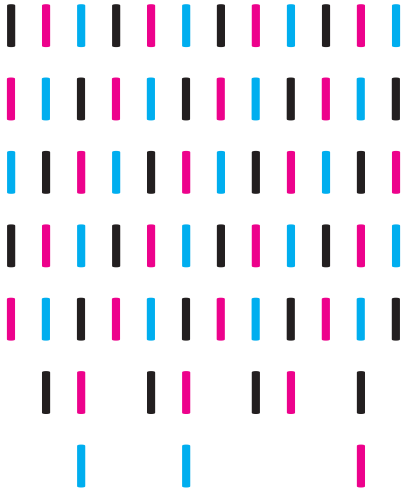

NEPS SURVEY PAPERS

Timo Gnambs

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

NEPS Survey Paper No. 63
Bamberg, January 2020

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

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).

The NEPS Survey Papers are edited by a review board consisting of the scientific management of LIfBi and 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 available at https://www.neps-data.de (see section “Publications“).

Editor-in-Chief: Corinna Kleinert, LIfBi/University of Bamberg/IAB Nuremberg

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

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

Timo Gnambs

Leibniz Institute for Educational Trajectories, Bamberg

E-mail address of lead author:

timo.gnambs@lifbi.de

Bibliographic data:

Gnambs, T. (2020). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 4 for Grade 9 in Special Schools* (NEPS Survey Paper No. 63). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP63:1.0

Acknowledgements:

This report is an extension to NEPS Working Paper 16 (Haberkorn, Pohl, Hardt, & Wiegand, 2012) that presents the scaling results for reading competence of Starting Cohort 4 for Grade 9 in general school students. Therefore, various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* to facilitate the understanding of the presented results.

NEPS Technical Report for Reading: Scaling Results of Starting Cohort 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 reading 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 five different test versions. Version 1 was identical to the test administered in the main sample to students from general schools, version 2 represented a shorter version of the test with fewer items, whereas in version 3 the most difficult items were replaced with items designed for a younger age cohort, thus, resulting in a considerably easier test as compared to version 1. Moreover, to examine potential item order effects, versions 4 and 5 presented the test versions 2 and 3 in reversed order. These five test versions were randomly distributed among a sample of N = 976 students (44% girls) from special schools. As a control group, a matched sample of N = 500 students (43% girls) attending lower secondary schools ("Hauptschule") was drawn from the main study. The responses of the two samples were scaled using the partial credit model. Item fit statistics, differential item functioning, and Rasch-homogeneity were evaluated to examine the quality of the tests. In particular, differential item functioning analyses between the five test versions were conducted to evaluate whether a common reading score can be estimated. These analyses showed that standard competence tests are too long for students in special schools; items at the end of the administered tests were finished by rather few students, resulting in large missing rates. Moreover, the tests exhibited somewhat limited variances and reliabilities, thus, allowing only rather crude analyses of interindividual differences between students with special educational needs. Nevertheless, a common reading score was estimated for test versions 1, 2, and 3, which allow cross-sectional analyses of students reading abilities. Importantly, there was substantial differential item functioning between special schools and lower secondary schools. Therefore, comparative analyses between the two school types using the administered reading competence test are not recommended. Overall, these results highlight substantial difficulties in assessing reading 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 R syntax for scaling the data.

Keywords

item response theory, special educational needs, reading 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, reading competence, mathematical 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. In Grade 9 of Starting Cohort 4 (ninth grade), a feasibility study was conducted to evaluate whether and if how students from special schools could be validly and meaningfully included in the NEPS. In this paper the results of these analyses are presented for a reading competence test administered to students with special educational needs attending special schools. First, the main concepts of the reading competence test and the test design are introduced. Then, the reading competence data of Starting Cohort 4 and the analyses performed to estimate competence scores and 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 Reading Competence

The framework and test development for the reading competence test are described by Weinert and colleagues (2011) and Gehrer, Zimmermann, Artelt, and Weinert (2013). In the following, specific aspects of the reading competence test will be pointed out that are necessary for understanding the scaling results presented in this paper.

In this study five different test versions were administered. These reading competence tests included either four or five texts and respective item sets referring to these texts. Each of these texts represented one text type or text function, namely, a) information, b) commenting or argumenting, c) literary, d) instruction, and e) advertising (see Gehrer et al., 2013, and Weinert et al., 2011, for the description of the framework). Furthermore, the tests assessed three cognitive requirements. These are a) finding information in the text, b) drawing text-related conclusions, and c) reflecting and assessing. The cognitive requirements do not depend on the text type, but each cognitive requirement is usually assessed within each text type (see Gehrer and Artelt, 2013, Gehrer et al., 2013, and Weinert et al., 2011, for a detailed description of the framework).

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

items had four response options. One response option represented a correct solution, whereas the other three were distractors (i.e., they were incorrect). In CMC items a number of subtasks with two response options were presented. MA items required the test taker to match a number of responses to a given set of statements. 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, among others, reading competence and general cognitive functioning (cf. Nusser & Messingschlager, 2018). For each participant, the reading test was administered as the first test. The study adopted an experimental design and administered five different versions of the reading competence test:

- Test version 1 (*standard test*) was identical to the test administered to the main sample in Grade 9 of Starting Cohort 4 (see Haberkorn, Pohl, Hardt, & Wiegand, 2012). The test included five texts including 33 items. Preliminary analyses identified excessive missing rates for the last text and severe misfit of items reg90130_c, reg90140_c, and reg9047s_c. Therefore, these items were excluded from the analyses, resulting in a test with four texts including 23 items.

- Test version 2 (*easy test*) was a shorter version of the test administered to the main sample in Grade 9 of Starting Cohort 4 (Haberkorn et al., 2012) that excluded the last text (text function: commenting or argumenting) as well as three difficult items. Again, item reg9047s_c was excluded from the analyses due to severe misfit, resulting in a test with four texts including 20 items.

- Test version 3 (*out-of-level test*) included two texts from the test administered to the main sample in Grade 9 of Starting Cohort 4 (Haberkorn et al., 2012) as well as three texts that were designed for a younger age cohort. Thus, the test was considerably easier as compared to the standard test version. Preliminary analyses identified excessive missing rates for the last text and severe misfit for item reg90710_c. Therefore, these items were removed, resulting in a test with four texts including 23 items.

- Test version 4 (*out-of-level test reversed*) was identical to test version 3. However, the five texts were presented in reversed order. Again, the last text and item reg90710_c were excluded from the analyses, resulting in a test with four texts including 23 items.

- Test version 5 (*easy test reversed*) was identical to test version 2, albeit presenting the four texts in reversed order. After excluding item reg9047s_c due to misfit, the test included 20 items.

The present analyses refer to these five test versions that included four texts with 20 or 23 items referring to these texts. The number of items for the different text types, cognitive requirements, and response formats are summarized in Tables 1, 2, and 3. The allocation of the items to the text types and cognitive requirements is given in Appendix A.

Table 1

Number of Items for the Different Text Types by Test Version

Text types	**Standard (special schools)**	**Standard (general schools)**	**Easy**	**Easy (reversed)**	**Out-of-level**	**Out-of-level (reversed)**
Information text	5	5	5	5	6	0
Instruction text	5	5	4	4	5	5
Advertising text	7	7	7	7	7	7
Commenting text	0	0	0	0	0	6
Literary text	6	6	4	4	5	5
Total number of items	23	23	20	20	23	23

Table 2

Number of Items for the Cognitive Requirements by Test Version

Cognitive requirements	**Standard (special schools)**	**Standard (general schools)**	**Easy**	**Easy (reversed)**	**Out-of-level**	**Out-of-level (reversed)**
Finding information	8	8	7	7	10	8
Drawing text-related conclusions	10	10	8	8	8	9
Reflecting and assessing	5	5	5	5	5	6
Total number of items	23	23	20	20	23	23

Table 3

Number of Items for the Different Response Formats by Test Version

Response format	Standard (special schools)	Standard (general schools)	Easy	Easy (reversed)	Out-of-level	Out-of-level (reversed)
Simple multiple choice items	20	20	17	17	21	19
Complex multiple choice items	2	2	2	2	1	3
Matching items	1	1	1	1	1	1
Total number of items	23	23	20	20	23	23

2.2 Samples

Overall, a total of 990[1] students from special schools received the reading competence tests. For 14 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 976 individuals (44% girls) from special schools. Moreover, a comparison group was selected by drawing a random sample of N = 500 students (43% girls) from general schools with lower secondary education ("Hauptschule") from the main sample in Grade 9 of Starting Cohort 4. These students were matched on selected socio-demographic information (sex, age, migration background, number of books at home) and basic reasoning abilities. Thus, by design the comparison sample was rather similar on these background variables to the students from special schools, albeit attending a different type of school. A summary of basic descriptive statistics for these samples is given in Table 4. The five test versions (see section 2.1) were randomly distributed among students from special schools; all students in the comparison sample from lower secondary schools received the standard test (i.e., test version 1). A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

[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.

Table 4

Sample Description by Test Version

	Standard (special schools)	Standard (general schools)	Easy	Easy (reversed)	Out-of-level	Out-of-level (reversed)
Sample size (*N*)	199	500	198	185	202	192
Median age	16	15	16	16	16	16
Girls (%)	40%	43%	48%	44%	49%	39%
Migration background (%)	22%	23%	21%	18%	18%	17%
100+ books at home (%)	22%	20%	22%	16%	20%	15%

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 CMC and MA items that are not determined.

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. All missing responses after the last valid response given were coded as not-reached. As CMC and MA items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A CMC or MA 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 partial credit model (PCM; Masters, 1982). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

CMC and MA items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC or MA item, indicating the number of correctly responded subtasks within that item. Categories of polytomous variables with less than $N = 20$ responses were collapsed in the analyses in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items. For four of the seven CMC and MA items categories were collapsed.

Reading competences were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989). To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC 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). Person parameter estimation in the NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in section 6.

3.3 Checking the Quality of the Tests

The reading 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 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 fit of the dichotomous MC and polytomous CMC and MA items to the partial credit model (PCM; Masters, 1982) 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 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 reading competence test should measure the same construct for all students. If some items favored certain subgroups (e.g., they were easier for students from general schools than for students from special schools), 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

different test versions administered in special schools to determine whether a common reading score might be derived. Moreover, test fairness was also evaluated for students from special schools and students from lower secondary schools to evaluate whether group comparisons across school types might be conducted. 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. Based on experiences with preliminary data, we considered absolute standardized differences in estimated difficulties between the subgroups that were greater than 0.5 as strong DIF, differences between 0.25 and 0.50 as small but not severe, and differences smaller than 0.25 as negligible DIF. Minimum hypothesis tests (see Fischer, Rohm, Gnambs, & Carstensen, 2016) were used to statistically test whether the observed standardized differences were significantly larger than 0.25 and, thus, was at least small in size. 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 reading competence test was scaled using the PCM (Masters, 1982), which assumes Rasch-homogeneity. The PCM 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 generalized partial credit model (GPCM; Muraki, 1992) was also fitted to the data and compared to the PCM. The independence assumption of the residuals in the PCM was examined using Yen's (1984) Q_3. Because in case of locally independent items, the Q_3 statistic tends to be slightly negative, the corrected Q_3 (aQ_3) is reported that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) absolute values of aQ_3 falling below .20 indicate essential unidimensionality.

3.4 Software

The IRT models were estimated in TAM version 3.2-24 (Robitzsch, Kiefer, & Wu, 2019) in R version 3.6.1 (R Core Team, 2019) using the Gauss-Hermite quadrature method with 21 nodes.

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. Between 78% and 85% of the students in special schools did not have any invalid response at all, whereas nearly 95% of the students from general school exhibited no invalid response. More than one invalid response was observed for 8% to 10% of students in special schools and less than two percent of students in general schools. Thus, although the overall rate of missing responses was small, students with special educational needs produced more invalid responses as compared to students from lower secondary schools.

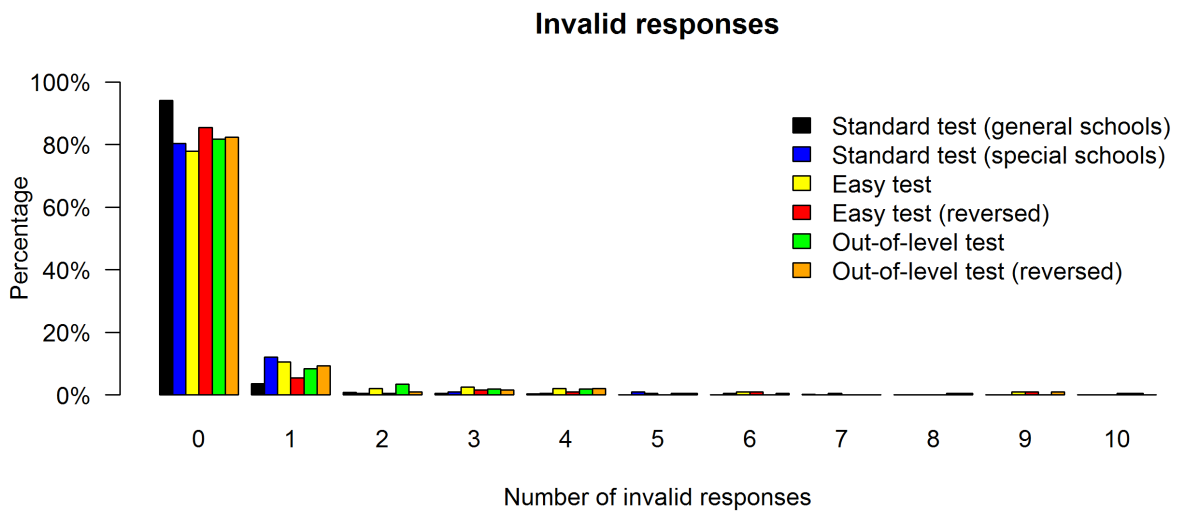

Figure 1. Number of invalid responses by test version

Missing responses also occurred when respondents omitted items. As illustrated in Figure 2, most respondents in special schools (65% to 71%) did not skip any item and about five to nine percent omitted more than two items. Again, students in general schools had fewer omitted responses. About 82% of respondents in general schools omitted no item and less than four percent had more than two omitted items.

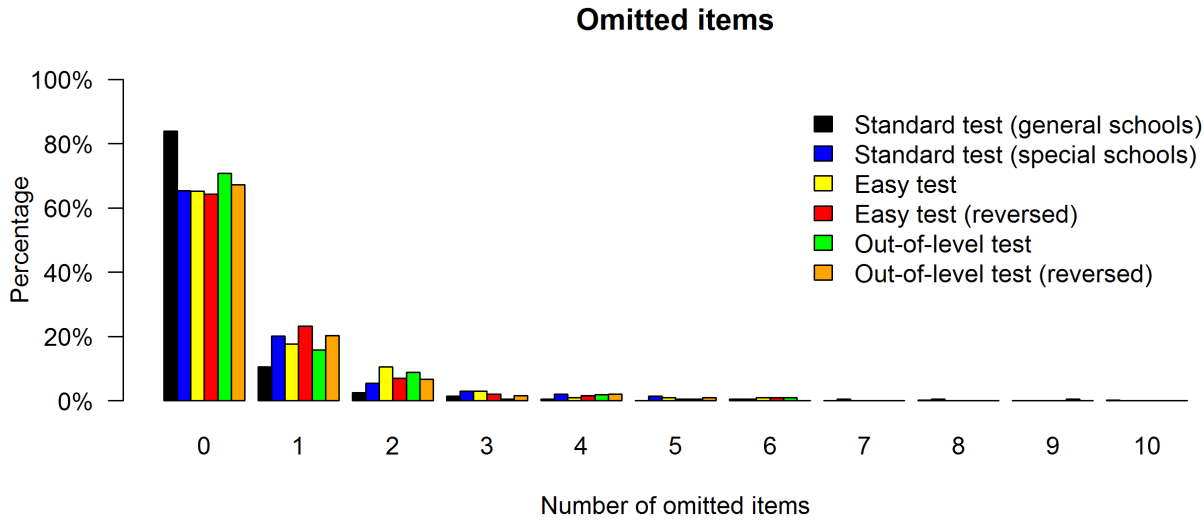

Figure 2. Number of omitted items by test version

Another source of missing responses was items that were not reached by the respondents; these are all missing responses after the last valid response. The number of not-reached items was rather high because many students from special schools were unable to finish the test within the allocated time limit. Therefore, only four texts were examined for all test versions (see section 2.1). Between 59% and 67% of students in special school finished all items referring to these four texts, whereas the respective percentage was 90% for students from general schools (Figure 3).

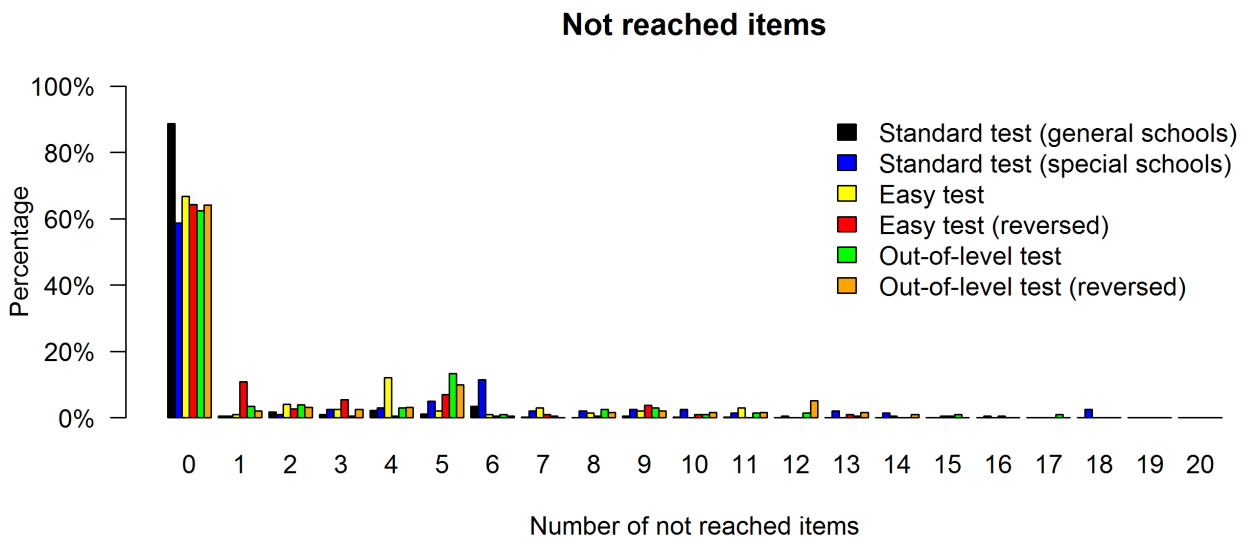

Figure 3. Number of not-reached items by test version

The aggregated polytomous variables were coded as not-determinable missing response when the subtasks of CMC or MA items contained different kinds of missing responses. Because not-determinable missing responses only occur in CMC and MA items, the maximum number of not-determinable missing responses was two to four (see Table 3). However, there were no substantial missing responses that were not determinable (Figure 4). The respective percentage fell between 2% and 3% in special schools and at 1% in general schools.

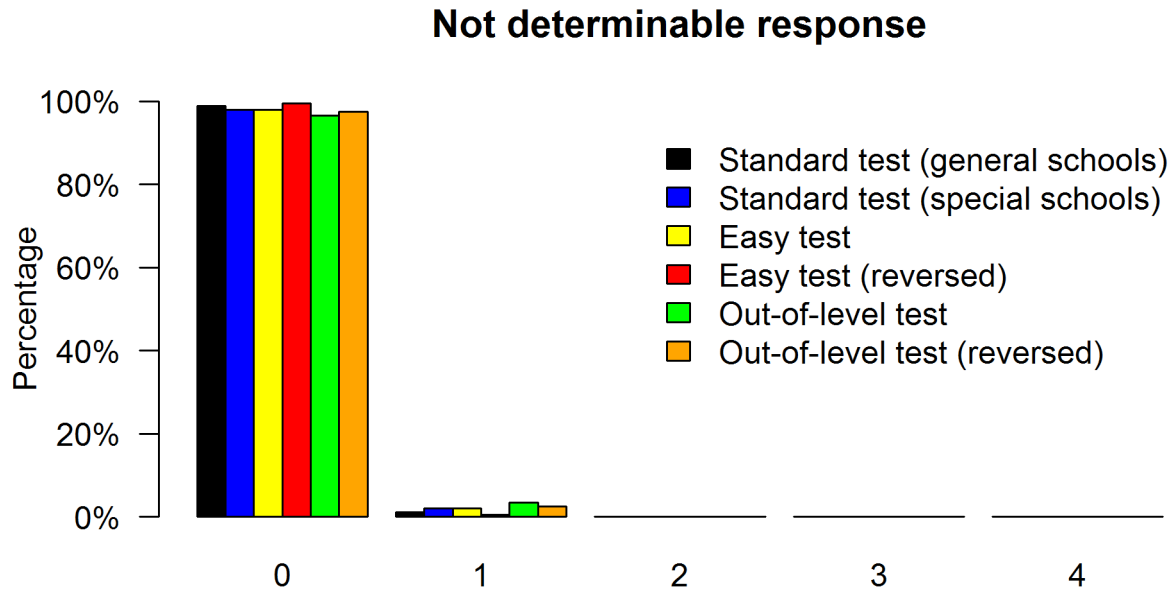

Figure 4. Number of not-reached items by test version

The total number of missing responses, aggregated over invalid, omitted, not-reached, and not-determinable missing responses per person, is illustrated in Figure 5. Students from special school had a rather large amount of missing values. About 29% to 37% of them had no missing response at all, whereas about 25% to 43% of these participants had five or more missing responses. Among students from general schools, about 70% had no missing response at all and only 9% had five or more missing responses.

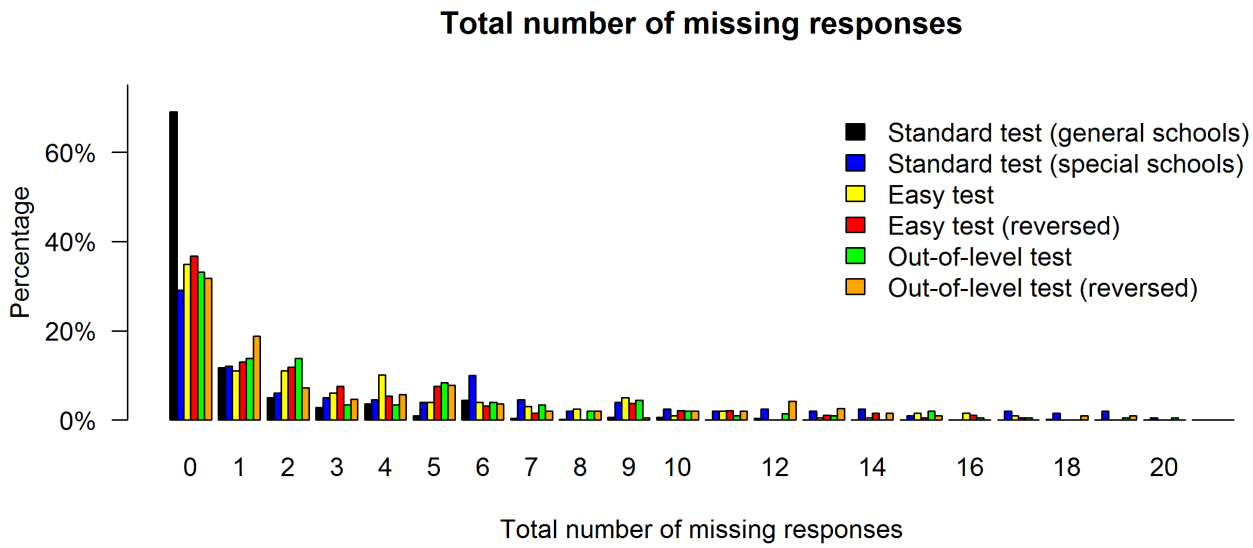

Figure 5. Total number of missing responses by test version

In sum, students in special schools had a large amount of omitted and not-reached items, whereas invalid and not-determinable missing responses were rare. This resulted in rather large overall missing rates for the different test versions. This is particularly notable because the test was already shorter as compared to the standard test administered in Grade 9 of Starting Cohort 4 (see Haberkorn et al., 2012) and included only four (instead of five) texts. In contrast, the comparison group of low-achieving students from lower secondary schools exhibited markedly lower missing rates. This was primarily a consequence of fewer not-reached items, but to a lesser degree also a result of fewer invalid and omitted responses. These results indicate that for students from special schools competence tests need to be substantially shorter for these students to be able to finish a test in the allocated time span or, alternatively, the available testing time needs to be increased.

4.1.2 Missing responses per item

Tables 5, 6, and 7 provide information on the occurrence of different kinds of missing responses per item for the different test versions. In special schools, the omission rates were rather low for most items; the median percentage of omitted responses across items fell between2.1% and 4.6%. However, polytomous CMC and MA items exhibited substantially larger omission rates around 10% and more. In contrast, students from general schools omitted polytomous items less frequently (2% to 5%). This indicates potential difficulties of students with special educational needs for more complex response formats. The percentage of invalid response showed few systematic differences between items. There was a slight tendency for more invalid responses for the first administered item, potentially, because the students in special school had to familiarize themselves with the response formats of the tests. With an item's progressing position in the tests, the number of students in special schools that did not reach an item (columns "NR" in Tables 5 to 7) rose to a considerable amount of 33% to 41% for the different test versions (see Figure 6). In contrast, for students in general school the respective number was 10%. Thus, for students with special educational needs the available testing time seemed to be too short.

Table 5

Percentage of Missing Values for Standard Test by School Type

	Standard (special schools)					*Standard (general schools)*				
Item	***N***	**NR**	**OM**	**NV**	**ND**	***N***	**NR**	**OM**	**NV**	**ND**
reg90110_c	176	0.0	2.0	9.6	0.0	490	0.0	0.8	1.2	0.0
reg90120_c	197	0.0	0.5	0.5	0.0	500	0.0	0.0	0.0	0.0
reg90150_c	190	0.0	1.5	3.0	0.0	494	0.0	1.0	0.2	0.0
reg9016s_c	165	0.0	11.1	4.5	1.5	471	0.0	3.8	1.0	1.0
reg9017s_c	177	0.0	11.1	0.0	0.0	475	0.0	5.0	0.0	0.0
reg90210_c	180	2.5	0.0	7.0	0.0	490	0.0	0.2	1.8	0.0
reg90220_c	181	2.5	0.5	6.0	0.0	492	0.0	0.8	0.8	0.0
reg90230_c	183	3.0	1.0	4.0	0.0	493	0.2	0.2	1.0	0.0
reg90240_c	180	3.0	2.5	4.0	0.0	492	0.2	1.0	0.4	0.0
reg90250_c	175	4.5	3.0	4.5	0.0	492	0.2	0.8	0.6	0.0
reg90310_c	173	6.5	2.0	4.5	0.0	490	0.4	0.8	0.8	0.0
reg90320_c	171	7.0	3.0	4.0	0.0	488	0.4	1.2	0.8	0.0
reg9033s_c	158	8.5	11.1	0.5	0.5	484	0.6	2.2	0.0	0.4
reg90340_c	160	11.1	4.5	4.0	0.0	492	0.6	0.6	0.4	0.0
reg90350_c	156	13.6	3.0	5.0	0.0	487	1.0	0.6	1.0	0.0
reg90360_c	159	15.6	2.5	2.0	0.0	490	1.0	1.0	0.0	0.0
reg90370_c	150	17.6	3.0	4.0	0.0	483	1.2	1.6	0.6	0.0
reg90410_c	131	29.2	2.0	3.0	0.0	473	4.6	0.6	0.2	0.0
reg90420_c	120	34.2	2.5	3.0	0.0	464	5.8	1.0	0.4	0.0
reg90430_c	113	37.2	3.0	3.0	0.0	444	7.4	3.6	0.2	0.0
reg90440_c	106	39.7	4.0	3.0	0.0	445	8.2	2.6	0.2	0.0
reg90450_c	106	40.7	2.5	3.5	0.0	441	9.2	1.8	0.8	0.0
reg90460_c	105	41.2	2.5	3.5	0.0	433	10.4	2.6	0.4	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.

Table 6

Percentage of Missing Values for Easy Test Versions

	Easy test					*Easy test (reversed)*				
Item	***N***	**NR**	**OM**	**NV**	**ND**	***N***	**NR**	**OM**	**NV**	**ND**
reg90110_c	175	0.0	1.0	10.6	0.0	140	16.2	1.6	6.5	0.0
reg90120_c	193	0.0	1.0	1.5	0.0	149	16.8	0.5	2.2	0.0
reg90150_c	184	0.0	3.5	3.5	0.0	30	22.2	2.7	4.9	0.0
reg9016s_c	155	0.0	18.2	2.0	1.5	120	24.9	8.1	1.6	0.5
reg9017s_c	173	0.0	12.1	0.5	0.0	116	35.9	1.6	0.0	0.0
reg90210_c	179	0.5	1.0	8.1	0.0	159	7.0	1.1	6.0	0.0
reg90220_c	177	1.0	1.0	8.6	0.0	155	7.6	0.5	8.1	0.0
reg90230_c	179	1.0	3.0	5.6	0.0	155	8.7	3.8	3.8	0.0
reg90240_c	179	1.0	1.5	7.1	0.0	153	9.2	3.2	4.9	0.0
reg90310_c	174	4.0	3.0	5.1	0.0	171	0.5	1.1	6.0	0.0
reg90320_c	175	4.0	2.5	5.1	0.0	171	1.1	0.5	6.0	0.0
reg9033s_c	163	5.1	10.1	1.0	0.5	150	1.1	17.8	0.0	0.0
reg90340_c	170	7.6	1.5	5.1	0.0	169	2.2	2.7	3.8	0.0
reg90350_c	166	10.6	2.0	3.5	0.0	169	2.2	2.7	3.8	0.0
reg90360_c	167	11.6	2.0	2.0	0.0	178	2.2	1.1	0.5	0.0
reg90370_c	159	13.6	2.5	3.5	0.0	163	3.2	1.1	7.6	0.0
reg90410_c	142	25.8	0.5	2.0	0.0	172	0.0	1.6	5.4	0.0
reg90420_c	136	28.3	0.5	2.5	0.0	170	0.0	3.2	4.9	0.0
reg90440_c	128	32.3	1.0	2.0	0.0	176	0.0	3.2	1.6	0.0
reg90450_c	126	33.3	0.5	2.5	0.0	178	0.0	1.1	2.7	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.

Table 7

Percentage of Missing Values for Out-of-Level Test Versions

	Out-of-level test					*Out-of-level test (reversed)*				
Item	***N***	**NR**	**OM**	**NV**	**ND**	***N***	**NR**	**OM**	**NV**	**ND**
reg90610_c	185	0.0	3.5	5.0	0.0					
reg90620_c	187	0.0	1.5	5.9	0.0					
reg90630_c	196	0.0	1.0	2.0	0.0					
reg90640_c	190	0.0	4.6	1.0	0.0					
reg90650_c	195	0.0	1.5	2.0	0.0					
reg90660_c	195	0.0	0.5	3.0	0.0					
reg90210_c	191	1.0	0.5	4.0	0.0	139	25.0	0.0	2.6	0.0
reg90220_c	189	1.0	1.0	4.5	0.0	135	28.1	0.0	1.6	0.0
reg90230_c	191	2.0	1.0	2.5	0.0	126	30.7	1.6	2.1	0.0
reg90240_c	188	2.0	1.5	3.5	0.0	118	33.9	2.6	2.1	0.0
reg90250_c	187	2.5	1.0	4.0	0.0	116	35.9	2.1	1.6	0.0
reg90310_c	188	4.0	1.5	1.5	0.0	170	7.8	1.0	2.6	0.0
reg90320_c	179	5.5	3.5	2.5	0.0	168	9.4	1.0	2.6	0.0
reg9033s_c	166	6.4	8.4	1.5	1.5	153	10.9	7.3	1.1	1.0
reg90340_c	167	9.4	4.0	4.0	0.0	156	13.0	3.1	1.0	0.0
reg90350_c	165	11.9	3.5	3.0	0.0	156	14.6	2.1	2.6	0.0
reg90360_c	172	12.4	2.0	0.5	0.0	156	14.6	3.1	2.1	0.0
reg90370_c	165	13.4	2.0	3.0	0.0	150	15.1	3.1	1.0	0.0
reg90720_c	144	26.7	1.0	1.0	0.0	184	0.0	2.1	2.1	0.0
reg90730_c	127	29.7	1.5	1.0	0.0	179	0.0	3.7	3.1	0.0
reg90740_c	127	30.2	1.5	0.5	0.0	181	0.0	2.1	2.6	0.0
reg90750_c	125	34.2	2.0	2.0	0.0	180	1.0	2.7	1.6	0.0
reg9076s_c	101	37.6	7.4	3.0	2.0	167	2.6	6.3	2.6	1.6
reg9081s_c						177	0.0	7.8	0.0	0.0
reg90820_c						183	0.0	0.0	4.7	0.0
reg90830_c						181	0.0	0.0	5.7	0.0
reg9084s_c						172	0.0	9.9	0.5	0.0
reg90850_c						172	0.0	3.7	6.8	0.0
reg90860_c						185	0.0	0.0	3.7	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.

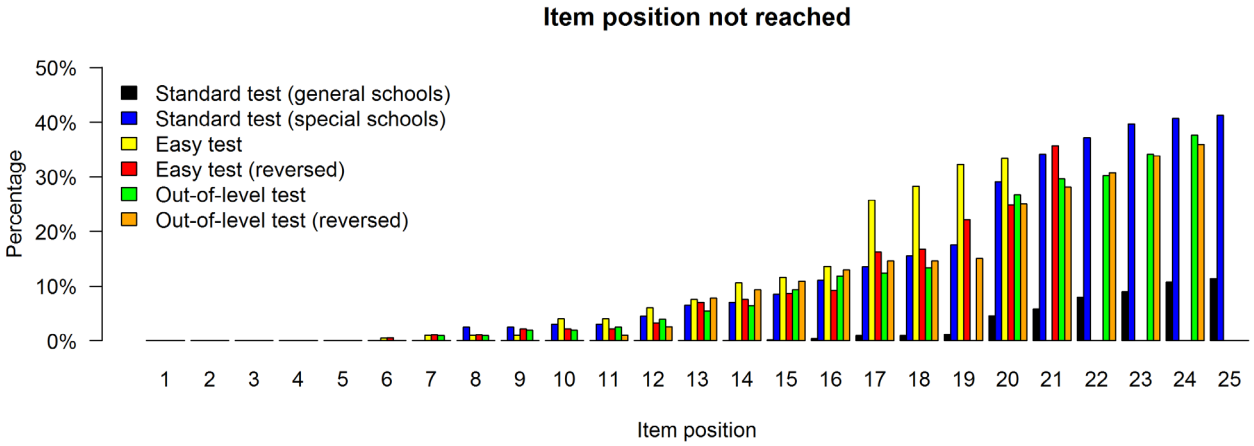

Figure 6. Item position not reached by test version

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 8). The median point-biserial correlations for the distractors fell between -.23 and -.17 for the different test versions. In contrast, the correlations of the correct responses with the total scores varied between *Mdn* = .38 and .46. These results indicate that the distractors functioned well in all five test versions.

Table 8

Distractor Analyses for Test Versions in Special Schools

	Distractors			***Correct response***		
Test version	***Mdn***	***Min***	***Max***	***Mdn***	***Min***	***Max***
Standard	-.17	-.42	.13	.40	.19	.62
Easy	-.19	-.35	-.01	.38	.28	.56
Easy (reversed)	-.19	-.39	.01	.38	.23	.60
Out-of-level	-.23	-.39	.10	.46	.22	.65
Out-of-level (reversed)	-.20	-.45	.12	.45	.26	.57

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 9. 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 did not vary substantially between the five test versions administered in special schools. Because there was a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index for item difficulty. The median percentage of correct responses within dichotomous MC items varied between 45% and 57%. In general schools, slightly more correct responses were observed (*Mdn* = 66%). The item difficulties (for dichotomous variables) and location parameters (for polytomous variables) in Table 9 were estimated by constraining the mean of the ability distribution to be zero. The median item difficulties (or location parameters for polytomous variables) were comparable between test versions in special schools and varied between -0.3 and 0.1. Similar, the respective range of these parameters fell between -3.2 and 1.6 and indicated no pronounced differences between test versions. In contrast, for low-achieving students in lower secondary schools the test was slightly easier resulting in median item difficulties (or location parameters) of -0.8 with a range of [-3.2, 1.0].

Table 9

Summary of Item Parameters for Different Test Versions

Test version	**Percentage correct**	**ξ**	**WMNSQ**	***t***	**Item-rest correlation**	**Discr.**
Standard (special schools)	46 [22, 87]	0.0 [-2.1, 1.4]	1.0 [0.9, 1.1]	0.1 [-2.4, 1.4]	.2 [.0, .5]	0.7 [0.2, 1.9]
Standard (general schools)	66 [28, 94]]	-0.8 [-3.2, 1.0]	1.0 [0.9, 1.2]	-0.2 [-2.1, 2.9]	.3 [.1, .5]	0.9 [0.2, 1.7]
Easy	47 [35, 88]	0.0 [-2.2, 0.7]	1.0 [0.9, 1.1]	-0.1 [-2.1, 1.8]	.3 [.1, .6]	0.7 [0.3, 1.5]
Easy (reversed)	57 [34, 79]	-0.3 [-1.5, 0.8]	1.0 [0.9, 1.2]	0.2 [-2.2, 2.3]	.2 [-.1, .4]	0.7 [-0.2, 1.7]
Out-of-level	45 [22, 60]	0.1 [-0.7, 1.5]	1.0 [0.8, 1.2]	-0.3 [-2.8, 3.2]	.3 [.1, .6]	1.0 [0.2, 2.3]
Out-of-level (reversed)	54 [18, 85]	-0.2 [-2.0, 1.6]	1.0 [0.9 1.1]	0.0 [-1.9, 2.0]	.3 [.1, .5]	0.7 [0.4, 1.6]

Note. Reported are median values across all items with minimum and maximum value in parentheses. ξ = Item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, Discr. = Discrimination parameter of a generalized partial credit model. Percent correct scores are not informative for polytomous CMC and MA 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 good (see Table 9). The median values of the WMNSQ for the five test versions fell around 1.0, few items exhibited considerable misfit greater than 1.15. The respective *t*-values indicated no substantial misfit ($|t| > 6$) at all. 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 or .3 and, thus, did not indicated substantial differences between test versions. Although some items had rather low item-total

correlations (e.g., reg9017s_c in the easy test reversed; see Appendix B) no systematic differences were identified between test versions. Moreover, all item characteristic curves showed an acceptable fit of the items. Similar results were observed for the standard test administered in general schools (see Table 9). Most items exhibited good to very good fit. On average, the item-total correlations seemed to be slightly larger.

4.2.4 Test targeting and reliability

Table 10 summarizes the estimated population standard deviations and the reliabilities of the different test versions. Generally, the standard deviations were satisfactory falling between 0.73 and 0.94. Notably, the out-of-level test version reflected substantially more interindividual differences (*SD* = 0.94) that was even larger than the standard test administered in general schools (*SD* = 0.76). Given the rather easy tests, the EAP and WLE reliabilities were somewhat compromised falling between .57 and .76 for the tests administered in special schools. Again, the out-of-level tests exhibited the largest reliabilities (EAP/PV reliability = .76 / .76, WLE reliability = .71 / .72) which was similar to the standard test administered in general schools (EAP/PV reliability = .75, WLE reliability = .76). Overall, the out-of-level test that included items designed for a younger age cohort seemed to perform better as compared to the alternative test versions.

Table 10

Reliabilities of the Different Test Versions

Test version	*SD*	EAP Rel.	WLE Rel.
Standard (special schools)	0.78	.67	.60
Standard (general schools)	0.76	.75	.71
Easy	0.83	.68	.63
Easy (reversed)	0.73	.63	.57
Out-of-level	0.94	.76	.71
Out-of-level (reversed)	0.94	.76	.72

4.3 Parameter Estimates for Concurrently Scaled Tests

4.3.1 Item parameters

Because the standard, out-of-level, and easy test versions administered in special schools presented most items at roughly the same position within each test, the three test versions 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 11, whereas the step parameters are given in Table 12. The estimated item difficulties and location parameters ranged from -2.2 (item reg90120_c) to 1.5 (item reg90250_c) with a median of 0.1 and, thus, covered a rather broad range. However, the standard errors (SE) of the estimated parameters were rather large with a *Mdn* = 0.14 and a range of [0.05, 0.23]. Thus, the reported item parameters had a somewhat limited precision.

Table 11

Item Parameters for Combined Scaling of Standard, Easy, and Out-of-Level Tests in Special Schools

	Item		Pos.		*N*	Percentage correct	ξ	$SE_ξ$	WMNSQ	*t*	Item-rest correlation	Discr.	aQ_3
1	reg90110_c	1	1		351	78.35	-1.46	0.14	1.01	0.17	0.20	0.82	0.05
2	reg90120_c	2	2		390	87.44	-2.18	0.16	0.96	-0.37	0.23	1.51	0.05
3	reg90150_c	5	3		374	38.24	0.54	0.11	0.99	-0.11	0.18	0.81	0.06
4	reg9016s_c	6	4		320		0.61	0.05	0.95	-0.61	0.34	0.50	0.04
5	reg9017s_c	7	5		350		-0.87	0.07	1.01	0.14	0.08	0.39	0.05
6	reg90210_c	8	6	7	550	45.64	0.20	0.09	0.95	-1.64	0.32	1.11	0.05
7	reg90220_c	9	7	8	547	30.16	0.97	0.10	1.04	0.81	0.18	0.70	0.06
8	reg90230_c	10	8	9	553	51.54	-0.07	0.09	0.97	-1.03	0.26	1.01	0.04
9	reg90240_c	11	9	10	347	36.01	0.65	0.10	1.08	2.12	0.18	0.49	0.05
10	reg90250_c	12		11	362	21.82	1.45	0.14	1.08	1.05	0.07	0.46	0.06
11	reg90310_c	13	10	12	535	45.42	0.19	0.09	0.99	-0.23	0.29	0.82	0.05
12	reg90320_c	14	11	13	525	53.90	-0.20	0.09	0.90	-3.38	0.44	1.58	0.06
13	reg9033s_c	15	12	14	487		-0.66	0.05	0.98	-0.31	0.30	0.53	0.05
14	reg90340_c	16	13	15	497	57.95	-0.41	0.10	0.99	-0.34	0.30	0.97	0.06
15	reg90350_c	17	14	16	487	57.49	-0.39	0.10	1.00	-0.14	0.28	0.85	0.06
16	reg90360_c	18	15	17	498	59.84	-0.50	0.10	1.08	2.28	0.15	0.48	0.06
17	reg90370_c	19	16	18	474	39.24	0.47	0.10	1.07	1.83	0.16	0.55	0.07
18	reg90410_c	20	17		273	67.40	-0.91	0.14	1.06	1.16	0.14	0.49	0.06
19	reg90420_c	21	18		256	39.45	0.39	0.14	1.05	0.90	0.24	0.63	0.06
20	reg90430_c	22			113	33.63	0.61	0.21	1.14	1.55	0.06	0.25	0.06
21	reg90440_c	23	19		234	46.58	0.01	0.14	1.02	0.46	0.28	0.72	0.04
22	reg90450_c	24	20		232	61.64	-0.68	0.14	1.05	0.91	0.21	0.55	0.05
23	reg90460_c	25			105	29.52	0.75	0.23	1.03	0.32	0.13	0.58	0.08
24	reg90610_c			1	185	57.30	-0.33	0.16	0.93	-1.25	0.36	1.36	0.08
25	reg90620_c			2	187	38.50	0.54	0.16	0.94	-0.99	0.24	1.21	0.07
26	reg90630_c			3	196	47.45	0.12	0.15	0.98	-0.35	0.28	1.01	0.07
27	reg90640_c			4	190	49.47	0.04	0.16	0.97	-0.57	0.29	1.03	0.07
28	reg90650_c			5	195	45.13	0.23	0.16	0.97	-0.51	0.29	1.06	0.08
29	reg90660_c			6	195	47.69	0.11	0.15	0.94	-1.17	0.34	1.26	0.06
30	reg90720_c			20	144	45.14	0.14	0.18	0.81	-3.18	0.60	2.36	0.10
31	reg90730_c			21	137	48.91	-0.08	0.18	0.97	-0.42	0.36	1.11	0.08
32	reg90740_c			22	137	41.61	0.27	0.19	0.98	-0.25	0.36	0.93	0.06
33	reg90750_c			23	125	44.00	0.14	0.20	1.00	0.01	0.35	0.96	0.05
34	reg9076s_c			24	101		0.46	0.11	0.99	-0.06	0.33	0.47	0.08

Note. Pos. = Item position in standard, easy, and out-of-level tests, *N* = Number of valid responses for item, ξ = Item difficulty / location parameter, $SE_ξ$ = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, Discr. = Discrimination parameter of a generalized partial credit model, aQ_3 = Adjusted average absolute residual correlation for item (Yen, 1984, 1993).
Percent correct scores are not informative for polytomous CMC and MA items and, thus, are not reported.

Table 12

Step Parameters (with Standard Errors) for Combined Scaling of Standard, Easy, and Out-of-Level Tests in Special Schools

Item	Step 1	Step 2	Step 3	Step 4
reg9016s_c	-0.41 (0.11)	0.12 (0.14)	0.16 (0.20)	0.14
reg9017s_c	1.32 (0.19)	-1.32		
reg9033s_c	0.32 (0.10)	0.40 (0.13)	-0.72	
reg9076s_c	-0.26 (0.20)	-0.66 (0.22)	0.92	

Note. The last step parameter is a constrained parameter for model identification and, thus, has no standard error.

4.3.2 Item fit

For the concurrently scaled test versions in special schools (see Table 11) no item exhibited a noteworthy WMNSQ exceeding 1.15. The values of the WMNSQ fell between 0.81 and 1.14 (*Mdn* = 0.99). All *t*-values indicated good fit (*Max* = 2.28). Although three items exhibited item-rest correlations less than .10, most items had adequate discriminations with a median of .28.

4.3.3 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 generalized partial credit model (GPCM) that estimates discrimination parameters was fitted to the data. The estimated discriminations for the concurrently scaled test versions in special schools differed moderately among items (see Table 11), ranging from 0.25 (item reg90430_c) to 2.36 (item reg90720_c). The median discrimination parameter fell at 0.82. Model fit indices suggested a slightly better model fit of the PCM (AIC =14,558, BIC = 14,747, number of parameters = 43) as compared to the GPCM (AIC = 14,516, BIC = 14,850, number of parameters = 76). In line with the theoretical conception underlying the test construction (see Pohl & Carstensen, 2012 and 2013, for a discussion of this issue) the PCM seemed an adequate scaling model for the test.

4.3.4 Unidimensionality

The dimensionality of the test was investigated by evaluating the correlations between the residuals of the PCM. The adjusted Q_3 statistics for the concurrently scaled test versions in special schools (see Table 11) were quite low (*Mdn* = .06, *Min* = .04, *Max* = 0.10) and, thus, indicated an essentially unidimensional test. Because the reading test is constructed to measure a single dimension, a unidimensional reading competence score was estimated.

4.3.5 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. Because some items in the reading competence tests were polytomous, we calculated Thurstonian thresholds for each response category (Wu, Tam, & Jen, 2016). These indicate the location at the latent dimension at which the probability of achieving a score above the respective threshold is 50%. Thus, it is similar to the item difficulties of dichotomous items. In Figure 7,

the category thresholds of the concurrently scaled reading items from the standard, easy, and out-of-level test versions administered in special schools and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of category thresholds. The respective thresholds ranged from -2.18 (item reg90120_c) to 3.16 (item reg9076s_c) and, thus, spanned a rather broad range. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 0.72, which implies a somewhat limited differentiation between students. The reliability of the test (EAP/PV reliability = .71, WLE reliability = .65) was acceptable. The mean of the item distribution was about 0.00 logits and, thus, comparable to the mean person ability distribution. Thus, the items covered a wide range of the ability distribution and, thus, had an acceptable difficulty.

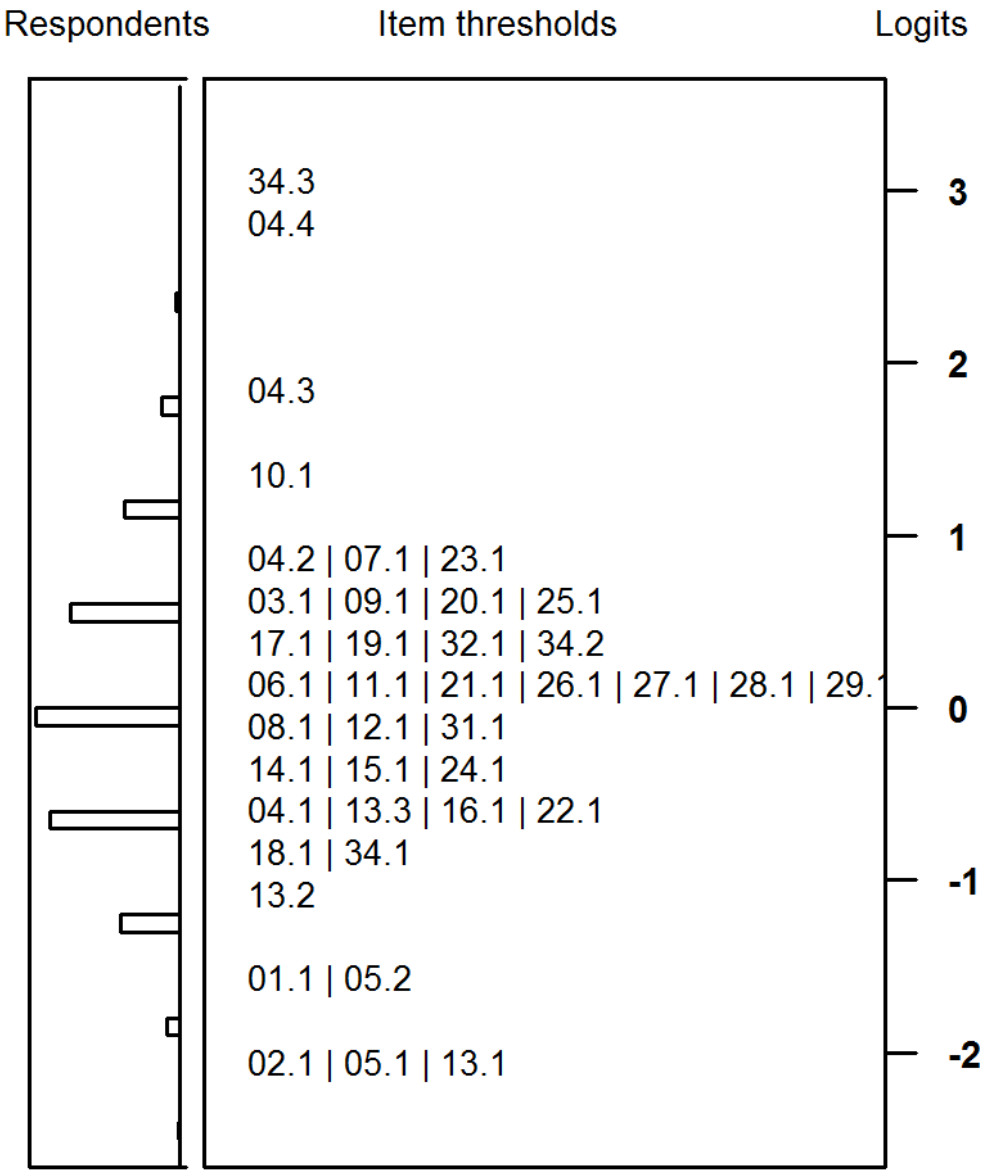

Figure 7. Test targeting. The distribution of person ability in the sample is given on the left-hand side of the graph. The category thresholds of the items are given on the right-hand side of the graph. Each number represents one threshold with the first part (before the dot) corresponding to the item number in Table 10 and the second part indicating the threshold.

4.4 Differential Item Functioning

Differential item functioning (DIF) was used to evaluate test fairness with regard to the different test versions administered in special schools. Additionally, the comparability of the standard test version across special and general schools was examined. The differences between the estimated item difficulties in the various groups are summarized in Tables 13, 14, and 15. For example, the column "Special schools vs. General schools" reports the differences in item difficulties between the two school types; a positive value would indicate that the item was more difficult for students from special schools, whereas a negative value would highlight a lower difficulty for students from general schools. In contrast, the main effect is to be interpreted on a group level. As such, a positive value indicates that students from general schools, on average, had a higher ability as compared to students from special schools; whereas a negative value would highlight a lower ability, on average, for students from general schools as compared to students from special schools. 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 16).

4.4.1 Text order effects

In special schools, the easy and the out-of-level test versions were administered in two different formats that varied the order of the reading texts (see section 2). In order to evaluate, whether changing the text positions (and, thus, also the item positions) within the tests DIF was evaluated separately for the easy and the out-of-level test versions. The respective differences in item difficulties (or location parameters) are summarized in Table 13. Because the out-of-level test originally included five texts but only four texts were analyzed (see section 2), the analyses for the out-of-level test version refer to the common items included in the first four presented texts.

Text order effects for the easy test version resulted in substantial DIF effects exceeding |Cohen's *d*| = 0.50 for 11 items. Three of them were significantly greater than *d* = 0.25 (i.e., our threshold for non-negligible DIF). DIF slightly affected the main effect for the test version: When ignoring DIF effects respondents receiving the easy test had, on average, *d* = -0.11 lower reading abilities as compared to respondents receiving the texts in the reversed order. In contrast, acknowledging DIF effects identified no group differences, *d* = -0.03. An overall test for DIF (see Table 16) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). A model comparison using Akaike's (1974) information criterion (AIC) favored the model estimating DIF. In contrast, the Bayesian information criterion (BIC; Schwarz, 1978) that takes the number of estimated parameters into account and, thus, guards against overparameterization of models, indicated more support for the main effect model.

Text order effects for the out-of-level test version resulted in substantial DIF effects exceeding |Cohen's *d*| = 0.50 for 9 items. Four of them were significantly greater than *d* = 0.25 (i.e., our threshold for non-negligible DIF). DIF strongly affected the main effect for the test version: When ignoring DIF effects respondents receiving the out-of-level test had, on average, *d* = -0.22 lower reading abilities as compared to respondents receiving the texts in the reversed order. In contrast, acknowledging DIF effects identified no group differences, *d* = -0.08. The overall test for DIF (see Table 16) using the AIC favored the model estimating DIF, whereas the BIC favored the main effect model.

Table 13

Differential Item Functioning for Original Easy and Out-of-Level Tests versus Reversed Text Order Versions

Item	Easy tests	Item	Out-of-Level tests
reg90110_c	0.16 (0.21)	reg90210_c	-0.20 (-0.22)
reg90120_c	-0.03 (-0.04)	reg90220_c	-0.04 (-0.05)
reg90150_c	0.08 (0.10)	reg90230_c	-0.02 (-0.02)
reg9016s_c	-0.06 (-0.07)	reg90240_c	0.02 (0.02)
reg9017s_c	-0.33 (-0.42)	reg90250_c	-0.21 (-0.23)
reg90210_c	-0.19 (-0.24)	reg90310_c	0.01 (0.01)
reg90220_c	0.42 (0.54)	reg90320_c	0.04 (0.04)
reg90230_c	0.07 (0.09)	reg9033s_c	-0.17 (-0.18)
reg90240_c	0.03 (0.04)	reg90340_c	-0.25 (-0.28)
reg90250_c	-0.15 (-0.20)	reg90350_c	-0.12 (-0.13)
reg90310_c	-0.18 (-0.23)	reg90360_c	0.18 (0.19)
reg90320_c	-0.04 (-0.05)	reg90370_c	-0.48 (-0.53)
reg9033s_c	-0.36 (-0.46)	reg90610_c	-0.33 (-0.37)
reg90340_c	0.32 (0.41)	reg90620_c	-1.32* (-1.44)
reg90350_c	-0.16 (-0.20)	reg90630_c	-0.68 (-0.74)
reg90360_c	0.17 (0.22)	reg90640_c	-0.04 (-0.04)
reg90370_c	0.10 (0.13)	reg90650_c	-0.61 (-0.67)
reg90410_c	-0.33 (-0.42)	reg90660_c	-0.78 (-0.86)
reg90420_c	-0.18 (-0.23)	reg90720_c	0.19 (0.21)
reg90430_c	0.54 (0.68)	reg90730_c	0.22 (0.24)
reg90440_c	-0.46 (-0.58)	reg90740_c	0.42 (0.46)
reg90450_c	0.93* (1.19)	reg90750_c	0.03 (0.03)
reg90460_c	-0.99* (-1.27)	reg9076s_c	-0.05 (-0.06)
reg90610_c	-0.12 (-0.15)	reg9081s_c	0.66* (0.72)
reg90620_c	-0.42 (-0.54)	reg90820_c	1.11* (1.21)
reg90630_c	-0.43 (-0.55)	reg90830_c	0.94* (1.03)
reg90640_c	0.63 (0.81)	reg9084s_c	0.54 (0.59)
reg90650_c	-0.01 (-0.02)	reg90850_c	0.40 (0.44)
reg90660_c	0.44 (0.57)		
reg90720_c	0.96* (1.22)		
reg90730_c	-0.44 (-0.56)		
reg90740_c	0.06 (0.07)		
reg90750_c	0.11 (0.14)		
reg9076s_c	0.14 (0.18)		
Main effects:			
DIF model	-0.09 (-0.11)		-0.08 (-0.08)
Main effect model	-0.02 (-0.03)		-0.21 (-0.22)

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

* Absolute standardized difference was significantly ($p < .05$) greater than 0.25 (see Fischer, Rohm, Gnambs, & Carstensen, 2016).

Taken together, these analyses indicate that text order effects distorted group comparisons to some degree. Therefore, the reversed test versions were not included in the concurrent scaling model (see Table 10).

4.4.2 Test version effects

In special schools, the standard, easy, and the out-of-level test versions administered a subsample of items at roughly the same item position within each test (see Table 11). Therefore, these items might be used as anchor items (cf. Fischer et al., 2016) to link the different test versions and estimate a common reading competence score. However, to do so these anchor items must not exhibit substantial DIF; otherwise, the estimated reading competence scores might be distorted. Therefore, DIF was evaluated for the common items included in the three concurrently scaled test versions (Table 11). The respective differences in item difficulties (or location parameters) are summarized in Table 14.

Table 14

Differential Item Functioning for Standard versus Easy versus Out-of-Level Test Versions

Item	Standard versus easy	Item	Standard versus out-of-level	Item	Easy versus out-of-level
reg90110_c	-0.12 (-0.14)	reg90210_c	-0.09 (-0.11)	reg90210_c	-0.08 (-0.09)
reg90120_c	0.09 (0.11)	reg90220_c	0.08 (0.09)	reg90220_c	-0.36 (-0.43)
reg90150_c	-0.24 (-0.28)	reg90230_c	-0.09 (-0.11)	reg90230_c	0.27 (0.32)
reg9016s_c	0.11 (0.13)	reg90240_c	0.38 (0.44)	reg90240_c	-0.14 (-0.16)
reg9017s_c	0.10 (0.12)	reg90250_c	0.01 (0.01)	reg90310_c	-0.10 (-0.12)
reg90210_c	-0.04 (-0.04)	reg90310_c	-0.10 (-0.12)	reg90320_c	0.28 (0.33)
reg90220_c	0.42 (0.51)	reg90320_c	0.05 (0.06)	reg9033s_c	0.07 (0.09)
reg90230_c	-0.38 (-0.45)	reg9033s_c	0.04 (0.05)	reg90340_c	0.14 (0.16)
reg90240_c	0.49 (0.59)	reg90340_c	0.22 (0.26)	reg90350_c	-0.11 (-0.13)
reg90310_c	-0.02 (-0.02)	reg90350_c	0.04 (0.04)	reg90360_c	-0.31 (-0.37)
reg90320_c	-0.24 (-0.28)	reg90360_c	-0.56 (-0.65)		
reg9033s_c	-0.03 (-0.03)				
reg90340_c	0.08 (0.10)				
reg90350_c	0.13 (0.15)				
reg90360_c	-0.27 (-0.32)				
reg90370_c	-0.33 (-0.39)				
reg90410_c	0.28 (0.33)				
reg90420_c	-0.02 (-0.02)				
reg90440_c	-0.09 (-0.11)				
Main effects:					
DIF model	-0.02 (-0.03)		-0.03 (-0.03)		-0.03 (-0.04)
Main effect model	-0.02 (-0.03)		-0.03 (-0.03)		-0.03 (-0.04)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses.
* Absolute standardized difference was significantly ($p < .05$) greater than 0.25 (see Fischer et al., 2016).

Substantial DIF effects exceeding |Cohen's *d*| = 0.50 for the three test versions were rare and observed for only three items (reg90220_c, reg90240_c, reg90360_c). However, none of these effects were significantly greater than *d* = 0.25 (i.e., our threshold for non-negligible DIF). Moreover, DIF did not affect the main effect for the test version: When ignoring DIF effects the main effects were close to 0.0 and did not change substantially when acknowledging DIF (see Table 14). Finally, the overall tests for DIF (see Table 16) favored the main effect models. Taken together, these results do not indicate substantial DIF effects that might have distorted estimates of students reading competences based on their concurrently scaled responses.

Table 15

Differential Item Functioning for Standard Tests in Special Schools versus General Schools

Item	Difference
reg90110_c	0.17 (0.19)
reg90120_c	0.28 (0.32)
reg90150_c	0.25 (0.28)
reg9016s_c	0.07 (0.08)
reg9017s_c	-0.32 (-0.36)
reg90210_c	0.43 (0.47)
reg90220_c	0.03 (0.03)
reg90230_c	0.66* (0.74)
reg90240_c	0.76* (0.84)
reg90250_c	-0.39 (-0.43)
reg90310_c	0.37 (0.41)
reg90320_c	0.74* (0.82)
reg9033s_c	-0.50* (-0.56)
reg90340_c	0.01 (0.01)
reg90350_c	0.23 (0.26)
reg90360_c	-0.77* (-0.86)
reg90370_c	-0.26 (-0.28)
reg90410_c	-0.20 (-0.22)
reg90420_c	-0.12 (-0.13)
reg90430_c	-0.28 (-0.31)
reg90440_c	-0.41 (-0.45)
reg90450_c	-0.58 (-0.64)
Main effects:	
DIF model	-0.73 (-0.81)
Main effect model	-0.60 (-0.67)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses.
* Absolute standardized difference was significantly ($p < .05$) greater than 0.25 (see Fischer et al., 2016).

4.4.3 Special schools versus general schools

The standard test version was administered to students with special educational needs in special schools and low-achieving students attending general secondary schools. Group comparisons between the two school types require measurement invariance; otherwise results might be biased. The respective differences in item difficulties (or location parameters) are summarized in Table 15. The analyses showed substantial DIF effects exceeding |Cohen's *d*| = 0.50 for six items. Four of them were significantly greater than *d* = 0.25 (i.e., our threshold for non-negligible DIF). DIF also affected the main effect for the school type: When ignoring DIF effects respondents in special schools had, on average, *d* = -0.67 lower reading abilities as compared to respondents in general schools. In contrast, acknowledging DIF effects identified a larger group difference of *d* = -0.81. The overall test for DIF (see Table 16) using the AIC favored the model estimating DIF, whereas the BIC favored the main effect model. Taken together, these analyses suggest non-negligible DIF between the two school types that might compromise group comparisons. However, it might be possible to identify subgroups of students from special schools for which a comparable measurement model can be identified (see Pohl, Südkamp, Hardt, Carstensen, & Weinert, 2016, for a respective approach).

Table 16

Comparisons of Models with and without DIF

Comparison	Model	*N*	Deviance	Number of parameters	AIC	BIC
Easy tests [a]	DIF	383	8633	50	8733	8930
	Main effects	383	8686	31	8748	8871
Out-of-level test [a]	DIF	394	12015	69	12153	12427
	Main effects	394	12117	41	12199	12362
Standard tests [b]	DIF	699	18188	56	18300	18555
	Main effects	699	18300	34	18368	18522
Standard versus easy	DIF	397	8773	50	8873	9072
	Main effects	397	8791	31	8853	8977
Standard versus out-of-level	DIF	401	5545	29	5603	5719
	Main effects	401	5555	18	5591	5662
Easy versus out-of-level	DIF	400	5287	27	5341	5449
	Main effects	400	5297	17	5331	5399

Note. [a] Original versus reversed text order, [b] Special versus general 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 different versions of a reading

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:

- Students with special educational needs required substantially more time for the reading test as compared to low-achieving students from lower secondary schools. As a result, large numbers of not-reached items were observed. Even a shortened test version including only four (instead of five) reading texts exhibited increased missing rates for the last items in the test. Thus, students with special educational needs either require longer testing times to finish tests of the same length as in general schools or, alternatively, their competence tests need to be substantially shorter.

- Students in special schools omitted substantially more items as compared to students from lower secondary schools. Notably, the omission rates were unrelated to the difficulty of the items (which was in contrast to lower secondary schools). This might indicate difficulties in understanding the instruction or the content of some items. Particularly, items with more complex response formats (CMC, MA) exhibited larger omission rates in special schools. For future assessments in special schools, it might be beneficial to limit the response formats to simple formats such as MC items.

- The reading competence tests administered in special schools exhibited somewhat limited variances and reliabilities. The most appropriate test version in terms of test targeting represented the out-of-level version that included items designed for a younger age cohort. Thus, reading competence tests for students with special educational needs should target a pronouncedly lower average ability as compared to low achieving students in lower secondary schools.

- Comparisons between students from different school types using the administered reading competence test cannot be recommended. Despite receiving identical test versions, substantial DIF suggested that the test functioned rather differently for students from special schools and students from general schools. This makes the analysis of schooling effects across different school types rather infeasible. However, it might be possible to identify subgroups of students from special schools for which measurement invariance can be achieved. A respective approach is outlined in Pohl et al. (2016).

Despite the large amount of missing responses the five test versions represented essentially unidimensional scales conforming to the partial credit model (Masters, 1982). Although some items exhibited a marginal misfit, most items had satisfactory psychometric properties allowing the estimation of reading competence scores. Moreover, the different test versions that presented the items at roughly the same position exhibited no pronounced DIF. Thus, measurement invariance between the three test versions administered in special schools could be established. As a result, linked competence scores for the concurrently scaled tests could be estimated that can be used in future research on reading competence in special schools.

In conclusion, these results indicate that reading competences can be measured in special schools, provided appropriate tests with shorter length, easier items, and simple response

formats are administered. However, even then comparisons with students from general schools might not be feasible because the tests lack measurement invariance.

6. Data in the Scientific Use File

6.1 Naming conventions

The data in the SUF contains 51 items of which 34 items were included in the reported analyses. Forty-four items were scored dichotomously (MC items) with 0 indicating an incorrect response and 1 indicating a correct response, whereas seven items were scored polytomously (CMC and MA items). MC items are marked with a '0_c' at the end of the variable name, whereas the variable names of the CMC and MA items end in 's_c'. In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category. For further details on the naming conventions of the variables see Fuß and colleagues (2019).

6.2 Reading competence scores in special schools

In the SUF, manifest reading competence scores are provided in the form of WLEs ("reg9_sc1") including their respective standard errors ("reg9_sc2"). These scores are based on the concurrently scaled standard, easy, and out-of-level test versions (see section 2). Thus, the WLEs are located on a common scale and, thus, can be compared across different test versions. Importantly, these scores are not linked to the scale of the main sample in Starting Cohort 4. Therefore, they must not be used to compare reading competences of students from special schools and respective competences of students from general schools.

The R Syntax for estimating the WLEs is provided in Appendix C. In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category. For persons who either did not take part in the reading test or who did not give enough valid responses, no WLE is estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values. Alternatively, users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values. A description of these approaches can be found in Pohl and Carstensen (2012).

References

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

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. (2016). *Linking the Data of the Competence Tests* (NEPS Survey Paper No. 1). Bamberg: Leibniz Institute for Educational Trajectories, 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.

Gehrer, K., & Artelt, C. (2013). Literalität und Bildungslaufbahn: Das Bildungspanel NEPS. In A. Bertschi-Kaufmann, & C. Rosebrock (Eds.), *Literalität erfassen: bildungspolitisch, kulturell, individuell* (pp. 168-187). Weinheim, Germany: Juventa.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online, 5*, 50-79.

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.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174. https://doi.org/10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. ETS *Applied Psychological Measurement, 16*, 159-176. https://doi.org/10.1177/014662169201600206

Nusser, L., & Messingschlager, M. (2018). Erfassung kognitiver Grundfähigkeiten bei Schülerinnen und Schülern an Förderschulen in Startkohorte 4 (Klasse 9) [*Measurement of general cognitive functioning in sutdents at special schools in starting cohort 4 (grade 9)*] (NEPS Survey Paper No. 33). Bamberg, Germany: Leibniz-Institute for Educational Trajectories, 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.

Pohl, S., Südkamp, A., Hardt, K., Carstensen, C. H., & Weinert, S. (2016). Testing students with special educational needs in large-scale assessments: Psychometric properties of test scores and associations with test taking behavior. *Frontiers in Psychology, 7*, 154. https://doi.org/10.3389/fpsyg.2016.00154

R Core Team (2016). *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.

Robitzsch, A., Kiefer, T., & Wu, M. (2019). *TAM: Test analysis modules*. R package version 3.2-24. URL: https://CRAN.R-project.org/package=TAM

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

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational Measurement for Applied Researchers*. Singapore, Singapore: Springer.

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*, 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*, 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Appendices

Appendix A: Allocation of items to text types and cognitive requirements

Item	Response format	Text number	Text type	Cognitive requirement
reg90110_c	MC	1	Information text	Drawing text-related conclusions
reg90120_c	MC	1	Information text	Finding information in the text
reg90130_c	MC	1	Information text	Drawing text-related conclusions
reg90140_c	MC	1	Information text	Reflecting and assessing
reg90150_c	MC	1	Information text	Finding information in the text
reg9016s_c	MA	1	Information text	Reflecting and assessing
reg9017s_c	CMC	1	Information text	Drawing text-related conclusions
reg90210_c	MC	2	Instruction text	Reflecting and assessing
reg90220_c	MC	2	Instruction text	Drawing text-related conclusions
reg90230_c	MC	2	Instruction text	Drawing text-related conclusions
reg90240_c	MC	2	Instruction text	Drawing text-related conclusions
reg90250_c	MC	2	Instruction text	Drawing text-related conclusions
reg90310_c	MC	3	Advertising text	Reflecting and assessing
reg90320_c	MC	3	Advertising text	Drawing text-related conclusions
reg9033s_c	CMC	3	Advertising text	Finding information in the text
reg90340_c	MC	3	Advertising text	Reflecting and assessing
reg90350_c	MC	3	Advertising text	Finding information in the text
reg90360_c	MC	3	Advertising text	Finding information in the text
reg90370_c	MC	3	Advertising text	Finding information in the text
reg90410_c	MC	4	Literary text	Finding information in the text
reg90420_c	MC	4	Literary text	Drawing text-related conclusions
reg90430_c	MC	4	Literary text	Drawing text-related conclusions
reg90440_c	MC	4	Literary text	Reflecting and assessing
reg90450_c	MC	4	Literary text	Drawing text-related conclusions
reg90460_c	MC	4	Literary text	Finding information in the text
reg9047s_c	CMC	4	Literary text	Reflecting and assessing
reg90510_c	MC	5	Commenting text	Reflecting and assessing
reg90520_c	MC	5	Commenting text	Finding information in the text
reg90530_c	MC	5	Commenting text	Drawing text-related conclusions
reg90540_c	MC	5	Commenting text	Finding information in the text
reg90550_c	MC	5	Commenting text	Reflecting and assessing
reg90560_c	MC	5	Commenting text	Finding information in the text
reg90570_c	MC	5	Commenting text	Finding information in the text
reg90610_c	MC	6	Information text	Finding information in the text
reg90620_c	MC	6	Information text	Finding information in the text
reg90630_c	MC	6	Information text	Finding information in the text
reg90640_c	MC	6	Information text	Finding information in the text
reg90650_c	MC	6	Information text	Drawing text-related conclusions
reg90660_c	MC	6	Information text	Drawing text-related conclusions
reg90710_c	MC	7	Literary text	Reflecting and assessing
reg90720_c	MC	7	Literary text	Finding information in the text

Item	Response format	Text number	Text type	Cognitive requirement
reg90730_c	MC	7	Literary text	Reflecting and assessing
reg90740_c	MC	7	Literary text	Finding information in the text
reg90750_c	MC	7	Literary text	Drawing text-related conclusions
reg9076s_c	MA	7	Literary text	Reflecting and assessing
reg9081s_c	CMC	8	Commenting text	Drawing text-related conclusions
reg90820_c	MC	8	Commenting text	Finding information in the text
reg90830_c	MC	8	Commenting text	Finding information in the text
reg9084s_c	CMC	8	Commenting text	Drawing text-related conclusions
reg90850_c	MC	8	Commenting text	Reflecting and assessing
reg90860_c	MC	8	Commenting text	Drawing text-related conclusions

Note. MC = Simple multiple-choice, CMC = Complex multiple-choice, MA = Matching.

Appendix B: Item parameters for different test versions

Table B.1

Item Parameters for Standard Test in Special Schools

	Item	Pos.	Percentage correct	ξ	$SE_ξ$	WMNSQ	*t*	Item-rest correlation	Discr.	aQ_3
1	reg90110_c	1	78.98	-1.49	0.19	1.03	0.32	0.14	0.69	0.08
2	reg90120_c	2	86.80	-2.09	0.22	0.95	-0.33	0.22	1.74	0.07
3	reg90150_c	5	40.53	0.44	0.16	1.02	0.39	0.15	0.68	0.07
4	reg9016s_c	6		0.73	0.07	0.95	-0.48	0.30	0.46	0.07
5	reg9017s_c	7		-0.78	0.1	1.03	0.30	0.03	0.29	0.07
6	reg90210_c	8	46.11	0.16	0.16	0.89	-2.16	0.39	1.48	0.07
7	reg90220_c	9	27.07	1.12	0.18	0.94	-0.65	0.28	1.20	0.09
8	reg90230_c	10	54.64	-0.22	0.16	0.99	-0.11	0.19	0.84	0.05
9	reg90240_c	11	30.00	0.94	0.17	1.08	1.01	0.26	0.44	0.07
10	reg90250_c	12	21.71	1.43	0.19	1.09	0.87	0.02	0.33	0.07
11	reg90310_c	13	46.24	0.16	0.16	1.02	0.32	0.24	0.73	0.06
12	reg90320_c	14	54.97	-0.25	0.16	0.88	-2.40	0.48	1.92	0.07
13	reg9033s_c	15		-0.64	0.08	1.00	0.06	0.25	0.36	0.06
14	reg90340_c	16	55.62	-0.29	0.17	0.95	-0.84	0.29	0.98	0.08
15	reg90350_c	17	55.77	-0.31	0.17	0.99	-0.15	0.31	0.77	0.07
16	reg90360_c	18	65.41	-0.76	0.18	0.99	-0.18	0.22	0.83	0.08
17	reg90370_c	19	40.67	0.37	0.18	1.05	0.73	0.19	0.63	0.09
18	reg90410_c	20	64.12	-0.74	0.19	1.09	1.29	0.12	0.37	0.07
19	reg90420_c	21	39.17	0.39	0.20	1.02	0.26	0.25	0.71	0.07
20	reg90430_c	22	33.63	0.62	0.21	1.12	1.42	0.06	0.24	0.06
21	reg90440_c	23	46.23	-0.02	0.21	0.99	-0.07	0.31	0.79	0.07
22	reg90450_c	24	59.43	-0.62	0.21	0.98	-0.32	0.32	0.92	0.08
23	reg90460_c	25	29.52	0.77	0.23	1.03	0.30	0.13	0.62	0.08

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, Discr. = Discrimination parameter of a generalized partial credit model, aQ_3 = Adjusted average absolute residual correlation for item (Yen, 1984, 1993).
Items at position 3, 4, and 26 were excluded from the analyses (see section 2). Percent correct scores are not informative for polytomous CMC and MA items and, thus, are not reported.

Table B.2

Step Parameters (with Standard Errors) in Standard Test for Special Schools

Item	Step 1	Step 2	Step 3	Step 4
reg9016s_c	-0.25 (0.16)	-0.28 (0.18)	0.15 (0.28)	0.39
reg9017s_c	1.62 (0.30)	-1.62		
reg9033s_c	0.23 (0.18)	0.51 (0.23)	-0.73	

Note. The last step parameter is a constrained parameter for model identification and, thus, has no standard error.

Table B.3

Item Parameters for Standard Test in General Schools

	Item	Pos.	Percentage correct	ξ	$SE_ξ$	WMNSQ	*t*	Item-rest correlation	Discr.	aQ_3
1	reg90110_c	1	89.12	-2.30	0.15	1.00	0.06	0.19	0.73	0.04
2	reg90120_c	2	95.19	-3.24	0.21	0.97	-0.11	0.21	1.21	0.05
3	reg90150_c	5	58.25	-0.37	0.10	0.99	-0.29	0.30	0.79	0.05
4	reg9016s_c	6		-0.23	0.04	1.18	2.93	0.40	0.43	0.08
5	reg9017s_c	7		-1.39	0.08	1.17	1.71	0.14	0.32	0.05
6	reg90210_c	8	70.49	-0.97	0.10	0.96	-0.95	0.35	1.00	0.04
7	reg90220_c	9	40.37	0.44	0.10	1.02	0.49	0.22	0.67	0.04
8	reg90230_c	10	78.41	-1.44	0.11	0.96	-0.68	0.32	0.98	0.04
9	reg90240_c	11	63.31	-0.61	0.10	0.96	-0.98	0.31	0.97	0.05
10	reg90250_c	12	28.69	1.03	0.10	1.11	2.26	0.11	0.20	0.04
11	reg90310_c	13	67.89	-0.84	0.10	1	0.05	0.28	0.87	0.06
12	reg90320_c	14	80.20	-1.56	0.12	0.91	-1.39	0.40	1.71	0.06
13	reg9033s_c	15		-1.16	0.06	1.01	0.18	0.41	0.70	0.06
14	reg90340_c	16	74.75	-1.20	0.11	0.89	-2.10	0.45	1.68	0.06
15	reg90350_c	17	78.03	-1.41	0.11	0.95	-0.86	0.33	1.21	0.05
16	reg90360_c	18	66.73	-0.78	0.10	1.04	0.89	0.21	0.58	0.05
17	reg90370_c	19	51.65	-0.06	0.10	1.04	1.33	0.21	0.58	0.05
18	reg90410_c	20	76.69	-1.34	0.11	0.97	-0.50	0.32	0.93	0.05
19	reg90420_c	21	54.62	-0.22	0.10	0.95	-1.62	0.37	1.13	0.06
20	reg90430_c	22	46.14	0.16	0.10	1.01	0.19	0.27	0.76	0.06
21	reg90440_c	23	58.09	-0.38	0.10	0.94	-1.89	0.38	1.17	0.07
22	reg90450_c	24	67.66	-0.85	0.11	0.95	-1.06	0.36	1.17	0.06
23	reg90460_c	25	45.56	0.18	0.10	0.99	-0.21	0.29	0.74	0.04

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, Discr. = Discrimination parameter of a generalized partial credit model, aQ_3 = Adjusted average absolute residual correlation for item (Yen, 1984, 1993).
Items at position 3, 4, and 26 were excluded from the analyses (see section 2). Percent correct scores are not informative for polytomous CMC and MA items and, thus, are not reported.

Table B.4

Step Parameters (with Standard Errors) in Standard Test for General Schools

Item	Step 1	Step 2	Step 3	Step 4
reg9016s_c	-0.52 (0.10)	0.07 (0.10)	-0.12 (0.11)	0.57
reg9017s_c	1.22 (0.18)	-1.22		
reg9033s_c	-0.13 (0.11)	0.59 (0.13)	-0.47	

Note. The last step parameter is a constrained parameter for model identification and, thus, has no standard error.

Table B.5

Item Parameters for Easy Test

	Item	Pos.	Percentage correct	ξ	SE_ξ	WMNSQ	*t*	Item-rest correlation	Discr.	aQ_3
1	reg90110_c	1	77.71	-1.41	0.19	0.99	-0.11	0.26	0.98	0.06
2	reg90120_c	2	88.08	-2.23	0.23	0.97	-0.14	0.25	1.30	0.07
3	reg90150_c	3	35.87	0.66	0.16	0.96	-0.68	0.23	1.08	0.07
4	reg9016s_c	4		0.50	0.07	0.91	-0.74	0.39	0.59	0.06
5	reg9017s_c	5		-0.96	0.11	0.98	-0.13	0.15	0.53	0.06
6	reg90210_c	6	46.37	0.18	0.16	0.96	-0.78	0.25	0.98	0.07
7	reg90220_c	7	35.03	0.69	0.17	1.04	0.65	0.13	0.61	0.09
8	reg90230_c	8	46.93	0.14	0.16	0.93	-1.38	0.27	1.35	0.06
9	reg90240_c	9	40.22	0.45	0.16	1.11	1.80	0.16	0.40	0.06
10	reg90310_c	10	45.98	0.16	0.16	1.00	0.07	0.27	0.74	0.07
11	reg90320_c	11	50.29	-0.03	0.16	0.89	-2.10	0.42	1.51	0.06
12	reg9033s_c	12		-0.61	0.08	0.96	-0.40	0.35	0.54	0.08
13	reg90340_c	13	57.65	-0.39	0.17	1.00	-0.06	0.30	0.95	0.06
14	reg90350_c	14	59.64	-0.46	0.17	0.96	-0.61	0.27	1.05	0.08
15	reg90360_c	15	60.48	-0.52	0.17	1.05	0.81	0.16	0.58	0.05
16	reg90370_c	16	35.22	0.69	0.18	1.11	1.52	0.06	0.37	0.06
17	reg90410_c	17	70.42	-1.05	0.19	1.03	0.33	0.18	0.69	0.07
18	reg90420_c	18	39.71	0.40	0.19	1.04	0.54	0.24	0.58	0.06
19	reg90440_c	19	46.88	0.05	0.19	1.04	0.58	0.26	0.68	0.05
20	reg90450_c	20	63.49	-0.71	0.20	1.10	1.39	0.12	0.30	0.06

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, Discr. = Discrimination parameter of a generalized partial credit model, aQ_3 = Adjusted average absolute residual correlation for item (Yen, 1984, 1993).
Item at position 21 was excluded from the analyses (see section 2). Percent correct scores are not informative for polytomous CMC and MA items and, thus, are not reported.

Table B.6

Step Parameters (with Standard Errors) in Easy Test

Item	Step 1	Step 2	Step 3	Step 4
reg9016s_c	-0.56 (0.16)	0.53 (0.20)	0.12 (0.30)	-0.09
reg9017s_c	1.06 (0.25)	-1.06		
reg9033s_c	0.37 (0.18)	0.65 (0.24)	-1.02	

Note. The last step parameter is a constrained parameter for model identification and, thus, has no standard error.

Table B.7

Item Parameters for Easy Test (Reversed)

	Item	Pos.	Percentage correct	ξ	$SE_ξ$	WMNSQ	*t*	Item-rest correlation	Discr.	aQ_3
1	reg90410_c	1	75.58	-1.25	0.19	1.03	0.32	0.08	0.52	0.08
2	reg90420_c	2	54.71	-0.22	0.16	0.91	-1.81	0.31	1.24	0.08
3	reg90440_c	3	63.07	-0.61	0.16	1.02	0.41	0.14	0.64	0.09
4	reg90450_c	4	74.72	-1.20	0.18	0.96	-0.48	0.26	1.02	0.06
5	reg90310_c	6	41.52	0.40	0.16	0.94	-1.07	0.20	1.06	0.08
6	reg90320_c	7	65.50	-0.72	0.17	0.90	-1.64	0.35	1.60	0.08
7	reg9033s_c	8		-0.84	0.09	1.02	0.15	0.14	0.35	0.06
8	reg90340_c	9	57.99	-0.37	0.16	0.89	-2.21	0.37	1.73	0.08
9	reg90350_c	10	64.50	-0.68	0.17	1.04	0.59	0.09	0.55	0.06
10	reg90360_c	11	56.74	-0.31	0.16	1.12	2.29	0.12	0.16	0.08
11	reg90370_c	12	42.94	0.34	0.17	1.02	0.39	0.17	0.67	0.09
12	reg90210_c	13	45.28	0.22	0.17	0.90	-2.05	0.35	1.70	0.10
13	reg90220_c	14	33.55	0.77	0.18	1.05	0.70	0.11	0.63	0.08
14	reg90230_c	15	50.97	-0.05	0.17	0.99	-0.17	0.25	0.71	0.07
15	reg90240_c	16	36.60	0.62	0.18	1.05	0.79	0.15	0.42	0.06
16	reg90110_c	17	78.57	-1.47	0.21	1.02	0.23	0.13	0.53	0.07
17	reg90120_c	18	73.83	-1.16	0.19	0.98	-0.15	0.22	0.89	0.07
18	reg90150_c	19	40.00	0.43	0.19	1.00	0.00	0.27	0.67	0.08
19	reg9016s_c	20		0.74	0.09	1.03	0.24	0.22	0.31	0.07
20	reg9017s_c	21		-0.63	0.12	1.16	1.48	-0.10	-0.23	0.10

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, Discr. = Discrimination parameter of a generalized partial credit model, aQ_3 = Adjusted average absolute residual correlation for item (Yen, 1984, 1993).
Item at position 5 was excluded from the analyses (see section 2). Percent correct scores are not informative for polytomous CMC and MA items and, thus, are not reported.

Table B.8

Step Parameters (with Standard Errors) in Easy Test (Reversed)

Item	Step 1	Step 2	Step 3	Step 4
reg9033s_c	0.47 (0.21)	0.66 (0.27)	-1.13	
reg9016s_c	-0.63 (0.15)	0.57 (0.25)	-0.30 (0.35)	0.35
reg9017s_c	1.36 (0.32)	-1.36		

Note. The last step parameter is a constrained parameter for model identification and, thus, has no standard error.

Table B.9

Item Parameters for Out-of-Level Test

	Item	Pos.	Percentage correct	ξ	$SE_ξ$	WMNSQ	*t*	Item-rest correlation	Discr.	aQ_3
1	reg90610_c	1	57.30	-0.34	0.16	0.94	-1.11	0.36	1.33	0.08
2	reg90620_c	2	38.50	0.54	0.16	0.95	-0.85	0.24	1.2	0.07
3	reg90630_c	3	47.45	0.11	0.16	0.99	-0.26	0.28	1.00	0.06
4	reg90640_c	4	49.47	0.03	0.16	0.98	-0.28	0.29	1.01	0.07
5	reg90650_c	5	45.13	0.23	0.16	0.98	-0.27	0.29	1.06	0.08
6	reg90660_c	6	47.69	0.11	0.16	0.95	-1.01	0.34	1.22	0.06
7	reg90210_c	7	44.50	0.26	0.16	0.99	-0.21	0.32	0.93	0.08
8	reg90220_c	8	28.57	1.10	0.17	1.12	1.42	0.15	0.46	0.05
9	reg90230_c	9	52.88	-0.13	0.16	0.98	-0.29	0.33	1.00	0.06
10	reg90240_c	10	37.77	0.59	0.16	1.06	0.96	0.16	0.63	0.08
11	reg90250_c	11	21.93	1.48	0.19	1.07	0.66	0.11	0.6	0.06
12	reg90310_c	12	44.15	0.26	0.16	0.97	-0.50	0.35	0.97	0.07
13	reg90320_c	13	56.42	-0.34	0.16	0.91	-1.54	0.43	1.37	0.06
14	reg9033s_c	14		-0.73	0.08	0.95	-0.40	0.31	0.68	0.06
15	reg90340_c	15	60.48	-0.56	0.17	1.00	0.04	0.33	1.00	0.08
16	reg90350_c	16	56.97	-0.39	0.17	1.04	0.65	0.27	0.72	0.07
17	reg90360_c	17	54.07	-0.24	0.17	1.20	3.22	0.07	0.2	0.07
18	reg90370_c	18	41.82	0.35	0.17	1.06	0.96	0.24	0.68	0.08
19	reg90720_c	20	45.14	0.13	0.18	0.82	-2.84	0.60	2.27	0.10
20	reg90730_c	21	48.91	-0.10	0.19	0.97	-0.42	0.36	1.09	0.08
21	reg90740_c	22	41.61	0.26	0.19	1.00	-0.04	0.36	0.93	0.06
22	reg90750_c	23	44.00	0.12	0.20	1.01	0.09	0.35	0.95	0.05
23	reg9076s_c	24		0.46	0.11	1.01	0.14	0.33	0.47	0.08

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, Discr. = Discrimination parameter of a generalized partial credit model, aQ_3 = Adjusted average absolute residual correlation for item (Yen, 1984, 1993).
Item at position 19 was excluded from the analyses (see section 2). Percent correct scores are not informative for polytomous CMC and MA items and, thus, are not reported.

Table B.10

Step Parameters (with Standard Errors) in Out-of-Level Test

Item	Step 1	Step 2	Step 3
reg9033s_c	0.37 (0.17)	0.08 (0.20)	-0.45
reg9076s_c	-0.28 (0.20)	-0.66 (0.22)	0.94

Note. The last step parameter is a constrained parameter for model identification and, thus, has no standard error.

Table B.11

Item Parameters for Out-of-Level Test (reversed)

	Item	Pos.	Percentage correct	ξ	$SE_ξ$	WMNSQ	*t*	Item-rest correlation	Discr.	aQ_3
1	reg9081s_c	1		-0.44	0.08	0.93	-0.78	0.34	0.72	0.08
2	reg90820_c	2	84.70	-1.96	0.22	0.98	-0.08	0.22	1.03	0.08
3	reg90830_c	3	54.14	-0.15	0.16	0.93	-1.22	0.40	1.35	0.07
4	reg9084s_c	4		-0.38	0.08	1.01	0.11	0.23	0.5	0.06
5	reg90850_c	5	58.14	-0.36	0.17	1.13	1.98	0.12	0.40	0.07
6	reg90860_c	6	56.22	-0.27	0.16	1.06	0.96	0.26	0.63	0.08
7	reg90720_c	8	53.26	-0.14	0.16	0.93	-1.18	0.39	1.44	0.09
8	reg90730_c	9	58.10	-0.39	0.16	1.00	-0.02	0.34	0.93	0.06
9	reg90740_c	10	55.25	-0.24	0.16	0.93	-1.23	0.38	1.21	0.09
10	reg90750_c	11	49.44	0.02	0.16	1.00	-0.03	0.33	0.96	0.06
11	reg9076s_c	12		0.33	0.08	1.02	0.27	0.27	0.43	0.06
12	reg90310_c	13	46.47	0.17	0.17	1.10	1.56	0.22	0.56	0.08
13	reg90320_c	14	58.93	-0.44	0.17	0.91	-1.40	0.44	1.55	0.09
14	reg9033s_c	15		-0.58	0.08	0.96	-0.37	0.36	0.60	0.08
15	reg90340_c	16	57.05	-0.37	0.18	0.95	-0.85	0.42	1.32	0.08
16	reg90350_c	17	56.41	-0.33	0.18	0.88	-1.90	0.51	1.61	0.07
17	reg90360_c	18	59.62	-0.49	0.18	1.10	1.48	0.23	0.46	0.09
18	reg90370_c	19	34.00	0.76	0.19	1.11	1.38	0.21	0.47	0.06
19	reg90210_c	20	39.57	0.39	0.19	0.93	-0.94	0.42	1.38	0.10
20	reg90220_c	21	26.67	1.07	0.21	1.04	0.38	0.24	0.73	0.07
21	reg90230_c	22	51.59	-0.19	0.19	0.99	-0.08	0.35	1.03	0.09
22	reg90240_c	23	37.29	0.50	0.21	1.13	1.46	0.19	0.40	0.12
23	reg90250_c	24	18.10	1.62	0.26	1.12	0.80	0.18	0.40	0.08

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, Discr. = Discrimination parameter of a generalized partial credit model, aQ_3 = Adjusted average absolute residual correlation for item (Yen, 1984, 1993).
Item at position 7 was excluded from the analyses (see section 2). Percent correct scores are not informative for polytomous CMC and MA items and, thus, are not reported.

Table B.12

Step Parameters (with Standard Errors) in Out-of-Level Test (Reversed)

Item	Step 1	Step 2	Step 3
reg9081s_c	-0.75 (0.16)	0.11 (0.16)	0.64
reg9084s_c	-0.17 (0.15)	-0.43 (0.16)	0.60
reg9076s_c	-0.45 (0.16)	-0.07 (0.18)	0.52
reg9033s_c	0.70 (0.20)	0.53 (0.27)	-1.23

Note. The last step parameter is a constrained parameter for model identification and, thus, has no standard error.

Appendix C: R-Syntax for estimating WLEs in Grade 9 of special schools for Starting Cohort 4

```
# load packages
library(haven) # to import SPSS files
library(TAM)   # for IRT analyses

# load competence data
dat <- read_sav("SC4_xTargetSpecialNeedsCompetencies.sav")

# items of reading competence tests
items <- c("reg90110 c", "reg90120 c", "reg90150 c",
           "reg9016s c", "reg9017s c", "reg90210 c",
           "reg90220 c", "reg90230 c", "reg90240 c",
           "reg90250 c", "reg90310 c", "reg90320 c",
           "reg9033s c", "reg90340 c", "reg90350 c",
           "reg90360 c", "reg90370 c", "reg90410 c",
           "reg90420 c", "reg90430 c", "reg90440 c",
           "reg90450 c", "reg90460 c", "reg90610 c",
           "reg90620 c", "reg90630 c", "reg90640 c",
           "reg90650 c", "reg90660 c", "reg90720 c",
           "reg90730 c", "reg90740 c", "reg90750 c",
           "reg9076s_c")

# define Q-matrix for 0.5 scoring of PCM
Q <- matrix(1, nrow = length(items), ncol = 1)
Q[c(4, 5, 13, 34), 1] <- 0.5    # score of 0.5

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

# item fit
tam.fit(mod)

# WLE
tam.wle(mod)
```