NEPS SURVEY PAPERS

Sven Rieger, Nicolas Hübner, and Wolfgang Wagner

NEPS TECHNICAL REPORT FOR ENGLISH READING: SCALING RESULTS FOR THE ADDITIONAL STUDY THURINGIA

NEPS Survey Paper No. 39
Bamberg, June 2018

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

NEPS Technical Report for English Reading:

Scaling Results for the Additional Study Thuringia

Sven Rieger, Nicolas Hübner, & Wolfgang Wagner

Hector Research Institute of Education Sciences and Psychology,
University of Tübingen

E-mail address of lead author:

sven.rieger@uni-tuebingen.de

Bibliographic data:

Rieger, S., Hübner, N., & Wagner, W. (2018). *NEPS Technical Report for English Reading – Scaling Results of the Additional Study Thuringia* (NEPS Survey Paper No. 39). Bamberg: Leibniz-Institute for Educational Trajectories, National Educational Panel Study.

We would like to thank Steffi Pohl and Claus. H. Carstensen for developing and providing standards for the technical reports. Additionally, we thank Eve Sarah Müller for her excellent support and valuable feedback.

NEPS Technical Report for English Reading: Scaling Results for the Additional Study Thuringia

Abstract

The National Educational Panel Study (NEPS) aims to investigate the development of competences across the whole life span. It also develops tests to assess different competence domains. In order to evaluate the quality of these competence tests, a wide range of analyses are carried out by using item response theory (IRT). This paper describes the data and results of analyzing the English reading competence test that was used in the additional study Thuringia. The items were originally designed for Grade 10 students but – due to the lack of Grade 12 tests in this domain at the time when the first assessment took place – they were used in the English reading competence test in two consecutive waves (2009/10 and 2010/11). In sum, 2,252 students participated in the test in these two waves. The English test consisted of 33 items (distributed among two booklets), representing different levels of the Common European Framework of References, ranging from level B1 to C1. A Rasch model was used for scaling the data. Item fit statistics and differential item functioning were investigated. The results showed that the items exhibited good item fit and measurement invariance across various groups. The reliability was modest, which might be due to the fact that item difficulties were rather low compared to students' competences. The paper also provides some information about the data available in the Scientific Use File, ConQuest- and TAM-syntaxes for scaling the data.

Keywords

item response theory, scaling, English reading competence, scientific use file

Content

1 Introduction

Within the National Educational Panel Study (NEPS) different competences are measured coherently across the life span. Tests have been developed for different competence domains. These include, among other things, reading competence, mathematical competence, scientific literacy, information and communication technologies literacy, metacognition, vocabulary, and domain-general cognitive functioning.

Most of the competence data are scaled using models that are based on item response theory (IRT). Because most of the competence tests were developed specifically for implementation in NEPS, several analyses have been 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 scales are described in Pohl and Carstensen (2012).

This paper presents the results of the English reading competence test in two waves of the additional study Thuringia. In this study, items developed by the Institute of Quality Development in Education (IQB) were composed for the English reading test used over two consecutive years (2011 through 2013) to test secondary-school students' English reading competences in their final year of Gymnasium (type of school leading to upper secondary education and Abitur). More detailed information about the aims of this study as well as further information about the test can be found on the NEPS website[1].

The present report draws strongly on previous technical reports such as Hübner, Rieger, & Wagner (2016), Pohl, Haberkorn, Hardt, and Wiegand (2012) and Pohl and Carstensen (2012). It includes extracts from these previous reports.

2 Testing English Reading Competence

The framework and item development for the English reading competence was led by the Institute for Educational Quality Improvement (IQB) and is described in Rupp, Vock, Harsch & Köller (2008) and NEPS (2011a; 2011b). In the following, we will point out specific aspects of the English reading competence paper-and-pencil test that are necessary for understanding the scaling results presented in this paper.

The items are arranged in units. Thus, on the test, students must usually read one or more texts and must subsequently answer multiple test items related to it. All items were developed by trained experts and corresponded to the National Educational Standards and the Common European Framework of Reference (NEPS, 2011a; 2011b). Item difficulties range between the levels B1 and C1.

There are three types of response formats on the English reading test. These are simple multiple choice (MC), complex multiple choice (CMC), and multiple matching (MA) items. For MC items, the test taker has to choose the correct answer out of several—usually four—response options. For CMC tasks, a number of subtasks with three response options are presented. MA

[1] https://www.neps-data.de/en-us/datacenter/dataanddocumentation/additionalstudythuringia.aspx

items require the test taker to match a specific sentence, phrase or word to a text or part of a text.

Tables 1 and 2 show how the difficulty levels of the GER and response formats are distributed across the items as well as the booklets.

Table 1

Difficulty Levels of Items in the English Test

Difficulty Levels	**Frequency**
Level B1	5
Level B1/B2	4
Level B2	16
Level C1	8
Total number of items	33

Number of Items by Difficulty Levels and by Booklets	**1**	**2**
Level B1	5	1
Level B1/B2	-	4
Level B2	8	8
Level C1	8	8
Total number of items	21	21

Table 2

Response Formats of Items in the English Test

Response format	**Frequency**
Single multiple choice	5
Complex multiple choice	8
Multiple Matching	20
Total number of items	33

Response format	**1**	**2**
Single multiple choice	1	5
Complex multiple choice	8	-
Multiple Matching	12	16
Total number of items	21	21

3 Data

A description of the design of the study, the sample, as well as the instruments that were used can be found on the NEPS website.[2] A total of 2,252 particpants took the English reading test: 1368 in 2009/2010 (Wave 1) and 884 in 2010/2011 (Wave 2)[3]. All subjects gave at least one valid answer so that for every subject, a competence score was estimated.

4 Analyses

This section briefly describes the analyses that were computed; these included inspecting the various missing responses, scaling the data, and examining the psychometric quality of the test.

4.1 Missing Responses

There are different types of missing responses in competence test data. These include missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach, and d) items that are missing by design (e.g., due to the different booklets). Missing responses provide information about how well the test worked (e.g., time limits, whether participants understood the instructions, how participants handled different response formats), and they need to be accounted for in the estimation of item and person parameters. We thoroughly inspected the occurrence of missing responses per person. This provided an indication of how well the test takers coped with the test. We then examined the occurrence of missing responses per item in order to obtain some information about how well the items performed. In addition, information was available about whether students did not take the English reading test (e.g., due to student tardiness) but did take at least one of the other competence tests (mathematics, biology, or physics). This missing code is referred to as e) missing by non-participation.

4.2 Scaling Model

In order to estimate the item and person parameters for English reading competence, a Rasch model (Rasch, 1960) was used and estimated in ConQuest 4.2 (Wu, Adams, Wilson, & Haldane 2007).

Item parameters are estimated difficulties for dichotomous variables in the Rasch model. Ability estimates for English competence were estimated as weighted maximum likelihood estimates (WLEs; Warm, 1989). Person parameter estimation in NEPS is described by Pohl and Carstensen (2012a), whereas the data available in the SUF are described in Section 7.

Plotting the item parameters in relation to the ability estimates of the persons was used in order to judge how well the item difficulties were targeted toward the test persons' abilities (see Figure 5). The test targeting provides some information about the precision of the ability estimates at different levels of ability.

[2] https://www.neps-data.de/en-us/datacenter/dataanddocumentation/additionalstudythuringia.aspx

[3] The dataset contains 2,260 persons.

4.3 Checking the Quality of the Scale

The items used on the English reading competence test were originally constructed for Grade-10 students. To ensure that the test featured appropriate psychometric properties in the sample of secondary-school students as well, the quality of the test was examined again with several analyses.

The item fit of dichotomous items was examined by analyzing them via a Rasch model (Rasch, 1960/1980), the weighted (or "infit") mean square (WMNSQ), the respective t-value, and correlations between the item scores and the total score. In accordance with Pohl and Carstensen (2012), items with a WMNSQ > 1.15 (t-value > |6|) were considered to have a noticeable item misfit, and items with a WMNSQ > 1.20 (t-value > |8|) were considered to have a considerable item misfit, and their performance was further investigated. Correlations between an item score and the total score (equal to the discrimination as computed in ConQuest) greater than 0.3 were considered good, greater than 0.2 acceptable, and below 0.2 problematic. Overall, the judgment of item fit was based on all fit indicators.

Our aim was to construct an English reading competence test that measured the same construct in all participants. If any items favored a certain subgroup (e.g., items that were easier for males than for females), measurement invariance would be violated, and a comparison of competence scores between the subgroups (e.g., males and females) would be biased and thus unfair.[4] We addressed the issue of measurement invariance by investigating test fairness for the variables gender, books at home (as a proxy for socioeconomic status; see Pohl and Carstensen, 2012 for a description of these variables), and wave (i.e., to which of the two waves do subjects belong?). Differential item functioning (DIF) was estimated by applying a multifaceted IRT model in ConQuest in which the main effects of the subgroups and the differential effects of the subgroups on item difficulty were modeled. Differences in the estimated item difficulties between the subgroups were evaluated. On the basis of our experiences with the preliminary data (e.g., Pohl & Carstensen, 2012), we judged absolute differences in estimated difficulties that were greater than 1 logit as having very strong DIF, absolute differences between 0.6 and 1 as worthy of further investigation, differences between 0.4 and 0.6 as considerable but not significant, and differences smaller than 0.4 as not having any considerable DIF. In addition to computing DIF analyses at the item level, we investigated test fairness by comparing a model that included differential item functioning with a model that estimated only main effects but no DIF.

The English reading competence data were scaled with the Rasch model, which assumes Rasch homogeneity. Nonetheless, Rasch homogeneity is an assumption that might not hold for empirical data. We therefore checked for deviations from uniform discrimination. We estimated item discrimination by applying the Birnbaum model (2PL; Birnbaum, 1968) with the TAM package in R (Robitzsch, Kiefer, & Wu, 2017; R Core Team, 2017).

[4] It should be noted that differential item functioning may also reflect valid differences between subgroups – that is, item impact (Zumbo, 1999).

5 Results

In this section, the key scaling results of the three waves of the additional study Thuringia are presented.

5.1 Missing Responses

In this subsection, we first report the number of missing responses that can be categorized into the different types of missing responses as described in Chapter 4.1 per person and the total number of missing responses per person. Afterwards, we describe the missing responses per item.

5.1.1 Missing responses per person

Figure 1 shows the number of *invalid responses* per person. As can be seen, only few of the participants (5.16%) produced any invalid responses. The maximum number of invalid responses was, however, 21.

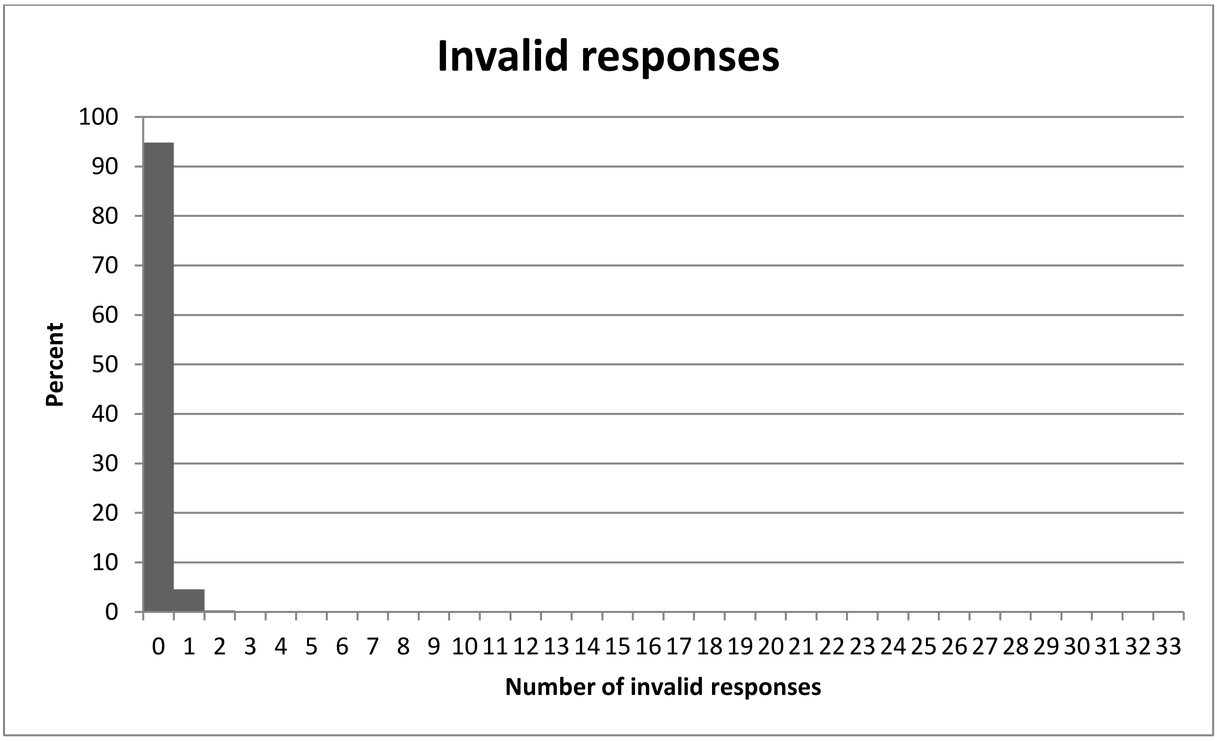

Figure 1. Number of invalid responses per person.

Figure 2 shows the number of *omitted responses* per person. As can be seen in Figure 2, only 3.59% of the participants skipped at least one item. Overall, 0.89% of the participants omitted five or more items.

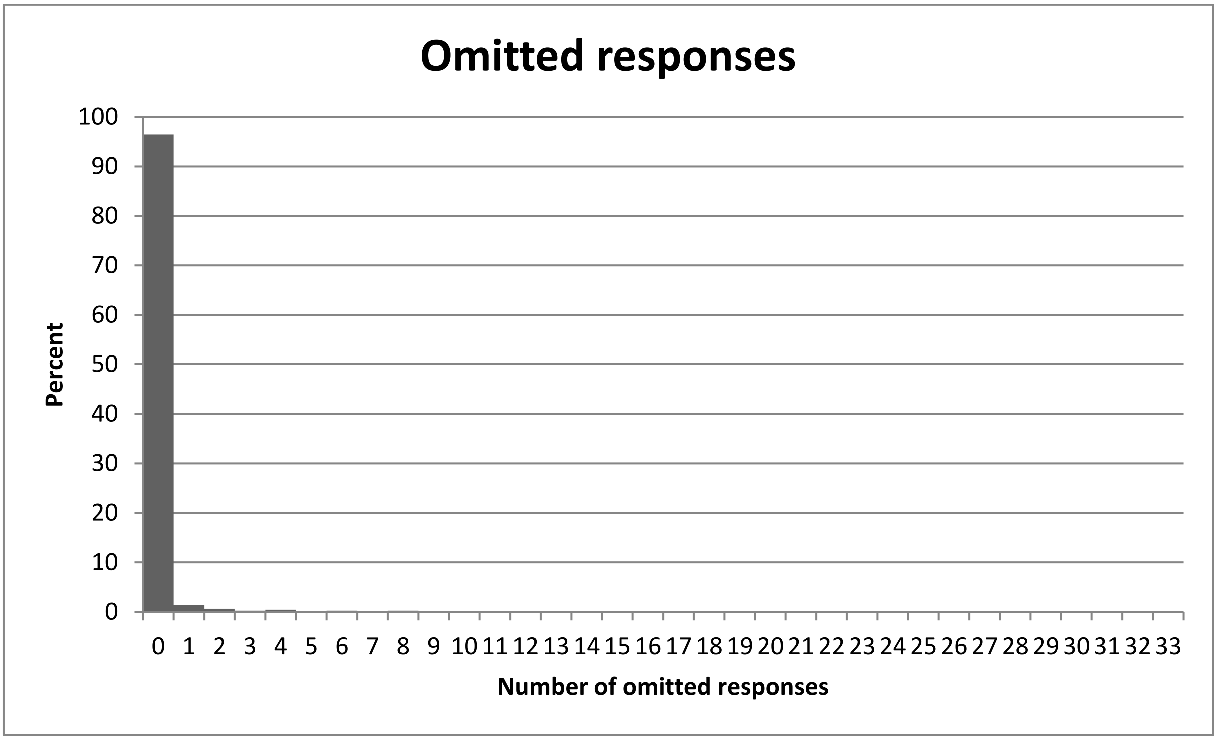

Figure 2. Number of omitted responses per person.

By definition, every item after the last item that was completed is labeled *not reached*. As Figure 3 shows, most participants (98.98%) reached the end of the test.

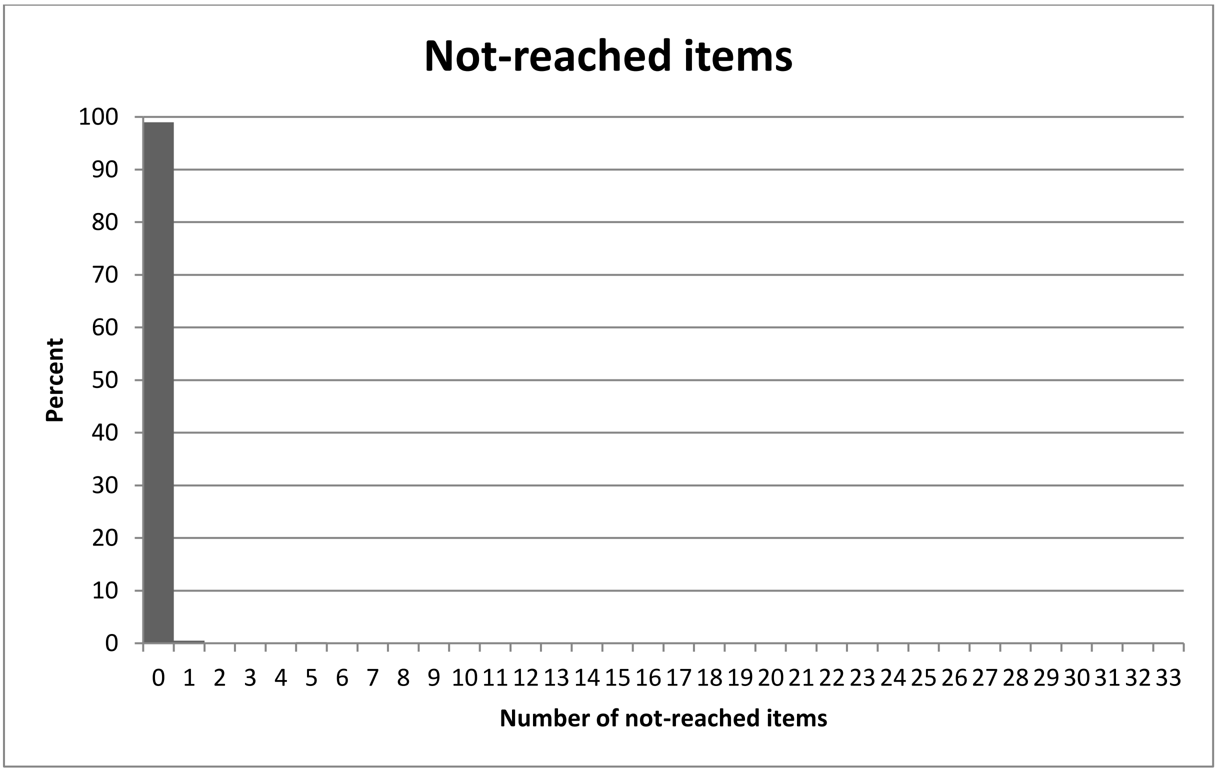

Figure 3. Number of not-reached items per person.

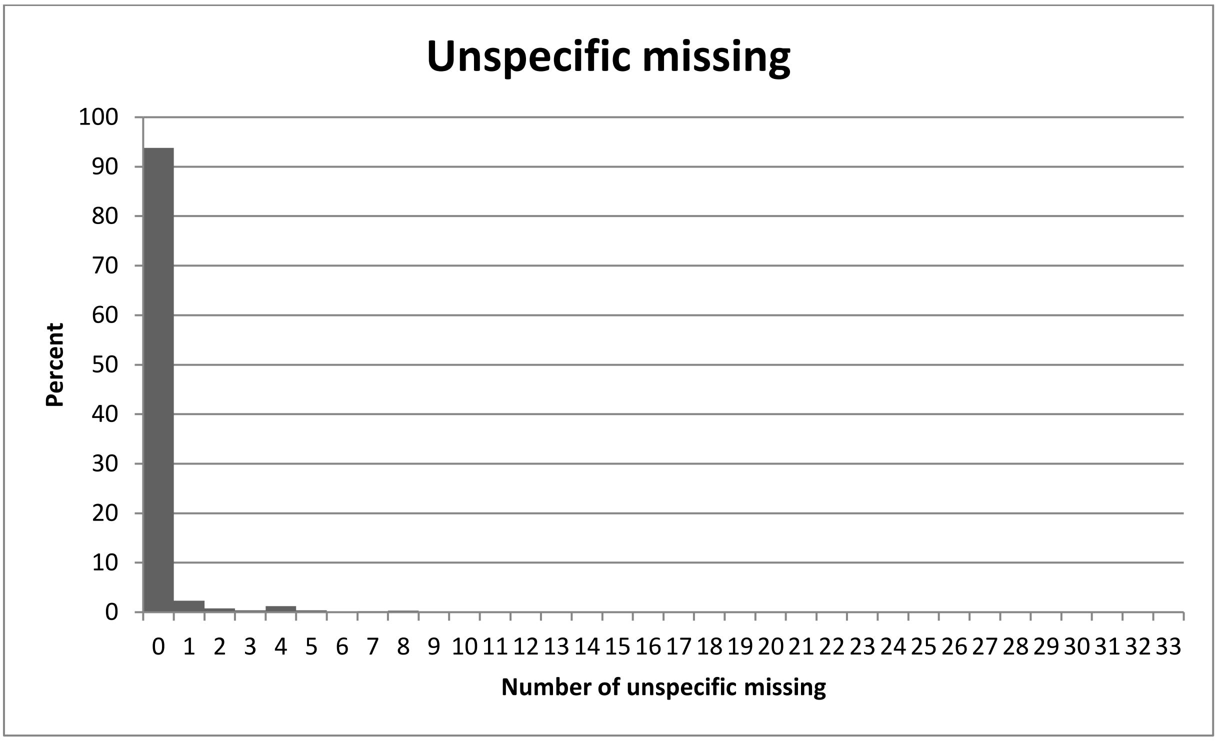

Figure 4. Number of unspecific missing per person.

Figure 4 shows the number of *unspecific missing* per person. As can be seen, 6.19% of the participants had this type of missing. Overall, 1.45% of the participants had more than five unspecific missings.

Overall, 99.78% of the participants had no items that were missing by *non-participation*. Only 0.22% of the students did not take the English reading test but did take at least one of the other tests.

The total number of missing responses (excluding those missing by non-participation, missing by design, and unspecific missing) aggregated across the invalid, omitted, and not-reached missing responses per person is illustrated in Figure 4. On average, the participants produced 0.24 (SD = 1.20) missing responses. Moreover, 91.02% of the participants had no missing responses at all. Only 1.44% of the participants had five or more missing responses.

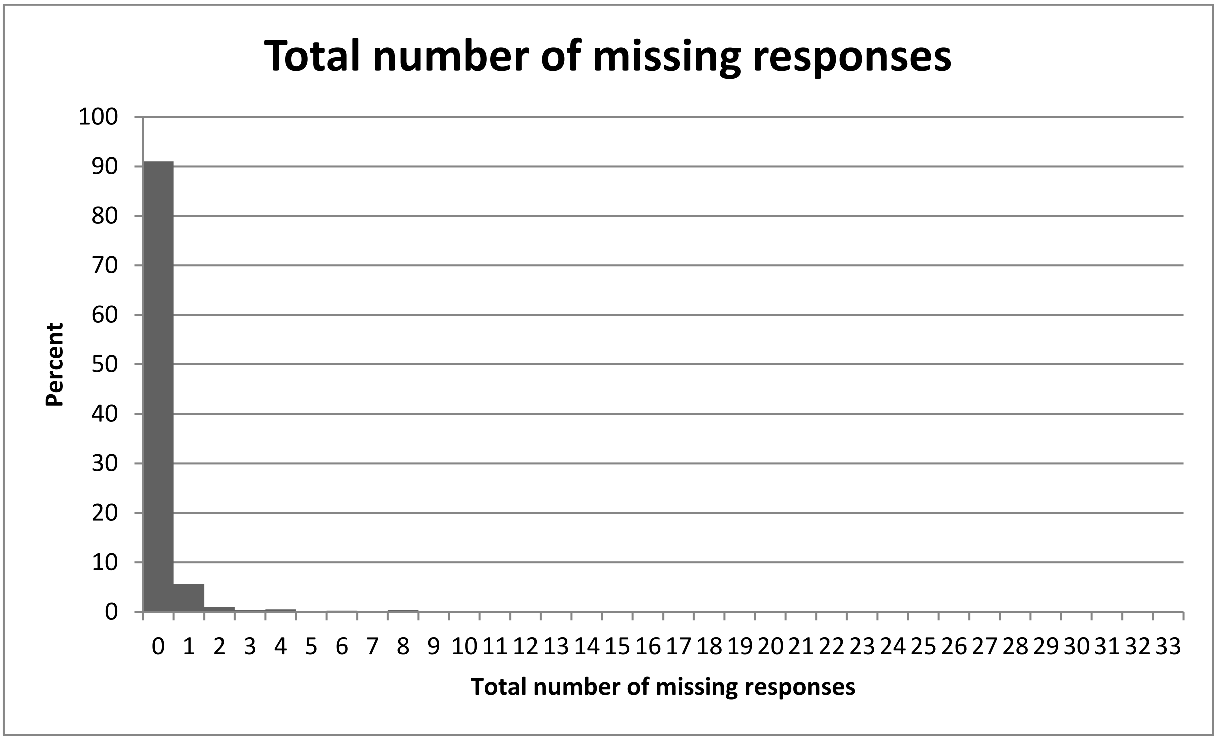

Figure 5. Total number of missing responses.

5.1.2 Missing responses per item

Table 3 provides information about the occurrence of the different kinds of responses that were missing per item. A maximum of 1.0% of the participants failed to reach items (column 5). No item had an omission rate that exceeded 1.1% (column 6). Overall, the percentage of invalid responses per item (column 7) was very low (the maximum was 0.6% for item efl008a_c). The percentage of items that were missing by non-participation (column 9) was very low (the maximum was 0.2%). All students who took the test had 12 items that were missing by design (column 10).

5.2 Parameter Estimates

5.2.1 Item parameters

The second column in Table 4 shows the percentage of correct responses relative to all valid responses for each item. Please note that, because there is a nonnegligible number of missing responses, this probability cannot be interpreted as an index of item difficulty. The percentage of correct responses varied from 39.14% to 88.52% with an average of 63.47% (*SD* = 13.36%) correct responses.

For reasons of model identification, in the Rasch model, the mean of the ability distribution was constrained to be zero. The estimated item difficulties (for dichotomous variables) are given in the third column of Table 4. The item difficulties ranged from -2.441 (item efl108a_c) to 0.582 (item efl059e_c) logits with an average difficulty of -0.76 logits (SD = 0.79). Altogether, the item difficulties are somewhat low. The 2PL discrimination parameters ranged from 0.341 to 2.232 (see again Table 4). Item efl065c_c had a negative discrimination (see Table S1 in the Appendix), paradoxically indicating that students with lower ability had a higher probability of solving the item. Therefore, after we rechecked the coding procedure, this item was excluded from further analyses.

Table 3

Missing Values

	Item	Booklet	Position in the test	Number of valid responses	Percentage of not-reached responses	Percentage of omitted responses	Percentage of invalid responses	Percentage of unspecific missing	Percentage of missing by non-participation	Percentage of missing by design
1	efl108a_c	1	9	1116	0.0	0.1	0.1	0.2	0.2	50.0
2	efl108b_c	1	10	1115	0.0	0.1	0.0	0.3	0.2	50.0
3	efl108c_c	1	11	1094	0.0	0.4	0.3	0.7	0.2	50.0
4	efl108d_c	1	12	1109	0.0	0.1	0.2	0.4	0.2	50.0
5	efl022b_c	1	13	1123	0.1	-	-	-	0.2	50.0
6	efl022c_c	1	14	1120	0.2	-	0.1	-	0.2	50.0
7	efl022d_c	1	15	1121	0.2	-	0.0	-	0.2	50.0
8	efl022e_c	1	16	1122	0.2	-	-	-	0.2	50.0
9	efl022f_c	1	17	1120	0.3	-	-	-	0.2	50.0
10	efl022g_c	1	18	1119	0.3	-	-	-	0.2	50.0
11	efl022h_c	1	19	1118	0.3	-	0.0	-	0.2	50.0
12	efl022i_c	1	20	1119	0.3	-	-	-	0.2	50.0
13	efl008a_c	1,2	1 / 5	2211	-	0.5	0.6	0.9	0.2	-
14	efl008b_c	1,2	2 / 6	2217	-	0.4	0.6	0.7	0.2	-

	Item	Booklet	Position in the test	Number of valid responses	Percentage of not-reached responses	Percentage of omitted responses	Percentage of invalid responses	Percentage of unspecific missing	Percentage of missing by non-participation	Percentage of missing by design
15	efl008c_c	1,2	3 / 7	2209	-	0.5	0.4	1.1	0.2	-
16	efl008e_c	1,2	4 / 8	2187	-	1.1	0.2	1.7	0.2	-
17	efl075a_c	1,2	8 / 20	2196		0.6	0.3	1.5	0.2	-
18	efl075b_c	1,2	7 / 19	2192	0.2	0.7	0.2	1.7	0.2	-
19	efl075c_c	1,2	6 / 18	2189	0.2	0.7	0.3	1.7	0.2	-
20	efl075d_c	1,2	5 / 17	2190	0.2	0.7	0.3	1.7	0.2	-
21	efl057a_c	1,2	21 / 21	2196	1.0	-	-	-	0.2	-
22	efl065a_c	2	1	1126	-	-	0.1	-	0.2	49,8
23	efl065b_c	2	2	1125	-	-	0.2	-	0.2	49,8
24	efl065c_c	2	3	1127	-	-	0.1	-	0.2	49.8
25	efl065d_c	2	4	1127	-	-	0.1	-	0.2	49.8
26	efl059a_c	2	9	1094	0.1	0.5	0.2	0.8	0.2	49.8
27	efl059b_c	2	10	1059	0.1	1.1	0.4	1.5	0.2	49.8
28	efl059c_c	2	11	1093	0.1	0.6	0.1	0.8	0.2	49.8
29	efl059d_c	2	12	1099	0.1	0.4	0.1	0.8	0.2	49.8
30	efl059e_c	2	13	1069	0.1	1.0	0.2	1.4	0.2	49.8

	Item	Booklet	Position in the test	Number of valid responses	Percentage of not-reached responses	Percentage of omitted responses	Percentage of invalid responses	Percentage of unspecific missing	Percentage of missing by non-participation	Percentage of missing by design
31	efl059f_c	2	14	1086	0.1	0.4	0.2	1.2	0.2	49.8
32	efl059g_c	2	15	1061	0.1	1.1	0.2	1.6	0.2	49.8
33	efl059i_c	2	16	1085	0.1	0.7	0.2	1.0	0.2	49.8

Table 4

Item Parameters of the English Test

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	t-value of WMNSQ	Correlation of item score with total score	Discrimination-2 PL
1	efl108a_c	88.52	-2.441	0.104	1.00	0.1	0.32	0.784
2	efl108b_c	84.20	-2.025	0.093	1.11	2.0	0.25	0.477
3	efl108c_c	44.92	0.257	0.074	1.01	0.5	0.48	0.804
4	efl108d_c	77.89	-1.543	0.084	0.99	-0.2	0.44	0.957
5	efl022b_c	46.52	0.164	0.073	1.22	7.7	0.29	0.334
6	efl022c_c	82.75	-1.911	0.090	1.03	0.6	0.35	0.636
7	efl022d_c	68.84	-0.992	0.077	1.00	0.1	0.46	0.775
8	efl022e_c	71.01	-1.118	0.078	1.12	3.4	0.34	0.459

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	t-value of WMNSQ	Correlation of item score with total score	Discrimination-2 PL
9	efl022f_c	55.59	-0.289	0.073	1.21	7.2	0.30	0.341
10	efl022g_c	48.93	0.045	0.073	1.08	2.8	0.43	0.612
11	efl022h_c	83.71	-1.990	0.092	0.98	-0.3	0.41	0.895
12	efl022i_c	62.16	-0.629	0.074	1.05	1.8	0.44	0.702
13	efl008a_c	53.69	-0.200	0.055	0.98	-1.0	0.53	1.149
14	efl008b_c	67.54	-0.939	0.057	0.96	-2.0	0.53	1.262
15	efl008c_c	58.13	-0.427	0.055	1.00	0.1	0.51	1.069
16	efl008e_c	47.83	0.101	0.055	0.95	-2.4	0.55	1.251
17	efl075a_c	51.87	-0.102	0.055	0.91	-4.4	0.58	1.828
18	efl075b_c	51.51	-0.086	0.055	0.99	-0.8	0.52	1.532
19	efl075c_c	61.55	-0.606	0.056	0.95	-2.6	0.56	1.751
20	efl075d_c	52.74	-0.148	0.055	0.90	-5.4	0.60	1.946
21	efl057a_c	81.63	-1.860	0.066	1.06	1.8	0.35	0.775
22	efl065a_c	71.82	-1.235	0.080	1.07	2.0	0.42	0.880
23	efl065b_c	80.96	-1.867	0.089	1.02	0.3	0.41	1.036
24	efl065c_c	-	-	-	-	-	-	-

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	t-value of WMNSQ	Correlation of item score with total score	Discrimination-2 PL
25	efl065d_c	49.73	-0.012	0.074	1.14	4.6	0.42	0.736
26	efl059a_c	65.23	-0.824	0.078	0.89	-3.5	0.60	1.770
27	efl059b_c	58.32	-0.452	0.077	0.93	-2.3	0.58	1.445
28	efl059c_c	64.74	-0.800	0.078	1.01	0.3	0.50	1.124
29	efl059d_c	73.86	-1.351	0.082	0.85	-4.3	0.61	2.232
30	efl059e_c	39.14	0.582	0.078	0.95	-1.6	0.55	1.328
31	efl059f_c	50.32	-0.017	0.075	1.05	1.8	0.50	0.960
32	efl059g_c	59.53	-0.507	0.077	0.89	-3.7	0.61	1.663
33	efl059i_c	68.08	-0.991	0.079	1.05	1.4	0.46	0.991

Note. Difficulty = Item difficulty / location parameter, SE = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, t = t-value for WMNSQ. Items 24 was excluded from the analyses due to an unsatisfactory item fit.

5.2.2 Person parameters

The person parameters were estimated as WLEs (Pohl & Carstensen, 2012). WLEs will be provided in the next release of the SUF. A description of the data in the SUF can be found in Section 7. An overview of how to work with competence data is presented by Pohl and Carstensen (2012).

5.2.3 Test targeting and reliability

Test targeting focuses on how well item difficulties and person abilities are matched; this is an important criterion for evaluating the appropriateness of the test for the target group. In Figure 5, the item difficulties and person abilities are plotted on the same scale. The items covered the lower part of the ability distribution very well but, in general, they were somewhat too easy. Hence, the test can measure person abilities in the low-ability regions relatively precisely, whereas high person abilities are measured with larger standard errors of measurement.

The mean of the ability distribution was constrained to be zero, and its variance was estimated to be 1.41[5], indicating a reasonable differentiation between the subjects. The reliability of the test was EAP/PV reliability = .81 and WLE reliability = .77.

[5] One item (i.e., efl065c_c) was excluded due to a negative item discrimination (see also below).

Scale (in logits)	**Person ability**	**Item difficulty**
	XX	
	X	
3	X	
	XXX	
	XXX	
	XXXXXXX	
	XXXXXXXXXX	
	XXXXXXXX	
2	XXXXXXXXXXX	
	XXXXXXXXXX	
	XXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXX	
1	XXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXX	29
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	11 13
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	4 18
0	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	5 6 8 24 30
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	1 17
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	3 26
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	7 20 31
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	2 25 27
-1	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	15 32
	XXXXXXXXXXXXXXXXXXXXX	16 22
	XXXXXXXXXXXXXXXXXXXXXX	28
	XXXXXXXXXXXXXXXXXXXX	12
	XXXXXXXXXXXXXX	
	XXXXXXXXXX	21 23
-2	XXXXXXXXXXXX	10 14 19
	XXXXXXX	
	XXXX	
	XXX	9
	XX	
	XX	
-3	X	
	X	
	X	

Figure 6. Test targeting. The distribution of person abilities in the sample is depicted on the left-hand side, with each 'X' representing 3.2 cases. The item difficulties (or location parameters) are depicted on the right-hand side. Each number represents one item with a corresponding position in the test, cf. Table 3.

5.3 Quality of the Test

5.3.1 Item fit

Altogether, the item fit could be considered moderate, with values of the WMNSQ ranging from 0.85 (item efl059d_c) to 1.22 (efl022b_c), cf. column 5 of Table 4. Point-biserial correlations between the item scores and the total scores ranged from 0.25 (item efl108b_c) to 0.61 (item efl059d_c resp. efl059c_c and efl059g_c). Discriminations estimated in the 2PL-model with the TAM package in R ranged from 0.334 (item efl022b_c) to 2.232 (item efl059d_c), cf. Table 4, column 8.

5.3.2 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i. e., measurement invariance with regard to item difficulties). For this purpose, DIF was examined for the variables gender, books, and wave (see Pohl & Carstensen, 2012, for a description of these variables). Table 5 provides a summary of the results of the DIF analyses. According to Pohl and Carstensen (2012), absolute difficulty differences greater than 1 logit can be considered to show very strong DIF. For the current test, no item exceeded this threshold.

The table depicts the differences in the estimated item difficulties between the respective groups. “Male vs. female,” for example, indicates the difference in difficulty $ß_{male}$ - $ß_{female}$. A positive value indicates a higher difficulty for males, whereas a negative value indicates a lower difficulty for males as opposed to females.

Gender: On average, female participants had a higher English reading competence (main effect = 0.174 logits, Cohen’s *d* = 0.146).[6] No item showed DIF greater than 0.6 logits.

Wave: On average, participants of the two waves did not differ in their English reading competence (main effect = 0.064, Cohen’s *d* = 0.054). No item showed DIF greater than 0.6 logits.

Books: On average, participants with many books at home performed better on the English reading competence test (0-200 vs 201-500: main effect = 0.240, Cohen’s *d* = 0.202; 0-200 vs 501-: main effect = 0.630, Cohen’s *d* = 0.530; 201-500 vs 501-: main effect = 0.390, Cohen’s *d* = 0.328). One item (efl108b_c) showed DIF greater than 0.6 logits.

[6] To estimate the effect size the variance of the Rasch model was used.

Table 5

Differential Item Functioning

		Gender	Wave	Books		
	Item	male vs female	1 vs 2	0-200 vs 201-500	0-200 vs 501-	201-500 vs 501-
1	efl108a_c	0.142	-0.138	0.016	-0.038	-0.054
2	efl108b_c	0.208	0.108	-0.398	-0.922	-0.524
3	efl108c_c	-0.318	0.008	-0.320	-0.343	-0.023
4	efl108d_c	-0.332	0.004	-0.174	-0.190	-0.016
5	efl022b_c	-0.038	0.100	-0.187	-0.335	-0.148
6	efl022c_c	-0.148	0.122	0.378	-0.207	-0.585
7	efl022d_c	-0.088	0.190	-0.459	-0.396	0.063
8	efl022e_c	0.010	0.010	0.049	-0.324	-0.373
9	efl022f_c	-0.320	-0.036	-0.215	-0.469	-0.254
10	efl022g_c	-0.166	0.190	0.092	-0.110	-0.202
11	efl022h_c	-0.134	0.052	0.086	-0.318	-0.404
12	efl022i_c	0.030	-0.026	-0.312	-0.402	-0.090
13	efl008a_c	0.312	0.048	-0.169	-0.155	0.014
14	efl008b_c	0.324	-0.070	-0.087	-0.136	-0.049

		Gender	**Wave**	**Books**		
	Item	male vs female	1 vs 2	0-200 vs 201-500	0-200 vs 501-	201-500 vs 501-
15	efl008c_c	-0.082	-0.024	-0.178	0.067	0.245
16	efl008e_c	-0.086	-0.080	-0.162	0.162	0.324
17	efl075a_c	-0.060	-0.242	0.120	0.300	0.180
18	efl075b_c	-0.090	-0.158	-0.041	0.119	0.160
19	efl075c_c	0.022	-0.150	-0.073	0.022	0.095
20	efl075d_c	-0.118	-0.046	-0.035	0.200	0.235
21	efl057a_c	0.176	-0.008	0.256	0.104	-0.152
22	efl065a_c	0.250	-0.216	0.111	0.301	0.190
23	efl065b_c	0.498	0.126	0.258	-0.133	-0.391
24	efl065c_c	-	-	-	-	-
25	efl065d_c	0.166	-0.004	-0.162	0.117	0.279
26	efl059a_c	-0.320	0.040	-0.019	-0.125	-0.106
27	efl059b_c	0.090	0.328	0.106	-0.088	-0.194
28	efl059c_c	0.034	-0.010	0.244	0.336	0.092
29	efl059d_c	-0.298	0.244	-0.075	0.120	0.195
30	efl059e_c	-0.022	0.270	0.193	0.098	-0.095

		Gender	**Wave**	**Books**		
	Item	male vs female	1 vs 2	0-200 vs 201-500	0-200 vs 501-	201-500 vs 501-
31	efl059f_c	0.288	0.128	0.241	0.123	-0.118
32	efl059g_c	0.042	-0.084	0.075	0.108	0.033
33	efl059i_c	-0.056	0.194	0.063	-0.155	-0.218
		0.174	0.064	0.240	0.630	0.390

In Table 6, the models with DIF are compared with those that included only the main effect of the respective variable. Regarding Akaike's (1974) information criterion (AIC), the more parsimonious models including only main effects were preferred over the ones containing differential effects for the variables wave and books. The Bayesian information criterion (BIC; Schwarz, 1978) takes into account the number of estimated parameters and thus prevents the overparameterization of models. Using BIC, the more complex model including DIF was preferred for none of the variables.

Table 6

Comparison of Models With and Without DIF

DIF variable	Model	Number of parameters	AIC	BIC
Gender	main effect	34	50,849.80	50,895.79
	DIF	66	50,843.59	50,932.86
Wave	main effect	34	50,878.32	50,924.31
	DIF	66	50,908.35	50,997.62
Books	main effect	35	42,288.51	42,335.85
	DIF	99	42,327.53	42,461.43

5.3.3 Rasch-homogeneity

One essential assumption of the Rasch (1960) model is Rasch homogeneity. Rasch homogeneity implies that all item-discrimination parameters are equal. In order to test this assumption, a Birnbaum model (2PL; Birnbaum, 1968) was specified. In this model, discrimination parameters are freely estimated and not fixed to 1. The estimated discriminations differed across the items (see Table 4), ranging from 0.334 (item efl022b_c) to 2.232 (item efl059d_c). Despite the empirical preference for the 2PL (AIC = 50303.18, BIC = 50674.95, number of parameters = 65) model, the Rasch model (AIC = 50877.60, BIC = 51066.35, number of parameters = 33) more adequately matched the theoretical conceptions underlying the construction of the test (see Pohl & Carstensen, 2012, 2013 for a discussion of this issue). For this reason, the 1PL model was chosen as the scaling model.

5.3.4 Unidimensionality and local item independence

The unidimensionality and assumption of local item independency of the test was further investigated by comparing the unidimensional model with a testlet model (Wang, & Wilson, 2005; see Figure 6) in which the factor loadings were constrained to 1. The testlet model, which was based on the seven texts, was estimated with the Monte Carlo estimation algorithm implemented in ConQuest. Covariances between the testlet-specific factors and between testlet-specific factors and the general factor were fixed to zero in this model. Comparing the model fit indices of the unidimensional model (see section 5.3.3) with the testlet model (AIC: 50,339.45, BIC: 50,303.81, number of parameters = 39) suggests that the testlet model better fits the data. However, for theoretical reasons, we used the unidimensional Rasch model for estimating the WLEs. We encourage the reader to further investigate the potential use of such models over the course of running their analyses. The variance of the testlet factors ranged from 0.20 to 0.87. The variance of the common factor was 1.31.

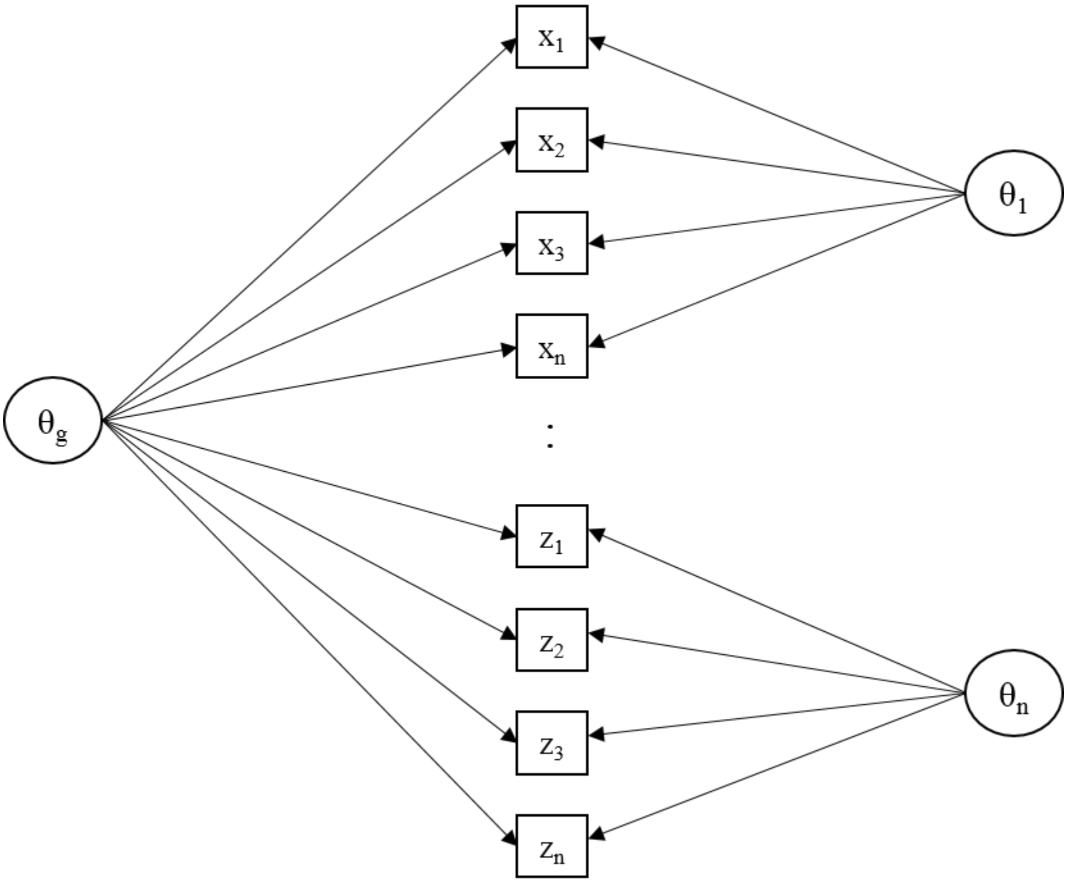

Figure 7. The testlet model that was specified and tested against the unidimensional model. The testlet model consists of one general latent variable θ_g and testlet-specific latent variables ($\theta_1 – \theta_n$) as well as testlet-specific indicators (X_1-X_n, Z_1-Z_n).

6 Discussion

Descriptions and analyses presented in the previous sections were aimed at documenting the quality of the English reading competence test used in the additional study Thuringia. The occurrence of different kinds of missing responses was evaluated, and item as well as test quality was examined. Furthermore, measurement invariance with regard to item difficulties was examined for various grouping variables. The item fit statistics provided evidence of items with good fit that were measurement invariant across these subgroups. The test was found to be reasonably reliable. As shown, ability estimates for participants with low performance were found to be precise but less precise for medium- and high-performing participants.

7 Data in the Scientific Use File

The data in the Scientific Use File contain 33 items, all of which are scored as dichotomous variables with 0 indicating an incorrect response and 1 indicating a correct response. MC items are marked with a '_c' at the end of the variable name. Appendix A provides the syntax that was used to generate the person estimates with the ConQuest 4.2 software (Wu, Adams, Wilson, & Haldane 2007). Appendix B provides an alternative syntax for use with the TAM package (Robitzsch, Kiefer, & Wu, 2017) in the software R (R Core Team, 2017).

Manifest English competence scores are provided in the form of WLEs (e_sc1) along with their corresponding standard errors (e_sc2). As described in Section 5, these person estimates were derived from the joint scaling of both waves of the study. For persons who did not take the English test, no WLE was estimated. WLEs were estimated for all items delivered in the Scientific Use File. Items with negative discriminations in the 2PL were excluded, therefore the delivered WLE is based on 32 items (item e065c_c was excluded). In order to allow the users to estimate their own WLEs by considering different item selection standards, all test items are delivered in the Scientific Use File. For researchers interested in analyses that require one of the variables that showed DIF > 0.6 logits, we emphasize that models should be considered on the basis of partial measurement invariance (e.g. Byrne, Shavelson, & Muthén, 1989).

We recommend the use of plausible values to investigate latent relationships between competence scores and other variables. Users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values themselves. A description of these approaches can be found in Pohl and Carstensen (2012)

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–722.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. & Novick, M. R. (Eds.). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structure: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Hübner, N., Rieger, S., & Wagner, W. (2016). *NEPS Technical Report for English Reading: Scaling Results for the Additional Study Baden-Württemberg* (NEPS Survey Paper No. 10). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

NEPS (2011a). *Curricular Reform Study in Thuringia Main Study 2009/10 (A70) Students, 12th grade Information on the Competence Test*. Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/TH/1-0-0/C_A70_EN.pdf

NEPS (2011b). *Curricular Reform Study in Thuringia Main Study 2010/11 (A71) Students, 12th grade Information on the Competence Test*. Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/TH/1-0-0/C_A71_EN.pdf

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., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading– Scaling results of Starting Cohort 3 in fifth grade* (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: Mesa Press.

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

Robitzsch, A., Kiefer, T., & Wu, M. (2017). *TAM: Test analysis modules. R package version 2.7-56*. Retrieved from 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 Verlag GmbH.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126-149. doi: 10.1177/0146621604271053

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalized item response modelling software*. Camberwell, AUS: ACER Press.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters.

This paper used data from the National Educational Panel Study (NEPS): Additional Study Thuringia doi:10.5157/NEPS:TH:2.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Appendix

Appendix A: ConQuest Syntax for generating WLE estimates in the Additional Study Thuringia

```
title Additional Study Thuringia, English, Waves 1-2;


datafile filename.dat;

format pid 1-7 responses 11-42;

labels << labels.nam;


codes 0,1;


model item;

set constraints=cases;


estimate ! stderr=empirical;

itanal ! form=long >> filename.itn;

export parameters >> filename.prm;

show cases ! estimates=wle >> filename.wle;

show ! estimates=latent, tables=1:2:3:4:5 >> filename.shw;
```

Appendix B: TAM Syntax for generating WLE estimates in the Additional Study Thuringia

```
setwd (“Your/Working/Directory”)
data <- # data read
items <- # column positions of the english items in the SUF
library (TAM)


# Compute 1 PL - Modell
ONEPL <- tam.mml(data[,items], irtmodel="1PL", pid=data$id)
summary (ONEPL)


# Compute 2 PL - Modell
TWOPL <- tam.mml.2pl(data[,items], irtmodel="2PL", est.variance = TRUE, pid=data$id)
summary (TWOPL)
```

Appendix C: Item Parameters based on all Items

Table S1

Item Parameters of the English Test (all Items)

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	t-value of WMNSQ	Correlation of item score with total score	Discrimination-2 PL
1	efl108a_c	88.52	-2.421	0.099	0.995	-0.061	0.32	0.859
2	efl108b_c	84.20	-2.007	0.088	1.104	1.914	0.25	0.523
3	efl108c_c	44.92	0.261	0.068	0.997	-0.087	0.48	0.881
4	efl108d_c	77.89	-1.527	0.078	0.980	-0.480	0.44	1.049
5	efl022b_c	46.52	0.169	0.067	1.194	7.016	0.29	0.366
6	efl022c_c	82.75	-1.893	0.085	1.021	0.425	0.35	0.698
7	efl022d_c	68.84	-0.979	0.071	0.996	-0.112	0.46	0.849
8	efl022e_c	71.01	-1.104	0.072	1.093	2.768	0.34	0.503
9	efl022f_c	55.59	-0.281	0.067	1.190	6.959	0.30	0.374
10	efl022g_c	48.93	0.051	0.067	1.056	2.142	0.43	0.672
11	efl022h_c	83.71	-1.972	0.087	0.977	-0.443	0.41	0.981
12	efl022i_c	62.16	-0.618	0.068	1.031	1.130	0.44	0.770

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	t-value of WMNSQ	Correlation of item score with total score	Discrimination-2 PL
13	efl008a_c	53.69	-0.193	0.048	0.965	-1.888	0.53	1.259
14	efl008b_c	67.54	-0.920	0.050	0.947	-2.441	0.52	1.385
15	efl008c_c	58.13	-0.416	0.048	0.981	-1.017	0.51	1.169
16	efl008e_c	47.83	0.102	0.048	0.937	-3.421	0.55	1.370
17	efl075a_c	51.87	-0.097	0.048	0.910	-4.944	0.58	2.005
18	efl075b_c	51.51	-0.081	0.048	0.969	-1.676	0.52	1.678
19	efl075c_c	61.55	-0.592	0.049	0.932	-3.484	0.56	1.917
20	efl075d_c	52.74	-0.142	0.048	0.891	-6.044	0.60	2.129
21	efl057a_c	81.63	-1.827	0.060	1.039	1.162	0.35	0.850
22	efl065a_c	71.82	-1.202	0.073	1.042	1.220	0.42	0.967
23	efl065b_c	80.96	-1.820	0.083	0.987	-0.262	0.42	1.127
24	efl065c_c	71.14	-1.163	0.073	1.442	11.498	0.01	-0.239
25	efl065d_c	49.73	-0.011	0.067	1.088	3.195	0.43	0.802
26	efl059a_c	65.23	-0.802	0.071	0.886	-3.887	0.59	1.943
27	efl059b_c	58.32	-0.440	0.070	0.917	-2.967	0.58	1.589
28	efl059c_c	64.74	-0.778	0.071	0.996	-0.123	0.50	1.233

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	t-value of WMNSQ	Correlation of item score with total score	Discrimination-2 PL
29	efl059d_c	73.86	-1.315	0.076	0.842	-4.525	0.61	2.442
30	efl059e_c	39.14	0.564	0.071	0.935	-2.199	0.55	1.455
31	efl059f_c	50.32	-0.017	0.069	1.024	0.895	0.49	1.054
32	efl059g_c	59.53	-0.494	0.070	0.882	-4.253	0.61	1.822
33	efl059i_c	68.08	-0.964	0.073	1.019	0.594	0.46	1.086