NEPS SURVEY PAPERS

Maike Krannich, Odin Jost, Theresa Rohm, Ingrid Koller, Steffi Pohl, Kerstin Haberkorn, Claus H. Carstensen, Luise Fischer, and Timo Gnambs

NEPS TECHNICAL REPORT FOR READING: SCALING RESULTS OF STARTING COHORT 3 FOR GRADE 7

LIfBi

LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS Technical Report for Reading:

Scaling Results of Starting Cohort 3 for Grade 7

Maike Krannich[1], Odin Jost[2], Theresa Rohm[2,3], Ingrid Koller[4], Steffi Pohl[5], Kerstin Haberkorn[2], Claus H. Carstensen[3], Luise Fischer[2,3], & Timo Gnambs[2]

[1]University of Konstanz, Germany
[2]Leibniz Institute for Educational Trajectories, Germany
[3]University of Bamberg, Germany
[4]Alpen-Adria-Universität Klagenfurt, Austria
[5]Freie Universität Berlin, Germany

Email address of the lead author:

timo.gnambs@lifbi.de

NEPS Technical Report for Reading:
Scaling Results of Starting Cohort 3 for Grade 7

Abstract

The National Educational Panel Study (NEPS) investigates the development of competences across the life span and develops tests for the assessment of different competence domains. In order to evaluate the quality of the competence tests, a range of analyses based on item response theory (IRT) are performed. This paper describes the data and scaling procedure for the reading competence test in grade 7 of starting cohort 3 (fifth grade). The reading competence test contained 42 items (distributed among an easy and a difficult booklet) with different response formats representing different cognitive requirements and text functions. The test was administered to 6,194 students. Their responses were scaled using the partial credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, the test`s dimensionality, and local item independence were evaluated to ensure the quality of the test. These analyses showed that the test exhibited an acceptable reliability and that the items fitted the model in a satisfactory way. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test were the large percentage of items at the end of the difficult test that were not reached due to time limits and minor differential item functioning between the easy and difficult test version for some items. Overall, the reading competence test had acceptable psychometric properties that allowed for an estimation of reliable reading competence scores. Besides the scaling results, this paper also describes the data in the Scientific Use File and presentes the ConQuest syntax for scaling the data.

Keywords

item response theory, scaling, 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, information and communication technologies literacy, metacognition, vocabulary, and domain general cognitive functioning. An overview of the competences measured in the NEPS is given by Weinert and colleagues (2011) as well as Fuß, Gnambs, Lockl, and Attig (2016).

Most of the competence data are scaled using models that are based on item response theory (IRT). Because most of the competence tests were developed specifically for implementation in the NEPS, several analyses were conducted to evaluate the quality of the tests. The IRT models chosen for scaling the competence data and the analyses performed for checking the quality of the scale are described in Pohl and Carstensen (2012).

In this paper the results of these analyses are presented for reading competence in grade 7 of starting cohort 3 (fith grade). The study represents a follow-up to the reading competence test administered in grade 5 of starting cohort 3 (see Pohl, Haberkorn, Hardt, & Wiegand, 2012). First, the main concepts of the reading competence test are introduced. Then, the reading competence data of starting cohort 3 and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the scientific use file 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 scientific use file (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, we briefly describe specific aspects of the reading competence test that are necessary for understanding the scaling results presented in this paper.

The reading competence test included five texts and five item sets referring to these texts. Each of these texts represented one text type or text function, namely, a) information, b) commenting or arguing, 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 (MA) items. 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 respondents to match a number of responses to a given set of statements. MA items were usually used to assign headings to paragraphs of a text. Examples of the different response formats are given in Pohl and Carstensen (2012) and Gehrer, Zimmermann, Artelt and Weinert (2012).

The competence test for reading that was administered in the present study included 42 items. In order to evaluate the quality of these items, extensive preliminary analyses were conducted. These preliminary analyses identified a poor fit for two items, one unique to the easy test version and one unique to the difficult test version. Therefore, these items were removed from the final scaling procedure. Thus, the analyses presented in the following sections and the competence scores derived for the respondents are based on the remaining 40 items.

3. Data

3.1 The Design of the Study

The study followed a two-factorial experimental design. These factors referred to (a) the position of the reading test within the competence assessment of grade 7 and (b) the difficulty of the administered test.

Table 1

Number of Items by Text Types and Difficulty of the Test

Text type/functions	Easy test	Difficult test
Advertising text	6	6
Information text	5	6
Instruction text	6	6
Literary text	5	6
Commenting or arguing text	5	5
Total number of items	27	29

The study assessed different competence domains including reading competence and mathematical competence. The competence tests for these domains were always presented first within the test battery. In order to control for test position effects, the tests were administered to participants in different sequence. For each participant the reading test was either administered at the first or the second position (i.e., after the mathematics test). For students that had already participated in grade 5 the test order remained unchanged; thus, students that had received the reading competence test before any other tests in grade 5 also received the reading competence test at the first position in grade 7. Students that participated for the first time in grade 7 were randomly assigned to one of the two test

order conditions. There was no multi-matrix design regarding the order of the items *within* a specific test. All students received the test items in the same order.

Table 2

Number of Items by Cognitive Requirements and Difficulty of the Test

Cognitive requirements	Easy test	Difficult test
Finding Information in text	6	7
Drawing text-related conclusions	14	13
Reflecting and assessing	7	9
Total number of items	27	29

In order to measure participants' reading competence with great accuracy, the difficulty of the administered items should adequately match the participants' abilities. Therefore, the study adopted the principles of longitudinal multistage testing (Pohl, 2013). Based on preliminary studies two different versions of the reading competence test were developed that differed in their average difficulty (i.e., an easy and a difficult test). Both tests included five texts that represented the five text functions (see Table 1) and three cognitive requirements (see Table 2) as described above. Three texts (*information*, *instruction*, and *literary*) with 17 items were identical in both test versions, whereas two texts with 13 items were unique to the easy and the difficult test. Moreover, one additional item referring to one of the three common texts was only included in the difficult test version. In total, the reading competence test in grade 7 consisted of 42 items with different response formats (see Table 3). The number of subtasks within CMC items varied between two and five. Participants were assigned either to the easy or the difficult test based on their estimated reading competence in the previous assessment (Haberkorn et al., 2012). Participants with an ability estimate below the sample's mean ability received the easy test, whereas participants with a reading competence above the sample's mean received the difficult test. Participants, who did not take part in the grade 5 assessment, received the difficult version of the reading test.

Table 3

Number of Items by Different Response Formats and Difficulty of the Test

Response format	Easy test	Difficult test
Simple multiple choice	17	20
Complex multiple choice	7	5
Matching	3	4

Total number of items	27	29

3.2 Sample

A total of 6,194[1] individuals received the reading competence test. For eight respondents less than three valid item responses were available. Because no reliable ability score can be estimated based on such few valid responses, these cases were exclude from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 6,186 individuals. The number of participants within each experimental condition is given in Table 4. 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 4

Number of Participants by Experimental Condition

Test order	Easy test	Difficult test	Total
First position	882	2,231	3,113
Second position	889	2,184	3,073
Total	1,771	4,415	6,186

4. Analyses

Some of the following analyses are based on both test versions whereas other analyses examined the two test versions separately. Results that are based on separate analyses are explicitly indicated in the text and are reported in separate tables for the two test versions. Otherwise, the results refer to both test versions. These analyses did neither correct for the position of the reading competence test nor for the difficulty of the different test versions.

4.1 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 items that are not determinable.

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

---

[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.

time. All missing responses after the last valid response given were coded as not-reached. Because of the multi-stage testing design, 23 items were not administered to all participants. For respondents receiving the easy test 12 difficult items were missing by design, whereas 11 easy items were missing by design for respondents answering the difficult test (see Table 1). As CMC items and matching 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 one or more subtasks contained a missing response. If just one kind of missing response occurred, the item was coded according to the corresponding missing response. If 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.

4.2 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 items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly responded subtasks within that item. If at least one of the subtasks contained a missing response, the CMC item was scored as missing. Categories of polytomous variables with less than $N = 200$ responses were collapsed in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items; in these cases, the lower categories were collapsed into one category.

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).

Reading competences were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989) and will later also be provided in form of plausible values (Mislevy, 1991). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in section 7.

4.3 Checking the quality of the test

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.

Before aggregating the subtasks of CMC and MA items to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks

were analyzed together with the MC items in a Rasch model (Rasch, 1960). The fit of the subtasks was evaluated based on the weighted mean square (WMNSQ), the respective *t*-value, point-biserial correlations of the correct responses with the total correct score, and the item characteristic curves. Only if the subtasks exhibited a satisfactory item fit, they were used to construct polytomous CMC and MA variables that were included in the final scaling model.

The MC items consisted of one correct response option and one or more 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 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).

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC and polytomous CMC items to the partial credit model (Masters, 1982) was evaluated using three indices (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (*t*-value > |6|) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 (*t*-value > |8|) were judged as having a considerable item misfit and their performance was further investigated. Correlations of the item score with the corrected total score (equal to the corrected discrimination as computed in ConQuest) greater than .30 were considered as good, greater than .20 as acceptable, and below .20 as problematic. Overall judgment of the fit of an item was based on all fit indicators.

The reading competence test should measure the same construct for all students. If some items favored certain subgroups (e.g., they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., males and females) would be biased and, thus, unfair. For the present study, test fairness was investigated for the variables sex, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Moreover, in light of the experimental design, measurement invariance analyses were also conducted for the test position and the difficulty of the test. 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 differences in estimated difficulties between the subgroups that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as considerable and noteworthy of further investigation, differences between 0.4 and 0.6 as small but not severe, and differences smaller than 0.4 as negligible DIF. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

The 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 dimensionality of the test was evaluated by two different multidimensional analyses. The different subdimensions of the multidimensional models were specified based on different construction criteria. First, a model with three different subdimensions representing the three cognitive requirements, and, second, a model with five different subdimensions based on the five text functions were fitted to the data. The correlations among the dimensions as well as differences in model fit between the unidimensional model and the respective multidimensional models were used to evaluate the unidimensionality of the test.

Since the reading test consisted of item sets that referred to one of five texts, the assumption of local item dependence (LID) may not necessarily hold. However, the five texts were perfectly confounded with the five text functions. Thus, multidimensionality and local item dependence cannot be evaluated separately with these data.

4.4 Software

The IRT models were estimated in ConQuest version 4.2.5 (Adams, Wu, & Wilson, 2015).

5. Results

5.1 Missing responses

5.1.1 Missing responses per person

The number of invalid responses per person is shown in Figure 1. The number of invalid responses was very low for both test versions. In the easy test version 93% of the students had no invalid responses at all and only about two percent of the students had more than one invalid response. In the difficult test version 95% of the students had no invalid responses at all and only about one percent of the students had more than one invalid response.

Figure 1: Number of invalid responses

Missing responses may also occur when respondents omit some items. As can be seen in Figure 2, there was a non-negligible amount of omitted items even if the number of omitted items was not remarkable. In the easy test version 74% of the students omitted no item at all, whereas only four percent of the students omitted more than three items. In the difficult test version 73% of the students omitted no item at all and four percent of the students omitted more than three items.
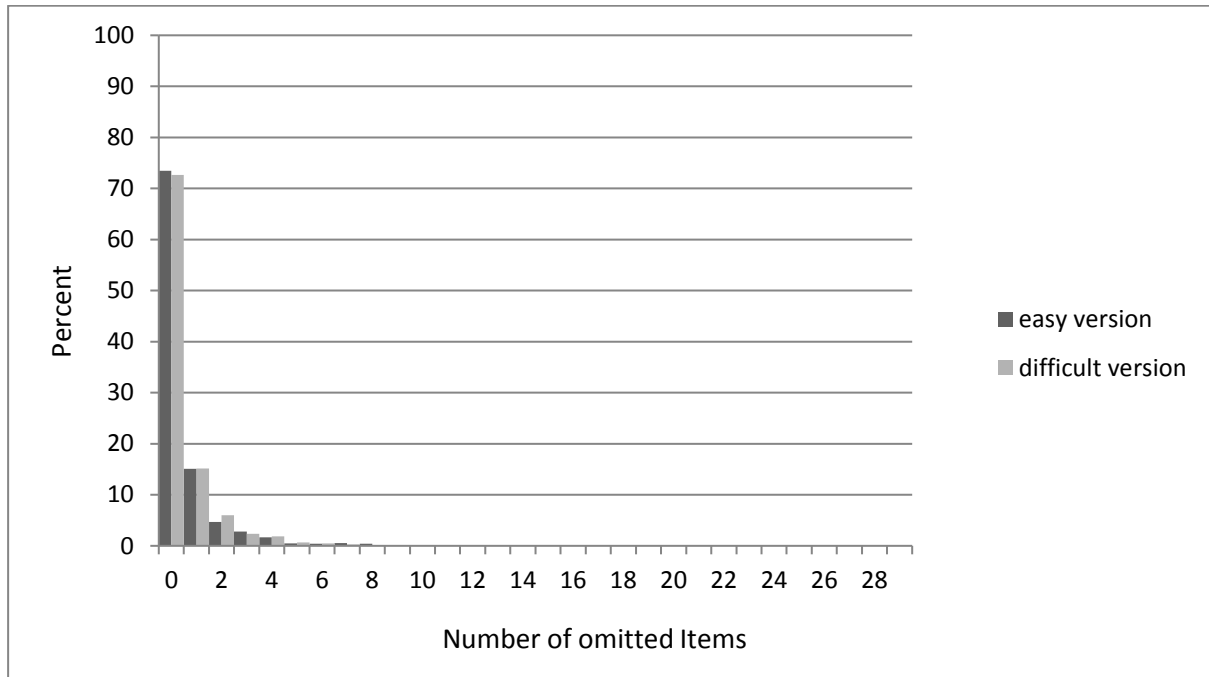
Figure 2: Number of omitted items

Per definition, all missing responses after the last valid response were not reached. The number of not-reached items for the easy test version was acceptable whereas the number of the not-reached items for the difficult test version was rather high. This is illustrated in Figure 3. About 81% of the students reached the end of the easy reading competence test; 9.5% of the students did not reach the items of the last text and 1.8% did not reach the last two of the five texts. Note that only 54% of the students reached the end of the difficult reading test. In this case, 18% of the students did not reach the items of the last text, 10.5% did not reach the last two of the five texts, and 5% only reached the first two texts. These figures are comparable to the amount of 48% of respondents who reached the end of the test in the fifth grade assessment (Pohl et al., 2012).
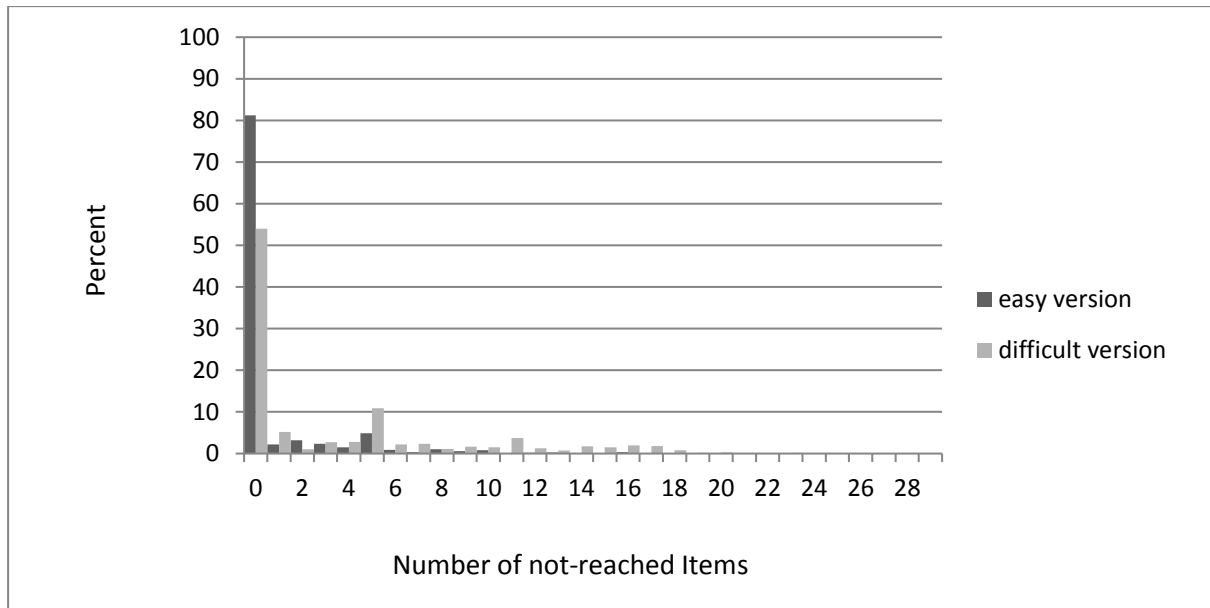
Figure 3: Number of not-reached items
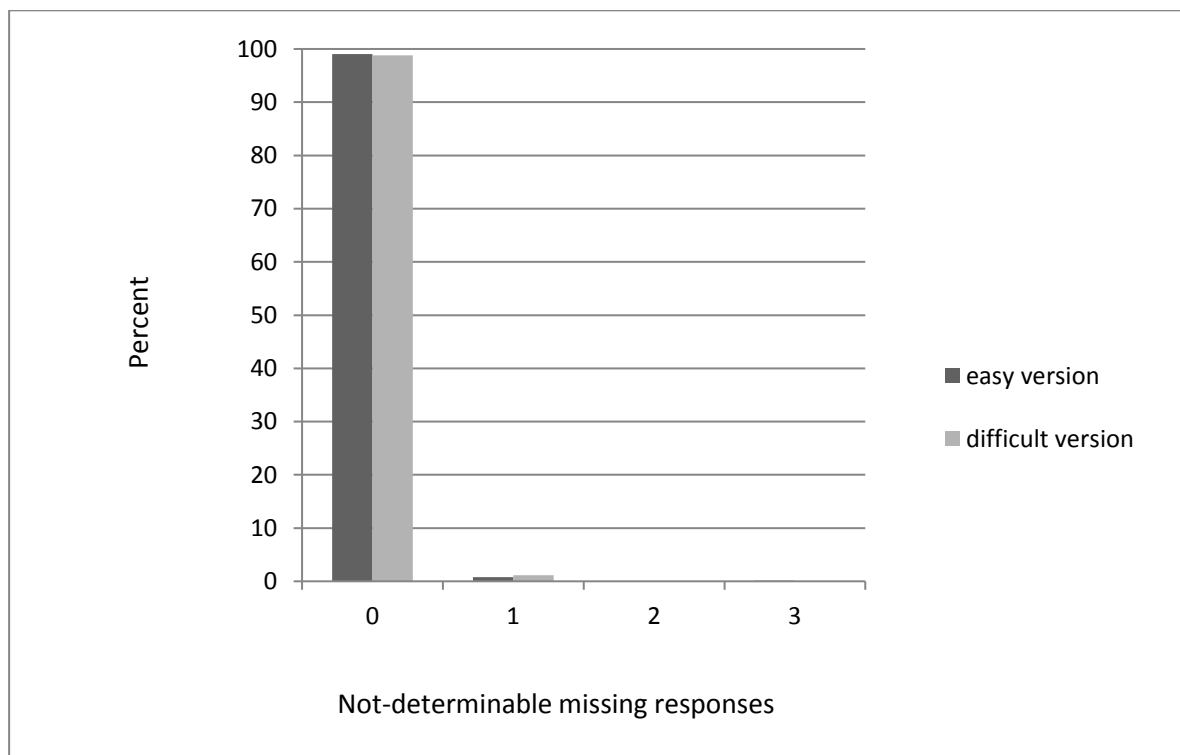
Figure 4: Number of not-determinable missing responses

The aggregated polytomous variables were coded as not-determinable missing response when the subtasks of CMC and MA items contained different kinds of missing responses. Because not-determinable missing responses may only occur in CMC and MA items, the maximum number of not-determinable missing responses was nine (for the difficult test version) or ten (for the easy test version). There was only a very small amount of not-

determinable missing responses for both test versions (see Figure 4). About 99% of the students in both test versions did not have a single not-determinable missing response.

The total number of missing responses aggregated over invalid, omitted, not-reached, and not-determinable missing responses per person is illustrated in Figure 5. It can be seen that 56% of the students that were administered the easy test version had no missing response at all. Only about 9% of these tested students had more than five missing responses. In the difficult test version, there were about 39% of the students who had no missing response at all. Almost 29% of these tested students had more than five missing responses and about 8% of the students had missing responses to more than 14 items (i.e., 50% of the items).
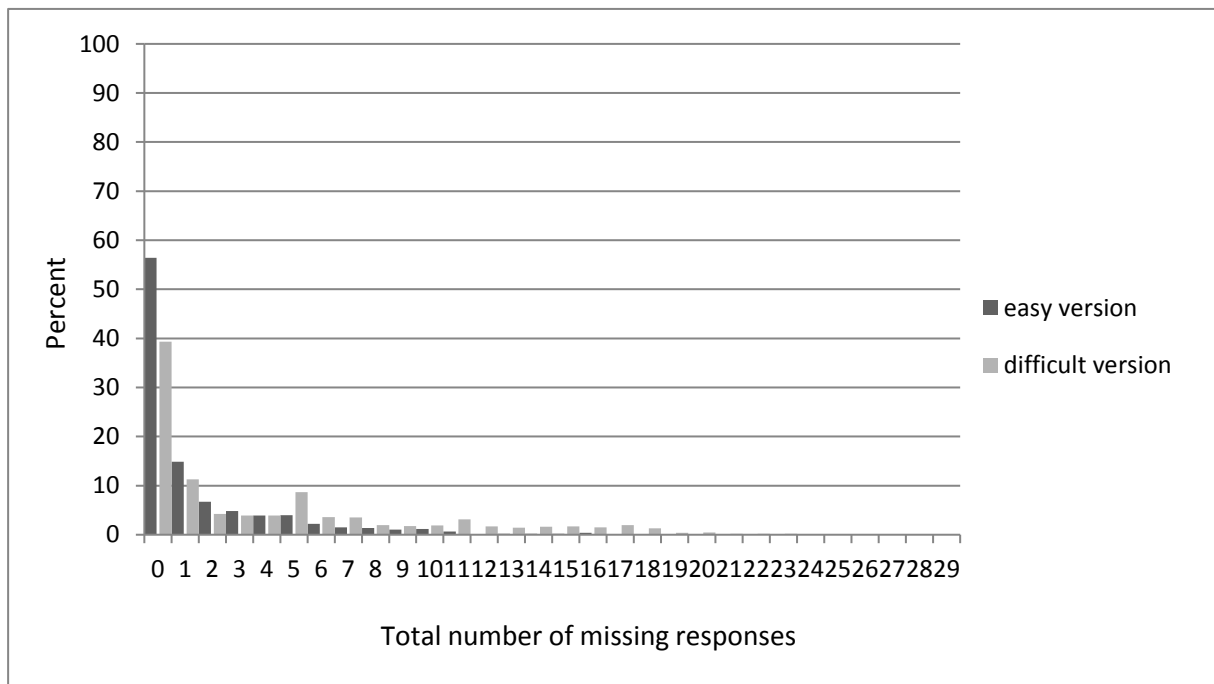
Figure 5: Total number of missing responses

In sum, there was a small amount of invalid and not-determinable missing responses for both test versions and a reasonable amount of omitted items. The number of not-reached items was—at least for the difficult test version—rather large and, therefore, the higher amount of total missing responses in the difficult version was primarily explained by the not-reached items.

Table 5

Missing Values for the Different Test Versions

	Position in booklet		Number of valid responses		Relative frequency of not reached items in %		Relative frequency of omitted items in %		Relative frequency of invalid items in %		Relative frequency of not determinable items in %	
	easy	difficult	easy	difficult	easy	difficult	easy	difficult	easy	difficult	easy	difficult
reg70110_c	1		1,740		0.00		0.51		1.24			
reg70120_c	2		1,738		0.00		1.75		0.11			
reg7013s_c	3		1,736		0.00		1.81		0.11		0.06	
reg70140_c	4		1,760		0.00		0.45		0.17			
reg7015s_c	5		1,722		0.00		2.65		0.06		0.06	
reg7016s_c	6		1,645		0.00		5.59		1.07		0.45	
reg70610_c		1		4,407		0.00		0.11		0.07		
reg70620_c		2		4,370		0.00		0.91		0.11		
reg7063s_c		3		4,313		0.00		2.24		0.07		
reg70640_c		4		4,214		0.00		4.51		0.05		
reg70650_c		5		4,322		0.00		1.68		0.43		0.00
reg7066s_c		6		4,102		0.02		6.27		0.66		0.14
reg70210_c	7	7	1,753	4,375	0.00	0.25	0.68	0.57	0.34	0.09		
reg70220_c	8	8	1,732	4,275	0.00	0.29	1.81	2.79	0.40	0.09		
reg7023s_c	9	9	1,720	4,248	0.00	0.39	2.82	3.19	0.06	0.16	0.00	0.05
reg7024s_c	10	10	1,689	4,255	0.00	0.63	4.52	2.54	0.11	0.45	0.00	
reg70250_c	11	11	1,710	4,301	0.00	0.86	1.64	1.61	1.81	0.11		
reg7026s_c		12		4,203		1.68		2.56		0.20		0.36
reg70310_c	12	13	1,744	4,239	0.34	3.49	0.62	0.39	0.56	0.11		

Note. The items of the easy test version are denoted by white color, the items of the difficult test version are denoted by dark grey color, and the common items are denoted by light grey color.

Table 5 *(continued)*

	Position in booklet		Number of valid responses		Relative frequency of not reached items in %		Relative frequency of omitted items in %		Relative frequency of invalid items in %		Relative frequency of not determinable items in %	
	easy	difficult	easy	difficult	easy	difficult	easy	difficult	easy	difficult	easy	difficult
reg70320_c	13	14	1,708	4,134	0.40	5.44	2.60	0.86	0.56	0.07		
reg7033s_c	14	15	1,690	4,072	0.68	6.91	3.27	0.61	0.45	0.07	0.17	0.18
reg70340_c	15	16	1,687	3,986	0.90	8.67	2.99	1.00	0.85	0.05		
reg70350_c	16	17	1,702	3,818	1.02	9.45	2.20	3.13	0.68	0.95		
reg70360_c	17	18	1,678	3,860	1.02	10.69	3.44	1.70	0.79	0.18		
reg70410_c	18	19	1,729	3,759	1.81	14.45	0.23	0.20	0.34	0.20		
reg70420_c	19	20	1,703	3,575	2.37	15.97	1.19	2.27	0.28	0.79		
reg70430_c	20	21	1,690	3,568	3.39	17.60	0.90	1.52	0.28	0.07		
reg70440_c	21	22	1,673	3,458	3.73	18.75	1.52	2.56	0.28	0.36		
reg7045s_c	22	23	1,579	3,357	4.63	21.13	5.19	2.33	0.85	0.25	0.17	0.25
reg70460_c		24		3,135		25.19		2.94		0.86		
reg7051s_c	23		1,554		9.54		2.48		0.17		0.06	
reg70520_c	24		1,541		11.07		1.36		0.56			
reg7053s_c	25		1,487		13.44		2.48		0.11		0.00	
reg7055s_c	27		1,417		16.60		2.20		0.90		0.28	
reg70560_c	26		1,426		18.80		0.00		0.68			
reg7071s_c		25		2,770		34.16		2.99		0.11		0.00
reg70720_c		26		2,664		36.99		2.49		0.18		
reg70730_c		27		2,614		39.71		0.97		0.11		
reg70740_c		28		2,588		40.77		0.48		0.14		
reg7075s_c		29		2,319		45.98		0.75		0.52		0.23

Table 6

Item Parameters

Item	Percentage correct	Difficulty / location parameter	*SE* (difficulty/location parameter)	WMNSQ	*t*-value of WMNSQ	Correlation of item score with total score	Discrimination 2PL
reg70110_c	45.50	-0.375	0.058	1.27	11.6	0.25	0.37
reg70120_c	82.50	-2.524	0.072	1.09	2.1	0.32	0.80
reg7013s_c	n.a.	-2.594	0.080	0.93	-1.9	0.43	1.70
reg70140_c	91.40	-3.456	0.093	0.90	-1.7	0.43	2.03
reg7015s_c	87.20	-2.940	0.081	1.04	0.9	0.29	0.94
reg7016s_c	n.a.	-1.099	0.063	1.11	3.8	0.29	0.80
reg70610_c	92.70	-2.847	0.064	0.94	-1.3	0.40	1.71
reg70620_c	66.30	-0.613	0.040	1.08	4.5	0.43	0.91
reg7063s_c	n.a.	-2.706	0.069	0.96	-1.4	0.37	1.45
reg70640_c	46.10	0.464	0.039	1.14	9.9	0.36	0.65
reg70650_c	50.50	0.229	0.039	1.05	3.9	0.44	0.95
reg7066s_c	n.a.	-1.208	0.043	0.98	-1.1	0.53	1.33
reg70210_c	90.40	-2.792	0.050	0.89	-3.6	0.46	2.02
reg70220_c	82.20	-1.941	0.041	1.02	0.8	0.44	1.26
reg7023s_c	n.a.	-1.932	0.046	1.01	0.8	0.35	1.22
reg7024s_c	n.a.	-0.754	0.040	0.97	-2.3	0.43	1.43
reg70250_c	68.70	-1.003	0.036	1.12	7.6	0.41	0.88
reg7026s_c	n.a.	-1.419	0.044	0.88	-5.3	0.60	1.98
reg70310_c	89.10	-2.629	0.048	1.03	1.0	0.36	1.10

Note. The items of the easy test version are denoted by white color, the items of the difficult test version are denoted by dark grey color, and the common items are denoted by light grey color. For the dichotomous items, the correlation with the total score corresponds to the point-biserial correlation between the correct response and the total correct score, whereas for polytomous items it corresponds to the product-moment correlation between the corresponding categories and the total correct score (discrimination value as computed in ConQuest). Percent correct scores are not informative for polytomous CMC and MA item scores. These are denoted by n.a.

Table 6 *(continued)*

Item	Percentage correct	Difficulty / location parameter	*SE* (difficulty/location parameter)	WMNSQ	*t*-value of WMNSQ	Correlation of item score with total score	Discrimination 2PL
reg70320_c	77.90	-1.627	0.039	1.11	5.7	0.38	0.85
reg7033s_c	n.a.	-1.215	0.040	0.93	-3.7	0.41	1.62
reg70340_c	76.80	-1.533	0.039	0.96	-2.1	0.51	1.52
reg70350_c	83.20	-2.040	0.043	0.87	-5.7	0.56	2.10
reg70360_c	72.50	-1.252	0.038	0.97	-1.5	0.52	1.42
reg70410_c	86.30	-2.350	0.047	0.95	-1.8	0.46	1.55
reg70420_c	81.20	-1.892	0.043	0.91	-4.3	0.54	1.85
reg70430_c	86.80	-2.403	0.048	0.87	-5.0	0.54	2.32
reg70440_c	81.50	-1.917	0.043	0.89	-5.0	0.55	1.90
reg7045s_c	n.a.	-0.469	0.038	0.97	-2.0	0.45	1.36
reg70460_c	40.10	0.801	0.044	0.99	-0.6	0.46	1.18
reg7051s_c	n.a.	-1.963	0.072	1.01	0.2	0.42	1.25
reg70520_c	62.30	-1.292	0.063	0.96	-1.5	0.53	1.40
reg7053s_c	n.a.	-1.164	0.070	1.04	1.6	0.39	1.03
reg7055s_c	n.a.	0.124	0.064	1.11	3.4	0.28	0.78
reg70560_c	27.60	0.522	0.070	1.07	2.4	0.37	0.89
reg7071s_c	n.a.	-1.482	0.061	1.08	3.2	0.31	0.93
reg70720_c	37.30	0.918	0.049	1.13	6.4	0.35	0.72
reg70730_c	42.00	0.631	0.049	1.12	6.5	0.38	0.74
reg70740_c	70.30	-0.911	0.052	1.14	5.7	0.39	0.80
reg7075s_c	n.a.	0.318	0.048	0.93	-3.1	0.49	1.49

5.1.2 Missing responses per item

Table 5 gives information on the number of valid responses for each item, as well as the percentage of missing responses. Overall, the omission rate was quite good. In the easy test version there were only two items with an omission rate above 5%; in the difficult test version none of the items had an omission rate above 5%. The highest omission rate occurred for item reg7016s_c (5.56% of the students omitted this item). The number of students that did not reach an item increased with the position of the item in the test to up to 18.8% (easy test version) or 45.98% (difficult test version). This is a rather large amount, especially for the difficult test version. The number of invalid responses per item was small. The highest number was 1.81 % for item reg70250_c (easy test version) or 0.95% for item reg7045s_c (difficult test version). The total number of missing responses per item varied between 0% and almost 46% (item reg7075s_c in the difficult test version).

5.2 Parameter estimates

5.2.1 Item parameters

The percentage of correct responses relative to all valid responses for each item is summarized in Table 6 (second column). Because there was a non-negligible amount of missing responses this value cannot be interpreted as an index of item difficulty. The percentage of correct responses within dichotomous items varied between 27.60% and 92.70% with an average of 69.94% correct responses.

Table 7

Step Parameters (and Standard Errors) of Polytomous Items

Item	Step 1 (*SE*)		Step 2 (*SE*)		Step 3 (*SE*)		Step 4 (*SE*)
reg7013s_c	0.418	(0.068)	-0.418				
reg7016s_c	-0.778	(0.066)	0.158	(0.069)	0.620		
reg7023s_c	-0.075	(0.033)	0.075				
reg7024s_c	-0.174	(0.029)	0.174				
reg7026s_c	-0.476	(0.072)	0.264	(0.069)	0.259	(0.060)	-0.047
reg7033s_c	-0.016	(0.050)	-0.936	(0.048)	0.952		
reg7045s_c	-0.954	(0.036)	0.817	(0.043)	0.137		
reg7051s_c	0.807	(0.074)	-0.807				
reg7053s_c	0.374	(0.062)	-0.374				
reg7055s_c	-0.413	(0.060)	0.535	(0.088)	-0.121		
reg7063s_c	0.045	(0.048)	-0.045				
reg7066s_c	-0.956	(0.065)	0.137	(0.054)	1.200	(0.058)	-0.380
reg7071s_c	0.351	(0.051)	-0.351				
reg7075s_c	-0.376	(0.049)	0.437	(0.065)	-0.061		

The item parameters were estimated based on the final scaling model, the partial credit model, with concurrent calibration (i.e., the easy and difficult test were scaled together). The

estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 6, whereas the step parameters (for polytomous variables) are summarized in Table 7. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties (or location parameters for polytomous variables) varied between -3.46 (item reg70140_c) and 0.92 (item reg70720_c) with a mean of -1.31. Overall, the item difficulties ranged from low to medium difficulty; however, there were no items with a high difficulty. Due to the large sample size, the standard errors (*SE*) of the estimated item difficulties (column 4 in Table 6) were rather small, *SE(ß)* ≤ 0.10.

5.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. This was done separately for the easy and the difficult test versions. In Figures 6a and 6b, the item difficulties of the reading competence items 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 item difficulties. In these analyses the mean of the item difficulties was constrained to be zero. The variance was estimated to be 1.124 for the easy and 1.462 for the difficult test version, which indicates good differentiation between the students. The reliabilities of the easy (EAP/PV reliability = 0.807, WLE reliability = 0.780) and for the difficult version (EAP/PV reliability = 0.807, WLE reliability = .761) were good. The mean of the person competence distribution was about 0.98 logits above the mean item difficulty of zero for the easy and 1.54 logits above the mean item difficulty of zero for the difficult test version. Subsequently, we replicated these analyses for the concurrently scaled easy and difficult test (i.e., both tests were scaled together; see Table 6). In this analysis, the variance was estimated to be 1.566. The reliability was good with EAP/PV reliability = 0.827 and WLE reliability = 0.791. The mean of the person competence distribution was about 1.36 logits above the mean item difficulty of zero. Although the items covered a wide range of the ability distribution, on average, the items were slightly too easy. As a consequence, person abilities in medium- and low-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors.

5.3 Quality of the test

5.3.1 Fit of the subtasks of complex multiple choice and matching items

Before the subtasks of CMC and MA items were aggregated and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the single subtasks together with the simple MC items in a Rasch model. Counting the subtasks of CMC and MA items separately, there were 48 items in the easy and 50 items in the difficult test version. The probability of a correct response ranged from 27% to 92% across all items. Thus, the number of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.85 to 1.21, the respective *t*-value from -10.6 to 12.4, and there were no noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. Due to the satisfying model fit of the subtasks, their aggregation to polytomous variables seemed justified.

scale in logits	person ability	item difficulty
	X	
	X	
3	XX	
	XX	
	XXX	
	XXX	
	XXXX	
	XXXXX	
2	XXXXXX	17
	XXXXXXX	
	XXXX	27
	XXXXXXXX	
	XXXXXXXX	
	XXXXXXXXX	24
	XXXXXXXXXX	1 22
1	XXXXXX	
	XXXXXXXX	
	XXXXXXX	
	XXXXXXXX	6 23
	XXXXXXX	11 20 26
	XXXXX	9 16
	XXXXXXX	
0	XXXXX	21
	XXXX	5 8 13
	XXXXX	10 15
	XXXX	25
	XXX	
	XXX	4 12 14
	XX	
-1	XX	2 18
	XX	7
	X	
	X	
	X	
	X	
		3
-2		
-3		19

Figure 6a: Test targeting for the easy test version. Distribution of person ability (left side of the graph) and item difficulties (right side of the graph). Each 'X' represents 11 cases. Each number represents an item (which corresponds to the item position given in Table 5).

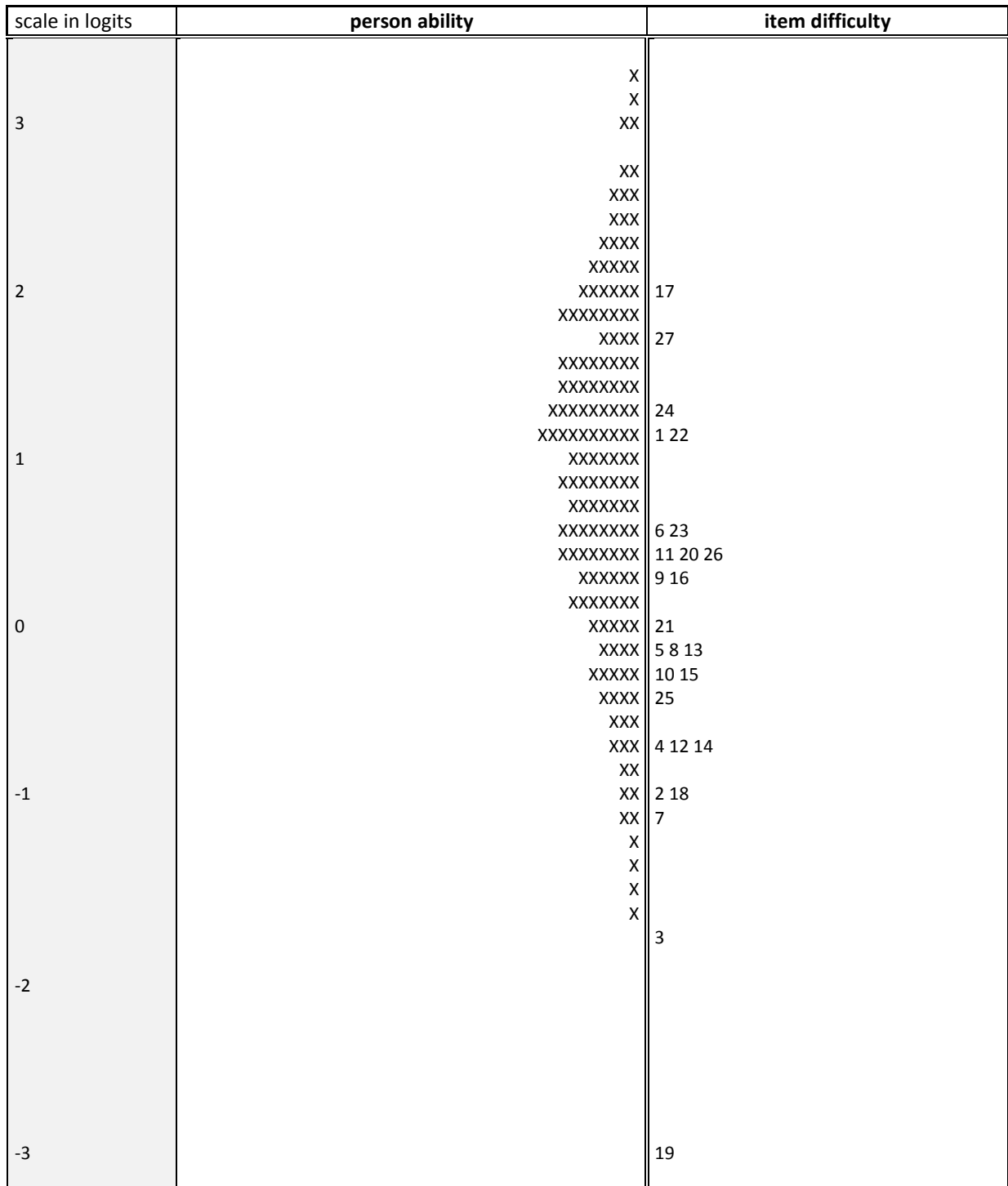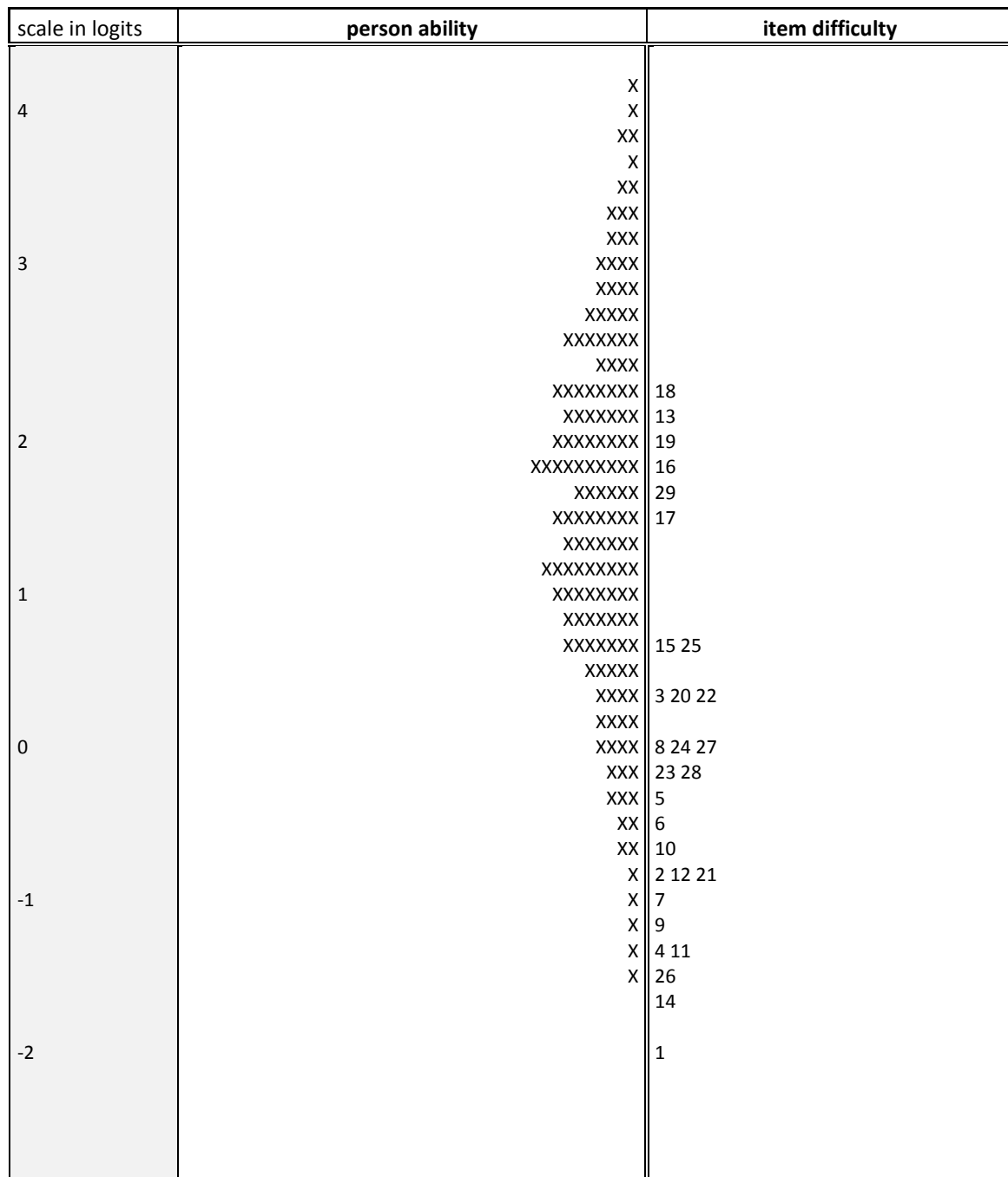| scale in logits	person ability	item difficulty
4	X X XX X XX XXX XXX	
3	XXXX XXXX XXXXX XXXXXX XXXX	
2	XXXXXXX XXXXXXX XXXXXXX XXXXXXXXX XXXXX XXXXXXXX XXXXXX XXXXXXXXX	18 13 19 16 29 17
1	XXXXXXX XXXXXX XXXXXX XXXXX	15 25
0	XXXX XXXX XXXX XXX XXX XX XX	3 20 22 8 24 27 23 28 5 6 10
-1	X X X X X	2 12 21 7 9 4 11 26 14
-2		1

Figure 6b: Test targeting for difficult test version. Distribution of person ability (left side of the graph) and item difficulties (right side of the graph). Each 'X' represents 28.4 cases. Each number represents an item (which corresponds to the item position depicted in Table 5).

5.3.2 Distractor analyses

In addition to the overall item fit, we specifically investigated how well the distractors performed in the test by evaluating the point-biserial correlation between selecting an incorrect response (distractor) and the students' total correct score. The distractors consistently yielded negative point-biserial correlations ranging from -.44 to .00 for the easy and between -.41 to .00 for the difficult test version. These results indicate that the distractors functioned well.

5.3.3 Item fit

The evaluation of item fit was performed based on the final scaling model, the partial credit model, with concurrent calibration (i.e., the easy and difficult test were scaled together). Altogether, the item fit can be considered good (see Table 6). Values of the WMNSQ were close to 1 with the lowest value being .87 (item reg70430_c) and the highest being 1.27 (item reg70110_c). Only two items exhibited a WMNSQ above 1.15 or a *t*-value above 8. There were no further indications of pronounced misfit of these items. Therefore, they were retained for estimating the reading competence scores. The correlations between the item scores and the total correct scores varied between .25 (item reg70110_c) and .60 (item reg7026s_c) with an average correlation of .42. All item characteristic curves showed a good fit of the items.

5.3.4 Differential item functioning

Differential item functioning (DIF) was used to evaluate the test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables sex, the number of books at home (as a proxy for socioeconomic status), migration background, and test position (see Pohl & Carstensen, 2012, for a description of these variables). In addition, for the common items that were administered to all participants we also studied their measurement invariance between the easy and difficult test version. The differences between the estimated item difficulties in the various groups are summarized in Table 9. For example, the column "Male vs. female" reports the differences in item difficulties between men and women; a positive value would indicate that the item was more difficult for males, whereas a negative value would highlight a lower difficulty for males as opposed to females. Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allowed for DIF to those that only estimated main effects (see Table 10).

Sex: The sample included 2,872 (48.3%) boys and 3,072 (51.7%) girls. 242 respondents that did not indicate their sex were excluded from the analysis. On average, male students had a lower reading ability than female students (main effect = -0.382 logits, Cohen's *d* = 0.307). Only one item (item reg70220_c) showed considerable DIF greater than 0.6 logits (-0.802 logits), whereas five items exhibited a small but not severe DIF between 0.4 and 0.6 logits.

Table 9: Differential Item Functioning

Item	Sex	Books			Migration			Position	Test version
	male vs. female	< 100 vs. ≥ 100	< 100 vs. missing	≥ 100 vs. missing	without vs. with	with vs. missing	without vs. missing	first vs. second	easy vs. difficult
reg70110_c	-0.344	-0.060	0.225	0.285	0.020	-0.183	-0.203	0.322	
	(-0.277)	(-0.05)	(0.188)	(0.238)	(0.016)	(-0.149)	(-0.165)	(0.259)	
reg70120_c	-0.102	-0.259	-0.319	-0.060	-0.028	-0.410	-0.382	-0.002	
	(-0.082)	(-0.216)	(-0.266)	(-0.050)	(-0.023)	(-0.334)	(-0.311)	(-0.002)	
reg7013s_c	-0.072	0.048	-0.222	-0.270	-0.016	0.004	0.020	-0.498	
	(-0.058)	(0.040)	(-0.185)	(-0.225)	(-0.013)	(0.003)	(0.016)	(-0.401)	
reg70140_c	0.150	0.058	-0.520	-0.578	-0.238	0.294	0.532	-0.630*	
	(0.121)	(0.048)	(-0.434)	(-0.483)	(-0.194)	(0.240)	(0.434)	(-0.508)	
reg7015s_c	-0.246	0.622	-0.604	-1.226	-0.348	-0.396	-0.048	0.066	
	(-0.198)	(0.519)	(-0.504)	(-1.024)	(-0.284)	(-0.323)	(-0.039)	(0.053)	
reg7016s_c	0.216	-0.115	0.142	0.250	-0.117	0.351	0.468	0.036	
	(0.174)	(-0.096)	(0.119)	(0.215)	(-0.095)	(0.286)	(0.381)	(0.029)	
reg70210_c	0.094	0.015	-0.411	-0.426	-0.079	-0.179	-0.100	-0.182	0.642*
	(0.076)	(0.013)	(-0.343)	(-0.356)	(-0.064)	(-0.146)	(-0.082)	(-0.147)	(0.480)
reg70220_c	-0.802*	-0.093	-0.031	0.062	0.026	0.007	-0.019	-0.016	0.146
	(-0.645)	(-0.078)	(-0.026)	(0.052)	(0.021)	(0.006)	(-0.015)	(-0.013)	(0.109)
reg7023s_c	-0.288	-0.058	0.000	0.058	0.015	0.084	0.069	-0.042	0.208
	(-0.232)	(-0.048)	(0.000)	(0.048)	(0.012)	(0.068)	(0.056)	(-0.034)	(0.156)
reg7024s_c	-0.560*	0.108	0.275	0.167	0.345	0.307	-0.038	-0.016	0.202
	(-0.450)	(0.090)	(0.230)	(0.139)	(0.281)	(0.250)	(-0.031)	(-0.013)	(0.151)
reg70250_c	-0.352	0.114	0.159	0.045	-0.200	-0.019	0.181	0.416*	-0.356
	(-0.283)	(0.095)	(0.133)	(0.038)	(-0.163)	(-0.015)	(0.148)	(0.335)	(-0.266)
reg7026s_c	-0.486*	0.206	-0.032	-0.238	0.113	-0.092	-0.205	-0.164	
	(-0.391)	(0.172)	(-0.027)	(-0.199)	(0.092)	(-0.075)	(-0.167)	(-0.132)	
reg70310_c	0.024	-0.161	0.050	0.211	-0.020	0.206	0.226	-0.042	-0.478*
	(0.019)	(-0.134)	(0.042)	(0.176)	(-0.016)	(0.168)	(0.184)	(-0.034)	(-0.358)
reg70320_c	0.138	-0.344	-0.259	0.085	0.140	-0.026	-0.166	-0.040	-0.420*
	(0.111)	(-0.287)	(-0.216)	(0.071)	(0.114)	(-0.021)	(-0.135)	(-0.032)	(-0.314)
reg7033s_c	-0.160	0.251	0.068	-0.183	-0.209	-0.203	0.006	-0.054	0.022
	(-0.129)	(0.210)	(0.057)	(-0.153)	(-0.170)	(-0.165)	(0.005)	(-0.044)	(0.016)

Item	Sex	Books			Migration			Position	Test version
reg70340_c	-0.176	0.095	0.034	-0.061	0.024	0.132	0.108	-0.052	0.150
	(-0.142)	(0.079)	(0.028)	(-0.051)	(0.020)	(0.108)	(0.088)	(-0.042)	(0.112)
reg70350_c	0.230	0.261	-0.117	-0.378	-0.363	-0.318	0.045	-0.152	0.178
	(0.185)	(0.218)	(-0.098)	(-0.316)	(-0.296)	(-0.259)	(0.037)	(-0.122)	(0.133)
reg70360_c	-0.396	0.237	0.216	-0.021	-0.300	-0.225	0.075	-0.004	-0.128
	(-0.318)	(0.198)	(0.180)	(-0.018	(-0.245)	(-0.183)	(0.061)	(-0.003)	(-0.096)
reg70410_c	0.326	-0.099	-0.255	-0.156	-0.032	0.200	0.232	0.058	-0.128
	(0.262)	(-0.083)	(-0.213)	(-0.130)	(-0.026)	(0.163)	(0.189)	(0.047)	(-0.096)
reg70420_c	0.244	0.023	-0.101	-0.124	0.149	-0.122	-0.271	-0.114	-0.060
	(0.196)	(0.019)	(-0.084)	(-0.104)	(0.121)	(-0.099)	(-0.221)	(-0.092)	(-0.045)
reg70430_c	0.306	0.159	-0.039	-0.198	-0.141	-0.046	0.095	-0.224	0.062
	(0.246)	(0.133)	(-0.033)	(-0.165)	(-0.115)	(-0.037)	(0.077)	(-0.180)	(0.046)
reg70440_c	0.146	0.144	-0.119	-0.263	-0.220	-0.275	-0.055	0.120	-0.030
	(0.117)	(0.120)	(-0.099)	(-0.220)	(-0.179)	(-0.224)	(-0.045)	(0.097)	(-0.022)
reg7045s_c	0.130	0.093	0.063	-0.030	0.259	-0.022	-0.281	0.018	-0.010
	(0.105)	(0.078)	(0.053)	(-0.025)	(0.211)	(-0.018)	(-0.229)	(0.015)	(-0.007)
reg70460_c	0.242	0.264	0.462	0.198	0.181	0.0380	-0.143	0.120	
	(0.195)	(0.220)	(0.386)	(0.165)	(0.148)	(0.031)	(-0.117)	(0.097)	
reg7051s_c	0.464	-0.006	0.570	0.576	0.221	-0.191	-0.412	0.008	
	(0.373)	(-0.005)	(0.476)	(0.481)	(0.180)	(-0.156)	(-0.336)	(0.006)	
reg70520_c	0.418	-0.004	0.070	0.074	0.150	0.285	0.135	0.094	
	(0.336)	(-0.003)	(0.058)	(0.062)	(0.122)	(0.232)	(0.110)	(0.076)	
reg7053s_c	-0.112	0.103	0.859*	0.756	-0.053	0.437	0.490	0.014	
	(-0.090)	(0.086)	(0.717)	(0.631)	(-0.043)	(0.356)	(0.399)	(0.011)	
reg7055s_c	0.232	-0.254	0.185	0.439	0.171	-0.201	-0.372	0.120	
	(0.187)	(-0.212)	(0.154)	(0.367)	(0.139)	(-0.164)	(-0.303)	(0.097)	
reg70560_c	-0.100	0.087	-0.030	-0.117	-0.069	-0.053	0.016	0.324	
	(-0.080)	(0.073)	(-0.025)	(-0.098)	(-0.056)	(-0.043)	(0.013)	(0.261)	
reg70610_c	0.272	0.120	0.046	-0.074	-0.299	-0.262	0.037	-0.340	
	(0.219)	(0.100)	(0.038)	(-0.062)	(-0.244)	(-0.214)	(0.030)	(-0.274)	
reg70620_c	-0.014	-0.325	-0.254	0.071	-0.032	0.083	0.115	0.008	
	(-0.011)	(-0.271)	(-0.212)	(0.059)	(-0.026)	(0.068)	(0.094)	(0.006)	
reg7063s_c	0.550	0.058	-0.280	-0.338	-0.175	-0.366	-0.191	-0.232	
	(0.442)	(0.048)	(-0.234)	(-0.282)	(-0.143)	(-0.298)	(-0.156)	(-0.187)	

Item	Sex	Books			Migration			Position	Test version
reg70640_c	-0.196	-0.358	-0.011	0.347	0.078	0.222	0.144	0.022	
	(-0.158)	(-0.299)	(-0.009)	(0.290)	(0.064)	(0.181)	(0.117)	(0.018)	
reg70650_c	-0.114	-0.003	0.153	0.156	-0.085	0.054	0.139	0.174	
	(-0.092)	(-0.003)	(0.128)	(0.130)	(-0.069)	(0.044)	(0.113)	(0.140)	
reg7066s_c	-0.116	0.028	-0.094	-0.122	-0.0900	-0.144	-0.054	0.036	
	(-0.093)	(0.023)	(-0.078)	(-0.102)	(-0.073)	(-0.117)	(-0.044)	(0.029)	
reg7071s_c	0.188	-0.287	0.025	0.312	0.263	-0.057	-0.320	0.206	
	(0.151)	(-0.240)	(0.021)	(0.261)	(0.214)	(-0.046)	(-0.261)	(0.166)	
reg70720_c	0.010	-0.183	0.304	0.487	0.177	0.339	0.162	0.230	
	(0.008)	(-0.153)	(0.254)	(0.407)	(0.144)	(0.276)	(0.132)	(0.185)	
reg70730_c	-0.024	-0.370	-0.044	0.326	0.602	0.346	-0.256	0.158	
	(-0.019)	(-0.309)	(-0.037)	(0.272)	(0.491)	(0.282)	(-0.209)	(0.127)	
reg70740_c	0.042	-0.102	0.008	0.110	0.003	0.246	0.243	0.174	
	(0.034)	(-0.085)	(0.007)	(0.092)	(0.002)	(0.201)	(0.198)	(0.140)	
reg7075s_c	0.236	-0.011	-0.168	-0.157	0.174	0.153	-0.021	0.074	
	(0.190)	(-0.009)	(-0.14)	(-0.131)	(0.142)	(0.125)	(-0.017)	(0.060)	
Main effect	-0.382	-0.719	0.107	0.826	0.499	0.611	0.112	0.318	-1.208
	(-0.307)	(-0.600)	(0.089)	(0.690)	(0.407)	(0.498)	(0.091)	(0.256)	(-0.904)

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

[*] Absolute standardized difference is significantly, $p < .05$, greater than 0.25 (see Fischer et al., 2016).

An overall test for DIF (see Table 10) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). Model comparisons using Akaike's (1974) information criterion (AIC) and also the Bayesian information criterion (BIC; Schwarz, 1978) that takes the number of estimated parameters into account and, thus, penalizes over-parameterized models favored the more complex DIF model. Ignoring the DIF for the items resulted in a slightly smaller mean difference between male and female students (main effect = -0.312 logits, Cohen's *d* = 0.252). Thus, the DIF regarding sex did not have a large impact.

<u>Number of books:</u> The number of books at home was used as a proxy for socioeconomic status. There were 2,473 (40.0%) students with 0 to 100 books at home, 3,225 (52.1%) students with more than 100 books at home, and 488 (7.9%) students without a valid response. DIF was investigated using these three groups (see Table 9). There were considerable average differences between the three groups. Participants with 100 or less books at home performed on average 0.719 logits (Cohen's *d* = 0.600) worse than participants with more than 100 books. Participants without a valid response performed 0.107 logits (Cohen's *d* = 0.089) or 0.826 logits (Cohen's *d* = 0.690) worse than participants with up to 100 and more than 100 books, respectively. One item (item reg7015s_c) showed considerable DIF between participants with many or fewer books (DIF = 0.622). Comparing the group without valid responses to the two groups with valid responses, DIF occurred up to -1.2 logits (item reg7015s_c). This is a rather large difference, which may, however, also be the result of the uncertainty in estimation due to the small number of students with missing responses. As a consequence, the overall test for DIF using the BIC favored the main effects model without DIF effects (Table 10).

<u>Migration:</u> There were 4,122 (66.7%) participants without a migration background, 1,363 (22.0%) participants with a migration background, and 701 (11.3%) participants without a valid response. There was a considerable difference in the average performance of participants with and without migration background. Participants without a migration background had a higher reading ability than participants with a migration background (main effect = 0.499 logits, Cohen's *d* = 0.407). Also, students with missing values on migration differed from those without a migration background (main effect = 0.611 logits, Cohen's *d* = 0.498) but only negligibly from those with a migration background (main effect = 0.112 logits, Cohen's *d* = 0.091). One item (item reg70730_c) exhibited considerable DIF between participants with and without migration background (DIF = 0.602 logits). However, the overall test for DIF using the BIC favored the main effects model that did not include item-level DIF. Therefore, reading competences were measured comparably in the three groups.

<u>Test version:</u> To estimate the participants' proficiency with greater accuracy the participants received different tests that either included a larger number of easy or a larger number of difficult items (see section 3.1 for the design of the study). Only a subset of 17 items that were included in both tests was administered to all participants. For these common items we examined potential DIF across the two test versions (easy versus difficult). A subsample of 1,771 participants (28.6%) received the easy test version and 4,415 participants (71.4%) received the difficult test version. As expected, students who were administered the difficult test version outperformed the participants receiving the easy test version (main effect = -1.208 logits, Cohen's *d* = -0.904). One item (item reg70210_c) showed considerable DIF (DIF = 0.642) and two items (item reg70310_c and item reg70320_c) exhibited small but no severe DIF (DIF = -0.478 and DIF = -0.420). The information criteria AIC and BIC showed

marginally lower values for the model including DIF effects (AIC = 101,673.60, BIC = 101,936.10, number of parameters = 39) as compared to a simpler model including only a main effect (AIC = 101,803.03, BIC = 101,964.54, number of parameters = 24). Therefore, we also examined whether the reading competence test exhibited measurement invariance between students who received the easy version based on their reading performance in grade 5, students who received the difficult version based on their reading performance in grade 5, and students who received the difficult version without having participated in grade 5. For the difficult test version, these analyses showed that students with a grade 5 score outperformed students without a grade 5 score (main effect = 0.597 logits, Cohen's *d* = .454). No item exhibited considerable DIF for these groups. Additionally, irrespective whether the students had a grade 5 score or not, students working on the difficult booklet outperformed students receiving the easy booklet (with grade 5 score: main effect = -1.497 logits, Cohen's *d* = -1.139; without grade 5 score: main effect = -0.900 logits, Cohen's *d* = -0.685). The item reg70210_c showed considerable DIF between students receiving the easy test version and the two groups receiving the difficult test version (DIF = 0.666 and 0.648 logits, respectively). Moreover, another item (item reg70320_c) showed considerable DIF between the easy test version and students with a grade 5 score (DIF = -0.620). However, a model comparison indicated a better fit for the simpler model without DIF effects (AIC = 101,664.60, BIC = 101,832.85, number of parameters = 25;) as compared to a more complex model that also acknowledged DIF effects between the three groups for the items (AIC = 101,511.40, BIC = 101,881.55, number of parameters = 55). Taking into account the detailed analysis and the minor model violations, measurement invariance between the booklets can be assumed.

Table 10

Comparison of models with and without DIF

DIF Variable	Model	Deviance	Number of parameters	AIC	BIC
Test version	main effect	101,755.026	24	101,803.026	101,964.543
	DIF	101,595.596	39	101,673.596	101,936.100
Sex	main effect	187,057.054	65	187,187.054	187,621.900
	DIF	186,615.286	104	186,823.286	187,519.100
Books	main effect	195,088.544	66	195,220.544	195,664.700
	DIF	194,845.883	144	195,133.883	196,103.000
Migration	main effect	195,353.122	66	195,485.122	195,929.300
	DIF	195,150.537	144	195,438.537	196,407.700

In summary, most of the differences in item difficulties across the different subgroups were (in absolute values) below 0.6. There were only two larger effects for the number of books and three larger values for the migration background of the participants. With regard to the model fit indices, the BIC indicated a better fit for the models without DIF; thus, there was no substantial indication of test unfairness.

5.3.5 Rasch-homogeneity

An essential assumption of the Rasch (1960) 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 6), ranging from 0.37 (item reg70110_c) to 2.32 (item reg70430_c). The average discrimination parameter fell at 1.28. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 193,864.00, BIC = 194,557.20, number of parameters = 103) as compared to the PCM (AIC = 195,687.60, BIC = 196,118.40, number of parameters = 64). Despite the empirical preference for the GPCM, the PCM more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the partial credit model was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

Table 11a

Results of three-dimensional scaling for the easy test version

	Dim 1	Dim 2	Dim 3
Finding information in the text (Dim 1) (10 items)	1.673		
Drawing Text-related conclusions (Dim 2) (18 items)	0.885	1.520	
Reflecting and assessing (Dim 3) (12 items)	0.858	0.931	0.635

Note. Variance of the dimensions are depicted in the diagonal, correlations are given in the off-diagonal.

Table 11b

Results of three-dimensional scaling for the difficult test version

	Dim 1	Dim 2	Dim 3
Finding information in the text (Dim 1) (10 items)	1.552		
Drawing Text-related conclusions (Dim 2) (18 items)	0.903	1.622	
Reflecting and assessing (Dim 3) (12 items)	0.919	0.936	1.552

Note. Variance of the dimensions are depicted in the diagonal, correlations are given in the off-diagonal.

5.3.6 Unidimensionality

The unidimensionality of the two test versions was investigated by specifying two different multidimensional models and comparing them to a unidimensional model. In the first multidimensional model, three different cognitive requirements were specified, whereas the

five different text types constituted the second multidimensional model. Both models were estimated separately for the easy and difficult test versions to examine the assumption of unidimensionality for both tests. Estimation of the models was carried out using the Gauss-Hermite quadrature method.

The estimated variances and correlations between the three dimensions representing the different cognitive requirements are reported in Table 11a and 11b. The correlations among the three dimensions were rather high and fell between .86 and .94. However, they deviated from a perfect correlation (i.e., they were lower than r = .95, see Carstensen, 2013). Moreover, according to model fit indices for the easy test version, the three-dimensional model (AIC = 61,976.93, BIC = 62,228.98, number of parameters = 46) fitted the data slightly better than the unidimensional model (AIC = 62,176.74, BIC = 62,401.39, number of parameters = 41). Similarly, for the difficult test version the three-dimensional model (AIC = 132,458.53, BIC = 132,784.60, number of parameters = 51) fitted the data slightly better than the unidimensional model (AIC = 132,520.66, BIC = 132,814.70, number of parameters = 46). These results indicate that the three cognitive requirements measure a common construct, albeit it is not completely unidimensional.

Table 12a

Results of five-dimensional scaling for the easy test version

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Information (Dim 1) (12 items)	1.111				
Instruction (Dim 2) (6 items)	0.818	1.052			
Advertising (Dim 3) (6 items)	0.731	0.813	1.811		
Commenting (Dim 4) (6 items)	0.695	0.763	0.783	2.763	
Literary (Dim 5) (10 items)	0.712	0.720	0.785	0.808	1.350

Note. Variance of the dimensions are given in the diagonal, correlations are given in the off-diagonal.

The estimated variances and correlations of the five-dimensional model based on the five text functions are given in Table 12a and 12b. The correlations between the dimensions varied between r = .70 and r = .89. All correlations deviated from a perfect correlation (i.e., they were considerably lower than r = .95, see Carstensen, 2013). For the easy test version, the five-dimensional model (AIC = 64,582.53, BIC = 64,883.89, number of parameters = 55) fitted the data worse than the unidimensional model (AIC = 62,176.74, BIC = 62,401.39, number of parameters = 41). Similarly, for the difficult test version the five-dimensional model (AIC = 138,615.80, BIC = 138,999.40, number of parameters = 60) fitted the data

worse than the unidimensional model (AIC = 132,520.66, BIC = 132,814.70, number of parameters = 46).

Table 12b

Results of five-dimensional scaling for the difficult test version

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Information (Dim 1) (12 items)	1.292				
Instruction (Dim 2) (6 items)	0.892	2.125			
Advertising (Dim 3) (6 items)	0.804	0.846	2.138		
Commenting (Dim 4) (6 items)	0.833	0.863	0.861	2.682	
Literary (Dim 5) (10 items)	0.768	0.823	0.742	0.795	1.143

Note. Variance of the dimensions are given in the diagonal, correlations are given in the off-diagonal.

As each text function corresponded to one of the five texts, local item dependence (LID) and the text functions were confounded. As a consequence, the deviation of the correlations from a perfect correlation shown in Tables 11a to 12b, may result from multidimensionality as well as from local item dependence. Given the testing design in the main studies, it was not possible to disentangle the two sources. In pilot studies (Gehrer et al., 2013), a larger number of texts were presented to test takers, so that the impact of text functions could be investigated independently of LID. The correlations estimated in the pilot study ranged from .78 to .91. As the correlations found in Gehrer and colleagues (2013) differed from a perfect correlation, it is concluded that text functions form subdimensions of reading competence. Comparing the correlations found in Gehrer et al. (2013), which were due to text functions, to those found in the main study (Table 12), which were due to both text functions and LID, allowed us to evaluate the impact of LID. The correlations found in the present study of starting cohort 3 were slightly lower (between 0.70 and 0.94) than those found in Gehrer et al. (between 0.78 and 0.91), indicating that there is some amount of local item dependence. However, according to the test developers a balanced assessment of reading competence can only be achieved by heterogeneity of text functions (Gehrer et al., 2013).

6. Discussion

The analyses in the previous sections provided detailed information on the quality of the reading test in starting cohort 3 for grade 7. We investigated different kinds of missing responses and examined the item and test parameters. We thoroughly checked item fit statistics for simple MC items, subtasks of CMC and MA items, as well as the aggregated polytomous CMC items, and examined the correlations between correct and incorrect responses and the total score. Further quality inspections were conducted by examining

differential item functioning, testing Rasch-homogeneity, investigating the tests' dimensionality as well as local item dependence.

Various criteria indicated a good fit of the items and measurement invariance across various subgroups. However, the number of not-reached items in the difficult test version was rather high, indicating that the test was too long for the allocated testing time and the difficulty of the items. However, the number of not-reached items was still comparable with other reading competence tests (see section 5.1.1). Other types of missing responses were reasonably small. The test had a high reliability and distinguished well between test takers. However, the test was mainly targeted at low-performing students and did not accurately measure reading competence of high-performing students. As a consequence, ability estimates will be precise for low-performing students but less precise for high performing students. Some degree of multidimensionality was present for different text functions. In combination with the high amount of missing responses at the end of the test (i.e., there were students with no valid responses to some of the text functions), the estimation of a single reading competence score might be challenged. This should be addressed in further studies. Nevertheless, Gehrer et al. (2013) argued that a balanced assessment of reading competence can only be achieved by heterogeneity of text functions and they provide theoretical arguments for a unidimensional measure of reading competence.

This was the first study within the NEPS that administered two difficulty-tiered tests and assigned students to one of the two test versions based on their previous performance on the reading competence test in grade 5. Because the complex design provided additional challenges as compared to previous studies, additional analyses were conducted that showed that the common items of the two test versions measured the same latent dimension as the test unique items. Moreover, the common items were largely measurement invariant across the two test versions. Dimensionality analyses showed that the latent associations between the content dimensions (cognitive requirements and text functions) were comparatively high.

In sum, it was shown that it is feasible to implement a macro-adaptive procedure that assigned students to an easy or a more difficult test version depending on their prior performance. Overall, the administered test had satisfactory psychometric properties that facilitated the estimation of a unidimensional reading competence score.

7. Data in the Scientific Use File

7.1 Naming conventions

The data in the Scientific Use File contain 40 items, of which 26 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. A total of 14 items were scored as polytomous variables (CMC items). MC items are marked with a '0_c' at the end of the variable name, whereas the variable names of CMC items end in 's_c'. In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category. Note that the values of the polytomous variables in the Scientific Use File do not necessarily indicate the number of correctly responded subtasks, as categories may have been collapsed within CMC and MA items.

7.2 Linking of reading competence scores of grade 5 and grade 7

In starting cohort 3, the reading competence tests administered in grades 5 (see Pohl et al., 2012) and 7 included different items that were constructed in such a way as to allow for an accurate measurement of reading competence within each age group. As a consequence, the competence scores derived in the different grades cannot be directly compared; differences in observed scores would reflect differences in competences as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competences across grades, we adopted the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016). The process of linking combines adjacent measurement points on the same scale. As such, the first wave of each competence scale within a cohort is used as reference scale that all subsequent measurement waves will refer to. For the domain of reading competence linking is achieved using an anchor-group design. In order to link the two tests of reading competence an independent link sample including students from grade 7 that were not part of starting cohort 3 were administered all items from the grade 5 and the grade 7 reading competence tests within a single measurement occasion. These responses were used to link the two tests administered in starting cohort 3 across the two grades. For detailed information on the process of linking competence data in the NEPS and the method for linking the tests of reading competence in starting cohort 3 (grade 5 and grade 7) see Fischer et al. (2016).

7.3 Reading competence scores

In the SUF manifest reading competence scores are provided in the form of two different WLEs, "reg7_sc1" and "reg7_sc1u", including their respective standard error, "reg7_sc2" and "reg7_sc2u". For "reg7_sc1u", person abilities were estimated using the linked item difficulty parameters. As a result, the WLE scores provided in "reg7_sc1u" can be used for longitudinal comparisons between grades 5 and 7. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in "reg7_sc1" are not linked to the underlying reference scale of grade 5. However, they are corrected for the position of the reading test within the booklet. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions. The ConQuest Syntax for estimating the WLE is provided in Appendix A. 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.

Plausible values that allow for an investigation of latent relationships of competence scores with other variables will be provided in future data releases. Alternatively, users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values themselves. A description of these approaches can be found in Pohl and Carstensen (2012).

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ConQuest 4*. Camberwell, Australia: Acer.

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

Carstensen, C. H. (2013). Linking PISA competencies over three cycles – Results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research Outcomes of the PISA Research Conference 2009* (pp. 199-214). New York, NY: Springer.

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

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2016). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6)*. Bamberg: 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. (2012). *The assessment of reading competence (including sample items for grade 5 and 9)*. Scientific Use File 2012, Version 1.0.0. Bamberg, Germany: University of Bamberg, National Educational Panel Study.

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.

Hardt, K., Pohl, S., Haberkorn, K., & Wiegand, E. (2013). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 6 for adults in main study 2010/11* (NEPS Working Paper No. 25). Bamberg: University of Bamberg, National Educational Panel Study.

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

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196. doi:10.1007/BF02294457

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

Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement, 50*, 447-468. doi:10.1111/jedm.12028

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-Universität, Nationales Bildungspanel.

Pohl, S., & Carstensen, C. (2013). Scaling the 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., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grade* (NEPS Working Paper No. 15). Bamberg, Germany: Otto-Friedrich-Universität, Nationales Bildungspanel.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen & Lydiche.

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

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

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011) Development of competencies across the life span. In H. P. Blossfeld, H. G. Roßbach, & J. v. Maurice (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67-86). (Zeitschrift für Erziehungswissenschaft, Sonderheft 14 . Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Appendix

Appendix A: ConQuest-Syntax for estimating WLE estimates in starting cohort 3

```
Title SC3 G7 READING: Partial Credit Model;


/* load data */
data [FILENAME].dat;
format pid 1-7 responses 10-49;
labels << [FILENAME].txt;

/* scoring */
codes 0,1,2,3,4;
score (0,1) (0,1)                   !items (1-25,27);
score (0,1,2) (0,0.5,1)             !item (26,29,30,34,35,37,39);
score (0,1,2,3) (0,0.5,1,1.5)       !item (28,32,33,36,40);
score (0,1,2,3,4) (0,0.5,1,1.5,2) !item (31,38);

/* model specification */
set constraint = cases;
model item + item*step;             /* uncorrected for test position*/
model item + item*step - rotation; /* corrected for test position*/


/* estimate model */
estimate ! method=gauss, nodes=15, iterations=1000, convergence=0.0001;

/* save results to file */
show ! estimate=latent    >> show.txt;
show cases ! estimate=wle >> wle.txt;
```