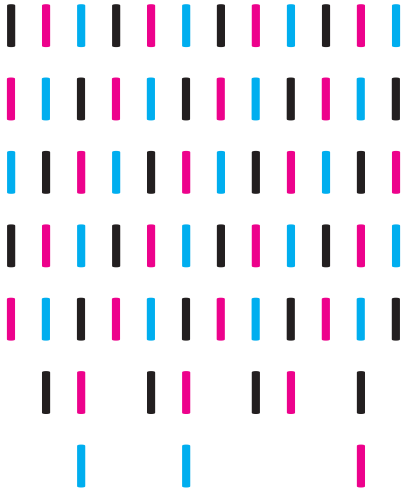

NEPS SURVEY PAPERS

Timo Gnambs

NEPS TECHNICAL REPORT FOR ENGLISH READING COMPETENCE: SCALING RESULTS OF STARTING COHORT 5 (WAVE 12)

NEPS Survey Paper No. 53
Bamberg, March 2019

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 5 (Wave 12)

Timo Gnambs

Leibniz Institute for Educational Trajectories, Bamberg, Germany

E-mail address of lead author:

timo.gnambs@lifbi.de

Bibliographic data:

Gnambs, T. (2019). *NEPS Technical Report for English Reading Competence: Scaling Results of Starting Cohort 5 (Wave 12)* (NEPS Survey Paper No. 53). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP53:1.0

Acknowledgements:

Various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (e.g., Gnambs, 2017a, 2017b) to facilitate the understanding of the presented results.

NEPS Technical Report for English Reading Competence: Scaling Results of Starting Cohort 5 (Wave 12)

Abstract

The National Educational Panel Study (NEPS) examines 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 for English as a foreign language that was administered in wave 12 of Starting Cohort 5 (students). The reading competence test in English included 23 items with multiple choice response formats. The test was administered to 3,490 individuals (60% women). About half of the respondents received the test in a proctored setting at their private homes (N = 1,666), whereas the remaining participants (N = 1,824) worked on unproctored, web-based tests. The responses of the participants 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 item response model. Furthermore, test fairness could be confirmed for different subgroups. A limitation of the test was that it included few difficult items, resulting in rather imprecise proficiency estimates for high-ability students. Overall, the English reading competence test had acceptable psychometric properties that allowed for an estimation of reliable competence scores. 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, 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 (2019).

Most of the competence data are scaled using models based on item response theory (IRT). Because these 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 test are described in Pohl and Carstensen (2012).

In this paper, the results of these analyses are presented for a reading competence test for English as a foreign language that was administered in wave 12 of Starting Cohort 5 (students). First, the main concepts of the competence test and the test design are introduced. Then, the competence data 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 Reading Competence for English as a Foreign Language

2.1 Conceptual Framework

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 five short texts that were accompanied by five 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 five texts were accompanied by 23 items with different response formats (see Table 1). Most items had simple multiple choice formats including four or five response options with one being correct and three or four response options functioning as distractors (i.e., they were incorrect). One item was a complex multiple choice (CMC) item consisting of several subtasks that had to be rated as true, false, or information not given in the text that was subsequently combined into a single polytomous variable. Examples of the different response formats are given in Pohl and Carstensen (2012) and Gehrer, Zimmermann, Artelt and Weinert (2012).

Table 1

Number of Items by Different Response Formats

Response format	**Number of Items**
Simple multiple choice items	22
Complex multiple choice items	1
Total number of items	23

The test administration followed an experimental design. About half of the respondents received the test as a computer-based test (CBT). The test administrators visited the respondents at their private homes and presented the competence test on a laptop. Thus, the respondents were administered the English reading competence test in a proctored setting. The remaining respondents were administered a web-based test (WBT). These respondents finished the competence test in an unproctored setting.

The study assessed different competence domains including, among others, mathematical competence, German reading competence, and English as a foreign language. 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 English test was either administered as the first or the second test (i.e., after the German reading test or the mathematics test). A detailed description of the study design is available on the NEPS website (http://www.neps-data.de).

Table 2

Sample Descriptions

	Computer-based	**Web-based**
Sample size	1,654	1,658
Women	59%	60%
Migration background	8%	9%
Mean age (*SD*)	28.05 (3.53)	28.13 (3.89)

3 Data

The test was administered to a total of 3,490[1] students (60% women). However, 178 respondents had less than three valid responses on the English reading competence test. Therefore, the psychometric analyses are based on a sample of 3,312 students (cf. Pohl & Carstensen, 2012). Basic sociodemographic information of the CBT and WBT samples is summarized in Table 2.

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 complex multiple choice items that are not determined. Invalid responses occurred, for example, 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 or lack of motivation, not all persons finished the test. All missing responses after the last valid response were coded as not reached. Because complex multiple choice 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 complex multiple choice 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). 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) with Gauss-Hermite quadrature (21 nodes). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

Complex multiple choice 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 partial credit item was scored as missing. Response categories of polytomous variables with less than $N = 200$ responses were collapsed in order to avoid possible estimation problems. This occurred for

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.

the two lower categories of the polytomous item; in this case, 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 multiple choice 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 3.

4.3 Checking the Quality of the Test

The reading competence test in English was specifically constructed for the 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 a complex multiple choice item to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the multiple choice items in a Rasch (1960) model. The fit of the subtasks was evaluated based on the weighted mean square (WMNSQ), the respective *t*-value, 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.

After aggregating the subtasks to polytomous variables, the fit of the dichotomous and polytomous items to the partial credit model (Masters, 1982) was evaluated using the weighted mean square (WMNSQ) statistic, the respective *t*-value, 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. 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, age, the number of books at home (as a proxy for cultural capital), migration background (see Pohl & Carstensen, 2012, for a description of these variables), and assessment mode (CBT versus WBT). Differential item functioning (DIF) was examined using a multigroup item response 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 were significantly larger than 0.4 and, thus, were 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 Q_3 falling below .20 indicate essential unidimensionality.

4.4 Software

The item response models were estimated with the *TAM* package version 3.0-21 (Robitzsch, Kiefer, & Wu, 2018) in *R* version 3.5.2 (R Core Team, 2018).

5 Results

5.1 Missing Responses

5.1.1 Missing responses per person

Missing responses can occur when respondents omit items. As illustrated in Figure 1 most respondents did not skip any item. However, there were notable differences between the two assessment modes. In the proctored CBT condition, more respondents omitted items (about 39%) as compared to the unproctored WBT condition (about 29%). About 28% and 19% skipped a single item. Participants with multiple omitted items were rare (about 11% and 10% of the two subsamples).

Missing responses that could not be determined (in the polytomous items) or invalid responses were not observed, neither in the CBT nor in the WBT condition.

Another source of missing responses is items that were not reached by the respondents because they aborted the test, for example, because the time limit was reached or a lack of motivation. These missing values refer to items after the last valid response. As illustrated in Figure 2, about 41% of the respondents in the procotored CBT condition and 66% of the respondents in the unproctored WBT condition did not abort the test and were administered all 23 items. About 67% and 81% of the two subsamples received at least 10 items. This indicates that the test was slightly too long for the limited testing time.

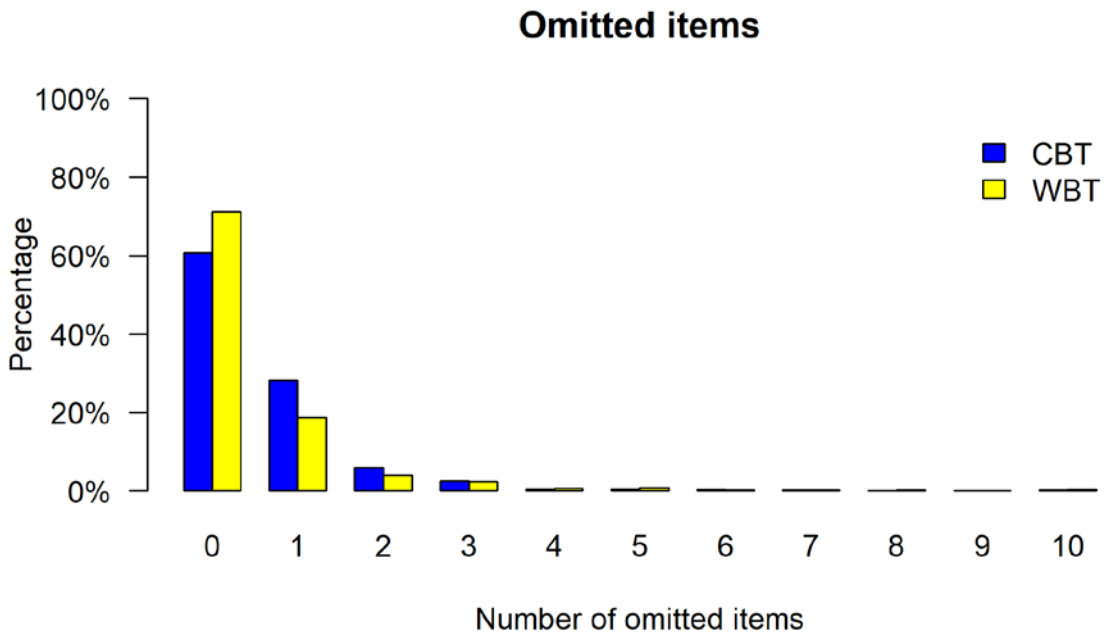

Figure 1. Number of omitted items by assessment mode (CBT = proctored computer-based testing, WBT = unproctored web-based testing)

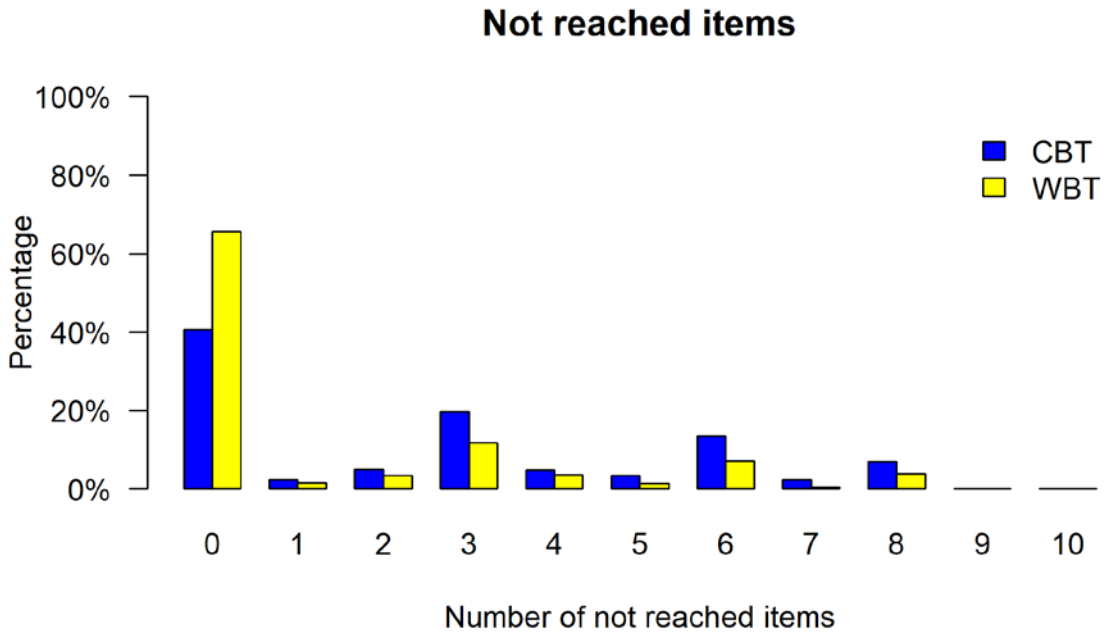

Figure 2. Number of not reached items by assessment mode (CBT = proctored computer-based testing, WBT = unproctored web-based testing)

With an item's progressing position in the test, the amount of persons that did not reach an item rose to about 59% in the CBT condition and 34% in the WBT condition (see Figure 3). The last items were reached by only few respondents. Thus, it seems that many respondents were

unable to finish the test within the allocated time span. This indicates that the testing time might have been too short for the difficulty of the administered test.

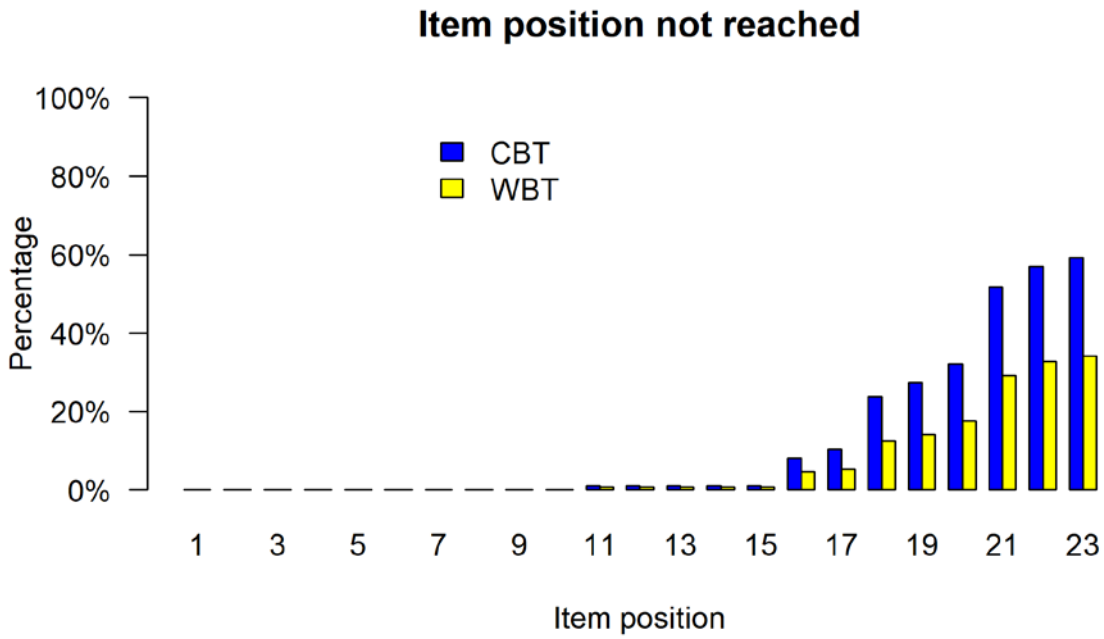

Figure 3. Item position not reached by assessment mode (CBT = proctored computer-based testing, WBT = unproctored web-based testing)

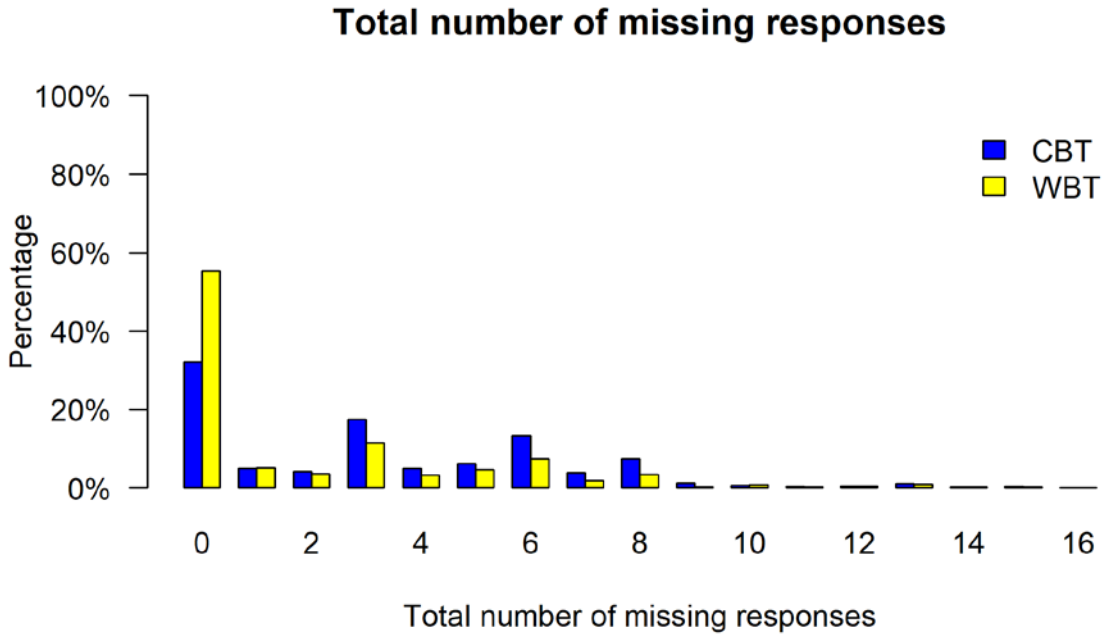

Figure 4. Total number of missing responses by assessment mode (CBT = proctored computer-based testing, WBT = unproctored web-based testing)

The total number of missing responses, aggregated over omitted and not reached missing responses per person, is illustrated in Figure 4. Because the majority of the sample did not reach the end of the test, there was a substantial number of missing values. In the unproctored CBT condition the median number of missing responses was 3; only about 32% had no missing response at all. In the WBT condition, about 55% of the respondents had no missing response at all.

In sum, the amount of missing responses was rather large (particularly in the CBT condition) because many respondents did not reach the end of the test.

5.1.2 Missing responses per item

Table 3 provides information on the occurrence of different kinds of missing responses per item for the two assessment modes. The number of omitted responses varied across items between 0.91% and 8.04% (*Mdn* = 2.05%) and were, thus, negligible. In contrast, there were substantially more missing responses because participants did not reach the item. Particularly, the last three items referring to the fifth text were frequently not reached.

Table 3

Percentage of Missing Values by Item

		CBT			WBT		
Pos.	**Item**	N_v	**OM**	**NR**	N_v	**OM**	**NR**
1	efs121010_c	1629	1.51	0.00	1635	1.09	0.00
2	efs121020_c	1632	1.33	0.00	1633	1.39	0.00
3	efs121030_c	1639	0.91	0.00	1631	1.39	0.00
4	efs121040_c	1630	1.45	0.00	1631	1.57	0.00
5	efs121050_c	1630	1.45	0.00	1629	1.75	0.00
6	efs121060_c	1625	1.75	0.00	1619	2.23	0.00
7	efs121070_c	1628	1.57	0.00	1622	1.99	0.00
8	efs121080_c	1638	0.97	0.00	1631	1.51	0.00
9	efs121090_c	1628	1.57	0.00	1624	2.05	0.00
10	efs121100_c	1615	2.36	0.00	1612	2.65	0.00
11	efs122010_c	1612	1.45	1.09	1609	1.39	0.78
12	efs122020_c	1610	1.57	1.09	1598	2.05	0.78
13	efs122030_c	1611	1.51	1.09	1598	1.99	0.78
14	efs122040_c	1601	2.12	1.09	1587	2.59	0.78
15	efs122050_c	1585	3.08	1.09	1566	3.68	0.78
16	efs123011_c	1467	3.26	8.04	1541	1.27	4.70
17	efs123012_c	1418	3.87	10.40	1511	2.53	5.25

		CBT			WBT		
Pos.	**Item**	N_v	**OM**	**NR**	N_v	**OM**	**NR**
18	efs124010_c	1182	4.66	23.88	1386	2.41	12.55
19	efs12402s_c	1069	8.04	27.33	1292	6.57	14.05
20	efs124030_c	1007	6.95	32.16	1265	4.64	17.61
21	efs125010_c	671	7.62	51.81	1043	6.03	29.31
22	efs125020_c	642	4.17	57.01	1033	3.14	32.81
23	efs125030_c	597	4.72	59.19	977	5.07	34.26

Note. Pos. = Item position within test. N_v = Number of valid responses, NR = Percentage of respondents that did not reach item, OM = Percentage of respondents that omitted the item.

5.2 Parameter Estimates

To avoid potentially biased parameter estimates resulting from mode effects (unproctored versus proctored settings), the following analyses are limited to the proctored CBT sample. Thus, the unproctored WBT sample was excluded from the scaling procedure. Information on the measurement invariance across assessment modes is given in section 5.3.3.

5.2.1 Item parameters

The fourth column in Table 5 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 45% and 89% with an average of 69% (*SD* = 14%) correct responses.

The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 5. The step parameters for the polytomous item 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 4 and 5) were rather small (all SEs ≤ 0.10). The estimated item difficulties and location parameters ranged from -2.51 (item `efs121030_c`) to 0.27 (item `efs125010_c`). Thus, there were rather few difficult items.

Table 4

Step Parameters (with Standard Errors) for Polytomous Item in CBT sample

Item	Step 1	Step 2
efs12402s_c	-0.43 (0.06)	-0.43 (0.04)

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

Table 5

Item Parameters for CBT sample

Item	Pos.	Item format	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q3
efs121010_c	1	MC	70.79	-1.18	0.06	0.97	-1.09	.35	1.13	.03
efs121020_c	2	MC	54.79	-0.25	0.05	1.09	-4.13	.20	0.62	.03
efs121030_c	3	MC	88.96	-2.51	0.09	1.06	0.97	.11	0.45	.03
efs121040_c	4	MC	81.32	-2.02	0.07	0.94	-1.24	.31	1.37	.04
efs121050_c	5	MC	66.25	-0.88	0.06	1.09	3.54	.19	0.55	.04
efs121060_c	6	MC	79.07	-1.63	0.07	0.93	-1.92	.33	1.45	.04
efs121070_c	7	MC	84.98	-2.26	0.08	0.96	-0.71	.31	1.22	.04
efs121080_c	8	MC	85.71	-2.35	0.08	0.98	-0.39	.24	1.17	.03
efs121090_c	9	MC	85.06	-2.24	0.08	0.90	-1.94	.33	1.74	.05
efs121100_c	10	MC	54.35	-0.21	0.05	1.06	3.17	.26	0.72	.04
efs122010_c	11	MC	86.72	-2.37	0.08	0.99	-0.08	.23	1.11	.03
efs122020_c	12	MC	74.10	-1.33	0.06	0.98	-0.54	.28	1.07	.03
efs122030_c	13	MC	73.20	-1.28	0.06	0.98	-0.50	.30	1.15	.04
efs122040_c	14	MC	57.06	-0.32	0.06	1.02	0.85	.29	0.90	.03
efs122050_c	15	MC	65.50	-0.84	0.06	0.93	-2.72	.37	1.43	.05
efs123011_c	16	MC	75.40	-1.37	0.07	0.90	-3.00	.43	1.60	.05
efs123012_c	17	MC	45.10	0.25	0.06	1.06	2.78	.28	0.71	.03
efs124010_c	18	MC	69.56	-1.12	0.07	0.97	-0.86	.39	1.16	.04
efs12402s_c	19	PC	NA	-0.24	0.04	0.93	-2.20	.39	0.77	.04
efs124030_c	20	MC	63.56	-0.67	0.07	1.00	-0.02	.38	1.03	.04
efs125010_c	21	MC	44.66	0.27	0.09	1.09	-2.67	.28	0.66	.03
efs125020_c	22	MC	66.57	-1.06	0.10	0.94	-1.37	.45	1.28	.04
efs125030_c	23	MC	51.27	-0.32	0.09	1.09	2.56	.28	0.66	.04

Note. Pos. = Item position, Format = Response format (MC = Multiple Choice, PC = Partial Credit), 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 item scores and, therefore, are not reported.

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 some items in the English test were polytomous, we calculated Thurstonian thresholds for each response (Wu, Adams, Wilson, & Haldane, 2007). 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 5, the category thresholds of the English 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 -2.51 (item `efs12103_c`) to 0.94 (item `efs12402s_c`) and, thus, spanned a rather broad range; albeit, there were rather few thresholds in the upper region of the proficiency distribution. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 1.00, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = .73, WLE reliability = .65) was acceptable. The mean of the category threshold distribution was about 1.11 logits below the mean person ability distribution. Thus, although the items covered a wide range of the ability distribution, the items were 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.

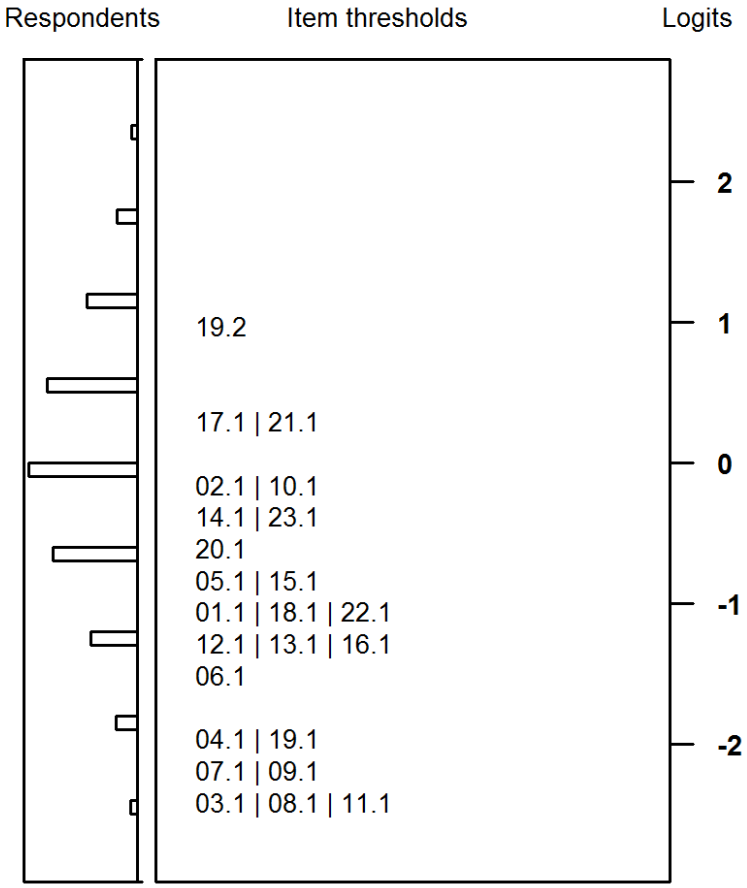

Figure 5. Test targeting. The distribution of person ability in the sample is given on the left-hand side of the graph. The category thresholds of the items are given on the right-hand side of the graph. Each number represents one threshold with the first part (before the dot) corresponding to the item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

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. Again, the test quality was examined for the CBT sample only, while excluding the unproctored WBT sample. Altogether, item fit was good (see Table 5). No item exhibited a WMNSQ greater than 1.10 or a *t*-value of the WMNSQ greater than 6. Moreover, a visual inspection of the item characteristic curves (ICC) showed no pronounced deviation from the expected ICC for the items. One item exhibited a somewhat small item-total correlation of r_{it} = .11 (`efs121030_c`). However, an inspection of the respective ICC did not identify a noteworthy misfit. The item-total correlations for the remaining items fell between .19 (`efs121050_c`) and .45 (`efs125020_c`).

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 scores. The point-biserial correlations for the distractors ranged from -.42 to -.07 with a mean of -.23. 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), migration background, age, and test position (see Pohl & Carstensen, 2012, for a description of these variables). In addition, we examined mode effects by comparing the proctored CBT sample and the unproctored WBT sample. All analyses were limited to items with at least 50 valid responses for each response category in each group. Because of varying sample sizes in the different subgroups, the reported DIF analyses included different item sets. The differences between the estimated item difficulties in the various groups are summarized in Table 6. 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 7).

Sex: The CAPI sample included 684 men and 970 women. There was no pronounced difference in the English reading competence between male and female participants (main effect = 0.07 logits, Cohen's d = 0.07). One item (`efs121080_c`) showed DIF greater than 0.50 logits (Cohen's *d* = 0.58). An overall test for DIF (see Table 7) 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 DIF model over the more parsimonious model including only the main effect. 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 favored the more parsimonious main effects model. Moreover, the estimated main effects for sex were rather similar in both models (Cohen's *d* = 0.07 versus 0.08). Thus, there was no pronounced DIF with regard to sex.

Table 6

Differential Item Functioning for CBT Sample

Item	Sex	Age	Books	Migration	Position	Mode
	male vs. female	< 27 vs. ≥ 27 years	< 100 vs. ≥ 100	without vs. with	first vs. second	CBT vs. WBT
efs121010_c	-0.22 (-0.22)	0.05 (0.05)	-0.01 (-0.01)		0.04 (0.04)	-0.07 (-0.07)
efs121020_c	-0.17 (-0.17)	0.00 (0.00)	0.07 (0.07)	0.02 (0.02)	0.03 (0.03)	0.09 (0.09)
efs121030_c	-0.23 (-0.23)	0.08 (0.08)	0.39 (0.40)		0.19 (0.19)	0.08 (0.08)
efs121040_c	0.03 (0.03)	0.12 (0.12)	0.19 (0.19)		-0.17 (-0.17)	-0.34 (-0.34)
efs121050_c	-0.28 (-0.28)	0.15 (0.15)	-0.02 (-0.02)	-0.17 (-0.22)	-0.11 (-0.11)	-0.01 (-0.01)
efs121060_c	0.00 (0.00)	-0.26 (-0.26)	0.01 (0.01)		0.06 (0.06)	0.08 (0.08)
efs121070_c	0.13 (0.13)	-0.26 (-0.26)	0.26 (0.26)		-0.01 (-0.01)	-0.22 (-0.22)
efs121080_c	0.58* (0.58)	-0.25 (-0.25)	-0.01 (-0.01)		0.08 (0.08)	-0.26 (-0.26)
efs121090_c	-0.28 (-0.28)	0.13 (0.13)	0.07 (0.07)		-0.20 (-0.20)	-0.17 (-0.17)
efs121100_c	-0.06 (-0.06)	0.05 (0.05)	-0.03 (-0.03)	0.12 (0.15)	0.20 (0.20)	0.11 (0.11)
efs122010_c	-0.01 (-0.01)	0.04 (0.04)	0.03 (0.03)		-0.23 (-0.23)	-0.12 (-0.12)
efs122020_c	0.26 (0.26)	0.04 (0.04)	-0.14 (-0.14)		-0.08 (-0.08)	0.02 (0.02)
efs122030_c	-0.45 (-0.45)	0.16 (0.16)	-0.19 (-0.19)		0.04 (0.04)	0.02 (0.02)
efs122040_c	-0.13 (-0.13)	0.15 (0.15)	-0.02 (-0.02)	0.22 (0.22)	-0.05 (-0.05)	0.16 (0.16)
efs122050_c	0.04 (0.04)	-0.16 (-0.16)	0.02 (0.02)	0.18 (0.18)	-0.12 (-0.12)	-0.01 (-0.01)
efs123011_c	0.08 (0.08)	-0.18 (-0.18)	-0.19 (-0.19)		-0.22 (-0.22)	0.11 (0.11)
efs123012_c	-0.04 (-0.04)	0.05 (0.05)	-0.08 (-0.08)		0.36 (0.26)	0.13 (0.13)
efs124010_c	0.38 (0.38)	-0.02 (-0.02)	-0.13 (-0.13)		0.16 (0.16)	-0.07 (-0.07)
efs12402s_c	0.19 (0.19)	0.10 (0.10)	0.01 (0.01)		0.17 (0.17)	0.18 (0.18)
efs124030_c	0.25 (0.25)	0.14 (0.14)	-0.06 (-0.06)		0.11 (0.11)	0.21 (0.21)
efs125010_c	-0.34 (-0.34)	0.03 (0.03)	0.01 (0.01)		-0.28 (-0.28)	0.27 (0.27)

Item	Sex	Age	Books	Migration	Position	Mode
	male vs. female	< 27 vs. ≥ 27 years	< 100 vs. ≥ 100	without vs. with	first vs. second	CBT vs. WBT
efs125020_c	0.04 (0.04)	-0.52 (-0.52)	0.03 (0.03)		0.11 (0.11)	-0.04 (-0.04)
efs125030_c	-0.25 (-0.25)	-0.35 (-0.35)	0.19 (0.19)		0.08 (0.08)	0.15 (0.15)
Main effect (DIF model)	-0.07 (-0.07)	0.34 (0.35)	0.01 (0.01)	0.28 (0.35)	0.25 (0.25)	0.15 (0.13)
Main effect (Main effect model)	-0.08 (-0.08)	0.31 (0.32)	0.02 (0.02)	0.28 (0.35)	0.20 (0.20)	0.09 (0.08)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses.
* Absolute standardized difference is significantly, $p < .05$, greater than 0.40 (see Fischer et al., 2016).

Age: There were 811 test takers younger than 27 years and 843 test takers aged 27 years or older. There were small average differences between the two groups. Younger participants performed on average 0.34 logits (Cohen's $d = 0.35$) better on the English test as compared to older respondents. One item exhibited a DIF effect of 0.51 (`efs125020_c`). However, the main effects did not differ substantially, whether DIF was modeled or not ($d = 0.35$ versus 0.31). Moreover, the information criteria also favored the more parsimonious main effect model that did not account for minor DIF effects (Table 7).

Books: The number of books at home was used as a proxy for cultural capital. There were 833 test takers with less than 100 books at home and 652 test takers with 100 or more books at home. There were no differences between the two groups; participants with fewer books at home performed comparably as participants with more books (Cohen's $d = 0.01$). There was no considerable DIF comparing participants with many or fewer books (highest DIF = 0.39 for item `efs121030_c`). As a consequence, also the overall test for DIF using the AIC and BIC favored the main effects model (Table 7).

Migration background: There were 1,517 participants without migration background and 137 respondents with a migration background. In comparison to subjects without migration background, participants with migration background had, on average, a slightly lower English reading competence (main effect = 0.28 logits, Cohen's $d = 0.35$). There was no noteworthy item DIF due to migration background. Therefore, the overall test for DIF using the information criteria also favored the main effects model that did not include item-level DIF.

Test position: There were 828 participants that received the English reading competence test first and 826 respondents that received the test after finishing another competence test. Participants receiving the English test second performed, on average, slightly worse on the English test (main effect = 0.25 logits, Cohen's $d = 0.25$). There was no noteworthy item DIF due to test position; the largest difference in estimated difficulties was 0.36 logits for item `efs123012_c`. Moreover, the overall test for DIF using the information criteria (see Table 7) also favored the main effects model that did not include item-level DIF.

Table 7

Comparisons of Models with and without DIF

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	DIF model	1654	33760	48	33856	34115
	Main effect	1654	33828	26	33880	34020
Age	DIF model	1654	33764	48	33860	34120
	Main effect	1654	33796	26	33848	33988
Books	DIF model	1485	30481	48	30577	30831
	Main effect	1485	30497	26	30549	30686
Migration	DIF model	1654	10520	11	10542	10602
	Main effect	1654	10523	7	10537	10575
Position	DIF model	1654	33784	48	33880	34140
	Main effect	1654	33816	26	33868	34009
Mode	DIF model	3312	70631	48	70727	71020
	Main effect	3312	70687	26	70739	70898

Assessment mode: Participants received either a proctored computerized test (CBT) or an unproctored web-based test (WBT). Therefore, mode effects were also examined. There were 1,654 respondents in the CBT condition and 1,658 respondents in the WBT condition. As expected, there were no pronounced differences in the subjects' mean abilities between the two modes (0.15 logits, Cohen's d = 0.13). There was also no noteworthy DIF (largest DIF = 0.34 logits for item `efs121040_c`). Also, the overall test for DIF using the BIC favored the main effects model that did not include item-level DIF (see Table 7).

5.3.4 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; Muraki, 1992) that estimates discrimination parameters was fitted to the data. The estimated discrimination parameters differed moderately among items (see Table 5). The average discrimination parameter fell at 1.04 (*SD* = 0.35). Particularly, the discrimination parameter of 0.45 for item `efs121030_c` was somewhat low. However, an inspection of the respective item characteristic curve of the PCM indicated an adequate fit. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 33,633, BIC = 33,888, number of parameters = 47) as compared to the PCM (AIC = 33,880, BIC = 34,015, number of parameters = 25). 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 5) were quite low (M = 0.04, SD =

0.01)—the largest individual residual correlation was 0.05 (item efs121090_c)—and, thus, indicated an essentially unidimensional test. Because the English test is constructed to measure a single dimension, a unidimensional English competence score was estimated.

6 Discussion

The analyses in the previous sections reported information on the quality of the English reading competence test that was administered in Starting Cohort 5. Different kinds of missing responses were examined, item fit statistics and item characteristic curves were evaluated, and item discriminations 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. However, the number of missing responses was somewhat large because many respondents did not finish the test in time. The test had a satisfactory reliability and distinguished well between test takers. However, the test was slightly better targeted at mediocre- and low-performing students and covered the high ability spectrum less well. 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 English reading competence score.

7 Data in the Scientific Use Files

7.1 Naming conventions

The SUF for Starting Cohort 5 contains 23 items, of which 22 were scored dichotomously (multiple choice items) with 0 indicating an incorrect response and 1 indicating a correct response and 1 being scored polytomously. The latter is marked with a 's_c' at the end of the variable name. For further details on the naming conventions of the variables see Fuß and colleagues (2019).

7.2 English competence scores

In the SUF, manifest English competence scores are provided in the form of WLEs (efs12_sc1) including their respective standard error (efs12_sc2). These WLEs are corrected for the position of the English test within the test battery. The R Syntax for estimating the WLEs is provided in the Appendix. Because no substantial DIF was found for the proctored CBT and the unproctored WBT conditions, WLEs for respondents receiving the WBT were estimated using the fixed item parameters from the CBT scaling (see Table 5)[2]. In the IRT scaling model, the polytomous variable was scored as 0.5 for each category. For respondents who did not take part in the English test or who did not give enough valid responses no WLEs were 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

2 The test taking behavior in unproctored testing cannot be properly supervised and, thus, might not be comparable to proctored settings (see Kröhne, Gnambs, & Goldhammer, 2019). Therefore, we inspected the response times in for respondents in the WBT condition. For 72 respondents exhibiting breaks (with no test interaction) of more than five minutes during the test no WLEs were estimated because they were suspected to adopt different test taking strategies.

their analyses or estimate plausible values. A description of these approaches can be found in Pohl and Carstensen (2012).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723. https://doi.org/10.1109/TAC.1974.1100705

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. (2019). *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.

Gnambs, T. (2017a). *NEPS Technical Report for English Reading Competence: Scaling Results of Starting Cohort 4 for Grade 10* (NEPS Survey Paper No. 26). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Gnambs, T. (2017b). *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.

Kröhne, U., Gnambs, T., & Goldhammer, F. (2019). *Disentangling setting and mode effects for online competence assessment*. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), Education as a lifelong process (2nd ed., pp. 171-193). Wiesbaden, Germany: Springer VS. https://doi.org/10.1007/978-3-658-23162-0_10

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

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

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests.* (NEPS Working Paper No. 14). Bamberg: 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 (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, DK: The Danish Institute of Education Research.

Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules*. R package version 2.12-18. URL: https://CRAN.R-project.org/package=TAM

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.

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

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

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

Wu, M., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0*. Camberwell, Australia: Acer Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145. https://doi.org/10.1177/014662168400800201

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

Appendix

R-Syntax for estimating WLEs

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

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

# 23 items of the English competence test
items <- c("efs121010_c", "efs121020_c",
           "efs121030_c", "efs121040_c",
           ...)

# identify polytomous item
f <- items %in% c("efs12402s_c")


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

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

# item fit
tam.fit(mod)

# WLE
tam.wle(mod)
```