NEPS SURVEY PAPERS

Nicolas Hübner, Sven Rieger, and Wolfgang Wagner

NEPS TECHNICAL REPORT FOR PHYSICS COMPETENCE: SCALING RESULTS FOR THE ADDITIONAL STUDY BADEN-WUERTTEMBERG

NEPS Survey Paper No. 11
Bamberg, November 2016

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg − Germany − contact@lifbi.de

NEPS Technical Report for Physics Competence:
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

Bibliographische Angaben:

Hübner, N., Rieger, S., & Wagner, W. (2016). NEPS Technical Report for Physics Competence: Scaling Results for the Additional Study Baden-Württemberg (NEPS Survey Paper No. 11). Bamberg: Leibniz-Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP11:1.0

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. Additionally, we like to thank the IEA for providing some unpublished items from the TIMSS II and TIMSS III (1995) study.

NEPS Technical Report for Physics Competence: 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 physics competence test that was used in the additional study Baden-Wuerttemberg. It is based on a subset of items from a test which was administered in the additional study Thuringia. In sum, 4,875 students took the test in these three waves. The physics competence test consisted of 41 items. 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. 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, physics 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.

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 physics competence test in three waves of the additional study Baden-Wuerttemberg. In this study, items were composed for the physics competence test used across three consecutive years (2011 through 2013) to test secondary-school students' physics 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 Hübner, Rieger, and Wagner (2016), Durchhardt (2015), Pohl, Haberkorn, Hardt, and Wiegand (2012) and Pohl and Carstensen (2012). It includes extracts from these previous reports.

2. Testing Physics Competence

The items for the physics competence consist of a subset of items from a test which was administered in the additional study Thuringia (Wagner et al., 2011). The framework and item development is therefore corresponded to the Thuringian curriculum for physics (Thüringer Kultusministerium, 1999). Furthermore, it takes the basic requirements for the Abitur in physics into account (Einheitliche Prüfungsanforderungen für die Abiturprüfung in Physik) (KMK, 2004). The items of the physics competence test are composed of a few different studies. Some of the items are unpublished. Table 1 depicts the sources where the items were obtained.

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

Table 1

Source of Items in the Physics Competence Test

Source	**Frequency**
TIMSS II	2
TIMSS III	17
Thermodynamik Testinventar[1]	4
BEMA[2]	2
Proprietary development[3]	16
Total number of items	41

References: [1]Einhaus, 2007; [2]Ding, Chabay, Sherwood, & Beichner, 2006; [3]Viering & Neumann, 2008; TIMSS II, 1995; TIMSS III, 1995

In the following, we will point out specific aspects of the physics competence paper-and-pencil test that are necessary for understanding the scaling results presented in this paper. The items are not arranged in units. Thus, on the test, students must usually read a certain situation and must subsequently answer only one task related to it.

There are three types of response formats in the physics competence test. These are simple multiple choice (MC), complex multiple choice (CMC), and short constructed response (SCR). For MC items, the test taker has to choose the correct answer out of several - usually four or five- response options. For CMC tasks, a number of subtasks with three response options are presented. SCR items require the test taker to fill in an answer into an empty field. Tables 2 and 3 show how the content areas and response formats are distributed across the items.

Table 2

Content Areas of the Items on the Physics Competence Test

Content area	**Frequency**
Electrical fields and interdependency	3
Magnetic fields and electromagnetic induction	6
Waves	4
Optics	7
Quantum physics: Quanta and matter	4
Dynamics: Vibrations	4
Dynamics: Mechanics of the Rigid Body	4
Thermodynamics	7

Special Theory of Relativity	2
Total number of items	41

Table 3

Response Formats of the Items on the Physics Competence Test

Response format	Frequency
Single multiple choice	32
Complex multiple choice	3
Short constructed response	6
Total number of items	41

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 4,875 participants took the physics competence test: 1,281 in 2011 (Wave 1), 2,388 in 2012 (Wave 2), and 1,206 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 physics competence test (e.g., due to student tardiness) but did take at least one of the other competence tests (mathematics, English, or biology). This missing code is referred to as e) missing by non-participation.

[2]https://www.neps-data.de/de-de/datenzentrum/datenunddokumentation/zusatzstudiebaden-w%C3%BCrttemberg/dokumentation.aspx

4.2 Scaling Model

In order to estimate the item and person parameters for physics 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 physics competence were estimated as weighted maximum likelihood estimates (WLEs; Warm, 1989). Person parameter estimation in NEPS is described by Pohl and Carstensen (2012), 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.

4.3 Checking the Quality of the Scale

To ensure that the test featured appropriate psychometric properties, the quality of the test was examined 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 score 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 tohave 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 a physics 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,

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

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 physics competence data were scaled with the Rasch model, which assumes Rasch homogeneity. Nonetheless, Rasch homogeneity is an assumption that may not hold for empirical data. We therefore checked for deviations from a uniform discrimination. We estimated item discrimination applying the Birnbaum model (2PL) (Birnbaum, 1986) using the TAM package in R (Kiefer, Robitzsch, & Wu, 2015; R Core Team, 2015).

5. Results

In this section, the key scaling results of the three waves of the additional study Baden-Wuerttemberg will be 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, 5.93% of the participants produced any invalid responses. The maximum number of invalid responses was 6.

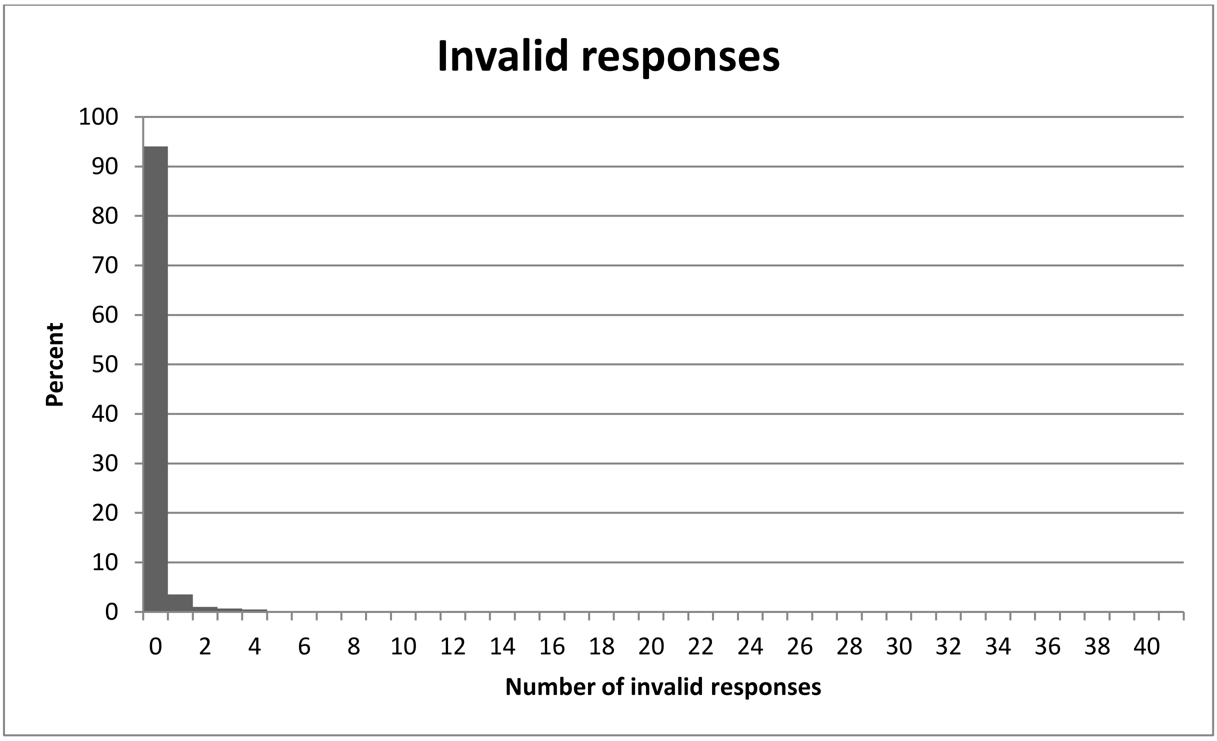

Figure 1. Number of invalid responses per person.

The largest source of missing responses on this test was the *omission of items*. As can be seen in Figure 2, almost half of the participants (49.70%) skipped at least one item. Overall, 10.28% of the participants omitted five or more items.

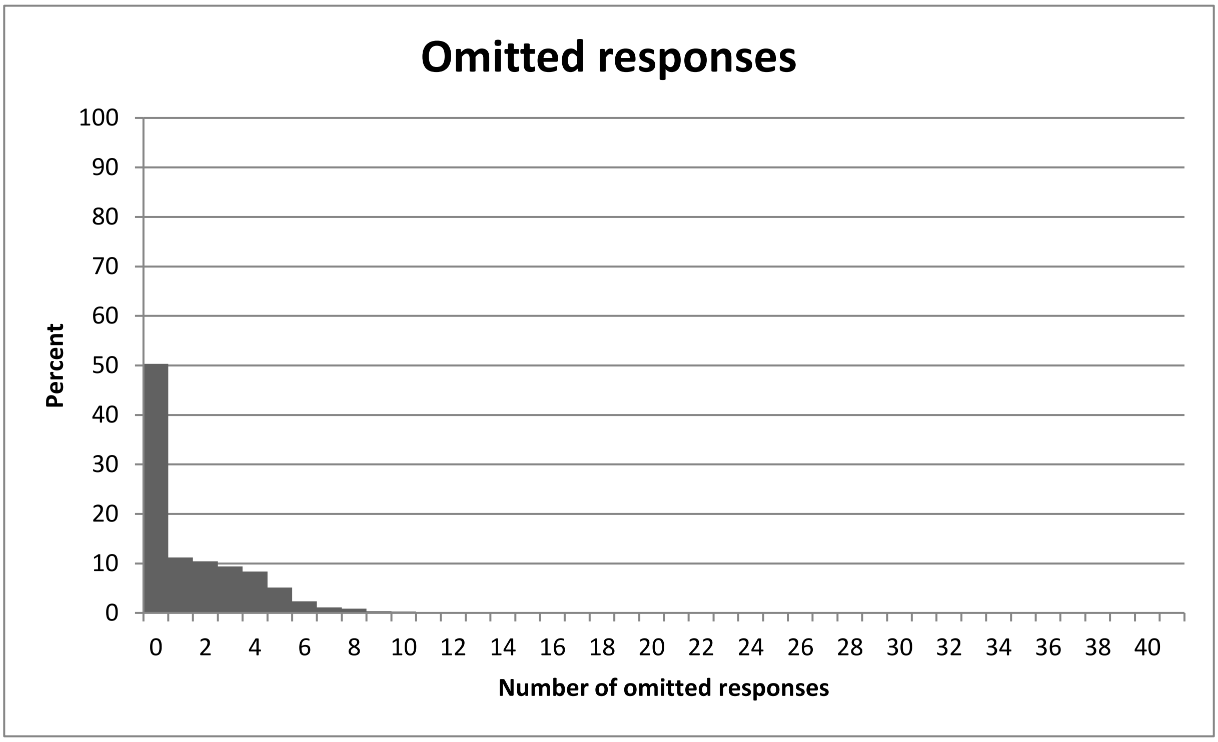

Figure 2. Number of omitted responses per person.

By definition, every item after the last item that was completed is labeled *not reached*. As Figure 3 shows, most participants (96.38%) reached the end of the test. Only 0.57% did not reach the fifth last item.

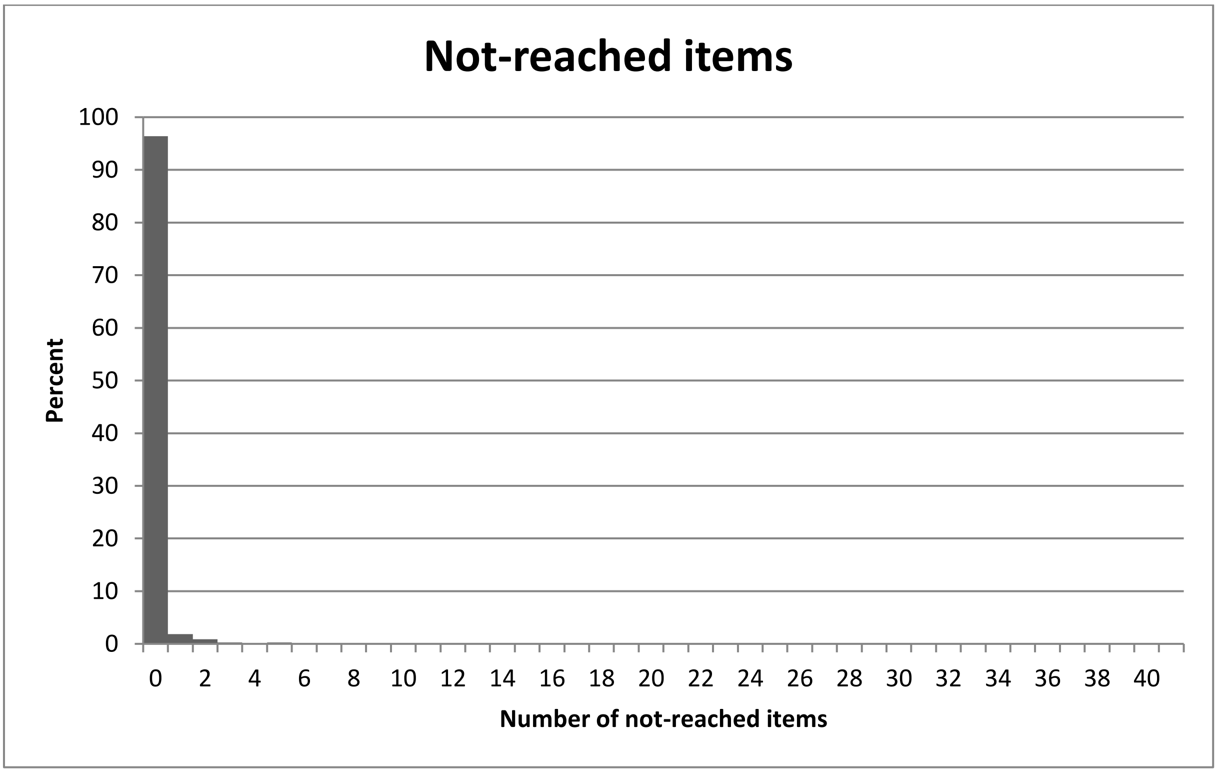

Figure 3. Number of not-reached items per person.

Overall, 99.63% of the participants had no items that were missing by *non-participation*. Only 0.37% (18) of the students did not take the physics competence 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 invalid, omitted, and not-reached missing responses per person is illustrated in Figure 4. On average, the participants produced 1.76 (SD = 2.32) missing responses. Moreover, 46.97% of the persons had no missing responses at all. Only 12.04% of the participants had five or more missing responses. Only ten students, who did not participate in the physics competence test, but in other achievement tests had to be excluded.

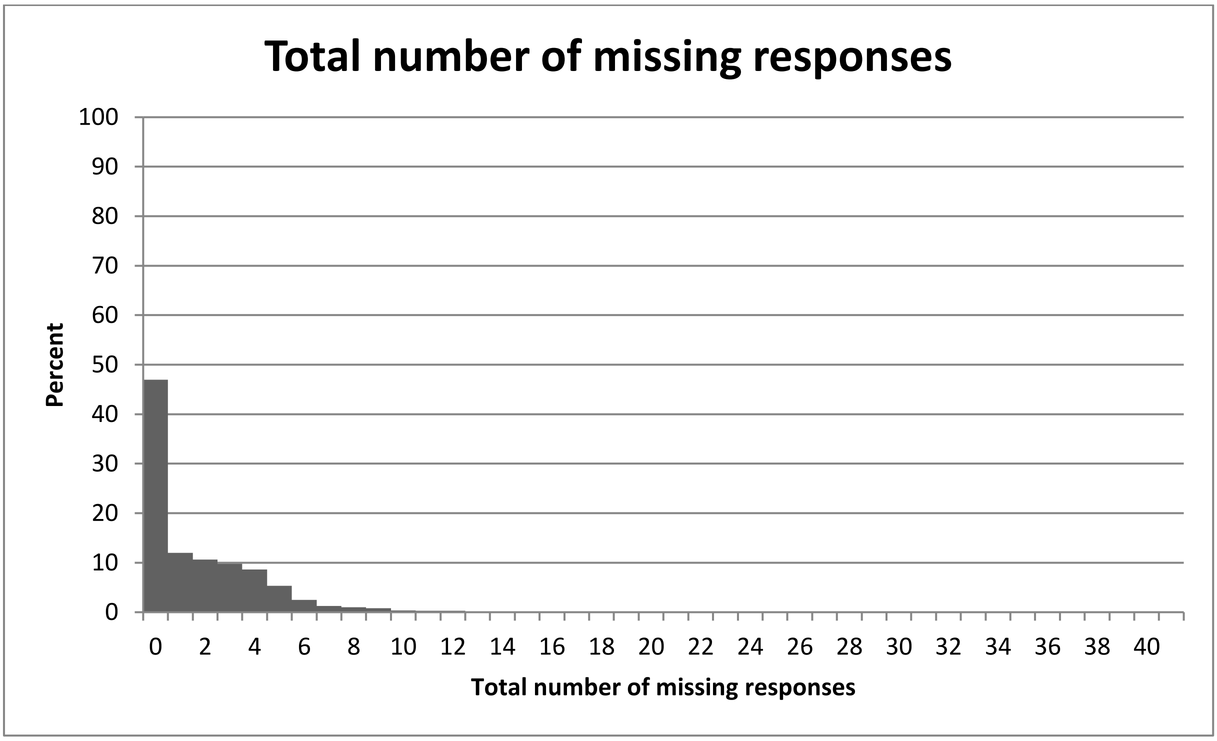

Figure 4. Total number of missing responses.

5.1.2 Missing responses per item

Table 4 provides information about the occurrence of the different kinds of responses that were missing per item. A maximum of 1.2% of the participants failed to reach items (column 5). 7 out of the 41 items had omission rates exceeding 5% (column 6). Item phyh6t_c (omitted by 18.0% of the participants), item phyn2t_c (28.5%), item phyn9t_c (25.9%), and item phyh5t_c (21.7%) were the most noticeable. Overall, the percentage of invalid responses per item (column 7) was very low (the maximum was 1.7% for item phyg2_c). The percentage of items that were missing by non-participation (column 8) was very low (the maximum was 0.4%). 0,4% of the persons who took the test had missing by design on 1 item, 24.5 % had missing by design on 20 items and 75,2% on 22 items (column 9).

5.2 Parameter Estimates

5.2.1 Item parameters

The second column in Table 5 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 3.0% to 88.0% with an average of 38.96% (SD = 21.58%) 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 5. The item difficulties ranged from -2.191 (item phye1_c) to 3.891 (item phyn2t_c) logits with an average difficulty of 0.61 logits (SD = 1.24). Altogether, the item difficulties were somewhat high. Owing to the large sample size, the corresponding standard errors of the estimated item difficulties (column 4) were small (SE(ß) ≤ 0.19).

Table 4

Item Parameters of the Physics Competence 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	phyh10_c	1,2,3,4	1	4703	-	3.4	0.1	0.4	-
2	phyg2_c	1,2,3,4	2	4785	-	0.2	1.7	0.4	-
3	phyg6_c	1,2,3,4	3	4723	-	2.5	0.6	0.4	-
4	phyg19_c	1,2,3,4	4	4790	-	1.0	0.7	0.4	-
5	phye1_c	1,2,3,4	5	4859	-	0.0	0.3	0.4	-
6	phyn14_c	1,2,3,4	6	4732	-	2.9	0.1	0.4	-
7	phyr1_c	1,2,3,4	7	4867	-	0.2	-	0.4	-
8	phyt1_c	1,2,3,4	8	4824	-	1.0	0.1	0.4	-
9	phyh12_c	1,2,3,4	9	4835	0.0	0.6	0.2	0.4	-
10	phyh6t_c	1,2,3	10	1554	0.1	18.0	1.0	0.2	48.9
11	phyn2t_c	1,2,3	11	1022	0.1	28.5	1.3	0.2	48.9
12	phyn9t_c	1,2,3	12	1148	0.1	25.9	1.4	0.2	48.9
13	phyn12t_c	1,2,3	13	2189	0.1	5.5	0.5	0.2	48.9
14	phyh5t_c	1,2,3	14	1368	0.2	21.7	1.1	0.2	48.9

	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	phyh2_c	1	15	1167	0.0	0.4	0.2	0.4	75.1
16	phyn11_c	1	16	1180	0.0	0.1	0.2	0.4	75.1
17	phyf5_c	1	17	1194	0.1	0.1	-	0.4	75.1
18	phyn6_c	1	18	1040	0.4	2.9	-	0.4	75.1
19	phyn7_c	1	19	1163	0.6	-	0.2	0.4	75.1
20	phyf7_c	2	15	1177	0.3	1.7	0.1	0.4	73.5
21	phyn5_c	2	16	1204	0.3	1.3	-	0.4	73.5
22	phyf13_c	2	17	1225	0.4	0.7	0.0	0.4	73.5
23	phyf9_c	2	18	1090	0.8	3.1	0.0	0.4	73.5
24	phyn3_c	2	19	1236	0.9	-	0.0	0.4	73.5
25	phyt4a_c	3	18	1096	0.2	1.8	0.0	0.4	75.2
26	phyt4b_c	3	19	1085	0.3	2.0	0.0	0.4	75.2
27	phyt4c_c	3	20	1145	0.3	0.8	-	0.4	75.2
28	phyn8_c	3	15	1146	0.1	1.0	-	0.4	75.2
29	phyb6_c	3	16	1139	0.1	1.0	0.1	0.4	75.2
30	phyh3_c	3	17	1084	0.2	2.1	0.0	0.4	75.2

	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 missing by non-participation	Percentage of missing by design
31	phyh8_c	3	21	1127	1.2	-	0.2	0.4	75.2
32	phyh6_c	4	10	2205	0.1	3.7	0.1	0.4	50.7
33	phyn2_c	4	11	2101	0.1	5.9	0.0	0.4	50.7
34	phyn9_c	4	12	2092	0.1	6.1	0.0	0.4	50.7
35	phyn12_c	4	13	2304	0.1	1.8	-	0.4	50.7
36	phyh5_c	4	14	2144	0.1	5.0	0.0	0.4	50.7
37	phyf4_c	4	15	1089	0.1	2.1	-	0.4	75.1
38	phyb24_c	4	16	1135	0.2	1.1	0.0	0.4	75.1
39	phym14_c	4	17	1170	-	0.5	0.1	-	75.5
40	phyg5_c	4	18	1163	0.3	0.2	0.2	0.4	75.1
41	phyg8_c	4	19	1152	0.9	-	-	0.4	75.1

Table 5

Item Parameters of the Physics Competence 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	phyh10_c	17.9	1.702	0.042	0.98	-0.9	0.36	0.82
2	phyg2_c	59.7	-0.449	0.033	0.97	-2.9	0.46	0.98
3	phyg6_c	57.0	-0.320	0.033	1.01	0.8	0.40	0.73
4	phyg19_c	44.9	0.226	0.033	0.96	-4.2	0.47	1.13
5	phye1_c	88.0	-2.191	0.047	1.03	0.8	0.24	0.56
6	phyn14_c	29.0	1.001	0.036	0.97	-1.7	0.42	0.91
7	phyr1_c	85.8	-1.994	0.044	0.99	-0.4	0.32	0.99
8	phyt1_c	35.0	0.690	0.034	1.01	0.6	0.39	0.67
9	phyh12_c	28.1	1.050	0.036	0.92	-5.0	0.50	1.33
10	phyh6t_c	36.7	0.764	0.058	1.04	2.1	0.35	0.53
11	phyn2t_c	3.0	3.891	0.187	0.96	-0.2	0.32	1.84
12	phyn9t_c	17.6	1.942	0.084	0.95	-1.0	0.43	1.16
13	phyn12t_c	16.2	1.865	0.063	0.93	-1.8	0.44	1.28
14	phyh5t_c	15.6	2.014	0.080	0.89	-2.4	0.53	1.67
15	phyh2_c	45.7	0.199	0.065	1.05	2.5	0.33	0.41

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
16	phyn11_c	42.8	0.333	0.065	0.95	-2.4	0.50	1.11
17	phyf5_c	45.8	0.192	0.064	1.03	1.4	0.39	0.62
18	phyn6_c	50.1	0.004	0.069	1.07	3.4	0.31	0.41
19	phyn7_c	52.7	-0.121	0.065	0.99	-0.7	0.45	0.94
20	phyf7_c	40.1	0.444	0.066	1.11	4.6	0.26	0.27
21	phyn5_c	46.7	0.134	0.064	0.96	-1.8	0.48	1.12
22	phyf13_c	53.2	-0.170	0.064	0.99	-0.8	0.44	0.83
23	phyf9_c	17.7	1.714	0.086	1.06	1.2	0.25	0.40
24	phyn3_c	57.4	-0.350	0.064	0.97	-1.5	0.47	1.13
25	phyt4a_c	76.6	-1.295	0.077	1.02	0.6	0.29	0.47
26	phyt4b_c	62.7	-0.557	0.068	1.03	1.4	0.35	0.53
27	phyt4c_c	19.8	1.570	0.080	1.12	2.6	0.12	-0.01
28	phyn8_c	9.5	2.487	0.105	1.03	0.4	0.19	0.38
29	phyb6_c	17.2	1.763	0.084	0.95	-1.1	0.43	1.13
30	phyh3_c	39.6	0.499	0.068	0.97	-1.5	0.46	1.01
31	phyh8_c	22.8	1.378	0.077	0.94	-1.5	0.45	1.15
32	phyh6_c	39.5	0.485	0.048	1.09	5.4	0.24	0.21

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
33	phyn2_c	20.5	1.493	0.058	1.07	2.2	0.20	0.24
34	phyn9_c	59.8	-0.440	0.049	1.11	6.9	0.21	0.16
35	phyn12_c	27.5	1.085	0.051	0.96	-1.7	0.43	0.89
36	phyh5_c	38.0	0.543	0.049	1.04	2.3	0.33	0.44
37	phyf4_c	22.4	1.363	0.078	0.97	-0.8	0.39	0.79
38	phyb24_c	15.1	1.901	0.088	0.99	-0.1	0.33	0.74
39	phym14_c	86.1	-2.003	0.089	1.04	0.7	0.21	0.36
40	phyg5_c	31.0	0.873	0.069	1.03	1.1	0.33	0.48
41	phyg8_c	22.4	1.370	0.076	0.96	-1.0	0.42	0.98

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 rather the medium and higher part of the ability distribution well but, in general, items were somewhat difficult. Hence, the test can measure person abilities in the medium and high-ability regions relatively precisely, whereas low 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 0.585, indicating a reasonable differentiation between the subjects. The reliability of the test (EAP/PV reliability = .63, WLE reliability = .61) was acceptable but not good. This should be related to the suboptimal test targeting described above.

Scale (in logits)	**Person ability**	**Item difficulty**
3		11
	X	28
	X	
	X	
	X	
2		12 14
	XXXX	13 38
	XXXX	29
	XXXX	1 23
	XXXXX	27
	XXXXXXX	33
	XXXXXX	31 37 41
	XXXXXXXXXXX	
	XXXXXXXXX	
	XXXXXXXXXXXXXX	9 35
1	XXXXXXXXXXXXXX	6
	XXXXXXXXXXXXXXXX	40
	XXXXXXXXXXXXXXXX	10
	XXXXXXXXXXXXXXXXXXXXX	8
	XXXXXXXXXXXXXXXXXXXXXXXXX	36
	XXXXXXXXXXXXXXXXXXXXXXXXX	20 32 30
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX	16
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	4
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	15 17 21
0	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	18
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	19 22
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	3 24
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	2 34
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	26
	XXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXX	
-1	XXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXX	
	XXXXXXXXXXXXX	
	XXXXXXXXXX	25
	XXXXXX	
	XXXX	
	XXXX	
	XX	
	X	
-2		39 7
	X	
		5

Figure 5. Test targeting. The distribution of person abilities in the sample is depicted on the left-hand side, with each 'X' representing 7.3 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 4.

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 0.89 (item phyh5t_c) to 1.12 (item phyt4c_c), cf. column 5 of Table 5. Point-biserial correlations between the item scores and the total scores ranged from 0.12 (item phyt4c_c) to 0.53 (item phyh5t_c). Discriminations estimated in the 2PL-model with the TAM package in R ranged from -0.01 (item phyt4c_c) to 1.84 (item phyn2t_c), cf. Table 5, column 8. In conclusion only item phyt4c_c showed considerably bad fit and was therefore excluded from further analyses.

5.3.2 Differential item funtioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i. e., measurement invariance). For this purpose, DIF was examined for the variables gender, immigration background, books, and wave (see Pohl & Carstensen, 2012, for a description of these variables). Table 6 provides a summary of the results of the DIF analyses. According to Pohl & 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, male participants had a considerably higher physics competence (main effect = -0.684 logits, Cohen’s *d* = -0.882).[4] Eight items (phyr1_c, phyn2t_c, phyn12t_c, phyh2_c, phyn6_c, phyn8_c, phyh6_c, phyn2_c) showed a DIF greater than 0.6 logits.

Immigration background: On average, participants without immigration background had a higher physics competence (main effect = 0.196 logits, Cohen’s *d* = 0.253). One item (phyn8_c) showed a DIF greater than 0.6 logits.

Wave: On average, participants in the three waves basically did not differ in their physics competence (1 vs 2: main effect = -0.009, Cohen’s d = -0.012; 1 vs 3: main effect = 0.009, Cohen’s d = 0.012; 2 vs 3: main effect = 0.018, Cohen’s *d* = 0.023). No item showed a DIF greater than 0.6 logits.

Books: On average, participants with many books at home performed better on the physics competence test (0-200 vs 201-500: main effect = 0.091, Cohen’s *d* = 0.117; 0-200 vs > 500: main effect = 0.231, Cohen’s *d* = 0.298; 201-500 vs > 500: main effect = 0.140, Cohen’s *d* = 0.180). 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 6

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	phyh10_c	-0.154	-0.168	0.047	-0.044	-0.091	-0.089	0.053	0.142
2	phyg2_c	-0.144	0.160	0.098	0.043	-0.055	0.160	0.380	0.220
3	phyg6_c	-0.136	0.010	0.116	0.027	-0.089	-0.107	-0.100	0.007
4	phyg19_c	-0.256	0.024	-0.005	-0.007	-0.002	0.016	0.014	-0.002
5	phye1_c	0.368	0.320	0.188	0.029	-0.159	0.013	0.066	0.053
6	phyn14_c	-0.308	0.120	-0.066	-0.066	0.000	0.022	0.092	0.070
7	phyr1_c	-0.640	0.314	0.115	-0.052	-0.167	0.155	0.146	-0.009
8	phyt1_c	0.062	-0.078	0.022	0.050	0.028	-0.154	-0.164	-0.010
9	phyh12_c	-0.292	0.194	0.101	0.001	-0.100	-0.029	0.072	0.101
10	phyh6t_c	0.446	0.180	-0.343	-0.311	0.032	0.005	-0.152	-0.157
11	phyn2t_c	-0.780	-0.230	-0.564	-0.513	0.051	-0.414	-0.105	0.309
12	phyn9t_c	0.140	-0.432	0.291	0.288	-0.003	-0.096	-0.150	-0.054
13	phyn12t_c	-0.820	0.394	-0.161	0.025	0.186	0.105	0.150	0.045
14	phyh5t_c	-0.420	-0.128	-0.420	-0.438	-0.018	0.018	0.089	0.071

		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	phyh2_c	0.772	-0.310	0.103	0.275	0.172	0.265	0.020	-0.245
16	phyn11_c	-0.372	-0.096	-0.358	-0.218	0.140	-0.013	0.076	0.089
17	phyf5_c	-0.276	-0.064	-0.232	0.035	0.267	-0.147	-0.002	0.145
18	phyn6_c	0.654	-0.082	0.101	0.101	0.000	0.101	-0.194	-0.295
19	phyn7_c	-0.340	-0.376	0.096	0.165	0.069	-0.078	-0.002	0.076
20	phyf7_c	0.562	-0.346	-0.090	0.165	0.255	0.284	0.126	-0.158
21	phyn5_c	-0.124	0.122	-0.049	0.026	0.075	-0.415	-0.337	0.078
22	phyf13_c	0.072	0.040	-0.120	-0.168	-0.048	-0.138	-0.063	0.075
23	phyf9_c	0.454	-0.166	0.274	0.119	-0.155	-0.144	-0.144	0.000
24	phyn3_c	-0.326	0.132	0.202	0.308	0.106	0.248	0.247	-0.001
25	phyt4a_c	-0.018	0.202	-0.363	-0.095	0.268	-0.075	-0.237	-0.162
26	phyt4b_c	0.006	-0.052	-0.359	0.026	0.385	-0.115	-0.500	-0.385
28	phyn8_c	0.642	-0.720	0.076	-0.205	-0.281	-0.080	0.206	0.286
29	phyb6_c	0.012	-0.106	-0.062	0.191	0.253	0.069	0.203	0.134
30	phyh3_c	-0.226	-0.084	-0.200	-0.441	-0.241	0.063	0.215	0.152
31	phyh8_c	0.178	0.108	0.149	0.040	-0.109	-0.123	-0.271	-0.148

		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	phyh6_c	0.696	0.072	-0.155	-0.014	0.141	0.069	-0.235	-0.304
33	phyn2_c	0.688	-0.140	-0.244	-0.402	-0.158	-0.166	-0.383	-0.217
34	phyn9_c	0.540	-0.060	0.050	0.389	0.339	-0.164	-0.181	-0.017
35	phyn12_c	-0.042	0.074	-0.026	-0.184	-0.158	0.064	0.032	-0.032
36	phyh5_c	0.488	-0.320	-0.103	-0.014	0.089	-0.140	-0.193	-0.053
37	phyf4_c	-0.012	-0.386	-0.073	0.005	0.078	-0.009	-0.174	-0.165
38	phyb24_c	0.368	-0.112	-0.039	-0.078	-0.039	-0.208	-0.137	0.071
39	phym14_c	0.376	-0.460	-0.016	0.301	0.317	0.062	-0.030	-0.092
40	phyg5_c	0.388	-0.302	0.323	-0.083	-0.406	0.346	0.092	-0.254
41	phyg8_c	-0.010	0.014	0.076	-0.253	-0.329	-0.089	-0.004	0.085
	main effect	-0.684	0.196	-0.009	0.009	0.018	0.091	0.231	0.140

In Table 7, 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 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 only for the variable gender.

Table 7

Comparison of Models With and Without DIF

DIF variable	Model	Number of parameters	AIC	BIC
Gender	main effect	42	93,061.95	93,132.85
	DIF	82	92,659.89	92,798.30
Immigration background	main effect	42	93,040.90	93,111.80
	DIF	82	93,021.29	93,159.70
Wave	main effect	43	94,064.59	94,137.17
	DIF	123	94,111.82	94,319.44
Books	main effect	43	93,655.52	93,728.11
	DIF	123	93,697.48	93,905.10

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 5), ranging from 0.16 (item phyn9_c) to 1.84 (item phyn2t_c). Item phyt4c_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 = 93331.21, BIC = 93850.56, number of parameters = 80) model, the Rasch model (AIC = 94060.89, BIC = 94327.06, number of parameters = 41) more adequately matches 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.

6. Discussion

Descriptions and analyses presented in the previous sections were aimed at documenting the quality of the physics 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 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 medium to good performance were found to be precise but less precise for low-performing participants.

7. Data in the Scientific Use File

The data in the Scientific Use File contain 41 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, 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 physics competence scores are provided in the form of WLEs (p_sc1) along with their corresponding standard errors (p_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 physics competence 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 40 items (phyt4c_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 Sstructure: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456-466.

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical review special Topics-Physics education research*, *2,* 010105.

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.

Einhaus, E. A. (2007). *Schülerkompetenzen im Bereich Wärmelehre: Entwicklung eines Testinstruments zur Überprüfung und Weiterentwicklung eines normativen Modells fachbezogener Kompetenzen*. Logos-Verlag.

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.

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

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK]. (2004). *Einheitliche Prüfungsanforderungen in der Abiturprüfung Physik (Beschluss der Kultusministerkonferenz 01.12.1989 i.d.F. vom 05.02.2004).* Retrieved from http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/1989/1989_12_01-EPA-Physik.pdf

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/

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

Thüringer Kultusministerium (1999). *Lehrplan für das Gymnasium – Physik.* Saalfeld: SATZ+DRUCK Centrum Saalfeld.

TIMSS II (1995). *IEA's Third International Mathematics and Science Study. TIMSS Science Items: Released Set for Population 2 (Seventh and Eighth Grades)*. Chestnut Hill, MA: Boston College.

TIMSS III (1995). *IEA's Third International Mathematics and Science Study. TIMSS Science Items: Released Item Set for the Final Year of Secondary School Mathematics and Science Literacy, Advanced Mathematics, and Physics*. Chestnut Hill, MA: Boston College.

Viering, T. & Neumann, K. (2008). Competence items for measuring physics competence. *Leipniz Insitute for Science and Mathematics Education.*

Wagner, W., Kramer, J., Trautwein, U., Lüdtke, O., Nagy, G., Jonkmann, K., Maaz, K., Meixner, S., & Schilling, J. (2011). Upper secondary education in academic school tracks and the transition from school to postsecondary education and the job market. *Zeitschrift für Erziehungswissenschaft*, *14*, 233-249.

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. (1997). *ConQuest: Multi-aspect test software [computer program]*. Camberwell, Vic.: Australian Council for Educational Research.

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, physics competence, Waves 1-3;


datafile filename.dat;

format pid 1-7 responses 12-51;

labels << labels.nam;


codes 0,1;


model item;

set constraint=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 physics competence 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 8

Item Parameters of the Physics Competence 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	phyh10_c	17.71	1.711	0.081	0.99	-0.1	0.34
2	phyg2_c	58.46	-0.393	0.064	0.98	-1.0	0.44
3	phyg6_c	55.59	-0.258	0.064	1.04	2.0	0.36
4	phyg19_c	44.99	0.219	0.064	0.96	-2.3	0.48
5	phye1_c	87.02	-2.095	0.089	1.02	0.3	0.25
6	phyn14_c	29.89	0.950	0.069	0.99	-0.2	0.40
7	phyr1_c	85.43	-1.953	0.085	1.00	-0.1	0.34
8	phyt1_c	34.57	0.710	0.066	1.00	0.2	0.38
9	phyh12_c	27.22	1.097	0.070	0.93	-2.4	0.49
10	phyh6t_c	42.12	0.517	0.112	1.08	2.0	0.33
11	phyn2t_c	4.60	3.511	0.305	0.97	-0.0	0.30
12	phyn9t_c	15.28	2.152	0.176	1.00	0.0	0.36
13	phyn12t_c	17.39	1.790	0.119	0.95	-0.7	0.41
14	phyh5t_c	20.22	1.705	0.144	0.90	-1.3	0.53

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
15	phyh2_c	42.95	0.318	0.129	1.04	1.0	0.36
16	phyn11_c	47.67	0.102	0.128	0.97	-0.8	0.48
17	phyf5_c	48.04	0.087	0.127	0.98	-0.4	0.44
18	phyn6_c	48.33	0.076	0.135	1.06	1.5	0.35
19	phyn7_c	50.84	-0.034	0.128	1.02	0.4	0.41
20	phyf7_c	41.42	0.436	0.128	1.09	2.2	0.26
21	phyn5_c	48.39	0.112	0.126	0.94	-1.6	0.50
22	phyf13_c	57.01	-0.274	0.126	1.00	-0.1	0.42
23	phyf9_c	16.61	1.870	0.170	1.00	0.0	0.32
24	phyn3_c	54.89	-0.178	0.125	0.95	-1.2	0.51
25	phyt4a_c	79.87	-1.503	0.153	0.99	-0.1	0.34
26	phyt4b_c	66.11	-0.728	0.132	1.02	0.4	0.36
28	phyn8_c	9.35	2.484	0.204	1.02	0.2	0.19
29	phyb6_c	16.61	1.788	0.162	0.96	-0.4	0.37
30	phyh3_c	43.48	0.298	0.128	0.96	-1.2	0.47
31	phyh8_c	21.38	1.462	0.150	0.93	-0.9	0.50
32	phyh6_c	40.57	0.406	0.092	1.06	2.1	0.27

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
33	phyn2_c	23.09	1.280	0.108	1.10	1.7	0.16
34	phyn9_c	56.51	-0.324	0.093	1.08	2.7	0.25
35	phyn12_c	27.86	1.028	0.098	0.98	-0.5	0.42
36	phyh5_c	38.41	0.489	0.094	1.03	0.9	0.34
37	phyf4_c	21.99	1.327	0.152	0.96	-0.4	0.44
38	phyb24_c	14.90	1.860	0.171	0.99	-0.0	0.34
39	phym14_c	84.86	-1.942	0.165	1.08	0.7	0.09
40	phyg5_c	27.44	1.015	0.136	0.99	-0.2	0.40
41	phyg8_c	21.94	1.347	0.147	0.99	-0.1	0.38

Table 9

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	phyh10_c	-0.284	-0.214	0.175	0.311	0.136
2	phyg2_c	-0.196	0.064	-0.047	0.179	0.226
3	phyg6_c	-0.100	0.016	-0.255	-0.145	0.110
4	phyg19_c	-0.264	0.014	-0.171	-0.036	0.135
5	phye1_c	0.150	0.184	0.105	-0.057	-0.162
6	phyn14_c	-0.326	-0.080	0.036	0.207	0.171
7	phyr1_c	-0.888	0.436	0.141	0.609	0.468
8	phyt1_c	0.136	0.110	-0.266	-0.040	0.226
9	phyh12_c	-0.428	0.222	0.057	0.195	0.138
10	phyh6t_c	0.600	0.270	0.206	-0.317	-0.523
11	phyn2t_c	-0.632	-0.620	-0.360	-0.477	-0.117
12	phyn9t_c	0.678	0.368	-0.328	-0.421	-0.093
13	phyn12t_c	-0.934	0.806	0.261	0.006	-0.255
14	phyh5t_c	-0.198	-0.306	0.282	0.213	-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
15	phyh2_c	0.906	0.152	0.857	0.106	-0.751
16	phyn11_c	-0.754	-0.958	-0.474	-0.516	-0.042
17	phyf5_c	-0.568	0.480	-0.050	0.323	0.373
18	phyn6_c	0.474	0.490	0.277	-0.246	-0.523
19	phyn7_c	-0.056	-0.926	-0.540	-0.254	0.286
20	phyf7_c	0.698	-0.128	0.376	-0.098	-0.474
21	phyn5_c	0.146	0.124	-0.540	-0.720	-0.180
22	phyf13_c	0.506	-0.302	-0.105	-0.204	-0.099
23	phyf9_c	-0.008	0.016	-0.144	-0.029	0.115
24	phyn3_c	-0.550	0.106	-0.338	-0.055	0.283
25	phyt4a_c	0.304	0.050	-0.482	-0.823	-0.341
26	phyt4b_c	0.040	-0.006	-0.195	-1.080	-0.885
28	phyn8_c	0.602	-0.290	-0.508	0.214	0.722
29	phyb6_c	-0.044	0.176	0.199	0.827	0.628
30	phyh3_c	-0.184	-0.176	0.108	0.315	0.207
31	phyh8_c	-0.182	0.450	-0.149	0.289	0.438

		Gender	Immigration background	Books		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
32	phyh6_c	0.790	-0.486	0.579	-0.368	-0.947
33	phyn2_c	0.982	0.164	-0.132	-0.367	-0.235
34	phyn9_c	0.546	0.052	-0.072	-0.114	-0.042
35	phyn12_c	0.108	-0.188	-0.162	0.048	0.210
36	phyh5_c	0.694	-0.272	-0.040	-0.377	-0.337
37	phyf4_c	-0.278	-0.692	0.011	0.073	0.062
38	phyb24_c	-0.002	-0.286	-0.118	0.364	0.482
39	phym14_c	0.442	-1.034	-0.134	-0.466	-0.332
40	phyg5_c	0.508	-0.014	0.482	-0.179	-0.661
41	phyg8_c	0.042	0.404	0.105	0.210	0.105
	main effect	-0.718	0.158	0.030	0.174	0.144

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

Table 10

Item Parameters of the Physics Competence 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	phyh10_c	18.21	1.678	0.059	0.96	-1.2	0.39
2	phyg2_c	60.42	-0.483	0.048	0.98	-1.3	0.44
3	phyg6_c	58.03	-0.367	0.047	1.01	0.9	0.40
4	phyg19_c	44.73	0.234	0.047	0.96	-3.0	0.47
5	phye1_c	88.80	-2.280	0.069	1.04	0.8	0.22
6	phyn14_c	28.55	1.028	0.051	0.98	-0.8	0.41
7	phyr1_c	86.59	-2.064	0.064	0.99	-0.2	0.31
8	phyt1_c	34.83	0.700	0.049	1.01	0.5	0.39
9	phyh12_c	28.92	1.008	0.051	0.93	-3.5	0.49
10	phyh6t_c	34.36	0.873	0.086	1.04	1.3	0.35
11	phyn2t_c	2.42	4.099	0.298	0.95	-0.1	0.33
12	phyn9t_c	18.18	1.882	0.118	0.93	-1.1	0.48
13	phyn12t_c	14.80	1.967	0.093	0.93	-1.3	0.43
14	phyh5t_c	13.86	2.143	0.120	0.87	-1.7	0.52

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
15	phyh2_c	45.49	0.224	0.094	1.06	2.1	0.34
16	phyn11_c	40.28	0.468	0.094	0.94	-2.1	0.52
17	phyf5_c	43.18	0.328	0.093	1.05	1.8	0.36
18	phyn6_c	50.60	-0.016	0.099	1.06	2.1	0.33
19	phyn7_c	52.93	-0.121	0.093	0.96	-1.4	0.48
20	phyf7_c	37.56	0.537	0.095	1.10	3.0	0.27
21	phyn5_c	45.24	0.171	0.091	0.96	-1.4	0.50
22	phyf13_c	51.96	-0.145	0.090	1.02	0.8	0.42
23	phyf9_c	18.52	1.614	0.120	1.09	1.3	0.23
24	phyn3_c	57.17	-0.371	0.090	0.98	-0.7	0.46
25	phyt4a_c	73.76	-1.135	0.107	1.07	1.4	0.25
26	phyt4b_c	58.45	-0.362	0.098	1.04	1.2	0.37
28	phyn8_c	10.20	2.423	0.147	1.04	0.4	0.19
29	phyb6_c	16.09	1.862	0.125	0.94	-0.8	0.44
30	phyh3_c	39.50	0.508	0.099	0.96	-1.3	0.47
31	phyh8_c	23.67	1.326	0.109	0.96	-0.7	0.41
32	phyh6_c	38.04	0.570	0.070	1.10	4.0	0.23

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
33	phyn2_c	20.14	1.535	0.084	1.09	1.8	0.20
34	phyn9_c	58.57	-0.368	0.071	1.12	5.3	0.21
35	phyn12_c	28.15	1.065	0.073	0.96	-1.3	0.45
36	phyh5_c	37.19	0.601	0.071	1.04	1.5	0.34
37	phyf4_c	22.22	1.410	0.112	0.97	-0.5	0.39
38	phyb24_c	15.32	1.911	0.125	1.01	0.1	0.32
39	phym14_c	85.46	-1.923	0.125	1.02	0.2	0.25
40	phyg5_c	34.98	0.702	0.096	1.06	1.8	0.29
41	phyg8_c	24.29	1.282	0.107	0.95	-0.9	0.45

Table 11

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	phyh10_c	-0.128	0.052	-0.292	-0.077	0.215
2	phyg2_c	-0.188	0.150	0.169	0.368	0.199
3	phyg6_c	-0.152	0.044	-0.042	-0.107	-0.065
4	phyg19_c	-0.220	-0.012	0.113	0.008	-0.105
5	phye1_c	0.550	0.378	0.104	0.136	0.032
6	phyn14_c	-0.286	0.226	0.181	0.182	0.001
7	phyr1_c	-0.592	0.302	0.332	0.183	-0.149
8	phyt1_c	0.078	-0.224	-0.141	-0.260	-0.119
9	phyh12_c	-0.120	0.106	-0.021	0.030	0.051
10	phyh6t_c	0.386	0.180	-0.253	-0.449	-0.196
11	phyn2t_c	-0.806	-0.480	-0.993	-0.174	0.819
12	phyn9t_c	-0.160	-0.458	-0.097	-0.341	-0.244
13	phyn12t_c	-0.968	0.018	0.071	0.139	0.068
14	phyh5t_c	-0.436	0.170	-0.133	0.064	0.197

		Gender	Immigration background	Books		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
15	phyh2_c	0.728	-0.416	-0.021	-0.021	0.000
16	phyn11_c	-0.160	0.082	0.145	0.350	0.205
17	phyf5_c	-0.186	-0.174	-0.272	-0.319	-0.047
18	phyn6_c	0.528	-0.100	0.134	0.007	-0.127
19	phyn7_c	-0.434	-0.242	-0.058	0.136	0.194
20	phyf7_c	0.428	-0.306	0.318	0.225	-0.093
21	phyn5_c	-0.266	-0.088	-0.368	-0.361	0.007
22	phyf13_c	0.162	0.126	-0.112	0.058	0.170
23	phyf9_c	0.586	-0.058	-0.337	-0.221	0.116
24	phyn3_c	-0.034	-0.018	0.384	0.422	0.038
25	phyt4a_c	-0.122	0.380	0.022	-0.214	-0.236
26	phyt4b_c	-0.116	-0.110	-0.309	-0.208	0.101
28	phyn8_c	0.688	-1.322	0.037	0.365	0.328
29	phyb6_c	-0.002	-0.072	0.026	-0.090	-0.116
30	phyh3_c	-0.196	-0.176	0.110	0.058	-0.052
31	phyh8_c	0.286	0.006	-0.286	-0.563	-0.277

		Gender	**Immigration background**	**Books**		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
32	phyh6_c	0.650	0.386	-0.097	-0.140	-0.043
33	phyn2_c	0.628	-0.184	-0.263	-0.338	-0.075
34	phyn9_c	0.438	-0.038	-0.159	-0.132	0.027
35	phyn12_c	-0.196	0.120	0.212	0.103	-0.109
36	phyh5_c	0.282	-0.282	-0.041	-0.010	0.031
37	phyf4_c	-0.112	-0.296	-0.038	-0.130	-0.092
38	phyb24_c	0.568	-0.046	-0.616	-0.469	0.147
39	phym14_c	0.434	-0.406	0.213	0.446	0.233
40	phyg5_c	0.380	-0.358	0.417	0.319	-0.098
41	phyg8_c	-0.072	-0.186	-0.334	-0.188	0.146
	main effect	-0.670	0.186	0.083	0.191	0.108

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

Table 12

Item Parameters of the Physics Competence 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	phyh10_c	17.40	1.763	0.086	1.01	0.3	0.33
2	phyg2_c	59.54	-0.451	0.068	0.96	-1.7	0.50
3	phyg6_c	56.36	-0.299	0.067	0.99	-0.4	0.44
4	phyg19_c	45.08	0.219	0.067	0.97	-1.5	0.47
5	phye1_c	87.30	-2.153	0.093	1.02	0.3	0.26
6	phyn14_c	29.00	1.017	0.073	0.94	-1.9	0.47
7	phyr1_c	84.80	-1.929	0.087	1.00	-0.1	0.33
8	phyt1_c	35.84	0.658	0.069	1.02	0.7	0.39
9	phyh12_c	27.53	1.097	0.073	0.92	-2.5	0.52
10	phyh6t_c	35.73	0.835	0.117	1.05	1.1	0.37
11	phyn2t_c	2.64	4.060	0.392	0.97	0.0	0.32
12	phyn9t_c	18.75	1.889	0.165	0.96	-0.4	0.43
13	phyn12t_c	17.54	1.777	0.122	0.92	-1.2	0.48
14	phyh5t_c	14.37	2.165	0.168	0.89	-1.1	0.54

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
15	phyh2_c	48.68	0.034	0.128	1.11	2.6	0.31
16	phyn11_c	42.76	0.311	0.130	0.96	-0.9	0.50
17	phyf5_c	48.54	0.042	0.127	1.02	0.6	0.40
18	phyn6_c	50.93	-0.034	0.136	1.12	2.7	0.25
19	phyn7_c	54.13	-0.210	0.129	1.02	0.4	0.45
20	phyf7_c	43.86	0.265	0.134	1.14	3.1	0.23
21	phyn5_c	47.80	0.077	0.131	1.00	-0.1	0.43
22	phyf13_c	51.84	-0.118	0.130	0.95	-1.2	0.50
23	phyf9_c	17.24	1.763	0.178	1.09	0.9	0.25
24	phyn3_c	60.33	-0.500	0.131	0.97	-0.7	0.48
25	phyt4a_c	78.65	-1.429	0.161	1.02	0.3	0.33
26	phyt4b_c	66.91	-0.769	0.142	0.99	-0.2	0.41
28	phyn8_c	8.30	2.704	0.229	1.04	0.3	0.21
29	phyb6_c	20.07	1.604	0.162	0.96	-0.4	0.43
30	phyh3_c	35.34	0.737	0.142	0.97	-0.6	0.46
31	phyh8_c	22.66	1.427	0.156	0.93	-0.8	0.49
32	phyh6_c	41.23	0.415	0.098	1.12	3.6	0.24

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score
33	phyn2_c	18.24	1.686	0.124	1.03	0.5	0.26
34	phyn9_c	66.02	-0.728	0.103	1.16	3.8	0.15
35	phyn12_c	25.67	1.214	0.107	0.99	-0.2	0.40
36	phyh5_c	39.23	0.498	0.100	1.06	1.7	0.31
37	phyf4_c	23.22	1.323	0.157	1.01	0.2	0.34
38	phyb24_c	14.75	1.944	0.181	1.00	0.1	0.33
39	phym14_c	88.65	-2.271	0.197	1.01	0.1	0.27
40	phyg5_c	27.14	1.096	0.147	1.02	0.4	0.34
41	phyg8_c	19.15	1.602	0.163	0.98	-0.1	0.39

Table 13

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	phyh10_c	-0.060	-0.532	0.060	0.072	0.012
2	phyg2_c	0.010	0.264	0.385	0.632	0.247
3	phyg6_c	-0.134	-0.080	-0.064	-0.038	0.026
4	phyg19_c	-0.308	0.092	0.038	0.082	0.044
5	phye1_c	0.284	0.336	-0.231	0.081	0.312
6	phyn14_c	-0.326	0.150	-0.317	-0.212	0.105
7	phyr1_c	-0.488	0.188	-0.146	-0.358	-0.212
8	phyt1_c	-0.038	-0.006	-0.055	-0.112	-0.057
9	phyh12_c	-0.502	0.324	-0.133	0.029	0.162
10	phyh6t_c	0.426	0.114	0.282	0.573	0.291
11	phyn2t_c	-0.898	n.a.	0.828	0.924	0.096
12	phyn9t_c	0.212	-1.062	0.170	0.412	0.242
13	phyn12t_c	-0.482	0.684	-0.026	0.275	0.301
14	phyh5t_c	-0.660	-0.400	-0.040	0.020	0.060

		Gender	Immigration background	Books		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
15	phyh2_c	0.744	-0.482	0.191	0.023	-0.168
16	phyn11_c	-0.452	0.222	0.065	0.086	0.021
17	phyf5_c	-0.196	-0.372	-0.036	0.264	0.300
18	phyn6_c	1.100	-0.558	-0.144	-0.537	-0.393
19	phyn7_c	-0.432	-0.186	0.341	-0.035	-0.376
20	phyf7_c	0.692	-0.722	0.110	0.133	0.023
21	phyn5_c	-0.120	0.554	-0.354	0.096	0.450
22	phyf13_c	-0.568	0.322	-0.197	-0.143	0.054
23	phyf9_c	0.634	-0.700	0.276	-0.042	-0.318
24	phyn3_c	-0.738	0.392	0.579	0.213	-0.366
25	phyt4a_c	-0.184	0.114	0.099	0.297	0.198
26	phyt4b_c	0.166	0.096	0.411	-0.474	-0.885
28	phyn8_c	0.616	-0.120	0.133	-0.064	-0.197
29	phyb6_c	0.068	-0.296	0.024	-0.012	-0.036
30	phyh3_c	-0.348	0.056	-0.114	0.432	0.546
31	phyh8_c	0.350	-0.080	0.225	-0.300	-0.525

		Gender	Immigration background	Books		
	Item	male vs female	without vs with	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
32	phyh6_c	0.684	0.122	-0.214	-0.293	-0.079
33	phyn2_c	0.422	-0.438	0.005	-0.443	-0.448
34	phyn9_c	0.770	-0.150	-0.257	-0.353	-0.096
35	phyn12_c	0.108	0.258	0.018	-0.120	-0.138
36	phyh5_c	0.676	-0.444	-0.451	-0.353	0.098
37	phyf4_c	0.460	-0.248	0.026	-0.527	-0.553
38	phyb24_c	0.376	-0.072	0.481	0.035	-0.446
39	phym14_c	0.172	0.094	-0.003	-0.519	-0.516
40	phyg5_c	0.246	-0.442	0.036	-0.162	-0.198
41	phyg8_c	0.046	0.156	0.246	0.184	-0.062
	main effect	-0.688	0.264	0.173	0.373	0.200