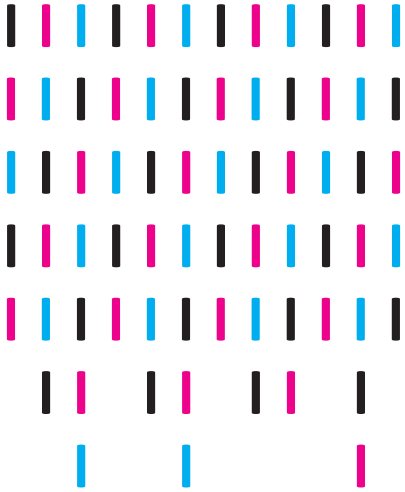

NEPS SURVEY PAPERS

Timo Gnambs

NEPS TECHNICAL REPORT FOR READING: SCALING RESULTS OF STARTING COHORT 4 FOR WAVE 10 IN SPECIAL SCHOOLS

NEPS Survey Paper No. 64
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 Wave 10 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 Wave 10 in Special Schools* (NEPS Survey Paper No. 64). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Acknowledgements:

This report is an extension to NEPS Survey Paper 63 (Gnambs, 2020) that presents the scaling results for reading competence of Starting Cohort 4 for Grade 9 in special schools. Therefore, various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* to facilitate the understanding of the presented results.

NEPS Technical Report for Reading: Scaling Results of Starting Cohort 4 for Wave 10 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 Wave 10 of Starting Cohort 4 (ninth grade) that was administered to former students in special schools. These analyses are part of a feasibility study to evaluate whether students with special educational needs can be included in large-scale assessments such as the NEPS. To this end, the same reading competence test that was previously administered in Grade 9 of special schools was again administered to these respondents at age 21. The present sample includes 293 respondents (46% women) previously attending special schools. Of these, *N* = 159 (44% women) had also finished the respective test in Grade 9. The responses of the sample were scaled using the partial credit model. Item fit statistics, differential item functioning, and Rasch-homogeneity were evaluated to examine the quality of the test. These analyses showed that the competence test was too long for the respondents; items at the end of the administered tests were finished by rather few participants, resulting in large missing rates. However, the test exhibited an acceptable variance and reliability, thus, allowing for analyses of interindividual differences between respondents with special educational needs. Importantly, there was substantial differential item functioning between the tests administered in Grade 9 and in Wave 10, making longitudinal analyses of reading competences challenging. Nevertheless, linked person scores were estimated that allow the examination of change trajectories in reading competence for former students in special schools. Overall, these results highlight substantial difficulties in assessing reading competence among respondents with special educational needs 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 types across Germany. In Grade 9 of Starting Cohort 4 (ninth grade), a feasibility study was conducted to evaluate whether students from special schools might be included in the NEPS (see Gnambs, 2020). A subsample of these respondents were assessed in a repeated measurement design and also provided responses to a reading competence test in Wave 10 of Starting Cohort 4. In this paper the results of the repeated test administration of a reading competence test are summarized. First, the main concepts of the reading competence test 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.

The reading competence test included 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 test 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 test 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 reading competence and general cognitive functioning (cf. Gnambs & Nusser, 2019). For each participant, the reading test was administered as the first test. There was no multi-matrix design regarding the order of the items *within* the test. All participants received the test items in the same order.

The administered reading competence test was identical to the standard test administered in Grade 9 to students from special schools (see Gnambs, 2020) and students from general schools (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 reg90140_c and reg9047s_c. Therefore, these items were excluded from the analyses, resulting in a test with four texts including 24 items. 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.

The tests were administered individually to respondents by trained interviewers at the respondents' private homes. 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 1

Number of Items for the Different Text Types

Text types	**Frequency**
Information text	6
Instruction text	5
Advertising text	7
Commenting text	0
Literary text	6
Total number of items	24

Table 2

Number of Items for the Cognitive Requirements

Cognitive requirements	**Frequency**
Finding information	8
Drawing text-related conclusions	11
Reflecting and assessing	5
Total number of items	24

Table 3

Number of Items for the Different Response Formats

Response format	**Frequency**
Simple multiple choice items	21
Complex multiple choice items	2
Matching items	1
Total number of items	24

2.2 Sample

A sample of 293[1] participants (46% women) that previously had attended special schools received the reading competence tests. All respondents provided at least three valid item responses and, thus, were included in the scaling procedure (see Pohl & Carstensen, 2012). The mean age of the sample was 21.93 years (*SD* = 0.60). About 15% of them had a migration background.

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

[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.

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 test was specifically constructed to be implemented in the NEPS. In order to ensure appropriate psychometric properties, the quality of the test was examined in several analyses.

The MC items consisted of one correct response option and 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.

For longitudinal comparisons, the reading competence test should measure the same construct at each measurement occasion. If the order of item difficulties changed over time, measurement invariance would be violated and a comparison of competence scores across time would be biased and, thus, unfair. For the present study, longitudinal differential item functioning (DIF) was examined by examining the differences in item difficulties following Fischer, Rohm, Gnambs, & Carstensen (2016). We considered absolute standardized differences in estimated difficulties between the measurement occasions 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 et al., 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.

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. Overall, there were very few invalid responses. Most respondents (88%) did not have any invalid response at all. More than one invalid response was observed for less than 6% of the sample.

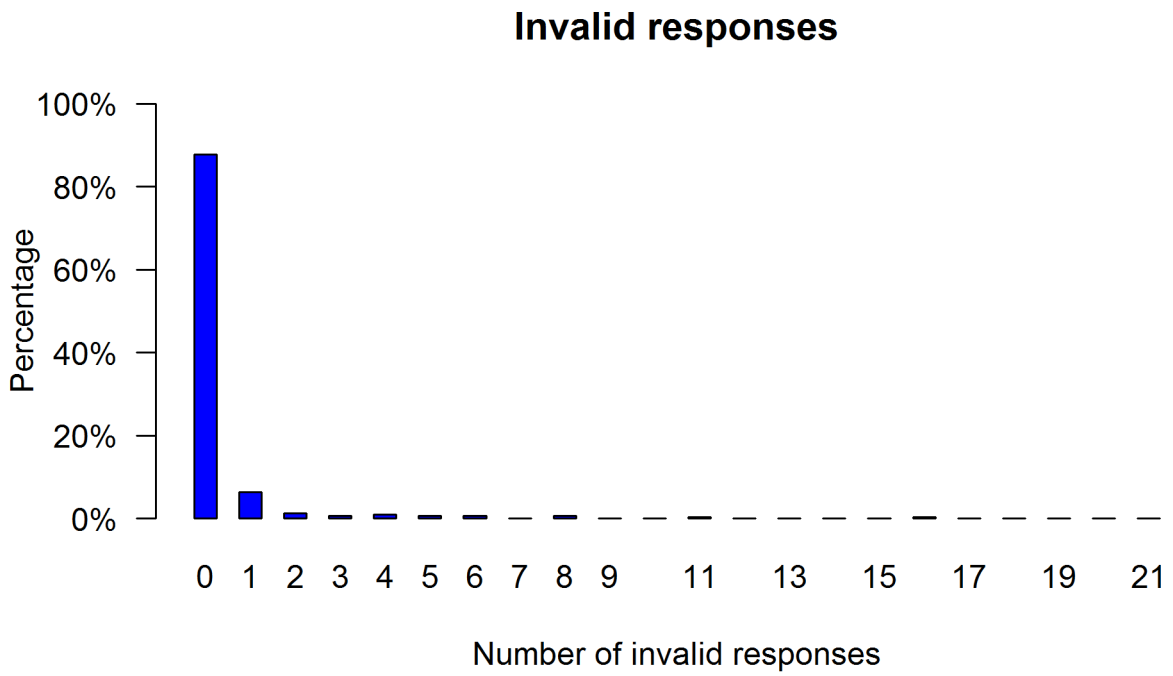

Figure 1. Number of invalid responses

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 three (see Table 3). However, there were no substantial missing responses that were not determinable. Only 0.34% of the respondents had not-determinable missing responses.

Missing responses also occurred when respondents omitted items. As illustrated in Figure 2, most respondents (72%) did not skip any item and about 11% omitted two or more items.

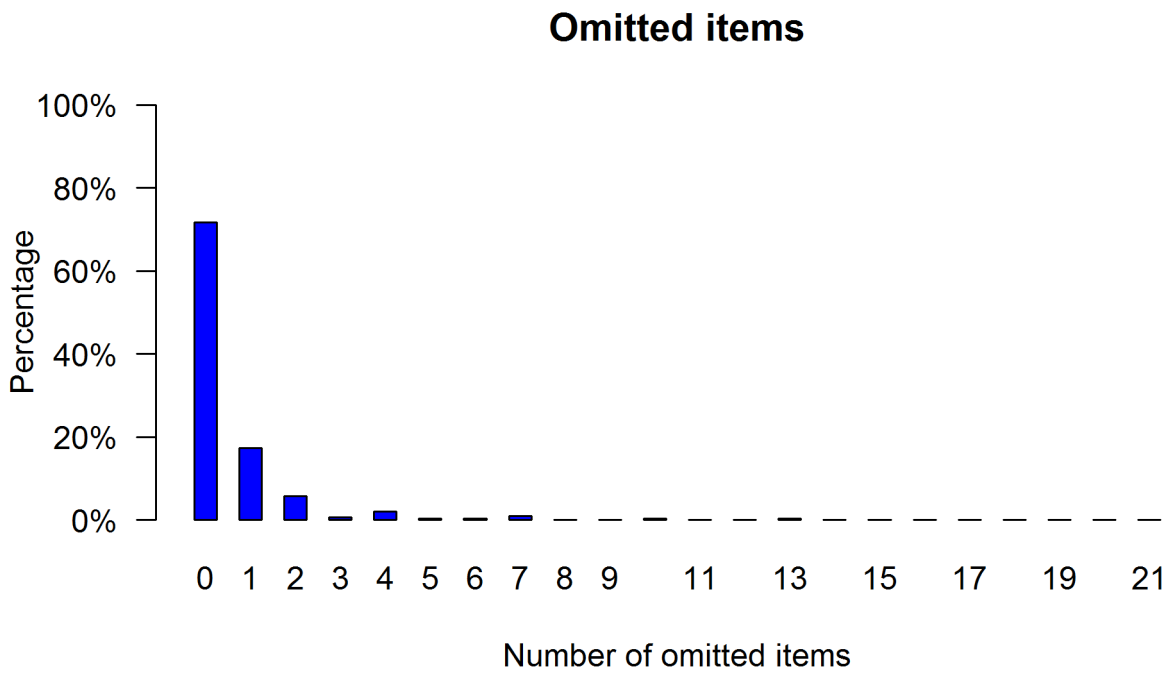

Figure 2. Number of omitted items

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 participants 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). About 64% of the respondents finished all items referring to these four texts (Figure 3). Thus, despite the shortened test a substantial proportion of the respondents did not reach the end of the test.

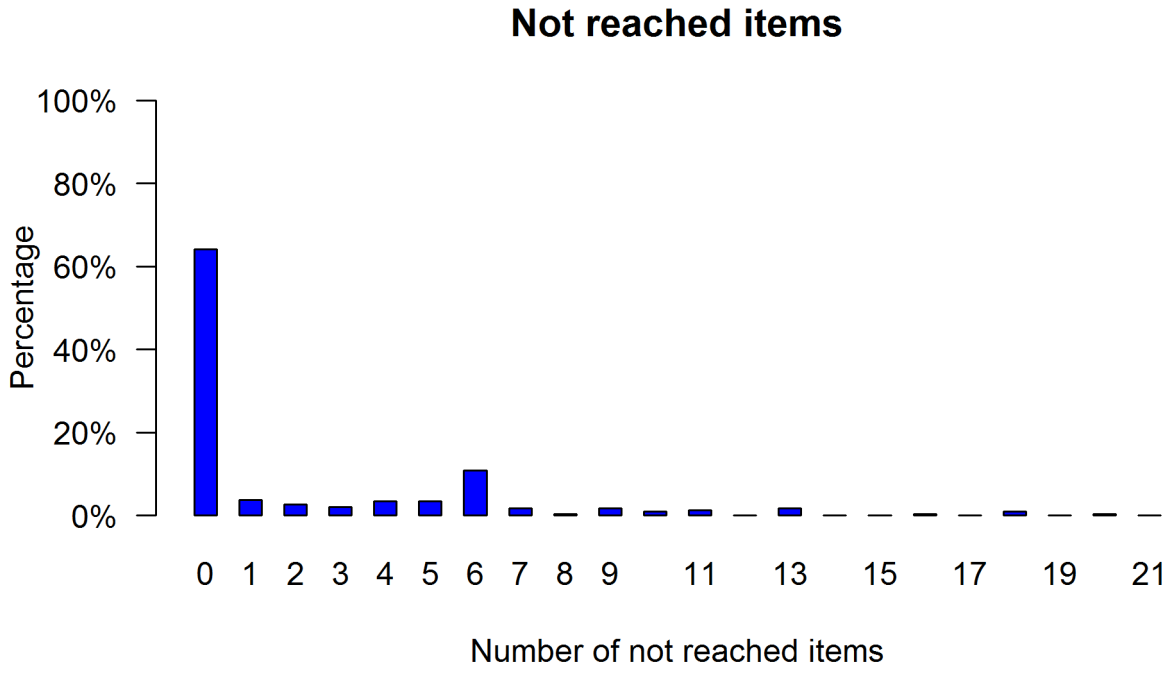

Figure 3. Number of not-reached items

The total number of missing responses, aggregated over invalid, omitted, not-reached, and not-determinable missing responses per person, is illustrated in Figure 4. Most participants had a rather large amount of missing values. Only 41% of them had no missing response at all, whereas about 31% had five or more missing responses.

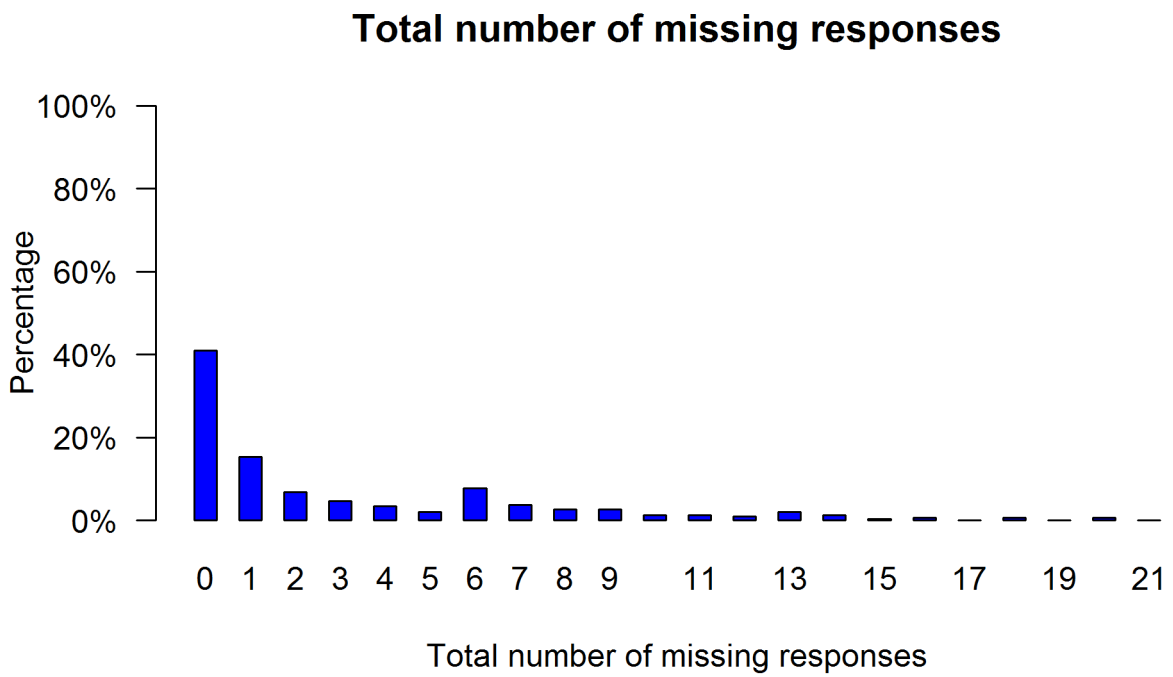

Figure 4. Total number of missing responses

In sum, the respondents 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. Notably, these results mirror the missing pattern previously reported in Grade 5 for students in special schools (Gnambs, 2020). Despite the higher age of the respondents, similar difficulties with the reading competence tests were observed. These results indicate that for participants previously attending special schools competence tests need to be substantially shorter for these respondents 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

Table 4 provides information on the occurrence of different kinds of missing responses per item. Generally, the omission rates were rather low for most items; the median percentage of omitted responses across items fell at 2.6%. However, polytomous CMC and MA items exhibited substantially larger omission rates up to 12%. 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 respondents had to familiarize themselves with the response formats of the tests. With an item's progressing position in the tests, the number of respondents that did not reach an item (columns "NR" in

Table 4) rose up to 36% (see Figure 5). Thus, for respondents with special educational needs the available testing time seemed to be too short.

Table 4

Percentage of Missing Values by Item

Item	*N*	NR	OM	NV	ND
reg90110_c	267	0.0	1.7	7.2	0.0
reg90120_c	293	0.0	0.0	0.0	0.0
reg90130_c	284	0.0	2.0	1.0	0.0
reg90150_c	283	0.0	1.7	1.7	0.0
reg9016s_c	268	0.3	7.8	0.3	0.0
reg9017s_c	258	0.3	11.6	0.0	0.0
reg90210_c	274	1.4	0.3	4.8	0.0
reg90220_c	274	1.4	1.0	4.1	0.0
reg90230_c	282	1.7	1.4	0.7	0.0
reg90240_c	283	1.7	0.7	1.0	0.0
reg90250_c	282	1.7	1.0	1.0	0.0
reg90310_c	269	3.4	2.7	2.0	0.0
reg90320_c	273	3.4	1.0	2.4	0.0
reg9033s_c	262	4.8	5.5	0.0	0.3
reg90340_c	265	5.8	1.4	2.4	0.0
reg90350_c	256	7.5	2.0	3.1	0.0
reg90360_c	265	7.8	1.4	0.3	0.0
reg90370_c	258	9.6	1.4	1.0	0.0
reg90410_c	226	20.5	2.0	0.3	0.0
reg90420_c	213	23.9	2.7	0.7	0.0
reg90430_c	202	27.3	3.1	0.7	0.0
reg90440_c	196	29.4	2.7	1.0	0.0
reg90450_c	193	32.1	1.4	0.7	0.0
reg90460_c	181	35.8	1.4	1.0	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.

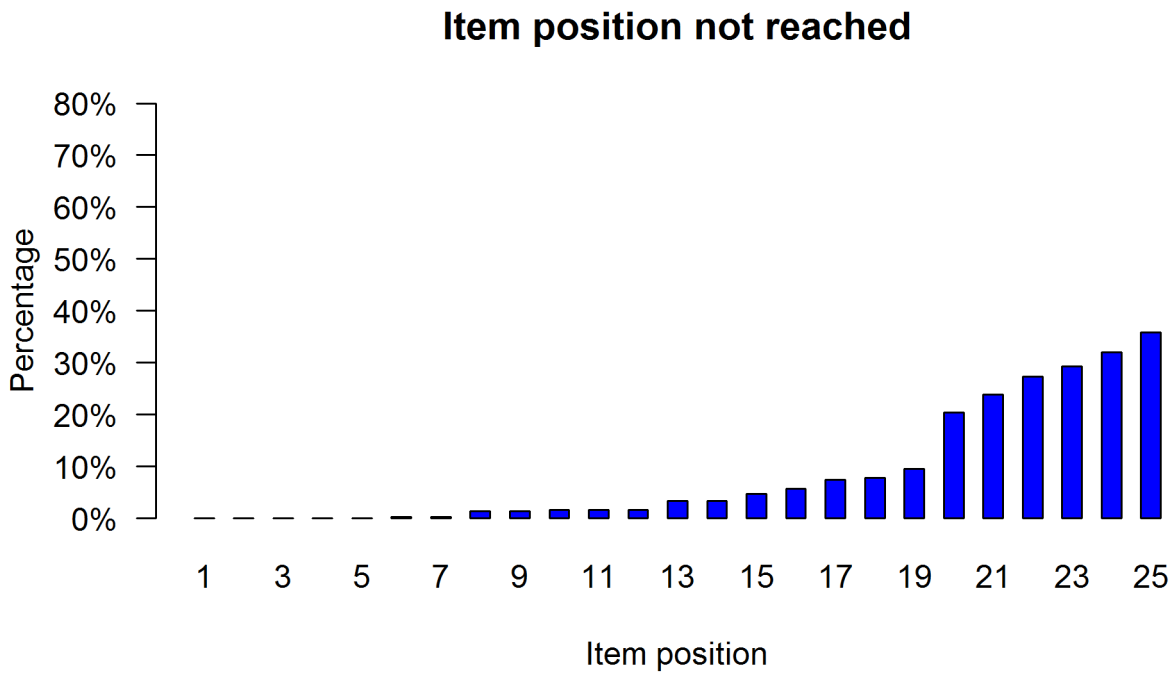

Figure 5. Item position not reached

4.2 Quality of the Test

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 respondents' total correct scores were calculated. The median point-biserial correlations for the distractors fell at -.22 (*Min* = -.37, *Max* = .00). In contrast, the correlations of the correct responses with the total scores varied between.25 and .54 (*Mdn* = .42). These results indicate that the distractors functioned well.

4.2.2 Item parameters

In Table 5 the percentage of correct responses (for simple multiple choice items) in relation to all valid responses are presented for each item. Because there was a non-negligible amount of missing responses, these probabilities cannot be readily interpreted as an index of item difficulty. The percentage of correct responses varied between 25% and 89% with an average of 61% (*SD* = 17%) correct responses.

The item parameters of the reading competence test are summarized in Table 5, whereas the step parameters are given in Table 6. The item difficulties (for dichotomous variables) and location parameters (for polytomous variables) were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties and location parameters ranged from -2.4 (item reg90110_sc4a10_c) to 1.3 (item reg90250_sc4a10_c) with a median of -0.9 and, thus, covered a rather broad range. However, the standard errors (SE) of the estimated parameters were rather large with a *Mdn* = 0.15 and a range of [0.05, 0.22]. Thus, the reported item parameters had a somewhat limited precision.

Table 5

Item Parameters

	Item	Pos.	*N*	Percentage correct	ξ	SE_ξ	WMNSQ	*t*	Item-rest correlation	Discr.	aQ_3
1	reg90110_sc4a10_c	1	267	89.14	-2.41	0.21	1.05	0.41	0.13	0.66	0.06
2	reg90120_sc4a10_c	2	293	87.03	-2.21	0.18	0.94	-0.46	0.27	1.52	0.06
3	reg90130_sc4a10_c	3	284	43.31	0.33	0.13	1.00	-0.06	0.33	0.97	0.06
4	reg90150_sc4a10_c	5	283	47.00	0.16	0.13	1.05	0.97	0.18	0.78	0.05
5	reg9016s_sc4a10_c	6	268		0.24	0.05	0.99	-0.06	0.35	0.47	0.06
6	reg9017s_sc4a10_c	7	258		-1.27	0.11	0.99	-0.04	0.16	0.56	0.06
7	reg90210_sc4a10_c	8	274	74.82	-1.28	0.15	0.92	-1.18	0.39	1.68	0.07
8	reg90220_sc4a10_c	9	274	41.24	0.42	0.13	1.04	0.72	0.28	0.88	0.06
9	reg90230_sc4a10_c	10	282	80.14	-1.62	0.16	1.00	0.01	0.25	0.99	0.08
10	reg90240_sc4a10_c	11	283	53.36	-0.16	0.13	0.98	-0.53	0.38	1.11	0.07
11	reg90250_sc4a10_c	12	282	25.18	1.28	0.15	1.07	0.91	0.18	0.66	0.05
12	reg90310_sc4a10_c	13	269	66.54	-0.81	0.14	0.95	-0.86	0.34	1.28	0.07
13	reg90320_sc4a10_c	14	273	84.25	-1.96	0.18	0.88	-1.12	0.39	2.18	0.07
14	reg9033s_sc4a10_c	15	262		-1.04	0.07	0.93	-0.74	0.33	0.69	0.08
15	reg90340_sc4a10_c	16	265	71.70	-1.11	0.15	0.92	-1.25	0.36	1.50	0.06
16	reg90350_sc4a10_c	17	256	68.75	-0.94	0.15	0.96	-0.70	0.33	1.14	0.06
17	reg90360_sc4a10_c	18	265	62.26	-0.62	0.14	1.14	2.55	0.10	0.38	0.05
18	reg90370_sc4a10_c	19	258	48.45	0.05	0.14	1.07	1.39	0.22	0.62	0.05
19	reg90410_sc4a10_c	20	226	71.68	-1.19	0.16	1.04	0.51	0.25	0.76	0.04
20	reg90420_sc4a10_c	21	213	52.11	-0.21	0.15	1.01	0.27	0.30	0.84	0.06
21	reg90430_sc4a10_c	22	202	42.57	0.22	0.16	1.05	0.82	0.29	0.72	0.05
22	reg90440_sc4a10_c	23	196	65.31	-0.89	0.16	0.96	-0.59	0.39	1.17	0.07
23	reg90450_sc4a10_c	24	193	70.98	-1.21	0.17	1.06	0.79	0.26	0.74	0.09
24	reg90460_sc4a10_c	25	181	44.20	0.09	0.16	1.07	1.10	0.25	0.59	0.06

Note. Pos. = Item position, *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.
Items at positions 4 and 26 were excluded due to poor item fit.

Table 6

Step Parameters (with Standard Errors)

Item	Step 1	Step 2	Step 3	Step 4
reg9016s_c	-0.35 (0.13)	0.09 (0.13)	0.21 (0.18)	0.05
reg9017s_c	1.33 (0.25)	-1.33		
reg9033s_c	-0.22 (0.14)	0.54 (0.17)	-0.32	

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

4.2.3 Item fit

Altogether, item fit can be considered to be good (see Table 5). The median value of the WMNSQ fell at 1.0, no item exhibited a considerable misfit greater than 1.15. Similar, the respective *t*-values indicated no substantial misfit ($|t| > 6$) at all (*Max* = 2.6). Overall, there was no indication of substantial item over- or underfit. The median correlation between the item scores and the total-rest scores was .29 (*Min* = .1, *Max* = .4) and, thus, suggested adequate item discriminations. Moreover, all item characteristic curves showed an acceptable fit of the items.

4.2.4 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 generalized partial credit model (GPCM) that estimates discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 5), ranging from 0.38 (item reg90360_sc4a10_c) to 2.18 (item reg90320_sc4a10_c). The median discrimination parameter fell at 0.81. Model fit indices suggested a slightly better model fit of the PCM (AIC =7,565, BIC = 7,679, number of parameters = 31) as compared to the GPCM (AIC = 7,548, BIC = 7,746, number of parameters = 54). 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.

4.2.5 Unidimensionality

The dimensionality of the test was investigated by evaluating the correlations between the residuals of the PCM. The adjusted Q_3 statistics (see Table 5) were quite low (*Mdn* = .06, *Min* = .04, *Max* = 0.09) 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.

4.2.6 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 reading competence test 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 -3.37 (item reg9033s_sc4a10_c) to 1.96 (item reg9016s_sc4a10_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 0.90, which implies adequate differentiation between students. The reliability of the test (EAP/PV reliability = .75, WLE reliability = .71) was satisfactory. The mean of the item distribution was about 0.67 logits below the mean person ability distribution. Thus, although the items covered a wide range of the ability distribution, the items were slightly too easy.

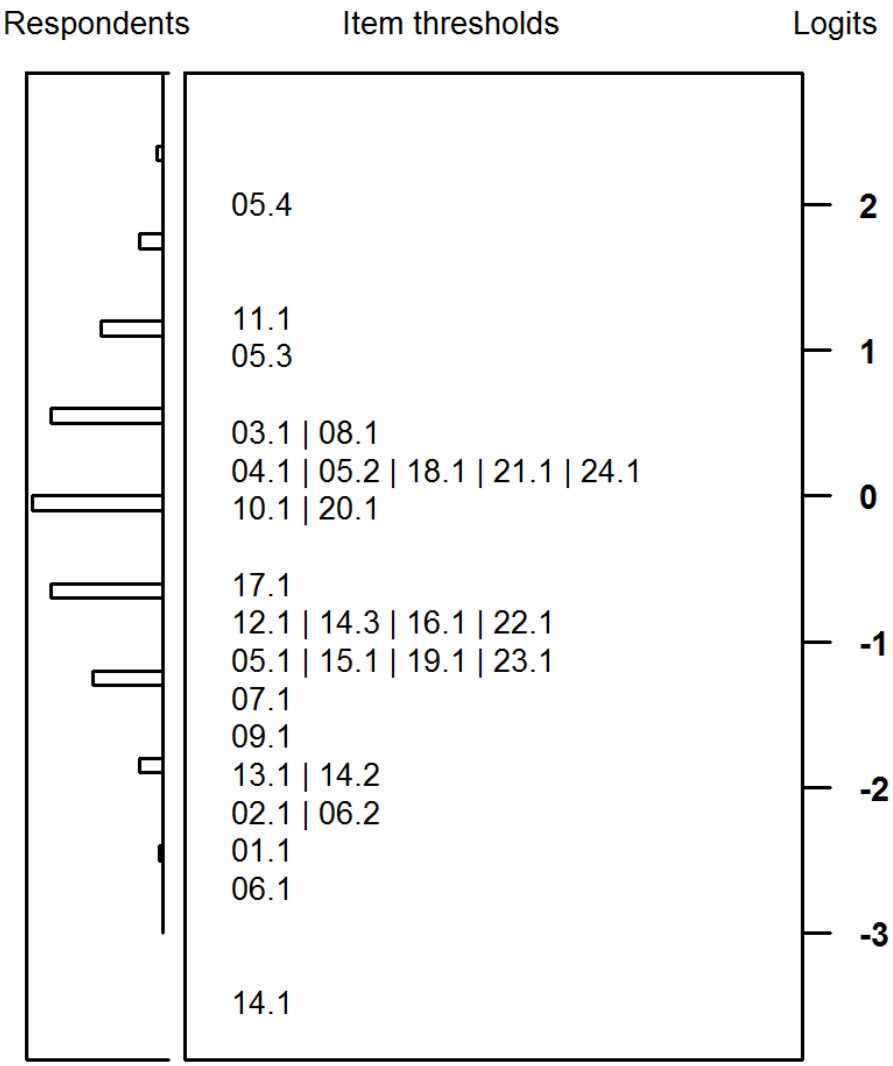

Figure 6. Test targeting. The distribution of person ability in the sample is given on the left-hand side of the graph. The category thresholds of the items are given on the right-hand side of the graph. Each number represents one threshold with the first part (before the dot) corresponding to the item number in Table 5 and the second part indicating the threshold.

4.3 Differential Item Functioning

In Grade 9 of Starting Cohort 4 different test versions were administered that included overlapping items with test administered in the present study (see Gnambs, 2020). This might allow examining changes in reading competence across the observational period. However, longitudinal comparisons require invariant measurements at both time points. Therefore, differential item functioning (DIF) analyses were conducted to evaluate test fairness and whether change analyses might be permissible. Therefore, it was tested whether the item parameters derived in the two waves showed a non-negligible shift in item difficulties. The differences in item difficulties between Grade 9 (= Wave 1) and Wave 10 and the tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 7. For these differences a positive value would indicate that the item was more difficult in Grade 9, whereas a negative value would highlight a lower difficulty in Grade 9.

Table 7

Differential Item Functioning Analyses between Waves 1 (Grade 9) and 10.

Item	*N*		Δσ	$SE_{\Delta\sigma}$	*F*
reg90110_sc4a10_c	100	148	0.61	0.39	2.48
reg90120_sc4a10_c	110	159	-0.68	0.47	2.08
reg90150_sc4a10_c	104	153	-0.10	0.28	0.12
reg9016s_sc4a10_c	90	144	0.08	0.13	0.43
reg9017s_sc4a10_c	99	140	0.20	0.21	0.87
reg90210_sc4a10_c	151	151	0.86	0.27	9.82
reg90220_sc4a10_c	147	148	0.01	0.27	0.00
reg90230_sc4a10_c	150	153	0.83	0.28	8.77
reg90240_sc4a10_c	153	151	0.65	0.26	6.11
reg90250_sc4a10_c	100	152	-0.25	0.34	0.54
reg90310_sc4a10_c	146	146	0.54	0.27	4.01
reg90320_sc4a10_c	142	148	1.29	0.32	16.12*
reg9033s_sc4a10_c	128	141	-0.08	0.15	0.31
reg90340_sc4a10_c	130	142	-0.08	0.29	0.08
reg90350_sc4a10_c	130	137	-0.05	0.29	0.04
reg90360_sc4a10_c	131	142	-0.75	0.28	6.84
reg90370_sc4a10_c	121	139	-0.10	0.28	0.13
reg90410_sc4a10_c	78	120	-0.32	0.35	0.82
reg90420_sc4a10_c	75	108	-0.11	0.34	0.12
reg90430_sc4a10_c	33	105	0.19	0.47	0.16
reg90440_sc4a10_c	66	101	0.93	0.36	6.50
reg90450_sc4a10_c	63	101	-0.38	0.38	0.96
reg90460_sc4a10_c	28	94	0.46	0.56	0.67

Note. *N* = Number of valid responses in the two waves; Δσ = Difference in item difficulty parameters between the two waves (negative values indicate easier items in Grade 9); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is F_{0154} (1, 157) = 10.53. A non-significant test indicates measurement invariance.

*$p < .05$

The minimum effects hypothesis test identified a lack of invariance for only one item (reg90320_sc4a10_c). However, given the small sample size this test might lack adequate power to identify meaningful differences. Descriptive analyses of the differences in item difficulties showed that only 11 items exhibited negligible differences smaller than 0.25. In contrast, nine items had considerable DIF exceeding 0.50. Taken together, these results indicate that the reading competence test seemed to have functioned rather differently at the two measurement occasions. At this point it remains unclear whether this difference is

attributable to the repeated measurement design and the age differences of the respondents. It is also conceivable that mode effects distorted the longitudinal comparisons to some degree because testing was conducted in a group setting in school classes in Grade 9, whereas respondents were assessed in an individual setting at their private homes in the present study.

5. Discussion

The presented analyses summarized information from a feasibility study to evaluate the possibility of including former students attending special schools in educational large-scale assessments such as the NEPS. The study administered the same test to a sample of young adults that was previously presented to them in Grade 9 of special schools. Overall, the analyses showed that the reading competence test represented an essentially unidimensional scale conforming to the partial credit model (Masters, 1982). Most items had satisfactory psychometric properties allowing the estimation of reading competence scores. Moreover, the population variance and the reliability of the test were satisfactory facilitating the analyses of interindividual differences in reading competence.

However, the results of the psychometric analyses also highlighted some challenges of administering standardized achievement tests to respondents with special educational needs. In line with previous reports (e.g., Gnambs, 2020), standard competence tests that are suitable for the general population seem to be too long for respondents with special educational needs. In the present study, a large number of not-reached items were observed because the respondents required more time for solving the test items. Even for the four reading texts analyzed in this study increased missing rates were observed for the last items in the test. Thus, future competence assessments for this target group need to make allowances for longer testing times.

Another limitation pertained to the substantial longitudinal differential item functioning. Less than half the administered items exhibited measurement invariance across time and might be used to link the two measurement waves to allow for change analyses. It might be suspected that competence measurements for respondents with special educational needs also capture a substantial degree of measurement error that limits the comparability of proficiency estimates over time. However, it is also conceivable that in the present study mode effects pertaining to the assessment setting contributed to the observed DIF effects. Unfortunately, in the present study these two components cannot be disentangled.

In conclusion, these results indicate that reading competences can be appropriately measured among former students from special schools, provided appropriate tests with shorter length are administered. However, even then longitudinal comparisons over time might not be feasible because the test lacks measurement invariance.

6. Data in the Scientific Use File

6.1 Naming conventions

The data in the SUF contains 33 items. Twenty-nine items were scored dichotomously (MC items) with 0 indicating an incorrect response and 1 indicating a correct response, whereas four 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 Linking of competence scores to Grade 9

In Starting Cohort 4, the same reading competence test was administered to students in special schools in Grade 9 and at age 21. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competences across grades, the two tests were linked using an anchor-item approach (cf. Fischer et al., 2016). Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. As indicated above (section 4.3), many items exhibited substantial DIF across the two measurement occasions. Therefore, only items were acknowledged that had at least 100 valid responses at each wave and negligible DIF less than 0.25. This resulted in seven items that were used to link the two tests across time using the mean/mean approach (see Fischer et al., 2016). The correction term was calculated as c = 0.25. This correction term was subsequently added to each difficulty parameter estimated in Wave 10 to derive the linked item parameters. The link error reflecting the uncertainty in the linking process was calculated according to equation 4 in Fischer et al. (2016) as 0.03.

6.3 Reading competence scores

In the SUF, manifest reading competence scores are provided in the form of two WLEs (“rea10_sc1” and “rea10_sc1u”) including their respective standard errors (“rea10_sc2” and “rea10_sc2u). For the variables ending with “u”, person abilities were estimated using the linked item difficulty parameters. As a result, these scores can be used for longitudinal comparisons between the two measurement occasions. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores without a “u” are not linked to the underlying reference scale of the respective previous wave. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions.

The R Syntax for estimating the WLEs is provided in Appendix B. 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

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.

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: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Gnambs, T., & Nusser, L. (2019). The longitudinal measurement of reasoning abilities in students with special educational needs. *Frontiers in Psychology, 10*:232, 88-92. https://doi.org/10.3389/fpsyg.2019.00232

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

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

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_sc4a10_c	MC	1	Information text	Drawing text-related conclusions
reg90120_sc4a10_c	MC	1	Information text	Finding information in the text
reg90130_sc4a10_c	MC	1	Information text	Drawing text-related conclusions
reg90140_sc4a10_c	MC	1	Information text	Reflecting and assessing
reg90150_sc4a10_c	MC	1	Information text	Finding information in the text
reg9016s_sc4a10_c	MA	1	Information text	Reflecting and assessing
reg9017s_sc4a10_c	CMC	1	Information text	Drawing text-related conclusions
reg90210_sc4a10_c	MC	2	Instruction text	Reflecting and assessing
reg90220_sc4a10_c	MC	2	Instruction text	Drawing text-related conclusions
reg90230_sc4a10_c	MC	2	Instruction text	Drawing text-related conclusions
reg90240_sc4a10_c	MC	2	Instruction text	Drawing text-related conclusions
reg90250_sc4a10_c	MC	2	Instruction text	Drawing text-related conclusions
reg90310_sc4a10_c	MC	3	Advertising text	Reflecting and assessing
reg90320_sc4a10_c	MC	3	Advertising text	Drawing text-related conclusions
reg9033s_sc4a10_c	CMC	3	Advertising text	Finding information in the text
reg90340_sc4a10_c	MC	3	Advertising text	Reflecting and assessing
reg90350_sc4a10_c	MC	3	Advertising text	Finding information in the text
reg90360_sc4a10_c	MC	3	Advertising text	Finding information in the text
reg90370_sc4a10_c	MC	3	Advertising text	Finding information in the text
reg90410_sc4a10_c	MC	4	Literary text	Finding information in the text
reg90420_sc4a10_c	MC	4	Literary text	Drawing text-related conclusions
reg90430_sc4a10_c	MC	4	Literary text	Drawing text-related conclusions
reg90440_sc4a10_c	MC	4	Literary text	Reflecting and assessing
reg90450_sc4a10_c	MC	4	Literary text	Drawing text-related conclusions
reg90460_sc4a10_c	MC	4	Literary text	Finding information in the text
reg9047s_sc4a10_c	CMC	4	Literary text	Reflecting and assessing
reg90510_sc4a10_c	MC	5	Commenting text	Reflecting and assessing
reg90520_sc4a10_c	MC	5	Commenting text	Finding information in the text
reg90530_sc4a10_c	MC	5	Commenting text	Drawing text-related conclusions
reg90540_sc4a10_c	MC	5	Commenting text	Finding information in the text
reg90550_sc4a10_c	MC	5	Commenting text	Reflecting and assessing
reg90560_sc4a10_c	MC	5	Commenting text	Finding information in the text
reg90570_sc4a10_c	MC	5	Commenting text	Finding information in the text

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

Appendix B: R-Syntax for estimating WLEs in Wave 10 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 sc4a10 c", "reg90120 sc4a10 c",
           "reg90130 sc4a10 c", "reg90150 sc4a10 c",
           "reg9016s sc4a10 c", "reg9017s sc4a10 c",
           "reg90210 sc4a10 c", "reg90220 sc4a10 c",
           "reg90230 sc4a10 c", "reg90240 sc4a10 c",
           "reg90250 sc4a10 c", "reg90310 sc4a10 c",
           "reg90320 sc4a10 c", "reg9033s sc4a10 c",
           "reg90340 sc4a10 c", "reg90350 sc4a10 c",
           "reg90360 sc4a10 c", "reg90370 sc4a10 c",
           "reg90410 sc4a10 c", "reg90420 sc4a10 c",
           "reg90430 sc4a10 c", "reg90440 sc4a10 c",
           "reg90450_sc4a10_c", "reg90460_sc4a10_c")

# define Q-matrix for 0.5 scoring of PCM
Q <- matrix(1, nrow = length(items), ncol = 1)
Q[c(5, 6, 14), 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)
```