NEPS SURVEY PAPERS

Insa Schnittjer

NEPS TECHNICAL REPORT FOR MATHEMATICS—SCALING RESULTS OF STARTING COHORT 2 IN KINDERGARTEN

NEPS Survey Paper No. 43
Bamberg, June 2018

LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

Survey Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

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 of Starting Cohort 2 in Kindergarten

Insa Schnittjer

IPN – Leibniz Institute for Science and Mathematics Education at Kiel University
University of Koblenz-Landau

Email address of the lead author:

schnittjer@uni-landau.de

Bibliographic Data:

Schnittjer, I. (2018): *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 2 in Kindergarten* (NEPS Survey Paper No. 43). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP43:1.0

Acknowledgements:

I would like to thank Steffi Pohl and Kerstin Haberkorn for developing and providing standards for the technical reports, and Anna-Lena Gerken for assistance in scaling the data. I also would like to thank Timo Gnambs for giving valuable feedback on previous drafts of this manuscript.

The present report has been modeled along previous reports published by NEPS. To facilitate the understanding of the presented results many text passages (e.g., regarding the introduction and the analytic strategy) are reproduces *verbatim* from previous working papers (e.g., Schnittjer & Gerken, 2017).

NEPS Technical Report for Mathematics—Scaling Results of Starting Cohort 2 in Kindergarten

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competencies across the whole life span and designs tests for assessing these different competence domains. In order to evaluate the quality of the competence tests, a wide range of analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedure for the mathematical competence test in kindergarten of starting cohort 2. The mathematics test contained 26 items representing different content areas as well as different cognitive components and using different response formats. The test was administered to 2,727 six-year old kindergarteners. Their responses were scaled using the Rasch model. Item fit statistics, differential item functioning, Rasch-homogeneity, and the test´s dimensionality were evaluated to ensure the quality of the test. These analyses showed that the test exhibited a high reliability, good item fit and that the items fitted the model in a satisfactory way. Furthermore, test fairness could be confirmed for different subgroups. As the correlations between the five content areas were very high in a multidimensional model, the assumption of unidimensionality seems adequate. Overall, the results revealed good psychometric properties of the mathematics test, thus supporting the estimation of a reliable mathematics competence score. Besides the scaling results, this paper also describes the data available in the Scientific Use File and provides ConQuest-Syntax for scaling the data.

Keywords

item response theory, scaling, mathematical competence, kindergarten, Scientific Use File

Contents

1 Introduction

Within the National Educational Panel Study (NEPS), different competencies 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 (ICT) literacy, metacognition, vocabulary, and domain-general cognitive functioning. An overview of the competence domains measured in the NEPS is given by Weinert et al. (2011) and Fuß, Gnambs, Lockl, and Attig (2016).

Most of the competence data are scaled using models based on item response theory (IRT). Because most of the competence tests were developed specifically for implementation in the NEPS, several analyses were conducted to evaluate the quality of the test. 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).

In this paper, the results of these analyses are presented for mathematical competence in starting cohort 2 (kindergarten), taking place in the last year of kindergarten before school starts. First, the main concepts of the mathematical competence test are introduced. Then, the mathematical competence data of the second wave of starting cohort 2 and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the Scientific Use File (SUF) is presented.

Please note that the analyses of this report are based on the data set available at some time different from data release. Due to data protection and data cleaning issues, the data set in the SUF may differ slightly from the dataset used for analyses in this paper. However, major changes in the presented results are not expected.

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. (2013), and Ehmke et al. (2009). In the following, there will be a brief description of specific aspects of the mathematics test that are necessary for understanding the scaling results presented in this paper.

The items are not arranged in units. Thus, in the test, children 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:

- sets, numbers, and operations,
- units and measuring,
- space and shape,
- change and relationships,
- data and chance.

Each item was constructed in such a way as to primarily address a specific content area. The framework also describes as a second and independent dimension six cognitive components required for solving the tasks. These are distributed across the items.

In the mathematics test there are four types of response formats. These are simple multiple-choice (MC), short constructed response (SCR), matching (M), and sorting (S). The most common response format for this age group is the short constructed response (SCR). SCR items require the test-taker to give mostly one-word answers, such as a number. All SCR items were scored dichotomously. Simple multiple-choice items (MC) are items where the children have to find the correct answer from several, usually three or four, response options presented as pictures. Another response format was given by the task to sort selection possibilities into its correct order (S). Items with this response format were scored dichotomously as well for there is only one true order in each item. In matching items (M) the children were asked to match or puzzle some picture cards to given response options. The tasks were constructed in such a way to enable a clear dichotomous scoring.

3 Data

3.1 The Design of the Study

The study assessed different competence domains including, among others, mathematical competence and basic cognitive competence. The test for mathematics competence was administered to all participants followed by a test for basic cognitive competence including perceptual speed. The mathematics test was always administered first on the first of two testing days. Therefore, there was always the same order of booklets. No multi-matrix design was applied regarding the choice and order of the items *within* the mathematics test. All subjects received the same mathematics items in the same order. The test for kindergarteners was conducted as an individual test and was administered in the premises of kindergartens.

The mathematics test for kindergarteners consisted of 26 items which represented different content-related and process-related components and used different response formats. The characteristics of the items are depicted in the following tables. Table 1 shows the distribution of the five content areas, whereas Table 2 shows the distribution of response formats.

Table 1: Number of Items by Content Areas

Content area	Frequency
Sets, numbers, and operations	12
Units and measuring	4
Space and shape	4
Change and relationships	4
Data and chance	2
Total number of items	26

Table 2: Number of items by Response Formats

Response format	Frequency
Short Constructed Response	16
Simple Multiple-Choice	7
Matching	1
Sorting	2
Total number of items	26

3.2 Sample

Overall, the test was administered to 2,727[1] children. Three of them gave less than three valid responses. Because no reliable competence scores can be estimated based on such few responses, these cases were excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 2,724 test takers. A detailed description of the study design, the sample, and the used during the test is available on the NEPS website (http://www.neps-data.de).

3.3 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach.

In this study, all respondents received the same set of items. As a consequence, there were no items that were not administered to a person. Invalid responses occurred if either a child selected more than one answer where only one was required or the person who administered the test did not understand the child's answer. Omitted items occurred if the child did not respond to an item. Due to reasons like exhaustion, it may have occurred that not every child finished the test completely. This resulted in missing responses due to items that have not been reached. However, there was no time limit for the test. Therefore, "not reached" items are missings produced by exhaustion or other reasons to discontinue the test. Still, there will be a report on the "not reached" missings by counting the missings from the end of the test as "not reached" even though their number is negligibly small.

Missing responses provide information on how well the test worked (e.g., exhaustion, understanding of instructions). They also need to be accounted for in the estimation of item and person parameters. Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the persons were coping with the test. Missing responses per item were examined in order to evaluate how well the items functioned.

3.4 Scaling Model

Item and person parameters were estimated using the Rasch model (Rasch, 1960), for all items were scored dichotomously as 0 for an incorrect and 1 for the correct response (see Pohl & Carstensen, 2013 for studies on the scoring of different response formats). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

[1] Note that these numbers may differ from those found in the SUF. This is due to still ongoing data protection and data cleaning issues.

Mathematical competencies were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989). Person parameter estimation in the NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in section 6.

3.5 Checking the Quality of the Scale

The mathematics test was specifically constructed to be implemented in NEPS. In order to ensure appropriate psychometric properties, the quality of the test was examined in several analyses.

The fit of the dichotomous variables to the Rasch model (Rasch, 1960) was evaluated using three indices (see Pohl & Carstensen, 2012). 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 a considerable item misfit, and their performance was further investigated. Correlations of the item score with the total correct score (equal to the discrimination value 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 the fit of an item was based on all fit indicators.

The mathematical competence test should measure the same construct for all participants. If some items favored certain subgroups (e.g., they 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. For the present study, test fairness was investigated for the variables gender, speaking, and understanding. Those last two variables gave an indication on how good the test takers command of German vocabulary and sentence structure for his or her age group was. Differential item functioning (DIF) was examined using a multi-group IRT model in which the main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties between the subgroups that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as considerable and noteworthy of further investigation, absolute differences between 0.4 and 0.6 as small and not severe, and differences smaller than 0.4 as negligible DIF. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

The dimensionality of the mathematics test was evaluated by specifying a five-dimensional model based on the five content areas. Each item was assigned to one content area (between-item-multidimensionality). To estimate this multidimensional model, Quasi Monte Carlo integration in TAM (Kiefer, Robitzsch, & Wu, 2017) was used. To guarantee the compatibility with the multidimensional model, the unidimensional model was estimated in TAM as well. The number of nodes in the multidimensional model was chosen in such a way as to obtain stable parameter estimates (15,000 nodes). The correlations between the subdimensions as well as differences in model fit between the unidimensional model and the respective multidimensional model were used to evaluate the unidimensionality of the test.

3.6 Software

The Rasch models were estimated in ConQuest version 4.2.5 (Adams, Wu & Wilson, 2015). The two-parametric logistic model was estimated in MDLTM (Matthias von Davier, 2005). To check the multidimensionality, the IRT models were also estimated in TAM version 2.4-9 (Kiefer et al., 2017) in R version 3.4.1 (R Core Team, 2017).

4 Responses

4.1 Missing Responses

4.1.1 Missing responses per person

As can be seen in Figure 1, the number of invalid responses per person was negligible. In fact, for 98.9 % of the test-takers no invalid response was determined. The maximum number of invalid responses was two.

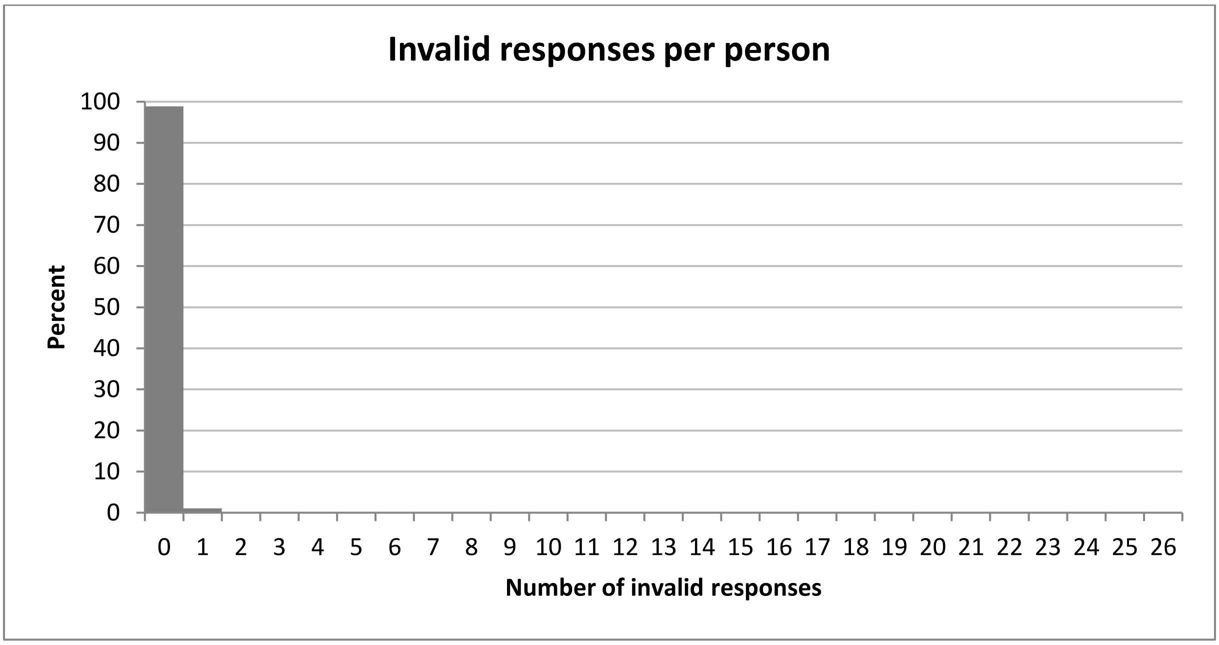

Figure 1. Number of invalid responses.

Missing responses may also occur when a child does not respond to an item (omit). The number of omitted responses per test taker is depicted in Figure 2. It shows that 53.7 % of the subjects omitted no item and only 6.1 % of the subjects omitted five or more items.

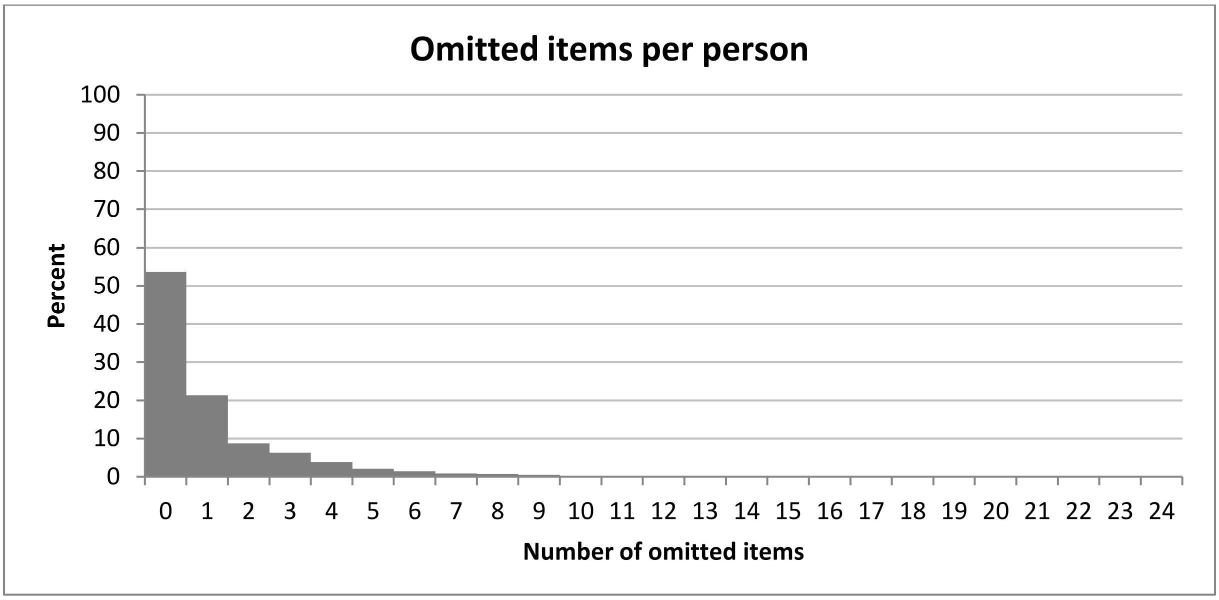

Figure 2. Number of omitted items.

All missing responses after the last valid response are defined as not reached. Figure 3 shows the number of items that were not reached by a person. As can be seen, 97.9 % reached the end of the test. Therefore, only 2.1 % of the subjects did not reach the last item.

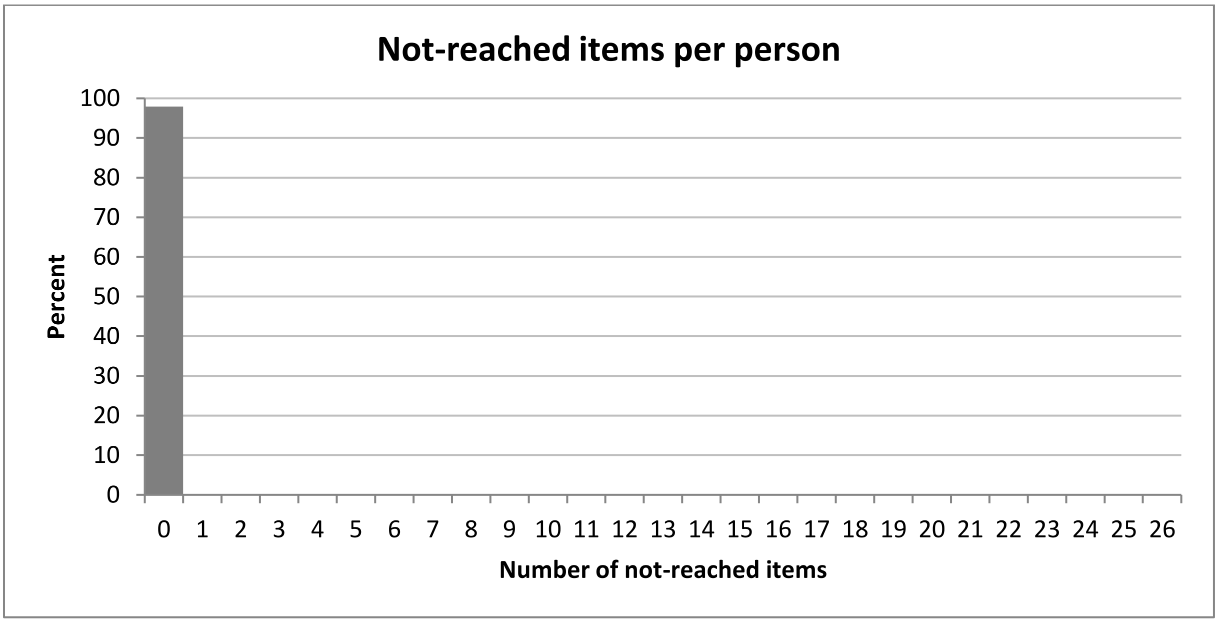

Figure 3. Number of not-reached items.

Figure 4 shows the total number of missing responses per person, which is the sum of invalid, omitted, and not-reached missing responses. In total, 53.1 % of the subjects show no missing response. Only 4.4 % show more than five missing responses at all. Overall, there was a negligible amount of invalid, not-reached, and omitted items.

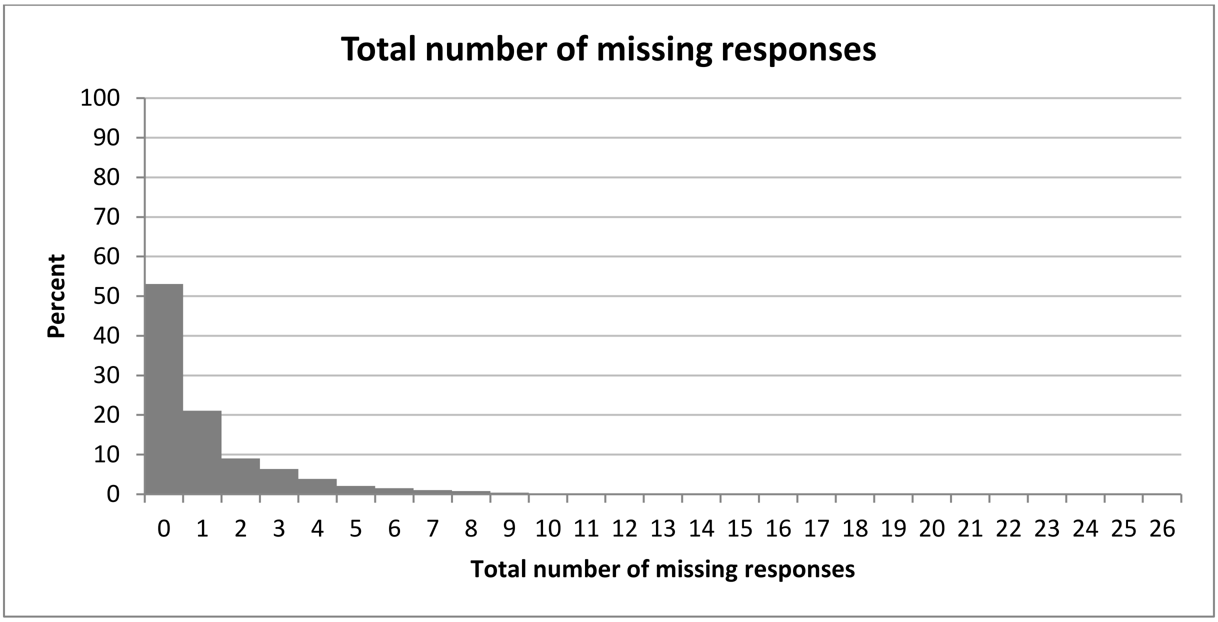

Figure 4. Total number of missing responses.

4.1.2 Missing responses per item

Table 3 shows the number of valid responses for each item as well as the percentage of missing responses. Overall, the number of invalid responses per item was negligible.

The omission rates were good, except for three noticeable items with an omission rate higher than 10%. The highest omission rate (19.53 %) occurred for item mak2z101_c. As this item was the first SCR item, without any supporting material for answering such as pictures or countable material, the children might have preferred to skip the item rather than to guess. Furthermore, it should be taken into account that it might be an age group specific behavior to react more reserved, indeed shy, than older test takers. The other two items with noticeable omission rates (10.61 % and 11.34 %) were the first item with the response format sorting (mak2g051_c), which also was one of the most difficult items, as well as the first SCR item in the test that asked the participants to provide an explanation for their response (mak2z161_c). This call for reasoning might have encouraged the children to skip rather than to guess an answer.

The number of persons that did not reach an item increased with the position of the item in the test to up to 2.09 %. The total number of missing responses per item varied between 0.33 % (mak2g051_c) and 19.53 % (mak2z101_c).

Table 3: Percentage of Missing Values

Item	Position in the test	Number of valid responses	Percentage of invalid responses	Percentage of omitted missings	Percentage of not-reached items
mak2z221_c	1	2,639	0.00	3.12	0.00
mak2z231_c	2	2,692	0.00	1.17	0.00
mak2z101_c	3	2,192	0.00	19.53	0.00
mak2r111_c	4	2,689	0.07	1.21	0.00
mak2g041_c	5	2,435	0.00	10.61	0.00
mak2g051_c	6	2,715	0.00	0.33	0.00
mak2v001_c	7	2,515	0.04	7.64	0.00
mak2r151_c	8	2,481	0.11	8.81	0.00
mak2z031_c	9	2,519	0.11	7.38	0.04
mak2d062_c	10	2,705	0.00	0.66	0.04
mak2z161_c	11	2,412	0.04	11.34	0.07
mak2z171_c	12	2,695	0.11	0.88	0.07
mak2g211_c	13	2,541	0.00	6.64	0.07
mak2r131_c	14	2,700	0.00	0.77	0.11
mak2z091_c	15	2,617	0.04	3.74	0.15
mak2v081_c	16	2,611	0.15	3.85	0.15
mak2z201_c	17	2,635	0.07	3.05	0.15
mak2d011_c	18	2,700	0.00	0.73	0.15
mak2z241_c	19	2,470	0.11	9.07	0.15
mak2z121_c	20	2,589	0.11	4.70	0.15
mak2v071_c	21	2,646	0.04	2.64	0.18
mak2g021_c	22	2,663	0.00	2.06	0.18
mak2z251_c	23	2,672	0.00	1.73	0.18
mak2r191_c	24	2,713	0.00	0.22	0.18
mak2v181_c	25	2,655	0.04	2.06	0.44
mak2z141_c	26	2,664	0.11	0.00	2.09

Note. In former versions of the SUF item mak2v081_c was named mak2v08s_c, scaling it dichotomously. The name had to be adapted to conform to the naming conventions in the NEPS.

4.2 Parameter Estimates

4.2.1 Item parameters

In order to get a first rough descriptive measure of item difficulties and check for possible estimation problems, the relative frequency of the responses given before performing IRT analyses were evaluated. The percentage of persons correctly responding to an item (relative to all valid responses) varied between 16.06 % and 90.98 % across all items. On average, the rate of correct responses was 47.09 % (*SD* = 19.75 %). From a descriptive point of view, the items covered a relatively wide range of difficulties.

The estimated item difficulties are depicted in Table 4. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties varied between -2.69 (item mak2z221_c) and 1.98 (item mak2g051_c) with a mean of 0.18. Due to the large sample size, the standard errors of the estimated item difficulties (column 4) were very small (*SE*(ß) ≤ 0.08).

Table 4: Item Parameters

Item	Posi-tion	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	Item total correlation (*r_{it}*)	Discr.
mak2z221_c	1	90.98	-2.688	0.075	0.90	-2.0	0.38	1.22
mak2z231_c	2	55.79	-0.269	0.047	0.90	-6.3	0.57	1.28
mak2z101_c	3	52.51	-0.031	0.052	0.88	-6.9	0.59	1.40
mak2r111_c	4	54.44	-0.209	0.047	1.13	7.7	0.34	0.45
mak2g041_c	5	21.03	1.624	0.058	1.01	0.4	0.37	0.70
mak2g051_c	6	16.06	1.978	0.060	0.97	-0.8	0.36	0.79
mak2v001_c	7	52.72	-0.096	0.049	1.02	1.0	0.45	0.72
mak2r151_c	8	46.39	0.174	0.049	1.13	7.7	0.33	0.43
mak2z031_c	9	43.19	0.386	0.049	0.95	-3.1	0.51	1.03
mak2d062_c	10	28.91	1.098	0.051	1.09	4.2	0.33	0.52
mak2z161_c	11	35.49	0.762	0.051	1.00	0.2	0.44	0.75
mak2z171_c	12	68.53	-0.934	0.050	0.99	-0.3	0.46	0.85
mak2g211_c	13	34.71	0.805	0.050	1.02	0.9	0.43	0.74
mak2r131_c	14	20.41	1.646	0.056	1.10	3.5	0.27	0.41
mak2z091_c	15	39.28	0.567	0.049	0.91	-5.9	0.56	1.24
mak2v081_c	16	48.53	0.101	0.048	0.96	-2.9	0.52	1.00
mak2z201_c	17	53.36	-0.147	0.048	0.92	-5.0	0.55	1.12
mak2d011_c	18	67.89	-0.903	0.050	1.12	5.7	0.32	0.45
mak2z241_c	19	22.59	1.514	0.056	0.90	-3.8	0.52	1.50
mak2z121_c	20	60.18	-0.466	0.049	0.92	-4.6	0.54	1.16
mak2v071_c	21	59.83	-0.468	0.048	1.06	3.6	0.40	0.58
mak2g021_c	22	75.59	-1.341	0.053	0.86	-5.6	0.56	1.37
mak2z251_c	23	24.48	1.370	0.053	1.01	0.5	0.40	0.76
mak2r191_c	24	67.45	-0.882	0.049	1.14	6.4	0.31	0.41
mak2v181_c	25	20.45	1.644	0.056	0.93	-2.3	0.47	1.22
mak2z141_c	26	63.59	-0.669	0.049	1.02	1.3	0.44	0.73

Note. In former versions of the SUF item mak2v081_c was named mak2v08s_c (see table 3).

4.2.2 Test targeting and reliability

Scale in logits	Person ability	Item difficulty
3	X	
	X	
	X	
	XXX	
	XX	
	XXX	
	XXXX	
2	XXXXX	6
	XXXXX	
	XXXXXXXX	
	XXXXXXXXXXXX	5 14 25
	XXXXXXXXXXXXXX	19
	XXXXXXXXXXXXXXXXXXXXXXX	23
	XXXXXXXXXXXXXXXXXXXXX	
1	XXXXXXXXXXXXXXXXXXXXXXX	10
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	11 13
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	15
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	9
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	8
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	16
0	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	3 7
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	2 4 17
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXX	20 21
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	26
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	12 18 24
-1	XXXXXXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXXXXXX	22
	XXXXXXXXXXXX	
	XXXXXXXXXXXXX	
	XXXXXXXXXXX	
	XXXXXXX	
-2	XXXXX	
	XXX	
	XXXXX	
	XX	
	XX	
	X	1
	XX	
-3	X	
	X	
	X	
	X	

Figure 5: Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 3.8 cases. The difficulty of the items is depicted on the right-hand side of the graph. Each number represents one item (see Table 4).

Test targeting focuses on comparing the item difficulties with the person´s abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. In Figure 5, item difficulties of the mathematics items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers´ ability is mapped onto the left side whereas the right side shows the distribution of item difficulties. The mean of the ability distribution was constrained to be zero. The respective item difficulties ranged from -2.69 (item mak2z221_c) to 1.98 (item mak2g051_c). Therefore, a rather broad range was covered. The variance was estimated to be 1.087, indicating that the test differentiated well between subjects. The reliability of the test (EAP/PV reliability = 0.820. WLE reliability = 0.804) was good. Although the items covered a wide range of the ability distribution, there were somewhat few very easy items. As a consequence, person abilities in high-ability regions will be measured relative precisely, whereas lower ability estimates will have larger standard errors.

4.3 Quality of the test

4.3.1 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the Rasch model. Altogether, item fit can be considered to be very good (see Table 4). Values of the WMNSQ were close to 1 with the lowest value being 0.86 (item mak2g021_c) and the highest 1.14 (item mak2r191_c). The items with the largest WMNSQ (mak2d011_c, mak2r111_c, mak2r151_c and mak2r191_c) showed acceptable, slightly flat item characteristic curves (ICC). Therefore, all ICC showed a good or very good fit of the items. Overall, there was no indication of severe item over- or underfit. The correlations of the item scores with the total scores varied between .27 (item mak2r131_c) and .59 (item mak2z101_c) with an average correlation of .44.

4.3.2 Differential item functioning

Differential item functioning (DIF) was used to evaluate the test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables gender, speaking, and understanding, the latter two being cohort specific variables that were considered important for this age group (see Pohl & Carstensen, 2012, for a more detailed description on DIF variables in the NEPS). These two variables gave an indication on how good the child´s command of German vocabulary and sentence structure for his or her age was (see https://www.neps-data.de/en-us/datacenter/overviewandassistance/nepsplorer.aspx#/search/q=e41260&y=2011*12 for more detailed descriptions on these two variables).

Table 6 shows the difference between the estimated difficulties of the items in different subgroups. Female versus male, for example, indicates the difference in difficulty between girls and boys, *ß(female)* – *ß*(male). A positive value indicates a higher difficulty for females, a negative value a lower difficulty for females compared to males.

Table 6: Differential Item Functioning

Item	**Gender**	**Speaking**			**Understanding**		
	Female vs. male	Not good vs. very good	Not good vs. missing	Very good vs. missing	Not good vs. very good	Not good vs. missing	Very good vs. missing
mak2z221_c	-0.264	0.744	-0.020	-0.770	0.596	0.012	-0.594
mak2z231_c	0.034	0.314	0.254	-0.068	0.196	0.214	0.008
mak2z101_c	0.240	0.374	0.228	-0.152	0.292	0.188	-0.114
mak2r111_c	-0.214	-0.374	-0.108	0.260	-0.316	-0.150	0.158
mak2g041_c	0.044	-0.036	-0.312	-0.276	-0.230	-0.456	-0.232
mak2g051_c	0.424	-0.114	-0.158	-0.048	-0.198	-0.192	-0.002
mak2v001_c	0.048	-0.032	-0.154	-0.126	0.014	-0.118	-0.142
mak2r151_c	-0.462	-0.172	-0.226	-0.058	-0.234	-0.344	-0.118
mak2z031_c	-0.150	0.040	0.032	-0.016	0.080	0.094	0.004
mak2d062_c	0.008	-0.172	0.064	0.230	-0.134	0.084	0.208
mak2z161_c	0,202	-0.142	-0.140	-0.002	-0.182	-0.134	0.040
mak2z171_c	0.358	-0.276	-0.272	0.000	-0.214	-0.296	-0.090
mak2g211_c	0.128	0.062	0.024	-0.042	0.276	0.196	-0.088
mak2r131_c	-0.082	-0.376	-0.100	0.270	-0.400	-0.142	0.250
mak2z091_c	-0.240	0.192	0.196	-0.004	0.232	0.304	0.060
mak2v081_c	-0.100	0.112	0.142	0.022	0.148	0.194	0.036
mak2z201_c	-0.092	-0.062	0.078	0.132	0.012	0.146	0.124
mak2d011_c	-0.058	-0.048	0.098	0.138	-0.128	0.026	0.144
mak2z241_c	0.118	0.316	-0.020	-0.340	0.406	0.062	-0.352
mak2z121_c	0.238	0.068	0.044	-0.030	0.036	0.030	-0.016
mak2v071_c	0.004	-0.088	-0.028	0.056	-0.092	0.024	0.106
mak2g021_c	-0.430	0.502	-0.018	-0.526	0.438	-0.004	-0.452
mak2z251_c	0.356	-0.148	-0.022	0.120	-0.106	-0.010	0.088
mak2r191_c	0.014	-0.322	-0.038	0.278	-0.252	-0.032	0.210
mak2v181_c	-0.100	0.500	0.226	-0.278	0.530	0.354	-0.186
mak2z141_c	0.054	-0.236	0.104	0.330	-0.312	-0.046	0.258
Main effect (Model with DIF)	**0.178**	**0.826**	**0.312**	**-0.510**	**0.786**	**0.398**	**-0.380**
Main effect (Model without DIF)	**0.178**	**0.820**	**0.312**	**-0.512**	**0.782**	**0.402**	**-0.384**

Overall, 1,339 (49.2 %) of the test takers were female and 1,385 (50.8 %) were male. On average, in kindergarten male children exhibited a higher mathematical competence than female children (main effect = 0.178 logits, Cohen´s *d* = 0.171). There were three items that showed DIF greater than 0.4 logits (mak2g051_c, mak2r151_c, and mak2g021_c). However, with DIFs being below 0.5 logits, the differences between the two groups were not considered severe.

In addition to the competence tests, the test takers were asked to answer some questions to categorize their speaking ability into speaking very well and speaking not very well. Overall, there were 1,196 (43.9 %) test takers categorized into speaking not very well, whereas 1,242 (45.6 %) test takers spoke very well. For 286 (10.5 %) test takers the administrating interviewers did not give any valid answers. All three groups were used for investigating DIF of speaking. On average, test takers with high speaking ability performed better than children with poor speaking ability (main effect = 0.820 logits, Cohen´s *d* = 0.856). Subjects with missing values for speaking differed from those who spoke well (main effect = 0.512 logits, Cohen's *d* = 0.559). Here, again, participants with a high speaking ability showed a higher mathematical competence. Subjects with a poor speaking ability performed better compared to participants with missings (main effect = 0.312 logits, Cohen's *d* = 0.302). There were two items with DIFs above 0.4 logits (mak2g021_c and mak2v181_c). However, the differences being close to 0.4 logits and showing good item fit in all other categories, DIFs were considered not severe. Furthermore, one item showed differences exceeding 0.6 logits (mak2z221_c). It showed a noticeably large DIF with a difference of 0.744 logits between the groups. This item was also the first item in the test, and therefore it seemed plausible that this first item was a major obstacle for test takers with lower speaking abilities. While going through the test, this obstacle might reduce slowly due to the test takers overcoming their shyness and adapting to the test situation, as well as the interviewers getting used to the pronunciation of the children. Furthermore, this item showed by far the lowest difficulty and therefore, the item showed overall good item fit.

Test takers participated in a short test on understanding the German language. Therefore, three subgroup categories were built: understanding very well, understanding not very well and missing. All three categories were analyzed through another DIF analysis. There were 890 (32.7 %) test takers that understood German not very well, 1,559 (57.2 %) test takers understood Germany well, and 275 (10.1 %) test takers with no valid answers. Group differences and DIF were investigated by using all three groups. On average, test takers with a high understanding ability performed better than children with a poor understanding ability (main effect = 0.786 logits, Cohen´s *d* = 0,806). Participants without a valid response in relation to the variable understanding performed 0.402 logits (Cohen's *d* = 0.394) better than participants with lower understanding ability and 0.384 logits (Cohen's *d* = 0.398) worse than participants with higher understanding ability. Overall, five items showed differences above 0.4 logits (mak2z221_c, mak2g041_c, mak2z241_c, mak2g021_c, mak2v181_c). Nevertheless, these items showed overall good item fit.

In Table 7, we compared the models that included only main effects on the three variables to models that additionally estimated DIF effects. Akaike's (1974) information criterion (AIC) favored the models estimating DIF for all three DIF variables. The Bayesian information criterion (BIC; Schwarz, 1978) takes the number of estimated parameters more strongly into account and, thus, prevents from overparametrization of models. Using the BIC, the more parsimonious models including only the main effects were preferred for all three DIF variables.

Table 7: Comparison of Models With and Without DIF

DIF variable	Model	Deviance	Number of parameters	AIC	BIC
Gender	Main effect	75525.38	28	75,581.376	75,746.852
	DIF	75391.34	54	75,499.342	75,818.474
Speaking	Main effect	75207.50	29	75,265.496	75,436.882
	DIF	75022.64	81	75,184.639	75,663.338
Understanding	Main effect	75263.28	29	75,321.280	75,492.666
	DIF	75093.62	81	75,255.620	75,734.319

4.3.3 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item discrimination parameters are equal. In order to test this assumption, a two-parametric logistic model (2PL; Birnbaum, 1968) that estimates different discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 4), ranging from 0.41 (item mak2r131_c and mak2r191_c) to 1.50 (item mak2z241_c). The average discrimination parameter fell at 0.88. Model fit indices suggested a slightly better model fit of the 2PL model (AIC = 74,699.24, BIC = 75,242.95, number of parameters = 92) as compared to the Rasch model (AIC = 75,659.72, BIC = 76,049.77, number of parameters = 66). Despite the empirical preference for the 2PL model, the Rasch model more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the Rasch model was chosen as our scaling model. Note that these calculations were performed in MDLTM (see Davier, 2005). As a consequence, other results for AIC and BIC using the Rasch model might differ from these results (see 4.3.5).

4.3.4 Unidimensionality

The unidimensionality of the test was investigated by specifying a five-dimensional model based on the five different content areas. Each item was assigned to one content area (between-item-multidimensionality).

To estimate this multidimensional model, the Quasi Monte Carlo method implemented in TAM in R version 3.4.1 (R Core Team, 2017) was used. (Due to convergence problems even with 25 nodes per dimension, model parameters could not be estimated in ConQuest using the Gauss-Hermite quadrature method. This might be caused by the fact that there is more than three dimensions as well as the high correlations between them.) The number of nodes used in TAM was set to 15,000.

The variances and correlations of the five dimensions are shown in Table 8. Three of the five dimensions exhibited a substantial variance. In dimension 4, three out of four items showed difficulties ranging from -0.468 to 0.101, so the difficulties were very homogenous in this dimension which could explain the small variance. A similar distribution of difficulties was found in dimension 3 which could also explain the rather small variance. As expected, the correlations between the five dimensions were rather high, varying between 0.700 and 0.945. However, they deviated from a perfect correlation (i.e., they were lower than *r* = .95, see Carstensen, 2013). Still, according to model fit indices, the five-dimensional model fitted the data slightly better (AIC =74,954.570, BIC = 75,196.874, number of parameters = 41) than the unidimensional model (AIC = 75,595.590, BIC = 75,755.156, number of parameters = 27).

These results indicate that the five content areas measure a common construct, although they are not completely unidimensional. Model fit between the unidimensional and the five-dimensional model is compared in Table 9.

Table 8: Results of Five-Dimensional Scaling

	Sets, numbers, and operations	Units and measurement	Space and shape	Change and relationships	Data and chance
Sets, numbers, and operations (12 items)	1.840				
Units and measuring (4 items)	0.906	1.585			
Space and shape (4 items)	0.800	0.886	0.567		
Change and relationships (4 items)	0.870	0.879	0.789	0.393	
Data and chance (2 items)	0.927	0.945	0.856	0.910	1.247

Note. Variances of the dimensions are depicted in the diagonal; correlations are given in the off-diagonal.

Table 9: Comparison of the Unidimensional and the Five-Dimensional Model

Model	Deviance	Number of parameters	AIC	BIC
Unidimensional	75,541.59	27	75,595.590	75,755.156
Five-dimensional	74,872.57	41	74,954.570	75,196.874

Note. Contrary to the calculations for the 1PL and 2PL models, results in this table were achieved by using TAM in R (see 3.6).

5 Discussion

The analyses in the previous sections aimed at providing information on the quality of the mathematics test in starting cohort 2 and at describing how the mathematics competence score was estimated.

The amount of different kinds of missing responses was evaluated and all kinds of missing responses were negligible. Furthermore, item as well as test quality were examined. Overall, there was a negligible amount of invalid, not-reached, and omitted items. As indicated by various fit criteria —WMNSQ, *t*-value of the WMNSQ, ICC— the items exhibited good fits. Moreover, discrimination values of the items (either estimated in a 2PL model or as a correlation of the item score with the total score) were acceptable. The test had a good reliability (EAP/PV-reliability = .820, WLE reliability = .804). It distinguished well between test takers, indicated by the test's variance (= 1.087). Different variables were used for testing measurement invariance. No considerable DIF became evident for any of these variables, indicating that the test was fair for the examined subgroups. Fitting a five-dimensional model (between-item-multidimensionality, the dimensions being the content areas) yielded a slightly better model-fit than the unidimensional model. However, high correlations of 0.7 and higher between the five dimensions indicated that the unidimensional model described the data reasonably well. In summary, the test had good psychometric properties that facilitated the estimation of a unidimensional mathematics competence score.

6 Data in the Scientific Use File

6.1 Naming conventions

The SUF contains 26 items that were scored as dichotomous variables with 0 indicating an incorrect response and 1 indicating a correct response. Dichotomous items are marked with a '_c' at the end of the variable name. Manifest scale scores are provided in the form of WLE estimates (mak2_sc1) including the respective standard error (mak2_sc2). The ConQuest Syntax for estimating the WLE scores from the items are provided in the Appendix. Test takers that did not take part in the test or that did not give enough valid responses to estimate a scale score will have a non-determinable missing value on the WLE score for mathematical competence. Users interested in investigating latent relationships may either include the measurement model in their analyses or estimate plausible values. A description of these approaches can be found in Pohl and Carstensen (2012).

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ConQuest 4*. Camberwell, Australia: Acer.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 397–479). Reading, MA: MIT Press.

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

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 (Eds.). Mathematiklernen vom Kindergarten bis zum Studium: Kontinuität und Kohärenz als Herausforderung für den Mathematikunterricht (pp. 313-327). Münster: Waxmann.

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2016). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Haberkorn, K., Pohl, S., Carstensen, C., & Wiegand, E. (2012). Incorporating different response formats in the IRT-scaling model for competence data. Manuscript submitted for publication.

Kiefer, T., Robitzsch, A. & Wu, M. (2016). TAM: Test Analysis Modules. [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=TAM (R package version 1.995-0).

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., & Carstensen, C. H. (2013). Scaling the competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online*, 5(2), 189-216.

R Core Team (2016). R: A language and environment for statistical computing (Version 3.2.4) [Software]. Retrieved from https://www.R-project.org/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche (Expanded edition, 1980, Chicago: University of Chicago Press).

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

Van den Ham, A.-K. (2016). *Ein Validitätsargument für den Mathematiktest der National Educational Panel Study für die neunte Klassenstufe.* Unpublished German dissertation, Leuphana University Lüneburg, Lüneburg.

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.

Appendix

Appendix A: ConQuest-Syntax for Estimating WLE Estimates in Starting Cohort II – Kindergarteners

```
Title Starting Cohort II. MATHEMATICS: Rasch Model;

data filename.dat;
format pid 1-10 responses 12-37; /* insert number of columns with data*/

labels << filename_with_labels.nam;

codes 0,1;
score (0,1)(0,1) !item(1-26);

set constraint=cases;

model item;
estimate;

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

Appendix B: Content Areas of Items in the Mathematics Test for Kindergarteners

Position	Item	Content area
1	mak2z221_c	Sets, numbers, and operations
2	mak2z231_c	Sets, numbers, and operations
3	mak2z101_c	Sets, numbers, and operations
4	mak2r111_c	Space and shape
5	mak2g041_c	Units and measuring
6	mak2g051_c	Units and measuring
7	mak2v001_c	Change and relationships
8	mak2r151_c	Space and shape
9	mak2z031_c	Sets, numbers, and operations
10	mak2d062_c	Data and chance
11	mak2z161_c	Sets, numbers, and operations
12	mak2z171_c	Sets, numbers, and operations
13	mak2g211_c	Units and measuring
14	mak2r131_c	Space and shape
15	mak2z091_c	Sets, numbers, and operations
16	mak2v081_c	Change and relationships
17	mak2z201_c	Sets, numbers, and operations
18	mak2d011_c	Data and chance
19	mak2z241_c	Sets, numbers, and operations
20	mak2z121_c	Sets, numbers, and operations
21	mak2v071_c	Change and relationships
22	mak2g021_c	Units and measuring
23	mak2z251_c	Sets, numbers, and operations
24	mak2r191_c	Space and shape
25	mak2v181_c	Change and relationships
26	mak2z141_c	Sets, numbers, and operations

Note. Up to now, the internal validity of the individual dimensions of mathematical competence as dependent measures has not yet been confirmed (van den Ham, 2016)