NEPS WORKING PAPERS

Christoph Duchhardt

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

NEPS Working Paper No. 59
Bamberg, July 2015

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 Mathematics:

Scaling Results for the

Additional Study Baden-Wuerttemberg

Christoph Duchhardt, IPN–Leibniz Institute for Science and Mathematics Education at Kiel University / University of Bremen

E-mail address of lead author:

Christoph.Duchhardt@uni-bremen.de

Bibliographic data:

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.

I would like to thank Steffi Pohl, Kerstin Haberkorn, Katinka Hardt, Elena Wiegand, Ingrid Koller, and Theresa Rohm for developing and providing standards for the technical reports.

I also thank Natalie Boonyaprasop for proofreading the English text.

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

Abstract

The National Educational Panel Study (NEPS) aims to investigate the development of competencies across the whole life span. It also develops tests to assess different competence domains. In order to evaluate the quality of these competence tests, a wide range of analyses are carried out by using item response theory (IRT). This paper describes the data and results of analyzing the mathematics competence test that was used in the additional study Baden-Wuerttemberg. The test was designed to test first-year students in higher education; here, three consecutive waves (2011–2013) of secondary-school students were tested in their final year of Gymnasium (type of school leading to upper secondary education and Abitur). In sum, 4,915 students participated in these three waves. The mathematics test consisted of 21 items representing different content areas as well as different cognitive components. A partial credit model was used for scaling the data. Item fit statistics and differential item functioning were investigated. The results show that the items exhibit good item fit and measurement invariance across various groups. However, the reliability is somewhat mediocre, which might be due to the fact that test targeting is not perfect. The paper also provides 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, mathematical competence, Scientific Use File

1. Introduction

Within the National Educational Panel Study (NEPS) different competencies are measured coherently across the life span, and tests have been developed for different competence domains. These include, among other things, reading competence, mathematical competence, and scientific literacy, as well as information and communication technologies (ICT) literacy. Weinert et al. (2011) give an overview of the competencies measured in NEPS.

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

This paper presents the results of the mathematics competence test in the three waves of the additional study Baden-Wuerttemberg. In this study, the mathematics test that was constructed to be implemented in NEPS Starting Cohort 5–First-Year Students in Higher Education–was used over three consecutive years (2011 through 2013) to test secondary-school students' mathematical competencies in their final year of Gymnasium (type of school leading to upper secondary education and Abitur). More detailed information about the aims of this study can be found on the NEPS website.[1]

The present report is heavily modeled on previous technical reports such as Pohl, Haberkorn, Hardt, and Wiegand (2012); Hardt, Pohl, Haberkorn, and Wiegand (2013); Jordan and Duchhardt (2013); and Koller, Haberkorn, and Rohm (2014). It includes extracts from these previous reports.

2. Testing Mathematical Competence

The framework and test development for the test of mathematical competence are described in Weinert et al. (2011), Neumann et al. (2012), and Ehmke et al. (2009). In the following, specific aspects of the mathematics test that are necessary for understanding the scaling results presented in this paper are briefly described.

The items are not arranged in units. Thus, in the test, students usually face a certain situation followed by only one task related to it; sometimes there are two tasks. Each of the items belongs to one of the following content areas:

- quantity,
- space and shape,
- change and relationships,
- data and chance.

The framework also describes as a second, independent dimension six cognitive components required for solving the tasks. These are distributed across the items. In the mathematics test, there are three types of response formats. These are simple multiple choice (MC),

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

complex multiple choice (CMC), and short constructed response (SCR). In MC items, the test taker has to find the correct answer from several—usually four—response options. In CMC tasks, a number of subtasks with two response options are presented. SCR items require the test taker to write down an answer into an empty field.

Tables 1 and 2 show how content areas and response formats are distributed among the items.[2]

Table 1

Content Areas of Items in the Mathematics Test

Content area	**Frequency**
Quantity	4
Space and shape	4
Change and relationships	6
Data and chance	6
Total number of items	20

Table 2

Response Formats of Items in the Mathematics Test

Response format	**Frequency**
Single multiple choice	16
Complex multiple choice	1
Short constructed response	3
Total number of items	20

3. Data and Sample Size

A description of the design of the study, the sample, as well as the instruments used can be found on the NEPS website.[3] In total, 4,915 subjects took the mathematics test: 1,282 in

[2] One item is not presented here, as it was excluded from further analyses, cf. 5.1.

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

2011 (Wave 1), 2,422 in 2012 (Wave 2), and 1,211 in 2013 (Wave 3). All subjects gave at least three valid answers, so that for every subject one competence score was estimated.

4. Analyses

This section briefly describes the analyses that were conducted: inspection of various missing responses, scaling of data, and examining the quality of the test.

4.1 Missing Responses

There are different types of missing responses in competence test data. These are missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach, d) items that have not been administered, and e) different kinds of missing responses within a CMC item that make the missing indeterminable. We thoroughly inspected the occurrence of missing responses in the test. First, we looked at the occurrence of the different types of missing responses per person. This gave an indication of how well the test persons were coping with the test. We then examined the occurrence of missing responses per item in order to obtain some information on how well the items performed.

4.2 Scaling Model

In order to estimate item and person parameters for mathematical competence, a partial credit model (Masters, 1982) was used and estimated in ConQuest (Wu, Adams, & Wilson, 1997), which uses a marginal maximum likelihood approach. A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

The CMC item consisted of four subtasks that were aggregated to a polytomous variable, indicating the number of correctly solved subtasks. The two lowest categories were collapsed in order to avoid possible estimation problems (see also Pohl & Carstensen, 2012, for an explanation of this approach).

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

Plotting the item parameters to the ability estimates of the persons was required in order to judge how well the item difficulties were targeted toward the test persons' abilities. The test targeting gives some information about the precision of the ability estimates at different levels of ability.

4.3 Checking the Quality of the Test

The mathematics competence test was constructed to be implemented in NEPS Starting Cohort 5–First-Year Students in Higher Education. To ensure that the test featured appropriate psychometric properties also in the sample of secondary-school students, the quality of the test was examined again by several analyses.

Before aggregating the responses of the CMS item to a polytomous variable, the aggregation was justified by a preliminary analysis. For this purpose, the four subtasks—together with the other items—were analyzed using a Rasch model (Rasch, 1960). The fit of the subtasks was evaluated based on the weighted mean square error (WMNSQ), its *t*-value, and the point biserial correlations of the correct responses with the total score. Only if the subtasks exhibited a satisfactory item fit were they aggregated to the polytomous variable included in the final scaling model.

Afterwards, the item fit of dichotomous and polytomous items was examined by analyzing them via a partial credit model. The WMNSQ, the respective *t*-value, correlations of the item score with the total score, and the item characteristic curve were evaluated for each item. Items with a WMNSQ > 1.15 (*t*-value > |6|) were considered as having a noticeable item misfit and items with a WMNSQ > 1.2 (*t*-value > |8|) were judged as having a considerable item misfit, and their performance was further investigated. Correlations of the item score with the total score (equal to the discrimination as computed in ConQuest) greater than 0.3 were considered as good, greater than 0.2 as acceptable, and below 0.2 as problematic. Overall, judgment of item fit was based on all fit indicators.

Our aim was to construct a mathematics competence test that measured the same construct in all participants. If any items favored a certain subgroup (e.g., easier items 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. We addressed the issue of measurement invariance by investigating test fairness for the variables gender, migration 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 & Carstensen, 2012, for a description of these variables. Differential item functioning was estimated using a multigroup IRT model, in which the main effects of the subgroups as well as 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 experiences with preliminary data we judged absolute differences in estimated difficulties that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as noteworthy for further investigation, differences between 0.4 and 0.6 as considerable but not significant, and differences smaller than 0.4 as no considerable DIF. In addition to DIF analyses at item level, test fairness was investigated by comparing a model including differential item functioning to a model that only estimated main effects and no DIF.

The mathematics competence data were scaled using the partial credit model (1PL), which assumes Rasch-homogeneity. The partial credit model was chosen because it preserves the weighting of the different aspects of the framework as intended by test developers (Pohl & Carstensen, 2012). 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 generalized partial credit model (2PL) (Muraki, 1992) using the software mdltm (von Davier, 2005).

5. Results

In this section, the key scaling results of the three waves of the additional study Baden-Wuerttemberg will be presented. Some results of scaling each wave separately can be found in Appendices C1–C3.

5.1 Preliminary Analyses

As described in Subsection 4.3, we first found that aggregating the four subtasks of the CMC item was justified.

During the first cycle of main analyses, one item was found to exhibit too large DIF (with respect to gender) and so had to be excluded. In the following, the results of the final analyses are presented—that is, without that one problematic item.

5.2 Missing Responses

In this subsection, we first report the number of missing responses of the different types as described in 4.1 per person and the total number of missing responses per person. Then, missing responses per item will be described. Note that there was only one version of the mathematics test; hence, no missing by design could occur.

5.2.1 Missing responses per person

Figure 1 shows the number of *invalid responses* per person. As can be seen, almost none of the participants—only 1.6%—produced any invalid responses. The maximum number of invalid responses was 3.

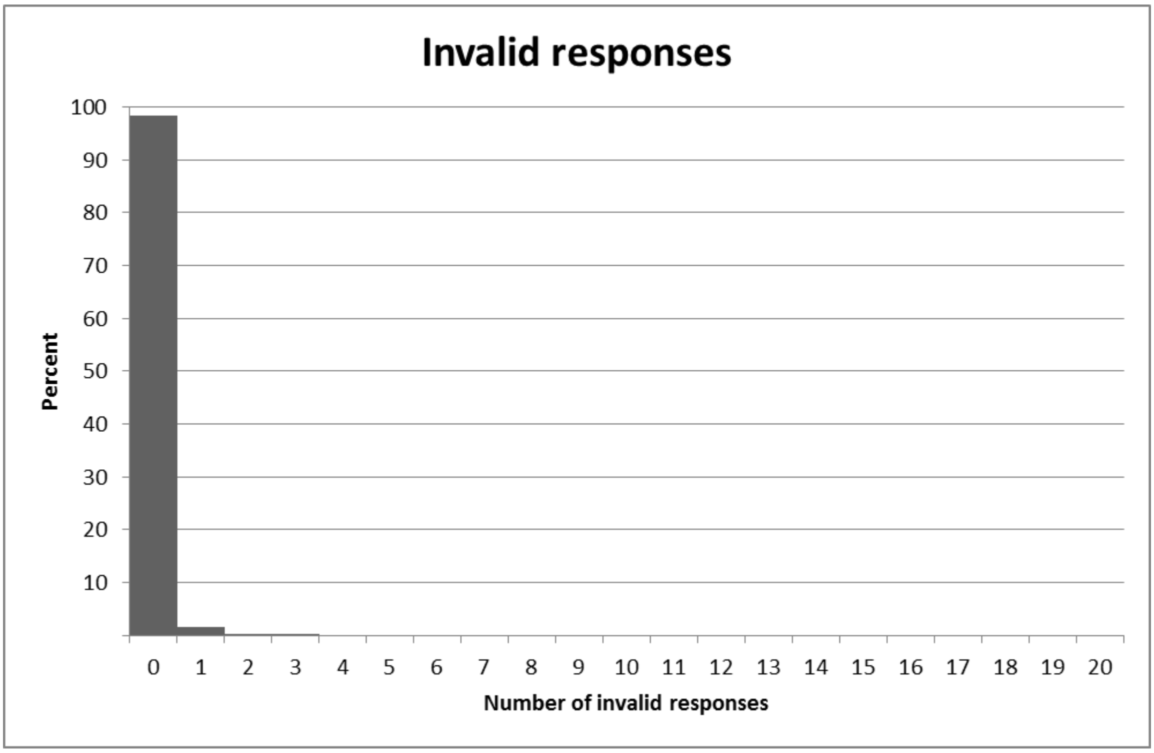

Figure 1. Number of invalid responses per person.

The largest source of missing responses by comparison in this test was the *omission of items*. As can be seen in Figure 2, more than half of the participants (52.9%) skipped at least one item. 5.4% of the participants omitted five or more items.

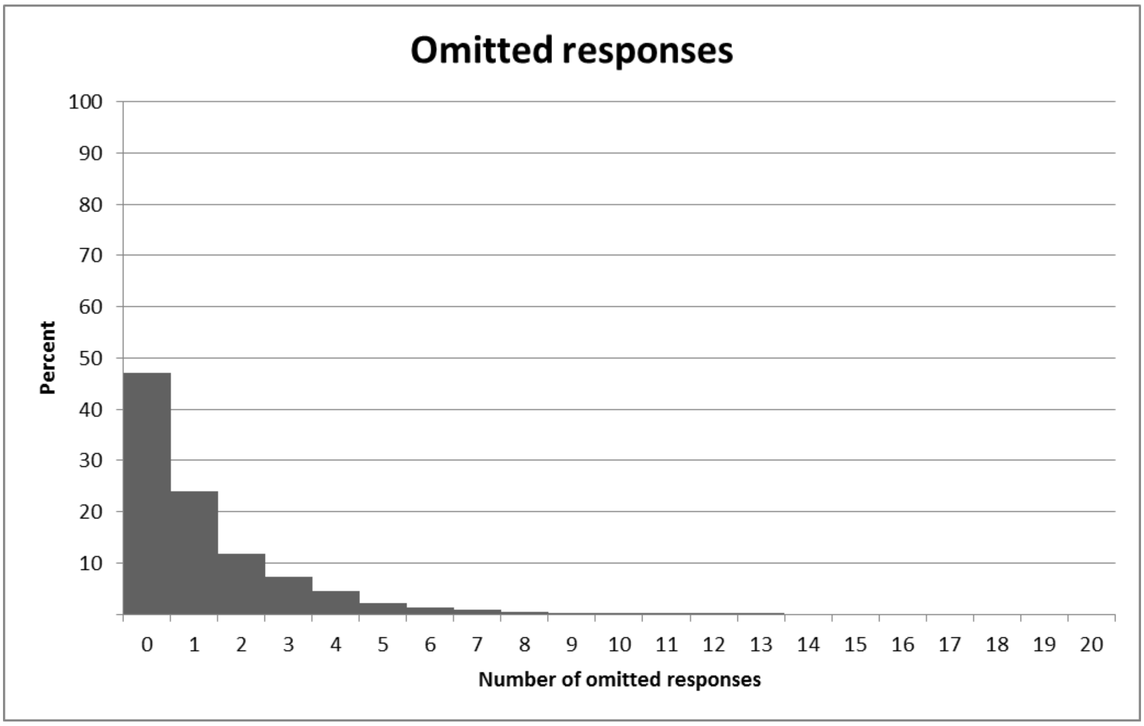

Figure 2. Number of omitted responses per person.

By definition, every item after the last item that was not omitted is labeled *not reached*. As Figure 3 shows, most participants (78.5%) reached the end of the test. Only 4% did not reach the fifth last item.

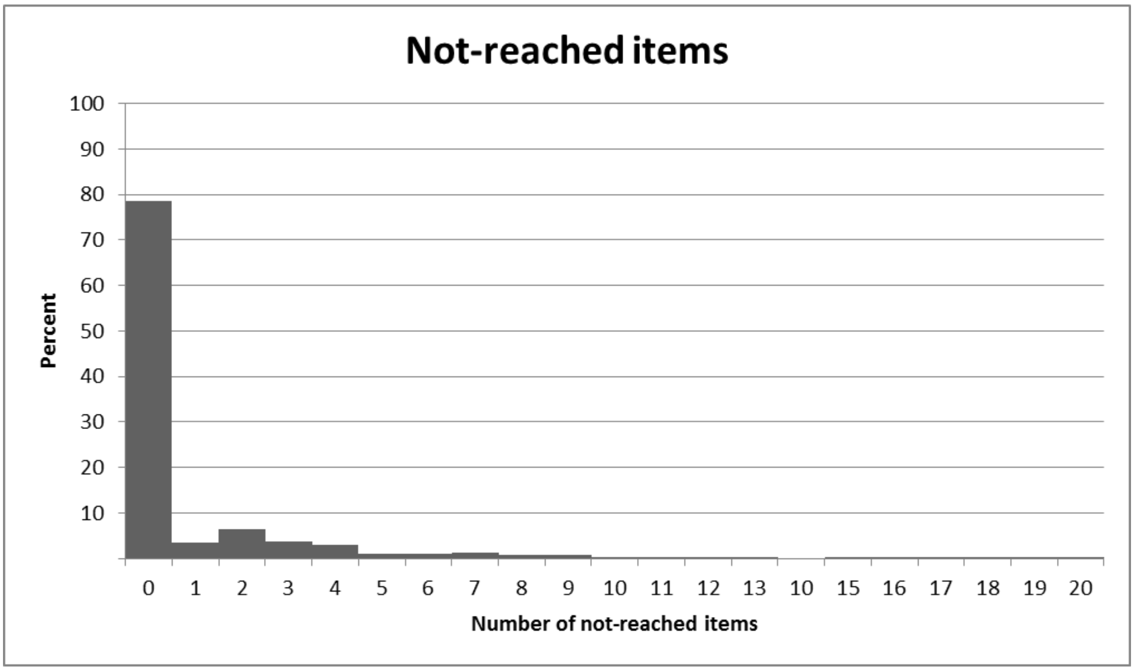

Figure 3. Number of not-reached items per person.

The CMC item was the only source of indeterminable missing responses. This missing occurred nine times only.

The total number of missing responses aggregated over invalid, omitted, not-reached, and indeterminable missing responses per person is illustrated in Figure 4. On average, the participants produced 1.98 (*SD* = 2.64) missing responses. Moreover, 41.1% of the persons had no missing response at all, and 15.7% of the participants gave five or more missing responses.

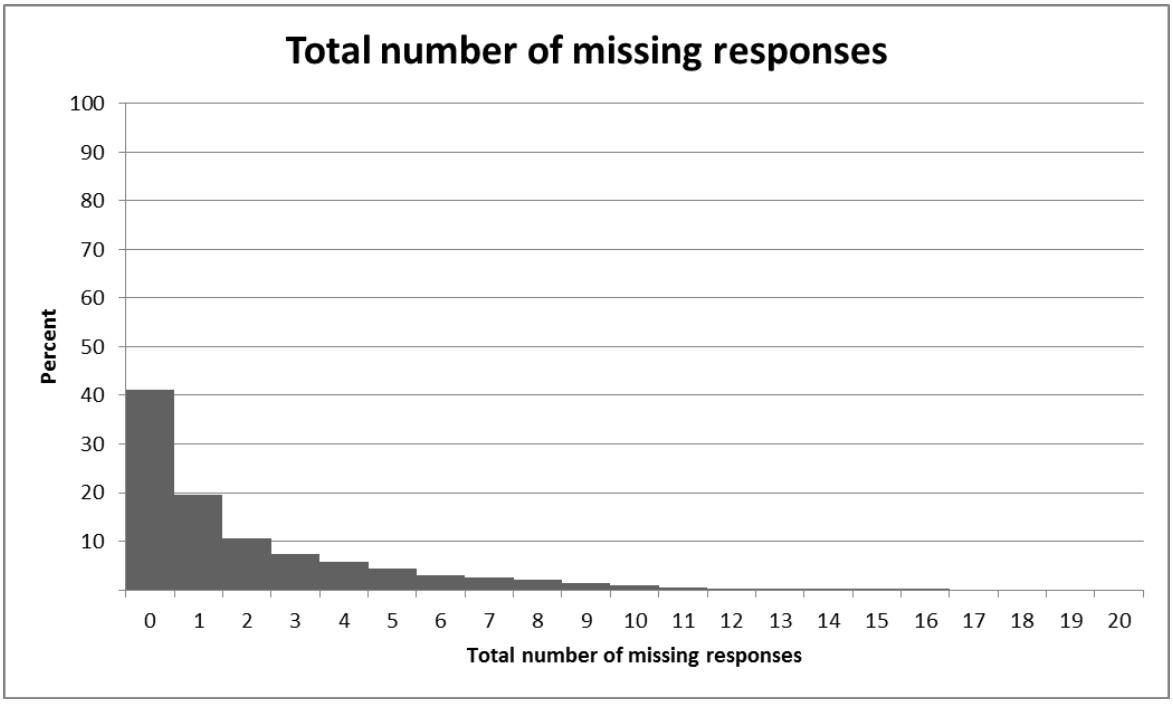

Figure 4. Number of omitted responses per person.

In sum, there is a very small amount of invalid, not-reached, and not-determinable missing responses and a reasonable amount of omitted items. No participant produced so many missing responses that they had to be excluded them from further analyses.

5.2.2 Missing responses per item

Table 3 provides information on the occurrence of different kinds of missing responses per item. The amount of persons failing to reach items rose successively—with increasing item position in the test—up to an amount of 21.5% (column 4). However, with about 95% of the participants reaching Item 17, the test can hardly be described as too long.

Omitting items was not an unusual occurrence in the test—9 out of the 20 items had omission rates exceeding 5% (column 5). Particularly noticeable are the items mag9r061_c (omitted by 11.8% of the participants), mas2v062_c (12.8%), and mas2v042_c (35.7%). Both mag9r061_c and mas2v042_c are short constructed response items—a format that might make it more appealing to skip these items.

Overall, the percentage of invalid responses per item (column 6) was very low (maximum of 0.8% for item mas2v042_c).

5.3 Parameter Estimates

5.3.1 Item parameters

The second column in Table 4 shows the percentage of correct responses relative to all valid responses for each item. Please note that, because there is a nonnegligible amount of missing responses, this probability cannot be interpreted as an index of item difficulty. The percentage of correct responses within MC or SCR items varied between 22.3% and 86.7% with an average of 60.3% (*SD* = 17.6%) correct responses.

For reasons of model identification, in the partial credit model the mean of the ability distribution was constrained to be zero. The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in the third column of Table 4. The step parameters for the polytomous variable are depicted in Table 5. The item difficulties ranged from -2.183 (item maa2d131_c) to 1.471 (item mas2r092_c) logits with an average difficulty of -0.607 logits (*SD* = 0.963). Altogether, the item difficulties are somewhat low. Owing to the large sample size, the corresponding standard errors of the estimated item difficulties (Column 4) are small ($SE(ß) \leq 0.044$).

Table 3

Item Parameters of the Mathematics Test

Item	**Position in the test**	**Number of valid responses**	**Percentage of not-reached responses**	**Percentage of omitted responses**	**Percentage of invalid responses**
maa2q071_c	1	4,847	–	1.4	–
mas2r092_c	2	4,892	–	0.5	–
mas2v093_c	3	4,866	–	1.0	0.0
mas2v032_c	5	4,650	–	5.4	–
maa2d131_c	6	4,881	0.0	0.7	–
maa2d132_c	7	4,768	0.0	3.0	–
mas2v062_c	8	4,280	0.0	12.8	0.1
mas2v063_c	9	4,719	0.1	3.9	0.1
maa2r081_c	10	4,630	0.1	5.7	–
maa2v082_c	11	4,482	0.2	8.6	–
mas2q041_c	12	4,598	0.4	6.0	0.0
mas2v042_c	13	3,071	1.0	35.7	0.8
mas2q02s_c	14	4,452	1.7	7.4	0.1

maa2d111_c	15	4,570	2.9	3.9	0.2
maa2d112_c	16	4,327	4.0	8.0	–
maa2r011_c	17	4,588	5.1	1.6	0.0
mas2q011_c	18	4,366	7.9	3.1	0.2
mag9r061_c	19	3,746	11.7	11.8	0.3
mas2d071_c	20	3,985	18.1	0.8	–
mas2d072_c	21	3,857	21.5	–	–

Note. The item in position 4 was excluded from the analyses due to inacceptable differential item functioning across the two gender groups (see 5.1).

Table 4

Item Parameters of the Mathematics Test

Item	Percentage correct	Difficulty / location parameter	*SE* (difficulty / location parameter)	WMNSQ	*t*-value of WMNSQ	Correlation of item score with total score	Discrimi-nation–2 PL
maa2q071_c	77.24	-1.443	0.037	0.99	-0.3	0.43	1.10
mas2r092_c	22.38	1.471	0.037	0.99	-0.5	0.41	0.88
mas2v093_c	73.33	-1.199	0.035	1.00	-0.2	0.44	1.04
mas2v032_c	72.43	-1.151	0.035	1.05	2.7	0.38	0.76
maa2d131_c	86.70	-2.183	0.044	0.99	-0.5	0.37	1.20
maa2d132_c	65.50	-0.765	0.033	0.93	-5.1	0.54	1.46
mas2v062_c	37.66	0.605	0.035	1.05	3.3	0.39	0.71
mas2v063_c	42.59	0.352	0.032	0.98	-1.6	0.48	1.01
maa2r081_c	70.50	-1.034	0.035	0.97	-1.8	0.48	1.19
maa2v082_c	60.55	-0.518	0.033	1.02	1.5	0.45	0.93
mas2q041_c	63.70	-0.675	0.033	1.03	2.0	0.43	0.86
mas2v042_c	65.19	-0.601	0.041	0.92	-5.0	0.56	1.45

mas2q02s_c	n. a.	-2.014	0.037	0.98	-1.0	0.46	1.26
maa2d111_c	71.03	-1.061	0.035	1.03	1.9	0.41	0.91
maa2d112_c	40.65	0.456	0.034	1.01	0.5	0.45	0.87
maa2r011_c	74.46	-1.274	0.036	0.93	-3.8	0.51	1.55
mas2q011_c	79.18	-1.574	0.040	0.98	-0.8	0.43	1.22
mag9r061_c	61.05	-0.468	0.037	1.04	2.8	0.41	0.77
mas2d071_c	52.47	-0.149	0.035	1.08	6.1	0.38	0.62
mas2d072_c	28.23	1.084	0.039	1.13	6.8	0.26	0.36

Table 5

Step Parameters (and Standard Errors) of the Polytomous Item

Item	**Step 1 (*SE*)**	**Step 2 (*SE*)**	**Step 3 (*SE*)**
mas2q02s_c	-0.055 (0.033)	0.047 (0.036)	0.007

5.3.2 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, item difficulties and person abilities are plotted on the same scale. The items cover the lower and medium part of the ability distribution very well, but in general, they are somewhat too easy. Hence, the test can measure person abilities in the low- and medium-ability regions relatively precisely, whereas high person abilities are measured with larger standard errors of measurement.

The mean of the ability distribution was constrained to be zero, its variance was estimated to be 0.937 indicating a reasonable differentiation between the subjects. The reliability of the test (EAP/PV reliability = .736, WLE reliability = .699) was acceptable, but not good. This should be related to the suboptimal test targeting described above.

Scale (in logits)	**Person ability**	**Item difficulty**
3		
	X	
	X	
	X	
2	XX	
	XX	
	X	
	XX	
	XXX	2
	XXXX	
	XXXX	
	XXXXX	21
1	XXXXXX	
	XXXXX	
	XXXXXX	
	XXXXXXX	8
	XXXXXXXX	16
	XXXXXXXX	9
	XXXXXXXXXX	
	XXXXXXXX	
0	XXXXXXXXX	20
	XXXXXXXXX	
	XXXXXXXXX	19
	XXXXXXXXXX	11 13
	XXXXXXXXX	7 12
	XXXXXXXX	
	XXXXXXXX	
-1	XXXXX	
	XXXXX	5 10 15
	XXXXX	3 17
	XXXX	
	XX	1
	XXX	18
	XX	
	X	
-2	X	14
	X	6
	X	

Figure 5. Test targeting. The distribution of person abilities in the sample is depicted on the left-hand side, with each ‘X’ representing 27.9 cases. The item difficulties (or location parameters) are depicted on the right-hand side. Each number represents one item with a corresponding position in the test, cf. Table 3.

5.4 Test Quality

5.4.1 Item fit

Altogether, item fit can be considered as good, with values of the WMNSQ ranging from 0.92 (item mas2v042_c) to 1.13 (item mas2d072_c), cf. column 5 of Table 4. This latter item also has the largest absolute *t*-value of the WMNSQ (6.8). It might be considered to be slightly underfitting.

Point-biserial correlations between the item scores and the total scores ranged from 0.26 (item mas2d072_c) to 0.52 (item mas2v042_c) and had a mean of 0.43, cf. column 7 of Table 4.

Discriminations estimated in the 2PL-model ranged from 0.36 (item mas2d072_c) to 1.55 (item maa2r011_c), cf. Table 4, column 8.

5.4.2 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, migration 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.

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. Positive main effects indicate that the first group, for example, “male”, scores higher on average.

Gender: On average, male participants had a considerably higher mathematical competence (main effect = 0.668 logits, Cohen’s *d* = 0.757). Only one item (mag9r061_c) showed DIF greater than 0.4 logits.

Migration background: On average, participants without migration background had a higher mathematical competence (main effect = 0.272 logits, Cohen’s *d* = 0.284). No item showed DIF greater than 0.4 logits.

Wave: On average, participants of the three waves basically do not differ in their mathematical competence (main effects between 0.007 and 0.020 logits, Cohen’s *d* < 0.021). Only one item (maa2d111_c) showed DIF greater than 0.4 logits.

Books: On average, participants with many books at home performed better in the test on mathematical competence (highest main effect = 0,290 logits for the group with 500+ books compared to the group with up to 200 books, Cohen’s *d* = 0.302). No item showed DIF greater than 0.4 logits.

Table 6

Differential Item Functioning

Item	Gender	Migration background	Wave			Books		
	male vs female	without vs with	1 vs 2	1 vs 3	2 vs 3	0-200 vs 201-500	0-200 vs 501-	201-500 vs 501-
maa2q071_c	-0.096	-0.01	-0.008	-0.167	-0.159	-0.064	-0.104	-0.04
mas2r092_c	0.024	0.154	-0.155	-0.277	-0.122	-0.028	0.055	0.083
mas2v093_c	-0.052	-0.096	0.069	-0.021	-0.09	-0.172	-0.215	-0.043
mas2v032_c	0.264	0.028	0.208	0.398	0.19	-0.044	-0.037	0.007
maa2d131_c	0.1	0.02	-0.156	-0.151	0.005	0.141	0.248	0.107
maa2d132_c	-0.268	-0.116	0.154	0.05	-0.104	0.077	0.09	0.013
mas2v062_c	0.144	0.178	0.114	0.267	0.153	-0.004	0.136	0.14
mas2v063_c	0.01	0.054	-0.045	-0.096	-0.051	-0.131	0.113	0.244
maa2r081_c	-0.21	-0.074	0.012	0.004	-0.008	0.087	-0.048	-0.135
maa2v082_c	-0.166	-0.008	-0.153	-0.258	-0.105	-0.147	-0.105	0.042
mas2q041_c	-0.158	-0.038	-0.208	-0.23	-0.022	-0.053	0.023	0.076

mas2v042_c	0.096	-0.056	-0.071	0.206	0.277	0.03	-0.05	-0.08
mas2q02s_c	-0.21	-0.064	-0.051	0.123	0.174	0.118	0.032	-0.086
maa2d111_c	0.114	-0.24	0.223	0.419	0.196	0.233	0.214	-0.019
maa2d112_c	0.084	-0.022	0.023	-0.059	-0.082	0.054	-0.042	-0.096
maa2r011_c	-0.19	-0.172	-0.158	-0.538	-0.38	0.082	0.092	0.01
mas2q011_c	-0.204	0.026	0.034	0.074	0.04	-0.026	0.104	0.13
mag9r061_c	0.418	0.024	0.113	0.334	0.221	0.014	-0.149	-0.163
mas2d071_c	0.164	0.132	-0.03	-0.011	0.019	0.088	0	-0.088
mas2d072_c	0.244	0.314	-0.134	-0.199	-0.065	-0.112	-0.257	-0.145
main effect	0.668	0.272	0.020	0.007	-0.009	-0,154	-0,290	-0.136

In Table 7, the models with DIF are compared to those that include only the main effect of the respective variable. Regarding Akaike's (1974) information criterion (AIC), the more parsimonious models including only main effects are preferred for the variables migration background 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 as well 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	24	103,342.61	103,383.20
	DIF	44	103,255.86	103,330.28
Migration Background	main effect	24	101,737.48	101,777.86
	DIF	44	101,739.32	101,813.35
Wave	main effect	25	103,790.13	103,832,90
	DIF	65	103,735.95	103,839.80
Books	main effect	25	103,323.71	103,365.95
	DIF	65	103,338.54	103,448.35

6. Discussion

Descriptions and analyses presented in the previous sections have aimed to document the quality of the mathematics 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 with regard to various grouping variables. Item fit statistics provided evidence of well-fitting items that are measurement invariant across these subgroups. The test is reasonably reliable. However, because the test targets mainly low- and medium-performing participants, ability estimates for these participants will be very precise but less precise for high-performing participants.

Regarding main effects between subgroups, it seems noteworthy that the test shows practically no differences between the three waves of the study.

7. Data in the Scientific Use File

Data in the Scientific Use File contain 20 items, of which 19 items were scored as dichotomous variables (MC or SCR items) with 0 indicating an incorrect response and 1 indicating a correct response. One CMC item was scored as polytomous variable. MC items are marked with a '_c' at the end of the variable name, whereas the variable name of the CMC item ends in 's_c'. Note that the values of the polytomous variable in the Scientific Use File do not necessarily correspond to the number of correctly solved subtasks. This is due to the collapsing of categories (cf. Section 4.2 for a description of the aggregation of CMC and

MA items). In the IRT scaling model, the polytomous CMC variable was scored as 0.5 for each category. Appendix A provides the syntax that was used to generate person estimates using the ConQuest software (Wu, Adams, & Wilson, 1997). Appendix B provides an alternative syntax using the TAM package (Kiefer, Robitzsch, & Wu, 2015) of the R software (R Core Team, 2014).

Manifest mathematical competence scores are provided in the form of WLEs (magd_sc1) together with their corresponding standard errors (magd_sc2). As described in Section 5, these person estimates are from the joint scaling of all three waves of the study. For persons who did not take part in the mathematics test, no WLE was estimated. The value of the WLE and the respective standard error for these persons were denoted as not-determinable missing values.

We recommend using plausible values to investigate latent relationships of competence scores with 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.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft, 14.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Ehmke, T., Duchhardt, C., Geiser, H., Grüßing, M., Heinze, A., & Marschick, F. (2009). Kompetenzentwicklung über die Lebensspanne: Erhebung von mathematischer Kompetenz im Nationalen Bildungspanel. In A. Heinze, & M. Grüßing (Hrsg.), Mathematiklernen vom Kindergarten bis zum Studium: Kontinuität und Kohärenz als Herausforderung für den Mathematikunterricht. (S. 313-327). Münster: Waxmann 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.

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

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

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Neumann, I., Duchhardt, C., Ehmke, T., Grüßing, M., Heinze, A., & Knopp, E. (2013). Modeling and assessing of mathematical competence over the lifespan. *Journal for Educational Research Online, 5*(2), 80–102.

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 (2014). *R: A language and environment for statistical computing.* R Foundation for 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*(2), 461–464.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETSResearch Report No. RR-05-16). Princeton, NJ: ETS.

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

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

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ACER Conquest: Generalised item response modelling software.* Melbourne: ACER Press.

This paper uses data from the National Educational Panel Study (NEPS): Additional Study Baden-Wuerttemberg, doi:10.5157/NEPS:BW: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 in the additional study Baden-Wuerttemberg

```
title Additional Study Baden-Wuerttemberg, Mathematics, Waves 1-3, Partial credit model;


datafile filename.dat;
format pid 4-10 responses 12-31;
labels << labels.nam;


codes 0,1,2,3;
score (0,1,2,3) (0,0.5,1,1.5) !item(13);


model item + item*step;
set constraint=cases;


estimate;
show !estimates=latent >> filename.shw;
itanal >> filename.ita;
show cases !estimates=wle >> filename.wle;
```

Appendix B: TAM Syntax for generating WLE estimates in the additional study Baden-Wuerttemberg

```
setwd("Your/Working/Directory")

data <- # data read

items <- # positions of the math items ordered as in the SUF

library(TAM)

# First: Generate Design Matrix

Des <- designMatrices(modeltype = "PCM", resp = data[,items])

B <- Des$B

# Score the polytomous item according to NEPS conventions

# The item on position 13 is the only polytomous one

B[,4,][13] <- 1.5

B[,3,][13] <- 1

B[,2,][13] <- 0.5

B

# Compute PCM

PCM <- tam.mml(data[,items], B=B)

summary(PCM)

# Generate WLE estimates

PCM.wle <- tam.wle(PCM)

WLE <- PCM.wle$theta

WLE.SE <- PCM.wle$error
```

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

Table 7

Item Parameters of the Mathematics Test–Wave 1

Item	**Percentage correct**	**Difficulty / location parameter**	***SE*** **(difficulty / location parameter)**	**WMNSQ**	***t*-value of WMNSQ**	**Correlation of item score with total score**
maa2q071_c	77.94	-1.501	0.073	1	-0.1	0.42
mas2r092_c	24.78	1.329	0.070	0.97	-0.7	0.45
mas2v093_c	72.91	-1.181	0.068	0.98	-0.5	0.45
mas2v032_c	69.00	-0.963	0.068	1.01	0.2	0.45
maa2d131_c	87.81	-2.313	0.090	0.97	-0.4	0.38
maa2d132_c	63.78	-0.686	0.064	0.93	-2.8	0.55
mas2v062_c	35.48	0.727	0.069	1.09	3.2	0.34
mas2v063_c	43.79	0.302	0.063	0.99	-0.6	0.47
maa2r081_c	70.41	-1.038	0.068	0.95	-1.5	0.51
maa2v082_c	63.45	-0.666	0.066	1.02	0.8	0.46

mas2q041_c	66.91	-0.845	0.067	1.03	1.1	0.42
mas2v042_c	65.51	-0.587	0.083	0.88	-3.5	0.60
mas2q02s_c	n. a.	-2.055	0.073	0.98	-0.5	0.46
maa2d111_c	67.23	-0.86	0.068	1.03	0.9	0.43
maa2d112_c	40.69	0.449	0.067	1.01	0.5	0.44
maa2r011_c	77.84	-1.505	0.075	0.95	-1.2	0.47
mas2q011_c	78.57	-1.552	0.078	1	0	0.42
mag9r061_c	58.39	-0.336	0.072	1.06	2.2	0.41
mas2d071_c	52.91	-0.174	0.068	1.07	2.6	0.39
mas2d072_c	30.24	0.968	0.075	1.11	3.3	0.30

Table 8

Step Parameters (and Standard Errors) of the Polytomous Item–Wave 1

Item	**Step 1 (*SE*)**	**Step 2 (*SE*)**	**Step 3 (*SE*)**
mas2q02s_c	-0.133 (0.064)	0.119 (0.071)	0.014

Table 9

Differential Item Functioning–Wave 1

Item	Gender	Migration background	Books		
	male vs female	without vs with	0-200 vs 201-500	0-200 vs 501-	201-500 vs 501-
maa2q071_c	0.048	0	-0.139	-0.28	-0.141
mas2r092_c	0.028	0.214	0.033	0.124	0.091
mas2v093_c	-0.038	-0.066	-0.026	0.017	0.043
mas2v032_c	0.112	0.154	0.039	0.006	-0.033
maa2d131_c	0.082	-0.062	0.235	0.329	0.094
maa2d132_c	-0.252	-0.272	-0.135	-0.117	0.018
mas2v062_c	0.242	0.226	-0.156	0.045	0.201
mas2v063_c	0.126	-0.15	0.008	0.199	0.191
maa2r081_c	-0.092	-0.132	0.239	-0.158	-0.397
maa2v082_c	-0.374	0.016	-0.189	-0.01	0.179
mas2q041_c	-0.272	0.096	-0.172	-0.005	0.167

mas2v042_c	0.004	-0.252	-0.031	-0.164	-0.133
mas2q02s_c	-0.252	0.07	-0.012	0.114	0.126
maa2d111_c	-0.016	-0.03	0.148	0.182	0.034
maa2d112_c	0.124	-0.052	0.078	-0.045	-0.123
maa2r011_c	0.032	-0.152	-0.105	0.165	0.27
mas2q011_c	-0.234	0.076	0.172	-0.019	-0.191
mag9r061_c	0.26	-0.004	-0.201	-0.363	-0.162
mas2d071_c	0.226	0.304	0.202	0.103	-0.099
mas2d072_c	0.328	-0.046	0.017	-0.191	-0.208
main effect	0.792	0.298	-0.066	-0.160	-0.094

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

Table 10

Item Parameters of the Mathematics Test–Wave 2

Item	Percentage correct	Difficulty / location parameter	*SE* (difficulty / location parameter)	WMNSQ	*t*-value of WMNSQ	Correlation of item score with total score
maa2q071_c	77.77	-1.471	0.053	1	0.1	0.42
mas2r092_c	22.04	1.489	0.053	1.01	0.3	0.38
mas2v093_c	73.93	-1.23	0.050	0.99	-0.3	0.44
mas2v032_c	72.47	-1.152	0.050	1.05	2	0.38
maa2d131_c	86.24	-2.132	0.062	0.99	-0.3	0.36
maa2d132_c	66.68	-0.822	0.047	0.94	-3.1	0.53
mas2v062_c	37.28	0.623	0.049	1.03	1.7	0.40
mas2v063_c	42.37	0.358	0.046	0.96	-2.2	0.50
maa2r081_c	70.46	-1.03	0.050	0.99	-0.2	0.45
maa2v082_c	60.08	-0.495	0.048	1.01	0.6	0.45

mas2q041_c	62.57	-0.619	0.047	1.03	1.7	0.43
mas2v042_c	63.18	-0.503	0.059	0.91	-3.6	0.56
mas2q02s_c	n. a.	-1.943	0.052	0.98	-0.6	0.47
maa2d111_c	71.11	-1.064	0.050	1.03	1.3	0.40
maa2d112_c	41.12	0.437	0.048	1	0.2	0.45
maa2r011_c	75.36	-1.325	0.052	0.94	-2.2	0.50
mas2q011_c	79.09	-1.563	0.057	0.98	-0.6	0.43
mag9r061_c	60.45	-0.434	0.052	1.03	1.5	0.43
mas2d071_c	52.06	-0.128	0.050	1.08	4.3	0.38
mas2d072_c	27.77	1.109	0.056	1.13	4.7	0.25

Table 11

Step Parameters (and Standard Errors) of the Polytomous Item–Wave 2

Item	**Step 1 (*SE*)**	**Step 2 (*SE*)**	**Step 3 (*SE*)**
mas2q02s_c	-0.048 (0.046)	0.060 (0.052)	-0.012

Table 12

Differential Item Functioning–Wave 2

Item	Gender	Migration background	Books		
	male vs female	without vs with	0-200 vs 201-500	0-200 vs 501-	201-500 vs 501-
maa2q071_c	-0.178	0.032	-0.064	-0.164	-0.1
mas2r092_c	0.308	0.312	-0.095	-0.019	0.076
mas2v093_c	0	-0.19	-0.238	-0.23	0.008
mas2v032_c	-0.278	-0.044	-0.048	-0.042	0.006
maa2d131_c	0.11	0.04	0.082	0.221	0.139
maa2d132_c	-0.046	-0.024	0.156	0.123	-0.033
mas2v062_c	-0.292	0.148	0.093	0.321	0.228
mas2v063_c	-0.074	0.102	-0.18	0.111	0.291
maa2r081_c	0.032	0.01	0.009	-0.054	-0.063
maa2v082_c	0.088	0.026	-0.14	-0.259	-0.119
mas2q041_c	-0.238	-0.21	0.056	-0.002	-0.058

mas2v042_c	0.198	0.058	0.02	-0.023	-0.043
mas2q02s_c	0.026	-0.02	0.02	-0.042	-0.062
maa2d111_c	-0.122	-0.386	0.247	0.284	0.037
maa2d112_c	-0.236	-0.07	0.063	-0.057	-0.12
maa2r011_c	0.426	-0.164	0.128	0.079	-0.049
mas2q011_c	0.096	-0.066	-0.142	0.286	0.428
mag9r061_c	0.232	0.032	0.133	-0.169	-0.302
mas2d071_c	-0.178	0.132	0.052	-0.061	-0.113
mas2d072_c	0.308	0.394	-0.087	-0.16	-0.073
main effect	0.636	0.234	-0.180	-0.342	-0.162

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

Table 13

Item Parameters of the Mathematics Test–Wave 3

Item	Percentage correct	Difficulty / location parameter	*SE* (difficulty / location parameter)	WMNSQ	*t*-value of WMNSQ	Correlation of item score with total score
maa2q071_c	75.46	-1.331	0.072	0.98	-0.5	0.45
mas2r092_c	20.53	1.598	0.077	0.95	-1.2	0.44
mas2v093_c	72.55	-1.158	0.070	1.02	0.6	0.42
mas2v032_c	75.95	-1.36	0.074	1.09	2.3	0.32
maa2d131_c	86.45	-2.156	0.089	0.98	-0.4	0.36
maa2d132_c	64.95	-0.735	0.067	0.92	-3.1	0.55
mas2v062_c	40.63	0.455	0.068	1.02	0.7	0.44
mas2v063_c	41.79	0.394	0.065	1	0	0.45
maa2r081_c	70.67	-1.04	0.070	0.96	-1.4	0.49
maa2v082_c	58.34	-0.407	0.067	1.04	1.4	0.43

mas2q041_c	62.53	-0.614	0.067	1.02	0.8	0.44
mas2v042_c	68.56	-0.796	0.082	0.96	-1.2	0.51
mas2q02s_c	n. a.	-2.128	0.078	0.98	-0.3	0.42
maa2d111_c	74.80	-1.277	0.074	1.03	0.7	0.39
maa2d112_c	39.68	0.504	0.069	1.01	0.5	0.44
maa2r011_c	69.10	-0.963	0.070	0.9	-3.2	0.56
mas2q011_c	80.00	-1.621	0.081	0.97	-0.7	0.44
mag9r061_c	64.94	-0.672	0.074	1.03	1.1	0.40
mas2d071_c	52.80	-0.164	0.069	1.08	3.2	0.38
mas2d072_c	27.05	1.161	0.077	1.13	3.5	0.26

Table 14

Step Parameters (and Standard Errors) of the Polytomous Item–Wave 3

Item	**Step 1 (*SE*)**	**Step 2 (*SE*)**	**Step 3 (*SE*)**
mas2q02s_c	0.017 (0.067)	-0.064 (0.073)	0.047

Table 15

Differential Item Functioning Wave 3

Item	Gender	Migration background	Books		
	male vs female	without vs with	0-200 vs 201-500	0-200 vs 501-	201-500 vs 501-
maa2q071_c	-0.202	-0.076	0.003	0.186	0.183
mas2r092_c	-0.204	-0.25	0.05	0.157	0.107
mas2v093_c	0.174	0.056	-0.197	-0.413	-0.216
mas2v032_c	0.358	0.026	-0.145	-0.107	0.038
maa2d131_c	0.314	0.052	0.161	0.22	0.059
maa2d132_c	-0.268	-0.108	0.156	0.246	0.09
mas2v062_c	0.108	0.186	-0.049	-0.14	-0.091
mas2v063_c	0.008	0.178	-0.186	0.03	0.216
maa2r081_c	-0.16	-0.168	0.088	0.079	-0.009
maa2v082_c	-0.146	-0.092	-0.112	0.112	0.224
mas2q041_c	-0.432	0.146	-0.153	0.099	0.252

mas2v042_c	0.178	-0.09	0.119	-0.011	-0.13
mas2q02s_c	-0.098	-0.31	0.497	0.091	-0.406
maa2d111_c	0.106	-0.202	0.32	0.093	-0.227
maa2d112_c	0.166	0.108	0.009	-0.015	-0.024
maa2r011_c	-0.534	-0.19	0.162	0.081	-0.081
mas2q011_c	-0.098	0.152	0.008	-0.099	-0.107
mag9r061_c	0.572	0.044	0.004	0.098	0.094
mas2d071_c	0.234	-0.038	0.033	0.007	-0.026
mas2d072_c	0.192	0.53	-0.31	-0.497	-0.187
main effect	0.606	0.318	-0.197	-0.328	-0.131