NEPS
National Educational Panel Study

NEPS Working Papers

Katrin Schöps

NEPS Technical Report for Science – Scaling Results of Starting Cohort 2 in Kindergarten

NEPS Working Paper No. 24

Bamberg, April 2013

SPONSORED BY THE
Federal Ministry
of Education
and Research

Contact: German National Educational Panel Study (NEPS) − University of Bamberg − 96045 Bamberg − Germany − contact.neps@uni-bamberg.de

NEPS Technical Report for Science – Scaling Results of Starting Cohort 2 in Kindergarten

Katrin Schöps

National Educational Panel Study
Leibniz Institute for Science and Mathematics Education (IPN), Kiel

Email address of the author:

schoeps@ipn.uni-kiel.de

Bibliographic data:

Schöps, K. (2013). *NEPS Technical Report for Science – Scaling results of Starting Cohort 2 in Kindergarten*. (NEPS Working Paper No. 24). Bamberg: University of Bamberg, National Educational Panel Study.

I thank Steffi Pohl and Kerstin Haberkorn for developing and providing standards for the technical reports and for giving valuable feedback on previous drafts of this manuscript. I thank Steffani Sass for producing the graphs and for proof-reading the manuscript.

NEPS Technical Report for Science – Scaling Results of Starting Cohort 2 in Kindergarten

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competences 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 have been performed based on item response theory (IRT). This paper describes the data on scientific literacy for starting cohort 2 in kindergarten. Next to descriptive statistics of the data, the scaling model applied to estimate competence scores, analyses performed to investigate the quality of the scale, as well as the results of these analyses are presented. The test consisted of 26 items and was administered to 2,955 children. A partial credit model was used for scaling the data. Item fit statistics, differential item functioning, Rasch-homogeneity, and the tests' dimensionality were evaluated to ensure the quality of the test. The results show that the items, except for one, exhibited good item fit and measurement invariance across various subgroups. Moreover, the test showed a high reliability. As the correlations between the two knowlegde domains are very high in a multidimensional model, the assumption of unidimensionality seems adequate. However, a justified point of criticism is that the test lacks very difficult items. But overall, the results revealed good psychometric properties of the science test, thus supporting the estimation of a reliable scientific literacy score. This paper describes the data available in the Scientific Use File and provides ConQuest-Syntax for scaling the data.

Keywords

Scientific Literacy, Kindergarten, Picture Based Assessment, Differential Item Functioning, Item Response Theory, Scientific Use File

Content

1. Introduction

Within the National Educational Panel Study (NEPS) different competences are measured coherently across the life span. Tests have been developed for different domains including scientific literacy. Weinert et al. (2011) give an overview of the competence domains measured in NEPS.

Most of the competence data are scaled using models based on Item Response Theory (IRT). Since most of the competence tests were developed solely for implementation in NEPS, several analyses have been performed 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 scale are described in Pohl and Carstensen (2012a). In this paper the results of these analyses are presented for scientific literacy in the starting cohort 2.

The present report has been modeled along the technical reports of Pohl, Haberkorn, Hardt, & Wiegand (2012) and Haberkorn, Pohl, Hardt, & Wiegand (2012). Note that some of the analyses presented in this report are based on preliminary data releases. Due to data protection and data cleaning issues the data set in the Scientific Use File (SUF) may differ slightly from the data set used for the analyses in this paper. We do, however, not expect severe changes in results.

2. Testing scientific literacy

The science test aims at assessing two types of scientific sub-competencies. These are a) knowledge of science (KOS) and b) knowledge about science (KAS). Using the definition by PISA (OECD, 2007, Prenzel et al. 2007) – (KOS) is specified as knowledge of basic scientific concepts and facts, whereas knowledge about science (KAS) can be regarded as the understanding of scientific processes (Prenzel et al., 2007).

KOS is divided into content-related components: matter, system, development and interaction. KAS is divided in the process-related components scientific enquiry and scientific reasoning. KAS and KOS are implemented in three contexts: health, environment, and technology (see Hahn et al. (2012) and Weinert et al. (2011) for the description of the framework). The test items are organized in units (testlets). Thus one unit consists of two or three items. Each unit refers to one context-component-combination.

There are three types of response formats. These are simple multiple choice (MC), complex multiple choice (CMC) in the special form of true false items and short-constructed response (SCR) items. In MC items the test taker has to find the correct answer out of four response options. In CMC items the test taker has to decide at each answer option whether the answer is correct or not. In the short-constructed response format the test taker is free to answer, whatever she/he wants.

In kindergarten the response categories were represented by picture cards. In order to answer a MC item the children had to choose between four different pictures. For answering the CMC items the children were successively confronted with single pictures (usually four) and gave an oral response to a question by saying either “yes” or “no”. In addition to the special item structure, the science test in kindergarten was also characterized by a special setting with an interviewer testing only one child at a time and by the fact that the items were embedded in a story.

3. Data

3.1 The Design of the study

Overall 2,955 took part in the test. On the testing day only the science test was administered and all children received the same test items.

The science test in kindergarten consists of 26 Items which represent different KOS/KAS, content, components and use different response formats (see below).

Due to unsatisfactory item fit parameters one item was eventually excluded when determining the WLEs. However the item was part of our initial analyses and only due to these dissatisfying results it was eventually excluded. The characteristics of the 26 items are depicted in table 1 (knowledge domains), table 2 (response format) and table 3 (content-components).

Table 1: Numbers of items for the two *scientific sub-competencies* KOS and KAS i*n the kindergarten science test*

Knowledge domains	**Frequency**
Knowledge of Science (KOS)	18
Knowledge about Science (KAS)	8
Total number of items	26

Table 2: Response formats of the items in the kindergarten science test

Response format	**Frequency**
Simple Multiple-Choice	19
Complex Multiple-Choice	4
Short-constructed response	3
Total number of items	26

Table 3: Number of items for the different contexts in the kindergarten science test

Context	Frequency
Health	4
Environment	14
Technology	8
Total number of items	26

3.2 Sample

A description of the design of the study, the sample, as well as the instruments used can be found on the NEPS-website[1].

Overall there 3,006 children[2] were supposed to take the science test. However, only data for 2,955 children are available. Either the persons -or in this case their parents- withdrew their approval or the children did not turn up for the test.

The data sets from four persons were omitted from the analyses as they only contained missing values. Hence, data from 2,951 persons are included in the descriptive analyses. For further analyses only persons with more than two valid responses were taken into account. The results of the remaining n=2,947 test takers are presented in the following sections.

4. Analyses

4.1 Missing responses

In the kindergarten science test only two kinds of missing values occurred. An answer was regarded as omitted, if the child did not respond to the question. Invalid responses occurred, for example, when a child selected more than one answer or the person who administered the test did not understand the answer. In this study, all subjects received the same set of items. As a consequence, there are no items that were not administered to a person. As this test was administered in individual test settings by an interviewer, every child who took part in the test completed it and the "not reached"-missing did not occur.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions) and need to be accounted for in the estimation of item and person parameters. We, therefore, thoroughly investigated the occurrence of missing responses in the test. First we investigated the occurrence of the different types of missing responses per person. This gives an indication on how well the persons got along with the

[1] www.neps-data.de

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

test. We will then look at the occurrence of missing responses per item, in order to get some information on how well the items worked.

4.2 Scaling model

For estimating item and person parameters for scientific literacy, a partial credit model (Masters, 1982) was used and estimated in ConQuest (Wu, Adams, & Wilson, 1997). A detailed description of the scaling model can be found in Pohl and Carstensen (2012a).

CMC items consist of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly responded subtasks within that item[3]. If at least one of the subtasks contains a missing response, the whole CMC item was scored as missing. When categories of the polytomous variables had less than N=200, in order to avoid possible estimation problems, the categories were collapsed. This usually occurred for the lower categories of polytomous items; especially when the item consisted of many subtasks. In these cases the lower categories were collapsed to one category. However, for none of the Kindergarten items, categories had to be collapsed (cell frequency >2%).

In the following analyses each category of the polytomous items (CMC) was scored with 0.5 points, while simple MC items were scored dichotomously as 0 for an incorrect and 1 for the correct response (see Haberkorn, Pohl, Carstensen, & Wiegand, 2012; and Pohl & Carstensen, 2012, for studies on the scoring of different response formats).

Item difficulties for dichotomous variables and location parameters for polytomous parameters are estimated using the partial credit model. Ability estimates for scientific literacy were estimated as weighted maximum likelihood estimates (WLEs, Warm, 1989) and later also in form of plausible values (Mislevy, 1991). Person parameter estimation in NEPS is described in Pohl & Carstensen (2012a), while the data available in the SUF are described in section 7.

4.3 Checking the quality of the scale

The kindergarten science test was specifically constructed to be implemented in NEPS. In order to ensure appropriate psychometric properties, the quality of the test was evaluated in pilot studies but also checked in several analyses for the data from the main study.

The responses on the subtasks of CMC items are aggregated to a polytomous variable for each CMC item. In order to justify such an aggregation, the fit of the single subtasks is checked in analyses. For this the single subtasks are separately included in a Rasch model together with the MC items and the fit of the subtasks is evaluated based on the weighted mean square error (WMNSQ), the respective t-value, point biserial correlations of the responses with total correct score and the item characteristic curve. Only if the subtasks have a satisfactory item fit, they were used to construct polytomous CMC item variables.

In MC and CMC items there are a number of distractors (incorrect response options). We investigated whether the distractors worked well, that is, whether they are chosen by the

[3] As described later, due to collapsing of categories, this interpretation does not necessarily hold for the variables in the SUF.

children with a lower general ability in science more often than by those with a higher general ability in science. For this we evaluated the point biserial correlation of giving a certain incorrect response and the total number correct score estimated in the analysis treating all subtasks of CMC items as single items. We judged correlations below zero as very good, correlations below 0.05 as acceptable and correlations above 0.05 as problematic.

Item fit was then evaluated for the MC items and the polytomous CMC items based on results of a partial credit model. Again the weighted mean square error (WMNSQ), the respective t-value, correlations of 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 items 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 with the total 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.

We aim at constructing a science literacy test that measures the same construct for all children. If there are items that favor certain subgroups (e.g., that are easier for boys than for girls), measurement invariance would be violated and a comparison of literacy scores between the subgroups (e.g., males and females) would be biased and, thus, unfair. Test fairness was investigated for the variables test position, gender, the number of books at home (as a proxy for socio-economic status), and migration background (see Pohl & Carstensen, 2012a, for a description of these variables). In order to test for measurement invariance, differential item functioning is estimated using a multi-group IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty are estimated. Differences in the estimated item difficulties between the subgroups are evaluated. Based on experiences with preliminary data, we consider absolute differences in estimated difficulties that are greater than 1 logit as very strong DIF (differential item functioning), absolute differences between .6 and 1 noteworthy to further investigate, and differences smaller than .4 as no considerable DIF. Additionally model fit was investigated by comparing a model including differential item functioning to a model that only includes main effects and no DIF.

The competence data in NEPS are scaled using the partial credit model (1PL), in which Rasch criterion of item homogeneity is assumed. The partial credit model was chosen because it preserves the weighting of the different aspects of the framework intended by the test developers (Pohl & Carstensen, 2012a). Nevertheless, Rasch-homogeneity is an assumption that may not hold for empirical data. We therefore checked for deviations from a uniform discrimination by estimating item discrimination with the generalized partial credit model (2PL) (Muraki, 1992) using the software mdltm (von Davier, 2005), and by comparing model fit indices of the 2PL model to those obtained when applying the partial credit model.

The science test is constructed to measure a unidimensional science literacy score (Hahn et al., 2012). The assumption of unidimensionality was, nevertheless, tested in the data by specifying a two dimensional model with KAS items representing one and KOS the other dimension. The correlation between the subdimensions as well as differences in model fit between the unidimensional model and the two dimensional model were used to evaluate the unidimensionality of the scale.

5. Results

5.1 Descriptive statistics of the responses

In order to a) get a first rough descriptive measure of item difficulty and b) check for possible estimation problems, before performing IRT-analyses we evaluated the relative frequency of the responses given. Regarding each subtask of CMC item as a single variable, the percent of persons correctly responding to an item (relative to all valid responses) varies over items from 30.4% to 89.9% (regarding only the simple MC items the correct response probabilities vary from 31.3% to 89.9) On average the rate of correct responses for an item is 65.04% (SD=17.26%). From a descriptive point of view, the items cover a relatively wide range of difficulties. Overall there are no very difficult items and the majority of items show a low difficulty.

5.2 Missing responses

5.2.1 Missing responses per person

The number of invalid responses per person is shown in Figure 1. As the test was administered in a one to one testing situation the number of invalid responses is very small. For 93.6 % of the persons all responses were valid. The same is true for omitted items since hardly any child skipped an item. All missing responses could be categorized as invalid or omitted. Hence, no not-determinable missing responses occurred.

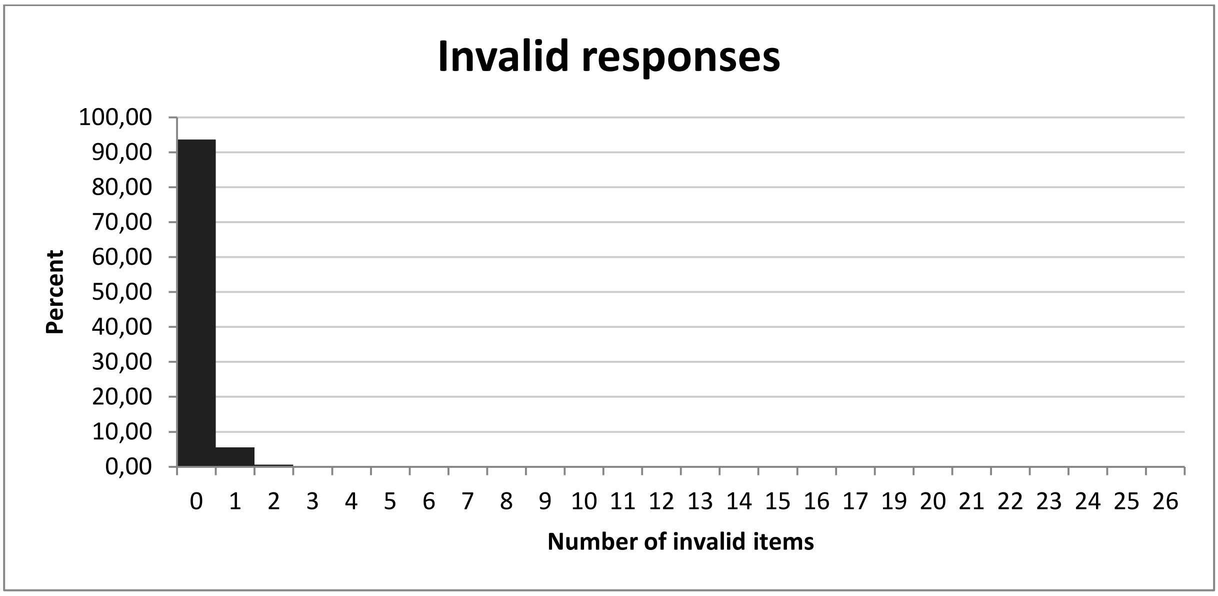

Figure 1: Number of invalid responses

The number of omitted responses per person is depicted in Figure 2. 85.5 percent of the subjects did not omit any of the items.

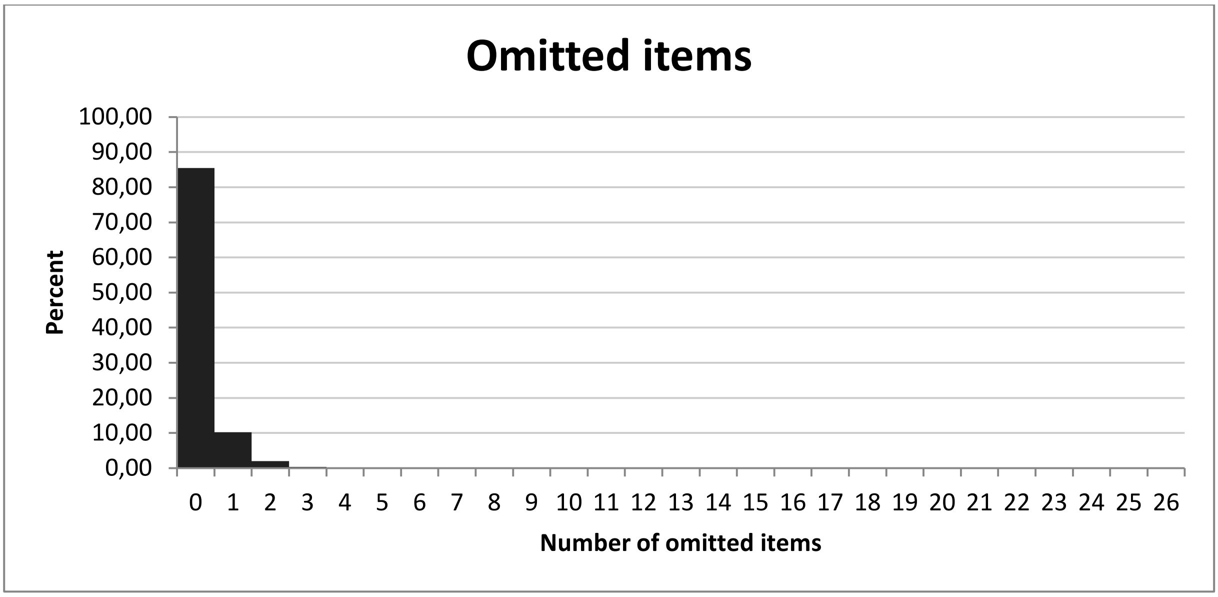

Figure 2: Number of omitted items

Figure 3 shows the total number of missing responses per person. The total number of missing responses is the sum of invalid and omitted responses. Figure 3 shows that for 81.2% of the children no missing responses occured.

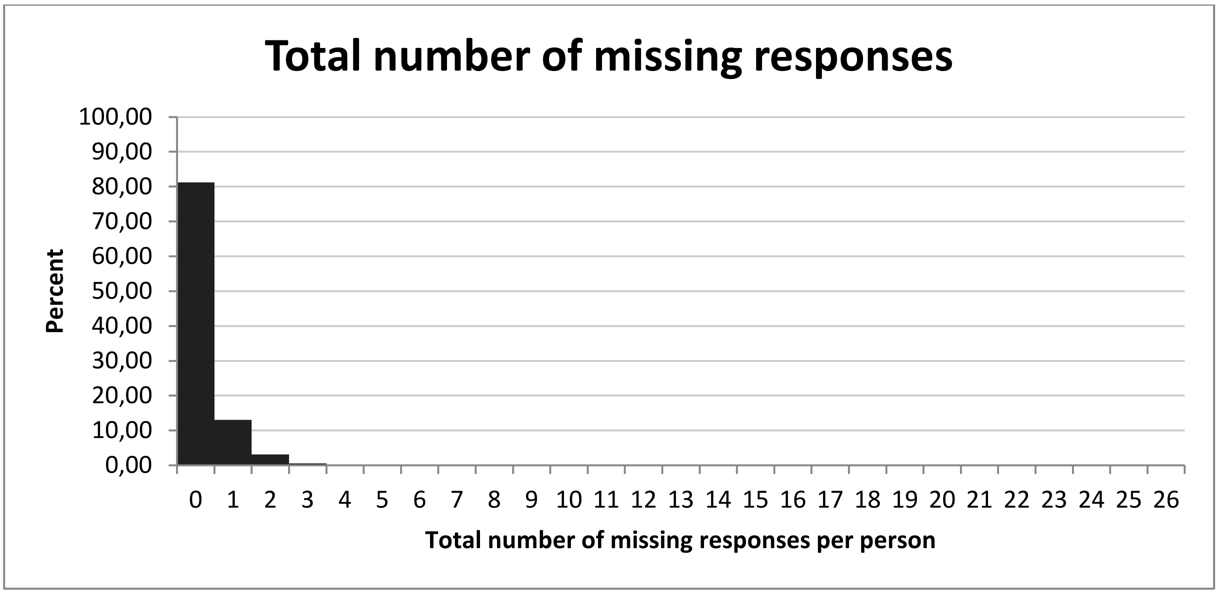

Figure 3: Total number of missing responses

5.2.2 Missing responses per item

Table 4 shows the number of valid responses for each item, as well as the number and percentage of missing responses. Overall, the number of persons that omit an item is small. There is only one item with an omission rate above 5% (sck16220_c: 5.18% of the persons omitted this item). This is probably due to the fact that the item was positioned right at the very end of the test. The number of missing responses is correlated to .46 with the difficulty of the item. This result indicates that the test takers tend to omit items that are more difficult. The number of invalid responses per item is small. The highest number is 1.76% for item sck1023s_c. The total number of missing responses per item varies between 0.37% and 5.18%.

Table 4: missing values

Item	Number of valid responses	Position in the test	Relative frequency of not-reached items in %	Relative frequency of omitted items in %	Relative frequency of invalid responses in %
sck10420_c	2954	1	0.03	0.88	0.91
sck10430_c	2944	2	0.37	0.78	1.16
sck16120_c	2955	3	0.00	0.37	0.37
sck16130_c	2955	4	0.00	0.61	0.61
sck1102s_c	2920	5	1.18	0.71	1.94
sck11030_c	2955	6	0.00	1.05	1.05
sck10320_c	2955	7	0.00	3.86	3.86
sck1033s_c	2910	8	1.52	0.88	2.45
sck10210_c	2953	9	0.07	1.15	1.22
sck1023s_c	2903	10	1.76	0.95	2.77
sck11110_c	2954	11	0.03	1.22	1.25
sck11120_c	2954	12	0.03	1.25	1.29
sck16010_c	2954	13	0.03	1.83	1.86
sck16020_c	2953	14	0.07	1.73	1.80
sck10510_c	2954	15	0.03	1.08	1.12
sck10530_c	2954	16	0.03	1.49	1.52
sck11610_c	2954	17	0.03	1.66	1.69
sck1162s_c	2908	18	1.59	1.56	3.20
sck10710_c	2951	19	0.14	1.73	1.87
sck10720_c	2955	20	0.00	1.90	1.90
sck11310_c	2955	21	0.00	4.40	4.40
sck11330_c	2946	22	0.30	3.21	3.53
sck10910_c	2955	23	0.00	2.27	2.27
sck10920_c	2955	24	0.00	2.27	2.27
sck16210_c	2955	25	0.00	2.20	2.20
sck1622b_c	2955	26	0.00	5.18	5.18

Table 5: Item parameters

Item	Difficulty/ location parameter	SE (difficulty/ location parameter)	Weighted MNSQ	Weight t-value	Pt.bis of correct response	Discrimination (2PL)
sck10420_c	-1.481	0.048	1.00	0.2	0.38	1.05
sck10430_c	-0.515	0.041	0.99	-0.9	0.45	1.10
sck16120_c	-1.749	0.051	0.98	-0.6	0.38	1.14
sck16130_c	-0.568	0.041	0.98	-1.0	0.45	1.12
sck1102s_c	-1.337	0.038	0.96	-1.7	0.48	0.63
sck11030_c	0.026	0.040	1.07	4.8	0.35	0.71
sck1033s_c	-2.061	0.044	0.89	-4.8	0.54	2.40
sck10210_c	-2.369	0.061	0.90	-2.4	0.44	0.48
sck1023s_c	-2.435	0.041	1.01	0.4	0.41	0.52
sck11110_c	-1.181	0.045	1.09	3.9	0.29	0.41
sck11120_c	-0.185	0.040	1.13	8.9	0.27	0.85
sck16010_c	-0.935	0.043	1.04	1.9	0.37	1.13
sck16020_c	0.758	0.042	0.97	-1.8	0.44	0.89
sck10510_c	-1.487	0.048	1.02	0.8	0.35	0.90
sck10530_c	-0.717	0.042	1.04	2.0	0.38	1.62
sck11610_c	-2.473	0.064	0.95	-1.1	0.36	0.55
sck1162s_c	-0.652	0.046	0.99	-0.5	0.39	1.21
sck10710_c	-1.683	0.050	0.97	-0.9	0.39	1.36
sck10720_c	-0.703	0.042	0.95	-3.3	0.49	0.36
sck11310_c	0.926	0.044	1.14	7.1	0.23	1.04
sck11330_c	-0.931	0.044	1.00	-0.1	0.41	0.97
sck10910_c	-0.270	0.041	1.01	0.6	0.42	1.44
sck10920_c	-1.994	0.055	0.96	-1.2	0.39	1.51
sck16210_c	-0.388	0.041	0.92	-5.3	0.52	0.96
sck1622b_c	0.923	0.044	1.01	0.5	0.39	0.75

5.3 Parameter estimates

5.3.1 Item parameters

Only 25 out of the 26 items had good fit values and were included in the analyses. Item sck10320_c had a weighted mean square of 1.34, a weighted t-value of 10.7 and a negative point biserial correlation. Also the item characteristic curve supported the misfit of item sck10320_c and consequently it was excluded from further analyses.

A few subtasks of the polytomous items had a discrimination that was clearly below 0.2. However, as the low values for certain subtasks did not have a markedly negative effect on the overall fit of the items, these subtasks were not excluded from the analyses.

The estimated item difficulties for polytomous variables (CMC items) and location parameters for dichotomous variables (MC and SCR items) are listed in in table 4. The step parameters (for polytomous variables) are depicted in table 6. The item difficulties were estimated by constraining the mean of the ability distribution to zero. The estimated item difficulties (or location parameters for polytomous variables) vary between -2.18 (item sck1023s_c) and 0.90 (item sck16220_c) with a mean of -0.95. Overall, the item difficulties are low and particularly difficult items are absent. Due to the large sample size, the standard error of the estimated item difficulties is very small (SE(ß) ≤ 0.06).

Table 6: Step parameters for the CMC items

Item	Step 1 (SE)	Step 2 (SE)	Step 3 (SE)	Step 4 (SE)
sck1102s_c	-2.20 (0.05)	-0.05 (0.04)	1.42 (0.05)	0.84
sck1033s_c	-1.27 (0.04)	-1.23 (0.04)	1.16 (0.04)	1.33
sck1023s_c	-1.35 (0.04)	-0.52 (0.04)	1.04 (0.04)	0.83
sck1162s_c	-2.09 (0.07)	-0.79 (0.04)	1.00 (0.04)	1.88

5.3.2 Person parameters

Person parameters are estimated as WLEs and plausible values (Pohl & Carstensen, 2012a). WLEs will be provided in the first release of the SUF. Plausible values will be provided in later analyses. A description of the data in the SUF can be found in section 7. An overview of how to work with competence data is given in Pohl and Carstensen (2012a).

5.4 Test Targeting and Reliability

Test targeting was investigated in order to evaluate the measurement precision of the estimated ability scores and to judge the suitability of the test for the specific target population. In the analyses, the mean ability is constrained to be zero. The variance was estimated to be 0.80, indicating that the test has reasonable potential to differentiate between children of different science abilities. The reliability of the test (WLE reliability = .751) is good.

The amount to which the item difficulties and location parameters are targeted to the ability of the persons is shown in Figure 4. The Figure shows that the items cover a great range of the ability distribution. However, only few items cover medium person ability and there are no items available for persons with high science ability. Instead the majority of items are easy or of medium difficulty. As a consequence, persons with a medium and low ability will be measured relatively precisely with a low standard error while ability estimates for children with high science ability will have a larger standard error.

scale logits	person ability	item difficulty
		each "x" represents 22.3 cases
3		
	X	
	X	
	X	
2		
	XX	
	XX	
	XXX	
	XXX	
	XXXX	
1	XXXX	
	XXXX	20 25
	XXXXX	13
	XXXXXX	
	XXXXX	
	XXXXX	
	XXXXX	
	XXXXXX	
0	XXXXXX	6
	XXXXX	
	XXXXX	11
	XXXXXXXXXX	22
	XXXXX	24
	XXXX	2 4
	XXXXX	17 19
	XXXXXXX	15
	XXX	12 21
-1	XXXX	
	XXX	10
	XXXX	
	XX	5
	XX	1 14
	XX	
	XX	3 18
	X	
-2	XX	23
	X	7
		8
		9 16
-3		

Figure 4: Test targeting. The distribution of person ability in the sample is depicted on the left side of the graph. Each 'X' represents 22.3 cases. The difficulty of the items is depicted on the right side of the graph. Each number represents an item (see table 1).

5.5 Quality of the test

5.5.1 Fit of the subtasks of complex multiple-choice items

Before the responses on the subtasks of CMC items were aggregated and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the single subtasks together with the simple MC and SCR items in a Rasch model.

All subtasks showed a satisfactory item fit. For the items that entered the analyses, the WMNSQ ranged from 0.92 to 1.34, the respective t-value from -6.0 to 7.6, and there were no noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves except for the items that were eventually excluded from the analyses. Due to the good model fit of the subtasks their aggregation to polytomous variables seem to be justified.

5.5.2 Distractor analyses

In addition to the overall item fit, we specifically investigated how well the distractors performed in the test by evaluating the point biserial correlation between each incorrect response (distractor) and the children's' total score. All distractors had a point biserial correlation with the total score below zero. The results indicate that the distractors work well.

5.5.3 Item fit

The item fit can be regarded as very good as WMNSQs are close to 1 with the lowest value being 0.92 (item sck16210_c) and the highest 1.11 (item sck11120_c). All item characteristic curves showed a reasonable or good fit. Overall, the correlations of the item score with the total score vary between .23 and .54 with an average correlation of .37.

5.5.4 Differential item functioning

We checked the test fairness for different groups (i.e., measurement invariance) by estimating the amount of differential item functioning (DIF). Differential item functioning was investigated for gender. As we did not have any information on the migration status of the children, we used information provided by the kindergarten teachers as to how well the child was able to understand and speak German (two items) and whether the main language at home was German (one item) as indicators for the child's migration status.

Table 8 shows the difference between the estimated item difficulties in different groups. Female vs. male, for example, indicates the difference in difficulty ß(male) – ß(female). A positive value indicates a higher difficulty for males, a negative value a lower difficulty for males as opposed to females. 1,432 (48.6%) of the test takers were female and 1,501 (50.9%) were male. There were 14 missing responses on the variable gender. These cases were excluded from the analysis. On average, boys have a slightly (but not significantly) lower science score than girls (main effect = -0.039 logits, Cohen's d = -.041). However, there is no item with a considerable gender DIF. The highest difference in difficulties between the two groups is .307 logits.

There were 476 (16.2%) children who, according to the kindergarten teacher, did not speak German at home. For 2,464 (83.6%) children German was stated to be the language spoken at home and for 7 (0.2%) children no information was provided. On average there were large differences in scientific literacy between children who spoke German at home and those who did not (Cohen’s d = -1.24).

483 (16.4%) children were not able to understand German well, whereas the majority of 2,209 (75%) children were regarded as competent in terms of understanding German. For 255 (8.7%) no information in this category was provided. On average there were large differences in scientific literacy between children who understood German well and those who did not (Cohen’s d = -1.26).

202 (6.9%) children were not able to speak German reasonably well, whereas the majority of 2,497 (84.7%) children were regarded as competent in terms of speaking German. For 248 children (8.4%) no information in this category was provided. On average there were large differences in scientific literacy between children who spoke German well and those who did not (Cohen’s d =-1.34).

However, most of the differences in item difficulties estimated via the DIF-analyses (see Table 8) are in absolute values below 0.5. The comparison of a model that allows for DIF to a model without modeling DIF supports these findings. Overall, the results indicate that the test is reasonably fair for the considered groups.

Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF with those that allow only for main effects. In Table 7 the models including only main effects are compared with those that additionally estimate DIF. Two different information criteria were used in the analyses: The Akaike's (1974) information criterion (AIC) favors the models estimating DIF for all four DIF variables. The Bayesian information criterion (BIC, Schwarz, 1978) takes the number of estimated parameters into account and, thus, prevents from overparameterization of models. Using BIC, the more complex models (including the main effect and the diff effect) have a slightly better fit for three out of the four variables than the more parsimonious model.(table 7). However, as the differences between the models are small and no item had a diff larger than 0.5 we decided to use the more parsimonious model for estimating WLEs.

Table 7: Comparison of models with and without DIF

	N	Number of parameters	BIC	AIC
Gender main effect	2933	39	94893.907	98501.103
Gender DIF	2933	64	94703.452	98363.514
Language spoken at home main effect	2940	39	94684.269	98312.752
Language spoken at home DIF	2940	64	94500.066	97984.445
Child understands German main effect	2692	39	86364.405	90261.958
Child understands German DIF	2692	64	86553.146	90058.030
Child speaks German main effect	2699	39	86910.038	89910.504
Child speaks German DIF	2699	64	86799.835	89522.556

Table 8: Differential Item functioning

Item	Male vs. female	Language spoken at home is not German vs German spoken at home	Child understands German poorly vs well	Child speaks German poorly vs well
sck10420_c	-0.020	-0.015	-0.049	-0.088
sck10430_c	-0.031	0.037	0.016	-0.050
sck16120_c	0.127	-0.053	-0.001	-0.084
sck16130_c	-0.032	0.096	0.022	0.050
sck1102s_c	-0.037	0.084	0.074	0.052
sck11030_c	0.006	-0.118	-0.228	-0.286
sck1033s_c	0.060	0.100	0.296	0.244
sck10210_c	-0.045	0.432	0.389	0.423
sck1023s_c	-0.045	0.002	0.031	0.013
sck11110_c	0.255	-0.220	-0.139	-0.090
sck11120_c	0.271	-0.216	-0.244	-0.264
sck16010_c	0.021	0.009	-0.037	-0.109
sck16020_c	-0.105	-0.023	-0.144	-0.049
sck10510_c	-0.038	-0.088	-0.145	-0.169
sck10530_c	0.097	-0.209	0.001	-0.103
sck11610_c	-0.008	0.337	0.255	0.219
sck1162s_c	-0.107	-0.061	-0.099	-0.075
sck10710_c	-0.104	0.143	0.191	0.274
sck10720_c	0.118	0.118	0.131	0.209
sck11310_c	-0.313	-0.368	-0.399	-0.392
sck11330_c	-0.191	0.056	0.002	-0.008
sck10910_c	0.069	-0.017	0.057	0.058
sck10920_c	-0.021	-0.043	0.012	-0.028
sck16210_c	0.024	0.049	0.119	0.114
sck1622b_c	-0.014	0.246	0.122	0.258

5.5.5 Rasch-homogeneity

In order to test for the assumption of Rasch-homogeneity we used the 25 items that met the selection criteria for estimating the person parameters. The subtasks of CMC items were aggregated according to the procedure described in chapter 4.2. A generalized partial credit model (2PL) was used to test for Rasch-homogeneity. The estimated discrimination parameters are depicted in table 4. They range from 0.36 (item sck10720_c) to 2.4 (item sck1033s_c). The discriminations differ considerably among the items and the 2PL model (BIC = 95170.67, number of parameters = 75) fits the data slightly better than the partial credit model (1PL) (BIC = 95449.39, number of parameters = 38). Nevertheless, the theoretical aim was to construct a test that equally represents the different aspects of the framework (see Pohl & Carstensen, 2012a, 2012b, for a discussion of this issue), and thus the partial credit model was used to preserve the item weightings intended in the constructional framework.

5.5.6 Unidimensionality of the test

The unidimensionality of the test was investigated by specifying a one- and a two-dimensional model.

The first model is based on the assumption that scientific literacy is a one-dimensional construct that measures one distinct competence whereas the second model distinguishes between the two sub-competencies knowledge about science (KAS) and knowledge of science (KOS) (for more details see Hahn et al., 2012). For estimating a two-dimensional model based on the Gauss Hermite quadrature estimation implemented in ConQuest (nodes were chosen in such a way that stable parameter estimation was obtained) was used. The two-dimensional model (BIC= 95425.99204, number of parameters = 40) fits the data better than a unidimensional model (BIC=95449.39, number of parameters =38). As the correlations of the two dimensions is 0.915 we decided to use the one-dimensional construct and consequently a single competence score for scientific literacy.

6. Discussion

The analyses in the previous sections aimed at providing information on the quality of the science test in kindergarten and at describing how the scientific literacy score is estimated.

The amount of not non-valid responses was low which was most likely due to the fact that the test was administered on a one on one basis. Some items show higher omission rates, although in general, the amount of omitted items is also low.

Indicated by various fit indices all items, except for one, show a good item fit and measurement invariance across various subgroups. The test has a good reliability. The items distinguish well between the testtakers. However, the test is mainly targeted to the low or medium performing children and does not sufficiently measure scientific literacy of high performing children. As a consequence, ability estimates will be very precise for low performing children but less precise for high performing children.

Summarizing the results, the test has good psychometric properties that facilitate the estimation of a one-dimensional scientific literacy score.

7. Data in the Scientific Use file

There are 25 items in the data set for estimating person parameters that are either scored as dichotomous variables (MC or SCR items) with 0 indicating an incorrect response and 1 indicating a correct response, or scored as a polytomous variable (CMC items) indicating the (partial) credit. The dichotomous variables are marked with a '0_c' at the end of the variable name, the CMC items are marked with a 's_c' at the end of the variable name. Note that the value of the polytomous variable does not necessarily indicate the number of correctly responded subtasks (see section 4.2 aggregation of CMC items). In the scaling model each category of CMC items is scored with 0.5 points. Manifest scale scores are provided in form of WLE estimates (sc_wle) including the respective standard error (sc_wle_se). Please note that when categories of the polytomous variables had less than N=200, the categories were collapsed. The ConQuest Syntax for estimating the WLE scores from the items is provided in Appendix A. Children that did not take part in the test or those that do not have enough valid responses to estimate a scale score will have a non-determinable missing value on the WLE score for scientific literacy.

Plausible values, that allow investigating latent relationships of competence scores with other variables, will be provided in later data releases. User interested in investigating latent relationships may alternatively 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 (2012a).

References

Akaike, H. (1974). A new look at the statistical model indentification. *IEEE Transactions on Automatic Control*, 19, 716-722.

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.

Haberkorn, K., Pohl, S., Hardt, K., & Wiegand, E. (2012a). Technical Report of Reading– Scaling Results of Starting Cohort 4 in Ninth Grade (NEPS Working Paper No. 16). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Hahn, I. Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., Dalehefte, I.M., & Prenzel, M. (2012). Assessing scientific literacy over the lifespan - A description of the NEPS science framework and the test development. Manuscript submitted for publication.

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

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177-196.

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

OECD (2007). PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*, PISA, OECD Publishing.*

Pohl, S. & Carstensen, C. H. (2012a). 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. (2012b). Scaling the competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. Manuscript submitted for publication.

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

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C., & Hamann, M. (2007). Naturwissenschaftliche Kompetenz im internationalen Vergleich. In PISA-Konsortium Deutschland (Hrsg.), PISA 2006 - Die Ergebnisse der dritten internationalen Vergleichsstudie (S. 63-105). Münster, Waxmann.

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 (ETS Research Rep. 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 & J. v. Maurice & (Eds.). Education as a Lifelong Process: The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft, Sonderheft 14* . 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 WLE estimates in starting cohort 2

```
Title Starting Cohort 2, SCIENCE: Partial Credit Model;

data filename.dat;
format pid 4-10 responses* /* insert number of columns with data*/

codes 0,1,2,3,4;

score (0,1)           (0,1)                  !item (1-4,6,8,10-16,18-25);
score (0,1,2,3,4)     (0,0.5,1,1.5,2)        !item (5,7,9,17);

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

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