was evaluated for both test versions. The respective differences in item difficulties are summarized in Table 11.

In total, 554 persons received test version 1, whereof 270 persons (48.7 %) received the test in original order and 284 persons (51.3 %) received the reversed test. On average, test takers, who received the reversed test version 1, exhibited a slightly higher mathematical competence than test takers, who received the original order (main effect = 0.014 logits, Cohen's *d* = 0.019). There were four items for which the rotation DIF exceeded |0.4| logits (mag5d041_sc4g9_c, mag5q140_sc4g9_c, mag5r101_sc4g9_c, mag5v321_sc4g9_c). These differences were considered not to be severe.

Table 11. Differential Item Functioning for test version 1 and test version 2 versus reversed test versions

	Item	Test version 1		Item	Test version 2
		original vs. reverse			original vs. reverse
1	mag5d041_sc4g9_c	0.496	1	mag5d041_sc4g9_c	0.328
2	mag5q291_sc4g9_c	-0.220	2	mag5q291_sc4g9_c	-0.330
3[e]	mag5v271_sc4g9_c		3	mag5v271_sc4g9_c	-0.030
4	mag5r171_sc4g9_c	0.110	4	mag5r171_sc4g9_c	-0.126
5	mag9r111_c	-0.156	5[e]	mag9r111_c	
6	mag5q301_sc4g9_c	0.064	6	mag5q301_sc4g9_c	0.302
7	mag9d151_c	-0.106	7	mag9d151_c	0.194
8[e]	mag9r051_c		8	mag5q221_sc4g9_c	-0.038
9	mag9v011_c	-0.034	9	mag9v011_c	-0.182
10	mag9v012_c	0.200	10	mag9v012_c	0.244
11	mag5q140_sc4g9_c	0.488	11	mag5q140_sc4g9_c	-0.080
12[e]	mag9d241_c		12	mag5r201_sc4g9_c	0.246
13	mag9r191_c	-0.040	13	mag9r191_c	0.328
14	mag5r101_sc4g9_c	-0.434	14[e]	mag5r101_sc4g9_c	
15	mag9q181_c	-0.320	15	mag9q181_c	-0.088
16	mag9q221_c	0.012	16	mag5q131_sc4g9_c	0.038
17	mag9d260_c	0.084	17	mag5d02s_sc4g9_c	-0.416
18[e]	mag9q081_c		18	mag5v024_sc4g9_c	-0.138
19	mag5v321_sc4g9_c	0.536	19	mag5v321_sc4g9_c	0.378
20	mag5v091_sc4g9_c	-0.056	20	mag5r191_sc4g9_c	-0.424
	Main effects:			*Main effects*:	
	DIF model	0.018		DIF model	0.008
	Main effect model	0.014		Main effect model	0.010

There were 550 persons, who received test version 2, whereof 297 (54.0 %) persons received test version 2 in original order and 253 persons (46.0 %) received the reversed test version 2. Again, test takers, who received the reversed test version, exhibited a slightly higher mathematical competence than those, who received the original order (main effect = 0.010

logits, Cohen's *d* = 0.012). Two items exceeded |0.4| logits (mag5d02s_sc4g9_c, mag5r191_sc4g9_c) but again, these differences were considered not to be severe.

Taken together, these analyses indicate that test order effects hardly affected group comparisons and, thus, the two groups could be scaled together.

4.4.2 Test version effects

Test version 1 and test version 2 administered a subsample of items at the same item position. Therefore, these items might be used as common items (cf. Fischer et al., 2016) to link the test versions and estimate a common mathematical competence score. However, to do so these common items must not exhibit substantial DIF; otherwise, the estimated mathematical competence scores might be distorted. Therefore, DIF was evaluated for the common items included in the concurrently scaled test versions (Table 8). The respective differences in item difficulties (or location parameters) are summarized in Table 12.

554 test takers received test version 1 (51.0 %) and 532 test takers received test version 2 (49.0 %). Test version 2 was slightly easier for the respondents than test version 1 (main effect = 0.108 logits, Cohen's *d* = 0.132). One item exceeded |0.4| logits (mag5q301_sc4g9_c) but this was considered not to be severe. These results do not indicate substantial DIF effects that might have distorted estimates of students' mathematical competences based on their concurrently scaled responses.

Table 12. Differential Item Functioning for Test version 1 versus Test version 2

Item	Test version 1 vs. Test version 2
mag5d041_sc4g9_c	0.002
mag5q291_sc4g9_c	-0.216
mag5r171_sc4g9_c	-0.164
mag5q301_sc4g9_c	0.452
mag9d151_c	0.374
mag9v011_c	-0.120
mag9v012_c	0.274
mag5q140_sc4g9_c	-0.170
mag9r191_c	-0.116
mag9q181_c	0.062
mag5v321_sc4g9_c	-0.040
Main effects:	
DIF model	0.108
Main effect model	0.108

4.4.3 Additional DIF-Variables

For the concurrently scaled test versions, DIF was examined for the variables gender, migration background, and the number of books.

Overall, 595 (54.9 %) of the test takers were male and 488 (45.1 %) were female. On average, male students exhibited a higher mathematical competence than female students (main effect = 0.438 logits, Cohen's *d* = 0.639). DIF exceeded |0.6| logits for item mag5q291_sc4g9_c.

There were 648 (59.7 %) participants without migration background, 297 (27.3 %) participants with migration background and 141 (13.0 %) students that gave no valid answer. On average, participants without migration background performed better in the mathematics test than those with migration background (main effect = 0.302 logits, Cohen's *d* = 0.431). DIF slightly exceeded |0.4| logits for the items mag5q301_sc4g9_c and mag9d260_c. These DIFs were considered not to be severe. For the item mag5q221_sc4g9_c DIF exceeded |1| logit.

Table 13. Differential Item Functioning of additional variables for the concurrently scaled test versions

	Item	Gender	Migration status	Books
		male vs. female	without vs. with	<=100 vs. >100
1	mag5d041_sc4g9_c	0.174	0.092	-0.052
2	mag5q291_sc4g9_c	-0.846	-0.270	0.256
3	mag5v271_sc4g9_c	0.296	0.284	-0.290
4	mag5r171_sc4g9_c	0.214	0.232	-0.064
5	mag9r111_c	0.294	0.050	0.258
6	mag5q301_sc4g9_c	-0.052	-0.458	0.230
7	mag9d151_c	0.112	-0.066	-0.152
8	mag9v011_c	-0.172	0.058	-0.204
9	mag9v012_c	-0.102	0.058	0.054
10	mag5q140_sc4g9_c	0.284	-0.136	0.112
11	mag9r191_c	0.022	0.178	-0.400
12	mag5r101_sc4g9_c	0.146	0.140	0.082
13	mag9q181_c	-0.300	0.162	-0.116
14	mag9q221_c	-0.088	-0.108	0.112
15	mag9d260_c	-0.062	-0.438	0.214
16	mag5v321_sc4g9_c	-0.090	0.268	-0.174
17	mag5v091_sc4g9_c	0.122	-0.152	0.234
18	mag5q221_sc4g9_c	-0.162	-1.028	0.162
19	mag5r201_sc4g9_c	0.070	0.394	-0.292
20	mag5q131_sc4g9_c	-0.288	-0.088	0.318
21	mag5d02s_sc4g9_c	0.222	-0.370	0.418
22	mag5v024_sc4g9_c	-0.060	0.024	-0.216
23	mag5r191_sc4g9_c	0.030	-0.182	0.254
	Main effects:			
	DIF model	-0.388	-0.314	0.160
	Main effect model	-0.438	-0.302	0.142

The number of books at home was used as a proxy for socioeconomic status. There were 836 (77.0 %) test takers with 0 to 100 books at home, 197 (18.1 %) test takers with more than 100 books at home, and 53 (4.9 %) test takers that did not provide this information. Participants with 100 or fewer books at home performed worse than participants with more than 100 books (main effect = 0.142 logits, Cohen's *d* = 0.226). DIF exceeded |0.4| logits only for item mag5d02s_sc4g9_c. Since this difference was very small, this DIF was considered not to be severe.

4.4.4 Sample effects

To test the feasibility of including students with special educational needs in the main field, a subsample from the main field including all children from secondary schools (*N* = 3519, 43.9 % female students, 36.0 % students with migration background) was chosen. The test from general schools and both test versions from special schools administered a subsample of six items at the same item position. These items might be used as common items (cf. Fischer et al., 2016) to link the samples and estimate a common mathematical competence score. To do so, these common items must not exhibit substantial DIF; otherwise, the estimated mathematical competence scores might be distorted. Therefore, DIF was evaluated for these common items. The respective differences in item difficulties (or location parameters) are summarized in Table 14.

The common items were substantially easier for the students from general schools than for students from special schools (main effect = 1.130 logits, Cohen's *d* = 1.307). Two items (mag9r111_c and mag9q181_c) exceeded substantial DIF effects of |0.6| logits.

Table 14. Differential Item Functioning for Test version 1 versus Test version 2

Item	**General schools vs. Special Schools**
mag9r111_c	0.864
mag9d151_c	-0.272
mag9v011_c	-0.210
mag9v012_c	-0.040
mag9r191_c	0.200
mag9q181_c	-0.634
Main effects:	
DIF model	-1.114
Main effect model	-1.130

Overall, test fairness could be confirmed for all tested subgroups. However, the group difference between general students and special needs students was to be expected and shows that it is important to consider special needs students and general students separately in educational studies. In Table 15, we compared the models that only included main effects to models that additionally estimated DIF effects. Overall, Akaike's (1974) information criterion (AIC) favored the main effect model for all DIF variables over the more complex DIF models, except for the variable sample. The Bayesian information criterion (BIC, Schwarz, 1978) takes the number of estimated parameters more strongly into account and, thus, prevents an overparameterization of models. Using BIC, the more parsimonious models including only the main effects were again preferred for all variables, except for the variable sample.

Table 15. Comparisons of Models with and without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Test version 1[a]	DIF	8,396.86	34	8,464.68	8,611.46
	Main effects	8,421.13	18	8,457.13	8,534.84
Test version 2[a]	DIF	9,241.03	38	9,317.03	9,479.54
	Main effects	9,263.03	20	9,303.03	9,388.56
Version 1 vs. Version 2	DIF	11,074.18	24	11,122.18	11,241.95
	Main effects	11,090.70	13	11,116.70	11,181.57
Gender	DIF	19,492.48	48	19,588.48	19,827.88
	Main effects	19,546.52	25	19,596.52	19,721.21
Migration status	DIF	17,093.25	48	17,189.25	17,422.11
	Main effects	17,130.07	25	17,180.07	17,301.35
Books	DIF	18,792.49	48	18,888.49	19,125.62
	Main effects	18,814.79	25	18,864.79	18,988.29
Sample[b]	DIF	33,553.71	14	33,581.71	33,671.69
	Main effects	33,755.83	8	33,771.83	33,823.25

Note. [a] Original versus reversed text order, [b] General versus special schools.

5. Discussion

The presented analyses summarized information from a feasibility study to evaluate the possibility of including students attending special schools in educational large-scale assessments such as the NEPS. The study included two different versions of a mathematical competence test for students in Grade 9 to examine how to best accommodate the special needs of these students. The results highlighted several challenges of administering standardized achievement tests in special schools.

The mathematical competence tests administered in special schools exhibited limited variances and reliabilities. The analyses of the distractors showed rather low point-biserial correlations with a mean value of almost 0 for both test versions (M = -.06 and -.08). Moreover, the difference in point-biserial correlations for distractors and correct responses was rather small. Other mathematical competence tests (e.g., SC4, grade 9; Duchhardt & Gerdes, 2013) showed larger mean differences in correlations between distractors and correct responses. As all items have been developed for students from general schools and have already shown good item fit in previous studies, this seems to be a specific issue for students with special educational needs.

Comparisons between students from different school types using the administered mathematical competence tests cannot be recommended. Substantial DIF for two of the six common items suggested that these items functioned rather differently for students from special schools and students from general schools (see 4.4.4). Additionally, the majority of the common items were too difficult for students with special educational needs. Aside from the common items, the easier test version 2 only consisted of items constructed for 5th grade students and was still too difficult. Due to this fact, it will be challenging to find enough common items to link the samples. This makes the analysis of schooling effects across different school types rather infeasible.

In conclusion, it seems that items developed for the implementation in mathematical competence tests in general schools (whether they are easy items or have been constructed for significantly younger students) cannot be used to accurately measure the mathematical competence of students in special schools. Due to their versatile learning difficulties, students from special schools cannot simply be compared to younger students from general schools. This study suggests the necessity to develop items especially for students with special educational needs. Therefore, the inclusion of students with special educational needs in the main field is still challenging.

6. Data in the Scientific Use File

6.1 Naming conventions

The data in the SUF contains 26 items of which 21 items were included in the reported analyses. All items were scored dichotomously with 0 indicating an incorrect response and 1 indicating a correct response. Items that were already administered in other grades kept their original names and a suffix was added in front of the '_c' to specify the current test administration ('sc4g9' referring to Starting Cohort 4, Grade 9). For further details on the naming conventions of the variables see Fuß and colleagues (2019).

6.2 Mathematical competence scores in special schools

In the SUF, mathematical competence scores in the form of WLEs are not provided because it is not recommended to compare students' competences using the administered test. Overall, both test versions did not perform well and did not show the necessary fit for the calculation of WLEs.

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER Conquest: Generalised Item Response Modelling Software* [Computer software]. Version 4. Camberwell, Victoria: Australian Council for Educational Research.

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

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Davier, M. von, (2005). A general diagnostic model applied to language testing data (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

Duchhardt, C. & Gerdes, A. (2012): NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 3 in Fifth Grade (NEPS Working Paper No. 19). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Duchhardt, C. & Gerdes, A. (2013): NEPS Technical Report for Mathematics – Scaling results of Starting Cohort 4 in ninth grade (NEPS Working Paper No. 22). Bamberg: University of Bamberg, National Educational Panel Study.

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2019). *Competence Data in NEPS: Overview of Measures and Variable Naming Conventions* (Starting Cohorts 1 to 6). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Gnambs, T., & Freund, M. (2019). *NEPS Technical Report for Attention: Administration of the Frankfurt Attention Inventory (FAIR) in Starting Cohort 4 (Grade 9) for Students with Special Educational Needs* (NEPS Survey Paper No. 57). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study

Haberkorn, K., Pohl, S., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 4 in Ninth Grade* (NEPS Working Papers No. 16). Bamberg, Germany: Otto-Friedrich-University, National Educational Panel Study.

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests* (NEPS Working Paper No. 14). Bamberg, Germany: Otto-Friedrich-University, Nation Educational Panel Study.

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online, 5*, 189-216.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

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

Warm, T. A., (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450. https://doi.org/10.1007/BF02294627

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft, 14*, 67-86. https://doi.org/10.1007/s11618-011-0182-7

Appendices

Appendix A: Allocation of items to response formats and content areas

	Item	Response format	Content area
1	mag5d041_sc4g9_c	MC	Data and chance
2	mag5q291_sc4g9_c	SCR	Quantity
3	mag5v271_sc4g9_c	MC	Change and relationships
4	mag5r171_sc4g9_c	MC	Space and shape
5	mag9r111_c	MC	Space and shape
6	mag5q301_sc4g9_c	SCR	Quantity
7	mag9d151_c	MC	Data and chance
8[e]	mag9r051_c	MC	Space and shape
9	mag9v011_c	MC	Change and relationships
10	mag9v012_c	MC	Change and relationships
11	mag5q140_sc4g9_c	SCR	Quantity
12[e]	mag9d241_c	MC	Data and chance
13	mag9r191_c	MC	Space and shape
14	mag5r101_sc4g9_c	MC	Space and shape
15	mag9q181_c	MC	Quantity
16	mag9q221_c	MC	Quantity
17	mag9d260_c	SCR	Data and chance
18[e]	mag9q081_c	MC	Quantity
19	mag5v321_sc4g9_c	SCR	Change and relationships
20	mag5v091_sc4g9_c	MC	Change and relationships
21	mag5q221_sc4g9_c	SCR	Quantity
22	mag5r201_sc4g9_c	MC	Space and shape
23	mag5q131_sc4g9_c	SCR	Quantity
24	mag5d02s_sc4g9_c	SCR	Data and chance
25	mag5v024_sc4g9_c	SCR	Change and relationships
26	mag5r191_sc4g9_c	MC	Space and shape

Note. MC = Simple multiple-choice, SCR = Short constructed response,
[e] Excluded from the analyses due to unsatisfactory item fit.

Appendix B: Item parameters for different test versions

Table B.1. Item Parameters for Test Version 1

Pos.	Item	Percentage correct	ξ	$SE_ξ$	WMNSQ	*t*	r_{it}	Discr.
1	mag5d041_sc4g9_c	28.83	0.994	0.109	1.01	0.2	0.38	0.69
2	mag5q291_sc4g9_c	38.52	0.563	0.115	0.95	-1.3	0.53	1.49
3[e]	mag5v271_sc4g9_c							
4	mag5r171_sc4g9_c	35.50	0.667	0.103	0.99	-0.3	0.44	1.05
5	mag9r111_c	40.75	0.406	0.103	1.04	1.3	0.34	0.42
6	mag5q301_sc4g9_c	4.84	3.245	0.221	0.98	0.0	0.27	1.11
7	mag9d151_c	22.97	1.352	0.114	0.98	-0.3	0.40	0.99
8[e]	mag9r051_c							
9	mag9v011_c	23.15	1.322	0.114	1.01	0.2	0.35	0.63
10	mag9v012_c	12.42	2.122	0.114	1.00	0.1	0.28	0.72
11	mag5q140_sc4g9_c	21.59	1.430	0.124	0.94	-1.0	0.50	1.75
12[e]	mag9d241_c							
13	mag9r191_c	33.33	0.790	0.106	1.01	0.2	0.38	0.80
14	mag5r101_sc4g9_c	38.50	0.526	0.099	1.03	0.8	0.37	0.51
15	mag9q181_c	37.57	0.573	0.100	1.01	0.2	0.41	0.79
16	mag9q221_c	29.78	0.944	0.107	1.01	0.2	0.40	0.75
17	mag9d260_c	46.65	0.200	0.109	0.96	-1.2	0.49	1.27
18[e]	mag9q081_c							
19	mag5v321_sc4g9_c	14.25	1.967	0.145	0.98	-0.1	0.34	0.91
20	mag5v091_sc4g9_c	15.75	1.848	0.131	1.00	0.0	0.32	0.65

Note. Pos. = Item position in test, ξ = Item difficulty / location parameter, $SE_ξ$ = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, r_{it} = Item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model (2PL).

[e] Excluded from the analyses due to unsatisfactory item fit.

Table B.2. Item Parameters for Test Version 2

Pos.	Item	Percentage correct	ξ	*SE*ξ	WMNSQ	*t*	r_{it}	Discr.
1	mag5d041_sc4g9_c	31.00	0.910	0.110	1.01	0.1	0.40	0.67
2	mag5q291_sc4g9_c	35.89	0.692	0.118	0.92	-1.9	0.55	1.51
3	mag5v271_sc4g9_c	22.69	1.391	0.120	1.07	1.2	0.29	0.42
4	mag5r171_sc4g9_c	34.48	0.744	0.108	1.00	0.0	0.44	0.81
5[e]	mag9r111_c							
6	mag5q301_sc4g9_c	7.89	2.775	0.182	0.98	-0.1	0.33	0.98
7	mag9d151_c	31.09	0.909	0.109	1.01	0.3	0.41	0.72
8	mag5q221_sc4g9_c	38.66	0.552	0.114	1.00	0.0	0.43	0.80
9	mag9v011_c	23.03	1.367	0.120	1.03	0.5	0.36	0.64
10	mag9v012_c	17.04	1.785	0.132	1.07	0.9	0.25	0.33
11	mag5q140_sc4g9_c	20.82	1.534	0.131	0.94	-0.8	0.46	1.12
12	mag5r201_sc4g9_c	60.69	-0.500	0.103	0.96	-1.1	0.47	1.04
13	mag9r191_c	33.06	0.819	0.111	1.06	1.4	0.34	0.51
14[e]	mag5r101_sc4g9_c							
15	mag9q181_c	40.97	0.416	0.103	1.00	0.0	0.45	1.84
16	mag5q131_sc4g9_c	50.35	0.042	0.111	0.92	-2.3	0.55	1.54
17	mag5d02s_sc4g9_c	60.88	-0.474	0.116	0.98	-0.5	0.46	0.99
18	mag5v024_sc4g9_c	28.65	1.123	0.131	0.97	-0.5	0.47	1.12
19	mag5v321_sc4g9_c	15.46	1.949	0.149	0.96	-0.4	0.39	0.85
20	mag5r191_sc4g9_c	27.99	1.086	0.116	1.04	0.9	0.35	0.55

Note. Pos. = Item position in test, ξ = Item difficulty / location parameter, *SE*ξ = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, r_{it} = Item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model (2PL).

[e] Excluded from the analyses due to unsatisfactory item fit.

Appendix C: Conquest-Syntax for scaling the data in Grade 9 of special schools for Starting Cohort 4

```
datafile filename.dat;
format id 4-10 responses 12-34;
labels << labels.nam;

codes 0,1;

score (0,1)    (0,1)     !item (1-23);

model item + item*step;
set constraint=cases;
estimate;
show cases! estimate=wle, filetype=spss >> %name%_wle.sav;
show cases! estimate=latent >> %name%.plv;
show cases! estimate=mle >> %name%.mle;
show cases! estimate=eap >> %name%.eap;
show >> %name%.shw;
itanal >> %name%.itn;
plot icc;
```