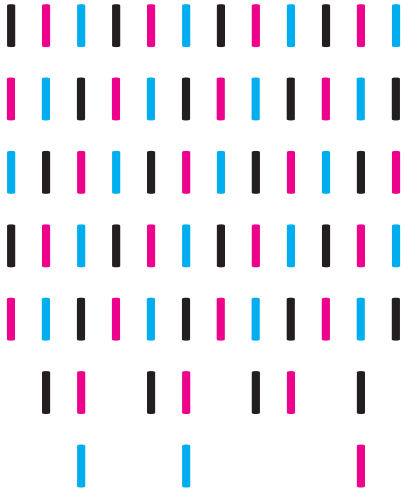

NEPS SURVEY PAPERS

Luise Fischer and Tabea Durda

NEPS TECHNICAL REPORT FOR RECEPTIVE VOCABULARY: SCALING RESULTS OF STARTING COHORT 2 FOR KINDERGARTEN (WAVE 1), GRADE 1 (WAVE 3) AND GRADE 3 (WAVE 5)

NEPS Survey Paper No. 65
Bamberg, January 2020

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).

The NEPS Survey Papers are edited by a review board consisting of the scientific management of LIfBi and 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 available at https://www.neps-data.de (see section “Publications“).

Editor-in-Chief: Corinna Kleinert, LIfBi/University of Bamberg/IAB Nuremberg

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

NEPS Technical Report for Receptive Vocabulary: Scaling Results of Starting Cohort 2 for Kindergarten (Wave 1), Grade 1 (Wave 3) and Grade 3 (Wave 5)

Luise Fischer[1,2] and Tabea Durda[1]

[1]*Leibniz Institute for Educational Trajectories, Bamberg, Germany*

[2]*University of Bamberg, Germany*

E-mail address of lead author:

luise.fischer@lifbi.de

Bibliographic data:

Fischer, L., & Durda, T. (2020). *NEPS Technical Report for Receptive Vocabulary: Scaling Results of Starting Cohort 2 for Kindergarten (Wave 1), Grade 1 (Wave 3) and Grade 3 (Wave 5)* (NEPS Survey Paper No. 65). Bamberg, Germany: Leibniz Institute for Educational Trajecto-ries, National Educational Panel Study. doi:10.5157/NEPS:SP65:1.0

Acknowledgments:

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 2, Kindergarten, Grades 1 and 3, doi:10.5157/NEPS:SC2:8.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

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

NEPS Technical Report for Receptive Vocabulary: Scaling Results of Starting Cohort 2 for Kindergarten (Wave 1), Grade 1 (Wave 3) and Grade 3 (Wave 5)

Abstract

The National Educational Panel Study (NEPS) investigates the development of competencies across the life span and develops tests for the assessment of different competence domains. In order to evaluate the quality of the competence tests, various analyses based on item response theory are performed. This paper describes the data and scaling procedures for the receptive vocabulary test that was administered in waves 1, 3, and 5 of Starting Cohort 2 (Kindergarten) to children attending Kindergarten and, subsequently, grades 1 and 3 in elementary school. The three receptive vocabulary tests contained 77, 66, and 72 items, respectively, that were adapted from the *Peabody Picture Vocabulary Test*. Each item included four black and white pictures (three distractors and one correct option) that had to be evaluated with regard to a spoken target word. The tests were administrated to 2,859 (50% girls), 6,471 (49% girls), and 5,602 (51% girls) children from Starting Cohort 2. The children's responses were scaled using the Rasch model. Because the receptive vocabulary tests comprises of a subset of items belonging to a well-established and validated instrument, only items exhibiting a large misfit were excluded from the scaling procedure. Item fit statistics, differential item functioning, Rasch-homogeneity, and local item independence were evaluated to ensure the quality of the tests. These analyses showed that the tests exhibited acceptable reliabilities and satisfactory model fits. The items covered primarily the lower and middle range of the samples' ability distributions. However, the variances implied good to satisfactory differentiations between the children. Furthermore, test fairness could be confirmed for different subgroups. Analyses of missing values revealed no shortcomings of the tests. Overall, the receptive vocabulary tests had satisfactory psychometric properties that allowed for an estimation of reliable competence scores. Importantly, the tests were also linked across measurement occasions, thus, allowing for longitudinal comparisons of changes in the children's vocabulary skills. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the R code for estimating the manifest ability scores.

Keywords

item response theory, scaling, receptive vocabulary, scientific use file

Contents

1 Introduction 5

2 The Measurement of Receptive Vocabulary 5

2.1 Item Development 6

2.2 Administration 6

3 Data 7

3.1 Design of the Study 7

3.2 Sample 7

4 Analyses 8

4.1 Missing Responses 8

4.2 Scaling Model 8

4.3 Checking the Quality of the Scale 9

4.4 Software 10

5 Results 10

5.1 Missing Responses 10

5.1.1 Missing responses per person 10

5.1.2 Missing responses per item 12

5.2 Parameter Estimates 13

5.2.1 Item parameters 13

5.2.2 Test targeting and reliability 20

5.3 Quality of the Test 22

5.3.1 Item fit 22

5.3.2 Differential item functioning 22

5.3.3 Rasch homogeneity 34

6 Discussion 34

7 Data in the Scientific Use File 34

7.1 Naming conventions 34
7.2 Linking of competence scores 34
7.2.1 Linking of waves 1 and 3 35
7.2.2 Linking of waves 3 and 5 37
7.2.3 Correcting for sample drop out 39
7.3 Receptive vocabulary scores 39
7.3.1 Cross-sectional WLE 39
7.3.2 Correcting for test position 40
References 41
Appendix A 44
Appendix B 51

1 Introduction

Within the National Educational Panel Study (NEPS) different competences 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 competences measured in the NEPS is given by Weinert and colleagues (2011) as well as Fuß, Gnambs, Lockl, and Attig (2019).

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 tests. 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 (2012).

In this paper, the results of these analyses are presented for three receptive vocabulary tests that were administered in waves 1, 3, and 5 of Starting Cohort 2 (Kindergarten) to children attending Kindergarten and, subsequently, grades 1 and 3 in elementary school. First, the main concept of the receptive vocabulary tests and the test designs are introduced. Then, the competence data of the three waves of Starting Cohort (SC) 2 and the analyses performed on these 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.

2 The Measurement of Receptive Vocabulary

The framework for the receptive vocabulary test is described in Berendes, Weinert, Zimmermann, and Artelt (2013). In the following, there will be a brief description of specific aspects of the receptive vocabulary test that are necessary for understanding the scaling results presented in this paper.

Receptive vocabulary represents a simple, internationally comparable indicator of language competencies reflecting children's accumulated knowledge and crystallized intelligence (Berendes et al., 2013). In the international context, the *Peabody Picture Vocabulary Test* (PPVT; Dunn, 1959; Dunn & Dunn, 1981, 1997, 2007) is certainly the most popular instrument for measuring receptive vocabulary. Because a published German version of the PPVT was not available for young children up to an age of 13 years at the time of administration (Dunn & Dunn, 2004), a modified test version that was comparable to the original PPVT was developed for the NEPS.

The PPVT is a picture selection task. Each item of the PPVT, and thus each item of the adapted NEPS version, consists of a set of four pictures. One picture represents the correct answer while the other three pictures show incorrect response options, so-called distractors. The child must select one picture out of four that best illustrates the meaning of a spoken target word.

2.1 Item Development

For the first wave conducted in kindergarten, existing data of the PPVT from the longitudinal *BiKS-3-10* ("Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter") study (see Mudiappa & Artelt, 2014) was analysed to shorten the administered research version. Using a sample of 504 children between the age of 3;10 and 5;7 years (*M*= 4.6, *SD* = 0.4; cf. von Maurice, Artelt, Blossfeld, Faust, Roßbach, & Weinert, 2007), 77 items were selected from a pool of 175 items that were particularly discriminative for this age range (see Berendes et al., 2013). In the thus adapted version of the PPVT, the items were arranged by increasing difficulty resulting in easier items being presented first and more difficult items later in the test. For the subsequent waves in elementary schools, 66 items (wave 3) and 72 items (wave 5) were selected based on pilot studies conducted for the NEPS including 566 and 638 children, respectively. In extensive preliminary analyses to evaluate the quality of these items, the present study identified some limitations with regard to the item difficulties (i.e., some items were too easy for the age range) or poor item fit. Therefore, these items were excluded from the present analyses resulting in tests including 57, 66, and 49 items, respectively, for the three waves.

2.2 Administration

In kindergarten, children were tested individually and all tests were introduced as playful games by well-trained test administrators. For each item, the test administrator showed the corresponding four pictures presented on one page of a test booklet (the four pictures filling the page completely; printed in black and white). Due to a ring binding, the pages could easily be turned to go to the next item. The target words were presented orally using a CD player. The test administrator played all the items one after another and paused after each item to observe and protocol the child's reaction, that is, which picture had been selected by the child to match the presented item. The target word had to be replayed if the child did not understand it clearly, if he/she pointed on several pictures simultaneously, or if he/she did not react for a duration of more than five seconds. All items were presented in a predetermined order until the termination criterion was met: the test was aborted after six consecutive items were not correctly solved. The child did not receive any aid or feedback on their performance.

In waves 3 and 5 in primary school receptive vocabulary was assessed in a group setting in the first graders' classrooms. The vocabulary test was presented in one test booklet, printed in black and white, and showing four items on one page. The target word for each item was presented orally using a CD player. The test administrator played all the items one after another and paused after each item, guaranteeing that the children had enough time to mark the corresponding picture in their exercise books. Once the test administrator was sure that all children were listening carefully, the presentation of the next item resumed. All items were presented in a predetermined order, and in contrast to wave 1, there was no predetermined termination criterion. The test booklets were distributed in two variants. In order to prevent cheating within the group setting, the response options were presented in different order, thus, displaying the pictures in each item in different sequence. Importantly, the item positions within the tests did not vary between groups.

3 Data

3.1 Design of the Study

The studies in waves 1, 3, and 5 of Starting Cohort 2 (Kindergarten) assessed different competence domains including technological and information literacy (ICT), mathematical literacy (MA), scientific literacy (SC), declarative metacognition (MD), receptive grammatical competence (GR) as well as receptive vocabulary competence (VO). Children in wave 1 were tested individually within their respective institution, while testing in waves 3 and 5 was performed in a group setting. The position of the vocabulary test within the test batteries varied over the three waves: In wave 1, the receptive vocabulary test was administered at second position (after the science test). In wave 3, it was administered at different positions, either as the first test or second after working on a test of declarative metacognition; the test sequence was randomly assigned to the children. Finally, in wave 5 the vocabulary test was administered on the first position to all children. There was no multi-matrix design regarding the order of the items *within* a specific test and all children received the test items in the same order. However, in order to prevent cheating, in the school setting (waves 3 and 5), two different test forms varying the order of the four response options *within* an item were randomly administered to the children (see above). A detailed description of the study designs is available on the NEPS website (http://www.neps-data.de).

3.2 Sample

A total of 9,095 children (50% girls) answered at least three items on the receptive vocabulary competence test forms in waves 1, 3, or 5 and, thus, were used for the psychometric analyses (cf. Pohl & Carstensen, 2012). Of these, *N* = 2,859 (50% girls) were tested in wave 1 (attending Kindergarten), *N* = 6,471 (49% girls) were tested in wave 3 (attending grade 1), and *N* = 5,602 (51% girls) were tested in wave 5 (attending grade 3). While *N* = 445 children were administered all three test forms, *N* = 528 children participated in waves 1 and 3, and *N* = 5,281 children participated in waves 3 and 5. Basic sociodemographic information of the three subsamples (waves 1, 3, and 5) is summarized in Table 1.

Table 1

Number of Children and Basic Sociodemographic Information

	Wave 1	**Wave 2**	**Wave 3**
Sample size	2,859	6,471	5,602
Girls	50%	49%	51%
Migration background	12%	9%	9%
≥ 100 Books at home	43%	56%	57%
Presented first before other tests within the test battery	0%	51%	100%

4 Analyses

This section briefly describes the analyses that were conducted to evaluate the test. These included inspecting various types of missing responses, scaling the data, and examining the quality of the test.

4.1 Missing Responses

There are different types of missing responses in competence test data. These include missing responses due to a) invalid responses and b) items that test takers did not reach. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response given were coded as not-reached. Omitted items (i.e., when test takers skipped some items) where recoded as wrong answers following the approach applied in the PPVT (Lenhard et al., 2015). 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 persons were coping with the test. Missing responses per item were examined in order to evaluate how well each of the items functioned.

4.2 Scaling Model

Item and person parameters were estimated using a Rasch model (Rasch, 1960). A detailed description of the scaling model can be found in Pohl and Carstensen (2012). Items with less than 200 correct responses or less than 200 incorrect responses were excluded from further analyses in order to avoid possible estimation problems. This occurred for nine items in wave 1 ("vok10001_c", "vok10003_c", "vok10004_c", "vok10005_c", "vok10006_c", "vok10029_c", "vok10030_c", "vok10044_c", "vok10059_c") and two items in wave 5 ("vok10042_sc2g3_c", "vok10028_sc2g3_c"). Receptive vocabulary scores were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in section 7 of the present report.

4.3 Checking the Quality of the Scale

The receptive vocabulary test was specifically adapted for the samples' age range. In order to ensure appropriate psychometric properties, the quality of the test was examined in several analyses.

The items were analyzed in a Rasch (1960) model. The fit of the items was evaluated based on the weighted mean square (WMNSQ), the respective *t*-value, corrected point-biserial correlations of the correct responses with the total correct score, and the item characteristic curves. As the receptive vocabulary test comprises of a subset of items belonging to the PPVT (Lenhard et al., 2015), only items exhibiting a severe misfit were excluded from the scaling model. In wave 1, 11 items presented at the end of the test ("vok10067_c", "vok10068_c", "vok10069_c", "vok10070_c", "vok10071_c", "vok10072_c", "vok10073_c", "vok10074_c", "vok10075_c", "vok10076_c" "vok10077_c") were excluded. These might reflect fatigue effect because the children were rather young. No items were excluded in wave 3. In wave 5, 21 items with low discriminations were excluded. ("vok10045_sc2g3_c", "vog60001_sc2g3_c", "vog90047_sc2g3_c", "vok10069_sc2g3_c", "vog90022_sc2g3_c", "vog60009_sc2g3_c", "vog10046_sc2g3_c", "vog30016_c", "vog60051_sc2g3_c", "vog90039_sc2g3_c", "vog60038_sc2g3_c", "vog60050_sc2g3_c", "vog90053_sc2g3_c", "vok10070_sc2g3_c", "vog90010_sc2g3_c", "vok10068_sc2g3_c", "vog90005_sc2g3_c", "vog90007_sc2g3_c", "vog60022_sc2g3_c", "vog90020_sc2g3_c", "vog30068_c").

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

The fit of the dichotomized (correct = 1 and incorrect = 0) multiple-choice items 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 > 1.20 (*t*-value > |8|) were judged as having a considerable item misfit and their performance was further investigated. Correlations of the item score with the corrected total greater than .30 were considered as good, greater than .20 as acceptable, and below .20 as problematic. Overall judgment of the fit of an item was based on all fit indicators.

The receptive vocabulary test should measure the same construct for all respondents. If some items favored certain subgroups (e.g., items were easier for males than for females, although being equally proficient), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., males and females) would be biased and, thus, unfair. For the present studies, test fairness was investigated for the variables test position (wave 3 only), gender, migration background, and the number of books at home (as a proxy for socioeconomic status; see Pohl & Carstensen, 2012, for a description of these variables). Differential item functioning (DIF) was examined using a multigroup IRT model, in which 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, differences between 0.4 and 0.6 as small but 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.

To test the assumption of equal item discrimination parameters as implied by the Rasch model, a two parameter logistic model (2PL; Birnbaum, 1968) was also fitted to the data and compared to the Rasch model. Moreover, we examined whether the residuals of the one-dimensional model exhibited approximately zero-order correlations as indicated by Yen's (1984) Q_3. Because in case of locally independent items, the Q_3 statistic tends to be slightly negative, we report the corrected Q_3 that has an expected value of 0. Following prevalent rules-of-thumb (Yen, 1993) values of Q_3 falling below .20 indicate essential unidimensionality.

4.4 Software

The item response models were estimated with the *TAM* package version 3.0-21 (Robitzsch, Kiefer, & Wu, 2018) in *R* version 3.5.2 (R Core Team, 2019).

5 Results

5.1 Missing Responses

5.1.1 Missing responses per person

Almost none of the participants (i.e., 0.0% in wave 1, 0.3% in wave 3, 0.1% in wave 5) produced an invalid response.

Another source of missing responses are items that were not reached by the children because they reached the termination criterion (wave 1) or ran out of time (waves 3 and 5); these are all missing responses after the last valid response. In waves 1, 3, and 5 about 80%, 90%, and 92% of the children finished the entire test (see Figure 1). In wave 1, about 17% of the children did not reach 11 or more items due to the termination criterion (i.e., after 6 successive incorrect responses). In waves 3 and 5, about 6% and 2% did not reach 11 or more items due to time limits. Thus, testing time did not seem to be a major issue for the children.

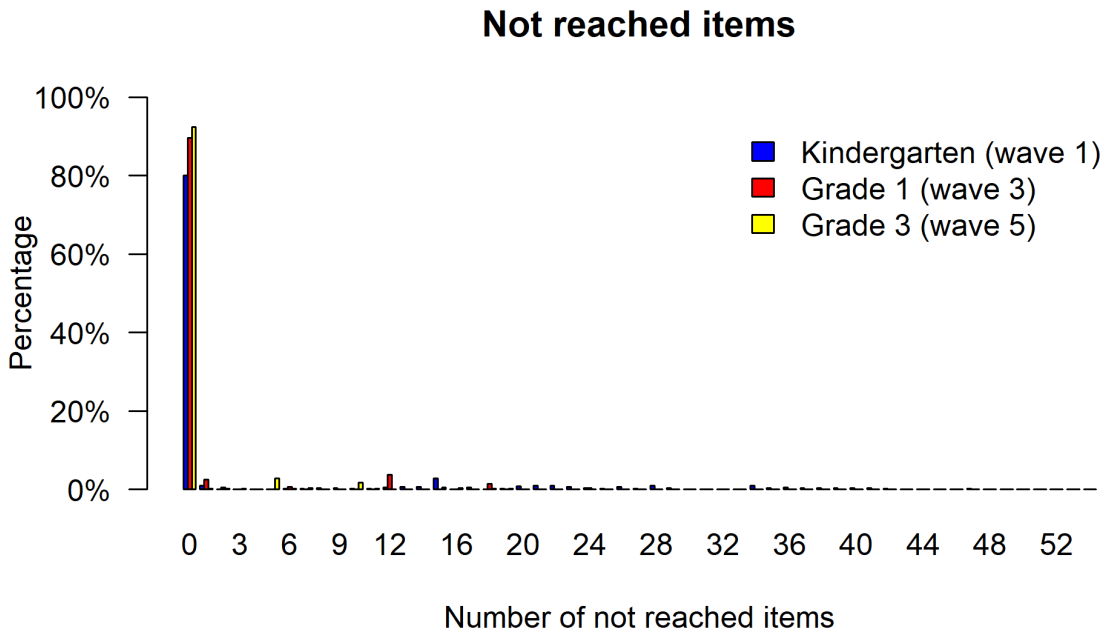

Figure 1. Number of not-reached items by wave.

With an item's progressing position in the test, the number of persons that did not reach the item rose in all waves (see Figure 2).

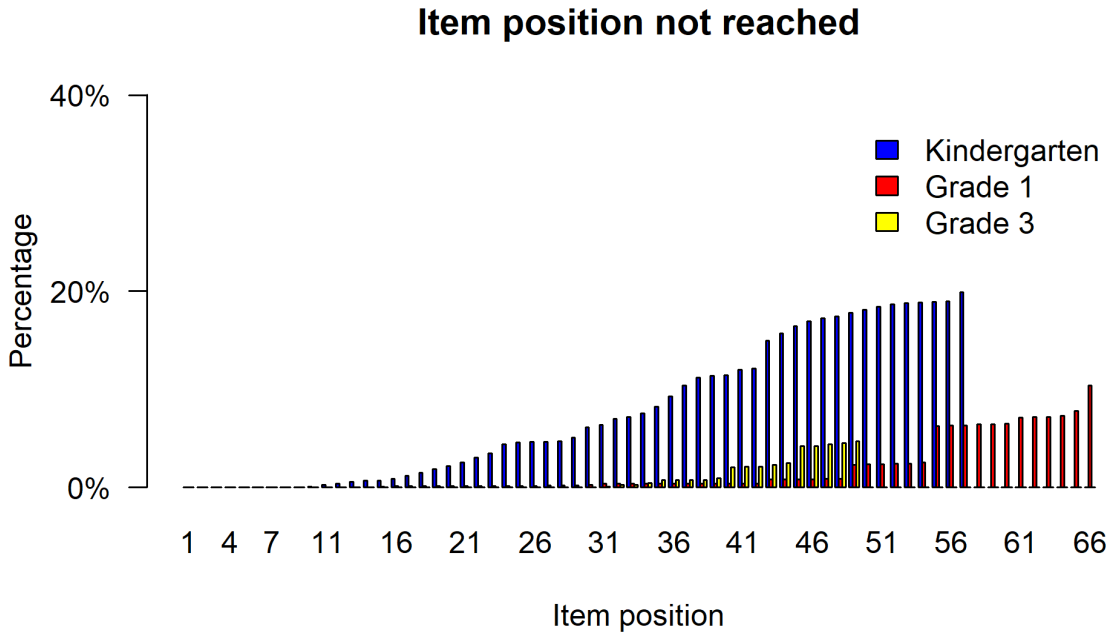

Figure 2. Item position not reached by waves. Note that the scale on the x-axis was adapted (i.e., the scale was cut off at 40%).

About 20% of the children did not reach the last item in wave 1 and about 10% of the children did not reach the last item in wave 3. However, item selection in wave 5 during the scaling process had no influence on the number of persons that did not reach the item. As such, about 8% of the children in wave 5 did not reach the last item (see Figure 2).

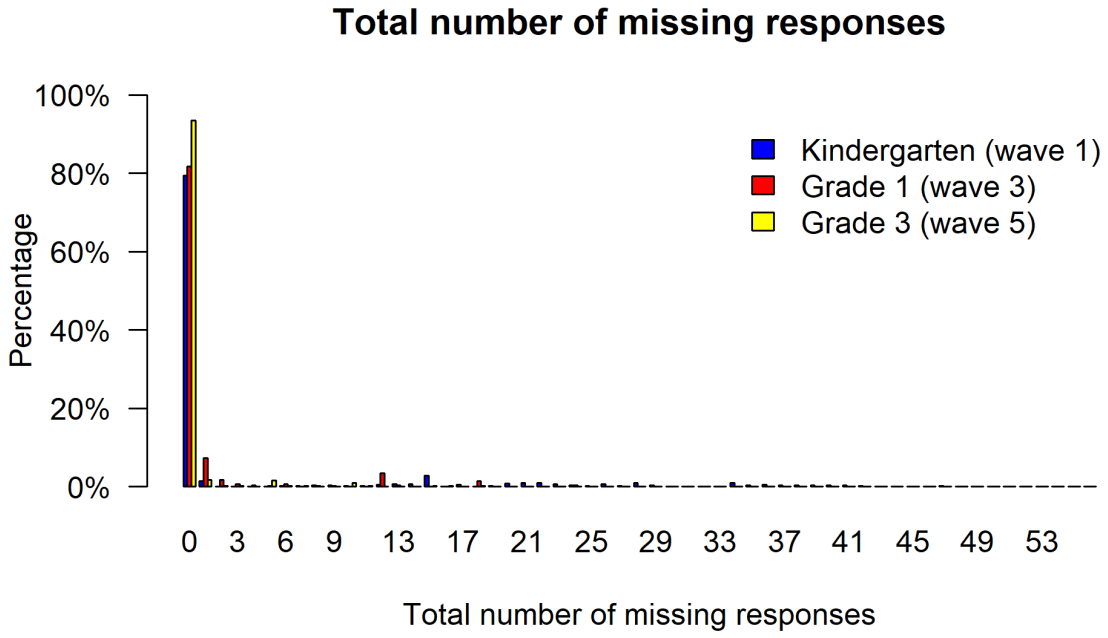

Figure 3. Total number of missing responses by wave.

The total number of missing responses, aggregated over not-reached and not valid missing responses per person, is illustrated in Figure 3. Children in waves 1, 3, and 5 had *M* = 4.34 (*SD* = 10.04), *M* = 1.36 (*SD* = 4.74), and *M* = 0.41 (*SD* = 1.99) missing responses, respectively. About 79%, 82%, and 93% of the test takers had no missing response at all and about 17%, 7%, and 1% had more than ten missing responses.

5.1.2 Missing responses per item

Tables 2, 3 and 4 (see Appendix A) provide information on the occurrence of two kinds of missing responses per item. Over waves 1, 3, and 5, the number of missing values per item was negligible. In wave 1, a maximum of 19.94% (*M* = 7.64%, *SD* = 5.07) of the children failed to reach items due to the termination criterion. The number of invalid responses varied across items between 0.00% and 0.07% (*M* = 0.02%, *SD* = 0.00). In wave 3, a maximum of 10.42% (*M* = 1.70%, *SD* = 0.37) of the children failed to reach items due to time constraints, while the number of invalid responses varied across items between 0.11% and 1.17% (*M* = 0.34%, *SD* = 0.32; see Table 3).In wave 5, a maximum of 7.66% (*M* = 1.27%, *SD* = 0.00) of the children failed to reach items due to time constraints, while the number of invalid responses varied across items between 0.00% and 0.37% (*M* = 0.11%, *SD* = 0.11; see Table 4).

5.2 Parameter Estimates

5.2.1 Item parameters

The second column in Tables 5 (wave 1), 6 (wave 3), and 7 (wave 5) presents the percentage of correct responses in relation to all valid responses for each item. Because there is a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index for item difficulty. The probabilities of a correct response ranged in wave 1 from about 20% to 93% (*M* = 67%, *SD* = 19.97), in wave 3 from about 17% to 94% (*M* = 62%, *SD* = 20.61) and in wave 5 from about 22% to 95% (*M* = 67%, *SD* = 18.32) across all items. Thus, the range of correct and incorrect responses was reasonably large.

The responses of waves 1, 3, and 5 were analyzed separately. The estimated item difficulty parameters are given in Tables 5 (wave 1), 6 (wave 3), and 7 (wave 5). The item difficulty parameters were estimated by constraining the means of the ability distributions to be zero and ranged from -3.15 (item vok10002_c) to 1.89 (item vok10065_c) with an average difficulty of -0.89 (*SD* = 1.27) in wave 1, from -2.95 (item vok10039_sc2g1_c) to 1.75 (item vok10053_sc2g1_c) with an average difficulty of -0.65 (*SD* = 1.12) in wave 3 and from -3.29 (item vok10040_sc2g3_c) to 1.43 (item vog60045_sc2g3_c) with an average difficulty of -1.01 (*SD* = 1.12) in wave 5. Due to the large sample sizes, the standard errors (SE) of the estimated item difficulties (column 4 in Tables 5, 6, and 7) were rather small (all SEs ≤ 0.08).

Table 5

Item Parameters for Wave 1

Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
vok10002_c	0.94	-3.15	0.08	1.01	0.15	0.25	1.08	0.03
vok10007_c	0.85	-2.06	0.06	1.05	1.38	0.31	1.00	0.02
vok10008_c	0.87	-2.29	0.06	1.04	1.07	0.28	0.89	0.02
vok10009_c	0.91	-2.75	0.07	0.91	-1.92	0.40	1.85	0.04
vok10010_c	0.92	-2.87	0.07	0.97	-0.57	0.33	1.43	0.02
vok10011_c	0.88	-2.40	0.06	0.93	-1.62	0.40	1.42	0.03
vok10012_c	0.87	-2.29	0.06	0.93	-1.7	0.41	1.51	0.02
vok10013_c	0.84	-1.97	0.06	0.96	-1.23	0.40	1.32	0.03
vok10014_c	0.84	-2.02	0.06	0.99	-0.22	0.36	1.17	0.02
vok10015_c	0.89	-2.46	0.06	0.89	-2.71	0.47	2.04	0.03
vok10016_c	0.87	-2.26	0.06	0.93	-1.77	0.40	1.52	0.03
vok10017_c	0.87	-2.23	0.06	0.95	-1.31	0.39	1.33	0.02
vok10018_c	0.87	-2.22	0.06	0.91	-2.43	0.43	1.55	0.02
vok10019_c	0.91	-2.70	0.07	0.90	-2.12	0.41	1.65	0.03
vok10020_c	0.83	-1.87	0.05	0.93	-2.25	0.43	1.43	0.03
vok10021_c	0.83	-1.90	0.05	0.88	-3.85	0.49	1.81	0.03
vok10022_c	0.73	-1.19	0.05	0.97	-1.33	0.41	1.17	0.03
vok10023_c	0.74	-1.22	0.05	1.09	3.78	0.28	0.69	0.02
vok10024_c	0.83	-1.83	0.05	0.96	-1.13	0.39	1.29	0.03
vok10025_c	0.81	-1.67	0.05	0.88	-4.02	0.49	1.83	0.03
vok10026_c	0.48	0.19	0.04	1.14	8.87	0.22	0.54	0.02
vok10027_c	0.72	-1.04	0.05	0.85	-7.40	0.54	2.05	0.04
vok10028_c	0.66	-0.72	0.04	0.88	-6.72	0.52	1.73	0.03
vok10031_c	0.54	-0.08	0.04	0.98	-1.59	0.40	1.10	0.02
vok10032_c	0.86	-1.99	0.06	0.88	-3.26	0.46	1.92	0.03
vok10033_c	0.35	0.83	0.04	1.07	4.11	0.25	0.70	0.02
vok10034_c	0.64	-0.59	0.04	1.15	8.11	0.21	0.49	0.03
vok10035_c	0.65	-0.63	0.04	1.03	1.55	0.34	0.91	0.03
vok10036_c	0.78	-1.40	0.05	0.99	-0.22	0.36	1.07	0.02
vok10037_c	0.71	-0.95	0.05	0.99	-0.55	0.37	1.12	0.02

Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
vok10038_c	0.55	-0.08	0.04	0.94	-4.14	0.43	1.29	0.03
vok10039_c	0.79	-1.40	0.05	1.03	1.08	0.30	0.86	0.02
vok10040_c	0.55	-0.07	0.04	0.98	-1.50	0.39	1.11	0.02
vok10041_c	0.58	-0.21	0.04	0.93	-4.59	0.45	1.34	0.03
vok10042_c	0.70	-0.84	0.05	1.00	0.20	0.36	1.03	0.02
vok10043_c	0.25	1.46	0.05	1.08	3.19	0.18	0.55	0.02
vok10045_c	0.40	0.67	0.04	0.99	-0.74	0.35	1.05	0.03
vok10046_c	0.37	0.85	0.04	1.02	1.18	0.30	0.85	0.02
vok10047_c	0.29	1.27	0.05	1.08	3.44	0.21	0.61	0.02
vok10048_c	0.44	0.50	0.04	1.01	0.98	0.33	0.89	0.02
vok10049_c	0.63	-0.39	0.04	0.91	-5.58	0.48	1.49	0.03
vok10050_c	0.74	-0.97	0.05	1.12	4.85	0.21	0.41	0.03
vok10051_c	0.56	-0.02	0.04	1.08	5.05	0.25	0.63	0.02
vok10052_c	0.38	0.82	0.05	1.04	2.43	0.27	0.76	0.02
vok10053_c	0.54	0.07	0.04	0.97	-1.82	0.37	1.09	0.02
vok10054_c	0.90	-2.23	0.07	0.94	-1.18	0.33	1.50	0.03
vok10055_c	0.85	-1.66	0.06	0.93	-1.86	0.38	1.56	0.03
vok10056_c	0.61	-0.22	0.05	1.04	2.68	0.27	0.76	0.02
vok10057_c	0.45	0.54	0.04	1.08	5.37	0.21	0.57	0.03
vok10058_c	0.50	0.28	0.04	1.02	1.42	0.29	0.86	0.03
vok10060_c	0.38	0.88	0.05	1.10	5.53	0.18	0.49	0.02
vok10061_c	0.50	0.31	0.04	1.02	1.52	0.29	0.83	0.03
vok10062_c	0.87	-1.82	0.06	1.05	1.08	0.16	0.65	0.02
vok10063_c	0.65	-0.39	0.05	1.07	3.68	0.22	0.62	0.02
vok10064_c	0.48	0.38	0.04	1.07	4.59	0.23	0.65	0.02
vok10065_c	0.20	1.89	0.05	1.08	2.53	0.14	0.47	0.02
vok10066_c	0.68	-0.54	0.05	1.07	3.46	0.21	0.63	0.02

Note. *SE* = Standard error of item difficulty parameter. WMNSQ = Weighted mean square. *t* = *t*-value for WMNSQ. r_{it} = Corrected item-total correlation. Discr. = Discrimination parameter of a two-parametric logistic model. Q_3 = Average absolute residual correlation for item (Yen, 1983). Estimated parameters are based on *N* = 2,859 (Starting Cohort 2, wave 1). The item-total correlation corresponds to the point-biserial correlation between the correct response and the total score.

Table 6

Item Parameters of Wave 3

Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
vok10067_sc2g1_c	0.80	-1.54	0.03	1.03	1.37	0.24	0.66	0.02
vok10043_sc2g1_c	0.55	-0.21	0.03	0.90	-12.48	0.45	1.40	0.04
vok10053_sc2g1_c	0.83	-1.75	0.03	0.96	-1.99	0.31	0.97	0.02
vok10049_sc2g1_c	0.87	-2.14	0.04	0.89	-4.07	0.40	1.54	0.04
vog60001_sc2g1_c	0.36	0.66	0.03	1.07	6.47	0.20	0.44	0.02
vok10025_sc2g1_c	0.89	-2.31	0.04	0.94	-2.06	0.31	1.16	0.03
vok10076_sc2g1_c	0.36	0.65	0.03	1.08	7.66	0.17	0.42	0.03
vok10050_sc2g1_c	0.81	-1.60	0.03	1.05	2.58	0.19	0.52	0.02
vog10009_c	0.24	1.30	0.03	1.08	4.78	0.13	0.35	0.02
vog60009_sc2g1_c	0.32	0.86	0.03	1.12	10.53	0.09	0.23	0.03
vok10060_sc2g1_c	0.57	-0.30	0.03	1.04	4.83	0.25	0.60	0.02
vok10066_sc2g1_c	0.88	-2.18	0.04	1.04	1.55	0.16	0.54	0.02
vok10063_sc2g1_c	0.84	-1.83	0.04	0.96	-1.98	0.33	1.09	0.03
vok10040_sc2g1_c	0.88	-2.15	0.04	0.90	-3.93	0.40	1.70	0.04
vok10074_sc2g1_c	0.44	0.28	0.03	1.08	8.76	0.21	0.48	0.01
vok10033_sc2g1_c	0.67	-0.79	0.03	0.96	-3.23	0.35	0.97	0.02
vog90015_sc2g1_c	0.49	0.08	0.03	1.07	8.53	0.21	0.50	0.03
vok10051_sc2g1_c	0.74	-1.18	0.03	1.03	1.94	0.26	0.70	0.02
vok10061_sc2g1_c	0.65	-0.70	0.03	0.89	-10.67	0.45	1.46	0.04
vog60051_sc2g1_c	0.38	0.58	0.03	1.12	12.19	0.12	0.28	0.02
vog90007_sc2g1_c	0.64	-0.62	0.03	1.10	9.19	0.18	0.39	0.03
vog60015_sc2g1_c	0.50	0.01	0.03	1.07	7.96	0.21	0.47	0.03
vok10057_sc2g1_c	0.64	-0.66	0.03	1.01	0.55	0.30	0.78	0.02
vok10072_sc2g1_c	0.57	-0.29	0.03	1.03	3.31	0.27	0.65	0.02
vog90016_sc2g1_c	0.44	0.27	0.03	1.09	9.90	0.18	0.43	0.02
vog90032_sc2g1_c	0.17	1.75	0.03	0.94	-2.72	0.28	1.03	0.03
vog60010_sc2g1_c	0.81	-1.58	0.03	0.86	-7.96	0.49	1.95	0.05
vok10041_sc2g1_c	0.80	-1.55	0.03	0.97	-1.67	0.31	0.93	0.02
vok10052_sc2g1_c	0.63	-0.60	0.03	0.92	-8.28	0.42	1.21	0.04
vog60032_sc2g1_c	0.41	0.43	0.03	1.06	6.82	0.21	0.50	0.02

Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
vok10031_sc2g1_c	0.87	-2.08	0.04	0.97	-1.15	0.29	1.03	0.02
vok10045_sc2g1_c	0.83	-1.79	0.03	0.91	-4.16	0.39	1.31	0.03
vok10039_sc2g1_c	0.94	-2.95	0.05	0.94	-1.39	0.28	1.32	0.03
vog10034_c	0.64	-0.66	0.03	0.94	-5.85	0.40	1.17	0.03
vok10034_sc2g1_c	0.85	-1.94	0.04	1.02	0.87	0.22	0.67	0.02
vok10058_sc2g1_c	0.67	-0.77	0.03	0.92	-7.15	0.43	1.21	0.04
vog90031_sc2g1_c	0.39	0.53	0.03	1.03	3.22	0.25	0.66	0.02
vog60049_sc2g1_c	0.35	0.72	0.03	1.05	4.66	0.21	0.55	0.02
vok10065_sc2g1_c	0.43	0.33	0.03	0.95	-5.33	0.37	1.03	0.02
vog10040_c	0.30	0.99	0.03	0.95	-4.07	0.33	1.04	0.02
vok10071_sc2g1_c	0.54	-0.17	0.03	1.02	2.82	0.28	0.66	0.02
vok10069_sc2g1_c	0.38	0.56	0.03	1.05	5.50	0.22	0.52	0.02
vog60025_sc2g1_c	0.45	0.24	0.03	1.02	2.14	0.28	0.71	0.02
vog10044_c	0.68	-0.85	0.03	0.90	-8.89	0.45	1.36	0.03
vok10028_sc2g1_c	0.90	-2.48	0.04	0.89	-3.38	0.40	1.82	0.04
vog10046_c	0.71	-1.01	0.03	0.98	-1.49	0.33	0.86	0.01
vog60027_sc2g1_c	0.28	1.09	0.03	1.02	1.21	0.24	0.66	0.02
vog60047_sc2g1_c	0.49	0.07	0.03	1.06	6.84	0.23	0.55	0.02
vok10022_sc2g1_c	0.80	-1.54	0.03	1.04	2.24	0.22	0.61	0.02
vok10038_sc2g1_c	0.82	-1.66	0.03	0.90	-5.03	0.41	1.47	0.04
vog90028_sc2g1_c	0.81	-1.58	0.03	0.96	-2.18	0.34	1.05	0.02
vok10047_sc2g1_c	0.55	-0.23	0.03	0.99	-1.45	0.33	0.84	0.02
vok10046_sc2g1_c	0.72	-1.03	0.03	0.94	-4.60	0.39	1.13	0.02
vog60019_sc2g1_c	0.60	-0.44	0.03	0.96	-4.20	0.37	1.06	0.03
vok10048_sc2g1_c	0.77	-1.31	0.03	0.92	-4.87	0.42	1.29	0.03
vog10056_c	0.37	0.62	0.03	1.16	14.76	0.08	0.17	0.03
vog90020_sc2g1_c	0.47	0.15	0.03	1.12	13.15	0.16	0.32	0.03
vok10037_sc2g1_c	0.91	-2.58	0.05	0.96	-1.10	0.28	1.09	0.02
vog60030_sc2g1_c	0.35	0.74	0.03	0.95	-4.41	0.34	1.00	0.03
vog10060_c	0.33	0.81	0.03	1.03	2.58	0.25	0.60	0.02
vok10077_sc2g1_c	0.67	-0.79	0.03	1.04	3.44	0.25	0.58	0.02
vog10062_c	0.81	-1.64	0.03	0.93	-3.61	0.39	1.26	0.03
vog10063_c	0.55	-0.23	0.03	1.06	6.58	0.23	0.53	0.02

Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
vok10042_sc2g1_c	0.91	-2.49	0.05	0.92	-2.48	0.35	1.46	0.03
vok10064_sc2g1_c	0.75	-1.23	0.03	0.93	-4.31	0.40	1.22	0.03
vok10026_sc2g1_c	0.71	-0.99	0.03	0.98	-1.54	0.34	0.93	0.03

Note. *SE* = Standard error of item difficulty parameter. WMNSQ = Weighted mean square. *t* = *t*-value for WMNSQ. r_{it} = Corrected item-total correlation. Discr. = Discrimination parameter of a two-parametric logistic model. Q_3 = Average absolute residual correlation for item (Yen. 1983). Estimated parameters are based on *N* = 6,471 (Starting Cohort 2, wave 3). The item-total correlation corresponds to the product-moment correlation between the corresponding categories and the total score.

Table 7

Item Parameters of Wave 5

Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
vog10034_sc2g3_c	0.87	-2.19	0.04	0.92	-3.07	0.39	1.58	0.03
vok10043_sc2g3_c	0.80	-1.57	0.04	0.87	-6.75	0.47	1.76	0.05
vog90031_sc2g3_c	0.48	0.09	0.03	1.04	4.42	0.27	0.69	0.02
vog10060_sc2g3_c	0.70	-0.96	0.03	1.05	3.74	0.26	0.65	0.02
vog10009_sc2g3_c	0.56	-0.29	0.03	1.03	2.52	0.29	0.76	0.01
vog60041_sc2g3_c	0.37	0.62	0.03	1.09	8.11	0.19	0.47	0.02
vog60025_sc2g3_c	0.64	-0.65	0.03	0.99	-0.47	0.34	0.94	0.02
vok10075_sc2g3_c	0.50	0.02	0.03	1.06	5.99	0.25	0.65	0.03
vok10033_sc2g3_c	0.92	-2.69	0.05	0.95	-1.41	0.30	1.31	0.03
vog90015_sc2g3_c	0.50	-0.01	0.03	1.00	-0.16	0.33	0.87	0.02
vok10061_sc2g3_c	0.81	-1.62	0.04	0.88	-6.07	0.46	1.74	0.04
vok10065_sc2g3_c	0.62	-0.57	0.03	0.99	-0.77	0.35	0.89	0.02
vog60015_sc2g3_c	0.64	-0.67	0.03	1.06	4.68	0.26	0.64	0.03
vok10072_sc2g3_c	0.80	-1.55	0.04	1.07	3.57	0.20	0.54	0.02
vog60030_sc2g3_c	0.58	-0.35	0.03	0.90	-9.76	0.46	1.46	0.03
vog60029_sc2g3_c	0.60	-0.48	0.03	1.09	8.20	0.21	0.50	0.03
vog90003_sc2g3_c	0.59	-0.40	0.03	1.11	10.26	0.19	0.46	0.04
vog10062_sc2g3_c	0.94	-3.10	0.06	0.91	-1.85	0.33	1.80	0.03
vok10026_sc2g3_c	0.91	-2.57	0.05	0.94	-1.86	0.33	1.38	0.03
vog60037_sc2g3_c	0.51	-0.03	0.03	1.00	-0.49	0.34	0.90	0.02

Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
vok10058_sc2g3_c	0.82	-1.73	0.04	0.88	-5.54	0.44	1.61	0.04
vog60049_sc2g3_c	0.51	-0.02	0.03	1.07	7.26	0.24	0.60	0.02
vok10076_sc2g3_c	0.59	-0.39	0.03	1.06	5.5	0.26	0.63	0.02
vok10040_sc2g3_c	0.95	-3.29	0.06	0.90	-1.88	0.34	2.12	0.04
vog10040_sc2g3_c	0.46	0.18	0.03	0.93	-7.28	0.42	1.22	0.03
vok10071_sc2g3_c	0.77	-1.41	0.03	1.00	0.15	0.30	0.83	0.03
vok10060_sc2g3_c	0.76	-1.33	0.03	1.06	3.43	0.22	0.56	0.02
vog10044_sc2g3_c	0.87	-2.11	0.04	0.86	-5.35	0.46	1.95	0.04
vog60045_sc2g3_c	0.22	1.43	0.03	1.03	1.62	0.23	0.68	0.02
vog90035_sc2g3_c	0.67	-0.82	0.03	1.09	7.2	0.21	0.51	0.02
vok10074_sc2g3_c	0.62	-0.57	0.03	1.08	7.15	0.22	0.56	0.02
vog60027_sc2g3_c	0.58	-0.36	0.03	1.00	-0.44	0.34	0.87	0.01
vok10051_sc2g3_c	0.87	-2.15	0.04	1.03	1.15	0.22	0.73	0.02
vog60047_sc2g3_c	0.56	-0.28	0.03	1.05	5.15	0.27	0.67	0.02
vok10073_sc2g3_c	0.51	-0.05	0.03	1.06	6.21	0.25	0.65	0.02
vog90037_sc2g3_c	0.35	0.72	0.03	1.05	3.77	0.24	0.64	0.02
vok10038_sc2g3_c	0.93	-2.83	0.05	0.91	-2.12	0.34	1.59	0.03
vok10047_sc2g3_c	0.79	-1.54	0.04	0.97	-1.74	0.35	1.06	0.03
vok10057_sc2g3_c	0.69	-0.93	0.03	1.03	2.44	0.28	0.71	0.02
vok10046_sc2g3_c	0.85	-1.92	0.04	0.94	-2.44	0.37	1.29	0.02
vog60019_sc2g3_c	0.73	-1.11	0.03	0.91	-6.16	0.45	1.47	0.04
vok10048_sc2g3_c	0.83	-1.82	0.04	0.96	-1.67	0.34	1.11	0.03
vog90016_sc2g3_c	0.59	-0.4	0.03	1.04	3.56	0.29	0.72	0.02
vog90032_sc2g3_c	0.38	0.56	0.03	0.88	-11.77	0.47	1.64	0.03
vog60010_sc2g3_c	0.91	-2.56	0.05	0.86	-4.12	0.46	2.44	0.05
vog60032_sc2g3_c	0.67	-0.82	0.03	1.04	3.32	0.28	0.69	0.02
vog60054_sc2g3_c	0.41	0.41	0.03	1.06	5.55	0.26	0.62	0.02
vok10064_sc2g3_c	0.89	-2.41	0.05	0.90	-3.15	0.40	1.71	0.04
vog90028_sc2g3_c	0.92	-2.73	0.05	0.93	-1.69	0.34	1.55	0.03

Note. *SE* = Standard error of item difficulty parameter. WMNSQ = Weighted mean square. *t* = *t*-value for WMNSQ. r_{it} = Corrected item-total correlation. Discr. = Discrimination parameter of a generalized partial credit model. Q_3 = Average absolute residual correlation for item (Yen. 1983). Estimated parameters are based on *N* = 5,602 (Starting Cohort 2, wave 5). The item-total correlation corresponds to the product-moment correlation between the corresponding categories and the total score.

5.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulty parameters with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. In Figure 4, the item difficulties of the receptive vocabulary items and the ability of the test takers are plotted on the same scale for waves 1, 3, and 5. The distributions of the estimated test takers' ability are mapped onto the left side of each graph, whereas the right side shows the distributions of the item difficulty parameters.

In wave 1, the respective item difficulty parameters ranged from -3.15 (vok10002_c) to 1.89 (vok10065_c) with a mean of -0.89 and, thus, spanned a rather broad range. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 1.11, which implies good differentiation between children. The reliability of the test (EAP/PV reliability =.89 WLE reliability = .89) was good. Thus, although the items covered a wide range of the ability distribution, the items were slightly too easy. As a consequence, person ability in medium- and low-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors of measurement.

In wave 3, the respective item difficulty parameters ranged from -2.95 (vok10039_sc2g1_c) to 1.75 (vog90032_sc2g1_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 0.61, which implies acceptable differentiation between children. The reliability of the test (EAP/PV reliability = .87 WLE reliability = .87) was good. Thus, although the items covered a wide range of the ability distribution, the differentiation among able and less able children was limited. Additionally, the items were slightly too easy. As a consequence, person ability in medium- and low-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors of measurement.

In wave 5, the respective item difficulty parameters ranged from -3.29 (vok10040_sc2g3_c) to 1.43 (vog60045_sc2g3_c) and, thus, spanned an acceptably broad range. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 0.76, which implies acceptable differentiation between children. The reliability of the test (EAP/PV reliability =.51 WLE reliability = .84) was good. Thus, although the items covered a wide range of the ability distribution, the items were slightly too easy. As a consequence, person ability in medium- and low-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors of measurement.

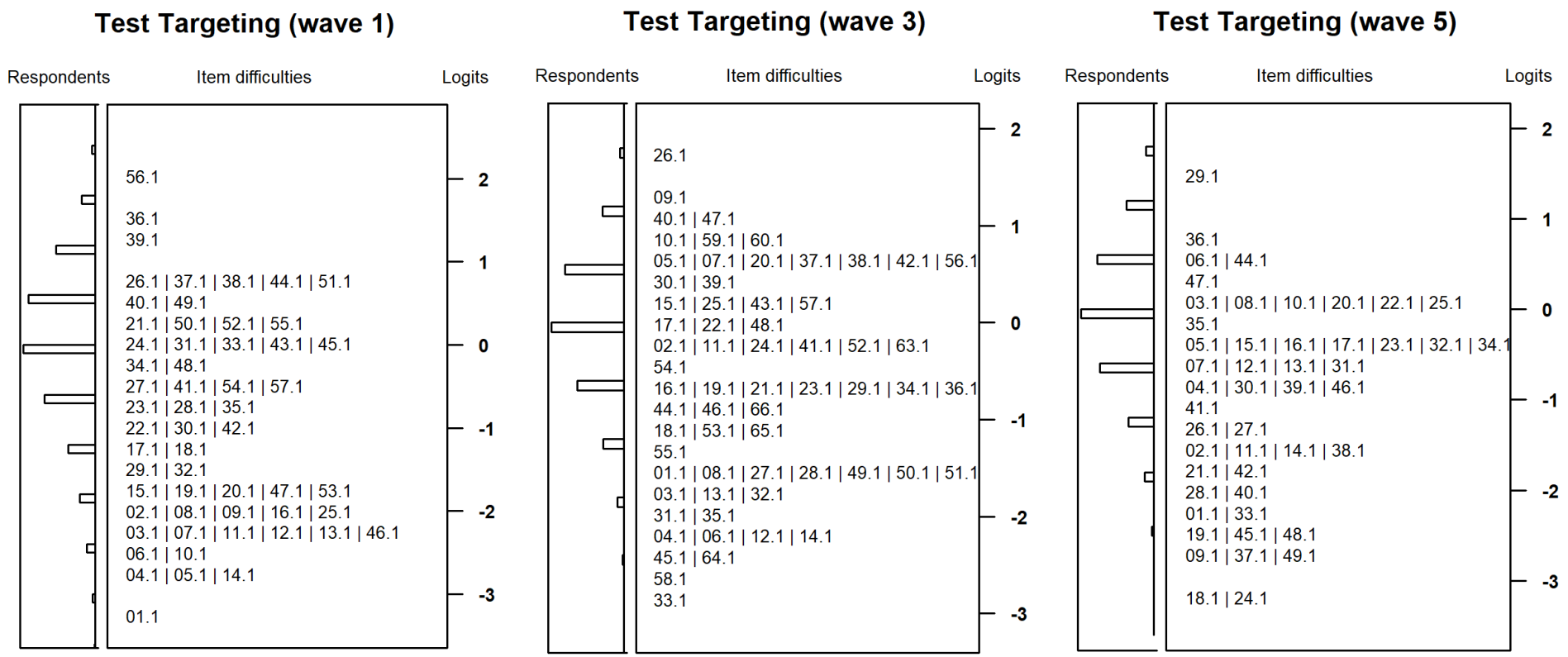

Figure 4. The distributions of person ability in the samples are given on the left-hand side of each graph. The item difficulties are given on the right-hand side of each graph. Each number represents one item corresponding to the sequential IDs in Tables 2, 3, and 4.

5.3 Quality of the Test

5.3.1 Item fit

Because the receptive vocabulary test comprises of a subset of items belonging to the PPVT (Lenhard et al., 2015), the scaling process focused on item maintenance rather than item selection in order to retain test forms as complete as possible.

All items showed a satisfactory item fit. WMNSQ ranged in wave 1 from 0.85 to 1.15 (the respective *t*-value ranged from -7.40 to 8.87), in wave 3 from 0.86 to 1.16 (the respective *t*-value ranged from -12.48 to 14.76), and in wave 5 from 0.86 to 1.11 (the respective *t*-value ranged from -11.77 to 10.26). There were no noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. Point-biserial correlations between the correct response and the total score ranged in wave 1 from .14 to .54 (*M* = .34, *SD* = .10), in wave 3 from .08 to .49 (*M* = .29, *SD* = .10), and in wave 5 from .19 to .47 (*M* = .32, *SD* = .08).

5.3.2 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables sex, the number of books at home (as a proxy for socioeconomic status), migration background, test position (wave 3 only), and the common items included in the tests administered in waves 3 and 5 (see Pohl & Carstensen, 2012, for a description of these variables). The differences between the estimated item difficulty parameters in the various groups are summarized in Tables 8 (wave 1), 10 (wave 3), and 12 (wave 5). For example, the column “Boys vs. girls” reports the differences in item difficulty parameters 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 boys as opposed to girls. Besides investigating DIF for each 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 9 for wave 1, Table 11 for wave 3, and Table 13 for wave 5).

5.3.2.1 Differential item functioning in wave 1

Sex: The sample included 1,451 (51%) boys and 1,408 (49%) girls. On average, boys had a comparable receptive vocabulary ability to girls (main effect = -0.02 logits, Cohen’s d = -0.02). However, there was considerable and statistical significant ($p < .05$, greater than 0.40; see Fischer et al., 2016) DIF for 5 out of the 57 items: “vok10002_c”, “vok10015_c”, “vok10037_c”, “vok10039_c” and “vok10054_c”. As the DIF found on item level did not affect the main effect of the DIF model (main effect = -0.02 logits, Cohen’s *d* = -0.02) the respective items showing DIF were not excluded from the model. An overall test for DIF (see Table 9) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). However, model comparisons using Akaike’s (1974) information criterion (AIC) and the Bayesian information criterion (BIC; Schwarz, 1978) both favored the model estimating DIF (Table 9). Nevertheless, the deviation was small in both cases. Thus, overall, there was negligible DIF regarding the gender of the children.

Table 8

Differential Item Functioning in Wave 1

Item	Sex	Books	Migration
	boys vs. girls	≤ 100 vs. > 100	without vs. with
vok10002_c	0.89* (0.84)	-0.20 (-0.22)	0.18 (0.19)
vok10007_c	0.31 (0.29)	0.28 (0.30)	-0.16 (-0.17)
vok10008_c	0.30 (0.28)	-0.30 (-0.33)	0.42 (0.46)
vok10009_c	-0.15 (-0.14)	0.23 (0.25)	-0.54 (-0.58)
vok10010_c	-0.32 (-0.30)	0.25 (0.27)	-0.01 (-0.01)
vok10011_c	0.30 (0.29)	-0.05 (-0.05)	-0.42 (-0.46)
vok10012_c	0.17 (0.16)	0.30 (0.32)	-0.45 (-0.48)
vok10013_c	0.04 (0.04)	-0.09 (-0.10)	-0.13 (-0.14)
vok10014_c	-0.49 (-0.47)	-0.08 (-0.09)	-0.15 (-0.16)
vok10015_c	-0.73* (-0.69)	0.72* (0.78)	-0.24 (-0.26)
vok10016_c	0.06 (0.06)	0.22 (0.24)	-0.10 (-0.10)
vok10017_c	-0.10 (-0.10)	-0.10 (-0.11)	-0.16 (-0.18)
vok10018_c	0.23 (0.22)	0.40 (0.43)	-0.27 (-0.29)
vok10019_c	0.05 (0.05)	0.10 (0.11)	-0.03 (-0.03)
vok10020_c	0.15 (0.14)	-0.02 (-0.03)	-0.10 (-0.11)
vok10021_c	0.50 (0.47)	0.42 (0.46)	-0.55 (-0.59)
vok10022_c	0.26 (0.25)	-0.23 (-0.25)	0.21 (0.22)
vok10023_c	0.33 (0.31)	-0.40 (-0.44)	0.66* (0.72)
vok10024_c	-0.29 (-0.27)	0.26 (0.29)	-0.13 (-0.14)
vok10025_c	0.26 (0.25)	0.49 (0.53)	-1.03* (-1.11)
vok10026_c	0.14 (0.14)	-0.19 (-0.20)	0.29 (0.31)
vok10027_c	0.26 (0.25)	0.51 (0.56)	-0.84* (-0.91)
vok10028_c	-0.21 (-0.20)	0.66* (0.71)	-0.74* (-0.79)
vok10031_c	-0.36 (-0.34)	0.19 (0.20)	0.28 (0.30)
vok10032_c	0.18 (0.17)	0.13 (0.14)	-0.69* (-0.75)
vok10033_c	0.02 (0.02)	-0.37 (-0.40)	0.80* (0.86)

Item	Sex	Books	Migration
vok10034_c	-0.51 (-0.48)	-0.63* (-0.68)	0.44 (0.48)
vok10035_c	-0.11 (-0.11)	-0.30 (-0.32)	0.49 (0.53)
vok10036_c	0.10 (0.09)	-0.15 (-0.17)	0.10 (0.11)
vok10037_c	0.79* (0.74)	-0.27 (-0.30)	0.17 (0.18)
vok10038_c	0.01 (0.00)	0.35 (0.38)	-0.30 (-0.32)
vok10039_c	-0.93* (-0.88)	-0.31 (-0.34)	0.14 (0.15)
vok10040_c	0.14 (0.13)	0.16 (0.17)	-0.07 (-0.08)
vok10041_c	0.16 (0.15)	0.14 (0.15)	-0.50 (-0.54)
vok10042_c	-0.14 (-0.14)	-0.07 (-0.08)	0.27 (0.29)
vok10043_c	-0.14 (-0.13)	-0.10 (-0.11)	0.34 (0.36)
vok10045_c	0.39 (0.37)	-0.14 (-0.15)	0.14 (0.16)
vok10046_c	0.27 (0.26)	-0.09 (-0.10)	0.15 (0.16)
vok10047_c	-0.13 (-0.12)	-0.10 (-0.11)	0.51 (0.55)
vok10048_c	-0.18 (-0.17)	0.13 (0.14)	-0.01 (-0.01)
vok10049_c	-0.40 (-0.37)	0.41 (0.45)	-0.57 (-0.61)
vok10050_c	-0.24 (-0.23)	-0.58 (-0.63)	0.67* (0.72)
vok10051_c	0.33 (0.32)	-0.28 (-0.30)	0.25 (0.26)
vok10052_c	0.15 (0.14)	-0.16 (-0.18)	0.10 (0.11)
vok10053_c	0.01 (0.01)	-0.07 (-0.07)	0.37 (0.40)
vok10054_c	-0.95* (-0.90)	-0.02 (-0.02)	0.08 (0.09)
vok10055_c	-0.20 (-0.19)	0.33 (0.36)	-0.53 (-0.57)
vok10056_c	0.15 (0.14)	0.03 (0.03)	0.08 (0.09)
vok10057_c	0.14 (0.14)	-0.16 (-0.17)	0.32 (0.35)
vok10058_c	-0.12 (-0.11)	0.45 (0.49)	-0.29 (-0.31)
vok10060_c	-0.45 (-0.42)	-0.18 (-0.19)	0.42 (0.45)
vok10061_c	0.21 (0.20)	-0.06 (-0.06)	-0.16 (-0.17)
vok10062_c	-0.01 (-0.01)	-0.43 (-0.46)	0.37 (0.40)
vok10063_c	-0.27 (-0.26)	-0.34 (-0.37)	0.20 (0.22)
vok10064_c	-0.18 (-0.17)	-0.15 (-0.16)	0.05 (0.05)
vok10065_c	-0.13 (-0.12)	-0.23 (-0.24)	0.25 (0.27)

Item	Sex	Books	Migration
vok10066_c	0.42 (0.40)	-0.30 (-0.32)	0.41 (0.44)
Main effect (DIF model)	-0.02 (-0.02)	0.79 (0.86)	-1.13 (-1.21)
Main effect (Main effect model)	-0.02 (-0.02)	0.76 (0.83)	-1.15 (-1.24)

Note. Raw differences between item difficulty parameters with standardized differences (Cohen's *d*) in parentheses. Absolute standardized differences marked with an asterisk are significantly (p < .05) greater than 0.40 (see Fischer et al., 2016).

Books: The number of books at home was used as a proxy for socioeconomic status. There were 1,159 (41%) test takers with 0 to 100 books at home, 1,226 (43%) test takers with more than 100 books at home, and 474 (17%) test takers without a valid response. There were considerable average differences between the two groups. For children with 100 or less books at home the receptive vocabulary test was on average 0.76 logits (Cohen's d = 0.83) more difficult than for children with more than 100 books. However, there was considerable and statistical significant (p < .05, greater than 0.40; see Fischer et al., 2016) DIF comparing children with 100 or less books at home and children with more than 100 books at home for three items: "vok10015_c", "vok10028_c" and "vok10034_c". Whereas the AIC favored the model estimating DIF, the BIC favored the main effects model (Table 9).

Migration background: There were 2,128 children (74%) with no migration background, 329 children (12%) with a migration background and 402 (14%) test takers without respective information. In comparison to children with migration background, the receptive vocabulary competence test was easier for children without migration background (main effect = -1.15 logits, Cohen's d = -1.24). However, there was considerable and statistical significant (p < .05, greater than 0.40; see Fischer et al., 2016) DIF comparing children with and without migration background for the items "vok10023_c", "vok10025_c", "vok10027_c", "vok10028_c", "vok10032_c", "vok10033_c" and "vok10050_c". As the DIF found on item level did not considerably affect the main effect of the DIF model (main effect = -1.13 logits, Cohen's *d* = -1.21) the respective items showing DIF were not excluded from the model. The overall test for DIF using the BIC favored the main effects model, while the AIC favored the model estimating DIF (Table 9).

Table 9

Comparison of Models with and without DIF in Wave 1

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	DIF	2,859	147,420	115	147,650	148,335
	Main effect	2,859	148,012	59	148,130	148,482
Books	DIF	2,457	126,019	115	126,249	126,917
	Main effect	2,457	126,403	59	126,521	126,864
Migration	DIF	2,385	122,112	115	122,342	123,006
	Main effect	2,385	122,524	59	122,642	122,983

5.3.2.2 Differential item functioning in wave 3

Sex: The sample included 3,180 (49%) boys and 3,291 (51%) girls. On average, the receptive vocabulary competence test was slightly more difficult for boys compared to girls (main effect = 0.08 logits, Cohen's d = 0.10). However, there was considerable and/or statistical significant (p < .05, greater than 0.40; see Fischer et al., 2016) DIF in 7 out of the 66 items: "vog90015_sc2g1_c", "vok10072_sc2g1_c", "vok10039_sc2g1_c", "vok10034_sc2g1_c", "vok10071_sc2g1_c", "vog90028_sc2g1_c" and "vok10037_sc2g1_c". As the DIF found on item level did not affect the model main effect (main effect = 0.08 logits, Cohen's *d* = 0.09) the respective items showing DIF were not excluded from the model. An overall test for DIF (see Table 9) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). However, model comparisons using the information criteria favored the model estimating DIF (Table 11). Nevertheless, the deviation was small in both cases. Thus, overall, there was acceptable DIF regarding the gender of the children.

Books: The number of books at home was used as a proxy for socioeconomic status. There were 2,029 (31%) test takers with 0 to 100 books at home, 3,633 (56%) test takers with more than 100 books at home, and 809 (13%) test takers without a valid response. There were considerable average differences between the two groups. For children with 100 or less books at home the receptive vocabulary competence test was on average 0.52 logits (Cohen's d = 0.73) more difficult than for children with more than 100 books. There was considerable and statistical significant (p < .05, greater than 0.40; see Fischer et al., 2016) DIF comparing children with 100 or less books at home and children with more than 100 books at home for 3 out of 66 items: "vok10040_sc2g1_c", "vok10038_sc2g1_c" and "vog90020_sc2g1_c". However, model comparisons using the information criteria favored the model estimating DIF (Table 11). Nevertheless, the deviation was small. Thus, overall, there was acceptable DIF regarding the number of books at home of the children.

Table 10

Differential Item Functioning in Wave 3

Item	Sex	Books	Migration	Position
	boys vs. girls	≤ 100 vs. > 100	without vs. with	
vok10067_sc2g1_c	0.15 (0.19)	-0.24 (-0.34)	0.42 (0.58)	-0.10 (-0.13)
vok10043_sc2g1_c	0.02 (0.03)	0.50 (0.70)	-0.29 (-0.40)	0.00 (0.00)
vok10053_sc2g1_c	0.13 (0.16)	-0.19 (-0.27)	0.14 (0.20)	0.29 (0.37)
vok10049_sc2g1_c	-0.08 (-0.11)	0.35 (0.49)	-0.65* (-0.90)	-0.02 (-0.03)
vog60001_sc2g1_c	-0.05 (-0.06)	-0.27 (-0.37)	0.47 (0.65)	0.09 (0.11)
vok10025_sc2g1_c	0.23 (0.29)	0.17 (0.24)	-0.36 (-0.49)	0.17 (0.21)
vok10076_sc2g1_c	0.21 (0.27)	-0.35 (-0.49)	0.56* (0.78)	0.06 (0.08)
vok10050_sc2g1_c	-0.06 (-0.08)	-0.28 (-0.39)	0.38 (0.53)	-0.05 (-0.06)
vog10009_c	0.05 (0.06)	-0.25 (-0.36)	0.71* (0.99)	0.10 (0.12)
vog60009_sc2g1_c	-0.10 (-0.13)	-0.33 (-0.46)	0.46 (0.63)	0.31 (0.40)
vok10060_sc2g1_c	-0.46 (-0.59)	-0.40 (-0.56)	0.17 (0.24)	0.08 (0.10)
vok10066_sc2g1_c	-0.13 (-0.17)	-0.40 (-0.57)	0.35 (0.49)	-0.12 (-0.16)
vok10063_sc2g1_c	-0.01 (-0.02)	0.24 (0.34)	-0.40 (-0.56)	-0.14 (-0.18)
vok10040_sc2g1_c	0.13 (0.16)	0.65* (0.91)	-0.78* (-1.08)	-0.03 (-0.04)
vok10074_sc2g1_c	0.21 (0.26)	-0.04 (-0.06)	0.27 (0.37)	-0.01 (-0.01)
vok10033_sc2g1_c	0.25 (0.31)	-0.06 (-0.08)	0.08 (0.11)	-0.06 (-0.07)
vog90015_sc2g1_c	-0.61*(-0.78)	-0.15 (-0.21)	0.23 (0.32)	0.22 (0.29)
vok10051_sc2g1_c	0.34 (0.43)	-0.06 (-0.08)	0.15 (0.21)	0.05 (0.07)
vok10061_sc2g1_c	0.35 (0.45)	0.29 (0.41)	-0.53* (-0.74)	0.11 (0.14)
vog60051_sc2g1_c	0.12 (0.15)	-0.45 (-0.62)	0.55* (0.76)	0.07 (0.09)
vog90007_sc2g1_c	-0.14 (-0.17)	-0.40 (-0.56)	0.73* (1.02)	0.11 (0.14)
vog60015_sc2g1_c	-0.01 (-0.02)	-0.40 (-0.56)	0.61* (0.85)	-0.04 (-0.05)
vok10057_sc2g1_c	0.21 (0.27)	-0.10 (-0.14)	0.05 (0.07)	-0.02 (-0.02)
vok10072_sc2g1_c	-0.64* (-0.82)	-0.23 (-0.32)	0.22 (0.31)	0.03 (0.04)
vog90016_sc2g1_c	0.30 (0.39)	-0.32 (-0.45)	0.32 (0.45)	-0.14 (-0.18)
vog90032_sc2g1_c	-0.43 (-0.55)	0.34 (0.48)	-0.06 (-0.08)	0.24 (0.31)

Item	Sex	Books	Migration	Position
vog60010_sc2g1_c	0.20 (0.25)	0.51 (0.72)	-0.96* (-1.33)	-0.11 (-0.15)
vok10041_sc2g1_c	0.34 (0.44)	0.00 (0.01)	-0.11 (-0.15)	-0.32 (-0.41)
vok10052_sc2g1_c	0.11 (0.14)	0.21 (0.30)	-0.76* (-1.06)	-0.07 (-0.09)
vog60032_sc2g1_c	-0.05 (-0.06)	-0.12 (-0.17)	0.34 (0.47)	-0.12 (-0.16)
vok10031_sc2g1_c	-0.22 (-0.29)	0.04 (0.06)	0.10 (0.15)	-0.07 (-0.09)
vok10045_sc2g1_c	0.51 (0.66)	0.15 (0.20)	-0.28 (-0.39)	-0.07 (-0.10)
vok10039_sc2g1_c	-0.60 (-0.77)	-0.05 (-0.07)	-0.44 (-0.61)	-0.16 (-0.20)
vog10034_c	0.25 (0.31)	0.46 (0.65)	-0.44 (-0.61)	0.07 (0.09)
vok10034_sc2g1_c	-0.68* (-0.87)	-0.20 (-0.28)	0.40 (0.56)	0.04 (0.05)
vok10058_sc2g1_c	0.01 (0.01)	0.38 (0.53)	-0.89* (-1.24)	0.06 (0.08)
vog90031_sc2g1_c	-0.08 (-0.10)	0.11 (0.15)	-0.15 (-0.21)	0.01 (0.02)
vog60049_sc2g1_c	0.12 (0.15)	-0.13 (-0.19)	0.20 (0.28)	0.03 (0.04)
vok10065_sc2g1_c	-0.24 (-0.31)	0.26 (0.36)	-0.41 (-0.58)	0.04 (0.05)
vog10040_c	-0.06 (-0.07)	0.25 (0.35)	-0.27 (-0.38)	0.01 (0.01)
vok10071_sc2g1_c	-0.66* (-0.85)	-0.34 (-0.47)	0.26 (0.36)	0.10 (0.13)
vok10069_sc2g1_c	-0.03 (-0.04)	-0.13 (-0.19)	0.47 (0.66)	-0.10 (-0.13)
vog60025_sc2g1_c	0.23 (0.30)	0.09 (0.12)	0.11 (0.15)	-0.05 (-0.07)
vog10044_c	-0.50 (-0.64)	0.09 (0.13)	-0.24 (-0.34)	-0.05 (-0.06)
vok10028_sc2g1_c	-0.01 (-0.01)	0.47 (0.66)	-0.92* (-1.28)	-0.01 (-0.02)
vog10046_c	-0.02 (-0.02)	-0.14 (-0.20)	0.02 (0.03)	0.06 (0.07)
vog60027_sc2g1_c	0.04 (0.05)	-0.10 (-0.14)	0.51 (0.70)	0.06 (0.08)
vog60047_sc2g1_c	-0.03 (-0.03)	-0.12 (-0.16)	0.24 (0.34)	-0.12 (-0.16)
vok10022_sc2g1_c	0.42 (0.53)	-0.34 (-0.48)	0.45 (0.63)	-0.19 (-0.25)
vok10038_sc2g1_c	0.17 (0.22)	0.64* (0.89)	-0.73* (-1.02)	0.06 (0.07)
vog90028_sc2g1_c	-0.66* (-0.85)	-0.05 (-0.07)	0.07 (0.09)	-0.07 (-0.09)
vok10047_sc2g1_c	-0.07 (-0.09)	-0.09 (-0.13)	0.25 (0.35)	0.06 (0.07)
vok10046_sc2g1_c	0.14 (0.18)	0.33 (0.46)	-0.26 (-0.36)	0.04 (0.05)
vog60019_sc2g1_c	0.20 (0.26)	0.47 (0.66)	-0.48 (-0.67)	-0.06 (-0.07)
vok10048_sc2g1_c	0.13 (0.16)	0.45 (0.63)	-0.38 (-0.53)	-0.02 (-0.02)
vog10056_c	-0.02 (-0.02)	-0.42 (-0.59)	0.46 (0.64)	0.02 (0.02)

Item	Sex	Books	Migration	Position
vog90020_sc2g1_c	-0.26 (-0.33)	-0.63* (-0.88)	0.84* (1.16)	-0.08 (-0.10)
vok10037_sc2g1_c	0.75* (0.96)	-0.01 (-0.01)	-0.10 (-0.14)	-0.18 (-0.23)
vog60030_sc2g1_c	-0.04 (-0.05)	0.30 (0.42)	-0.10 (-0.14)	0.06 (0.08)
vog10060_c	0.21 (0.27)	-0.32 (-0.45)	0.54* (0.75)	-0.09 (-0.11)
vok10077_sc2g1_c	0.04 (0.05)	-0.29 (-0.41)	0.19 (0.26)	-0.08 (-0.10)
vog10062_c	0.04 (0.05)	0.13 (0.18)	-0.52 (-0.73)	0.01 (0.01)
vog10063_c	0.04 (0.05)	-0.28 (-0.39)	0.47 (0.65)	-0.12 (-0.15)
vok10042_sc2g1_c	-0.20 (-0.26)	0.36 (0.50)	-0.46 (-0.64)	0.04 (0.05)
vok10064_sc2g1_c	0.06 (0.07)	0.15 (0.21)	-0.61* (-0.84)	0.12 (0.15)
vok10026_sc2g1_c	0.50 (0.64)	0.30 (0.42)	-0.21 (-0.29)	0.06 (0.08)
Main effect (DIF model)	0.08 (0.09)	0.55 (0.77)	-0.83 (-1.16)	0.04 (0.05)
Main effect (Main effect model)	0.08 (0.10)	0.52 (0.73)	-0.82 (-1.15)	0.03 (0.04)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. Absolute standardized differences marked with an asterisk are significantly ($p < .05$) greater than 0.40 (see Fischer et al., 2016).

Table 11

Comparison of Models with and without DIF in Wave 3

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	DIF	6,471	447,885	133	448,151	449,052
	Main effect	6,471	449,318	68	449,454	449,914
Books	DIF	5,662	386,250	133	386,516	387,399
	Main effect	5,662	387,577	68	387,713	388,164
Migration	DIF	5,823	398,065	133	398,331	399,218
	Main effect	5,823	399,349	68	399,485	399,939
Position	DIF	6,462	448,496	133	448,762	449,663
	Main effect	6,462	448,703	68	448,839	449,299

Migration background: There were 5,227 children (81%) with no migration background, 596 children (9%) with a migration background and 648 (10%) children without respective information. In comparison to children with migration background, the receptive vocabulary test was easier for children without migration background (main effect = -0.82 logits, Cohen's d = -1.15). However, there was considerable and/or statistical significant (p < .05, greater than 0.40; see Fischer et al., 2016) DIF comparing children with and without migration background in the items "vok10049_sc2g1_c", "vok10076_sc2g1_c", "vog10009_c", "vok10040_sc2g1_c", "vok10061_sc2g1_c", "vog60051_sc2g1_c", "vog90007_sc2g1_c", "vog60015_sc2g1_c", "vog60010_sc2g1_c", "vok10052_sc2g1_c", "vok10058_sc2g1_c", "vok10028_sc2g1_c", "vok10038_sc2g1_c", "vog90020_sc2g1_c", "vog10060_c" and "vok10064_sc2g1_c". As the DIF found on item level did not considerably change the main effect of the DIF model (main effect = -0.83 logits, Cohen's *d* =-1.16) the respective items showing DIF were not excluded from the model. However, model comparisons using the information criteria both favored the model estimating DIF (Table 11).

Test position: In wave 3, the receptive vocabulary test was administered in two different positions (see section 3 for the design of the study). A sample of 3,271 (51%) persons received the receptive vocabulary test on the first position and 3,191 (49%) children took the receptive vocabulary test after having completed the declarative metacognition test. Differential item functioning of the position of the test may, for example, occur if there are differential fatigue effects for certain items. The results show negligible effects of item position. In comparison to children who received the receptive vocabulary competence test second, the test was slightly more difficult for children who received the receptive vocabulary competence test first (main effect = 0.03 logits, Cohen's d = 0.04). Note that this main effect does not indicate a threat to measurement invariance. Instead, it may be an indication of fatigue effects that are similar for all items. There was no DIF due to the position of the test in the booklet. The largest difference in difficulty between the two design groups was |-0.32| logits (item "vok10041_sc2g1_c"). The overall test for DIF using the BIC favored the main effects model, while the AIC favored the model estimating DIF.

5.3.2.3 Differential item functioning in wave 5

Sex: The sample included 2,735 (49%) boys and 2,866 (51%) girls, and 1 (0%) participant without a valid response. On average, the receptive vocabulary test was slightly more difficult for boys as compared to girls (main effect = 0.09 logits, Cohen's d = 0.10). However, there was considerable and/or statistical significant (p < .05, greater than 0.40; see Fischer et al., 2016) DIF comparing boys and girls for 5 out of 49 items: "vog90015_sc2g3_c", "vog90003_sc2g3_c", "vok10026_sc2g3_c", "vog10044_sc2g3_c" and "vog60054_sc2g3_c". As the DIF found on item level did not change the main effect of the DIF model (main effect = 0.08 logits, Cohen's *d* = 0.09) the respective items showing DIF were not excluded from the model. An overall test for DIF (see Table 13) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). However, model comparisons using the information criteria favored the model estimating DIF. Nevertheless, the deviation was small. Thus, overall, there was acceptable DIF regarding the gender of the children.

Table 12

Differential Item Functioning in Wave 5

Item	Sex	Books	Migration
	boys vs. girls	≤ 100 vs. > 100	without vs. with
vog10034_sc2g3_c	0.23 (0.27)	0.62 (0.80)	-0.23 (-0.29)
vok10043_sc2g3_c	0.15 (0.17)	0.48 (0.61)	-0.42 (-0.53)
vog90031_sc2g3_c	-0.14 (-0.16)	0.07 (0.09)	-0.16 (-0.20)
vog10060_sc2g3_c	0.38 (0.44)	-0.60 (-0.77)	0.22 (0.28)
vog10009_sc2g3_c	0.20 (0.22)	-0.19 (-0.25)	0.03 (0.04)
vog60041_sc2g3_c	0.08 (0.09)	-0.42 (-0.54)	0.75* (0.94)
vog60025_sc2g3_c	0.22 (0.25)	0.05 (0.06)	0.13 (0.16)
vok10075_sc2g3_c	0.36 (0.41)	0.08 (0.10)	0.39 (0.50)
vok10033_sc2g3_c	0.03 (0.03)	0.08 (0.11)	0.08 (0.10)
vog90015_sc2g3_c	-0.79* (-0.90)	0.02 (0.03)	-0.05 (-0.06)
vok10061_sc2g3_c	0.42 (0.48)	0.27 (0.35)	-0.57 (-0.72)
vok10065_sc2g3_c	-0.23 (-0.26)	0.05 (0.06)	0.02 (0.03)
vog60015_sc2g3_c	-0.13 (-0.15)	-0.44 (-0.56)	0.81* (1.01)
vok10072_sc2g3_c	-0.42 (-0.48)	-0.42 (-0.53)	0.51 (0.65)
vog60030_sc2g3_c	0.18 (0.20)	0.23 (0.30)	-0.19 (-0.24)
vog60029_sc2g3_c	-0.15 (-0.17)	-0.47 (-0.60)	0.52 (0.65)
vog90003_sc2g3_c	-1.28* (-1.46)	-0.63* (-0.80)	0.65* (0.82)
vog10062_sc2g3_c	0.02 (0.02)	0.46 (0.59)	-0.58 (-0.73)
vok10026_sc2g3_c	0.87* (0.99)	0.28 (0.36)	0.03 (0.03)
vog60037_sc2g3_c	0.10 (0.12)	0.05 (0.06)	-0.50 (-0.63)
vok10058_sc2g3_c	0.01 (0.01)	0.43 (0.55)	-0.95* (-1.19)
vog60049_sc2g3_c	0.20 (0.22)	-0.17 (-0.22)	0.15 (0.19)
vok10076_sc2g3_c	0.23 (0.26)	-0.34 (-0.44)	0.91* (1.14)
vok10040_sc2g3_c	0.10 (0.11)	0.84 (1.07)	-0.59 (-0.74)
vog10040_sc2g3_c	-0.08 (-0.09)	0.00 (0.00)	-0.01 (-0.01)
vok10071_sc2g3_c	-0.62 (-0.71)	-0.51 (-0.65)	0.33 (0.41)

Item	Sex	Books	Migration
vok10060_sc2g3_c	-0.47 (-0.54)	-0.53 (-0.68)	0.31 (0.38)
vog10044_sc2g3_c	-0.70 (-0.80)	0.28 (0.36)	-0.33 (-0.41)
vog60045_sc2g3_c	0.06 (0.06)	-0.07 (-0.08)	0.36 (0.45)
vog90035_sc2g3_c	0.37 (0.43)	-0.26 (-0.34)	0.26 (0.33)
vok10074_sc2g3_c	0.16 (0.19)	-0.43 (-0.56)	0.43 (0.54)
vog60027_sc2g3_c	0.13 (0.15)	-0.14 (-0.18)	0.13 (0.16)
vok10051_sc2g3_c	0.53 (0.61)	0.03 (0.04)	0.34 (0.43)
vog60047_sc2g3_c	-0.07 (-0.08)	-0.17 (-0.22)	0.51 (0.64)
vok10073_sc2g3_c	0.48 (0.55)	-0.21 (-0.27)	0.20 (0.25)
vog90037_sc2g3_c	-0.28 (-0.32)	-0.20 (-0.26)	0.39 (0.49)
vok10038_sc2g3_c	0.22 (0.25)	0.33 (0.43)	-0.60 (-0.75)
vok10047_sc2g3_c	-0.20 (-0.23)	-0.12 (-0.15)	0.05 (0.07)
vok10057_sc2g3_c	0.08 (0.10)	-0.12 (-0.15)	-0.09 (-0.12)
vok10046_sc2g3_c	0.16 (0.19)	0.25 (0.32)	-0.39 (-0.48)
vog60019_sc2g3_c	0.33 (0.37)	0.54 (0.69)	-0.60 (-0.76)
vok10048_sc2g3_c	0.45 (0.51)	0.20 (0.26)	-0.13 (-0.16)
vog90016_sc2g3_c	0.34 (0.38)	-0.25 (-0.32)	0.18 (0.22)
vog90032_sc2g3_c	-0.21 (-0.24)	0.59 (0.76)	-0.34 (-0.42)
vog60010_sc2g3_c	0.18 (0.20)	0.71 (0.91)	-1.03* (-1.29)
vog60032_sc2g3_c	-0.01 (-0.01)	-0.16 (-0.20)	0.30 (0.37)
vog60054_sc2g3_c	-0.78* (-0.89)	-0.41 (-0.52)	-0.17 (-0.21)
vok10064_sc2g3_c	-0.01 (-0.01)	0.36 (0.46)	-0.92* (-1.15)
vog90028_sc2g3_c	-0.71 (-0.81)	-0.06 (-0.07)	-0.15 (-0.19)
Main effect (DIF model)	0.08 (0.09)	0.69 (0.88)	-0.94 (-1.18)
Main effect (Main effect model)	0.09 (0.10)	0.61 (0.78)	-0.89 (-1.11)

Note. Raw differences between item difficulty parameters with standardized differences (Cohen's *d*) in parentheses. Absolute standardized differences marked with an asterisk are significantly (p < .05) greater

Item	Sex	Books	Migration

than 0.40 (see Fischer et al., 2016).

Books: The number of books at home was used as a proxy for socioeconomic status. There were 1,736 (31%) test takers with 0 to 100 books at home, 3,213 (57%) test takers with more than 100 books at home, and 653 (12%) test takers without a valid response. There were considerable and/or statistical significant (p < .05, greater than 0.40; see Fischer et al., 2016) average differences between the two groups. For children with 100 or less books at home the receptive vocabulary competence test was on average 0.61 logits (Cohen's d = 0.78) more difficult than for children with more than 100 books at home. However, there was considerable DIF comparing children with 100 or less books at home and children with more than 100 books at home in two items: "vog90003_sc2g3_c" and "vog60010_sc2g3_c". However, model comparisons using the information criteria favored the model estimating DIF (Table 13).

Migration background: There were 4,589 children (82%) with no migration background, 504 children (9%) with a migration background and 509 (9%) test takers without respective information. In comparison to children with migration background, the receptive vocabulary competence test was easier for children without migration background (main effect = -0.89 logits, Cohen's d = -1.11). However, there was considerable and/or statistical significant (p < .05, greater than 0.40; see Fischer et al., 2016) DIF comparing children for 9 out of 49 items: "vog60041_sc2g3_c", "vog60015_sc2g3_c", "vog90003_sc2g3_c", "vok10058_sc2g3_c", "vok10076_sc2g3_c", "vok10038_sc2g3_c" "vog60019_sc2g3_c", "vog60010_sc2g3_c", and "vok10064_sc2g3_c". As the DIF found on item level did not considerably change the main effect of the DIF model (main effect (DIF model) = -0.94 logits, Cohen's *d* =-1.18) the respective items showing DIF were not excluded from the model. Model comparisons using the information criteria favored the model estimating DIF (Table 13).

Table 13

Comparison of Models with and without DIF in Wave 5

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	DIF	5,601	276,299	99	276,497	277,153
	Main effect	5,601	277,977	51	278,079	278,418
Books	DIF	4,949	239,903	99	240,101	240,745
	Main effect	4,949	240,859	51	240,961	241,293
Migration	DIF	5,093	248,040	99	248,238	248,885
	Main effect	5,093	248,806	51	248,908	249,241

5.3.3 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 (2PL; Birnbaum, 1968) model that estimates discrimination parameters was fitted to the data. The estimated discriminations varied between 0.41 and 2.05 (*M* = 1.11, *SD* = 0.45) in wave 1, between 0.17 and 1.95 (*M* = 0.88, *SD* = 0.41) in wave 3, and between 0.46 and 2.44 (*M* = 1.05, *SD* = 0.51) in wave 5. For all three waves model fit indices suggested a slightly better fit of the 2PL (wave 1: AIC = 146,387, BIC = 147,067; wave 3: AIC = 443,271, BIC = 444,165; wave 5: AIC = 274,225, BIC = 274,875) as compared to the Rasch model (wave 1: AIC = 148,129, BIC = 148,475; wave 3: AIC = 449,466, BIC = 449,920; wave 5: AIC = 278,126, BIC = 278,458). Despite the empirical preference for the 2PL model, the Rasch model more adequately matches the theoretical conceptions underlying the test construction (see Lenhard et al., 2015). For this reason, the Rasch model was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

6 Discussion

The analyses in the previous sections reported information on the quality of the receptive vocabulary competence tests that were administered in waves 1, 3, and 5 of Starting Cohort 2 to children attending Kindergarten, grade 1, and grade 3 in Germany. Furthermore, the estimation of the respective receptive vocabulary competence scores was described. 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. The number of missing responses was rather low. The tests had satisfactory reliabilities and distinguished acceptably between test takers. However, the tests were slightly better targeted at medium and low performing students and covered the high ability spectrum less well. As a consequence, ability estimates will be precise for low-performing children but less precise for high performing children. In summary, the tests had acceptable psychometric properties that allowed the estimation of unidimensional receptive vocabulary competence scores.

7 Data in the Scientific Use File

7.1 Naming conventions

The data in the Scientific Use File contains 77 items in wave 1, 66 items in wave 3, and 72 items in wave 5. All items (marked with a '_c' at the end of the variable name) were scored dichotomously, 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 (2019).

7.2 Linking of competence scores

In waves 1, 3, and 5 of Starting Cohort 2 the children attending Kindergarten, grade 1, and grade 3 were administered different test forms that were constructed in such a way as to allow for an accurate measurement of receptive vocabulary competence within the respec-

tive age group (Berendes et al., 2013). As a consequence, the competence scores derived in the different waves cannot be directly compared; differences in observed scores would reflect differences in competences as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for the longitudinal comparison of competences across waves, the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016) was adopted. As such, an anchor-group design was applied to link the vocabulary scores between waves 1 and 3. These relied on an independent link sample that was not part of the main study and received items from the receptive vocabulary test forms of waves 1 and 3 within a single measurement occasion. These responses were used to link the tests across the two waves. The anchor-group design was chosen to account for the change in test situations (individual versus group setting). To link waves 3 and 5, an anchor-items design was applied because the two tests shared a number of common items. In the following, the two linking procedures will be described in greater detail.

7.2.1 Linking of waves 1 and 3

A subsample of 528 children (49% girls) participated at both measurement occasions, in wave 1 (i.e., Kindergarten) and also in wave 3 (i.e., grade 1). Consequently, these children were used to link the two test forms across both waves (see Fischer et al., 2016). Moreover, an independent link sample of *N* = 437 children (49% girls) attending grade 1 received both tests within a single measurement occasion.

The test administered to children of the independent link sample included 32 items of the test form administered in wave 1 and 44 items of the test form administered in wave 3. Similar to the test setting in waves 1 and 3 of Starting Cohort 2, the test situation for the link sample used an individual setting for the test form of wave 1 and a group setting for the test form of wave 3.

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal subsample from the starting cohort. The differences in item difficulties between the link sample and Starting Cohort 2 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 14.

Analyses of differential item functioning between the waves 1 and 3 identified 11 items (wave 1: 7 items, wave 3: 4 items) with absolute differences in item difficulty parameters greater than 0.5 logits. For wave 1, the respective differences in logits fell between 0.52 and 0.91 and for wave 3 they ranged between 0.53 and 0.71. These items are marked with an asterisk in Table 14 and were excluded prior to linking the receptive vocabulary test forms using the “mean/mean” method for the anchor-group design (see Fischer et al., 2016).

The linking correction term was calculated as $c_{1,3} = 1.238$. This correction term was subsequently added to each difficulty parameter estimated in wave 3 (see Table 6) to derive the linked item parameters. The link error reflecting the uncertainty in the linking process was calculated according to equation 4 in Fischer et al. (2016) as 0.057 and has to be included into the *SE* when statistical tests are used to compare groups concerning their mean change of ability between two linked measurements.

Table 14

Differential Item Functioning Analyses between Waves 1 / 3 and the Link Sample.

	Wave 1				Wave 3			
	Item	**Δσ**	$SE_{\Delta\sigma}$	***F***	**Item**	**Δσ**	$SE_{\Delta\sigma}$	***F***
1.	vok10007_c	-0.08	0.27	0.08	vok10067_sc2g1_c	-0.28	0.22	1.72
2.	vok10008_c	0.36	0.25	2.11	vok10043_sc2g1_c	-0.19	0.17	1.35
3.	vok10012_c*	0.59	0.29	4.13	vok10053_sc2g1_c	0.45	0.21	4.69
4.	vok10023_c*	-0.83	0.25	11.29	vok10049_sc2g1_c	0.35	0.24	2.05
5.	vok10026_c	-0.05	0.19	0.06	vok10025_sc2g1_c	-0.08	0.28	0.09
6.	vok10027_c	0.06	0.23	0.08	vok10076_sc2g1_c	-0.19	0.18	1.15
7.	vok10031_c*	-0.84	0.26	10.35	vok10050_sc2g1_c	-0.37	0.22	2.98
8.	vok10033_c	-0.20	0.19	1.09	vok10060_sc2g1_c	-0.07	0.17	0.20
9.	vok10034_c	-0.12	0.22	0.27	vok10066_sc2g1_c	0.10	0.25	0.15
10.	vok10035_c	-0.44	0.19	5.35	vok10063_sc2g1_c	0.00	0.22	0.00
11.	vok10038_c	0.22	0.21	1.18	vok10040_sc2g1_c*	0.71	0.24	8.60
12.	vok10040_c	0.03	0.21	0.03	vok10074_sc2g1_c*	-0.53	0.17	10.28
13.	vok10041_c	0.18	0.21	0.79	vok10033_sc2g1_c	0.00	0.18	0.00
14.	vok10042_c	0.11	0.25	0.21	vog90015_sc2g1_c	0.02	0.13	0.01
15.	vok10043_c	-0.07	0.18	0.14	vok10051_sc2g1_c	-0.23	0.19	1.45
16.	vok10045_c*	-0.92	0.21	18.22	vok10061_sc2g1_c	0.27	0.17	2.40
17.	vok10046_c	-0.35	0.19	3.37	vog90007_sc2g1_c	0.25	0.14	3.03
18.	vok10047_c	0.00	0.19	0.00	vok10057_sc2g1_c	-0.01	0.17	0.01
19.	vok10048_c	0.27	0.19	2.10	vok10072_sc2g1_c	0.07	0.17	0.16
20.	vok10049_c	-0.08	0.25	0.10	vog90016_sc2g1_c	-0.10	0.13	0.59
21.	vok10050_c	0.39	0.23	2.97	vog90032_sc2g1_c	-0.32	0.18	3.35
22.	vok10051_c	-0.27	0.21	1.54	vok10041_sc2g1_c	-0.21	0.21	1.05
23.	vok10052_c	0.25	0.19	1.80	vok10052_sc2g1_c	0.27	0.17	2.40
24.	vok10053_c	-0.46	0.22	4.17	vok10031_sc2g1_c	-0.02	0.24	0.01
25.	vok10056_c	0.05	0.16	0.11	vok10045_sc2g1_c	-0.14	0.21	0.48
26.	vok10057_c*	0.80	0.19	18.13	vok10034_sc2g1_c	-0.21	0.26	0.69
27.	vok10058_c	0.01	0.20	0.00	vok10058_sc2g1_c	-0.01	0.18	0.01
28.	vok10060_c	0.05	0.19	0.08	vog90031_sc2g1_c	-0.27	0.15	3.39
29.	vok10061_c	0.14	0.20	0.53	vok10065_sc2g1_c	-0.08	0.17	0.25

	Wave 1				Wave 3			
	Item	**Δσ**	$SE_{\Delta\sigma}$	F	**Item**	**Δσ**	$SE_{\Delta\sigma}$	F
30.	vok10063_c*	0.51	0.21	5.92	vok10071_sc2g1_c	0.04	0.17	0.05
31.	vok10064_c*	0.81	0.19	18.59	vok10069_sc2g1_c	-0.30	0.17	3.11
32.	vok10065_c	-0.23	0.20	1.28	vok10028_sc2g1_c	0.19	0.26	0.50
33.					vok10022_sc2g1_c*	-0.54	0.21	6.64
34.					vok10038_sc2g1_c	0.31	0.21	2.05
35.					vog90028_sc2g1_c	0.00	0.18	0.00
36.					vok10047_sc2g1_c	0.01	0.17	0.00
37.					vok10046_sc2g1_c	0.22	0.19	1.27
38.					vok10048_sc2g1_c	0.24	0.20	1.40
39.					vog90020_sc2g1_c	-0.21	0.13	2.52
40.					vok10037_sc2g1_c	0.29	0.29	0.97
41.					vok10077_sc2g1_c	0.30	0.17	2.97
42.					vok10042_sc2g1_c*	0.58	0.28	4.37
43.					vok10064_sc2g1_c	0.05	0.20	0.05
44.					vok10026_sc2g1_c	-0.20	0.18	1.28

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in wave 1 or 3 and the link sample (positive values indicate more difficult items in the link sample); $SE_{\Delta\sigma}$ = Pooled standard error; F = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is F_{0154} (1, 964) = 30.72. A non-significant test indicates measurement invariance. Item suffixes refer to Starting Cohort 2.
*item excluded from the linking procedure.

7.2.2 Linking of waves 3 and 5

A subsample of 5,281 children (51% girls) participated at both measurement occasions, in wave 3 (i.e., grade 1) and also in wave 5 (i.e., grade 3). Consequently, these children were used to link the two test forms across both waves (see Fischer et al., 2016). As the test situation was a group setting in both waves, the linking was based on 11 common items among waves 3 and 5 applying an anchor-items design (Fischer et al., 2016).

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in waves 3 and 5 showed a non-negligible shift in item difficulties. The differences in item difficulties between waves 3 and 5 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 15.

Table 15

Differential Item Functioning Analyses between Wave 3 and Wave 5.

	Item	Δσ	$SE_{\Delta\sigma}$	*F*
1.	vok10043_c	0.27	0.05	29.55
2.	vog10009_c*	0.49	0.04	134.38
3.	vok10033_c*	0.80	0.06	189.09
4.	vog90015_c*	-0.95	0.04	499.47
5.	vok10061_c	-0.19	0.05	14.16
6.	vog60015_c	-0.42	0.04	97.15
7.	vok10072_c	0.16	0.05	10.47
8.	vok10058_c	-0.17	0.05	11.31
9.	vog10040_c	-0.30	0.04	49.40
10.	vok10071_c	0.15	0.05	9.22
11.	vog10044_c	0.14	0.05	8.04

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in waves 3 and 5 (positive values indicate more difficult items in wave 3); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is F_{0154} (1, 5,279) = 115.37. A non-significant test indicates measurement invariance. Item suffixes (i.e. to make the present item application identifiable) were not reported in the table.
*item excluded from the linking procedure.

Analyses of differential item functioning between the waves 3 and 5 identified 3 items with significant (α = .05) DIF. The relevant items are marked with an asterisk in Table 15 and were excluded prior to linking the receptive vocabulary competence test forms of waves 3 and 5 using the “mean/mean” method for the anchor-items design (see Fischer et al., 2016).

The linking correction term was calculated as $c_{3,5} = 1.045$. This correction term as well as the correction term resulting from the linking of waves 1 and 3 (*c* = 1.238) were subsequently added to each difficulty parameter estimated in wave 5 (see Table 7) to derive the linked item parameters. The link error reflecting the uncertainty in the linking process was calculated according to equation 2 in Fischer et al. (2016) as 0.090 and has to be included into the *SE* when statistical tests are used to compare groups concerning their mean change of ability between two linked measurements.

7.2.3 Correcting for sample drop out

Changes in the sample (e.g., due to drop out or sample refreshment) between waves may be an issue in longitudinal measurement as several subsamples may differ in their mean ability.

7.2.3.1 Drop out between waves 1 and 3

In wave 1, children of the longitudinal subsample that participated at both measurement occasions had a mean ability of 0.34 logits above the mean ability of the overall sample. As such, the longitudinal subsample was more able than the overall sample. Similar, in wave 3 the longitudinal subsample had a mean ability of 0.12 logits above the mean ability of the overall sample. Consequently, a drop out correction term of $d_{1,3} = 0.34 - 0.12 = 0.22$, that controls for the nonrandom dropout, needs to be added to each difficulty parameter estimated in wave 3 (see Table 6).

7.2.3.2 Drop out between waves 3 and 5

Children taking part in waves 3 and 5 (i.e., longitudinal subsample) had mean abilities of 0.02 logits and 0.01 logits, respectively, above the mean abilities of the overall samples. Consequently, drop out correction terms of $d_{1,3} = 0.22$ (see above) and $d_{3,5} = 0.01$, needed to be added to each difficulty parameter estimated in wave 5 (see Table 7).

7.3 Receptive vocabulary scores

In the SUF manifest receptive vocabulary competence scores are provided in the form of WLEs. In the following there will be differentiated between cross-sectional WLEs and linked WLEs (marked with a “u” at the end of the WLE variable.

The R code for estimating the WLEs is provided in the Appendix B. For persons who did not give enough valid responses (wave 1: $N = 27$, wave 3: $N = 16$, wave 5: $N = 8$) or for which the test position was unknown (wave 3: $N = 11$), no WLE was estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values.

7.3.1 Cross-sectional WLE

In wave 1, the respective variable is called “vok1_sc1”, including its respective standard error, “vok1_sc2”. As there was no difference in test position (all children received the test on second position), no correction for position effects was necessary.

In wave 3, the respective variable is called “vog1_sc1”, including its respective standard error, “vog1_sc2”. The estimated WLE scores were corrected for differences in the test position as the receptive vocabulary test was either presented as the first or the second test within the test battery (see page 7). To correct for differences in the test position, we added the main effect related to the test position (see Table 10) to the WLE scores of children that received the receptive vocabulary test after working on another test.

In wave 5, the respective variable is called “vog3_sc1”, including its respective standard error, “vog3_sc2”. As there was no difference in test position (all children received the test on first position), no correction for position effects was necessary.

7.3.2 Correcting for test position

In order to link the waves 1, 3, and 5, it was necessary to correct for differences in the test positions between all measurement points. Therefore, the first test position served as the baseline.

In wave 1, the respective variable is called “vok1_sc1u”, including its respective standard error, “vok1_sc2u”. As all children received the test on second position, we added the main effect related to the test position in wave 3 (see Table 10) to the WLE scores of “vok1_sc1”.

In wave 3, the respective variable is called “vog1_sc1u”, including its respective standard error, “vog1_sc2u”. The estimated WLE scores “vog1_sc1” of children that received the test on second position were corrected for the main effect related to the test position (see Table 10). In wave 5, the respective variable is called “vog3_sc1u”, including its respective standard error, “vog3_sc2u”. As all children received the test on first position, no correction for position effects was necessary.

Of course, correcting the WLEs for test position had no effect on their *SE* (i.e., “vok1_sc2u” and “vog1_sc2u”).

References

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

Berendes, K., Weinert, S., Zimmermann, S. & Artelt, C. (2013). Assessing language indicators across the lifespan within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online, 5*, 15-49.

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). Reading, MA: Addison-Wesley.

Carstensen, C. H. (2013). Linking PISA competencies over three cycles – Results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research Outcomes of the PISA Research Conference 2009* (pp. 199-214). New York, NY: Springer.

Dunn, L. M. & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised (PPVT-R)*. Circle Pines, MN: American Guidance Service.

Dunn, L.M. & Dunn, L.M. (1997). *Peabody Picture Vocabulary Test, Third Edition (PPVT-III)*. Circle Pines, MN: American Guidance Service.

Dunn, L. M. & Dunn, L. M. (2004). *Peabody Picture Vocabulary Test (PPVT) (German version)*. Göttingen: Hogrefe.

Dunn, L. M. & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4).* Upper Saddle River, NJ: Pearson.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. (2016). *Linking the Data of the Competence Tests* (NEPS Survey Paper No. 1). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2019). *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.

Lenhard, A., Lenhard, W., Segerer, R., & Suggate, S. (2015). *Peabody Picture Vocabulary Test - 4. Ausgabe.* Pearson.

Mudiappa, M., & Artelt, C. (2014). *BiKS – Ergebnisse aus den Längsschnittstudien: Praxisrelevante Beispiele aus dem Primar- und Sekundarschulbereich* [BiKS - Results of the longistudinal studies]. Bamberg, Germany: University of Bamberg Press.

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

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

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules. R package version 2.7-56*. Retrieved from https://CRAN.R-project.org/package=TAM

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

von Maurice, J., Artelt, C., Blossfeld, H.-P., Faust, G., Rossbach, H.-G., & Weinert, S. (2007). *Bildungsprozesse, Kompetenzentwicklung und Formation von Selektionsentscheidungen im Vor- und Grundschulalter. Überblick über die Erhebung in den Längsschnitten BiKS-3-8 und BiKS-8-12 in den ersten beiden Projektjahren*. Unpublished manuscript, Bamber, Germany: Otto-Friedrich University Bamberg. Retrieved from https://hdl.handle.net/20.500.11780/440

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. von Maurice & (Eds.), *Education as a lifelong process: The German National Education Panel Study (NEPS)* (pp. 67-86). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

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*(2), 125-145. doi:10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

Appendix A

Table 2

Percentage of Missing Values by Item for Wave 1

Item	ID	Position	*N*	Not reached	Not valid
vok10002_c	1	2	2,859	0.00	0.00
vok10007_c	2	7	2,859	0.00	0.00
vok10008_c	3	8	2,857	0.00	0.07
vok10009_c	4	9	2,857	0.00	0.07
vok10010_c	5	10	2,859	0.00	0.00
vok10011_c	6	11	2,858	0.00	0.03
vok10012_c	7	12	2,859	0.00	0.00
vok10013_c	8	13	2,859	0.00	0.00
vok10014_c	9	14	2,859	0.00	0.00
vok10015_c	10	15	2,857	0.07	0.00
vok10016_c	11	16	2,851	0.28	0.00
vok10017_c	12	17	2,847	0.42	0.00
vok10018_c	13	18	2,842	0.56	0.03
vok10019_c	14	19	2,839	0.70	0.00
vok10020_c	15	20	2,839	0.70	0.00
vok10021_c	16	21	2,832	0.87	0.07
vok10022_c	17	22	2,825	1.19	0.00
vok10023_c	18	23	2,815	1.54	0.00
vok10024_c	19	24	2,806	1.85	0.00
vok10025_c	20	25	2,797	2.17	0.00
vok10026_c	21	26	2,785	2.55	0.03
vok10027_c	22	27	2,769	3.08	0.07
vok10028_c	23	28	2,760	3.46	0.00
vok10031_c	24	31	2,732	4.44	0.00
vok10032_c	25	32	2,726	4.58	0.07
vok10033_c	26	33	2,725	4.65	0.03
vok10034_c	27	34	2,723	4.69	0.07
vok10035_c	28	35	2,723	4.72	0.03
vok10036_c	29	36	2,714	5.07	0.00
vok10037_c	30	37	2,684	6.12	0.00

Item	ID	Position	*N*	Not reached	Not valid
vok10038_c	31	38	2,676	6.37	0.03
vok10039_c	32	39	2,659	7.00	0.00
vok10040_c	33	40	2,653	7.21	0.00
vok10041_c	34	41	2,642	7.56	0.03
vok10042_c	35	42	2,624	8.22	0.00
vok10043_c	36	43	2,591	9.30	0.07
vok10045_c	37	45	2,561	10.39	0.03
vok10046_c	38	46	2,539	11.19	0.00
vok10047_c	39	47	2,534	11.37	0.00
vok10048_c	40	48	2,530	11.47	0.03
vok10049_c	41	49	2,516	12.00	0.00
vok10050_c	42	50	2,513	12.10	0.00
vok10051_c	43	51	2,430	14.97	0.03
vok10052_c	44	52	2,409	15.70	0.03
vok10053_c	45	53	2,387	16.47	0.03
vok10054_c	46	54	2,374	16.96	0.00
vok10055_c	47	55	2,366	17.24	0.00
vok10056_c	48	56	2,360	17.45	0.00
vok10057_c	49	57	2,350	17.80	0.00
vok10058_c	50	58	2,341	18.12	0.00
vok10060_c	51	60	2,333	18.40	0.00
vok10061_c	52	61	2,325	18.68	0.00
vok10062_c	53	62	2,321	18.82	0.00
vok10063_c	54	63	2,318	18.89	0.03
vok10064_c	55	64	2,318	18.92	0.00
vok10065_c	56	65	2,317	18.96	0.00
vok10066_c	57	66	2,289	19.94	0.00

Note. ID = item identification number, Position = Item position within test, *N* = Number of valid responses, Not reached = Percentage of children that did not reach the item, Not valid = Percentage of children who gave an invalid answer.

Table 3

Percentage of Missing Values by Item for Wave 3

Item	ID	Position	*N*	Not reached	Not valid
vok10067_sc2g1_c	1	1	6,421	0.00	0.77
vok10043_sc2g1_c	2	2	6,435	0.00	0.56
vok10053_sc2g1_c	3	3	6,433	0.00	0.59
vok10049_sc2g1_c	4	4	6,395	0.00	1.17
vog60001_sc2g1_c	5	5	6,444	0.00	0.42
vok10025_sc2g1_c	6	6	6,448	0.00	0.36
vok10076_sc2g1_c	7	7	6,432	0.00	0.60
vok10050_sc2g1_c	8	8	6,446	0.00	0.39
vog10009_c	9	9	6,424	0.00	0.73
vog60009_sc2g1_c	10	10	6,448	0.00	0.36
vok10060_sc2g1_c	11	11	6,433	0.00	0.59
vok10066_sc2g1_c	12	12	6,434	0.00	0.57
vok10063_sc2g1_c	13	13	6,452	0.03	0.26
vok10040_sc2g1_c	14	14	6,453	0.03	0.25
vok10074_sc2g1_c	15	15	6,429	0.06	0.59
vok10033_sc2g1_c	16	16	6,441	0.14	0.32
vog90015_sc2g1_c	17	17	6,442	0.14	0.31
vok10051_sc2g1_c	18	18	6,449	0.14	0.20
vok10061_sc2g1_c	19	19	6,444	0.15	0.26
vog60051_sc2g1_c	20	20	6,437	0.15	0.37
vog90007_sc2g1_c	21	21	6,440	0.15	0.32
vog60015_sc2g1_c	22	22	6,436	0.17	0.37
vok10057_sc2g1_c	23	23	6,437	0.17	0.36
vok10072_sc2g1_c	24	24	6,448	0.17	0.19
vog90016_sc2g1_c	25	25	6,438	0.17	0.34
vog90032_sc2g1_c	26	26	6,442	0.17	0.28
vog60010_sc2g1_c	27	27	6,451	0.19	0.12
vok10041_sc2g1_c	28	28	6,441	0.19	0.28
vok10052_sc2g1_c	29	29	6,436	0.22	0.32
vog60032_sc2g1_c	30	30	6,429	0.25	0.40
vok10031_sc2g1_c	31	31	6,420	0.37	0.42

Item	ID	Position	*N*	Not reached	Not valid
vok10045_sc2g1_c	32	32	6,435	0.37	0.19
vok10039_sc2g1_c	33	33	6,416	0.37	0.48
vog10034_c	34	34	6,423	0.37	0.37
vok10034_sc2g1_c	35	35	6,426	0.37	0.32
vok10058_sc2g1_c	36	36	6,436	0.37	0.17
vog90031_sc2g1_c	37	37	6,433	0.37	0.22
vog60049_sc2g1_c	38	38	6,427	0.37	0.31
vok10065_sc2g1_c	39	39	6,424	0.37	0.36
vog10040_c	40	40	6,431	0.37	0.25
vok10071_sc2g1_c	41	41	6,413	0.37	0.53
vok10069_sc2g1_c	42	42	6,423	0.40	0.34
vog60025_sc2g1_c	43	43	6,397	0.82	0.32
vog10044_c	44	44	6,408	0.82	0.15
vok10028_sc2g1_c	45	45	6,408	0.83	0.14
vog10046_c	46	46	6,398	0.83	0.29
vog60027_sc2g1_c	47	47	6,398	0.87	0.26
vog60047_sc2g1_c	48	48	6,392	0.90	0.32
vok10022_sc2g1_c	49	49	6,308	2.32	0.20
vok10038_sc2g1_c	50	50	6,310	2.35	0.14
vog90028_sc2g1_c	51	51	6,300	2.38	0.26
vok10047_sc2g1_c	52	52	6,303	2.41	0.19
vok10046_sc2g1_c	53	53	6,306	2.43	0.12
vog60019_sc2g1_c	54	54	6,295	2.53	0.19
vok10048_sc2g1_c	55	55	6,039	6.26	0.42
vog10056_c	56	56	6,042	6.31	0.32
vog90020_sc2g1_c	57	57	6,045	6.32	0.26
vok10037_sc2g1_c	58	58	6,037	6.43	0.28
vog60030_sc2g1_c	59	59	6,029	6.48	0.36
vog10060_c	60	60	6,028	6.54	0.31
vok10077_sc2g1_c	61	61	5,990	7.15	0.28
vog10062_c	62	62	5,991	7.19	0.23
vog10063_c	63	63	5,983	7.19	0.36
vok10042_sc2g1_c	64	64	5,991	7.29	0.12

Item	ID	Position	*N*	Not reached	Not valid
vok10064_sc2g1_c	65	65	5,952	7.83	0.19
vok10026_sc2g1_c	66	66	5,790	10.42	0.11

Note. ID = item identification number, Position = Item position within test, *N* = Number of valid responses, Not reached = Percentage of children that did not reach the item, Not valid = Percentage of children who gave an invalid answer.

Table 4

Percentage of Missing Values by Item for Wave 5

Item	ID	Position	*N*	Not reached	Not valid
vog10034_sc2g3_c	1	1	5,601	0.00	0.02
vok10043_sc2g3_c	2	4	5,594	0.00	0.14
vog90031_sc2g3_c	3	5	5,599	0.00	0.05
vog10060_sc2g3_c	4	8	5,600	0.00	0.04
vog10009_sc2g3_c	5	9	5,593	0.00	0.16
vog60041_sc2g3_c	6	10	5,594	0.00	0.14
vog60025_sc2g3_c	7	13	5,598	0.00	0.07
vok10075_sc2g3_c	8	17	5,592	0.00	0.18
vok10033_sc2g3_c	9	18	5,598	0.00	0.07
vog90015_sc2g3_c	10	19	5,597	0.00	0.09
vok10061_sc2g3_c	11	20	5,596	0.00	0.11
vok10065_sc2g3_c	12	21	5,594	0.00	0.14
vog60015_sc2g3_c	13	24	5,596	0.00	0.11
vok10072_sc2g3_c	14	25	5,600	0.00	0.04
vog60030_sc2g3_c	15	26	5,595	0.00	0.12
vog60029_sc2g3_c	16	27	5,596	0.00	0.11
vog90003_sc2g3_c	17	28	5,595	0.00	0.12
vog10062_sc2g3_c	18	29	5,595	0.00	0.12
vok10026_sc2g3_c	19	32	5,602	0.00	0.00
vog60037_sc2g3_c	20	33	5,593	0.00	0.16
vok10058_sc2g3_c	21	34	5,596	0.00	0.11
vog60049_sc2g3_c	22	35	5,594	0.00	0.14
vok10076_sc2g3_c	23	36	5,594	0.00	0.14
vok10040_sc2g3_c	24	37	5,600	0.00	0.04
vog10040_sc2g3_c	25	38	5,595	0.00	0.12
vok10071_sc2g3_c	26	39	5,598	0.00	0.07
vok10060_sc2g3_c	27	40	5,591	0.00	0.20
vog10044_sc2g3_c	28	42	5,599	0.00	0.05
vog60045_sc2g3_c	29	43	5,594	0.00	0.14
vog90035_sc2g3_c	30	45	5,596	0.00	0.11
vok10074_sc2g3_c	31	47	5,572	0.16	0.37

Item	ID	Position	*N*	Not reached	Not valid
vog60027_sc2g3_c	32	50	5,568	0.43	0.18
vok10051_sc2g3_c	33	51	5,576	0.43	0.04
vog60047_sc2g3_c	34	53	5,551	0.75	0.16
vok10073_sc2g3_c	35	55	5,524	1.23	0.16
vog90037_sc2g3_c	36	56	5,520	1.25	0.21
vok10038_sc2g3_c	37	57	5,530	1.27	0.02
vok10047_sc2g3_c	38	58	5,528	1.27	0.05
vok10057_sc2g3_c	39	60	5,510	1.55	0.09
vok10046_sc2g3_c	40	61	5,409	3.37	0.07
vog60019_sc2g3_c	41	62	5,407	3.43	0.05
vok10048_sc2g3_c	42	63	5,399	3.43	0.20
vog90016_sc2g3_c	43	64	5,382	3.78	0.14
vog90032_sc2g3_c	44	66	5,370	4.02	0.12
vog60010_sc2g3_c	45	67	5,217	6.84	0.04
vog60032_sc2g3_c	46	69	5,209	6.91	0.11
vog60054_sc2g3_c	47	70	5,195	7.18	0.09
vok10064_sc2g3_c	48	71	5,185	7.35	0.09
vog90028_sc2g3_c	49	72	5,169	7.66	0.07

Note. ID = item identification number, Position = Item position within test, *N* = Number of valid responses, Not reached = Percentage of children that did not reach the item, Not valid = Percentage of children who gave an invalid answer.

Appendix B

R-Syntax for estimating WLEs

```
# load packages
library(haven) # to import SPSS files
library(TAM)   # for IRT analyses

# load competence data
dat <- read_sav("SUF for competencies.sav")

# items of the receptive vocabulary test
items <- c("vok10002_c", "vok10007_c ",
           "vok10008_c", "vok10009_c ",
           ...)

# estimate Rasch model
mod <- tam.mml(resp = dat[, items], irtmodel = "1PL",
               pid = dat$ID_t)
summary(mod)

# item fit
tam.fit(mod)

# WLE
tam.wle(mod)
```