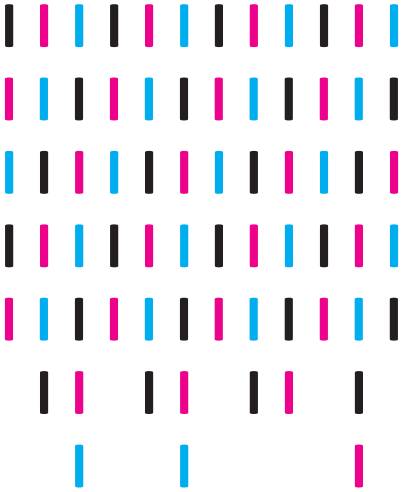

NEPS *SURVEY PAPERS*

Timo Gnambs

NEPS TECHNICAL REPORT FOR EARLY READING COMPETENCE: SCALING RESULTS OF STARTING COHORT 1 FOR EIGHT-YEAR-OLD CHILDREN (WAVE 9)

NEPS *Survey Paper* No. 96
Bamberg, June 2022

LIfBi
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

The NEPS *Survey Paper* series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

They are of particular relevance for the analysis of NEPS data as they describe data editing and data collection procedures as well as instruments or tests used in the NEPS survey. Papers that appear in this series fall into the category of 'grey literature' and may also appear elsewhere.

The NEPS *Survey Papers* are edited by a review board consisting of the scientific management of LIfBi and NEPS.

The NEPS *Survey Papers* are available at www.neps-data.de (see section “Publications“) and at www.lifbi.de/publications.

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

NEPS Technical Report for Early Reading Competence: Scaling Results of Starting Cohort 1 for Eight-Year-Old Children (Wave 9)

Timo Gnambs

Leibniz Institute for Educational Trajectories, Bamberg, Germany

E-mail address of the lead author:

timo.gnambs@lifbi.de

Bibliographic data:

Gnambs, T. (2022). *NEPS Technical Report for Early Reading Competence: Scaling Results of Starting Cohort 1 for Eight-Year-Old Children (Wave 9)* (NEPS Survey Paper No. 96). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP96:1.0

Acknowledgments:

Various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (e.g., Gnambs, 2020) to facilitate the understanding of the presented results.

NEPS Technical Report for Early Reading Competence: Scaling Results of Starting Cohort 1 for Eight-Year-Old Children (Wave 9)

Abstract

The National Educational Panel Study (NEPS) examines the development of competencies across the life span. Therefore, the NEPS develops tests for the assessment of various competence domains in different age cohorts. To evaluate the quality of these competence tests, several analyses based on item response theory (IRT) are performed. This paper describes the data and scaling procedures for an early reading competence test that was administered in Wave 9 of Starting Cohort 1 (newborns) to eight-year-old children. The early reading competence test included 26 items with multiple choice response formats that were administered as a proctored web-based test. The test was administered to a total of 1,588 children (50% girls). The responses of the children were scaled using a unidimensional Rasch model. Item fit statistics and differential item functioning were evaluated to ensure the quality of the test. These analyses showed that the test differentiated well between children, exhibited good reliability, and showed a satisfactory fit to the item response model. Furthermore, comparable measurement models could be confirmed for different subgroups. Limitations of the test pertained to a large number of missing values because many children were unable to finish the test in the available testing time. Overall, the early reading competence test had good psychometric properties that allowed for an estimation of reliable competence scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the R syntax for scaling the data.

Keywords

item response theory, scaling, early reading competence, remote testing, scientific use file

Content

1 Introduction

Within the National Educational Panel Study (NEPS) different competencies are measured coherently across the life span. These include, among others, reading competence, mathematical competence, scientific literacy, information and communication technologies literacy, metacognition, vocabulary, and domain-general cognitive functioning. An overview of the competencies measured in the NEPS is given by Weinert and colleagues (2011) as well as Fuß, Gnambs, Lockl, and Attig (2021). Most of the administered competence tests are developed specifically for implementation in the NEPS and, thus, are routinely evaluated using psychometric models based on item response theory (IRT; see Pohl & Carstensen, 2012).

In this paper, the psychometric properties of a commercial test ("ELFE II"; Lenhard, Lenhard, & Schneider, 2018) published by Hogrefe are summarized that measured early reading competence in Wave 9 of Starting Cohort 1 (newborns). In the following sections, the administered test of early reading competence and key aspects of the test design are introduced. Then, the sample and the psychometric analyses performed to check the quality of the test are described. Finally, an overview of the data that is available for public use in the scientific use file (SUF) is presented.

Please note that the analyses summarized in this report are based on the data available at some time before the public data release. Due to ongoing data protection and data cleansing issues, the data in the SUF may differ slightly from the data used for the analyses in this report. However, we do not expect pronounced differences in the presented results.

2 Testing Early Reading Competence

2.1 Construction Rationale and Test Design

Early reading competence was measured with the "ELFE II: Ein Leseverständnistest für Erst- bis Siebtklässler – Version II" (Lenhard et al., 2018) that is distributed by Hogrefe. The test measures children's reading comprehension of short texts and, thus, the ability to integrate information contained in single words and sentences into a coherent overall picture of the text. It captures a deductive reading proficiency that allows children to combine singular pieces of reading information, develop a mental model of the text, and draw further inferences that supplement or continue the information presented in the text. Further information on the theoretical background guiding the test development is given in Lenhard et al. (2018).

The test presented several short texts (including about two to eight sentences) that were accompanied by one to three items. Each multiple-choice item included four response options with one being correct and three response options functioning as distractors (i.e., they were incorrect). The item development was guided by a framework that specified three independent factors (see Lenhard et al., 2018). The text addressed by each item presented either a fictional or a non-fictional topic (factor genre: non-fiction versus fiction) that required retrieving a literal piece of information or drawing an analogy from the presented information (factor information: literal versus analogous). Moreover, each item required either drawing connections between neighboring sentences or between multiple sentences (factor coherence: local versus global). The items covered all combinations of the three factors to measure a unidimensional competence score. There was no multi-matrix design regarding the

order of the items; thus, all respondents received the test items in the same order. The items were roughly ordered by their estimated difficulty with easier items at the beginning of the test and more difficult items at the end of the test. The testing time was limited to 7 minutes after which the test was automatically terminated.

2.2 Assessment Procedure

The study was conducted in summer 2020 and assessed different competence domains including reading speed, early reading competence, and mathematical competence (cf. Petersen, Beyer, & Bednorz, 2022). The test for early reading competence was always presented second after the test for reading speed. There was no rotation design, thus, all children received the tests in the same order. A detailed description of the study design is available on the NEPS website (http://www.neps-data.de).

Originally, the test was supposed to be administered as a proctored computerized test (CBT) by test administrators visiting the children at their private homes and presenting the test on a dedicated tablet (comparably to previous assessments in Starting Cohort 1). However, due to the rise of the COVID-19 pandemic the administration mode had to be changed at short notice and was switched to a proctored web-based format (WBT). Here, the test administrators accompanied the computer-based testing via phone. The results reported in this technical report refer only to the students who were tested via WBT administration.

A couple of weeks before the test date a telephone interview was conducted with a parent to discuss the necessary computer equipment in the household that would allow the child to take the WBT. Although tablet devices were preferred (to keep as comparable as possible to the previous assessments), laptops with a minimum screen size were allowed as alternative assessment devices. At a prearranged test date and time, a trained test administrator called the parent by phone to assist in setting up the tablet or laptop (e.g., positioning the device on the table) and starting the web-based test (e.g., opening the browser, entering the correct link and password). Then, the children worked alone on the WBT. During the test administration, the test administrators supervised the child's progress on the test remotely using a dashboard that showed in real time the test page a child was currently visiting. Assistance and verbal support to the children were provided by phone. Thus, the test administrators had a continuous means of communication with the children during the entire test procedure. Although the test administrators could not directly see the child or the specific testing conditions such as the room a child was occupying or whether other people were present during the assessment, they could monitor the child's progress in the test, listen to voiced problems or background noise, and talk to the children. By design, direct assistance through test administrators was rarely required because the web-based test used standardized video instructions that introduced the different tests with prerecorded demonstrations and, thus, allowed a high level of standardization. The role of the test administrators was primarily limited to assisting in starting the test, motivating children between different tests, and helping with unforeseen problems during the test.

3 Data

From a total of 1,588[1] children that were administered the test 45 children were excluded because they had less than three valid responses on the early reading competence test (cf. Pohl & Carstensen, 2012) or serious problems were observed during the test administration (e.g., interference by a parent, lack of experience in using a mouse, technical errors) that invalidated the test scores. Moreover, 84 children were excluded from the psychometric analyses because of diagnosed special educational needs, dyslexia, or an attention deficit hyperactivity disorder. Because test performance might be influenced by the used computer system, only children working on a tablet or a laptop with a mouse were considered for the analyses. Therefore, 73 further children were excluded that worked on the test using a laptop with touch functionality or a touchpad. This resulted in an analysis sample of N = 1,386 (51% girls) with an average age of M = 8.26 years (SD = 0.12). About 11% of the children had a migration background, that is, at least one parent born abroad, but 19% of the sample reported using an interaction language at home other than German. Basic sociodemographic information of the children split by the used computer device is summarized in Table 1.

Table 1.

Sample Descriptions

	Administration device	
	Tablet	**Laptop**
Sample size	1,176	210
Girls	51%	49%
Migration background	11%	11%
Non-German interaction language at home	18%	21%
Mean age (*SD*)	8.26 (0.12)	8.26 (0.13)
Attended grade 2	89%	91%
Highest parental International Socioeconomic Index	69.68 (15.33)	68.28 (16.29)

[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.

About 89% of the children attended Grade 2, while about 10% were in Grade 3 (see Table 2). The remaining children went to first grade (1%) and one child attended the last month of school entry education before starting in school. Most children (96%) were tested during or shortly before the holiday season about one to three months before starting the new school year, while the remaining children were tested in the first month of the school year (see Table 2).

Table 2.

Number of Children by School Month

School month	School year 2019/20	School year 2020/21	Current grade [a]			
			0 [b]	1	2	3
1	0	59	0	0	11	48
10	178	0	0	2	170	6
11	636	0	0	6	598	31
12	513	0	1	7	457	48
Total	1,327	59	1	15	1236	133

Note. The school month does not refer to the month of the year, but the number of months since the beginning of the current school year (see Lenhard et al., 2018). Because the beginning of the school year differs between the German federal states, the same school month might refer to different months of the year. The beginning of the school year in each state was determined from https://www.schulferien.org/deutschland/ferien. [a] For one child in school month 11, no information on the grade was available; [b] School entry-level ("*Schuleingangsstufe*").

4 Psychometric Analyses

4.1 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) omitted items, b) items that test-takers did not reach, and c) technical difficulties. Omitted items occurred when test-takers skipped some items. Because of the time limit, not all persons finished the test. All missing responses after the last valid response were coded as not reached. Because the test was administered on the private computers of the children, unforeseen technical errors might have prevented the correct presentation of some items or the whole test. If the test had to be prematurely terminated by the test administrator, missing values for these items that were not administered were coded as a technical error. In case, the entire test could not be administered and, thus, no valid response was observed the test was considered as not administered. Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions). Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the children

were coping with the test. Missing responses per item were examined to evaluate how well each of the items functioned.

4.2 Scaling Model

The test manual (Lenhard et al., 2018) recommends calculating sum scores across the 26 items of the test treating missing values as incorrect. To reflect this scoring rule, the item and person parameters were estimated using a Rasch (1960) model with Gauss-Hermite quadrature (21 nodes). Omitted items and items that were not reached due to the time limit were scored as incorrect. In contrast, missing values resulting from a technical error leading to premature termination of the test were treated as missing values in the scaling model and, thus, did not contribute to the parameter estimation. Early reading competencies were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989).

4.3 Checking the Quality of the Test

The early reading competence test was thoroughly validated in several developmental samples. Details on the test construction and the psychometric properties of the test in these studies are given in Lenhard et al. (2018). To ensure appropriate psychometric properties in the present sample, several additional analyses were conducted to evaluate the quality of the test scores provided in the SUF.

The multiple-choice items consisted of one correct response option and three distractors (i.e., incorrect response options). The quality of the distractors within the multiple-choice items was examined using the point-biserial correlation between selecting an incorrect response option for a given item and the total correct score for the remaining items. Negative correlations indicate good distractors, whereas correlations between .00 and .05 were considered acceptable and correlations above .05 were viewed as problematic distractors (Pohl & Carstensen, 2012).

The fit of the dichotomous items to the Rasch (1960) model was evaluated using the weighted mean square (WMNSQ) statistic, the respective *t*-value, and a visual inspection of the item characteristic curves (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 > 1.20 (*t*-value > |8|) were judged as having a considerable item misfit. The overall judgment of the fit of an item was based on all fit indicators.

The early reading competence test should measure the same construct for all children. If some items favored certain subgroups (e.g., they were easier for boys than for girls), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., boys and girls) would be biased. For the present study, measurement invariance was investigated for the variables sex, highest parental international socioeconomic index (Ganzeboom, 2010; as a proxy for socioeconomic status), migration background, and administration device (i.e., tablet or laptop). Differential item functioning (DIF) was examined using a multigroup item response 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 DIF and noteworthy of further investigation,

differences between 0.4 and 0.6 as small but not severe DIF, and differences smaller than 0.4 as negligible DIF. Moreover, we report these differences also in a Cohen's *d*-like metric by dividing them by the population standard deviation. Additionally, an overall test for DIF using information criteria was conducted by comparing the fit of a model including DIF to a model that only included main effects and no DIF.

The early reading competence test was scaled using the Rasch (1960) model because it preserves the equal weighting of the test items as reflected in the scoring rule recommended by the test developers (Lenhard et al., 2018). Nonetheless, Rasch-homogeneity is an assumption that might not hold for empirical data. To test the assumption of equal item discrimination parameters, a two-parametric item-response model (2PL; Birnbaum, 1968) was also fitted to the data and compared to the Rasch model.

The dimensionality of the test was evaluated by examining the residuals of the Rasch model. Approximately zero-order correlations as indicated by Yen's (1984) Q_3 indicate essential unidimensionality. Because in the case of locally independent items, the Q_3 statistic tends to be slightly negative, we report the adjusted Q_3 (aQ_3) that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) values of aQ_3 falling below .20 indicate essential unidimensionality.

4.4 Software

The item response models were estimated with the *TAM* package version 3.7-16 (Robitzsch, Kiefer, & Wu, 2021) in *R* version 4.1.1 (R Core Team, 2021).

5 Results

5.1 Missing Responses

5.1.1 Missing responses per person

Omitted responses were extremely rare with less than 0.2% of the children skipping an item (see Figure 1). Similarly, missing values resulting from a premature test termination because of technical difficulties were observed for less than 0.5% of the sample. This indicates that for most children the test functioned as intended. In contrast, missing responses because items were not reached due to the time limit were substantially more prevalent. These missing values refer to items after the last valid response. As illustrated in Figure 2, less than 5% of the children finished the test and were administered all 26 items. About 50% of the sample received about half of the test, while 9 or more items were reached by 80% of the sample. This might indicate that the test was too difficult for the limited testing time. The results given in Figure 2 also show that there were no substantial differences in missing rates between children working on a tablet or a laptop.

With an item's progressing position in the test, the number of children that did not reach an item rose to about 96%. For both devices, the last items were reached by only a few children. As illustrated in Figure 3, children working on a laptop tended to reach slightly more items of the test as compared to children using a tablet. Thus, it seems that many children were unable to finish the test within the allocated time. This indicates that the testing time might have been too short for the difficulty of the administered test.

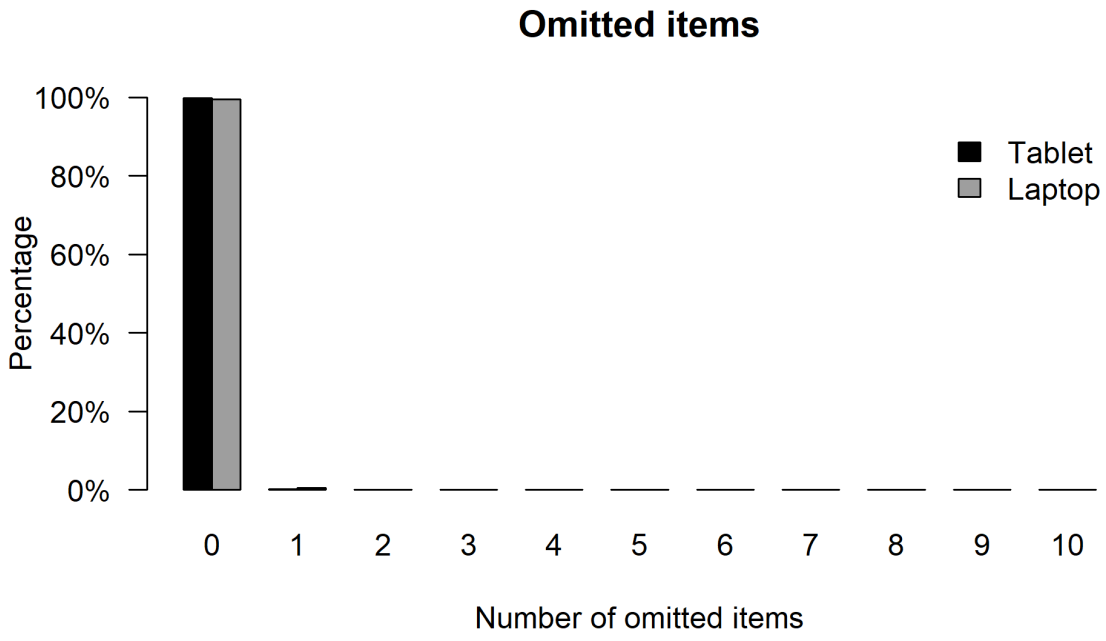

Figure 1. Number of omitted items by device

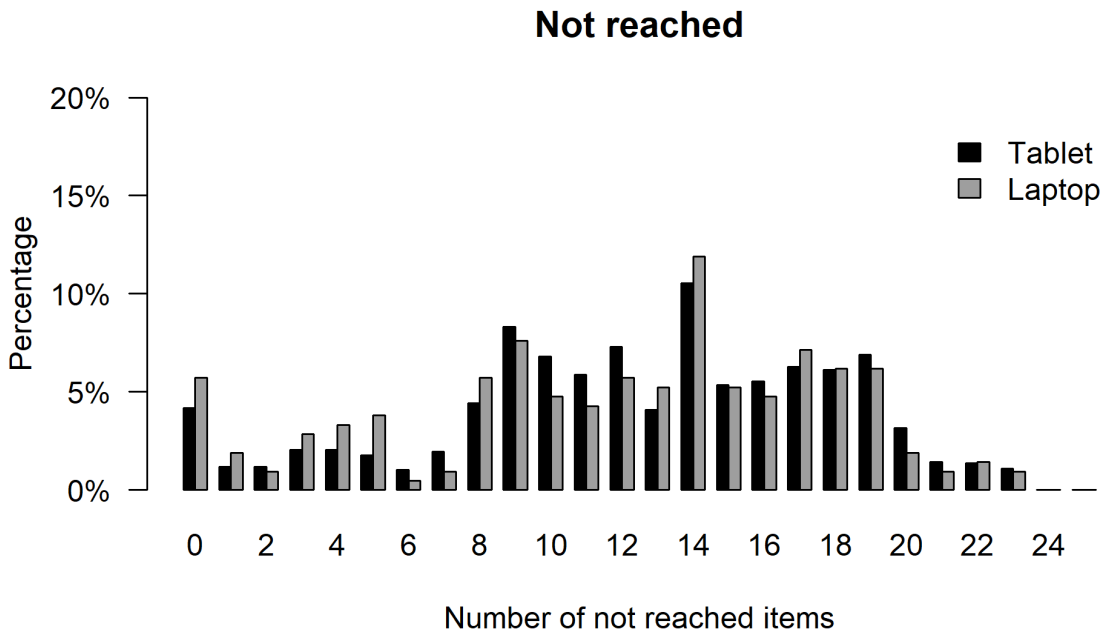

Figure 2. Number of not reached items by device

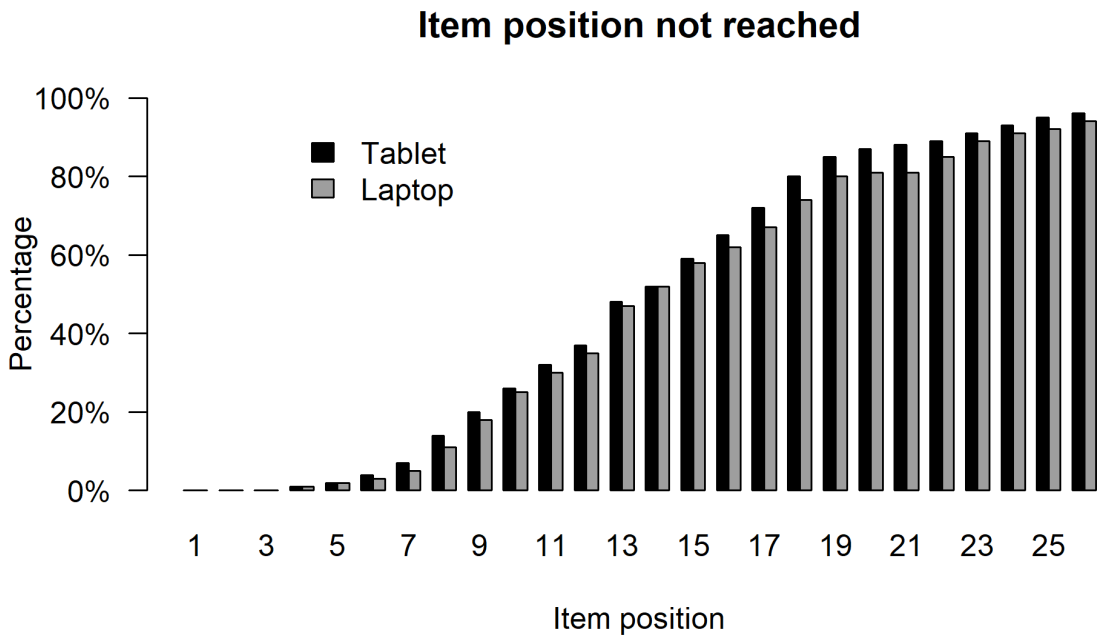

Figure 3. Item position not reached by device

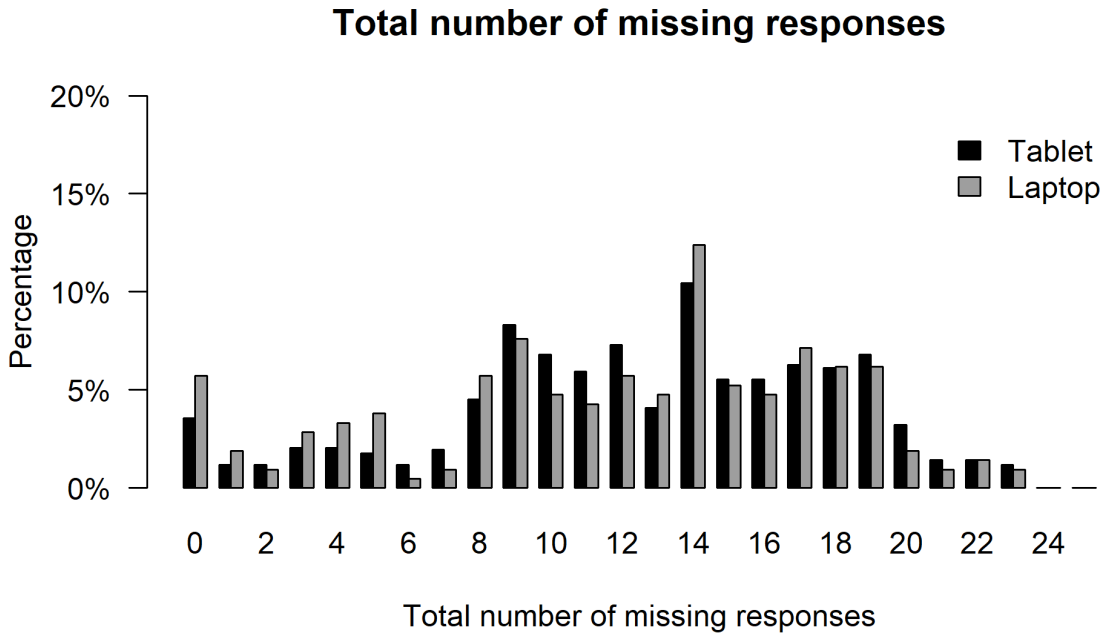

Figure 4. Total number of missing responses by device

Table 3.

Percentage of Missing Values by Item.

		Tablet			Laptop		
Pos.	**Item**	***N***	**OM**	**NR**	***N***	**OM**	**NR**
1	rxn90001_c	1176	0.00	0.00	210	0.00	0.00
2	rxn90002_c	1175	0.09	0.00	210	0.00	0.00
3	rxn90003_c	1175	0.09	0.00	210	0.00	0.00
4	rxn90004_c	1162	0.00	1.11	208	0.00	0.95
5	rxn90005_c	1145	0.00	2.47	205	0.00	2.38
6	rxn90006_c	1128	0.00	3.91	202	0.48	3.33
7	rxn90007_c	1091	0.00	7.06	199	0.00	5.24
8	rxn90008_c	1010	0.00	13.95	186	0.00	11.43
9	rxn90009_c	938	0.00	20.07	173	0.00	17.62
10	rxn90010_c	864	0.00	26.36	158	0.00	24.76
11	rxn90011_c	799	0.00	31.89	148	0.00	29.52
12	rxn90012_c	735	0.00	37.24	137	0.00	34.76
13	rxn90013_c	611	0.00	47.79	112	0.00	46.67
14	rxn90014_c	563	0.00	51.87	101	0.00	51.90
15	rxn90015_c	477	0.00	59.18	89	0.00	57.62
16	rxn90016_c	407	0.00	65.05	80	0.00	61.90
17	rxn90017_c	327	0.00	72.85	70	0.00	66.67
18	rxn90018_c	229	0.00	80.19	54	0.00	74.29
19	rxn90019_c	176	0.00	84.61	42	0.00	80.00
20	rxn90020_c	153	0.00	85.56	40	0.00	80.95
21	rxn90021_c	139	0.00	87.59	39	0.00	81.43
22	rxn90022_c	118	0.00	89.37	31	0.00	85.24
23	rxn90023_c	94	0.00	91.41	24	0.00	88.57
24	rxn90024_c	70	0.00	93.45	18	0.00	91.43
25	rxn90025_c	56	0.00	94.64	16	0.00	92.28
26	rxn90026_c	42	0.00	95.83	12	0.00	94.29

Note. Pos. = Item position within the test. *N* = Number of valid responses, NR = Percentage of respondents that did not reach an item, OM = Percentage of respondents that omitted the item.

The total number of missing responses, aggregated over omitted, not reached, and technical missing responses per person, is illustrated in Figure 4. Because the majority of the sample did not reach the end of the test, there was a substantial number of missing values. The median number of missing responses was 13; only about 3.9% of the children had no missing response at all.

In sum, the number of missing responses was rather large because many respondents did not reach the end of the test. However, there were no substantial differences in missing rates between children working on a tablet or a laptop.

5.1.2 **Missing responses per item**

Table 3 provides information on the occurrence of different kinds of missing responses per item and device. A few omitted responses were observed for items 2 and 3, while the remaining items exhibited no omitted items. In contrast, there were substantially more missing responses because children did not reach the item. On average, the items had a median of 49.83% missing values of this type. Particularly, items in the second half of the test were frequently not reached.

5.2 **Parameter Estimates**

To avoid potentially biased parameter estimates resulting from mode effects (tablet versus laptop), the following analyses are limited to children using a tablet. Thus, the subsample of children using a laptop was excluded from the scaling procedure. Information on the measurement invariance across device types is given in section 5.2.5.

5.2.1 **Item parameters**

The third column in Table 4 presents the percentage of correct responses in relation to all valid responses for each item. The percentage of correct responses varied between 1% and 97% with an average of 42% (*SD* = 33%) correct responses and, thus, spans a rather broad range. The estimated item difficulties are given in the fourth column of Table 4. The item difficulties were estimated by scoring all missing values (except technical missing values resulting from a premature test termination) as incorrect and constraining the mean of the ability distribution to zero. The standard errors (*SE*) of most difficulty parameters were rather small (*SE*s ≤ 0.09). However, for items with very large or very small percentages of correct responses (i.e., items with a limited response variability) the standard errors were substantially larger and reached up to *SE* = .30. Thus, items with difficulties matching the proficiency distribution of the samples were estimated more precisely, while extremely easy or difficult items exhibited larger uncertainties. The estimated item difficulties ranged from -4.48 (item rxn90001_c) to 6.09 (item rxn90026_c) and, thus, covered a rather wide range including easy as well as difficult items.

Table 4.

Item Parameters for Children using a Tablet

Item	Pos.	*N*	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	Discr.	*aQ3*
rxn90001_c	1	1176	96.68	-4.48	0.17	1.07	0.51	0.28	0.08
rxn90002_c	2	1176	80.61	-2.12	0.09	1.28	5.65	0.36	0.08
rxn90003_c	3	1176	84.27	-2.45	0.09	1.18	3.22	0.46	0.11
rxn90004_c	4	1175	84.00	-2.43	0.09	1.10	1.82	0.61	0.10
rxn90005_c	5	1174	76.49	-1.79	0.08	1.25	5.76	0.49	0.09
rxn90006_c	6	1174	80.75	-2.13	0.09	1.11	2.33	0.67	0.10
rxn90007_c	7	1174	79.98	-2.07	0.08	0.97	-0.58	1.06	0.10
rxn90008_c	8	1174	67.46	-1.16	0.07	0.99	-0.19	1.32	0.07
rxn90009_c	9	1174	63.63	-0.91	0.07	0.93	-2.17	1.67	0.08
rxn90010_c	10	1174	66.18	-1.07	0.07	0.84	-5.06	2.86	0.09
rxn90011_c	11	1174	50.26	-0.09	0.07	0.96	-1.05	1.83	0.10
rxn90012_c	12	1173	49.02	-0.01	0.07	0.82	-5.8	2.85	0.09
rxn90013_c	13	1173	42.03	0.42	0.07	0.74	-7.88	5.21	0.12
rxn90014_c	14	1173	36.83	0.76	0.08	0.81	-5.31	4.21	0.12
rxn90015_c	15	1173	31.97	1.09	0.08	0.79	-5.74	4.57	0.12
rxn90016_c	16	1172	24.06	1.69	0.08	0.78	-5.13	4.33	0.09
rxn90017_c	17	1172	13.4	2.75	0.10	1.02	0.27	2.35	0.07
rxn90018_c	18	1172	12.12	2.92	0.11	0.82	-2.77	3.19	0.08
rxn90019_c	19	1171	10.16	3.20	0.11	0.94	-0.79	2.21	0.13
rxn90020_c	20	1171	7.94	3.57	0.13	0.82	-2.20	3.08	0.13
rxn90021_c	21	1169	7.44	3.66	0.13	0.81	-2.22	3.31	0.14
rxn90022_c	22	1169	4.96	4.24	0.15	0.85	-1.34	2.68	0.12
rxn90023_c	23	1169	3.93	4.55	0.17	1.00	0.03	1.57	0.12
rxn90024_c	24	1169	3.08	4.87	0.19	0.97	-0.18	1.800	0.11
rxn90025_c	25	1169	1.80	5.53	0.24	0.89	-0.52	2.20	0.08
rxn90026_c	26	1169	1.11	6.09	0.30	1.12	0.56	0.81	0.08

Note. Pos. = Item position, *N* = Number of observed responses; Difficulty = Item difficulty, *SE* = Standard error of item difficulty, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, Discr. = Discrimination parameter of a two-parametric item response model (Birnbaum, 1968), aQ_3 =Average absolute residual correlation for item (Yen, 1983).

5.2.2 **Test targeting and reliability**

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. In Figure 5, the item difficulties of the early reading competence items and the ability of the children are plotted on the same scale. The distribution of the children's estimated abilities is mapped onto the left side whereas the right side shows the distribution of the item difficulties. The respective difficulties ranged from -4.47 (item rxn90001_c) to 6.09 (item rxn90026_c) and, thus, spanned a rather broad range. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 3.03, which implies good differentiation between children. The reliability of the test (EAP/PV reliability = .88, WLE reliability = .88) was good. The median of the item difficulty distribution was about 0.95 logits above the mean person ability distribution. Thus, although the items covered a wide range of the ability distribution, on average, the items were too difficult for the children. As a consequence, proficiency estimates in medium- and high-ability regions will be measured relative precisely, whereas lower ability estimates will have larger standard errors of measurement.

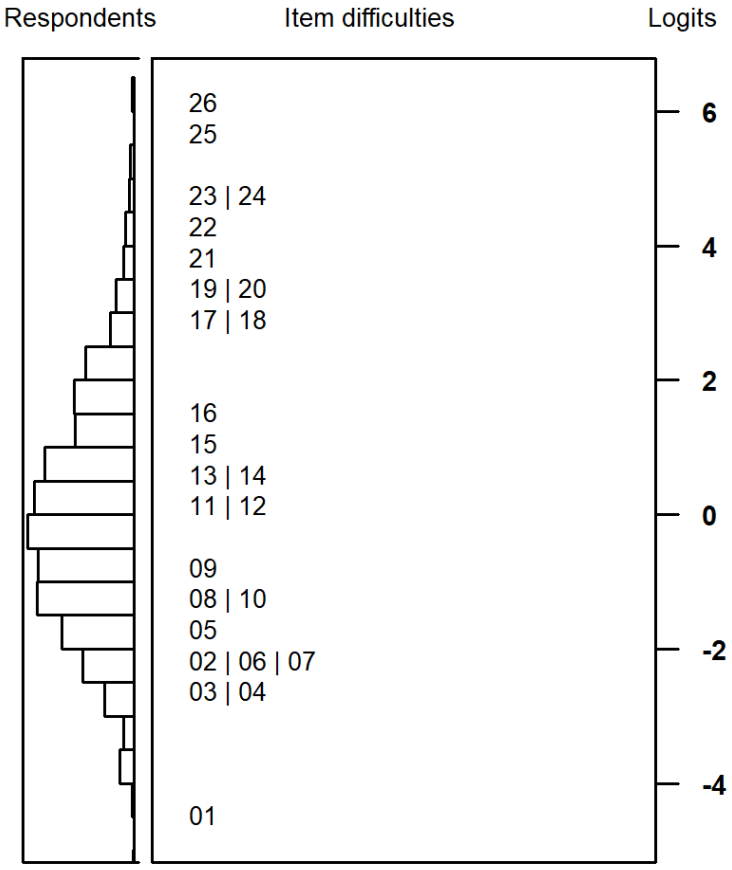

Figure 5. Test targeting. The distribution of the person abilities in the sample is given on the left-hand side of the graph, while the item difficulties are given on the right-hand side of the graph. Each number represents one item parameter corresponding to the item positions given in Tables 3 and 4.

5.3 Quality of the test

5.3.1 Distractor analyses

To investigate how well the distractors of the multiple-choice items performed the point-biserial correlations between each incorrect response (distractor) and the respondents' total correct scores for the remaining items were calculated. The median point-biserial correlations for the distractors fell at -.17 (*Min* = -.50, *Max* = .12). Positive correlations were limited to the last items in the test for which rather few valid responses were observed. In contrast, the correlations of the correct responses with the total scores varied between .13 and .72 (*Mdn* = .31). These results indicate that the distractors functioned well.

5.3.2 Item fit

The evaluation of item fit was based on the final scaling model presented above. Again, the test quality was examined for children working on a tablet only while excluding children using a laptop. Altogether item fit was satisfactory (see Table 4). Two items exhibited a WMSNQ greater than 1.20 (rxn90002_c, rxn90005_c) and one item had a WMNSQ greater than 1.15 (rxn90003_c). However, the respective *t*-values were smaller than 6 and, thus, did not indicate a serious misfit. Moreover, a visual inspection of the ICCs showed no pronounced deviation from the expected ICC for these items. For the remaining items, values of the WMNSQ ranged from 0.74 (item rxn90013_c) to 1.12 (item rxn90026_c).

5.3.3 Differential item functioning

DIF was used to evaluate whether the measurement models were comparable for several subgroups. For this purpose, DIF was examined for the variables sex, highest parental international socioeconomic index (HISEI), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Again, these analyses were limited to children using a tablet while excluding children working on a laptop. In addition, we also examined DIF effects between the two administration devices (tablet versus laptop). The differences between the estimated item difficulties (on the logit scale) in the various groups are summarized in Table 5. For example, the column "boys vs. girls" reports the differences in item difficulties between boys and girls; a positive value would indicate that the test was more difficult for boys, whereas a negative value would highlight a lower difficulty for girls as opposed to girls. Besides investigating DIF for every single item, an overall test for DIF was performed by comparing models which allow for DIF to those that only estimate main effects (see Table 6).

Sex: The sample included 578 boys and 598 girls. There were no substantial gender differences in early reading competence as indicated by the main effect of 0.07 logits (Cohen's *d* = 0.04). Two items (rxn90001_c, rxn90004_c) showed DIF greater than 0.60 logits (or a *d* greater than 0.34) and were more difficult for boys than for girls. The large DIF of 1.11 logits for the first item might indicate that the test instruction was slightly too complicated for boys, thus, requiring the first item as an exercise to grasp the test concept. However, an overall test for DIF (see Table 6) by comparing the DIF model to a model that only estimated the main effect (but ignored potential DIF) suggested that the observed DIF was negligible for the administered test. Although a model comparison using Akaike's (1974) information criterion (AIC) favored the DIF model over the more parsimonious model including only the main effect, the Bayesian information criterion (BIC; Schwarz, 1978) that also takes the number of estimated parameters into account and, thus, guards against overparameterization of models

suggested a superior fit for the main effects model. Moreover, the estimated main effects for sex were nearly identical in both models. These results indicated that there was no pronounced DIF concerning sex that might have distorted the parameter estimates.

HISEI: The HSIEI of the children's parents was used as a proxy for socioeconomic status and split at a value of 75 to create two approximately equally sized groups. This resulted in 630 children with low socioeconomic status and 546 children with high socioeconomic status. On average, children with lower socioeconomic status performed on average -0.51 logits (Cohen's *d* = -0.30) lower in early reading competence as compared to children with higher socioeconomic status. There was no considerable DIF comparing the two groups (see Table 5) with the highest DIF being 0.50 for item rxn90025_c. As a consequence, also the overall test for DIF using the AIC and BIC favored the main effect model that did not account for potential DIF (Table 6).

Migration background: There were 1,045 children without migration background and 127 children with a migration background. In comparison to children without a migration background, children with a migration background had, on average, a slightly lower early reading competence (main effect = -0.29 logits, Cohen's *d* = -0.16). Most items did not exhibit a noteworthy DIF due to migration background with differences in the estimated item difficulties less than 0.6 logits (highest DIF = 0.58 for item rxn90022_cc). Only the four items presented last in the test showed substantial DIF between -4.53 and 0.83 logits. However, this is likely a consequence of the small sample size of children with migration backgrounds and the previously described problems with missing values resulting in most children not reaching the last items of the test. Moreover, the DIF of -0.66 for the first item highlighting a larger difficulty for children with migration background suggests that these children required this item as a means to understand the test procedure. Consequently, the overall test for DIF using the AIC and BIC also favored the main effect model that did not include item-level DIF (see Table 6). Nevertheless, the DIF might have distorted mean level comparisons to some degree as evidenced by the different main effects observed in the main effect model and the DIF model (Cohen's *d*s of 0.29 versus 0.38).

Device: The children worked on the early reading competence test using either a tablet (with touch functionality) or a laptop (with a mouse). Therefore, we examined potential device effects. 1,176 children were using a table and 210 children were using a laptop. As expected, there were no pronounced differences in the children's mean abilities between the two modes (-0.18 logits, Cohen's *d* = -0.14). More importantly, there was no noteworthy DIF except for the last item (DIF = 0.80 logits for rxn90026_c). However, this seemed to be related to the small sample size in the group using a laptop, similarly to the DIF for migration background. Also, the overall tests for DIF favored the main effect model that did not include item-level DIF (see Table 6).

Table 5.

Differential Item Functioning

Item	Sex	HISEI	Migration	Device
	boys vs. girls	low vs. high	without vs. with	tablet vs. laptop
rxn90001_c	1.11 (0.63)	-0.08 (-0.05)	-0.66 (-0.37)	-0.12 (-0.07)
rxn90002_c	0.22 (0.13)	0.15 (0.09)	-0.13 (-0.07)	-0.05 (-0.03)
rxn90003_c	0.49 (0.28)	0.04 (0.02)	0.12 (0.07)	-0.25 (-0.15)
rxn90004_c	0.73 (0.42)	0.01 (0.01)	-0.35 (-0.20)	0.18 (0.10)
rxn90005_c	0.54 (0.31)	-0.16 (-0.09)	-0.13 (-0.08)	-0.19 (-0.11)
rxn90006_c	0.57 (0.33)	-0.24 (-0.14)	-0.15 (-0.08)	-0.46 (-0.27)
rxn90007_c	0.26 (0.15)	-0.36 (-0.21)	-0.01 (-0.00)	-0.13 (-0.07)
rxn90008_c	-0.05 (-0.03)	-0.15 (-0.08)	-0.13 (-0.07)	0.17 (0.10)
rxn90009_c	-0.27 (-0.15)	-0.08 (-0.05)	-0.32 (-0.18)	-0.06 (-0.03)
rxn90010_c	-0.33 (-0.19)	-0.22 (-0.13)	0.35 (0.20)	-0.16 (-0.09)
rxn90011_c	-0.15 (-0.08)	-0.14 (-0.08)	0.09 (0.05)	-0.27 (-0.16)
rxn90012_c	-0.23 (-0.13)	-0.14 (-0.08)	-0.04 (-0.02)	-0.26 (-0.15)
rxn90013_c	-0.07 (-0.04)	0.01 (0.01)	0.22 (0.13)	-0.02 (-0.01)
rxn90014_c	-0.14 (-0.08)	-0.03 (-0.02)	0.26 (0.15)	-0.30 (-0.17)
rxn90015_c	-0.11 (-0.06)	0.12 (0.07)	0.33 (0.19)	-0.20 (-0.11)
rxn90016_c	-0.06 (-0.03)	0.22 (0.13)	0.47 (0.27)	0.16 (0.09)
rxn90017_c	-0.19 (-0.11)	0.15 (0.09)	0.21 (0.12)	0.11 (0.06)
rxn90018_c	-0.36 (-0.20)	0.35 (0.20)	0.40 (0.23)	0.37 (0.22)
rxn90019_c	-0.62 (-0.35)	0.20 (0.11)	0.32 (0.18)	-0.12 (-0.07)
rxn90020_c	-0.12 (-0.07)	0.30 (0.18)	0.24 (0.14)	0.25 (0.15)
rxn90021_c	-0.30 (-0.17)	0.12 (0.07)	-0.04 (-0.03)	0.55 (0.32)
rxn90022_c	0.26 (0.15)	-0.12 (-0.07)	0.58 (0.33)	0.34 (0.20)
rxn90023_c	-0.18 (-0.10)	-0.06 (-0.03)	0.94 (0.54)	-0.36 (-0.20)
rxn90024_c	-0.10 (-0.05)	0.03 (0.02)	1.10 (0.63)	-0.43 (-0.24)
rxn90025_c	-0.37 (-0.21)	0.50 (0.29)	0.83 (0.489)	-0.45 (0.26)
rxn90026_c	-0.55 (-0.31)	-0.43 (-0.25)	-4.51 (-2.58)	0.80 (0.46)
Main effects:				
DIF model	-0.06 (-0.04)	-0.54 (-0.31)	0.38 (0.22)	-0.24 (-0.11)
Main effect model	-0.07 (-0.04)	-0.51 (-0.30)	0.29 (0.16)	-0.18 (-0.14)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. HISEI = Highest international socio-economic index of parents.

Table 6.

Comparisons of Models with and without DIF

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	DIF model	1176	21283	53	<u>21389</u>	21657
	Main effect	1176	21358	28	21414	<u>21556</u>
HISEI	DIF model	1176	21317	53	21423	21691
	Main effect	1176	21336	28	<u>21392</u>	<u>21534</u>
Migration	DIF model	1172	21185	53	21291	21559
	Main effect	1172	21209	28	<u>21265</u>	<u>21407</u>
Device	DIF model	1386	25302	53	25408	25785
	Main effect	1386	25327	28	<u>25383</u>	<u>25539</u>

Note. The best-fitting model according to each information criterion is underlined. HISEI = Highest international socio-economic index of parents.

5.3.4 **Rasch-homogeneity**

An essential assumption of the Rasch (1960) model is that all item-discrimination parameters are equal. To test this assumption, a 2PL that estimates discrimination parameters was fitted to the data. The estimated discrimination parameters differed substantially between items (see Table 4). The median discrimination parameter fell at 2.02 (*Min* = 0.28, *Max* = 5.21). Particularly, the first items with extremely high rates of correct responses exhibited lower discrimination parameters as compared to items in the middle of the test that more closely matched the ability distribution of the sample. Also, model fit indices suggested a better model fit of the 2PL (AIC = 20338, BIC = 20601, number of parameters = 52) as compared to the Rasch model (AIC = 21377, BIC = 21514, number of parameters = 27). However, an inspection of the respective ICCs of the Rasch model indicated an adequate fit of the observed ICCs to the expected ICCs. Despite the empirical preference for the 2PL, 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 to preserve the item weightings as intended in the theoretical framework (Lenhard et al., 2018).

5.3.5 **Unidimensionality**

The dimensionality of the test was investigated by evaluating the correlations between the residuals of the Rasch model. The aQ_3 statistics were quite low. The average aQ_3 statistic across all item pairs was M = 0.10 (SD = 0.02). About 10% of all pairwise residual correlations exceeded .20 and, thus, exhibited slight dependencies. But no item exhibited a noticeable average residual correlation (see the last column in Table 4). Overall, these results indicate an essentially unidimensional test. Because the early reading competence test was constructed to measure a single dimension, a unidimensional competence score was estimated.

6 Discussion

The analyses in the previous sections reported information on the quality of an early reading competence test (Lenhard et al., 2018) that was administered in Starting Cohort 2 of the NEPS. Different kinds of missing responses were examined, item fit statistics and item characteristic curves were evaluated, and item discriminations were investigated. Further quality inspections were conducted by examining differential item functioning and testing Rasch-homogeneity. Various criteria indicated a good fit of the items and measurement invariance across various subgroups. However, the number of missing responses was rather large because many children did not finish the test in time. The test had a good reliability and distinguished well between test-takers. However, the test was slightly better targeted at medium- to high-performing children and better covered the high ability spectrum. As a consequence, ability estimates will be more precise for high-performing children as compared to low-performing children. In summary, the test had good psychometric properties that allowed the estimation of a unidimensional early reading competence score.

7 Data in the Scientific Use Files

The SUF contains 26 dichotomously scored items with 0 indicating an incorrect response and 1 indicating a correct response. For further details on the naming conventions of the variables see Fuß and colleagues (2021). In the SUF, manifest early reading competence scores are provided in the form of sum scores (rxn9_sc3) as suggested in the test manual (Lenhard et al., 2018). For children that exhibited a premature test termination because of technical difficulties no sum scores are provided.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723. https://doi.org/10.1109/TAC.1974.1100705

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* (pp. 397-479). Addison-Wesley.

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2021). *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.

Ganzeboom, H. B. G. (2010, May). *A new international socio-economic index [ISEI] of occupational status for the International Standard Classification of Occupation 2008 [ISCO-08] constructed with data from the ISSP 2002-2007*. Annual Conference of International Social Survey Programme, Lisbon, Portugal.

Gnambs, T. (2020). *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohorts 4 (Wave 10), 5 (Wave 12), and 6 (Wave 9)* (NEPS Survey Paper No. 72). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP72:1.0

Lenhard, W., & Lenhard, A., & Schneider, W. (2018). *ELFE II: Ein Leseverständnistest für Erst- bis Siebtklässler – Version II*. Hogrefe.

Petersen, L. A., Beyer, T., & Bednorz, D. (2022): *NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 1 for Eight-Year-Old Children* (NEPS Survey Paper No. 97). Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests.* (NEPS Working Paper No. 14). Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5*, 189-216. https://doi.org/10.25656/01:8430

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. The Danish Institute of Education Research.

Robitzsch, A., Kiefer, T., & Wu, M. (2021). *TAM: Test analysis modules*. R package version 3.7-16. URL: https://CRAN.R-project.org/package=TAM

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464. https://doi.org/10.1214/aos/1176344136

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450. https://doi.org/10.1007/BF02294627

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft, 14*, 67-86. https://doi.org/10.1007/s11618-011-0182-7

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145. https://doi.org/10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x