NEPS SURVEY PAPERS

Nicolas Hübner, Sven Rieger, and Wolfgang Wagner

NEPS TECHNICAL REPORT FOR ENGLISH READING: SCALING RESULTS FOR THE ADDITIONAL STUDY BADEN-WUERTTEMBERG

NEPS Survey Paper No. 10
Bamberg, November 2016

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

NEPS Technical Report for English Reading:

Scaling Results for the

Additional Study Baden-Wuerttemberg

Nicolas Hübner, Sven Rieger, & Wolfgang Wagner

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

E-mail address of lead author:

nicolas.huebner@uni-tuebingen.de

Bibliographic data:

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.
doi:10.5157/NEPS:SP10:1.0

We would like to thank Steffi Pohl and Claus. H. Carstensen for developing and providing standards for the technical reports. In addition, we would like to thank Eve Sarah Müller for her excellent support and valuable feedback. We would also like to thank Natalie Boonyaprasop for proofreading a primary version of this manuscript.

NEPS Technical Report for English Reading: Scaling Results for the Additional Study Baden-Wuerttemberg

Abstract

The National Educational Panel Study (NEPS) is aimed at investigating the development of competences across the entire life span. It also develops tests for assessing different competence domains. In order to evaluate the quality of these competence tests, a wide range of item response theory (IRT) analyses were carried out. This paper describes the data and results of analyses of the English reading competence test that was used in the additional study Baden-Wuerttemberg. 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 – the items were used in the English reading competence test in all three consecutive waves (2011–2013). The test was based on a subset of items from a test that was administered in the additional study Thuringia. In sum, 4,885 students took the test in these three waves. The English test consisted of 33 items, representing different levels of the Common European Framework of References, ranging from level B1 to C1. A Rasch model was used to scale 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. However, the reliability was somewhat modest, which might be due to the fact that the item difficulties were rather low compared with the 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, and appendices that describe the scaling of each wave separately.

Keywords

item response theory, scaling, English reading competence, Scientific Use File

Contents

1. Introduction

In 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. Weinert et al. (2011) provide an overview of the competences measured in NEPS.

Most of the competence data are scaled with 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 to scale the competence data and the analyses performed to check the quality of the scales are described in Pohl and Carstensen (2012).

This paper presents the results of the English reading competence test in three waves of the additional study Baden-Wuerttemberg. In this study, items developed by the Institute of Quality Development in Education (IQB) were composed for the English reading test used across three consecutive years (2011 through 2013) to test secondary-school students' English reading competences in their final year of Gymnasium (the type of school that leads to upper secondary education and the Abitur). More detailed information about the aims of this study can be found on the NEPS website.[1] Further information about the test can be found in NEPS (2011; 2012).

The present report draws strongly on previous technical reports such as Durchhardt (2015), 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 tests was led by the Institute for Educational Quality Improvement (IQB) and is described in Rupp, Vock, Harsch, and Köller (2008) and NEPS (2011; 2012). 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, 2011; 2012). 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 (MM) items. For MC

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

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. MM 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.

Table 1

Content Areas of the Items on the English Test

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

Table 2

Response Formats of the Items on the English Test

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

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 4,885 participants took the English reading test: 1,283 in 2011 (Wave 1), 2,391 in 2012 (Wave 2), and 1,211 in 2013 (Wave 3). All subjects gave at least one valid answer so that for every subject, one 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. 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, 1960/1980) was used and estimated in ConQuest 4.2.5 (Wu, Adams, & Wilson, 1997).

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/de-de/datenzentrum/datenunddokumentation/zusatzstudiebaden-w%C3%BCrttemberg/dokumentation.aspx

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.[3] We addressed the issue of measurement invariance by investigating test fairness for the variables gender, immigration background, books at home (as a proxy for socioeconomic status), and wave (i.e., to which of the three waves do subjects belong?); see Pohl and Carstensen (2012) for a description of these variables. 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, 1986) with the TAM package in R (Kiefer, Robitzsch, & Wu, 2015; R Core Team, 2015).

[3] 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 Baden-Wuerttemberg are presented. Some results in which each wave was scaled separately can be found in Appendices C1–C3.

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, almost none of the participants—only 2.9%—produced any invalid responses. The maximum number of invalid responses was 6.

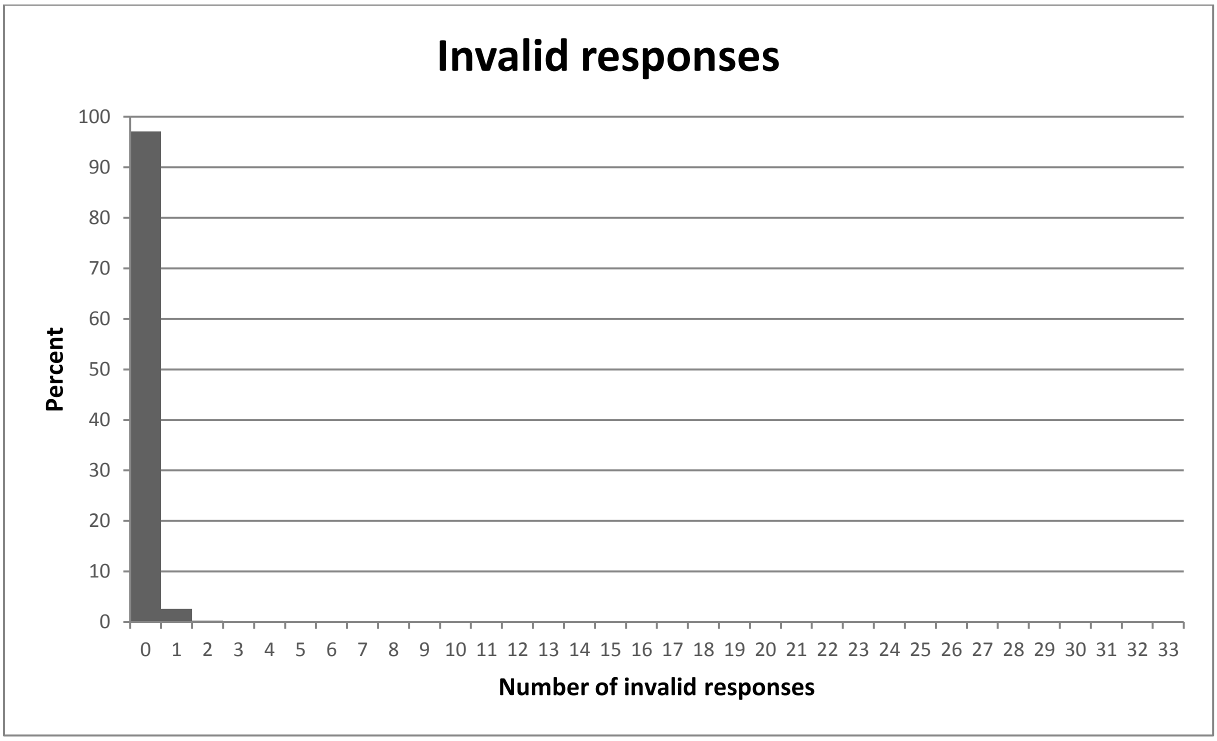

Figure 1. Number of invalid responses per person.

The largest source of missing responses on this test was the *omission of items*. As can be seen in Figure 2, only 11.7% of the participants skipped at least one item. Overall, 1.2% of the participants omitted five or more items.

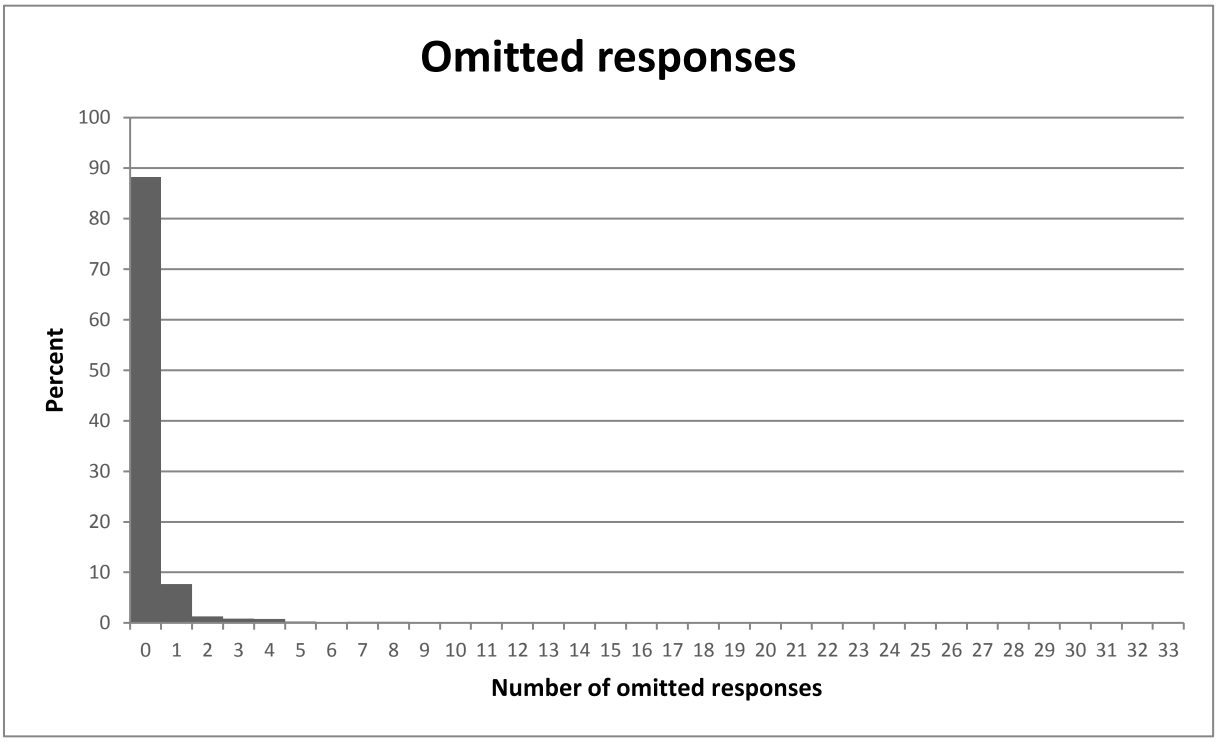

Figure 2. Number of omitted responses per person.

By definition, every item after the last item that was completed is labeled *not reached*. As Figure 3 shows, most participants (97.2%) reached the end of the test.

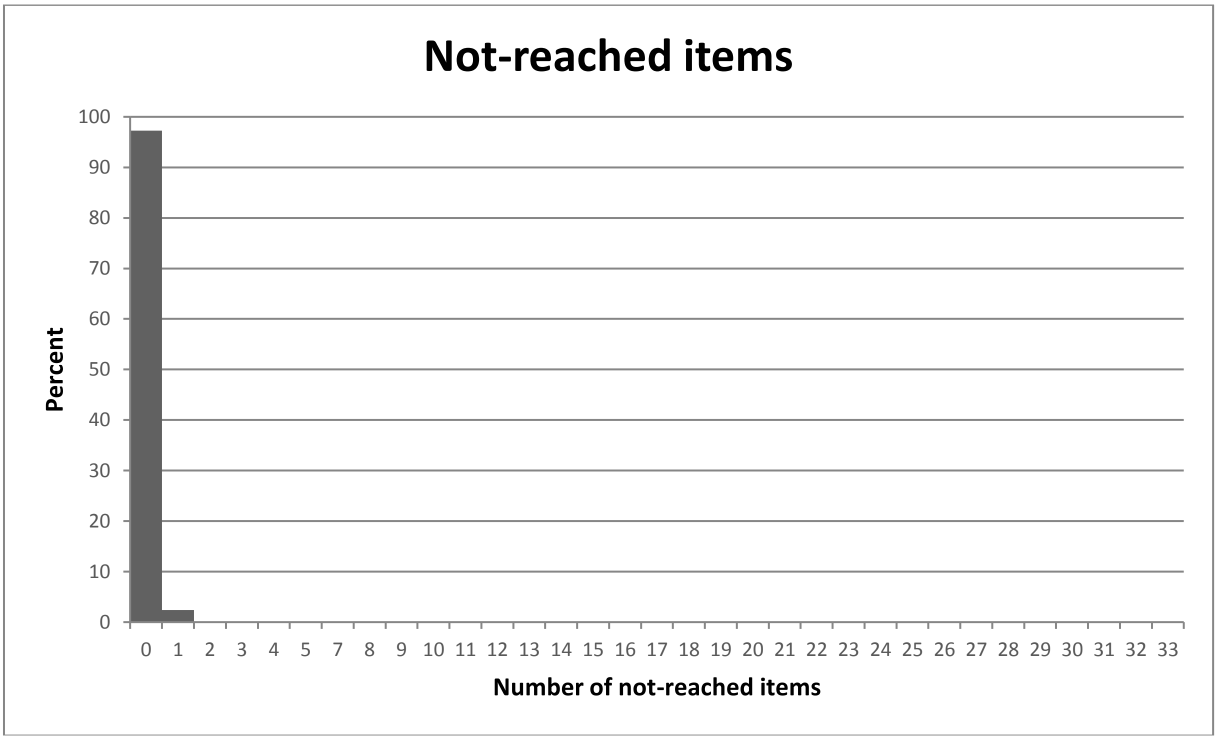

Figure 3. Number of not-reached items per person.

Overall, 99.8% of the participants had no items that were missing by *non-participation*. Only 0.2% (eight) 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 and missing by design) aggregated across the invalid, omitted, and not-reached missing responses per person is illustrated in Figure 4. On average, the participants produced 0.33 (SD = 1.20) missing responses. Moreover, 84.2% of the participants had no missing responses at all. Only 1.6% of the participants had five or more missing responses.

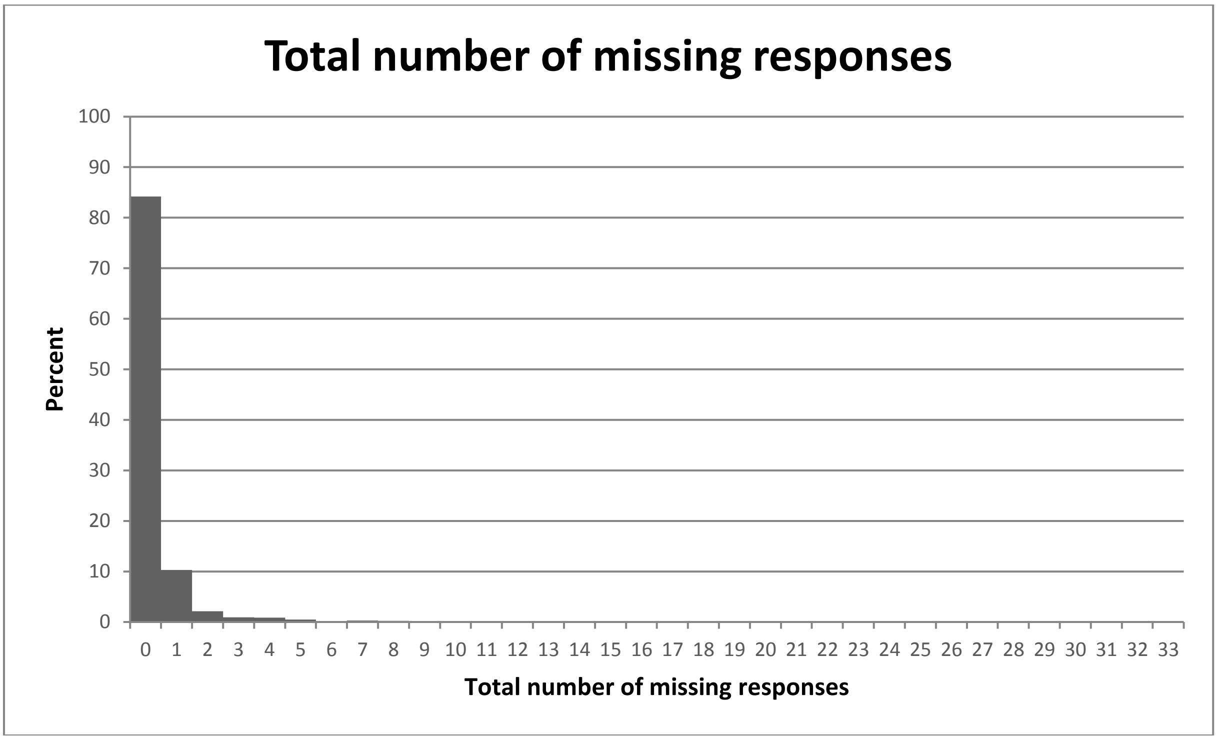

Figure 4. Total number of missing responses.

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 2.8% of the participants failed to reach items (column 5). No item had an omission rate that exceeded 5% (column 6). Item e022d_c (omitted by 2.3% of the participants) and item e008e_c (2.0%) were the most noticeable. Overall, the percentage of invalid responses per item (column 7) was very low (the maximum was 0.9% for item e057a_c). The percentage of items that were missing by non-participation (column 8) was very low (the maximum was 0.2%). All students who took the test had 12 items that were missing by design (column 9).

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 57.9% to 96.3% with an average of 80.2% (SD = 9.8%) 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 -3.775 (item e108a_c) to -0.380 (item e022b_c) logits with an average difficulty of -1.88 logits (SD = 0.83). Altogether, the item difficulties were somewhat low. Owing to the large sample size, the corresponding standard errors of the estimated item difficulties (column 4) were small (SE(ß) ≤ 0.117).

Table 3

Item Parameters of the English Test

	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 missing by non-participation	Percentage of missing by design
1	e108a_c	1	9	2450	-	0.3	0.1	0.1	49.5
2	e108b_c	1	10	2454	-	0.3	0.0	0.1	49.5
3	e108c_c	1	11	2389	-	1.5	0.1	0.1	49.5
4	e108d_c	1	12	2451	-	0.3	0.0	0.1	49.5
5	e022b_c	1	13	2432	0.0	0.6	0.0	0.2	49.5
6	e022c_c	1	14	2450	0.0	0.3	-	0.2	49.5
7	e022d_c	1	15	2350	0.1	2.3	0.0	0.2	49.5
8	e022e_c	1	16	2451	0.1	0.2	-	0.2	49.5
9	e022f_c	1	17	2434	0.1	0.6	-	0.2	49.5
10	e022g_c	1	18	2416	0.1	0.9	-	0.2	49.5
11	e022h_c	1	19	2437	0.2	0.4	-	0.2	49.5
12	e022i_c	1	20	2431	0.2	0.5	-	0.2	49.5
13	e008a_c	1,2	1 / 5	4847	-	0.7	0.1	0.1	-
14	e008b_c	1,2	2 / 6	4856	-	0.6	0.0	0.1	-

	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 missing by non-participation	Percentage of missing by design
15	e008c_c	1,2	3 / 7	4818	-	1.2	0.2	0.1	-
16	e008e_c	1,2	4 / 8	4781	-	2.0	0.2	0.1	-
17	e075a_c	1,2	8 / 20	4830	0.1	0.9	0.2	0.1	-
18	e075b_c	1,2	7 / 19	4820	0.1	1.1	0.1	0.1	-
19	e075c_c	1,2	6 / 18	4819	0.2	1.2	0.1	0.1	-
20	e075d_c	1,2	5 / 17	4815	0.2	1.3	0.0	0.1	-
21	e057a_c	1,2	21 / 21	4708	2.8	-	0.9	0.2	-
22	e065a_c	2	1	2411	-	0.2	0.0	0.2	50.4
23	e065b_c	2	2	2410	-	0.1	0.1	0.2	50.4
24	e065c_c	2	3	2401	-	0.2	0.1	0.2	50.4
25	e065d_c	2	4	2398	-	0.2	0.3	0.2	50.4
26	e059a_c	2	9	2365	-	1.1	0.1	0.1	50.4
27	e059b_c	2	10	2344	-	1.5	0.1	0.1	50.4
28	e059c_c	2	11	2393	-	0.6	0.0	0.1	50.4
29	e059d_c	2	12	2397	-	0.5	-	0.1	50.4
30	e059e_c	2	13	2348	-	1.4	0.1	0.1	50.4

	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 missing by non-participation	Percentage of missing by design
31	e059f_c	2	14	2383	-	0.7	0.2	0.1	50.4
32	e059g_c	2	15	2347	0.0	1.5	0.1	0.0	50.4
33	e059i_c	2	16	2382	0.0	0.7	0.1	0.1	50.4

Table 4

Item Parameters of the English Test

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
1	e108a_c	96.3	-3.775	0.112	1.01	0.1	0.21	0.82
2	e108b_c	88.0	-2.392	0.070	1.09	2.1	0.27	0.58
3	e108c_c	66.6	-0.834	0.052	1.01	0.7	0.49	0.85
4	e108d_c	89.8	-2.604	0.074	0.99	-0.1	0.37	0.96
5	e022b_c	57.9	-0.380	0.050	1.16	8.1	0.35	0.48
6	e022c_c	92.9	-3.048	0.085	1.01	0.3	0.29	0.77
7	e022d_c	87.7	-2.364	0.071	0.97	-0.6	0.41	0.97
8	e022e_c	82.6	-1.896	0.061	1.12	3.5	0.29	0.46
9	e022f_c	67.1	-0.875	0.052	1.20	8.7	0.30	0.38
10	e022g_c	66.2	-0.820	0.052	1.08	3.5	0.43	0.66
11	e022h_c	94.8	-3.397	0.097	0.96	-0.5	0.34	1.18
12	e022i_c	80.7	-1.740	0.060	1.13	4.0	0.30	0.51
13	e008a_c	79.6	-1.686	0.043	0.97	-1.5	0.49	1.13
14	e008b_c	86.0	-2.230	0.048	0.96	-1.4	0.46	1.21

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
15	e008c_c	76.5	-1.470	0.041	0.94	-3.1	0.54	1.33
16	e008e_c	71.6	-1.156	0.040	0.91	-5.5	0.59	1.44
17	e075a_c	75.1	-1.357	0.041	0.86	-8.0	0.62	4.33
18	e075b_c	73.5	-1.765	0.040	0.89	-6.6	0.60	3.97
19	e075c_c	80.6	-1.276	0.044	0.87	-5.9	0.59	3.62
20	e075d_c	74.8	-1.379	0.041	0.86	-7.8	0.62	3.96
21	e057a_c	90.2	-2.686	0.055	1.07	2.0	0.26	0.59
22	e065a_c	83.7	-2.058	0.064	1.06	1.7	0.37	0.80
23	e065b_c	90.2	-2.733	0.076	1.02	0.3	0.34	0.86
24	e065c_c	70.3	-1.118	0.054	1.47	17.3	0.05	-0.20
25	e065d_c	70.1	-1.109	0.054	1.14	5.5	0.37	0.61
26	e059a_c	85.8	-2.229	0.067	0.93	-1.9	0.51	1.43
27	e059b_c	87.2	-2.371	0.070	0.93	-1.8	0.50	1.37
28	e059c_c	84.9	-2.162	0.066	1.03	0.7	0.40	0.98
29	e059d_c	94.0	-3.321	0.093	0.92	-1.2	0.44	1.70
30	e059e_c	70.7	-1.125	0.055	0.96	-1.5	0.54	1.17

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
31	e059f_c	66.2	-0.867	0.053	1.13	5.8	0.38	0.56
32	e059g_c	83.9	-2.052	0.065	0.93	-1.9	0.51	1.30
33	e059i_c	80.9	-1.823	0.061	1.10	2.9	0.35	0.61

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 1.33, indicating a reasonable differentiation between the subjects. The reliability of the test (EAP/PV reliability = .74, WLE reliability = .60) was acceptable but not good. This should be related to the suboptimal test targeting described above.

Scale (in logits)	Person ability	Item difficulty
	XX	
3	X	
	X	
	XXX	
	XXX	
	XXXXXX	
	XXXXXXXX	
2	XXXXXX	
	XXXXXXXXX	
	XXXXXXXXXXX	
	XXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
1	XXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
0	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	5
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	9 10 31 3
-1	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXX	16 30 24 25
	XXXXXXXXXXXXXXXXXXXX	20 19 17
	XXXXXXXXXXXXXXXXXXX	15
	XXXXXXXXXXXXX	13
	XXXXXXXXXX	12 33 18
-2	XXXXXXXXXX	8
	XXXXXXX	32 22
	XXXX	14 26 28
	XXX	7 27 2
	XXX	4
	XXX	21 23
	X	
-3	X	6
	X	
	X	29
		11
		1
-4		

Figure 5. Test targeting. The distribution of person abilities in the sample is depicted on the left-hand side, with each 'X' representing 7.1 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 0.86 (items e075a_c and e075d_c) to 1.47 (item e065c_c), cf. column 5 of Table 4. This latter item also had the largest absolute WMNSQ t-value (17.3). It might be viewed as underfitting. Point-biserial correlations between the item scores and the total scores ranged from 0.05 (item e065c_c) to 0.62 (items e075a_c and e075d_c). Discriminations estimated in the 2PL-model with the TAM package in R ranged from -0.20 (item e065c_c) to 4.33 (item e075a_c), cf. Table 4, column 8. In conclusion, only item e065c_c showed considerably bad fit and was therefore excluded from further analyses.

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, immigration background, 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 = 0.112 logits, Cohen’s d = 0.091).[4] One item (e065d_c) showed a DIF greater than 0.6 logits.

Immigration background: On average, participants with and without an immigration background did not differ in their English reading competence (main effect = -0.010 logits, Cohen’s d = -0.008). No item showed a DIF greater than 0.6 logits.

Wave: On average, participants in the three waves differed in their English reading competence (1 vs 2: main effect = -0.143, Cohen’s d = -0.116; 1 vs 3: main effect = -0.224, Cohen’s d = -0.182; 2 vs 3: main effect = -0.081, Cohen’s d = -0.066). One item (e108a_c) showed a 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 = 0.325, Cohen’s d = 0.263; 0-200 vs > 500: main effect = 0.608, Cohen’s d = 0.493; 201-500 vs > 500: main effect = 0.283, Cohen’s d = 0.229). No item showed a DIF greater than 0.6 logits.

[4] The variance of the Rasch model was used to estimate the effect size.

Table 5

Differential Item Functioning

		Gender	**Immigration background**	**Wave**			**Books**		
	Item	male vs female	without vs with	1 vs 2	1 vs 3	2 vs 3	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
1	e108a_c	0.232	-0.110	-0.202	-0.644	-0.442	-0.378	-0.117	0.261
2	e108b_c	-0.114	0.048	0.198	0.168	-0.030	-0.023	-0.346	-0.323
3	e108c_c	0.018	0.392	0.054	0.024	-0.030	-0.111	-0.090	0.021
4	e108d_c	-0.062	0.308	-0.219	-0.215	0.004	0.026	-0.102	-0.128
5	e022b_c	-0.188	0.202	0.270	0.393	0.123	0.010	-0.100	-0.110
6	e022c_c	0.142	0.068	0.190	0.074	-0.116	-0.097	-0.224	-0.127
7	e022d_c	-0.172	0.240	-0.231	-0.234	-0.003	-0.186	0.162	0.348
8	e022e_c	0.066	0.102	-0.043	-0.008	0.035	0.105	-0.264	-0.369
9	e022f_c	0.084	0.054	-0.077	-0.022	0.055	-0.313	-0.302	0.011
10	e022g_c	0.056	-0.116	-0.157	-0.119	0.038	-0.080	-0.304	-0.224
11	e022h_c	-0.068	-0.292	-0.230	-0.307	-0.077	0.277	0.014	-0.263
12	e022i_c	0.136	-0.006	0.405	0.270	-0.135	-0.162	-0.014	0.148
13	e008a_c	0.124	-0.122	-0.121	-0.185	-0.064	0.099	0.096	-0.003
14	e008b_c	0.298	-0.038	0.102	0.057	-0.045	0.142	0.121	-0.021

		Gender	Immigration background	Wave			Books		
	Item	male vs female	without vs with	1 vs 2	1 vs 3	2 vs 3	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
15	e008c_c	-0.176	-0.116	-0.019	0.004	0.023	0.122	0.256	0.134
16	e008e_c	-0.374	-0.100	-0.084	0.048	0.132	0.219	0.390	0.171
17	e075a_c	-0.282	0.038	-0.110	-0.172	-0.062	-0.078	0.221	0.299
18	e075b_c	-0.014	0.014	-0.189	-0.234	-0.045	-0.105	0.086	0.191
19	e075c_c	-0.150	0.050	-0.059	-0.136	-0.077	-0.061	0.046	0.107
20	e075d_c	-0.090	-0.016	-0.063	-0.162	-0.099	-0.047	0.047	0.094
21	e057a_c	0.186	0.198	-0.246	0.058	0.304	0.292	0.185	-0.107
22	e065a_c	0.508	0.080	0.279	0.111	-0.168	-0.089	-0.047	0.042
23	e065b_c	0.580	0.030	0.549	0.576	0.027	0.013	-0.018	-0.031
25	e065d_c	0.632	-0.334	-0.157	-0.092	0.065	-0.081	-0.063	0.018
26	e059a_c	-0.198	0.034	-0.222	-0.267	-0.045	0.196	-0.013	-0.209
27	e059b_c	0.252	-0.096	-0.143	0.052	0.195	0.100	0.005	-0.095
28	e059c_c	-0.194	0.050	0.096	0.426	0.330	0.288	-0.015	-0.303
29	e059d_c	-0.068	-0.102	0.323	-0.032	-0.355	-0.133	-0.266	-0.133
30	e059e_c	0.142	-0.164	-0.223	-0.144	0.079	0.321	0.009	-0.312
31	e059f_c	0.088	-0.054	0.309	0.344	0.035	-0.122	-0.295	-0.173

		Gender	Immigration background	Wave			Books		
	Item	male vs female	without vs with	1 vs 2	1 vs 3	2 vs 3	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
32	e059g_c	0.066	-0.042	0.124	0.080	-0.044	0.174	0.072	-0.102
33	e059i_c	-0.076	-0.046	0.082	-0.232	-0.314	-0.325	-0.455	-0.130
	main effect	0.112	-0.010	-0.143	-0.224	-0.081	0.325	0.608	0.283

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 the variables immigration background and wave. 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 only for the variable gender.

Table 6

Comparison of Models With and Without DIF

DIF variable	Model	Number of parameters	AIC	BIC
Gender	main effect	35	83,890.82	83,948.24
	DIF	68	83,798.31	83,909.78
Immigration background	main effect	35	83,294.25	83,351.67
	DIF	68	83,319.74	83,431.20
Wave	main effect	36	84,198.51	84,257.62
	DIF	102	84,220.07	84,387.27
Books	main effect	36	83,673.29	83,732.40
	DIF	102	83,672.45	83,839.64

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, 1986) 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 0.38 (item e022f_c) to 4.33 (item e075a_c). Item e065c_c had a negative discrimination, 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. Despite the empirical preference for the 2PL (AIC = 81759.24, BIC = 82174.85, number of parameters = 64) model, the Rasch model (AIC = 84,212.74, BIC = 84,268.47, 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 the 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: 82,658.28, BIC: 82,724.14, number of parameters = 39) suggested that the testlet model fit the data better. 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 0.52 to 1.32. The variance of the common factor was 1.37.

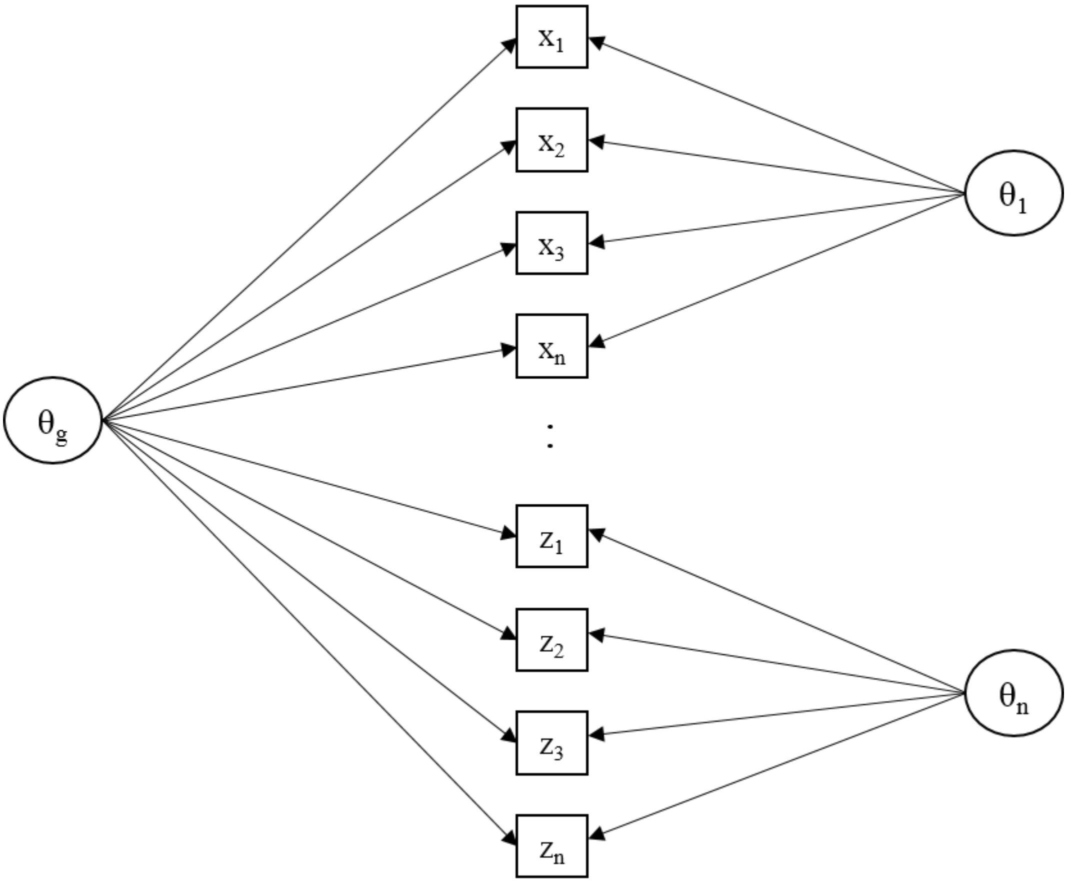

Figure 6. The testlet model that was specified and tested against the unidimensional model. The testlet model consisted of one general latent variable θ_g and testlet-specific latent variables (θ_1 – θ_n) as well as testlet-specific indicators (X_1-X_n, Z_1-Z_n).

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 Baden-Wuerttemberg. 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, 1997). Appendix B provides an alternative syntax for use with the TAM package (Kiefer, Robitzsch, & Wu, 2015) in the software R (R Core Team, 2015).

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 all three 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.

Duchhardt, C. (2015). *NEPS Technical Report for Mathematics: Scaling results for the additional study Baden-Wuerttemberg* (NEPS Working Paper No. 59). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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.

Jordan, A.-K., & Duchhardt, C. (2013). *NEPS Technical Report for Mathematics - Scaling results of Starting Cohort 6–Adults* (NEPS Working Paper No. 32). Bamberg: University of Bamberg, National Educational Panel Study.

Kiefer, T., Robitzsch, A., & Wu, M. (2015). *TAM: Test Analysis Modules (R package version 1.4-1)* [Computer software]. Retrieved from http://CRAN.R-project.org/package=TAM

Koller, I., Haberkorn, K., & Rohm, T. (2014). *NEPS Technical Report for Reading: Scaling results of Starting Cohort 6 for adults in main study 2012* (NEPS Working Paper No. 48). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

NEPS (2011). *G8-Reform in Baden-Württemberg, Haupterhebung 2010/11 (A72), Schüler/innen, Klasse 13. Informationen zum Kompetenztest.* Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/BW/2-0-0/C_A72_de.pdf.

NEPS (2012). *G8-Reform in Baden-Württemberg, Haupterhebung 2011/12 (A73), Schüler/innen, Klasse 12/13. Informationen zum Kompetenztest.* Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/BW/2-0-0/C_A73_de.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*. Copenhagen: Nielsen & Lydiche (Expanded edition, 1980, Chicago: University of Chicago Press).

R Core Team (2015). R: *A language and environment for statistical computing*. R Foundationfor Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

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.

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

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 Baden-Wuerttemberg, doi:10.5157/NEPS:BW:3.0.0. 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 Baden-Wuerttemberg

```
title Additional Study Baden-Wuerttemberg, English, Waves 1-3;

datafile filename.dat;
format pid 1-7 responses 12-13;
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 Baden-Wuerttemberg

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


# Compute Rasch
RASCH <- tam(data[,items], irtmodel="Rasch", pid=data$id)
summary (RASCH)


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

Appendix C1: Item Parameters and Differential Item Functioning for Wave 1 from the Additional Study Baden-Wuerttemberg only

Table 7

Item Parameters of the English Test – Wave 1

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
1	e108a_c	97.40	-4.227	0.255	0.99	0.0	0.19
2	e108b_c	87.58	-2.408	0.134	1.09	1.0	0.31
3	e108c_c	67.61	-0.947	0.103	0.96	-1.0	0.55
4	e108d_c	91.74	-2.925	0.156	0.98	-0.1	0.36
5	e022b_c	55.19	-0.294	0.098	1.19	5.1	0.32
6	e022c_c	92.79	-3.096	0.165	1.07	0.6	0.22
7	e022d_c	90.05	-2.694	0.149	1.00	0.0	0.38
8	e022e_c	83.90	-2.072	0.123	1.17	2.3	0.26
9	e022f_c	69.38	-1.063	0.103	1.20	4.2	0.30
10	e022g_c	69.58	-1.071	0.104	1.06	1.2	0.45
11	e022h_c	95.97	-3.754	0.212	0.98	-0.1	0.26
12	e022i_c	78.33	-1.630	0.113	1.12	2.0	0.34
13	e008a_c	82.67	-1.953	0.089	0.98	-0.3	0.48
14	e008b_c	86.77	-2.333	0.097	0.97	-0.6	0.46

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
15	e008c_c	78.60	-1.635	0.084	1.00	0.0	0.49
16	e008e_c	74.24	-1.336	0.080	0.92	-2.1	0.58
17	e075a_c	78.22	-1.609	0.083	0.88	-3.0	0.60
18	e075b_c	84.16	-2.081	0.092	0.90	-2.0	0.56
19	e075c_c	76.52	-1.492	0.082	0.90	-2.6	0.58
20	e075d_c	78.11	-1.605	0.083	0.89	-2.9	0.60
21	e057a_c	92.04	-2.972	0.119	1.10	1.3	0.22
22	e065a_c	84.00	-2.072	0.127	1.06	0.9	0.40
23	e065b_c	88.46	-2.513	0.143	1.04	0.5	0.36
25	e065d_c	74.64	-1.363	0.111	1.16	2.8	0.37
26	e059a_c	89.48	-2.596	0.151	0.91	-0.9	0.49
27	e059b_c	89.53	-2.614	0.151	0.90	-1.1	0.52
28	e059c_c	85.35	-2.188	0.132	1.01	0.2	0.44
29	e059d_c	94.35	-3.379	0.190	0.87	-0.9	0.48
30	e059e_c	76.09	-1.426	0.115	1.03	0.5	0.48
31	e059f_c	65.11	-0.773	0.104	1.17	3.8	0.37
32	e059g_c	85.08	-2.146	0.132	0.92	-1.0	0.53

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
33	e059i_c	83.63	-2.018	0.127	1.11	1.5	0.34

Table 8

Differential Item Functioning – Wave 1

		Gender	Immigration background	Books		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
1	e108a_c	-0.066	0.412	-0.217	0.386	0.603
2	e108b_c	-0.276	-0.128	-0.034	-0.572	-0.538
3	e108c_c	-0.068	0.544	0.281	-0.092	-0.373
4	e108d_c	-0.470	0.018	0.163	0.425	0.262
5	e022b_c	-0.132	0.154	0.206	-0.164	-0.370
6	e022c_c	0.632	0.212	-0.014	-0.010	0.004
7	e022d_c	-0.232	-0.028	-0.059	-0.011	0.048
8	e022e_c	-0.108	-0.272	-0.304	-0.518	-0.214
9	e022f_c	0.376	-0.264	0.020	-0.281	-0.301

		Gender	Immigration background	Books		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
10	e022g_c	0.166	-0.334	0.248	0.025	-0.223
11	e022h_c	-0.186	-0.466	0.405	-0.498	-0.903
12	e022i_c	0.334	0.038	0.091	0.021	-0.070
13	e008a_c	0.002	-0.280	0.131	0.127	-0.004
14	e008b_c	0.420	-0.190	-0.224	-0.382	-0.158
15	e008c_c	-0.026	-0.320	-0.024	0.076	0.100
16	e008e_c	-0.432	-0.086	0.084	0.082	-0.002
17	e075a_c	-0.252	0.002	-0.013	0.162	0.175
18	e075b_c	-0.030	0.060	-0.108	0.375	0.483
19	e075c_c	-0.226	-0.062	-0.024	0.132	0.156
20	e075d_c	0.036	-0.054	-0.135	0.129	0.264
21	e057a_c	0.042	0.212	-0.199	0.265	0.464
22	e065a_c	0.392	0.536	-0.252	-0.084	0.168
23	e065b_c	0.740	0.280	-0.150	-0.399	-0.249
25	e065d_c	0.556	-0.076	-0.494	-0.152	0.342
26	e059a_c	-0.526	0.466	0.284	0.215	-0.069

		Gender	Immigration background	Books		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
27	e059b_c	0.120	0.296	0.270	0.333	0.063
28	e059c_c	-0.258	0.248	0.139	0.126	-0.013
29	e059d_c	-0.026	0.642	0.002	-0.605	-0.607
30	e059e_c	0.104	-0.264	-0.023	-0.193	-0.170
31	e059f_c	-0.016	0.004	0.058	0.058	0.000
32	e059g_c	-0.128	0.586	0.111	0.015	-0.096
33	e059i_c	0.136	0.310	-0.462	-0.294	0.168
	main effect	-0.066	0.018	0.225	0.486	0.261

Appendix C2: Item Parameters and Differential Item Functioning for Wave 2 from the Additional Study Baden-Wuerttemberg only

Table 9

Item Parameters of the English Test – Wave 2

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
1	e108a_c	96.58	-3.892	0.167	1.02	0.2	0.19
2	e108b_c	88.50	-2.471	0.102	1.14	2.0	0.24
3	e108c_c	66.95	-0.861	0.076	1.06	2.0	0.45
4	e108d_c	89.40	-2.572	0.105	1.02	0.3	0.35
5	e022b_c	58.67	-0.422	0.073	1.16	5.5	0.35
6	e022c_c	93.41	-3.152	0.126	1.00	-0.0	0.31
7	e022d_c	87.25	-2.329	0.100	1.00	0.0	0.41
8	e022e_c	82.39	-1.893	0.088	1.14	2.9	0.29
9	e022f_c	66.53	-0.847	0.075	1.20	6.2	0.31
10	e022g_c	65.20	-0.776	0.075	1.10	3.1	0.43
11	e022h_c	94.64	-3.391	0.138	0.96	-0.4	0.34
12	e022i_c	82.59	-1.899	0.089	1.17	3.2	0.27
13	e008a_c	79.17	-1.696	0.062	1.00	0.0	0.49
14	e008b_c	86.24	-2.300	0.070	0.98	-0.5	0.45

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
15	e008c_c	76.07	-1.478	0.060	0.94	-2.1	0.55
16	e008e_c	70.38	-1.114	0.057	0.92	-3.5	0.59
17	e075a_c	74.33	-1.362	0.059	0.84	-6.2	0.65
18	e075b_c	79.92	-1.755	0.063	0.86	-4.6	0.61
19	e075c_c	73.29	-1.296	0.059	0.88	-4.6	0.61
20	e075d_c	74.96	-1.405	0.059	0.86	-5.3	0.63
21	e057a_c	89.05	-2.593	0.077	1.10	2.0	0.29
22	e065a_c	84.60	-2.214	0.094	1.11	2.0	0.35
23	e065b_c	91.02	-2.927	0.114	1.05	0.7	0.31
25	e065d_c	68.74	-1.067	0.078	1.19	5.3	0.36
26	e059a_c	85.09	-2.239	0.096	0.95	-0.9	0.49
27	e059b_c	86.10	-2.336	0.098	0.96	-0.7	0.49
28	e059c_c	84.00	-2.148	0.093	1.07	1.3	0.38
29	e059d_c	94.73	-3.569	0.141	0.92	-0.8	0.40
30	e059e_c	69.03	-1.064	0.079	0.96	-1.3	0.56
31	e059f_c	66.95	-0.942	0.077	1.17	4.9	0.39
32	e059g_c	84.01	-2.133	0.094	0.96	-0.8	0.50

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
33	e059i_c	81.88	-1.963	0.090	1.17	3.4	0.32

Table 10

Differential Item Functioning – Wave 2

		Gender	Immigration background	Books		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
1	e108a_c	0.440	-0.180	-0.579	-0.345	0.234
2	e108b_c	0.026	0.044	-0.132	-0.375	-0.243
3	e108c_c	-0.024	0.152	-0.170	-0.085	0.085
4	e108d_c	0.142	0.144	0.322	-0.130	-0.452
5	e022b_c	-0.172	-0.004	-0.084	-0.141	-0.057
6	e022c_c	-0.110	-0.032	-0.023	-0.295	-0.272
7	e022d_c	-0.188	0.160	-0.419	0.114	0.533
8	e022e_c	0.008	0.094	-0.003	-0.508	-0.505
9	e022f_c	-0.004	0.208	-0.462	-0.382	0.080

		Gender	Immigration background	Books		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
10	e022g_c	-0.024	-0.344	-0.348	-0.600	-0.252
11	e022h_c	-0.006	-0.356	-0.108	-0.072	0.036
12	e022i_c	-0.104	-0.090	-0.164	-0.061	0.103
13	e008a_c	0.228	-0.022	0.058	-0.010	-0.068
14	e008b_c	0.306	0.060	0.229	0.257	0.028
15	e008c_c	-0.230	-0.032	-0.007	0.229	0.236
16	e008e_c	-0.360	-0.120	0.125	0.448	0.323
17	e075a_c	-0.378	-0.016	-0.015	0.357	0.372
18	e075b_c	-0.092	-0.034	-0.069	0.066	0.135
19	e075c_c	-0.148	0.146	0.034	0.089	0.055
20	e075d_c	-0.154	0.006	0.086	0.007	-0.079
21	e057a_c	0.278	0.260	0.481	0.205	-0.276
22	e065a_c	0.664	0.182	0.012	0.006	-0.006
23	e065b_c	0.464	-0.120	0.086	0.064	-0.022
25	e065d_c	0.672	-0.318	0.029	-0.017	-0.046
26	e059a_c	-0.222	-0.006	0.080	-0.047	-0.127

		Gender	Immigration background	Books		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
27	e059b_c	0.302	-0.096	0.060	-0.018	-0.078
28	e059c_c	-0.152	0.290	0.333	-0.001	-0.334
29	e059d_c	-0.228	-0.170	-0.248	0.068	0.316
30	e059e_c	0.168	-0.042	0.495	0.255	-0.240
31	e059f_c	0.388	0.096	-0.162	-0.346	-0.184
32	e059g_c	0.058	-0.312	0.373	0.278	-0.095
33	e059i_c	0.086	0.108	-0.448	-0.577	-0.129
	main effect	0.178	-0.006	0.319	0.605	0.286

Appendix C3: Item Parameters and Differential Item Functioning for Wave 3 from the Additional Study Baden-Wuerttemberg only

Table 11

Item Parameters of the English Test – Wave 3

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
1	e108a_c	94.49	-3.379	0.193	1.00	0.1	0.26
2	e108b_c	87.38	-2.368	0.140	1.10	1.2	0.27
3	e108c_c	64.78	-0.753	0.107	1.02	0.4	0.50
4	e108d_c	88.65	-2.503	0.146	0.98	-0.2	0.40
5	e022b_c	59.27	-0.466	0.104	1.15	3.6	0.38
6	e022c_c	92.15	-2.964	0.167	1.01	0.1	0.31
7	e022d_c	86.25	-2.252	0.138	0.98	-0.2	0.44
8	e022e_c	81.70	-1.853	0.124	1.11	1.7	0.32
9	e022f_c	65.88	-0.824	0.107	1.26	5.4	0.26
10	e022g_c	64.47	-0.736	0.106	1.11	2.6	0.40
11	e022h_c	93.80	-3.242	0.184	0.94	-0.4	0.40
12	e022i_c	79.70	-1.688	0.121	1.14	2.2	0.33
13	e008a_c	77.00	-1.556	0.086	0.99	-0.1	0.50
14	e008b_c	84.79	-2.182	0.097	0.98	-0.3	0.47

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
15	e008c_c	75.13	-1.425	0.085	0.93	-1.8	0.56
16	e008e_c	71.17	-1.169	0.082	0.89	-3.2	0.61
17	e075a_c	71.96	-1.224	0.083	0.90	-2.8	0.60
18	e075b_c	78.14	-1.636	0.088	0.91	-2.2	0.59
19	e075c_c	70.73	-1.142	0.082	0.92	-2.5	0.59
20	e075d_c	72.15	-1.230	0.083	0.85	-4.2	0.64
21	e057a_c	90.58	-2.825	0.115	1.11	1.5	0.24
22	e065a_c	81.46	-1.974	0.126	1.10	1.5	0.40
23	e065b_c	90.41	-2.883	0.157	1.02	0.2	0.37
25	e065d_c	68.22	-1.055	0.110	1.18	3.5	0.39
26	e059a_c	83.31	-2.121	0.130	0.92	-1.1	0.55
27	e059b_c	86.99	-2.458	0.143	0.93	-0.8	0.50
28	e059c_c	86.26	-2.406	0.139	1.01	0.1	0.43
29	e059d_c	92.18	-3.144	0.170	0.89	-0.9	0.47
30	e059e_c	68.53	-1.066	0.111	0.97	-0.7	0.56
31	e059f_c	65.77	-0.899	0.109	1.18	3.6	0.40
32	e059g_c	82.24	-2.016	0.130	0.96	-0.6	0.52

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
33	e059i_c	76.30	-1.574	0.118	1.14	2.3	0.39

Table 12

Differential Item Functioning – Wave 3

		Gender	Immigration background	Books		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
1	e108a_c	0.094	-0.326	-0.228	-0.024	0.204
2	e108b_c	-0.178	0.262	0.210	-0.063	-0.273
3	e108c_c	0.204	0.694	-0.423	-0.123	0.300
4	e108d_c	-0.156	0.868	-0.632	-0.478	0.154
5	e022b_c	-0.252	0.656	-0.025	0.001	0.026
6	e022c_c	0.112	0.110	-0.264	-0.299	-0.035
7	e022d_c	-0.122	0.644	0.095	0.403	0.308
8	e022e_c	0.346	0.492	0.738	0.453	-0.285
9	e022f_c	-0.044	0.104	-0.399	-0.177	0.222

		Gender	Immigration background	Books		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
10	e022g_c	0.098	0.544	0.082	-0.025	-0.107
11	e022h_c	-0.132	-0.024	0.849	0.567	-0.282
12	e022i_c	0.412	0.096	-0.390	0.021	0.411
13	e008a_c	0.034	-0.154	0.153	0.273	0.120
14	e008b_c	0.176	-0.068	0.325	0.349	0.024
15	e008c_c	-0.214	-0.062	0.537	0.486	-0.051
16	e008e_c	-0.346	-0.054	0.551	0.585	0.034
17	e075a_c	-0.142	0.186	-0.270	0.029	0.299
18	e075b_c	0.130	0.078	-0.173	-0.097	0.076
19	e075c_c	-0.086	-0.002	-0.278	-0.108	0.170
20	e075d_c	-0.086	-0.008	-0.204	0.054	0.258
21	e057a_c	0.100	0.082	0.377	0.058	-0.319
22	e065a_c	0.358	-0.572	-0.124	-0.074	0.050
23	e065b_c	0.652	-0.076	0.054	0.429	0.375
25	e065d_c	0.636	-0.572	0.100	-0.088	-0.188
26	e059a_c	0.050	-0.228	0.390	-0.122	-0.512

		Gender	**Immigration background**	**Books**		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
27	e059b_c	0.266	-0.426	0.086	-0.227	-0.313
28	e059c_c	-0.190	-0.632	0.438	-0.162	-0.600
29	e059d_c	0.108	-0.820	-0.049	-0.347	-0.298
30	e059e_c	0.132	-0.272	0.307	-0.286	-0.593
31	e059f_c	-0.386	-0.382	-0.229	-0.546	-0.317
32	e059g_c	0.266	-0.250	-0.153	-0.234	-0.081
33	e059i_c	-0.540	-0.650	-0.017	-0.400	-0.383
	main effect	0.178	-0.068	0.475	0.779	0.304