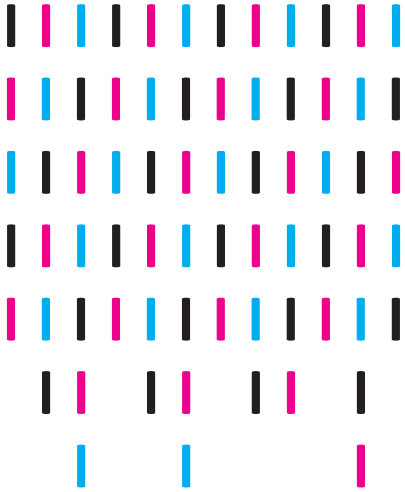

NEPS SURVEY PAPERS

Timo Gnambs

NEPS TECHNICAL REPORT FOR ENGLISH READING COMPETENCE: SCALING RESULTS OF STARTING COHORT 4 FOR GRADE 12

NEPS Survey Paper No. 27
Bamberg, August 2017

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 English Reading Competence: Scaling Results of Starting Cohort 4 for Grade 12

Timo Gnambs

Leibniz Institute for Educational Trajectories, Bamberg, Germany

E-mail address of lead author:

timo.gnambs@lifbi.de

Bibliographic data:

Gnambs, T. (2017). *NEPS Technical Report for English Reading Competence: Scaling Results of Starting Cohort 4 for Grade 12* (NEPS Survey Paper No. 27). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Acknowledgements:

Various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (e.g., Haberkorn et al., 2012; Hardt, K., Pohl, S., Haberkorn, K., & Wiegand, E., 2013) to facilitate the understanding of the presented results.

NEPS Technical Report for English Reading Competence: Scaling Results of Starting Cohort 4 for Grade 12

Abstract

The National Educational Panel Study (NEPS) investigates the development of competencies across the life span. Therefore, the NEPS develops tests for the assessment of various competence domains in different age cohorts. In order to evaluate the quality of these competence tests, several analyses based on item response theory (IRT) are performed. This paper describes the data and scaling procedures for a reading competence test on English as a foreign language in grade 12 of starting cohort 4 (ninth grade). The reading competence test included 10 items with multiple choice response formats and matching tasks that represented different levels of the Common European Framework of References. The test was administered to 3,898 students (56% girls). Their responses were scaled using a partial credit model. Item fit statistics and differential item functioning were evaluated to ensure the quality of the test. These analyses showed that the test exhibited an acceptable reliability and a satisfactory fit to the Rasch model. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test pertained to its difficulty that that did not adequately cover the upper range of the ability distribution. Overall, the English reading test had acceptable psychometric properties that allowed for an estimation of reliable competence scores. The competence scores derived in the present study were linked to the underlying scale of the English test administered in grade 10 to allow for meaningful longitudinal comparisons of English reading competence. 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, scaling, English as a foreign language, scientific use file

Content

1. Introduction ... 3
2. Testing English Reading Competence ... 3
3. Data ... 4
4. Analyses ... 4
4.1 Missing Responses ... 4
4.2 Scaling Model ... 5
4.3 Checking the Quality of the Test ... 5
4.4 Software ... 7
5. Results ... 7
5.1 Missing Responses ... 7
5.1.1 Missing responses per person ... 7
5.1.2 Missing responses per item ... 10
5.2 Parameter Estimates ... 12
5.2.1 Item parameters ... 12
5.2.2 Test targeting and reliability ... 12
5.3 Quality of the test ... 14
5.3.1 Item fit ... 14
5.3.2 Distractor analyses ... 14
5.3.3 Differential item functioning ... 14
5.3.4 Rasch-homogeneity ... 16
5.3.5 Unidimensionality ... 16
6. Discussion ... 17
7. Data in the Scientific Use File ... 17
7.1 Naming conventions ... 17
7.2 English reading competence scores ... 18

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 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 a competence test on English as a foreign language that was administered in grade 12 of starting cohort 4 (ninth grade). First, the main concepts of the English competence test are introduced. Then, the competence data of starting cohort 4 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, no fundamental changes in the presented results are expected.

2. Testing English Reading Competence

The framework and item development for the English reading competence tests was led by the Institute for Educational Quality Improvement (IQB) and is described in Rupp, Vock, Harsch, and Köller (2008). The reading competence test in English included four short texts that were accompanied by four item sets referring to these texts. All items were developed by trained experts and corresponded to the National Educational Standards and the Common European Framework of Reference (Council of Europe, 2001). The students had to read each text and, subsequently, answer multiple items related to this text.

The items were accompanied by different response formats (see Table 1). Simple multiple choice formats included four response options with one being correct and three response options functioning as distractors (i.e., they were incorrect). Complex multiple choice (CMC) items consisted of several subtasks that had to be rated as true, false, or information not given in the text. Matching (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) and Gehrer, Zimmermann, Artelt and Weinert (2012).

The competence test for English reading that was administered in the present study included 10 items. To evaluate the quality of these items, extensive preliminary analyses were conducted. These preliminary analyses identified a poor item fit for one subtask in items efg10022s_sc4g12_c and efg12b00s_c. Moreover, items efg12d006_c and efg12d007_c

exhibited an inadequate fit to the scaling model and low discriminations. Therefore, these subtasks and 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 8 items.

Table 1. Number of Items by Different Response Formats

Response format	Frequency
Simple multiple choice items	7
Complex multiple choice items	1
Matching items	2
Total number of items	10

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. A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

3. Data

A total of 3,898[1] students received the English reading competence test. For six respondents no valid item responses were available. These cases were excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 3,892 individuals (56% female). All participants attended higher secondary schools ("Gymnasium"). The test was not administered to students attending other school types or to school leavers.

4. Analyses

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 and MA items that are not determined. Invalid responses occurred, for example, when two response options were selected although 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

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.

finished the test within the given time. All missing responses after the last valid response given were coded as not reached. Because of the booklet design, the items unique to each booklet were not administered to participants receiving another booklet. These items were missing by design. Because 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. 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 and MA items consisted of a set of subtasks that were aggregated to a polytomous variable for each item, indicating the number of correctly solved subtasks within that item. If at least one of the subtasks contained a missing response, the CMC or MA item was scored as missing. Categories of polytomous variables with less than $N = 100$ 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.

English 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 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 English reading competence test was specifically constructed for administration in the NEPS. In order to ensure appropriate psychometric properties, the quality of the test was examined in several analyses.

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 (1960) model. 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 variables that were included in the final scaling model.

The MC items consisted of one correct response option and three distractors (i.e., incorrect response options). The quality of the distractors within the 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 and MA items to the partial credit model (Masters, 1982) was evaluated using the weighted mean square (WMNSQ) statistic, the respective *t*-value, point-biserial correlations of the correct responses with the total correct score, and the item characteristic curves (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) 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 English 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, 2012a, for a description of these variables). 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. Minimum hypothesis tests (see Fischer, Rohm, Gnambs, & Carstensen, 2016) were used to statistically test whether the observed differences was significantly larger than 0.4 and, thus, was at least small in size. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

The English reading competence test was scaled using the PCM (Masters, 1982) because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). 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 examining the residuals of the PCM. Approximately zero-order correlations as indicated by Yen's (1984) Q_3 indicate unidimensionality. Because in case of locally independent items, the Q_3 statistic tends to be slightly negative, we report the corrected Q_3 that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) values of aQ_3 falling below .20 indicate essential unidimensionality.

4.4 Software

The IRT models were estimated in TAM version 2.4-9 (Kiefer, Robitzsch, & Wu, 2017) in R version 3.4.1 (R Core Team, 2017) using the Gauss-Hermite quadrature method with 21 nodes.

5. Results

5.1 Missing Responses

5.1.1 Missing responses per person

Figure 1 shows the number of invalid responses per person by booklet. Overall, there were hardly any invalid responses; more than 95% of the respondents had no invalid response at all, whereas about 5% exhibited one invalid response.

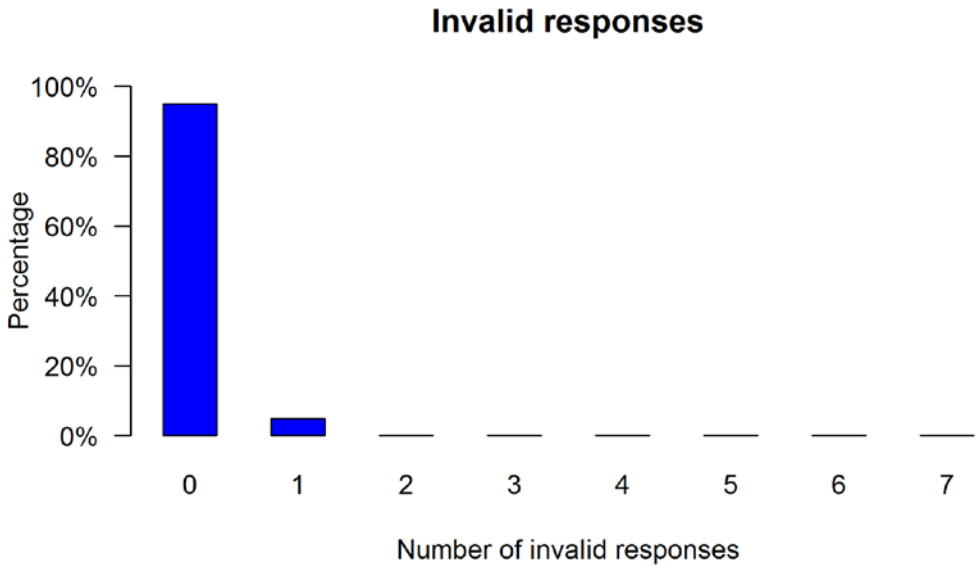

Figure 1. Number of invalid responses

Missing responses can also occur when respondents omit items. As illustrated in Figure 2 there was a considerable amount of missing responses. Only, 53% or the respondents omitted no items. In contrast, about 37% had one omitted response and 10% two or more missing responses. This indicates that at least some items (see below) were too complex for the respondents and, therefore, resulted in omitted responses.

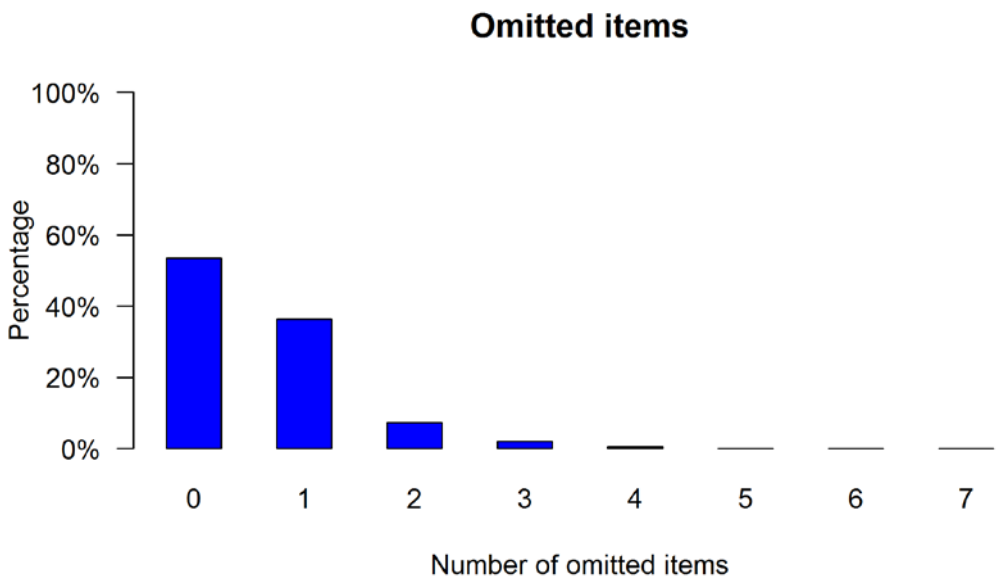

Figure 2. Number of omitted items

Another source of missing responses is 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 low; more than 96% of the respondents finished the entire test (Figure 3). Thus, most respondents were able to finish the test within the allocated time limit.

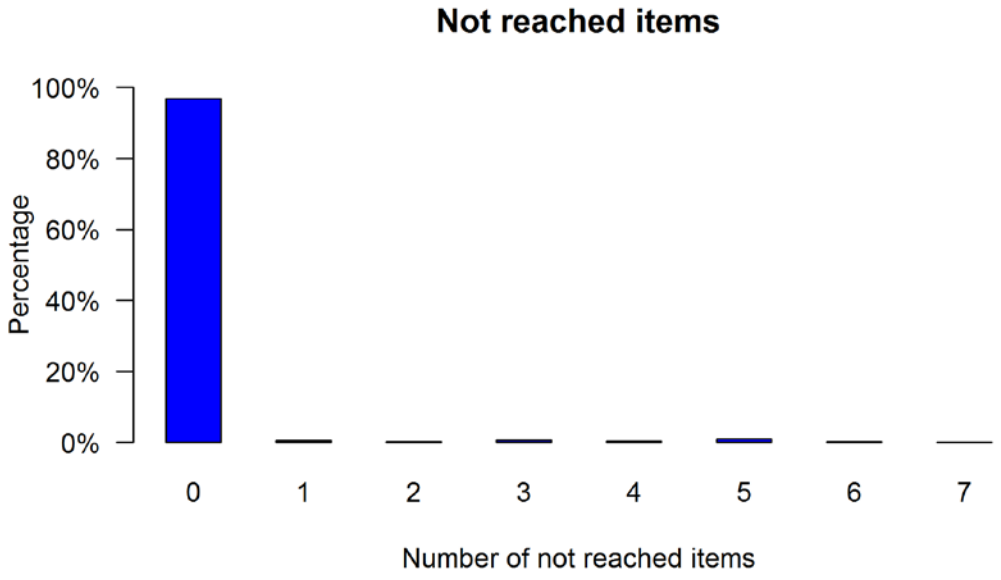

Figure 3. Number of not reached items

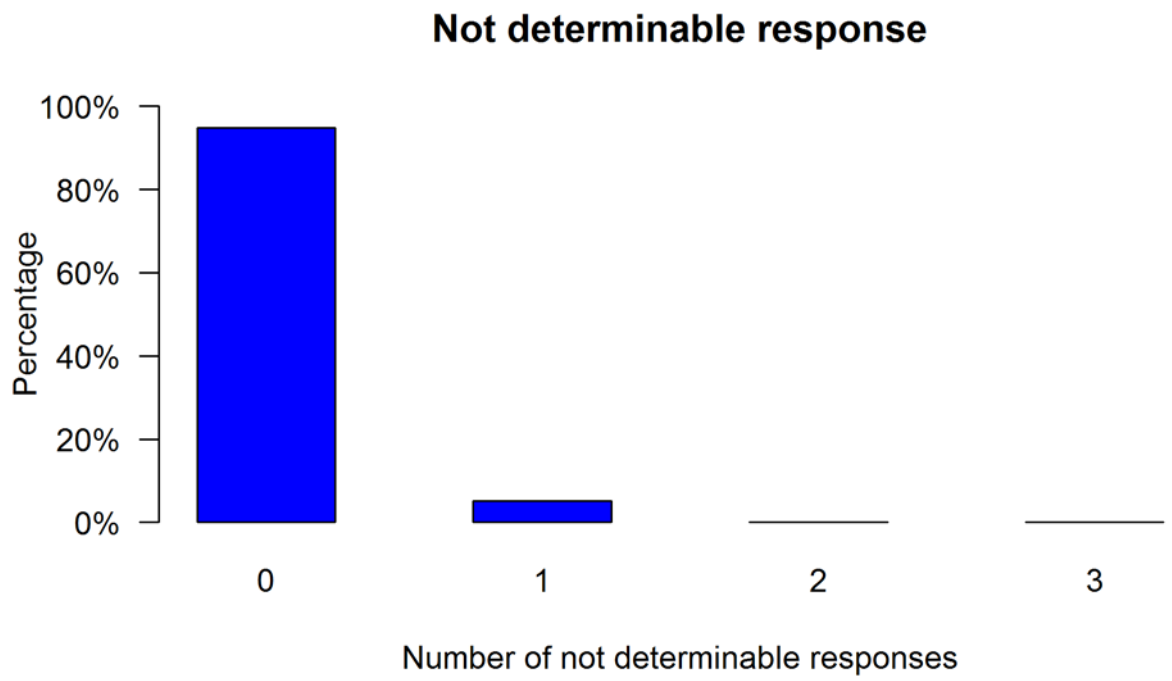

Figure 4. Number of not-determinable items

Because the CMC and MA items were aggregated from several subtasks, the missing type could not be determined for some of these items. About 5% of the respondents exhibited 1 not determinable missing value, whereas most of the respondents had no not determinable missing value at all (see Figure 4).

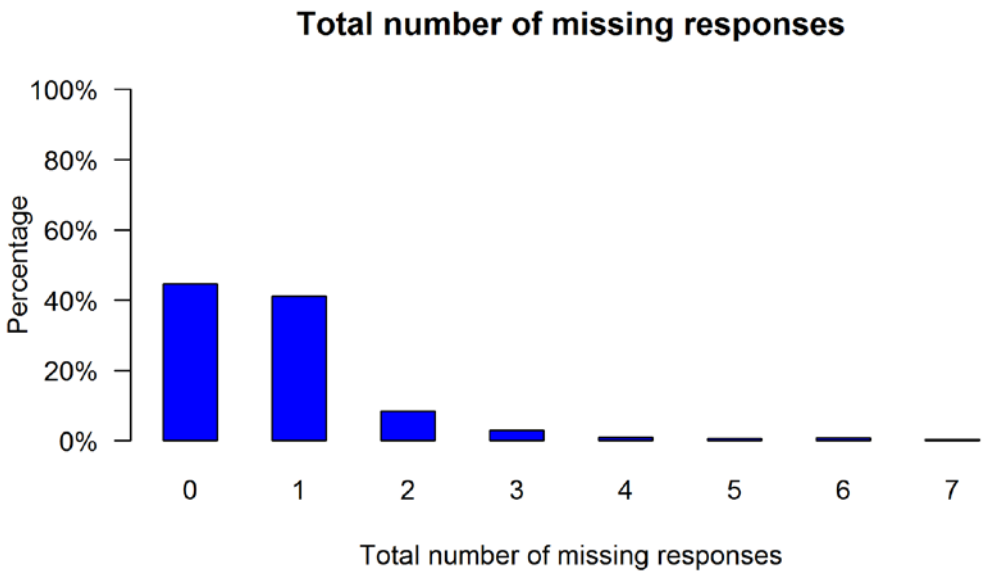

Figure 5. Total number of missing responses

The total number of missing responses, aggregated over invalid, omitted, not reached, and not determinable missing responses per person, is illustrated in Figure 5. Because of the large number of omitted responses, only about 53% of respondents had no missing value at all. In contrast, about 36% had one and 7% even two missing responses.

In sum, there was a considerable amount of omitted responses, particularly for one PCM item (see below). However, other types of missing values were less prevalent and did not indicate severe administration difficulties for the test.

5.1.2 Missing responses per item

Table 2 provides information on the occurrence of different kinds of missing responses per item. The percentage of invalid responses per item (column “NV”) was rather low and increased with the position of the test to a maximum of about 3%. Overall, the percentage of respondents that did not reach an item was rather low (see Figure 6). Similar, there were few invalid and not determinable responses. The largest missing rates for these types of missing responses were about 3% and 5% for item efg12b00s_c. In contrast, there were more omitted responses (column “OM"). In particular, item efg12b00s_c was omitted by about 40% of the respondents. Thus, the large amount of missing responses for item efg12b00s_c suggests that the respondents had difficulties in properly understanding and responding to this matching item. In contrast, for the remaining items no pronounced difficulties in terms of excessive missing rates were observed.

Table 2. Percentage of Missing Values by Item.

	Item	*N*	NR	OM	NV	ND
1.	efg10022s_sc4g12_c	3,764	0.00	3.21	0.08	0.00
2.	efg12b00s_c	2,003	0.05	40.26	3.16	5.06
3.	efg10108s_sc4g12_c	3,594	0.26	5.86	1.44	0.10
4.	efg12d001_c	3,812	1.26	0.59	0.21	0.00
5.	efg12d002_c	3,769	1.59	1.46	0.10	0.00
6.	efg12d003_c	3,555	2.29	6.27	0.10	0.00
7.	efg12d004_c	3,738	2.57	1.28	0.10	0.00
8.	efg12d005_c	3,706	3.16	1.54	0.08	0.00

Note. *N* = Number of valid responses, NR = Percentage of respondents that did not reach item, OM = Percentage of respondents that omitted the item, NV = Percentage of respondents with an invalid response, ND = Percentage of respondents with a not-determinable response.

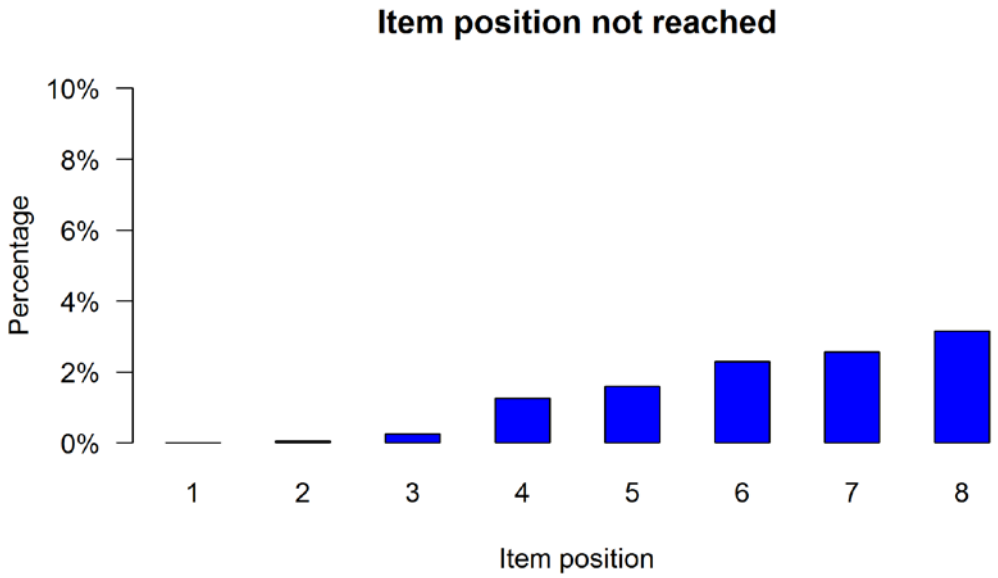

Figure 6. Item position not reached.

Table 3. Item Parameters

	Item	Position	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
1.	efg10022s_sc4g12_c	1		-0.85	0.01	1.04	0.39	.39	0.54	.06
2.	efg12b00s_c	2		-0.08	0.01	0.95	-1.59	.57	0.85	.06
3.	efg10108s_sc4g12_c	3		-1.12	0.02	0.99	-0.36	.37	0.71	.02
4.	efg12d001_c	4	59	-0.40	0.03	1.03	-2.73	.16	0.45	.03
5.	efg12d002_c	5	43	0.37	0.04	0.96	-2.80	.38	1.49	.07
6.	efg12d003_c	6	12	2.53	0.06	0.95	-1.40	.32	1.66	.07
7.	efg12d004_c	7	62	-0.66	0.04	0.98	-1.12	.36	1.46	.06
8.	efg12d005_c	8	42	0.43	0.04	1.05	2.96	.32	1.02	.04

Note. Difficulty = Item difficulty / location, *SE* = Standard error of item difficulty / location, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, r_{it} = Corrected item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model, Q_3 =Average absolute residual correlation for item (Yen, 1983).

Percent correct scores are not informative for polytomous CMC and MC item scores and, therefore, are not reported.

5.2 Parameter Estimates

5.2.1 Item parameters

The third column in Table 3 presents the percentage of correct responses (for simple multiple choice items) in relation to all valid responses for each item. Because there was a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index of item difficulty. The percentage of correct responses varied between 12% and 62% with an average of 44% (*SD* = 20%) correct responses.

Because of a low discrimination of item efg12d001_c, this item was scored with 0.5 (instead of 1.0) points in the PCM analyses. The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 3. The step parameters for polytomous variables are summarized in Table 4. The item difficulties and location parameters were estimated by constraining the mean of the ability distribution to be zero. Due to the large sample size, the standard errors (SE) of the estimated parameters (see Tables 3 and 4) were rather small (all SEs ≤ 0.07). The estimated item difficulties and location parameters ranged from -0.85 (item efg10108s_sc4g12_c) to 2.53 (item efg12d003_c). Thus, they covered a rather broad range; however, there were no items with low difficulty or location parameters.

Table 4. Step Parameters (with Standard Errors) for Polytomous Items

Item	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8
efg10022s_sc4g12_c	-0.81 (0.04)	-0.38 (0.04)	-0.25 (0.03)	0.00 (0.03)	0.31 (0.04)	1.13		
efg12b00s_c	-0.65 (0.07)	-0.87 (0.06)	-0.66 (0.06)	-0.07 (0.05)	0.03 (0.05)	0.07 (0.05)	0.49 (0.06)	1.67
efg10108s_sc4g12_c	-0.58 (0.04)	-0.84 (0.03)	0.57 (0.04)	0.85				

Note. The last step parameter for each item is not estimated and has, thus, no standard error because it is a constrained parameter for model identification.

5.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. Because most items in the English reading test 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 6, the category thresholds of the English reading 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 category thresholds. The respective thresholds ranged from -4.27 (item efg10022s_sc4g12_c) to 3.70 (item efg12b00s_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 1.70, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = .70, WLE reliability = .64) was slightly below established standards. The mean of the category threshold distribution was about 0.84 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. As a consequence, person ability in medium- and low-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors of measurement.

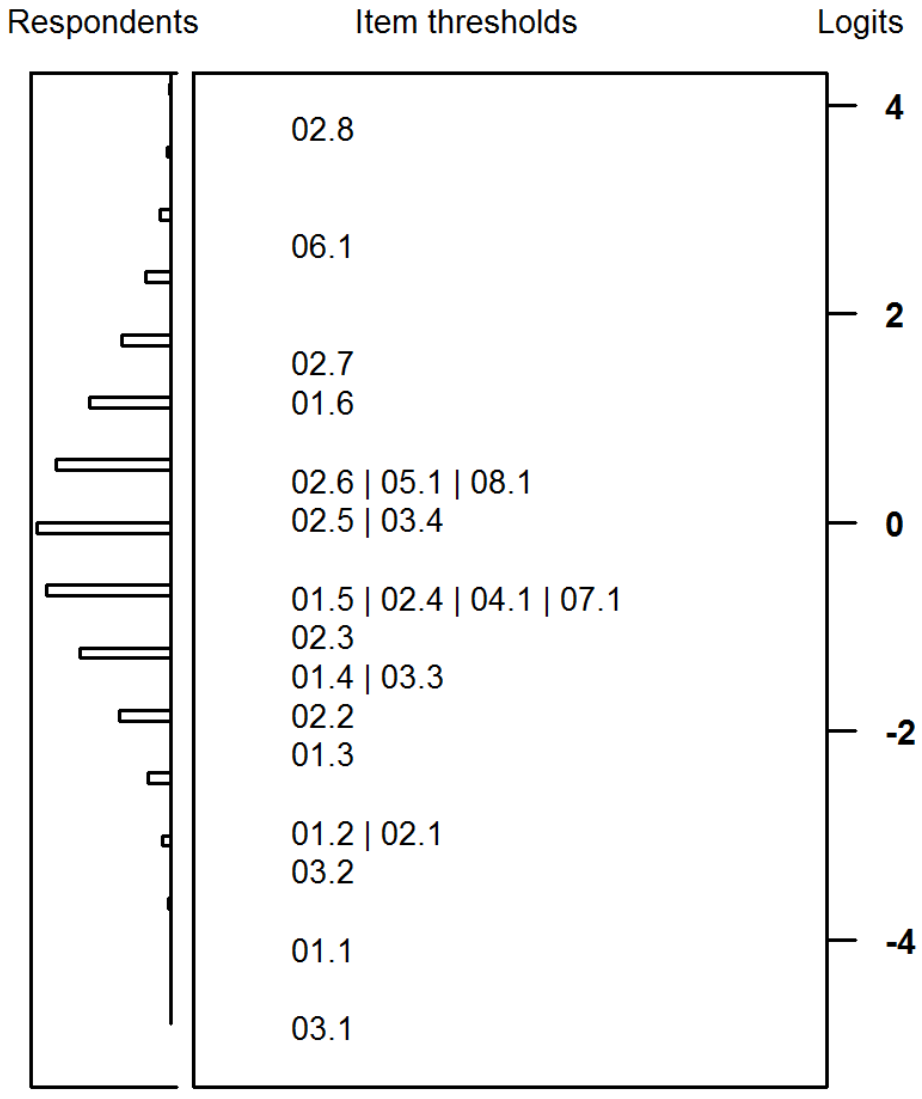

Figure 6. Test targeting. The distribution of person ability in the sample is given on the left-hand side of the graph. The category thresholds of the items are given on the right-hand side of the graph. Each number represents one threshold with the first part (before the dot) corresponding to the item number in Table 3 and the second part indicating the threshold.

5.3 Quality of the test

5.3.1 Item fit

The evaluation of the item fit was performed based on the final scaling model, the PCM. Altogether, item fit was good (see Table 3). For all items, values of the WMNSQ ranged from 0.95 (item efg12d003_c) to 1.05 (item efg12d005_c). Moreover, a visual inspection of the item characteristic curves indicated a good fit of the item to the chosen scaling model. Point-biserial correlations between the item scores and the total rest scores ranged from .16 (item efg12d001_c) to .57 (item efg12b00s_c) and had a mean of .36.

5.3.2 Distractor analyses

In addition to the overall item fit, it was investigated how well the distractors performed in the test by evaluating the point-biserial correlation between each incorrect response (distractor) and the students' total correct score. The point-biserial correlations for the distractors ranged from -.37 to .08 with a mean of -.18. These results indicate that the distractors functioned well.

5.3.3 Differential item functioning

Differential item functioning (DIF) was used to evaluate 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), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). The differences between the estimated item difficulties in the various groups are summarized in Table 5. For example, the column "Male vs. female" reports the differences in item difficulties between men and women; a positive value would indicate that the test 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 allow for DIF to those that only estimate main effects (see Table 6).

Sex: The sample included 1,720 boys and 2,172 girls. On average, female participants had a slightly higher estimated English reading ability than males (main effect = 0.16 logits, Cohen's d = 0.18). One item (efg12d003_c) showed DIF greater than 0.4 logits; however, with 0.57 logits the difference between the two groups was not considered severe. Moreover, the minimum effect test (see Fischer et al., 2016) did not identify significant DIF exceeding 0.4. An overall test for DIF (see Table 6) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). A model comparison using Akaike's (1974) information criterion (AIC) favored the more complex DIF model. In contrast, the Bayesian information criterion (BIC; Schwarz, 1978) that takes the number of estimated parameters into account and, thus, guards against overparameterization of models showed a better fit for the more parsimonious model including only the main effect. Thus, overall, there was no pronounced DIF with regard to sex.

Table 5. Differential Item Functioning

Item	Sex	Books	Migration
	male vs. female	< 100 vs. ≥ 100	without vs. with
efg10022s_sc4g12_c	0.10 (0.11)	-0.13 (-0.16)	-0.07 (-0.08)
efg12b00s_c	0.10 (0.12)	-0.01 (-0.01)	-0.01 (-0.02)
efg10108s_sc4g12_c	-0.03 (-0.04)	-0.03 (-0.03)	-0.10 (-0.12)
efg12d001_c	0.07 (0.09)	-0.39 (-0.47)	0.20 (0.23)
efg12d002_c	-0.03 (-0.04)	0.07 (0.09)	-0.14 (-0.17)
efg12d003_c	-0.57 (-0.66)	0.36 (0.44)	-0.07 (-0.08)
efg12d004_c	0.17 (0.19)	0.21 (0.25)	0.05 (0.05)
efg12d005_c	0.20 (0.23)	-0.09 (-0.10)	0.16 (0.18)
Main effect (DIF model)	-0.10 (-0.12)	-0.45 (-0.54)	0.13 (0.15)
Main effect (Main effect model)	-0.16 (-0.18)	-0.40 (-0.47)	0.16 (0.19)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses.
* Absolute standardized difference is significantly, $p < .05$, greater than 0.4 (see Fischer et al., 2016).

Books: The number of books at home was used as a proxy for socioeconomic status. There were 752 test takers with 0 to 100 books at home and 2,994 test takers with more than 100 books at home. There were considerable average differences between the two groups. Participants with 100 or less books at home performed on average 0.40 logits (Cohen's d = 0.47) lower in reading than participants with more than 100 books. There was no considerable DIF comparing participants with many or fewer books (highest DIF = 0.39 for item efg12d001_c). As a consequence, also the overall test for DIF using the BIC favored the main effects model (Table 7).

Migration background: There were 3,463 participants with no migration background and 410 subjects with a migration background. In comparison to subjects without migration background, participants with migration background had, on average, a slightly lower English reading ability (main effect = 0.16 logits, Cohen's d = 0.19). There was no noteworthy item DIF due to migration background; differences in estimated difficulties did not exceed 0.4 logits (highest DIF = 0.20 for item efg12d001_c). Moreover, the overall test for DIF also favored the main effects model that did not include item-level DIF.

Table 6. Comparisons of Models with and without DIF

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	DIF	3,892	49,562	36	49,640	49,884
	main effect	3,892	49,610	29	49,674	49,874
Books	DIF	3,746	47,622	36	47,700	47,943
	main effect	3,746	47,659	29	47,723	47,921
Migration	DIF	3,873	49,348	36	49,426	49,670
	main effect	3,873	49,357	29	49,421	48,622

5.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1980) model is that all item-discrimination parameters are equal. In order to test this assumption, a generalized partial credit model (GPCM; Muraki, 1992) that estimates discrimination parameters was fitted to the data. The estimated discrimination parameters differed moderately among items (see Table 3). The average discrimination parameter fell at 1.02. Particularly, the discrimination parameter of 0.45 for item efg12d001_c was rather low. Therefore, this item was scored with 0.5 (instead of 1.0) points in the PCM analyses. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 49,312, BIC = 49,507, number of parameters = 31) as compared to the PCM (AIC = 49,389, BIC = 49,539, number of parameters = 24). 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 PCM was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

5.3.5 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 3) were quite low (M = .05, SD = .02)—the largest individual residual correlation was .07—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.

6. Discussion

The analyses in the previous sections reported information on the quality of the English reading test in starting cohort 4 for grade 12 and described how the reading competence scores were estimated. Different kinds of missing responses were examined, item fit statistics were thoroughly checked, and the correlations between the responses and the total correct scores were investigated. Further quality inspections were conducted by examining differential item functioning and testing Rasch-homogeneity. Various criteria indicated a good fit of the items and measurement invariance across various subgroups. Moreover, for most items the number of missing responses were reasonably small. However, the large number of omitted responses for item efg12b00s_c indicates that the respondents had difficulties in understanding the instruction or properly using the matching response format. Because the test was rather short, it had a slightly impaired reliability and did not distinguish well between test takers in the upper ability range. The test was better targeted at mediocre- and low-performing. As a consequence, ability estimates will be precise for low-performing students but less precise for high performing students. In summary, the test had acceptable psychometric properties that allowed the estimation of a unidimensional reading competence score for English as a foreign language.

7. Data in the Scientific Use File

7.1 Naming conventions

The SUF contains 10 items, of which 7 were scored dichotomously (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. These items are marked with a '0_c' at the end of the variable name. A total of 3 items were scored as polytomous variables (CMC and MA items) that are marked with a 's_c' at the end of the variable names. For further details on the naming conventions of the variables see Fuß and colleagues (2016).

7.2 Linking of competence scores

In starting cohort 4, the English reading competence tests administered in grades 10 and 12 include 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, the two tests were linked using a fixed item parameter approach (cf. Fischer, Rohm, Gnambs, & Carstensen, 2016). Because two items referring to two texts were identical in both tests, we fixed the item parameters of these items in grade 12 to those item parameters derived in grade 10. In this way, the two tests are placed on a common scale that allows meaningful comparisons across grades.

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the two grades showed a non-negligible shift in item difficulties. The differences in item difficulties between grades 10 and 12 and the tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 7. For the two common items measurement invariance of the location and step parameters

was supported, that is, the minimum effects hypothesis test was not significant (α = .05). For the, N = 3,666 respondents that participated at both measurement occasions, linked English reading competence scores were estimated using the item parameters derived in grade 10 as fixed parameters in grade 12.

Table 7. Differential Item Functioning Analyses between Starting Cohorts 3 and 4.

	Item	Δσ	$SE_{\Delta\sigma}$	*F*
1.	efg10022s_sc4g12_c	0.08	0.02	14.71
	- step 1	-0.26	0.06	19.19
	- step 2	-0.24	0.06	19.70
	- step 3	0.12	0.05	5.43
	- step 4	0.08	0.05	2.57
	- step 5	0.14	0.05	6.55
3.	efg10108s_sc4g12_c	-0.08	0.03	7.14
	- step 1	-0.36	0.05	47.28
	- step 2	0.25	0.05	23.72
	- step3	-0.04	0.05	0.48

Note. Δσ = Difference in item difficulty parameters between the two grades (negative values indicate easier items in grade 12); $SE_{\Delta\sigma}$ = 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, 3666) = 59.87. A non-significant test indicates measurement invariance.

$^{*}p < .05$

7.3 English reading competence scores

In the SUF, manifest reading competence scores are provided in the form of two different WLEs (“efg12_sc1” and “efg12_sc1u”) including their respective standard error (“efg12_sc2” and “efg12_sc2u”). For “efg12_sc1u”, person abilities were estimated using the linked item difficulty parameters. As a result, the WLE scores provided in “efg12_sc1u” can be used for longitudinal comparisons between grades 10 and 12. The resulting differences in WLE scores can be interpreted as development trajectories across measurement points. In contrast, the WLE scores in “efg12_sc1” are not linked to the underlying reference scale of grade 10. 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 A. In the IRT scaling model, the polytomous CMC and MA variables (and item efg12d001_c) were

scored as 0.5 for each category. In case less than 50% of subtasks in a PCM item were missing, these missing values were imputed with the expected score from the Rasch analyses presented above. Subsequently, the PCM scores were recalculated based on the imputed values. No imputations were performed if more than 50% of subtasks were missing for a given respondent. 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

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, United Kingdom: University Press.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (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. (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., 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: University of Bamberg, National Educational Panel Study.

Haberkorn, K., Pohl, S., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 4 in Ninth Grade* (NEPS Working Paper No. 16). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

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.

Kiefer, T., Robitzsch, A. & Wu, M. (2017). *TAM: Test analysis modules*. R package version 1.99999-31. URL: https://CRAN.R-project.org/package=TAM

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. ETS *Applied Psychological Measurement, 16*, 159-176. doi: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: Otto-Friedrich-Universität, Nationales Bildungspanel.

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

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

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

Rupp, A. A., Vock, M., Harsch, C., & Köller, O. (2008*). Developing standards-based assessment tasks for English as a first foreign language: context, processes, and outcomes in Germany* (Vol. 1). Waxmann.

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. *Zeitschrift für Erziehungswissenschaft, 14*, 67-86. doi:10.1007/s11618-011-0182-7

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational Measurement for Applied Researchers*. Singapore, SG: 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. doi:10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

Appendix

Appendix A: R-Syntax for estimating WLEs grade 12 of starting cohort 4

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

# load competence data
dat <- read_sav("SUF for competencies in SC 4.sav")

# items of the English competence test
items <- c("efg10022s_sc4g12_c", "efg12b00s_c",
           "efg10108s_sc4g12_c", "efg12d001_c", "efg12d002_c",
           "efg12d003_c", "efg12d004_c", "efg12d005_c")


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

# 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)
```