NEPS SURVEY PAPERS

Sven Rieger, Nicolas Hübner, and Wolfgang Wagner

NEPS TECHNICAL REPORT FOR PHYSICS COMPETENCE: SCALING RESULTS FOR THE ADDITIONAL STUDY THURINGIA

NEPS Survey Paper No. 40
Bamberg, June 2018

LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

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 Thuringia

Sven Rieger, Nicolas Hübner, & Wolfgang Wagner

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

E-mail address of lead author:

sven.rieger@uni-tuebingen.de

Bibliographische Angaben:

Rieger, S., Hübner, N., & Wagner, W. (2018). *NEPS Technical Report for Physics Competence: Scaling Results for the Additional Study Thuringia* (NEPS Survey Paper No. 40). Bamberg: Leibniz-Institute for Educational Trajectories, National Educational Panel Study.

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

NEPS Technical Report for Physics Competence: Scaling Results for the Additional Study Thuringia

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 Thuringia. In sum, 2,254 students took the test in two waves. The physics competence test consisted of 64 items (distributed among nine booklets). A Rasch model was used to scale the data. Item fit statistics and differential item functioning were investigated. The results showed that a subset of 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 as well as Con-Quest- and TAM-syntaxes for scaling the data.

Keywords

item response theory, scaling, physics competence, scientific use file

Content

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 two waves of the additional study Thuringia. In this study, items were composed for the physics competence test used across two consecutive school years (2009/10 and 2010/11) 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 as well as further information about the test can be found on the NEPS website[1].

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

2 Testing Physics Competence

The framework and item development is 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/dataanddocumentation/additionalstudythuringia.aspx

Table 1

Source of Items in the Physics Competence Test

Source	**Frequency**
TIMSS II	3
TIMSS III	24
Thermodynamik Testinventar[1]	10
BEMA[2]	4
Proprietary development[3]	23
Total number of items	64

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 as well as booklets (for the content area of each item see Table S2 in the Appendix D).

Table 2

Content Areas of the Items on the Physics Competence Test

Content area	**Frequency**
Electrical fields and interdependency	6
Magnetic fields and electromagnetic induction	12
Waves	8
Optics	8
Quantum physics: Quanta and matter	5
Dynamics: Mechanics of the Rigid Body	7
Thermodynamics	16
Special Theory of Relativity	2
Total number of items	64

Number of Items by Content Area and Booklet	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Electrical fields and interdependency	2	2	3	2	2	2	2	1	2
Magnetic fields and electromagnetic induction	2	3	5	3	2	3	4	5	4
Waves	4	3	1	2	3	3	3	2	2
Optics	2	2	2	2	3	3	3	3	3
Quantum physics: Quanta and matter	1	1	1	1	1	2	2	1	-
Dynamics: Mechanics of the Rigid Body	1	2	2	2	2	1	1	2	1
Thermodynamics	5	3	4	8	10	6	3	2	4
Special Theory of Relativity	1	1	-	-	-	-	-	1	1
Total number of items	18	17	18	20	23	20	18	17	17

Table 3

Response Formats of the Items on the Physics Competence Test

Response format	**Frequency**
Single multiple choice	51
Complex multiple choice	7
Short constructed response	6
Total number of items	64

Number of Items by Response Format and Booklet	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Single multiple choice	17	18	18	17	16	17	18	12	12
Complex multiple choice	-	-	-	4	7	3	-	-	-
Short constructed response	1	-	-	-	-	-	-	5	5
Total number of items	18	18	18	21	23	20	18	17	17

3 Data

A description of the design of the study, the sample, as well as the instruments that were used can be found on the NEPS website[2]. A total of 2,254 participants took the physics competence test: 1,370 in 2009/2010 (Wave 1), and 884 in 2010/2011 (Wave 2)[3]. All subjects gave at least one valid answer so that for every subject, a competence score was estimated.

4 Analyses

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

4.1 Missing Responses

There are different types of missing responses in competence test data. These include (among others) missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach, and d) items that are missing by design (e.g., due to the different booklets). Missing responses provide information about how well the test worked (e.g., time limits, whether participants understood the instructions, how participants handled different response formats), and they need to be accounted for in the estimation of item and person parameters. We thoroughly inspected the occurrence of missing responses per person. This provided an indication of how well the test takers coped with the test. We then examined the occurrence of missing responses per item in order to obtain some information about how well the items performed. In addition, information was available about whether students did not take the physics competence test (e.g., due to student tardiness) but did take at least one of the other competence tests (mathematics, or biology). 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 physics competence, a Rasch model (Rasch, 1960) was used and estimated in ConQuest 4.2 (Wu, Adams, Wilson, & Haldane, 2007).

Item parameters are estimated difficulties for dichotomous variables in the Rasch model. Ability estimates for 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.

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

[3] The dataset contains 2,260 persons.

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

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

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

5 Results

In this section, the key scaling results of the two waves of the additional study Thuringia will be presented.

5.1 Missing Responses

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

5.1.1 Missing responses per person

Figure 1 shows the number of *invalid responses* per person. As can be seen, 5.75% of the participants produced any invalid responses. The maximum number of invalid responses was 6.

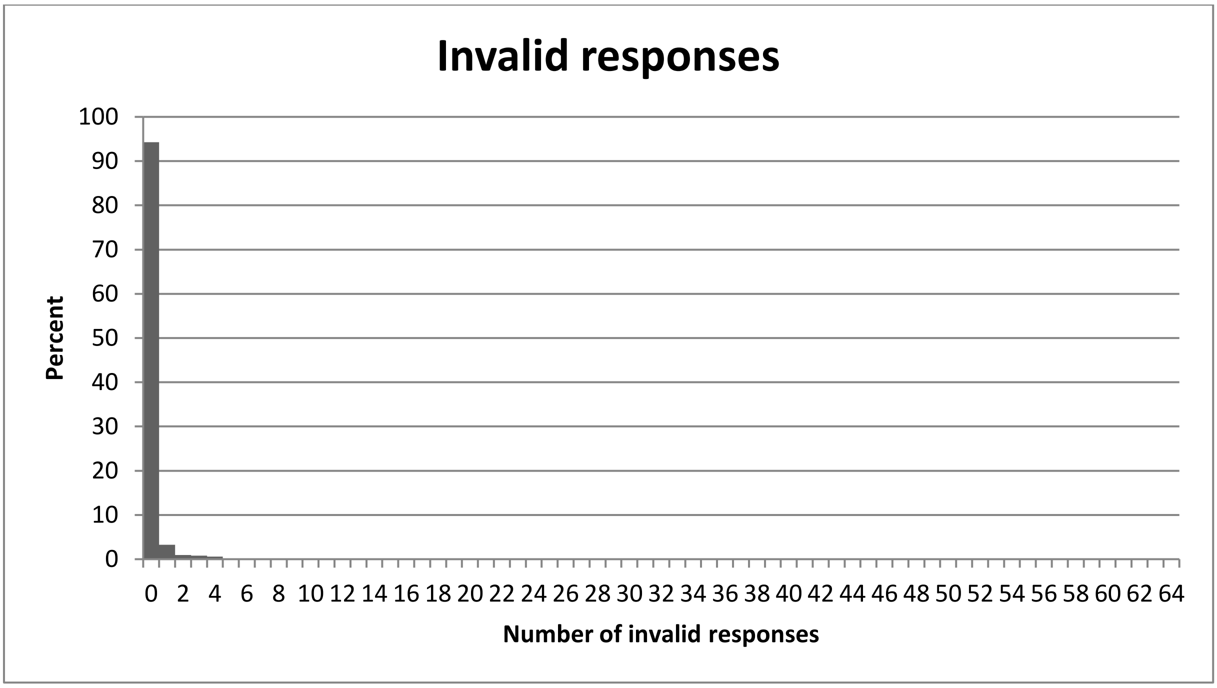

Figure 1. Number of invalid responses per person.

The largest source of missing responses on this test was the *omission of items*. As can be seen in Figure 2, almost one out of four of the participants (22.54%) skipped at least one item. Overall, 3.15% of the participants omitted five or more items.

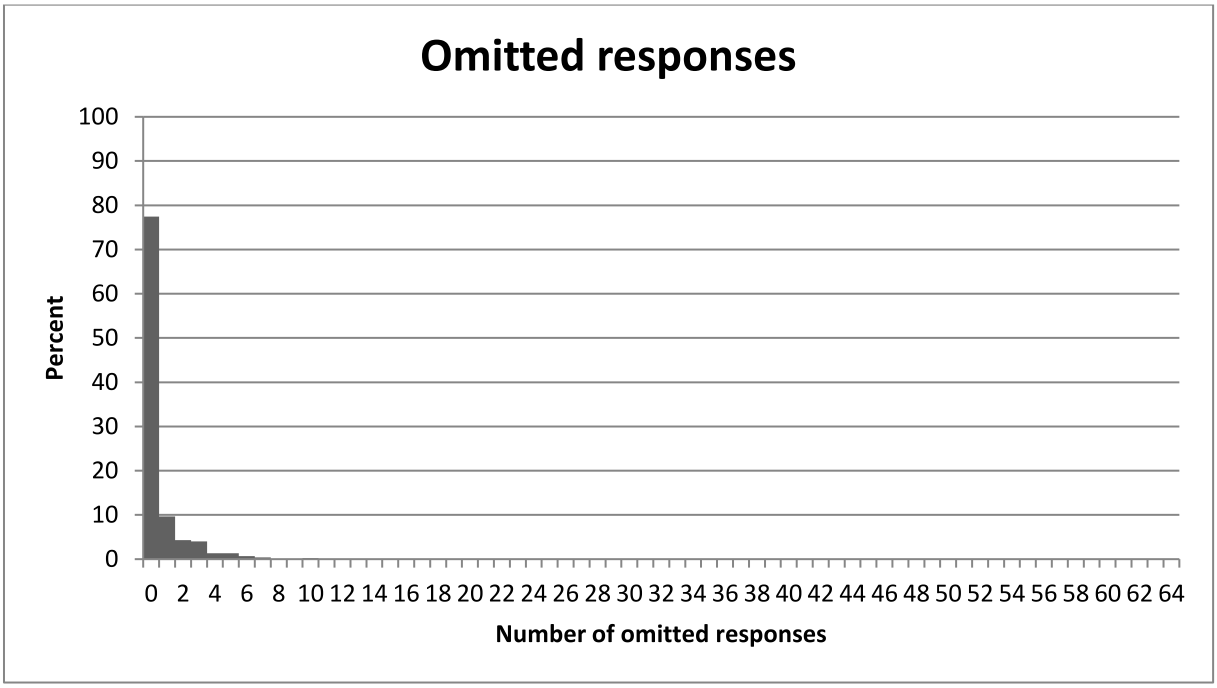

Figure 2. Number of omitted responses per person.

By definition, every item after the last item that was completed is labeled *not reached*. As Figure 3 shows, most participants (89.16%) reached the end of the test. Only 1.22% did not reach the fifth last item.

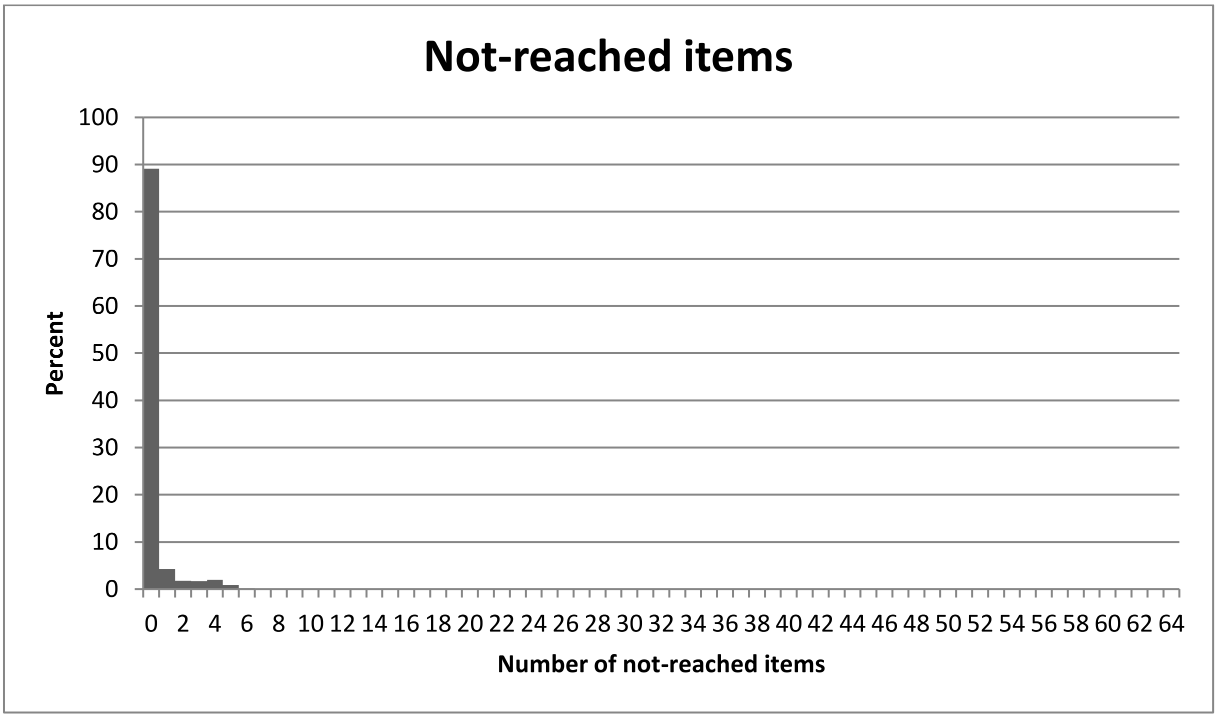

Figure 3. Number of not-reached items per person.

Overall, 89.16% of the participants had no items that were missing by *non-participation*. Only 0.27% 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 the invalid, omitted, and not-reached missing responses per person is illustrated in Figure 4. On average, the participants produced 0.95 (SD = 1.97) missing responses. Moreover, 68.27% of the persons had no missing response at all. Only 5.00% of the participants had five or more missing responses.

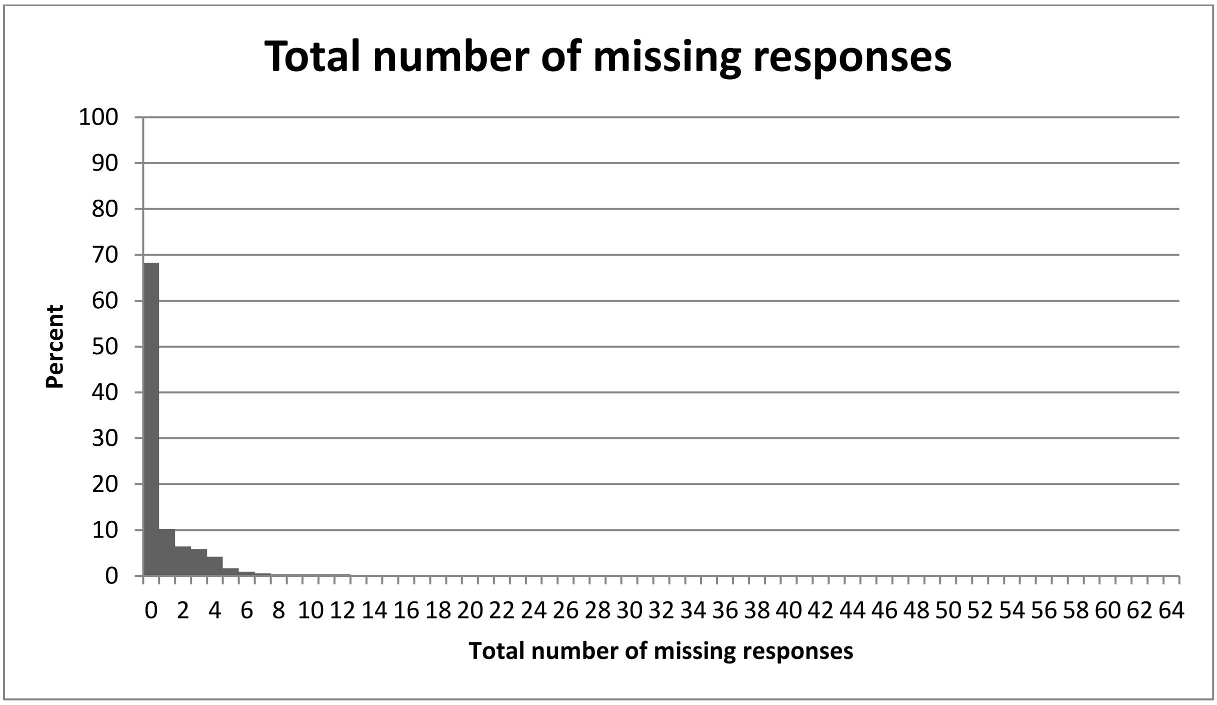

Figure 4. Total number of missing responses.

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 6.3% of the participants failed to reach items (column 5). None of the 64 items had an omission rate that exceeded 5% (column 6). Overall, the percentage of invalid responses per item (column 7) was very low (the maximum was 1.7% for item phyn9t_c). The percentage of items that were missing by non-participation (column 8) was very low (the maximum was 0.3%). The percentage of missing by designs per items is displayed in column 8. The percentages ranged from 76.8% to 81.2%.

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 9.6% to 88.2% with an average of 38.57 % (*SD* = 20.02%) 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 -2.187 (item phyr1_c) to 2.627 (item phyn2t_c) logits with an average difficulty of 0.64 logits (*SD* = 1.10). Altogether, the item difficulties were somewhat high. The 2PL discrimination parameters ranged from 0.040 to 3.695 (see again Table 5). The items phye6_c, phyt13b_c, phyt13c_c, phyg13_c, phyb18_c, phyn2_c, and phyt9_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, those items were excluded from further analyses (see Table S1 in Appendix C).

Table 4

Missing Values

	Item	Booklet	Position in the test	Number of valid responses	Percentage of not-reached responses	Percentage of omitted responses	Percentage of invalid responses	Percentage of missing by non-participation	Percentage of missing by design
1	phyh10_c	1-9	1	2210	-	1.7	0.2	0.3	-
2	phyg1_c	1-9	2	2198	-	2.5	-	0.3	-
3	phyn5_c	1-9	3	2168	-	3.7	0.1	0.3	-
4	phyr1_c	1-9	4	2248	-	0.2	0.0	0.3	-
5	phyg2_c	1-9	5	2212	0.0	0.5	1.3	0.3	-
6	phye2_c	1-9	6	2196	0.0	2.5	-	0.3	-
7	phyh8_c	1,9	7/12	510	-	0.3	0.1	0.3	76.8
8	phyn1_c	1,9	8/13	508	-	0.4	0.0	0.3	76.8
9	phyg8_c	1,9	9/14	505	0.0	0.5	0.0	0.3	76.8
10	phym14_c	1,9	10/15	506	-	0.4	0.1	0.3	76.8
11	phyt1_c	1,9	11/16	513	0.1	0.1	0.0	0.3	76.8
12	phyg6_c	1,9	12/17	496	0.6	0.3	0.1	0.3	76.8
13	phyh12_c	1,2	13/7	473	-	1.1	0.1	0.3	77.6
14	phyn12_c	1,2	14/8	487	0.0	0.5	-	0.3	77.6

	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	1,2	15/9	490	0.1	0.2	0.2	0.3	77.6
16	phyh5_c	1,2	16/10	477	0.1	0.9	-	0.3	77.6
17	phyn7_c	1,2	17/11	487	0.1	0.4	0.1	0.3	77.6
18	phyf3_c	1,2	18/12	473	0.4	0.7	0.0	0.3	77.6
19	phyb6_c	2,3	13/7	482	-	0.5	0.1	0.3	77.7
20	phyg4_c	2,3	14/8	487	-	0.4	-	0.3	77.7
21	phyn4_c	2,3	15/9	472	-	1.1	0.0	0.3	77.7
22	phyn10_c	2,3	16/10	488	-	0.4	-	0.3	77.7
23	phyf5_c	2,3	17/11	496	-	0.0	-	0.3	77.7
24	phyn13_c	2,3	18/12	486	0.2	0.3	0.0	0.3	77.7
25	phyb14_c	3,4	13/7	448	0.0	2.1	0.1	0.3	77.7
26	phyh6_c	3,4	14/8	476	0.0	0.9	0.0	0.3	77.7
27	phyn6_c	3,4	15/9	456	0.0	0.4	0.2	0.3	77.7
28	phyn15_c	3,4	16/10	485	0.0	0.4	0.2	0.3	77.7
29	phyt3_c	3,4	17/11	486	0.1	0.4	0.0	0.3	77.7
30	phyf1_c	3,4	18/12	466	0.4	0.8	0.2	0.3	77.7

	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	phye6_c	4,5	13/7	474	-	1.0	0.0	0.3	77.7
32	phye1_c	4,5	14/8	492	-	0.1	0.1	0.3	77.7
33	phyn9_c	4,5	15/9	447	-	2.2	-	0.3	77.7
34	phyo13_c	4,5	16/10	497	-	-	-	0.3	77.7
35	phyt13a_c	4,5	17/11	456	0.6	1.2	-	0.3	77.7
36	phyt13b_c	4,5	18/12	459	0.6	1.1	-	0.3	77.7
37	phyt13c_c	4,5	19/13	452	0.6	1.3	0.0	0.3	77.7
38	phyt13d_c	4,5	20/14	457	0.6	1.2	-	0.3	77.7
39	phyf9_c	4,5	21/15	448	1.2	0.7	0.2	0.3	77.7
40	phyf6_c	5,6	16/7	475	-	0.9	0.0	0.3	77.8
41	phyg13_c	5,6	17/8	493	-	0.1	-	0.3	77.8
42	phyn8_c	5,6	18/9	482	0.0	0.4	0.1	0.3	77.8
43	phyn14_c	5,6	19/10	483	0.1	0.4	0.1	0.3	77.8
44	phyt4a_c	5,6	20/11	471	0.1	1.0	-	0.3	77.8
45	phyt4b_c	5,6	21/12	467	0.1	1.2	-	0.3	77.8
46	phyt4c_c	5,6	22/13	477	0.2	0.6	0.0	0.3	77.8

	Item	Booklet	Position in the test	Number of valid responses	Percentage of not-reached responses	Percentage of omitted responses	Percentage of invalid responses	Percentage of missing by non-participation	Percentage of missing by design
47	phyf7_c	5,6	23/14	471	0.4	0.5	0.2	0.3	77.8
48	phyb18_c	6,7	15/7	462	-	0.9	-	0.3	78.4
49	phyn3_c	6,7	16/8	477	-	0.3	-	0.3	78.4
50	phyn2_c	6,7	17/9	443	-	1.8	-	0.3	78.4
51	phyg5_c	6,7	18/10	478	-	0.1	0.1	0.3	78.4
52	phyt9_c	6,7	19/11	468	0.1	0.5	0.1	0.3	78.4
53	phyh3_c	6,7	20/12	465	0.4	0.3	0.0	0.3	78.4
54	phyb24_c	7,8	13/7	475	0.1	0.6	0.2	0.3	77.8
55	phyg19_c	7,8	14/8	489	0.1	0.0	0.1	0.3	77.8
56	phyf13_c	7,8	15/9	479	0.1	0.6	-	0.3	77.8
57	phyn11_c	7,8	16/10	483	0.1	0.2	0.3	0.3	77.8
58	phyf4_c	7,8	17/11	475	0.4	0.6	-	0.3	77.8
59	phyh15_c	7,8	18/12	444	1.1	1.2	-	0.3	77.8
60	phyn12t_c	8,9	13/7	491	0.4	0.3	0.0	0.3	77.3
61	phyh5t_c	8,9	14/8	354	2.3	1.6	0.7	0.3	79.5
62	phyh6t_c	8,9	15/9	243	3.8	2.3	1.6	0.3	81.2

	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
63	phyn9t_c	8,9	16/10	212	5.2	2.5	1.7	0.3	81.0
64	phyn2t_c	8,9	17/11	189	6.3	2.3	1.6	0.3	81.2

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	16.22	1.799	0.062	0.97	-0.7	0.36	1.115
2	phyg1_c	35.31	0.666	0.049	1.07	3.9	0.26	0.318
3	phyn5_c	42.96	0.311	0.048	0.94	-4.5	0.50	1.632
4	phyr1_c	88.20	-2.187	0.069	0.97	-0.7	0.32	1.523
5	phyg2_c	60.93	-0.494	0.048	0.98	-1.1	0.43	1.145
6	phye2_c	60.47	-0.477	0.048	1.00	0.3	0.37	0.832
7	phyh8_c	12.97	2.081	0.140	0.98	-0.2	0.37	1.093
8	phyn1_c	28.80	0.965	0.107	1.05	1.1	0.30	0.478
9	phyg8_c	24.40	1.230	0.113	0.90	-1.8	0.54	1.890

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
10	phym14_c	82.18	-1.730	0.124	1.02	0.2	0.28	0.612
11	phyt1_c	39.45	0.436	0.099	0.98	-0.6	0.45	1.130
12	phyg6_c	64.24	-0.701	0.103	1.00	-0.0	0.38	0.824
13	phyh12_c	16.49	1.743	0.132	0.97	-0.3	0.40	1.307
14	phyn12_c	26.90	1.069	0.110	0.97	-0.6	0.42	1.109
15	phyh2_c	38.78	0.464	0.101	1.01	0.2	0.39	0.807
16	phyh5_c	39.83	0.412	0.102	1.01	0.4	0.38	0.787
17	phyn7_c	39.63	0.430	0.101	0.98	-0.6	0.46	1.078
18	phyf3_c	34.46	0.676	0.105	1.05	1.2	0.30	0.466
19	phyb6_c	16.60	1.774	0.129	1.00	-0.0	0.35	0.925
20	phyg4_c	33.06	0.792	0.104	1.09	2.3	0.19	0.040
21	phyn4_c	11.44	2.234	0.151	1.02	0.2	0.20	0.423
22	phyn10_c	28.89	1.008	0.107	1.01	0.1	0.36	0.886
23	phyf5_c	46.57	0.166	0.098	1.00	0.1	0.39	0.812
24	phyn13_c	35.60	0.667	0.102	1.09	2.4	0.21	0.070
25	phyb14_c	9.62	2.430	0.166	1.03	0.3	0.10	0.120
26	phyh6_c	35.79	0.683	0.103	1.08	2.3	0.21	0.084

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
27	phyn6_c	37.80	0.588	0.104	1.02	0.5	0.34	0.703
28	phyn15_c	24.17	1.278	0.113	1.03	0.6	0.29	0.433
29	phyt3_c	31.55	0.884	0.105	1.02	0.5	0.27	0.442
30	phyf1_c	51.18	-0.009	0.100	1.06	2.1	0.28	0.117
31	phye6_c	-	-	-	-	-	-	-
32	phye1_c	79.84	-1.486	0.119	1.04	0.6	0.25	0.575
33	phyn9_c	50.67	-0.033	0.103	1.08	2.9	0.26	0.214
34	phyo13_c	77.82	-1.355	0.115	1.04	0.7	0.25	0.401
35	phyt13a_c	81.76	-1.633	0.128	1.07	0.9	0.16	0.156
36	phyt13b_c	-	-	-	-	-	-	-
37	phyt13c_c	-	-	-	-	-	-	-
38	phyt13d_c	40.79	0.390	0.103	1.04	1.2	0.32	0.476
39	phyf9_c	19.69	1.515	0.126	0.99	-0.1	0.33	0.855
40	phyf6_c	18.11	1.630	0.126	0.97	-0.4	0.37	0.902
41	phyg13_c	-	-	-	-	-	-	-
42	phyn8_c	21.58	1.400	0.118	0.92	-1.3	0.47	1.777
43	phyn14_c	33.75	0.731	0.104	0.96	-1.0	0.43	1.103

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
44	phyt4a_c	74.95	-1.206	0.114	1.01	0.2	0.34	0.577
45	phyt4b_c	62.96	-0.593	0.104	1.03	0.9	0.32	0.526
46	phyt4c_c	22.43	1.356	0.117	1.05	0.9	0.21	0.254
47	phyf7_c	36.31	0.601	0.104	1.04	1.2	0.29	0.412
48	phyb18_c	-	-	-	-	-	-	-
49	phyn3_c	56.09	-0.280	0.101	0.96	-1.3	0.48	1.283
50	phyn2_c	-	-	-	-	-	-	-
51	phyg5_c	33.33	0.763	0.105	1.00	0.0	0.42	0.890
52	phyt9_c	-	-	-	-	-	-	-
53	phyh3_c	36.42	0.610	0.105	0.99	-0.2	0.43	0.971
54	phyb24_c	14.98	1.921	0.137	1.04	0.5	0.28	0.704
55	phyg19_c	47.54	0.112	0.100	0.98	-0.8	0.46	1.024
56	phyf13_c	50.42	-0.026	0.101	0.93	-2.5	0.53	1.540
57	phyn11_c	34.23	0.721	0.105	0.96	-1.1	0.48	1.317
58	phyf4_c	24.47	1.252	0.116	0.97	-0.6	0.47	1.365
59	phyh15_c	30.93	0.875	0.112	1.10	2.2	0.22	0.200
60	phyn12t_c	14.69	0.610	0.105	0.99	-0.2	0.49	1.759

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
61	phyh5t_c	20.40	1.972	0.136	0.93	-0.8	0.61	2.616
62	phyh6t_c	27.69	1.585	0.143	0.87	-1.7	0.49	1.384
63	phyn9t_c	10.43	1.270	0.157	0.95	-0.7	0.45	1.663
64	phyn2t_c	12.77	2.627	0.238	0.97	-0.1	0.62	3.695

Note. Difficulty = Item difficulty / location parameter, SE = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, t = t-value for WMNSQ. Items 31, 36, 37, 41, 48, 50, and 52 were excluded from the analyses due to an unsatisfactory item fit.

5.2.2 Person parameters

The person parameters were estimated as WLEs (Pohl & Carstensen, 2012). 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 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 0.497[5], indicating a reasonable differentiation between the subjects. The reliability of the test (EAP/PV reliability = .58, WLE reliability = .55) was modest. This should be related to the suboptimal test targeting described above.

[5] Seven items (i.e., phye6_c, phyt13b_c, phyt13c_c, phyg13_c, phyb18_c, phyn2_c, and phyt9_c) were excluded due to negative item discriminations (see also Section 5.2.1).

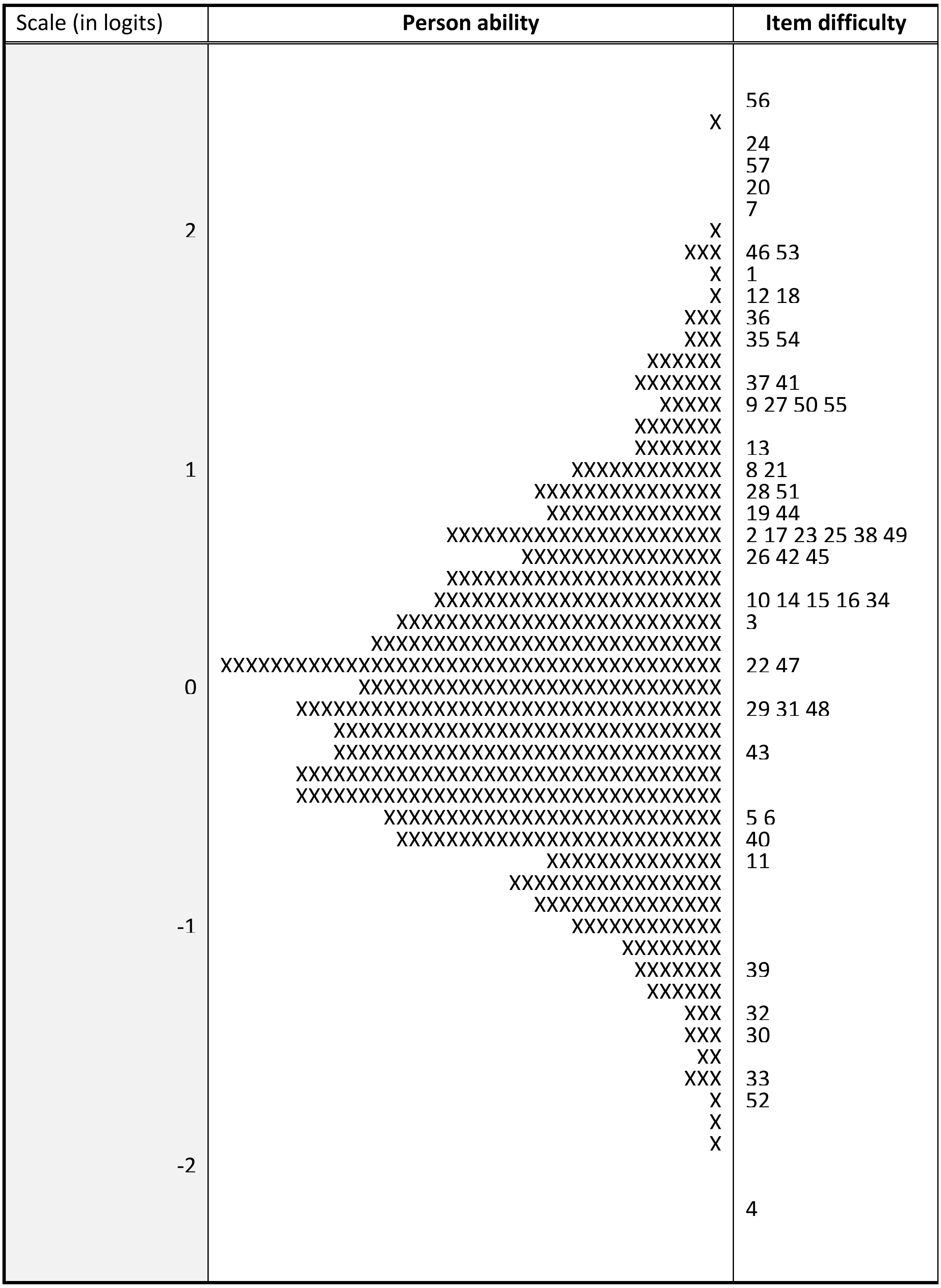

Figure 5. Test targeting. The distribution of person abilities in the sample is depicted on the left-hand side, with each 'X' representing 3.8 cases. The item difficulties (or location parameters) are depicted on the right-hand side. Each number represents one item with a corresponding position in the 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 0.90 (item phyg8_c) to 1.10 (item phyh15_c), cf. column 5 of Table 5. Point-biserial correlations between the item scores and the total scores ranged from 0.10 (items phyb14_c and phyb18_c) to 0.62 (item phyn2t_c). Discriminations estimated in the 2PL-model with the TAM package in R ranged from 0.040 (item phyg4_c) to 3.695 (item phyn2t_c), cf. Table 5, column 8.

5.3.2 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i. e., measurement invariance). For this purpose, DIF was examined for the variables gender, 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 and Carstensen (2012), absolute difficulty differences greater than 1 logit can be considered to show very strong DIF. For the current test, four items 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 = -0.594 logits, Cohen's *d* = -0.843).[6] Fourteen items (see Table 6) showed a DIF greater than 0.6 logits. Three items (phyn12t_c, phyn9t_c, and phyn2t_c) showed a very strong DIF reaching 1 logit.

Wave: On average, participants in the two waves basically did not differ in their physics competence (main effect = 0.020, Cohen's d = 0.028). 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 = 0.123, Cohen's *d* = 0.174; 0-200 vs > 500: main effect = 0.327, Cohen's *d* = 0.464; 201-500 vs > 500: main effect = 0.204, Cohen's *d* = 0.289). Ten items (phyt1_c, phyh12_c, phyn7_c, phyn4_c, phyn13_c, phyb14_c, phye1_c, phyn12t_c, phyh5t_c, phyn9t_c) showed a DIF greater than 0.6 logits. Item phyn2t_c showed a very strong DIF exceeding 1 logit.

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

Table 6

Differential Item Functioning

		Gender	Wave	Books		
	Item	male vs female	1 vs 2	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
1	phyh10_c	-0.228	0.024	0.086	0.040	-0.046
2	phyg1_c	0.272	-0.158	-0.030	-0.333	-0.303
3	phyn5_c	-0.252	0.116	-0.171	0.081	0.252
4	phyr1_c	-0.742	-0.074	-0.106	0.220	0.326
5	phyg2_c	-0.328	-0.112	0.098	0.323	0.225
6	phye2_c	-0.096	0.118	-0.111	-0.246	-0.135
7	phyh8_c	-0.258	0.110	-0.109	0.239	0.348
8	phyn1_c	0.756	-0.382	-0.093	-0.018	0.075
9	phyg8_c	-0.460	-0.060	-0.059	0.089	0.148
10	phym14_c	0.480	-0.144	0.327	0.209	-0.118
11	phyt1_c	-0.158	0.134	0.616	0.056	-0.560
12	phyg6_c	0.006	0.304	0.001	0.410	0.409
13	phyh12_c	-0.062	-0.334	-0.413	0.269	0.682
14	phyn12_c	-0.602	-0.302	0.000	0.036	0.036

		Gender	Wave	Books		
	Item	male vs female	1 vs 2	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
15	phyh2_c	0.288	-0.160	-0.120	-0.123	-0.003
16	phyh5_c	0.326	-0.032	-0.144	-0.216	-0.072
17	phyn7_c	-0.368	-0.060	0.402	0.957	0.555
18	phyf3_c	0.438	-0.052	-0.358	-0.266	0.092
19	phyb6_c	-0.350	-0.006	-0.204	-0.114	0.090
20	phyg4_c	0.820	-0.180	0.404	-0.047	-0.451
21	phyn4_c	0.214	0.384	0.868	0.801	-0.067
22	phyn10_c	0.402	0.180	-0.197	-0.139	0.058
23	phyf5_c	-0.556	-0.092	0.216	-0.300	-0.516
24	phyn13_c	0.936	0.246	-0.379	-0.680	-0.301
25	phyb14_c	0.634	0.366	0.767	0.499	-0.268
26	phyh6_c	0.928	0.430	-0.252	-0.306	-0.054
27	phyn6_c	0.226	0.174	-0.043	-0.362	-0.319
28	phyn15_c	0.540	-0.052	-0.174	-0.081	0.093
29	phyt3_c	0.406	0.180	-0.168	-0.301	-0.133
30	phyf1_c	0.812	-0.142	-0.270	-0.477	-0.207

		Gender	Wave	Books		
	Item	male vs female	1 vs 2	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
31	phye6_c	-	-	-	-	-
32	phye1_c	0.256	0.014	0.290	-0.323	-0.613
33	phyn9_c	0.340	0.100	0.472	-0.070	-0.542
34	phyo13_c	0.046	-0.248	-0.179	-0.464	-0.285
35	phyt13a_c	0.510	0.316	-0.526	-0.536	-0.010
36	phyt13b_c	-	-	-	-	-
37	phyt13c_c	-	-	-	-	-
38	phyt13d_c	0.832	0.090	0.173	-0.059	-0.232
39	phyf9_c	0.006	0.046	-0.305	0.017	0.322
40	phyf6_c	0.000	0.192	0.095	0.079	-0.016
41	phyg13_c	-	-	-	-	-
42	phyn8_c	-0.328	0.128	0.077	0.322	0.245
43	phyn14_c	-0.670	-0.358	-0.457	-0.266	0.191
44	phyt4a_c	-0.246	-0.124	0.025	-0.139	-0.164
45	phyt4b_c	0.242	0.092	-0.185	-0.223	-0.038
46	phyt4c_c	0.492	0.146	-0.409	-0.320	0.089

		Gender	Wave	Books		
	Item	male vs female	1 vs 2	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
47	phyf7_c	0.628	0.152	-0.458	-0.043	0.415
48	phyb18_c	-	-	-	-	-
49	phyn3_c	-0.760	-0.176	-0.471	0.039	0.510
50	phyn2_c	-	-	-	-	-
51	phyg5_c	0.366	0.286	0.380	0.250	-0.130
52	phyt9_c	-	-	-	-	-
53	phyh3_c	-0.058	-0.140	-0.150	-0.213	-0.063
54	phyb24_c	0.058	0.234	-0.132	-0.222	-0.090
55	phyg19_c	-0.342	-0.316	-0.037	0.127	0.164
56	phyf13_c	-0.322	0.284	-0.248	-0.220	0.028
57	phyn11_c	-0.402	0.204	0.266	0.330	0.064
58	phyf4_c	-0.300	-0.214	0.195	-0.126	-0.321
59	phyh15_c	0.682	-0.192	0.301	0.218	-0.083
60	phyn12t_c	-1.116	-0.266	0.406	0.808	0.402
61	phyh5t_c	-0.742	0.208	0.274	0.692	0.418
62	phyh6t_c	-0.148	-0.474	0.093	0.369	0.276

		Gender	Wave	Books		
	Item	male vs female	1 vs 2	0-200 vs 201-500	0-200 vs > 500	201-500 vs > 500
63	phyn9t_c	-1.126	0.336	0.163	-0.568	-0.731
64	phyn2t_c	-1.000	-0.446	0.795	1.287	0.492
	main effect	-0.594	0.020	0.123	0.327	0.204

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, but not gender. 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	59	41,134.55	41,214.34
	DIF	116	40,941.45	41,098.32
Wave	main effect	59	41,401.68	41,481.47
	DIF	116	41,461.86	41,618.74
Books	main effect	60	34,395.33	34,476.47
	DIF	174	34,487.20	34,722.51

5.3.3 Rasch homogeneity

One essential assumption of the Rasch (1960) model is Rasch homogeneity. Rasch homogeneity implies that all item-discrimination parameters are equal. In order to test this assumption, a Birnbaum model (2PL; Birnbaum, 1968) was specified. In this model, discrimination parameters are freely estimated and not fixed to 1. The estimated discriminations differed across the items (see Table 5), ranging from 0.006 (item phyt13a_c) to 3.714 (item phyn2t_c). Despite the empirical preference for the 2PL (AIC = 41021.54, BIC = 41679.24, number of parameters = 115) model, the Rasch model (AIC = 41399.92, BIC = 41679.24, number of parameters = 58) 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.

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 Thuringia. 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 acceptable to good fit and some items that were measurement invariant across these subgroups (but see Table 6). 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 64 items, all of which are scored as dichotomous variables with 0 indicating an incorrect response and 1 indicating a correct response. MC items are marked with a '_c' at the end of the variable name. Appendix A provides the syntax that was used to generate the person estimates with the ConQuest 4.2 software (Wu, Adams, Wilson, & Haldane, 2007). Appendix B provides an alternative syntax for use with the TAM package (Robitzsch, Kiefer, & Wu, 2017) in the software R (R Core Team, 2017).

Manifest physics competence scores are provided in the form of WLEs (phy_sc1) along with their corresponding standard errors (phy_sc2). As described in Section 5, these person estimates were derived from the joint scaling of all two 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; except items with negative discriminations in the 2PL were excluded (items phye6_c, phyt13b_c, phyt13c_c, phyg13_c, phyb18_c, phyn2_c, and phyt9_c were excluded). Therefore, the delivered WLE is based on 57 items. 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 or 1 logits, we emphasize that (latent variable) 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.

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physics Review Special Topics-Physics Education Research*, *2,* 010105.

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

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

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

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical report – Scaling the Data of the Competence Tests* (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

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

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

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

Robitzsch, A., Kiefer, T., & Wu, M. (2017). *TAM: Test analysis modules. R package version 2.7-56*. Retrieved from https://CRAN.R-project.org/package=TAM

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. *Leibniz Insitute for Science and Mathematics Education*.

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

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

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

This paper uses data from the National Educational Panel Study (NEPS): Additional Study Thuringia, doi:10.5157/NEPS:TH:2.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey 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

```
title XXX;


datafile filename.dat;

format pid 1-7 responses 11-67;

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

```
setwd(“Your/Working/Directory”)
data <- # data read
items <- # column positions of the 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 C: Item Parameters based on all Items

Table S1

Item Parameters of the Physics Competence Test (all Items)

	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	16.22	1.769	0.061	0.97	-0.9	0.37	1.109
2	phyg1_c	35.31	0.656	0.049	1.05	3.0	0.25	0.316
3	phyn5_c	42.96	0.306	0.047	0.94	-4.7	0.48	1.644
4	phyr1_c	88.20	-2.154	0.068	0.97	-0.6	0.30	1.470
5	phyg2_c	60.93	-0.484	0.047	0.98	-1.6	0.42	1.134
6	phye2_c	60.47	-0.467	0.048	0.99	-0.5	0.36	0.829
7	phyh8_c	12.97	2.054	0.139	0.98	-0.2	0.37	1.100
8	phyn1_c	28.80	0.956	0.106	1.05	1.0	0.30	0.480
9	phyg8_c	24.40	1.215	0.111	0.91	-1.6	0.54	1.898
10	phym14_c	82.18	-1.701	0.123	1.01	0.2	0.28	0.610
11	phyt1_c	39.45	0.435	0.098	0.97	-1.1	0.45	1.133
12	phyg6_c	64.24	-0.685	0.101	0.98	-0.6	0.38	0.826

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
13	phyh12_c	16.49	1.726	0.131	0.97	-0.3	0.40	1.313
14	phyn12_c	26.90	1.060	0.109	0.97	-0.7	0.42	1.114
15	phyh2_c	38.78	0.463	0.100	0.99	-0.2	0.39	0.810
16	phyh5_c	39.83	0.412	0.101	1.00	0.1	0.38	0.792
17	phyn7_c	39.63	0.429	0.100	0.96	-1.1	0.46	1.083
18	phyf3_c	34.46	0.672	0.104	1.04	1.0	0.30	0.467
19	phyb6_c	16.60	1.753	0.128	0.99	-0.1	0.35	0.932
20	phyg4_c	33.06	0.781	0.103	1.08	2.0	0.19	0.039
21	phyn4_c	11.44	2.210	0.150	1.01	0.1	0.20	0.425
22	phyn10_c	28.89	0.994	0.106	1.00	0.0	0.36	0.894
23	phyf5_c	46.57	0.163	0.097	1.00	-0.1	0.39	0.812
24	phyn13_c	35.60	0.658	0.101	1.07	1.9	0.21	0.072
25	phyb14_c	9.62	2.398	0.165	1.02	0.2	0.09	0.130
26	phyh6_c	35.79	0.667	0.102	1.06	1.7	0.19	0.118
27	phyn6_c	37.80	0.573	0.103	1.01	0.3	0.33	0.699
28	phyn15_c	24.17	1.256	0.112	1.03	0.5	0.28	0.402
29	phyt3_c	31.55	0.867	0.104	1.01	0.3	0.29	0.363

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
30	phyf1_c	51.18	-0.013	0.099	1.05	1.8	0.27	0.117
31	phye6_c	27.06	1.067	0.109	1.06	1.3	0.15	-0.018
32	phye1_c	79.84	-1.457	0.118	1.02	0.4	0.24	0.514
33	phyn9_c	50.67	-0.028	0.101	1.05	2.1	0.25	0.216
34	phyo13_c	77.82	-1.329	0.113	1.03	0.5	0.24	0.344
35	phyt13a_c	81.76	-1.600	0.127	1.03	0.4	0.23	0.006
36	phyt13b_c	60.70	-0.465	0.102	1.03	1.0	0.31	-0.219
37	phyt13c_c	66.30	-0.729	0.106	1.08	2.1	0.17	-0.533
38	phyt13d_c	40.79	0.388	0.102	1.07	2.4	0.19	0.782
39	phyf9_c	19.69	1.492	0.124	0.99	-0.1	0.30	0.902
40	phyf6_c	18.11	1.606	0.125	0.97	-0.4	0.34	0.914
41	phyg13_c	60.04	-0.441	0.098	1.09	3.1	0.13	-0.256
42	phyn8_c	21.58	1.378	0.116	0.93	-1.2	0.46	1.714
43	phyn14_c	33.75	0.724	0.102	0.97	-0.9	0.40	1.098
44	phyt4a_c	74.95	-1.171	0.112	1.00	0.1	0.31	0.604
45	phyt4b_c	62.96	-0.569	0.102	1.04	1.3	0.26	0.587
46	phyt4c_c	22.43	1.334	0.116	1.04	0.7	0.19	0.263

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
47	phyf7_c	36.31	0.600	0.102	1.04	1.2	0.26	0.424
48	phyb18_c	31.89	0.808	0.107	1.11	2.7	0.09	-0.358
49	phyn3_c	56.09	-0.272	0.099	0.96	-1.6	0.47	1.277
50	phyn2_c	28.28	0.993	0.112	1.09	1.7	0.16	-0.024
51	phyg5_c	33.33	0.743	0.104	0.98	-0.6	0.40	0.886
52	phyt9_c	44.75	0.220	0.100	1.10	3.7	0.16	-0.092
53	phyh3_c	36.42	0.596	0.103	0.98	-0.5	0.39	0.983
54	phyb24_c	14.98	1.876	0.135	1.02	0.3	0.28	0.702
55	phyg19_c	47.54	0.103	0.098	0.97	-1.0	0.45	1.022
56	phyf13_c	50.42	-0.029	0.099	0.94	-2.3	0.51	1.565
57	phyn11_c	34.23	0.700	0.103	0.96	-1.1	0.47	1.309
58	phyf4_c	24.47	1.220	0.114	0.97	-0.5	0.45	1.373
59	phyh15_c	30.93	0.854	0.110	1.08	1.7	0.23	0.206
60	phyn12t_c	14.69	1.939	0.135	0.94	-0.7	0.49	1.764
61	phyh5t_c	20.40	1.553	0.142	0.88	-1.6	0.61	2.632
62	phyh6t_c	27.69	1.231	0.155	0.95	-0.7	0.49	1.388
63	phyn9t_c	10.43	2.563	0.236	0.99	0.0	0.45	1.668

	Item	Percentage correct	Difficulty/ location parameter	*SE* (difficulty/ location parameter)	WMNSQ	WMNSQ t-value	Correlation of item score with total score	Discrimination-2 PL
64	phyn2t_c	12.77	2.245	0.231	0.90	-0.6	0.62	3.714

Appendix D: Content Area for each Item

Table S2.
Content Area for each Items

	Item	Content Area		Item	Content Area
1	phyh10_c	Electrical fields and interdependency	33	phyn9_c	Optics
2	phyg1_c	Magnetic fields and electromagnetic induction	34	phyo13_c	Dynamics: Mechanics of the Rigid Body
3	phyn5_c	Waves	35	phyt13a_c	Thermodynamics
4	phyr1_c	Optics	36	phyt13b_c	Thermodynamics
5	phyg2_c	Thermodynamics	37	phyt13c_c	Thermodynamics
6	phye2_c	Thermodynamics	38	phyt13d_c	Thermodynamics
7	phyh8_c	Electrical fields and interdependency	39	phyf9_c	Thermodynamics
8	phyn1_c	Magnetic fields and electromagnetic induction	40	phyf6_c	Magnetic fields and electromagnetic induction
9	phyg8_c	Waves	41	phyg13_c	Waves
10	phym14_c	Optics	42	phyn8_c	Optics
11	phyt1_c	Thermodynamics	43	phyn14_c	Dynamics: Mechanics of the Rigid Body
12	phyg6_c	Thermodynamics	44	phyt4a_c	Thermodynamics
13	phyh12_c	Waves	45	phyt4b_c	Thermodynamics
14	phyn12_c	Dynamics: Mechanics of the Rigid Body	46	phyt4c_c	Thermodynamics
15	phyh2_c	Thermodynamics	47	phyf7_c	Quantum physics: Quanta and matter
16	phyh5_c	Special Theory of Relativity	48	phyb18_c	Electrical fields and interdependency
17	phyn7_c	Waves	49	phyn3_c	Waves
18	phyf3_c	Quantum physics: Quanta and matter	50	phyn2_c	Magnetic fields and electromagnetic induction
19	phyb6_c	Electrical fields and interdependency	51	phyg5_c	Optics

Table S2.
Content Area for each Items

	Item	**Content Area**		**Item**	**Content Area**
	Item	**Content Area**		**Item**	**Content Area**
20	phyg4_c	Magnetic fields and electromagnetic induction	52	phyt9_c	Thermodynamics
21	phyn4_c	Magnetic fields and electromagnetic induction	53	phyh3_c	Quantum physics: Quanta and matter
22	phyn10_c	Optics	54	phyb24_c	Magnetic fields and electromagnetic induction
23	phyf5_c	Thermodynamics	55	phyg19_c	Magnetic fields and electromagnetic induction
24	phyn13_c	Dynamics: Mechanics of the Rigid Body	56	phyf13_c	Waves
25	phyb14_c	Electrical fields and interdependency	57	phyn11_c	Optics
26	phyh6_c	Magnetic fields and electromagnetic induction	58	phyf4_c	Dynamics: Mechanics of the Rigid Body
27	phyn6_c	Magnetic fields and electromagnetic induction	59	phyh15_c	Quantum physics: Quanta and matter
28	phyn15_c	Dynamics: Mechanics of the Rigid Body	60	phyn12t_c	Dynamics: Mechanics of the Rigid Body
29	phyt3_c	Thermodynamics	61	phyh5t_c	Special Theory of Relativity
30	phyf1_c	Quantum physics: Quanta and matter	62	phyh6t_c	Magnetic fields and electromagnetic induction
31	phye6_c	Electrical fields and interdependency	63	phyn9t_c	Optics
32	phye1_c	Waves	64	phyn2t_c	Magnetic fields and electromagnetic induction