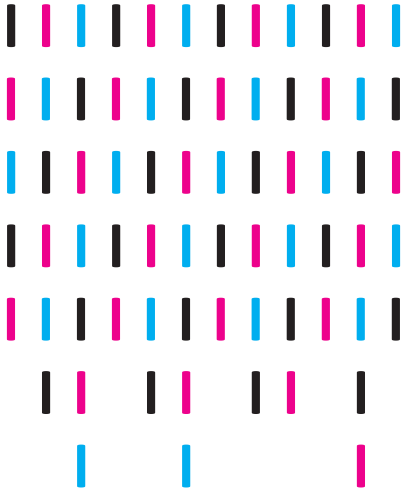

NEPS SURVEY PAPERS

Theresa Rohm, Anna Scharl, Jennifer Ettner, and Karin Gehrer

NEPS TECHNICAL REPORT FOR READING: SCALING RESULTS OF STARTING COHORTS 4 (WAVE 10), 5 (WAVE 12), AND 6 (WAVE 9)

NEPS Survey Paper No. 62
Bamberg, November 2019

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 Reading:

Scaling Results of Starting Cohorts

4 (Wave 10), 5 (Wave 12), and 6 (Wave 9)

Theresa Rohm, Anna Scharl, Jennifer Ettner, and Karin Gehrer

Leibniz Institute for Educational Trajectories, Bamberg, Germany

E-mail address of lead author:

theresa.rohm@lifbi.de

Bibliographic data:

Rohm, T., Scharl, A., Ettner, J., & Gehrer, K. (2019). *NEPS Technical Report for Reading: Scaling Results of Starting Cohorts 4 (Wave 10), 5 (Wave 12), and 6 (Wave 9)* (NEPS Survey Paper No. 62). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educa-tional Panel Study. doi:10.5157/NEPS:SP62:1.0

Acknowledgements:

Various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (e.g., Gnambs, Fischer & Rohm, 2017; Koller, Haberkorn & Rohm, 2014; Pohl, Haberkorn, & Hardt, 2014) to facilitate the understanding of the presented results.

We thank Katharina Krohmer for her assistance in scaling the data.

NEPS Technical Report for Reading: Scaling Results of Starting Cohorts 4 (Wave 10), 5 (Wave 12), and 6 (Wave 9)

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, a range of analyses based on item response theory (IRT) were performed. This paper describes the data and scaling procedures for the reading competence test that was administered in wave 10 of Starting Cohort 4 (ninth grade), wave 12 of Starting Cohort 5 (students), and wave 9 of Starting Cohort 6 (adults). The reading competence test contained 38 items, distributed among an easy booklet (27 items) and a difficult booklet (23 items in Starting Cohorts 4 and 6; 21 items in Starting Cohort 5) with different response formats representing different cognitive requirements and text functions. The test was finished by 17,972 individuals (53% women) from Starting Cohort 4 (*N* = 6,871), Starting Cohort 5 (*N* = 4,816), and Starting Cohort 6 (*N* = 6,441). In Starting Cohort 5 about half of the respondents received the test in a proctored setting at their private homes (*N* = 2,766), whereas the remaining participants (*N* = 2,050) worked on unproctored, web-based tests. Starting Cohorts 4 and 6 were limited to proctored computerized testing. The participants' responses were scaled using the partial credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, the test's dimensionality, and local item independence were evaluated to ensure the quality of the test. These analyses showed that the test exhibited an acceptable reliability and that the items fitted the model in a satisfactory way. The items covered a wide range of the ability distribution of the participants and the variance implies good differentiation between respondents. Furthermore, test fairness could be confirmed for different subgroups. A limitation of the test is the large percentage of items at the end of the test that were not reached due to time limits. Further challenges related to the dimensionality analyses based on both text functions and cognitive requirements. Overall, the reading test had satisfactory psychometric properties that allowed for an estimation of reliable reading competence scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the *R* syntax for scaling the data.

Keywords

item response theory, scaling, reading competence, scientific use file

Content

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 that are 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 a reading competence test that was administered in wave 10 of Starting Cohort 4 (ninth grade), wave 12 of Starting Cohort 5 (students), and wave 9 of Starting Cohort 6 (adults). First, the main concepts of the reading test and the test design are introduced. Then, the competence data of the three starting cohorts (SC) and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the scientific use file is presented.

Please note that the analyses in this report are based on the data available at some time before public data release. Due to ongoing data protection and data cleansing issues, the data in the scientific use file (SUF) may differ slightly from the data used for the analyses in this paper. However, we do not expect fundamental changes in the presented results.

2. Testing Reading Competence

2.1 Conceptual Framework

The framework and test development for the reading competence test are described by Weinert and colleagues (2011) and Gehrer, Zimmermann, Artelt, and Weinert (2012). In the following, there will be a brief description of specific aspects of the reading competence test that are necessary for understanding the scaling results presented in this paper.

Each test included five texts with respective item sets. Each of these texts represented one of the following text types or text functions: a) information, b) commenting or argumenting, c) literary, d) instruction, and e) advertising (see Gehrer & Artelt, 2013, and Gehrer, Zimmerman, Artelt, & Weinert, 2013, for a description of the framework). Furthermore, the test assessed three cognitive requirements. These are a) finding information in the text, b) drawing text-related conclusions, and c) reflecting and assessing. The cognitive requirements do not depend on the text type, but each cognitive requirement is usually assessed within each text type (see Gehrer et al., 2013; Gehrer & Artelt, 2013).

Four types of response formats were included in the reading competence test: simple multiple choice (MC) items, complex multiple choice (CMC) items, matching (MA) items, and text-enrichment-task (TET) items. Examples of the first three response formats are given in Pohl

and Carstensen (2012) and Gehrer et al. (2012). As a new format, the present study introduces, for the first time in the NEPS, text-enrichment-task (TET) items (see section 2.2.). MC items had four response options. One response option represented a correct solution, whereas the other three were distractors (i.e., they were incorrect). CMC items consisted of a number of subtasks with one correct answer out of two response options. MA items required the test taker to match a number of responses to a given set of statements. The number of potential responses always exceeded the amount of statement sets by two distractor items (i.e., they were incorrect). TET items required the test taker to locate the correct position of responses (i.e., additional sentences) in gaps within the text (see below for further details).

The reading competence test that was administered in the present study included 67 items. Extensive preliminary analyses were conducted to evaluate the quality of these items. These preliminary analyses identified a satisfactory fit for all items. Therefore, all 67 items were included in the final scaling procedure.

2.2 New Computer-based Item Type: Text-Enrichment-Task

The present test, for the first time in the NEPS, administered a new item format that made use of the possibilities of computerized assessments. The aim of this task is to enrich a text meaningfully with three to four additional sentences. The task requires the test taker to identify the correct gap between two sentences in a text to determine the correct position for each supplementary sentence within the text in order to meaningfully expand the text (see Figure 1). The new item format is called text-enrichment-task (TET).

To successfully solve TET items, test takers have to understand the story, line of arguments, or action described in the text before coming to a decision about the correct position of the additional argument (or theme or example or part of the story line). This new format is implemented using the drag and drop technique. For this purpose, each additional sentence is accompanied by a symbol. Within the text, each gap after the end of a sentence is marked with a circle. For a test taker to respond, he or she must drag the symbol to the correct gap within the text. A response can be changed or corrected (e.g., by inserting the sentence into another gap and, thus, deleting the initial response). If in a text that is presented on two separate pages one of the additional phrases has already been inserted on the previous page, but fits better on the second page, it can still be used here and will be automatically deleted on the previous page. The test taker is notified of this instance by means of a dialog box and confirms the changed response. Since this enrichment of the text with several sentences creates a slightly new text or a modified story, this item is always presented as the last item of the text. These text-enrichment-tasks are scored as a partial credit item within the scaling procedure. The three or four additional sentences are understood as subtasks that are aggregated for IRT scaling (see 4.2 Scaling Model).

For the development of this new task format, a development study (N = 937) was conducted that pretested the new presentation format in a sample of college students (n = 443) and adults (n = 458) of the same age and education groups as the samples in Starting Cohorts 4, 5, and 6. A pool with 16 TET items was developed that implemented the new format for each of the five text types. In a video instruction, both the task-to-task navigation and the drag-and-drop mouse technique were introduced. For computer novices a longer version of the test instructions was recorded, in which the required technique was described in more detail. A repetition of incomprehensible instruction parts was possible. In addition to the

item content and layout, the understanding of the instruction for the new format were checked. Feedback from the interviewers revealed that ambiguities about the deletion or insertion of the symbols were particularly evident in the case of two-page texts and that more technical assistance had to be provided. Therefore, the instruction for this new format was optimized and a dialog box was implemented for inserting the answers on the second page of the text, making the handling clearer and requiring less technical assistance. In addition, the item presentation was optimized in such a way that the associated sentence content also remains displayed until the symbol is inserted into one of the gaps.

The test quality parameters of all text-enrichment-tasks (TET) in the development study were satisfactory to very good. The items discriminated well between good and poor readers (r_{it} = .32 to .66). For the test administered in Starting Cohorts 4, 5, and 6, those text-enrichment-tasks were selected that best suited the five text types according to the framework of the NEPS reading tests. By means of log data analyses, the times required for the processing of the respective TET were calculated and used as a further selection criterion for the accuracy of fit for the test instrument. To achieve a total testing time of 28 minutes, TET items were not implemented for each of the five texts.

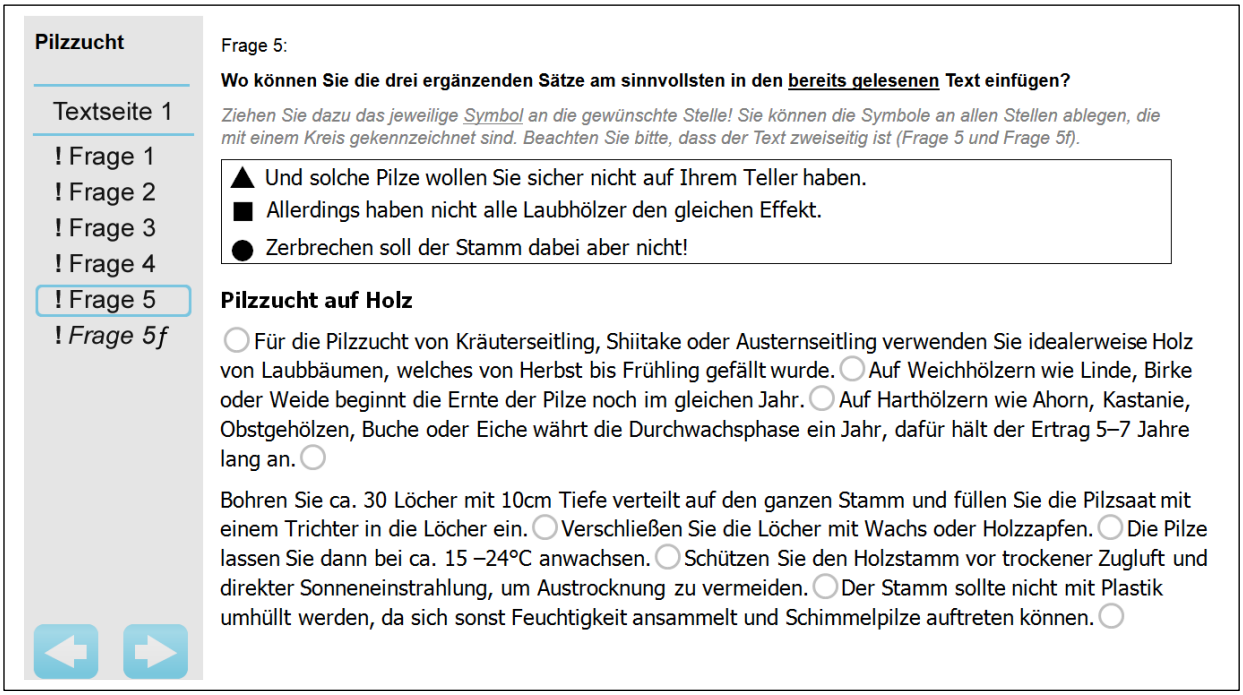

Figure 1. Example of the new item format text-enrichment-task[1].

1 This TET-example was part of a pilot study, but was not used in the main study because it proved to be too easy despite good fit statistics. Please note that the instruction (light gray) has been optimized for the main studies to the wording: “Drag the respective symbol to the desired position within a circle! The symbol will be deleted on the first page if you insert it again on the second page”.

3. Data

3.1 Design of the Study

The study followed a three-factorial (quasi-)experimental design. These factors referred to (a) the position of the reading test within the test battery, (b) the difficulty of the administered test, and (c) the assessment setting (i.e., the context of test administration).

Table 1

Number of Items for the Different Text Types by Difficulty of the Test

Text types	**Starting Cohorts 4 & 6**		**Starting Cohort 5**
	Easy test	**Difficult test**	**Difficult test**
Information text	7	6	6
Instruction text	5	4	4
Advertising text	4	4	3
Commenting text	5	5	5
Literary text	6	4	3
Total number of items	27	23	21

The study assessed different competence domains including reading competence, mathematical competence and English as a foreign language. The latter was only administered in Starting Cohort 5. In order to control for test position effects, the tests were administered to participants in different sequence. Each participant received the test in the same position as in the previous wave to allow for longitudinal comparsions of reading competences. For each participant the reading test was either administered as the first or the second test (i.e., after the mathematics test). There was no multi-matrix design regarding the order of the items *within* a specific test. All students received the test items in the same order. A detailed description of the study design is available on the NEPS website (http://www.neps-data.de).

In order to measure participants' reading competence with greater accuracy, the difficulty of the administered items should adequately match the participants' abilities. Therefore, the study adopted the principles of longitudinal multistage testing (Pohl, 2013). Based on preliminary studies three different versions of the reading competence test were developed that differed in their average difficulty (i.e., an easy and two difficult tests). All three tests included five texts representing all five text functions. Referring to the texts, the easy test contained 27 items, the difficult test for the adults and younger adults (Starting Cohorts 4 and 6) comprised of 23 items, and the difficult test for the students (Starting Cohort 5) had 21 items (see Table 1). Since difficult items and difficult texts often need more time to edit, in the difficult versions of the test fewer items could be administered. Moreover, the three cognitive requirements (see Table 2) were assessed as described above. The assignment of the items

to the different text types and cognitive requirements can be found in Appendix A. Three texts with 10 items were identical in all three test versions and acted as an anchor for linking the different test versions; 12 items (easy test), 1 item (difficult test in Starting Cohorts 4 and 6), and 2 items (difficult test in Starting Cohort 5) were unique to each test.

Table 2

Number of Items by Cognitive Requirements and Difficulty of the Test

Cognitive requirements	**Cohorts 4 & 6 Easy test**	**Cohorts 4 & 6 Difficult test**	**Cohort 5 Difficult test**
Finding information	5	2	2
Drawing text-related conclusions	11	8	9
Reflecting and assessing	11	13	10
Total number of items	27	23	21

Table 3

Number of Items by Different Response Formats and Difficulty of the Test

Response format	**Cohorts 4 & 6 Easy test**	**Cohorts 4 & 6 Difficult test**	**Cohort 5 Difficult test**
Simple multiple choice items	17	16	14
Complex multiple choice items	7	2	2
Matching items	3	3	3
Text-enrichment-tasks		2	2
Total number of items	27	23	21

The different response formats of the items are summarized in Table 3. Only the difficult tests contained the new TET format. In the easy test the new format was not administered. A development study has shown, that the text-enrichment-tasks require more time than simple multiple-choice items. Therefore, and due to time constraints, fewer items could be accommodated in the two difficult tests than in the easy test. The number of subtasks within CMC items varied between 3 and 5. Participants from Starting Cohorts 4 and 6 were assigned either to the easy or the difficult test version, based on their estimated reading competence in the previous assessment (Haberkorn, Pohl, Hardt & Wiegand, 2012). Participants with an ability estimate below the sample's mean ability received the easy test, whereas participants with a reading competence above the sample's mean received the difficult test.

All test takers from Starting Cohort 5 received the student version of the difficult test. About half of the respondents (*N* = 2,763) received the test in a proctored setting at their private homes, within a computer-based test mode. The remaining participants of Starting Cohort 5 (*N* = 1,930) took the test in an unproctored setting, working on web-based tests (see Table 4). Please note that by this design the assessment setting (proctored vs. unproctored) and the test administration mode (computer-based test vs. web-based test) are completely confounded. Further information on setting and mode effects is presented by Kröhne, Gnambs, and Goldhammer (2019).

Sample

A total of 17,972[2] participants (53% women) had at least three valid responses on the reading competence test and, thus, were used for the psychometric analyses (cf. Pohl & Carstensen, 2012). Of these, *N* = 6,866 (50% women) were from Starting Cohort 4, *N* = 4,693 (60% women) were from Starting Cohort 5, and *N* = 6,413 (51% women) were from Starting Cohort 6. The age of the respondents ranged from 18 to 73 years. The number of participants within each (quasi-)experimental condition is given in Table 4 and basic sociodemographic information of the different samples are summarized.

Table 4

Number of Participants by the (Quasi-)Experimental Conditions

	Starting Cohort 4		**Starting Cohort 5**		**Starting Cohort 6**	
	Easy test	**Difficult test**	**CBT**	**WBT**	**Easy test**	**Difficult test**
Sample size	3,003	3,863	2,763	1,930	3,295	3,118
Women	46%	54%	61%	59%	51%	50%
Migration	19%	8%	7%	9%	10%	7%
Mean age (*SD*) in years	21.21 (0.63)	20.99 (0.53)	27.91 (3.35)	27.95 (3.46)	54.97 (10.68)	51.12 (10.22)
First position	54%	49%	50%	52%	68%	64%

Note. CBT = Computer-based test, WBT = Web-based test.

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

4. Analyses

4.1 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) omitted items, b) items that test takers did not reach, c) test abortion, d) items that have not been administered, e) response resets, and, finally, f) multiple kinds of missing responses within CMC items that are not determined.

Omitted items occurred when test takers skipped some items. 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. Furthermore, test takers had the opportunity to end the test before the time limit was reached, leading to missing responses due to test abortion. Because of the shared anchor items across different booklets, 10 of 38 items were administered to all participants. Items unique to the other booklets were missing by design. Hence, for respondents receiving the easy test, 11 difficult items were missing by design, whereas 15 items were missing by design for respondents of the difficult test in Starting Cohorts 4 and 6, and 17 items were not administered to Starting Cohort 5 (see Table 1). Response resets occurred when participants deleted a given answer after submission. As CMC items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. When one subtask contained a missing response, the CMC item was coded as missing. If just one kind of missing response occurred, the item was coded according to the corresponding missing response. If the subtasks contained different kinds of missing responses, the item was labeled as a not-determinable missing response.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions, handling of different response formats). 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 partial credit model (PCM; Masters, 1982). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

CMC and TET items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC and TET item, indicating the number of correctly responded subtasks within that item. If at least one of the subtasks contained a missing response, the CMC or TET item was scored as missing. Categories of polytomous variables with less than $N = 200$ responses were collapsed in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items; in these cases, the lower categories were collapsed into one category. For 3 of the 7 CMC items and 1 of the 2 TET items categories were collapsed (see Appendix B).

To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items were scored dichotomously as 0 for

an incorrect and 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats).

Reading competences 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 7.3.

4.3 Checking the Quality of the Test

The reading competence test was specifically constructed to be implemented in the NEPS. In order to ensure appropriate psychometric properties, the quality of the test was examined in several analyses. These quality steps were also carried out in the process of developing the new items and taken into account in the selection from the development study (see Gehrer et al., 2013, for details of the test construction).

Before aggregating the subtasks of CMC, MA, and TET items to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the MC items in a Rasch model (Rasch, 1960). The fit of the subtasks was evaluated based on the weighted mean square (WMNSQ), the respective *t*-value, point-biserial correlations of the correct responses with the total correct score, and the item characteristic curves. Only if the subtasks exhibited a satisfactory item fit, they were used to construct polytomous CMC, MA, and TET variables that were included in the final scaling model.

The MC items consisted of one correct response option and one or more distractors (i.e., incorrect response options). The quality of the distractors within MC items 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).

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC and polytomous CMC, MA, and TET items to the partial credit model (Masters, 1982) 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 reading competence test should measure the same construct for all respondents. If some items favored certain subgroups (e.g., they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between 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, gender, migration background, the number of books at home (as a proxy for socioeconomic status), school degree (only Starting Cohort 6), and age (only Starting Cohort 6; see Pohl & Carstensen, 2012, for a description of these variables). Moreover, in light of the quasi-experimental design measurement invariance analyses were also conducted for the test difficulty and admin-

istration setting (e.g., CBT vs. WBT in Starting Cohort 5). Furthermore, measurement invariance was analysed between the starting cohorts. 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.

The reading competence test was scaled using the PCM (Masters, 1982), which assumes Rasch-homogeneity. The PCM was chosen because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that might not hold for empirical data. To test the assumption of equal item discrimination parameters, a generalized partial credit model (GPCM; Muraki, 1992) was also fitted to the data and compared to the PCM.

The dimensionality of the test was evaluated by two different multidimensional analyses. The different subdimensions of the multidimensional models were specified based on different construction criteria. First, a model with three different subdimensions representing the three cognitive requirements, and, second, a model with five different subdimensions based on the five text functions were fitted to the data. The correlations among the dimensions as well as differences in model fit between the unidimensional model and the respective multidimensional models were used to evaluate the unidimensionality of the test. 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.

Since the reading competence test consisted of item sets that referred to one of seven texts, the assumption of local item dependence (LID) may not necessarily hold. However, the seven texts were perfectly confounded with the five text functions. Please note, that each participant received only one text per type of text function. Thus, multidimensionality and local item dependence cannot be evaluated separately with these data.

4.4 Software

The item response models were estimated with the *TAM* package version 2.12-18 (Robitzsch, Kiefer, & Wu, 2018) in *R* version 3.5.0 (R Core Team, 2017).

5. Results

5.1 Missing Responses

5.1.1 Missing responses per person

There was not a single invalid response in this study, and less than one percent of items were missing due to response resets in all three cohorts. Furthermore, less than three per-

cent of responses were missing due to test abortion in all samples. Missing responses may also occur when respondents omit items. As illustrated in Figure 2 most respondents in the easy test, 70% to 84%, did not skip any item and less than five percent omitted more than one item in Starting Cohort 4, while less than five percent omitted more than four items in Starting Cohort 6. There is a difference in the amount of omitted items between the two Starting Cohorts for the easy test.

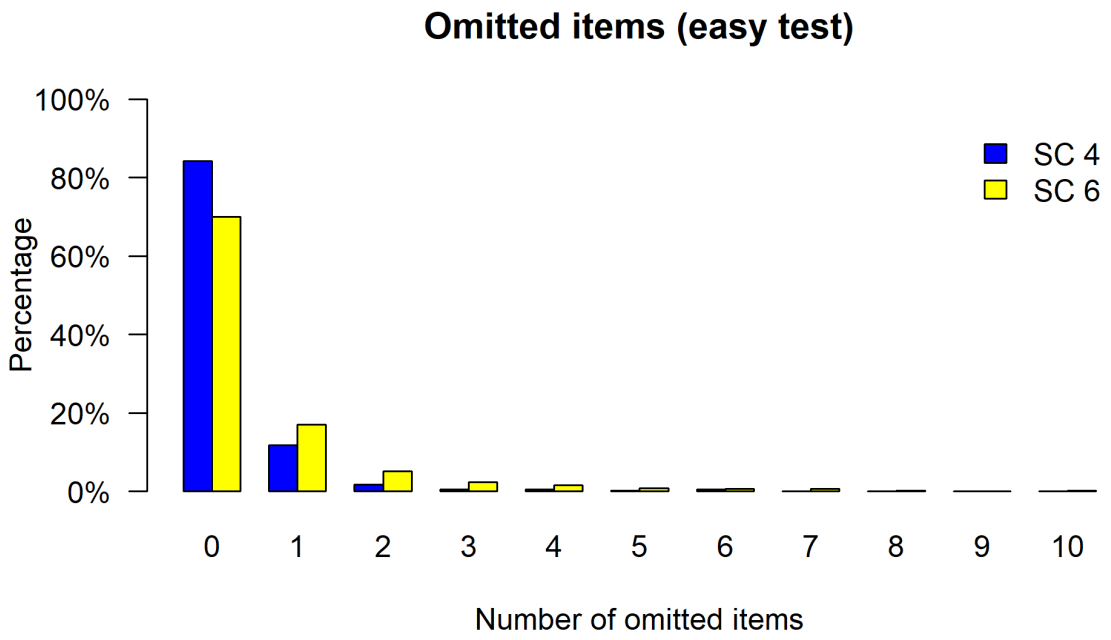

Figure 2. Number of mitted items by starting cohort in easy test.

As Figure 3 shows, between 71% and 86% of the respondents in the difficult test did not skip any item and less than five percent omitted more than one item in Starting Cohorts 4 and 5 (CBT and WBT), while less than five percent omitted more than two items in Starting Cohort 6.

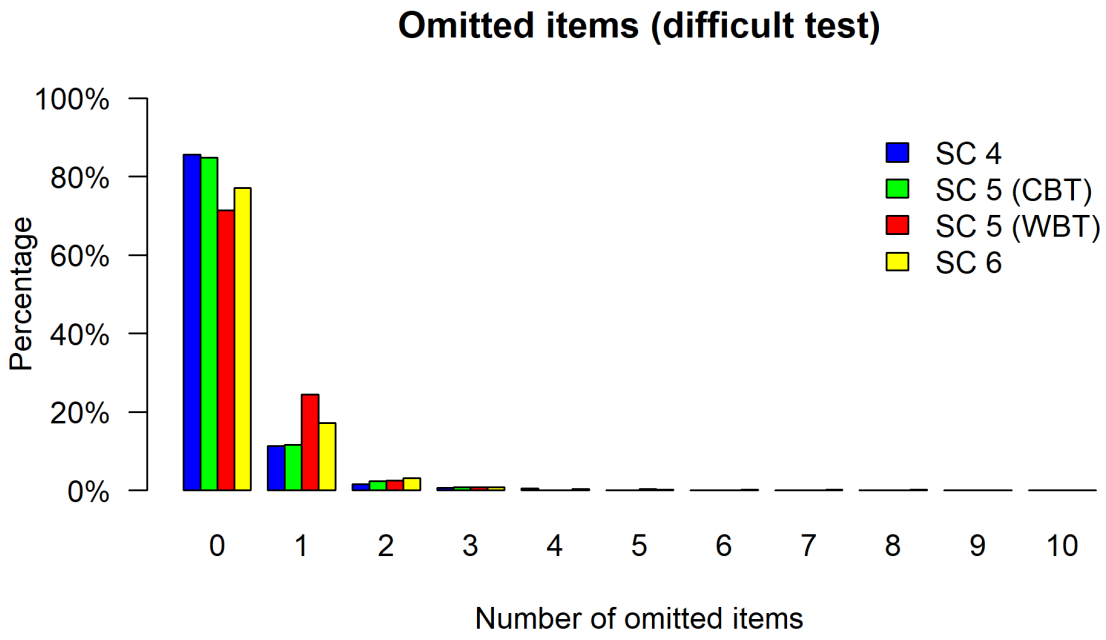

Figure 3. Number of omitted items by starting cohort in difficult test.

Another source of missing responses are items that were not reached by the respondents because they ran out of time; these are all missing responses after the last valid response. In the easy test (see Figure 4) between 56% and 65% of the respondents finished the entire test. The items of the last text were not reached by about 21% in Starting Cohort 4 and 27% in Starting Cohort 6.

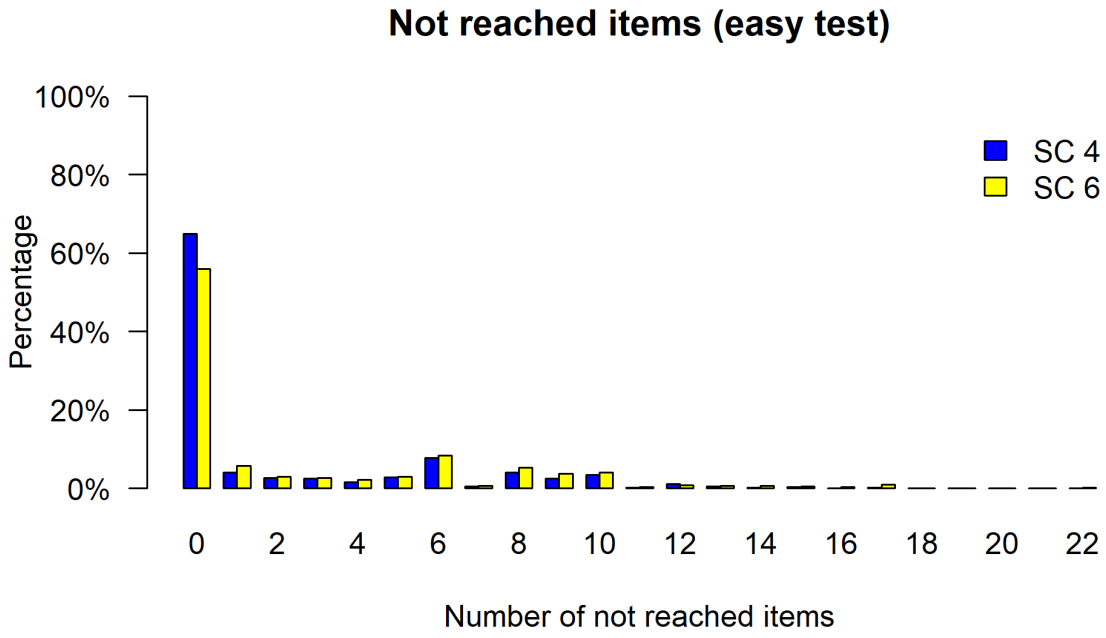

Figure 4. Number of not-reached items by starting cohort in easy test.

In the difficult test, between 36% and 77% of the respondents finished the entire test. The last text was not reached by about 24% of the test takers in SC 4, 34% in the proctored setting of SC 5, 19% in the web based setting of SC 5, and 38% in SC 6.

The aggregated polytomous variables were coded as not-determinable missing response when the subtasks of CMC, MA, or TET items contained different kinds of missing responses. Less than 0.3 percent of responses were not determinable in all starting cohorts and booklets.

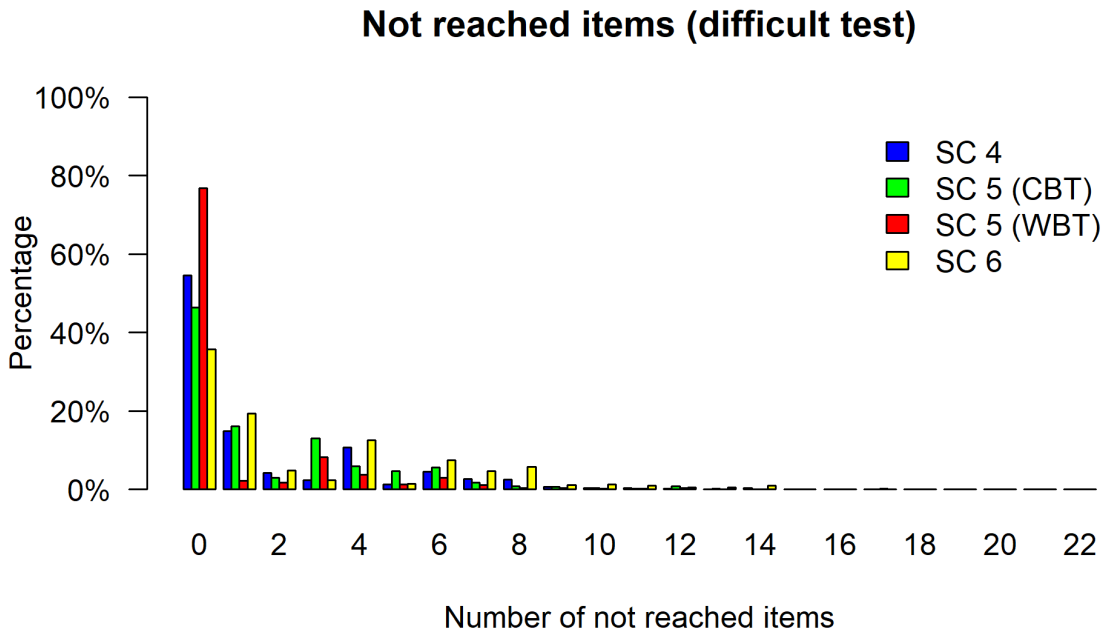

Figure 5. Number of not-reached items by starting cohort in difficult test.

With an item's progressing position in the test, the amount of persons that did not reach the item rose in all samples (see Figures 6, 7, and 8). However, in the difficult tests administered to Starting Cohorts 5 (CBT) and 6, about half of the respondents did reach the last item.

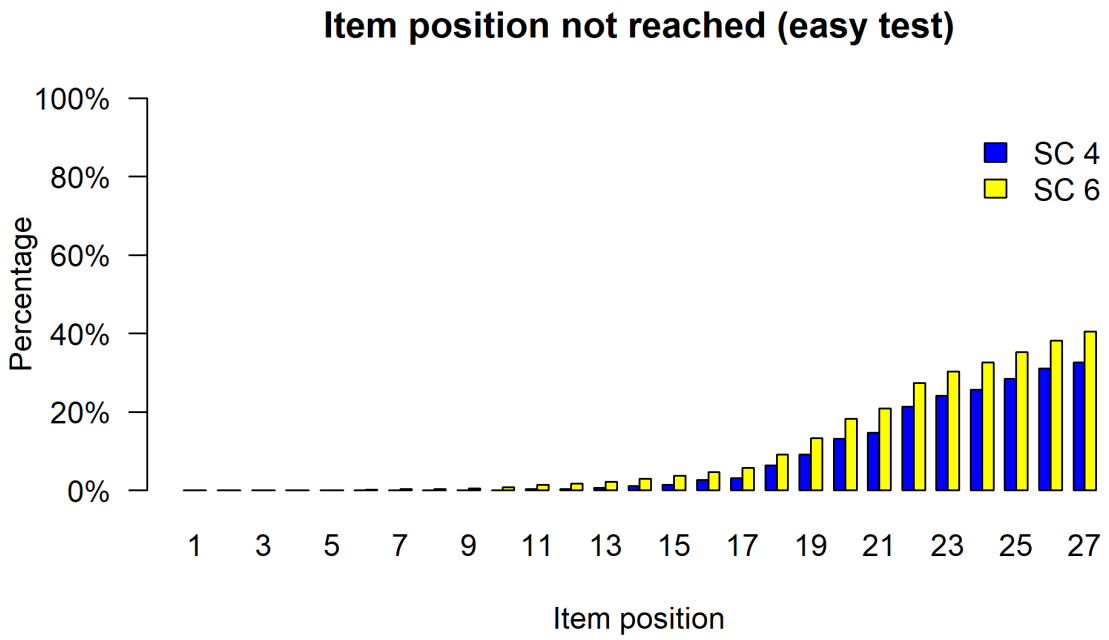

Figure 6. Item position not reached by sample in the easy test.

As illustrated in Figure 8, in the proctored CBT condition of Starting Cohort 5 substantially more persons did not reach the last item of the test (about 53%) as compared to the unproctored WBT condition (about 23%). Thus, it seems that some respondents were unable to finish the test within the allocated time span. This indicates that the testing time might have been too short, especially for test-takers of the CBT-setting.

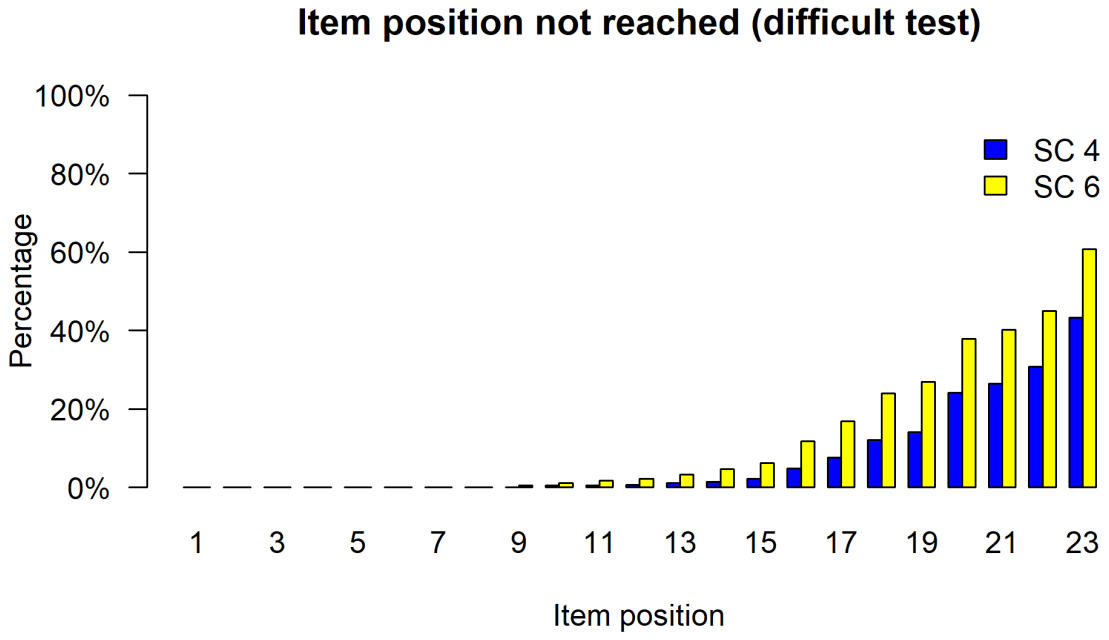

Figure 7. Item position not reached in Starting Cohorts 4 and 6 for the difficult test.

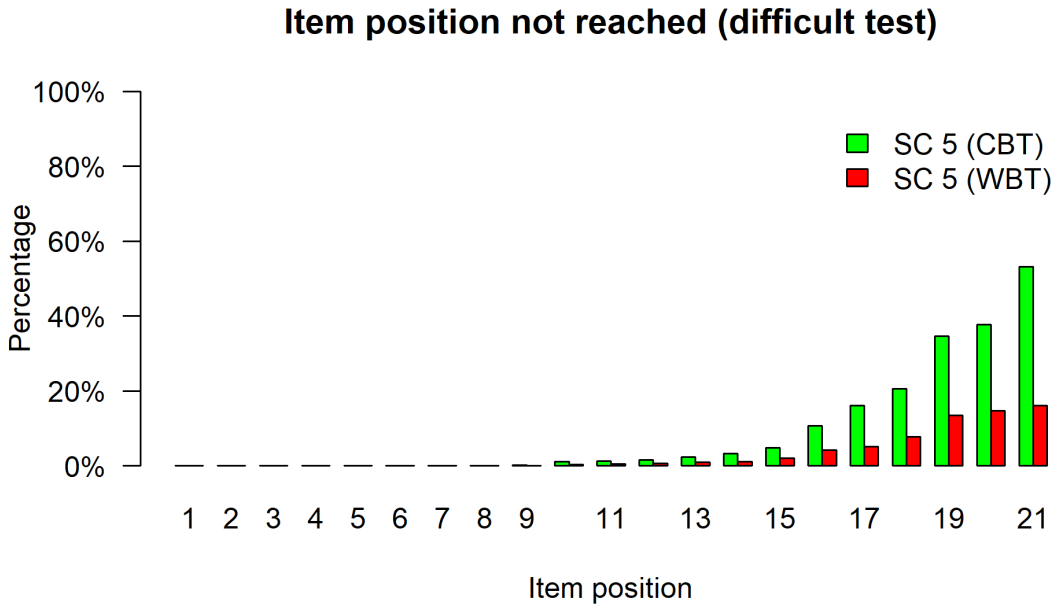

Figure 8. Item position not reached in Starting Cohort 5.

The total number of missing responses, aggregated over omitted, not-reached, response reset, and not-determinable missing responses per person, is illustrated in Figure 9 for the easy test and in Figure 10 for the difficult test. In the easy test, respondents of the Starting Cohort 6 had *M* = 4.54 (*SD* = 12.37) and respondents of the Starting Cohort 4 *M* = 4.00 (*SD* = 8.24) missing responses. About 41% (Starging Cohort 6) and 59% (Starging Cohort 4) of the test takers had no missing response at all and only about 24% (Starting Cohort 4) to 32% (Starting Cohort 6) of the participants had more than five missing responses.

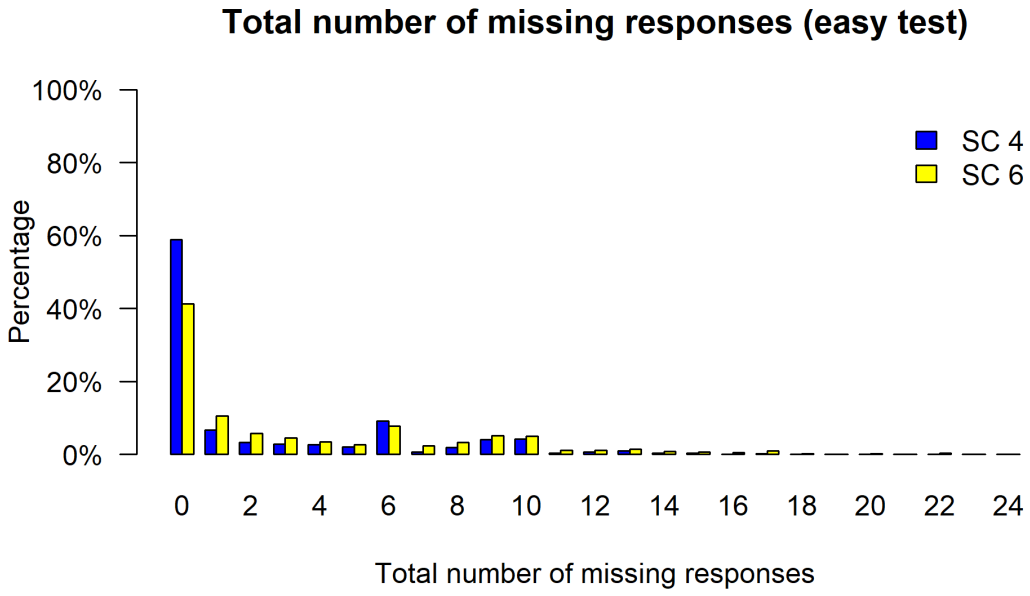

Figure 9. Total number of missing responses by starting cohort in the easy test.

In the difficult test, respondents of the Starting Cohort 4 showed *M* = 6.25 (*SD* = 12.34) missing responses while it were *M* = 5.26 (*SD* = 7.65) for the respondents of Starting Cohort 6. Respondents of the proctored setting in Starting Cohort 5 had *M* = 6.67 (*SD* = 11.02) missing responses compared to *M* = 5.88 (*SD* = 14.85) in the web-based setting. Not a single missing response was found for 28% of respondents from Starting Cohort 6, 42% of respondents in Starting Cohort 5 in the proctored setting, 49% of respondents of Starting Cohort 4, and 62% of respondents of Starting Cohort 5 in the web-based setting. Five or more missing responses were counted for 14% in Starting Cohort 4, 20% in the proctored setting of Starting Cohort 5, 13% in the web-based setting of Starting Cohort 5, and 28% in Starting Cohort 6. Particularly respondents of Starting Cohort 6 showed more missing responses because they did not reach the last of the five texts.

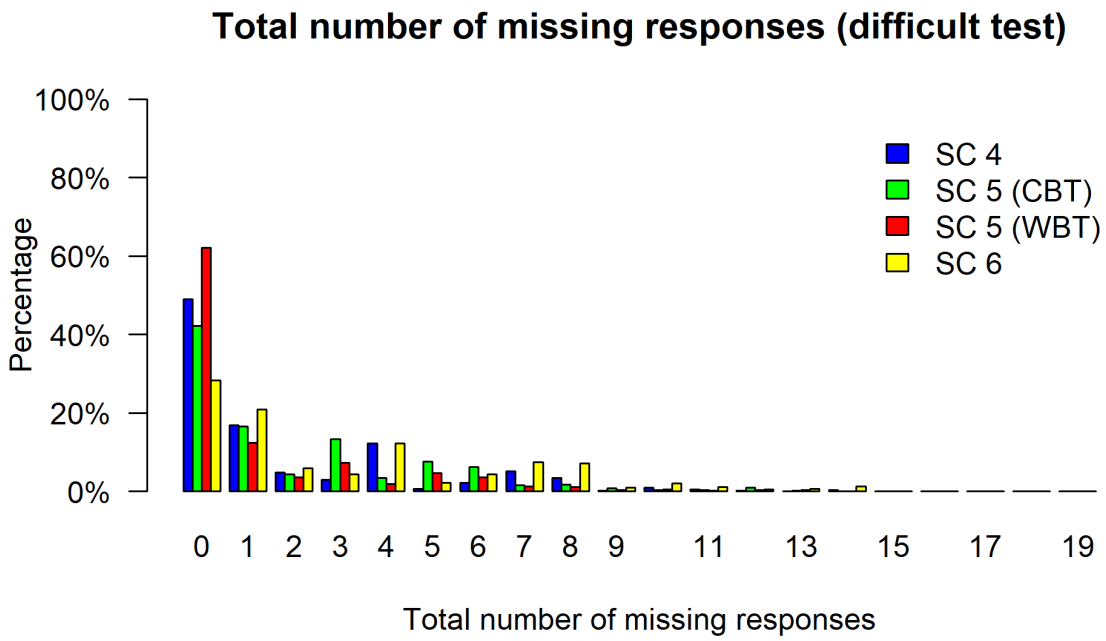

Figure 10. Total number of missing responses by Starting Cohort in the difficult test.

In sum, the amount of missing responses was rather large because some respondents did not reach the end of the test. On average, respondents from Starting Cohort 4 or the web-based condition of Starting Cohort 5 exhibited more valid responses than respondents in Starging Cohort 6 or in the proctored condition of Starting Cohort 5.

5.1.2 Missing responses per item

Tables 5, 6, and 7 provide information on the occurrence of different kinds of missing responses per item for the easy and difficult test versions, as well as the proctored and unproctored test setting of Starting Cohort 5. The number of omitted responses varied across items between 0.00% and 12.12% (*Mdn* = 1.16%) and were, thus, negligible. In contrast, there were substantially more missing responses because participants did not reach the item. On average, the items had *Mdn* = 6.76% missing values of this type in Starting Cohort 4, *Mdn* = 9.28% in Starting Cohort 6, *Mdn* = 12.04% in Starting Cohort 5 (CBT), and *Mdn* = 6.00% in Starting Cohort 5 (WBT). Particularly, some items of the last text were frequently

not reached. It is noticeable that the last items for students in the online mode (SC 5, WBT) have fewer missings.

Table 5

Percentage of Missing Values by Item: Easy Test by Starting Cohort

		SC 4				*SC 6*			
Item	**Position**	***N***	**NR**	**OM**	**TA**	***N***	**NR**	**OM**	**TA**
rea90101s_c	1	2991	0.00	0.40	0.00	3163	0.00	4.01	0.00
rea90102s_c	2	2996	0.00	0.23	0.00	3214	0.00	2.46	0.00
rea901030_c	3	2999	0.00	0.13	0.00	3265	0.00	0.91	0.00
rea90104s_c	4	2968	0.00	1.13	0.03	2986	0.00	9.38	0.00
rea90105s_c	5	2994	0.00	0.17	0.03	3234	0.00	1.79	0.00
rea90201s_c	6	2990	0.00	0.40	0.03	3187	0.24	3.00	0.03
rea902020_c	7	2996	0.00	0.20	0.03	3249	0.33	1.03	0.03
rea902030_c	8	2995	0.03	0.20	0.03	3247	0.39	1.03	0.03
rea902040_c	9	2995	0.07	0.17	0.03	3241	0.49	1.12	0.03
rea90205s_c	10	2959	0.13	1.30	0.03	3019	0.82	7.53	0.03
rea903010_c	11	2977	0.33	0.33	0.03	3177	1.49	1.61	0.06
rea903020_c	12	2976	0.37	0.33	0.03	3166	1.85	1.58	0.06
rea903030_c	13	2959	0.73	0.53	0.03	3146	2.31	1.73	0.06
rea903040_c	14	2948	1.10	0.53	0.03	3125	3.07	1.61	0.06
rea90305s_c	15	2913	1.53	1.23	0.03	3058	3.85	2.79	0.06
rea903060_c	16	2897	2.66	0.67	0.03	3061	4.73	1.88	0.06
rea90307s_c	17	2820	3.16	2.73	0.03	2767	5.77	9.77	0.06
rea904010_c	18	2724	6.43	1.50	0.13	2884	9.26	1.43	0.06
rea90402s_c	19	2579	9.26	3.33	0.13	2625	13.29	4.98	0.06
rea90403s_c	20	2517	13.15	1.67	0.13	2454	18.36	5.37	0.06
rea904040_c	21	2485	14.69	1.20	0.13	2494	20.85	1.67	0.06
rea905010_c	22	2182	21.35	1.93	0.13	2169	27.34	1.67	0.06
rea905020_c	23	2125	24.08	1.10	0.13	2091	30.35	1.03	0.06
rea905030_c	24	2059	25.74	1.63	0.13	2003	32.56	1.49	0.06
rea905040_c	25	1979	28.40	1.63	0.13	1899	35.27	1.94	0.06

		SC 4				*SC 6*			
Item	**Position**	***N***	**NR**	**OM**	**TA**	***N***	**NR**	**OM**	**TA**
rea905050_c	26	1910	31.07	1.27	0.13	1826	38.27	1.15	0.06
rea905060_c	27	1854	32.60	1.60	0.13	1751	40.52	1.18	0.06

Note. Position = Item position within test, N = Number of valid responses, NR = Percentage of respondents that did not reach the item, OM = Percentage of respondents that omitted the item, TA = Percentage of respondents who aborted the test. Item names refer to Starting Cohort 6; the corresponding variable names for Starting Cohorts 4 and 5 are given in Appendix C.

Table 6

Percentage of Missing Values by Item: Difficult Test by Starting Cohort

		SC 4				*SC 6*			
Item	**Position**	***N***	**NR**	**OM**	**TA**	***N***	**NR**	**OM**	**TA**
rea906010_c	1	3862	0.00	0.03	0.00	3115	0.00	0.10	0.00
rea906020_c	2	3862	0.00	0.03	0.00	3117	0.00	0.03	0.00
rea906030_c	3	3863	0.00	0.00	0.00	3113	0.00	0.16	0.00
rea90604s_c	4	3860	0.00	0.05	0.00	3110	0.00	0.26	0.00
rea90201s_c	5	3817	0.03	0.47	0.00	3034	0.06	2.60	0.00
rea902020_c	6	3829	0.03	0.16	0.00	3098	0.13	0.48	0.00
rea902030_c	7	3831	0.03	0.10	0.00	3105	0.10	0.29	0.00
rea902040_c	8	3832	0.03	0.08	0.00	3103	0.13	0.32	0.00
rea90206s_c	9	3760	0.10	2.33	0.00	2910	0.61	5.77	0.00
rea903010_c	10	3804	0.47	0.26	0.00	3028	1.15	0.90	0.00
rea903020_c	11	3804	0.49	0.23	0.00	3012	1.76	0.80	0.00
rea903030_c	12	3793	0.67	0.34	0.00	2987	2.21	1.15	0.00
rea903040_c	13	3772	1.09	0.47	0.00	2956	3.30	1.06	0.00
rea90305s_c	14	3735	1.40	1.09	0.00	2873	4.68	2.21	0.00
rea903060_c	15	3726	2.25	0.49	0.00	2847	6.32	1.54	0.00
rea904010_c	16	3601	4.84	1.16	0.00	2610	11.74	2.05	0.03
rea90402s_c	17	3375	7.64	3.93	0.00	2346	16.93	4.81	0.03
rea90403s_c	18	3281	12.04	2.25	0.00	2176	23.93	3.78	0.03
rea904040_c	19	3230	14.16	1.45	0.00	2147	26.97	1.67	0.03
rea907010_c	20	2722	24.13	1.55	0.00	1689	37.81	1.41	0.06

rea907020_c	21	2623	26.46	1.79	0.00	1579	40.22	2.53	0.06
rea907030_c	22	2479	30.81	1.16	0.00	1480	45.00	0.93	0.06
rea90704s_c	23	1914	43.36	3.00	0.03	904	60.71	3.53	0.06

Note. Position = Item position within test, *N* = Number of valid responses, NR = Percentage of respondents that did not reach the item, OM = Percentage of respondents that omitted the item, TA = Percentage of respondents who aborted the test. Item names refer to Starting Cohort 6; the corresponding variable names for Starting Cohorts 4 and 5 are given in Appendix C.

Table 7

Percentage of Missing Values by Item for Starting Cohort 5

		CBT				*WBT*			
Item	**Position**	***N***	**NR**	**OM**	**TA**	***N***	**NR**	**OM**	**TA**
rea906010_c	1	2762	0.00	0.04	0.00	1929	0.00	0.00	0.05
rea906020_c	2	2762	0.00	0.04	0.00	1928	0.00	0.05	0.05
rea906030_c	3	2761	0.00	0.07	0.00	1927	0.00	0.10	0.05
rea90604s_c	4	2761	0.00	0.07	0.00	1921	0.00	0.31	0.05
rea90201s_c	5	2750	0.00	0.47	0.00	1890	0.16	0.62	1.30
rea902020_c	6	2763	0.00	0.00	0.00	1898	0.16	0.21	1.30
rea902030_c	7	2763	0.00	0.00	0.00	1900	0.16	0.10	1.30
rea90205s_c	8	2745	0.00	0.65	0.00	1880	0.16	1.14	1.30
rea90206s_c	9	2712	0.22	1.48	0.00	1843	0.21	2.69	1.30
rea903010_c	10	2723	1.09	0.36	0.00	1872	0.57	0.36	2.07
rea903020_c	11	2718	1.30	0.33	0.00	1867	0.83	0.36	2.07
rea903030_c	12	2706	1.59	0.47	0.00	1856	1.04	0.73	2.07
rea903040_c	13	2685	2.39	0.43	0.00	1849	1.45	0.67	2.07
rea90305s_c	14	2637	3.33	1.19	0.00	1824	1.76	1.66	2.07
res1203080_c	15	2598	4.81	1.16	0.00	1808	3.01	1.24	2.07
rea904010_c	16	2424	10.71	1.56	0.00	1733	6.06	1.81	2.33
rea90402s_c	17	2198	16.11	4.05	0.00	1639	7.36	5.34	2.33
res1204050_c	18	2144	20.63	1.77	0.00	1617	11.19	2.69	2.33
rea907010_c	19	1752	34.64	1.95	0.00	1463	19.27	2.18	2.75
rea907020_c	20	1653	37.71	2.46	0.00	1412	21.04	3.06	2.75
rea90704s_c	21	1204	53.17	3.11	0.00	1194	23.11	12.12	2.75

Note. CBT = Computer-based test, WBT = Web-based test. Position = Item position within test, *N* = Number of valid responses, NR = Percentage of respondents that did not reach the item, OM = Percentage of respondents that omitted the item, TA = Percentage of respondents who aborted the test. Item names refer to Starting Cohort 6; the corresponding variable names for Starting Cohort 5 are given in Appendix C.

5.2 Parameter Estimates

To avoid potentially biased parameter estimates resulting from mode effects (unproctored versus proctored settings), the following analyses are limited to the proctored CBT samples. Thus, the unproctored WBT sample from Starting Cohort 5 was excluded from the scaling procedure. Information on the measurement invariance across assessment modes is given in section 5.3.4. Moreover, preliminary analyses identified rather high WMNSQ value for the dichotomous item rea901030_c. Therefore, we used a 0.5 point scoring for this item in our scaling procedure (similar to the scoring rule for PCM items, see section 4.2). Additionally, three items lacked measurement invariance across the starting cohorts and are therefore treated as unique items for each cohort in our scaling procedure (see items with subscript _SC4 and _SC6 in Tables 9 and 10).

5.2.1 Item parameters

Table 8 displays the position of items in the three different test booklets and the number of test takers included in the analyses. Please note that for Starting Cohort 5 only the results of the proctored setting were used for scaling.

Table 8

Item Position

Item	Position easy (SC 4, SC 6)	Position difficult (SC 4, SC 6)	Position difficult (SC 5 CBT)	*N*
rea90101s_c	1	NA	NA	6154
rea90102s_c	2	NA	NA	6210
rea901030_c	3	NA	NA	6264
rea90104s_c	4	NA	NA	5954
rea90105s_c	5	NA	NA	6228
rea906010_c	NA	1	1	9739
rea906020_c	NA	2	2	9741
rea906030_c	NA	3	3	9737
rea90604s_c	NA	4	4	9731
rea90201s_c	6	5	5	15778
rea902020_c	7	6	6	15935
rea902030_c	8	7	7	15941

Item	Position easy (SC 4, SC 6)	Position difficult (SC 4, SC 6)	Position difficult (SC 5 CBT)	*N*
rea902040_c	9	8	NA	13171
rea90205s_c	10	NA	8	8723
rea90206s_c	NA	9	9	9382
rea903010_c	11	10	10	15709
rea903020_c	12	11	11	15676
rea903030_c	13	12	12	15591
rea903040_c	14	13	13	15486
rea90305s_c	15	14	14	15216
rea903060_c	16	15	NA	12531
rea90307s_c	17	NA	NA	5587
res1203080_c	NA	NA	15	2598
rea904010_c	18	16	16	14243
rea90402s_c	19	17	17	13123
rea90403s_c	20	18	NA	10428
rea904040_c	21	19	NA	10356
res1204050_c	NA	NA	18	2144
rea905010_c	22	NA	NA	4351
rea905020_c	23	NA	NA	4216
rea905030_c	24	NA	NA	4062
rea905040_c	25	NA	NA	3878
rea905050_c	26	NA	NA	3736
rea905060_c	27	NA	NA	3605
rea907010_c	NA	20	19	6163
rea907020_c	NA	21	20	5855
rea907030_c	NA	22	NA	3959
rea90704s_c	NA	23	21	4022

Note. Item names refer to Starting Cohort 6; the corresponding variable names for Starting Cohorts 4 and 5 are given in Appendix C.

The third column in Table 9 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 percentage of correct responses within dichotomous items varied between 20% and 72% with an average of 47% (*SD* = 16%) correct responses.

Table 9

Item Parameters

Seq. Position	Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
1	rea90101s_c	NA	-0.10	0.01	1.01	0.61	0.22	0.44	0.02
2	rea90102s_c	NA	-0.68	0.02	0.96	-2.23	0.28	0.76	0.03
3	rea901030_c	26.82	0.80	0.03	1.02	1.57	0.06	0.23	0.02
4	rea90104s_c	NA	-0.76	0.02	1.00	-0.21	0.18	0.50	0.02
5	rea90105s_c	NA	-0.16	0.01	0.97	-2.20	0.31	0.57	0.03
6	rea906010_c	65.07	-0.42	0.02	1.00	0.35	0.22	0.93	0.02
7	rea906020_c	57.12	-0.03	0.02	1.00	-0.29	0.23	0.96	0.03
8	rea906030_c	25.98	1.51	0.02	1.01	1.05	0.19	0.85	0.02
9	rea90604s_c	NA	-0.62	0.01	0.95	-4.01	0.23	0.69	0.03
10	rea90201s_c	NA	-0.92	0.01	0.93	-5.29	0.33	0.73	0.02
11	rea90201s_c_SC 6	NA	-0.55	0.01	0.92	-4.67	0.35	0.72	0.04
12	rea902020_c	71.80	-1.09	0.02	0.98	-2.53	0.26	1.04	0.02
13	rea902030_c	33.23	0.82	0.02	1.03	4.23	0.24	0.73	0.02
14	rea902040_c	38.30	0.44	0.02	0.99	-1.10	0.21	0.95	0.02
15	rea90205s_c	NA	-0.34	0.01	0.99	-0.78	0.17	0.47	0.02
16	rea90206s_c	NA	0.96	0.01	0.98	-1.61	0.22	0.55	0.02
17	rea903010_c	47.46	0.00	0.03	1.07	8.11	0.20	0.58	0.02
18	rea903010_c_SC 6	30.59	0.86	0.03	1.07	5.11	0.21	0.60	0.03
19	rea903010_c_SC 4	53.62	0.37	0.04	1.05	3.30	0.18	0.65	0.02
20	rea903020_c	37.28	0.68	0.02	1.03	3.04	0.26	0.76	0.02
21	rea903020_c_SC 6	40.89	0.33	0.03	1.06	5.84	0.22	0.63	0.03
22	rea903030_c	25.93	1.23	0.02	1.07	7.17	0.17	0.56	0.02
23	rea903040_c	55.81	-0.27	0.02	0.96	-6.31	0.33	1.10	0.02
24	rea90305s_c	NA	0.32	0.01	0.95	-5.36	0.37	0.58	0.02
25	rea903060_c	53.32	-0.27	0.02	1.06	9.27	0.17	0.63	0.02
26	rea90307s_c	NA	-0.66	0.02	1.02	1.40	0.19	0.40	0.02
27	res1203080_c	40.26	1.01	0.04	1.01	0.72	0.24	0.82	0.02
28	rea904010_c	38.16	0.57	0.02	1.02	2.86	0.26	0.81	0.02
29	rea90402s_c	NA	0.28	0.01	0.95	-4.43	0.41	0.56	0.02
30	rea90403s_c	NA	-0.42	0.01	0.89	-8.87	0.33	0.84	0.03
31	rea904040_c	64.13	-0.81	0.02	1.06	7.42	0.17	0.62	0.01
32	res1204050_c	65.72	-0.20	0.05	1.09	4.28	0.14	0.42	0.03
33	rea905010_c	58.95	-0.92	0.03	1.01	0.91	0.26	0.88	0.02

Seq. Position	Item	Percentage correct	Item difficulty	*SE*	WMNSQ	*t*	r_{it}	Discr.	Q_3
34	rea905020_c	70.33	-1.48	0.04	0.95	-3.53	0.34	1.44	0.03
35	rea905030_c	54.92	-0.73	0.03	1.01	5.64	0.20	0.58	0.02
36	rea905040_c	46.54	-0.36	0.03	0.98	-1.68	0.32	1.08	0.03
37	rea905050_c	64.24	-1.18	0.04	1.01	1.02	0.24	0.90	0.02
38	rea905060_c	19.75	1.05	0.04	1.06	2.16	0.14	0.49	0.03
39	rea907010_c	25.62	1.49	0.03	1.05	3.31	0.18	0.58	0.02
40	rea907020_c	64.27	-0.41	0.03	1.01	1.13	0.26	0.84	0.01
41	rea907030_c	52.03	0.07	0.03	1.03	2.51	0.27	0.77	0.02
42	rea90704s_c	NA	0.55	0.02	0.96	-2.15	0.31	0.61	0.03

Note. Seq. Position = Sequential Position. Difficulty = Item difficulty / location parameter, *SE* = Standard error of item difficulty / location 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). Percent correct scores are not informative for polytomous CMC, TET, and MA item scores. These are denoted by n.a. For the dichotomous items, the item-total correlation corresponds to the point-biserial correlation between the correct response and the total score; for polytomous items it corresponds to the product-moment correlation between the corresponding categories and the total score. Item names refer to Starting Cohort 6; the corresponding variable names for Starting Cohorts 4 and 5 are given in Appendix C. As a consequence of DIF between the three starting cohorts, the items rea9021s_c and rea903020_c were separated and treated as unique items for Starting Cohort 6 (see subscript _SC 6) in the scaling model. The item rea903010_c was treated as unique for all three starting cohorts in the scaling model (see subscript _SC 6 and _SC 4).

The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 9. The step parameters for polytomous variables are depicted in Table 10. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated item difficulties (or location parameters for polytomous variables) ranged from -1.48 (item rea905020_c) to 1.51 (item rea906030_c) with an average difficulty of 0 (*Mdn* = -0.13). Overall, the item difficulties were quite balanced with a tendency of being too easy for the present sample. Due to the large sample size, the standard errors (SE) of the estimated item difficulties (column 4 in Table 9) were rather small (all SEs ≤ 0.06).

Table 10

Step Parameters (with Standard Errors) for Polytomous Items

Item	Step 1	Step 2	Step 3	Step 4
rea90101s_c	-1.07 (0.03)	0.08 (0.03)	1	
rea90102s_c	-1.28 (0.03)	1.28		
rea90104s_c	-0.32 (0.03)	0.32		
rea90105s_c	-0.41 (0.03)	0.19 (0.03)	0.23	
rea90201s_c	-1.31 (0.02)	-0.16 (0.02)	0.17 (0.02)	1.31
rea90201s_c_SC 6	-1.33 (0.03)	-0.62 (0.03)	0.42 (0.03)	1.54
rea90205s_c	-1.29 (0.02)	0.62 (0.02)	0.68	
rea90206s_c	-0.71 (0.02)	0.16 (0.03)	0.55	
rea90305s_c	-0.68 (0.02)	0.05 (0.02)	0.63	
rea90307s_c	1.08 (0.04)	-1.08		
rea90402s_c	-0.83 (0.02)	-0.13 (0.02)	0.11 (0.02)	0.86
rea90403s_c	-0.62 (0.02)	-0.69 (0.02)	0.09 (0.02)	1.23
rea90604s_c	-1.07 (0.02)	-0.15 (0.02)	1.21	
rea90704s_c	-0.91 (0.03)	-0.03 (0.04)	0.94	

Note. The last step parameter for each item is not estimated and has, thus, no standard error because it is a constrained parameter for model identification. Item names refer to Starting Cohort 6; the corresponding variable names for Starting Cohorts 4 and 5 are given in Appendix C.

5.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. Because some items in the reading test were polytomous, we calculated Thurstonian thresholds for each response category (Wu, Adams, Wilson, & Haldane, 2007). These indicate the location at the latent dimension at which the probability of achieving a score above the respective threshold is 50%. Thus, it is similar to the item difficulties of dichotomous items. In Figure 11, the category thresholds of the reading items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of category thresholds. The respective thresholds ranged from -4.99 (rea90201s_c) to 3.87 (item rea90206s_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 0.83, which implies good differentiation between respondents. The reliability of the test (EAP/PV reliability =.77, WLE reliability = .75) was good. The mean of the item distribution was about 0.15 logits below the mean person ability distribution. 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 more precisely, whereas higher ability estimates will have larger standard errors of measurement.

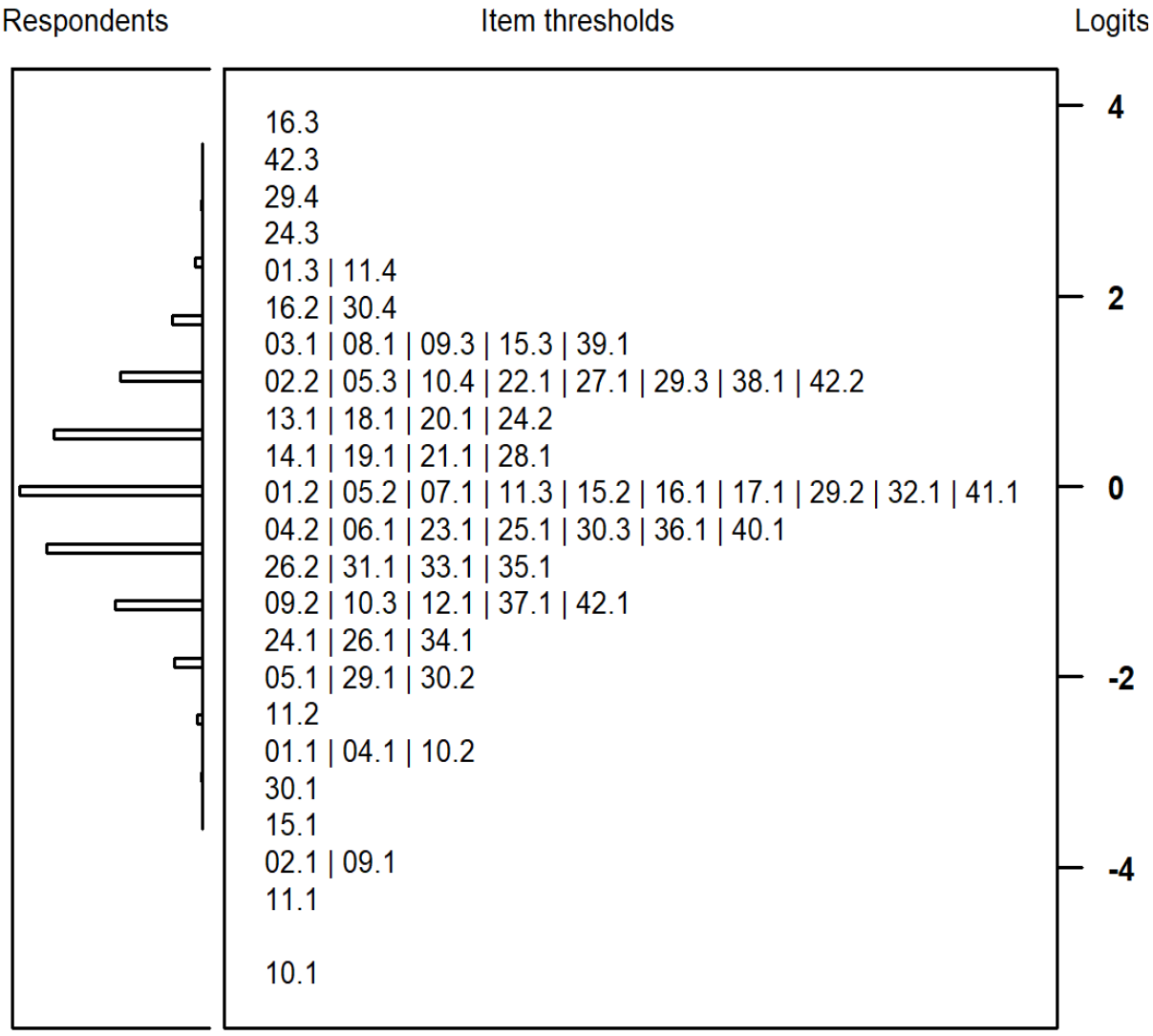

Figure 11. Test targeting. The distribution of person ability in the sample is given on the left-hand side of the graph. The category thresholds of the items are given on the right-hand side of the graph. Each number represents one threshold with the first part (before the dot) corresponding to the sequential position in Table 9 and the second part indicating the threshold.

5.3 Quality of the test

5.3.1 Fit of the subtasks of complex multiple choice items

Before the subtasks of CMC, MA, and TET items were aggregated and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the single subtasks together with the MC items in a Rasch model. Counting the subtasks of CMC, MA, and TET items separately, there were 42 items. The probability of a correct response ranged from 20% to 83% across all items (*Mdn* = 47%). Thus, the number of correct and incorrect responses was reasonably large. All subtasks showed a satisfactory item fit. WMNSQ ranged from 0.92 to 1.10, the respective *t*-value from -15.10 to 10.21, and there were no noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. Due to the good model fit of the subtasks, their aggregation to polytomous variables seemed justified.

5.3.2 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the partial credit model, using the MC and polytomous CMC, MA, and TET items. Altogether, item fit can be considered to be very good (see Table 9). Values of the WMNSQ ranged from 0.89 (item rea90403s_c) to 1.09 (item res1204050_c). Only three items exhibited a *t*-value of the WMNSQ greater than 8 and none exceeded a value of 9.3. Thus, there was no indication of severe item over- or underfit. Point-biserial correlations between the item scores and the total scores ranged from .06 (item rea901030_c) to .41 (item rea90402s_c) and had a mean of .24. All item characteristic curves showed a good fit of the items.

5.3.3 Distractor analyses

In addition to the overall item fit, we specifically investigated how well the distractors performed in the test by evaluating the point-biserial correlation between each incorrect response (distractor) and the students' total correct score. The point-biserial correlations for the distractors ranged from -.41 to .03 with a mean of -.17. These results indicate that the distractors functioned well.

5.3.4 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, school type, age, and test position (see Pohl & Carstensen, 2012, for a description of these variables). In addition, we compared the common items that were administered to all participants for the easy and difficult tests (booklet). The differences between the estimated item difficulties in the various groups are summarized in Table 11. For example, the column "Male vs. female" reports the differences in item difficulties between men and women; a positive value would indicate that the test was more difficult for males, whereas a negative value would highlight a lower difficulty for males as opposed to females. In addition, the effect of the three starting cohorts was also studied (see Table 12). 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 13).

Table 11

Differential Item Functioning

Item	Sex	Books	Migration	School	Age	Position	Booklet
	male vs. female	< 100 vs. ≥ 100	without vs. with	no sec. vs. sec.	< 54 vs. ≥ 54	first vs. second	easy vs. difficult
rea90101s_c	-0.121 (-0.174)	0.024 (0.036)	-0.052 (-0.076)	-0.155 (-0.264)	-0.013 (-0.020)	0.089 (0.129)	
rea90102s_c	-0.268 (-0.385)	0.101 (0.148)	0.001 (0.001)	0.077 (0.131)	0.082 (0.125)	0.095 (0.137)	
rea901030_c	-0.122 (-0.175)	-0.157 (-0.230)	0.297 (0.432)	-0.307 (-0.524)	0.205 (0.312)	0.057 (0.082)	
rea90104s_c	0.084 (0.121)	0.009 (0.013)	-0.005 (-0.007)	-0.191 (-0.326)	0.250 (0.380)	-0.083 (-0.119)	
rea90105s_c	-0.060 (-0.086)	-0.009 (-0.013)	0.118 (0.171)	0.019 (0.032)	-0.124 (-0.188)	0.143 (0.206)	
rea90201s_c	-0.039 (-0.057)	-0.091 (-0.134)	0.182 (0.265)	0.174 (0.296)	-0.160 (-0.242)	0.053 (0.076)	0.044 (0.075)
rea902020_c	0.060 (0.087)	0.286 (0.420)	-0.223 (-0.325)	0.362 (0.617)	-0.034 (-0.052)	0.124 (0.178)	0.410 (0.698)
rea902030_c	-0.051 (-0.073)	-0.184 (-0.269)	0.040 (0.058)	0.217 (0.371)	0.050 (0.076)	0.112 (0.161)	0.170 (0.290)
rea902040_c	-0.090 (-0.129)	0.270 (0.396)	0.089 (0.130)	0.479 (0.817)	-0.209 (-0.318)	0.331 (0.477)	0.262 (0.446)
rea90205s_c	-0.021 (-0.030)	-0.100 (-0.147)	0.143 (0.208)	-0.199 (-0.340)	0.129 (0.196)	0.062 (0.090)	
rea90206s_c	0.048 (0.069)	-0.125 (-0.183)	0.204 (0.297)	-0.050 (-0.086)	0.064 (0.097)	-0.007 (-0.010)	
rea903010_c	0.065 (0.094)	-0.086 (-0.125)	0.170 (0.247)	0.117 (0.200)	-0.279 (-0.423)	0.158 (0.227)	-0.090 (-0.153)
rea903020_c	0.110 (0.158)	-0.010 (-0.014)	-0.166 (-0.241)	-0.062 (-0.106)	0.367 (0.557)	-0.049 (-0.071)	-0.137 (-0.233)
rea903030_c	-0.093 (-0.134)	-0.163 (-0.239)	0.019 (0.028)	-0.126 (-0.214)	-0.096 (-0.146)	0.033 (0.048)	-0.231 (-0.393)
rea903040_c	-0.269 (-0.387)	0.090 (0.131)	-0.242 (-0.352)	0.214 (0.366)	-0.110 (-0.167)	-0.011 (-0.016)	0.182 (0.311)
rea90305s_c	-0.100 (-0.143)	-0.089 (-0.131)	0.126 (0.183)	0.035 (0.060)	-0.051 (-0.077)	0.011 (0.016)	-0.107 (-0.183)

Item	Sex	Books	Migration	School	Age	Position	Booklet
rea903060_c	-0.172 (-0.248)	-0.009 (-0.013)	-0.130 (-0.190)	-0.188 (-0.321)	-0.017 (-0.026)	-0.041 (-0.059)	-0.207 (-0.352)
rea90307s_c	-0.141 (-0.203)	-0.135 (-0.198)	0.173 (0.251)	-0.272 (-0.463)	0.169 (0.256)	-0.073 (-0.105)	
res1203080_c	-0.085 (-0.122)	-0.085 (-0.124)	0.036 (0.052)	*NV*		-0.072 (-0.104)	
rea904010_c	-0.207 (-0.297)	-0.051 (-0.075)	0.135 (0.196)	0.084 (0.143)	-0.035 (-0.053)	-0.068 (-0.098)	-0.046 (-0.079)
rea90402s_c	0.013 (0.019)	-0.082 (-0.121)	0.107 (0.156)	0.046 (0.078)	0.073 (0.111)	0.013 (0.019)	-0.102 (-0.174)
rea90403s_c	-0.147 (-0.212)	0.097 (0.142)	-0.097 (-0.141)	0.056 (0.095)	0.040 (0.060)	-0.096 (-0.137)	-0.046 (-0.078)
rea904040_c	-0.014 (-0.021)	-0.023 (-0.034)	0.057 (0.083)	-0.086 (-0.146)	0.007 (0.010)	0.047 (0.068)	-0.102 (-0.174)
res1204050_c	-0.170 (-0.244)	-0.027 (-0.040)	0.200 (0.291)	*NV*		-0.081 (-0.117)	
rea905010_c	0.305 (0.439)	0.189 (0.277)	-0.366 (-0.532)	0.029 (0.049)	0.145 (0.219)	-0.196 (-0.282)	
rea905020_c	0.298 (0.429)	0.342 (0.502)	-0.245 (-0.357)	0.143 (0.244)	-0.043 (-0.065)	-0.223 (-0.320)	
rea905030_c	0.200 (0.288)	-0.038 (-0.056)	0.138 (0.200)	-0.098 (-0.167)	-0.316 (-0.480)	-0.060 (-0.086)	
rea905040_c	0.129 (0.185)	0.144 (0.211)	-0.096 (-0.139)	0.149 (0.255)	0.034 (0.052)	-0.110 (-0.158)	
rea905050_c	-0.097 (-0.139)	0.110 (0.161)	-0.208 (-0.303)	-0.076 (-0.130)	0.015 (0.022)	-0.106 (-0.153)	
rea905060_c	-0.236 (-0.339)	-0.006 (-0.009)	0.054 (0.078)	-0.126 (-0.215)	0.037 (0.057)	0.113 (0.163)	
rea906010_c	-0.011 (-0.015)	0.013 (0.019)	-0.123 (-0.179)	-0.114 (-0.194)	-0.011 (-0.017)	-0.013 (-0.019)	
rea906020_c	-0.030 (-0.043)	-0.079 (-0.116)	-0.202 (-0.294)	-0.175 (-0.299)	0.000 (0.000)	0.144 (0.208)	
rea906030_c	-0.060	-0.039	0.037	-0.044	0.055	0.053	

Item	Sex	Books	Migration	School	Age	Position	Booklet
	(-0.086)	(-0.057)	(0.054)	(-0.075)	(0.083)	(0.076)	
rea90604s_c	0.295 (0.423)	-0.066 (-0.097)	-0.133 (-0.193)	-0.197 (-0.337)	0.022 (0.034)	0.001 (0.001)	
rea907010_c	0.256 (0.369)	-0.052 (-0.076)	0.074 (0.108)	0.161 (0.275)	0.003 (0.005)	-0.088 (-0.127)	
rea907020_c	0.111 (0.160)	-0.180 (-0.265)	0.085 (0.123)	-0.109 (-0.186)	-0.381 (-0.579)	-0.044 (-0.063)	
rea907030_c	0.527 (0.758)	0.283 (0.415)	-0.192 (-0.279)	0.130 (0.222)	0.097 (0.147)	-0.160 (-0.231)	
rea90704s_c	0.100 (0.143)	-0.073 (-0.108)	-0.003 (-0.004)	0.084 (0.144)	0.035 (0.053)	-0.059 (-0.086)	
Main effect (DIF model)	-0.084 (-0.121)	-0.279 (-0.409)	0.383 (0.557)	-0.791 (-1.350)	0.407 (0.617)	-0.010 (-0.014)	-0.935 (-1.595)
Main effect (Main effect model)	-0.064 (-0.092)	-0.242 (-0.354)	0.343 (0.499)	-0.815 (-1.384)	0.411 (0.624)	-0.030 (-0.043)	-0.920 (-1.565)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. Sec. = Secondary school (German: „Gymnasium"). Age DIF is only reported for SC 6, because this is the only cohort with considerable age differences. NV in the column School indicates items that were administered in SC 5 only and no DIF was calculated as nearly all participants in SC 5 finished secondary school. No absolute standardized difference is significantly, $p < .05$, greater than 0.25 (see Fischer et al., 2016).

Sex: The sample included 8,428 (47%) males and 9,544 (53%) females. On average, male participants had a lower estimated reading ability than females (main effect = -0.084 logits, Cohen's d = -0.121). None of the items showed a noticeable DIF above 0.6. logits. Only one item (rea907030_c) showed DIF between 0.4 and 0.6 logits. 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). Model comparisons using Akaike's (1974) information criterion (AIC) and the Bayesian information criterion (BIC; Schwarz, 1978) both favored the model estimating DIF. Nevertheless, the deviation was small in both cases. Thus, overall, there was no pronounced DIF regarding the gender of the participants.

Books: The number of books at home was used as a proxy for socioeconomic status. There were 6,764 (38%) test takers with 0 to 100 books at home, 10,271 (57%) test takers with more than 100 books at home, and 937 (5%) test takers without a valid response. There were considerable average differences between the two groups. Participants with 100 or less books at home performed on average 0.279 logits (Cohen's d = -0.409) lower in reading than participants with more than 100 books. There was no considerable DIF comparing participants with many or fewer books (highest DIF = 0.342 for item rea905020_c). Whereas the AIC favored the model estimating DIF, the BIC favored the main effects model (Table 13).

Migration background: There were 16,138 participants (90%) with no migration background, and 1,790 respondents (10%) with a migration background. In comparison to participants with migration background, participants without migration background had, on average, a higher reading ability (main effect = 0.383 logits, Cohen's d = 0.557). There was no noteworthy item DIF due to migration background; differences in estimated difficulties did not exceed 0.4 logits. The overall test for DIF using the BIC favored the main effects model, while the AIC favored the model estimating DIF.

School type: Overall, 11,803 respondents (66%) who took the reading test had attended secondary school (German: "Gymnasium") whereas 6,015 (33%) had not. There were 154 (1%) test takers without respective information. Participants from secondary schools showed on average a higher reading ability (0.791 logits, Cohen's d = 1.350). There was no noteworthy item DIF; no item exhibited DIF greater than 0.5 logits. However, the overall model test indicated a slightly better fit for the more complex DIF model, because several items showed DIF effects between 0.2 and 0.5. Nevertheless, these differences were not considered beeing severe.

Age: Age differences were only calculated for participants of starting cohort 6, because this is the only cohort with a considerable age range. 3,213 (50%) of the respondents were younger than 54 years and therefore grouped in the younger agegroup, while 3,200 (50%) participants were at least 54 years old and grouped in the older agegroup. Younger respondents showed on average a slightly higher reading ability (0.407 logits, Cohen's d = 0.617) than respondents of the older age group. There was no noteworthy item DIF; no item exhibited DIF greater than 0.4 logits. The overall test for DIF using the BIC favored the main effects model, while the AIC favored the model estimating DIF.

Position: The reading competence test was administered in two different positions (see section 3 for the design of the study). A sample of 10,161 (57%) persons received the reading test first and 7,811 (43%) respondents took the reading test after having completed the mathematics 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 showed no average effect of item position[3]. In this study, persons who received the reading test first performed on average 0.010 logits (Cohen's d = 0.014) better than respondents who received the reading test second. 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 0.331 logits (item rea902040_c). The overall test for DIF using the BIC favored the main effects model, while the AIC favored the model estimating DIF.

Booklet: To estimate the participants' proficiency with great accuracy, the participants received different tests that either included a larger number of easy or a larger number of difficult items (see section 3 for the design of the study). Only a subset of 14 items were included in both tests and, in effect, administered to all participants. For these common items we examined potential DIF across the two test versions (easy versus difficult). A subsample of 6,298 (35%) persons received the easy test and 11,674 (65%) persons received the difficult test. As expected, participants who were administered the easy test scored on average -

[3] 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.

0.935 logits (Cohen's d = -1.595) lower than participants who received the difficult test. There was no DIF for the common items regarding the test version. The largest difference in difficulties between the two groups was 0.410 logits (item rea902020_c).

Table 12

Differential Item Functioning between Starting Cohorts and Administration Mode

Item	**Sample**			**Mode SC5**
	SC 4 vs. SC 6	SC 5 vs. SC 6	SC 4 vs. SC 5	CBT vs. WBT
rea90201s_c	-0.373* (-0.569)	-0.273 (-0.418)	-0.099 (-0.152)	-0.021 (-0.034)
rea902020_c	0.156 (-0.238)	0.038 (0.057)	0.118 (0.181)	0.065 (0.105)
rea902030_c	-0.054 (-0.082)	-0.295 (-0.451)	0.241 (0.369)	-0.087 (-0.141)
rea90205s_c				0.092 (0.150)
rea90206s_c				0.041 (0.066)
rea903010_c	-0.737* (-1.127)	-0.247 (-0.378)	-0.490* (-0.749)	0.025 (0.041)
rea903020_c	0.487* (0.744)	0.347 (0.530)	0.140 (0.214)	-0.055 (-0.089)
rea903030_c	0.263 (0.402)	0.199 (0.304)	0.064 (0.097)	0.038 (0.062)
rea903040_c	0.276 (0.422)	0.054 (0.082)	0.222 (0.340)	0.003 (0.005)
rea90305s_c	-0.234 (-0.357)	-0.124 (-0.189)	-0.110 (-0.168)	-0.115 (-0.188)
res1203080_c				-0.032 (-0.053)
rea904010_c	0.388* (0.593)	0.315 (0.482)	0.073 (0.111)	0.052 (0.085)
rea90402s_c	-0.172 (-0.263)	-0.013 (-0.020)	-0.159 (-0.243)	-0.19 (-0.309)
res1204050_c				-0.084 (-0.137)
rea906010_c				0.046 (0.075)

Item	Sample			Mode SC5
	SC 4 vs. SC 6	SC 5 vs. SC 6	SC 4 vs. SC 5	CBT vs. WBT
rea906020_c				0.072 (0.117)
rea906030_c				0.125 (0.204)
rea90604s_c				-0.037 (-0.061)
rea907010_c				0.075 (0.122)
rea907020_c				0.067 (0.110)
rea90704s_c			0.07 (0.110)	-0.080 (-0.130)
Main effect (DIF model)	0.053 (0.205)	0.813 (0.965)	-0.760 (-1.037)	0.010 (0.016)
Main effect (Main effect model)	0.172 (0.361)	0.865 (1.055)	-0.693 (-0.979)	0.034 (0.055)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. *Absolute standardized difference is significantly, $p < .05$, greater than 0.25 (see Fischer et al., 2016).

Sample: Table 12 shows the DIF between the three starting cohorts (see column Sample) and 6,866 (43%) participants were from Starting Cohort 4, 2,763 (17%) from Starting Cohort 5 (only CBT participants), and 6,413 (40%) from Starting Cohort 6. The largest difference in difficulties between the three groups was -0.737 logits (item rea903010_c). As a consequence of these results, the items rea9021s_c and rea903020_c were separated and treated as unique items for Starting Cohort 6 in the scaling model. The item rea903010_c was treated as unique for all three starting cohorts in the scaling model.

Mode: The administration setting in Starting Cohort 5 was either WBT (41 %) or CBT (59 %). None of the items showed a noticeable DIF and all items presented a DIF below 0.2 logits. An overall test for DIF revealed that the AIC favored the model estimating DIF, while the BIC favored the main effects model (see Table 13). Nevertheless, the deviation was small and, overall, there was no pronounced DIF regarding the mode of the study.

Table 13

Comparisons of Models with and without DIF

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Sex	main effect	17,972	625,088.3	66	625,220.3	625,734.9
	DIF	17,972	624,491.2	103	624,697.2	625,500.2
Books	main effect	17,035	592,749.1	66	592,881.1	593,392.1
	DIF	17,035	592,406.6	103	592,612.6	593,410.2
Migration	main effect	17,928	623,040.0	66	623,172.0	623,686.4
	DIF	17,928	622,793.4	103	622,999.4	623,802.2
School type	main effect	17,818	604,380.2	64	604,508.2	605,006.6
	DIF	17,818	603,661.9	99	603,859.9	604,630.9
Age	main effect	6,413	220,072.6	64	220,200.6	220,633.7
	DIF	6,413	219,859.6	99	220,057.6	220,727.5
Position	main effect	17,972	625,111.6	66	625,243.6	625,758.2
	DIF	17,972	624,852.9	103	625,058.9	625,862.0
Booklet	main effect	17,972	353,667.4	27	353,721.4	353,931.9
	DIF	17,972	353,292.1	40	353,372.1	353,683.9
Sample	main effect	16,042	252,314.7	21	252,356.7	252,518.1
	DIF	16,042	251,133.3	39	251,211.3	251,510.9
Mode	main effect	4,693	145,571.3	39	145,649.3	145,901.0
	DIF	4,693	145,507.9	59	145,625.9	146,006.6

5.3.5 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 generalized partial credit model (GPCM; Muraki, 1992) that estimates discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 9), ranging from 0.23 (item rea901030_c) to 1.44 (item rea905020_c). The average discrimination parameter fell at 0.66 (*SD* = 0.17). Model fit indices suggested a slightly better model fit of the GPCM (AIC = 560,713, BIC = 561,581) as compared to the PCM model (AIC = 563,013, BIC = 563,567). Despite the empirical preference for the GPCM, the PCM more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012; 2013; for a discussion of this issue). For this reason, the PCM was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

5.3.6 Unidimensionality

The unidimensionality of the test was investigated by specifying two different multidimensional models and comparing them to a unidimensional model. In the first multidimensional model, three different cognitive requirements were specified, whereas the five different text types constituted the second multidimensional model. Both multidimensional models were estimated using Quasi Monte Carlo method with 10,000 nodes. The estimated variances and correlations between the three dimensions representing the different cognitive requirements are reported in Table 14. The correlations among the three dimensions were rather high and ranged from .844 to .926. However, they deviated from a perfect correlation (i.e., they were marginally lower than r = .95, see Carstensen, 2013). According to model fit indices, the unidimensional model fitted the data slightly better (AIC = 528,217, BIC = 528,717, number of parameters = 65) than the three-dimensional model (AIC = 530,035, BIC = 530,573, number of parameters = 70). Therefore, these results indicate that the three cognitive requirements measure an essentially unidimensional construct.

Table 14

Results of Three-Dimensional Scaling

	Dim 1	**Dim 2**	**Dim 3**
Finding information in the text (Dim 1)	(0.493)		
(6 items)			
Drawing text-related conclusions (Dim 2)	0.844	(0.723)	
(16 items)			
Reflecting and assessing (Dim 3)	0.853	0.926	(0.637)
(16 items)			

Note. Variances of the dimensions are given in the diagonal and correlations are presented in the off-diagonal.

The estimated variances and correlations of the five-dimensional model based on the five text functions are given in Table 15. The correlations between the dimensions varied between r = .652 and r = .873. The smallest correlation was found between dimension 2 (“instruction”) and dimension 5 (“information”). Dimension 1 (“literary”) and dimension 3 (“commenting”) showed the strongest correlation. All correlations deviated from a perfect correlation (i.e., they were considerably lower than r = .95, see Carstensen, 2013). The unidimensional model (AIC = 528,217, BIC = 528,717, number of parameters = 65) fitted the data better than the five-dimensional model (AIC = 529,848 , BIC = 530,455, number of parameters = 79).

As each text function corresponded to one of the five texts, local item dependence (LID) and the text functions were confounded. In consequence, the deviation of the correlations from a perfect correlation shown in Table 15, may result from multidimensionality as well as from local item dependence. Given the testing design in the main studies, it is not possible to disentangle the two sources. In pilot studies (Gehrer et al., 2013), a larger number of texts were

presented to test takers, so that the impact of text functions could be investigated independently of LID. The correlations estimated in the pilot study ranged from .78 to .91. As the correlations found in Gehrer and colleagues (2013) differ from a perfect correlation, it is concluded that text functions form subdimensions of reading competence. Comparing the correlations found in Gehrer et al. (2013), which are due to text functions, to those found in the main study (Table 15), which are due to both text functions and LID, allows us to evaluate the impact of LID. The correlations found in the present study were somewhat lower (between 0.65 and 0.87) than those found in Gehrer et al. (2013; between 0.78 and 0.91), indicating that there is some amount of local item dependence. However, according to the test developers a balanced assessment of reading competence can only be achieved by a heterogeneity of text functions (Gehrer et al., 2013).

For the unidimensional model the average absolute residual correlations, as indicated by the Q_3 statistic (see Table 9), were quite low (M = .02, SD = .01). The largest individual residual correlation was .04 and thus indicated an essentially unidimensional test. As all relevant parameters support the measurement of a single dimension, a unidimensional reading competence score was estimated.

Table 15

Results of Five-Dimensional Scaling

	Dim 1	**Dim 2**	**Dim 3**	**Dim 4**	**Dim 5**
Literary (Dim 1)	(0.548)				
(10 items)					
Instruction (Dim 2)	0.731	(1.260)			
(9 items)					
Commenting (Dim 3)	0.873	0.715	(0.770)		
(6 items)					
Advertising (Dim 4)	0.767	0.674	0.791	(0.802)	
(5 items)					
Information (Dim 5)	0.813	0.652	0.799	0.753	(1.097)
(8 items)					

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

6. Discussion

The analyses in the previous sections reported information on the quality of the reading test that was administered in Starting Cohorts 4, 5, and 6. Furthermore, the estimation of the respective reading 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. However, the number of missing responses was rather large because many respondents did not finish the test in time. The test had a satisfactory reliability and distinguished well between test takers. However, the test was slightly better targeted at mediocre- and low-performing students and covered the high ability spectrum less well. As a consequence, ability estimates will be precise for low-performing respondents but less precise for high performing respondents. Furthermore, some degree of multidimensionality is present for different text functions. In combination with the high amount of missing responses at the end of the test (i.e., there are participants with no valid responses to some of the text functions), the estimation of a single reading competence score is challenged. This should be addressed in further studies. Nevertheless, Gehrer et al. (2013) argue that a balanced assessment of reading competence can only be achieved by heterogeneity of text functions and they provide theoretical arguments for a unidimensional measure of reading competence. In summary, the test had acceptable psychometric properties that allowed the estimation of a unidimensional reading competence score.

7. Data in the Scientific Use File

7.1 Naming conventions

The data in the Scientific Use File contains 36 items in Starting Cohort 4 (Wave 10) and Starting Cohort 6 (Wave 9), of which 23 items were scored as dichotomous variables (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. In Starting Cohort 5 (Wave 12), the Scientific Use File contains 21 items of which 14 items were scored as dichotomous variables (MC items). MC items are marked with a ‘0_c’ at the end of the variable name, whereas the variable names of CMC, MA and TET items end in ‘s_c’. For further details on the naming conventions of the variables see Fuß and colleagues (2019). In the IRT scaling model, the polytomous CMC, MA, and TET variables were scored as 0.5 for each category.

7.2 Linking of competence scores

In all starting cohorts, reading competence was measured in the current wave and also in a previous wave. The tests in the different waves were constructed in such a way as to allow for an accurate measurement of reading competence within the respective age group (Gehrer 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. Following an anchor-group design, independent link samples that were

not part of the current waves, were administered all items from the current reading competence tests and their previous sample within a single measurement occasion. These responses were used to link the tests across the two waves of each starting cohort. In the following, the linking procedure will be described for the Starting Cohorts 4, 5, and 6 successively.

7.2.1 Linking in the SC 4

7.2.1.1 Sample

In Starting Cohort 4, a subsample of 3,450 respondents (54% women) participated at both measurement occasions, in wave 7 (i.e., grade 12; see Gnambs, Fischer,& Rohm, 2017) and also in wave 10 (young adults; see above). Consequently, these respondents were used to link the two tests across both waves (see Fischer et al., 2016). Moreover, an independent link sample of *N* = 813 young adults (48% women) received both tests within a single measurement occasion.

7.2.1.2 The design of the link study

The test administered to students in wave 7 included 29 items (see Gnambs, Fischer, & Rohm, 2017), whereas the test administered in wave 10 had 36 items (see above). Again, two versions of the test were used in the link study (easy and difficult).

A random sample of 408 students received the easy test version and 405 students were administered the difficult version. No multi-matrix design regarding the selection and order of the items within a test was established. Thus, all test takers were given the reading items in the same order.

7.2.1.3 Results

To examine whether the two tests administered in the link sample measured a common scale, we compared a one-dimensional model that specified a single latent factor for all items to a two-dimensional model that specified separate latent factors for the two tests. The information criteria slightly favored the two-dimensional model, AIC = 51,608, BIC = 52,158, over the one-dimensional model, AIC = 51,678, BIC = 52,219. However, an examination of the residual correlations for the one-dimensional model using the corrected Q_3 statistic (Yen, 1984) indicated a largely unidimensional scale— the average absolute residual correlation was $M = .05$ ($SD = .04$, $Max = .31$). This indicates that the reading competence tests administered in waves 7 and 10 were essentially unidimensional.

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 4 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 16.

Table 16

Differential Item Functioning Analyses between Wave 7, Wave 10, and the Link Sample.

	wave 7				wave 10			
	Item	**Δσ**	***SE*$_{Δσ}$**	***F***	**Item**	**Δσ**	***SE*$_{Δσ}$**	***F***
1.	reg120110_c	-0.38	0.15	6.16	rea90101s_c	-0.03	0.07	0.21
2.	reg120120_c	-0.15	0.14	1.18	rea90102s_c	-0.18	0.10	2.88
3.	reg120130_c	-0.40	0.22	3.44	rea901030_c	0.03	0.13	0.07
4.	reg12014s_c*	-2.50	0.29	73.65	rea90104s_c	-0.18	0.08	5.18
5.	reg120150_c	-0.22	0.14	2.56	rea90105s_c	-0.33	0.06	26.56
6.	reg120160_c*	-0.69	0.16	19.53	rea90201s_c	0.08	0.04	3.81
7.	reg120170_c	0.04	0.18	0.06	rea902020_c	0.15	0.09	2.74
8.	reg12021s_c*	1.27	0.09	193.31	rea902030_c	0.11	0.09	1.54
9.	reg120220_c	-0.17	0.09	3.63	rea902040_c	0.16	0.09	3.35
10.	reg120230_c	0.09	0.11	0.78	rea90206s_c	0.22	0.07	8.73
11.	reg12024s_c	0.06	0.04	2.10	rea903010_c	0.30	0.09	12.32
12.	reg120250_c	0.20	0.11	3.46	rea903020_c	-0.11	0.09	1.38
13.	reg12026s_c	0.04	0.03	1.89	rea903030_c	-0.03	0.10	0.12
14.	reg120310_c	-0.28	0.09	9.86	rea903040_c	0.06	0.09	0.53
15.	reg120320_c	0.11	0.11	1.16	rea90305s_c	-0.02	0.04	0.15
16.	reg120330_c	-0.18	0.10	3.25	rea903060_c	0.23	0.09	6.84
17.	reg120340_c	0.04	0.13	0.11	rea90307s_c	-0.09	0.08	1.27
18.	reg120350_c	0.00	0.09	0.00	rea904010_c	0.00	0.09	0.00
19.	reg120360_c	0.20	0.09	4.35	rea90402s_c	-0.01	0.04	0.04
20.	reg12042s_c*	1.91	0.09	422.89	rea90403s_c	-0.02	0.05	0.18
21.	reg120430_c	0.21	0.10	4.70	rea904040_c	0.06	0.10	0.41
22.	reg12044s_c	0.00	0.04	0.00	rea905010_c	0.04	0.16	0.07
23.	reg120450_c	-0.18	0.11	2.61	rea905020_c	-0.25	0.19	1.84
24.	reg120510_c	0.03	0.24	0.01	rea905030_c	-0.08	0.17	0.21
25.	reg12052s_c	-0.09	0.10	0.84	rea905040_c	-0.27	0.16	2.84
26.	reg120530_c*	-0.52	0.25	4.14	rea905050_c	-0.25	0.17	2.06
27.	reg120540_c	-0.17	0.20	0.70	rea905060_c	-0.16	0.18	0.79
28.	reg12055s_c	0.28	0.10	7.84	rea906010_c	0.03	0.12	0.07
29.	reg120560_c	-0.48	0.21	5.34	rea906020_c	0.27	0.12	5.23

	wave 7				wave 10			
	Item	**Δσ**	***SE*$_{Δσ}$**	***F***	**Item**	**Δσ**	***SE*$_{Δσ}$**	***F***
30.	reg120610_c	0.06	0.12	0.26	rea906030_c	-0.05	0.14	0.13
31.	reg120620_c	0.06	0.13	0.22	rea90604s_c	0.27	0.07	17.33
32.	reg120630_c	0.13	0.14	0.88	rea907010_c	0.11	0.18	0.37
33.	reg120640_c	-0.22	0.12	3.23	rea907020_c	-0.19	0.16	1.48
34.	reg12065s_c	-0.30	0.07	20.90	rea907030_c	-0.10	0.16	0.40
35.	reg120660_c	-0.15	0.13	1.33	rea90704s_c	0.20	0.10	3.93
36.	reg120670_c	0.35	0.13	6.83				
37.	reg12071s_c*	0.87	0.17	27.71				
38.	reg120720_c	0.06	0.20	0.08				
39.	reg120730_c	0.07	0.17	0.17				
40.	reg120740_c	0.10	0.18	0.29				
41.	reg12075s_c*	0.91	0.11	64.30				

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in wave 7 or 10 and the link sample (positive values indicate easier items in the link sample); *SE*$_{Δσ}$ = 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} (2, 4,261) = 48.75. A non-significant test indicates measurement invariance.
*item excluded from the linking procedure.

Analyses of differential item functioning between the link sample and Starting Cohort 4 identified 4 items with significant (α = .05) DIF for wave 7, and 3 items with a difference in item difficulty parameters greater than 0.5 logits (differences in logits between -0.52 and -2.50 for these 7 items). No significant (α = .05) differences were found for wave 10 (difference in logits: *Min* = 0.00, *Max* = -0.33). The relevant items are marked with an asterisk in Table 16 and were excluded prior to linking the reading competence tests using the “mean/mean” method for the anchor-group design (see Fischer et al., 2016).

The correction term was calculated as $c = 0.053$. This correction term was subsequently added to each difficulty parameter estimated in wave 10 (see Table 9) 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 0.045 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.

7.2.2 Linking in the SC 5

7.2.2.1 Sample

In Starting Cohort 5, a subsample of 2,277 young adults (64% women) participated at both measurement occasions, in wave 1 (see Pohl, Haberkorn & Hardt, 2014) and also in wave 12 (see above). These respondents were used to link the two tests across both waves. Please note, that the subsample of wave 12 included only respondents from the CBT condition. Moreover, an independent link sample of *N* = 536 young adults (52% women) received both tests within a single measurement occasion.

7.2.2.2 The design of the link study

The test administered to students in wave 1 included 29 items (see Pohl, Haberkorn & Hardt, 2014), whereas the test administered to adults in wave 12 included 21 items (see above). For SC 5, only the difficult reading test versions were used. No multi-matrix design regarding the selection and order of the items within a test was established. Thus, all test takers were given the reading items in the same order.

7.2.2.3 Results

To examine whether the two tests administered in the link sample measured a common scale, we compared a one-dimensional model that specified a single latent factor for all items to a two-dimensional model that specified separate latent factors for the two tests. The information criteria slightly favored the two-dimensional model, AIC = 28,668, BIC = 28,980, over the one-dimensional model, AIC = 28,681, BIC = 28,985. However, an examination of the residual correlations for the one-dimensional model using the corrected Q_3 statistic (Yen, 1984) indicated a largely unidimensional scale—the average absolute residual correlation was $M = .05$ ($SD = .04$, $Max = .20$). This indicates that the reading competence tests administered in waves 1 and 12 were essentially unidimensional.

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 5 and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 17.

Table 17

Differential Item Functioning Analyses between Wave 1, Wave 12, and the Link Sample.

	wave 1				wave 12			
	Item	**Δσ**	***SE*$_{Δσ}$**	***F***	**Item**	**Δσ**	***SE*$_{Δσ}$**	***F***
1.	res10110_c	0.08	0.12	0.42	rea90201s_c	-0.15	0.06	6.12
2.	res1012s_c	0.47	0.14	10.72	rea902020_c	0.05	0.13	0.14
3.	res10130_c	0.16	0.25	0.41	rea902030_c	0.05	0.10	0.28
4.	res10140_c	-0.04	0.11	0.11	rea90206s_c	-0.15	0.06	6.41

	wave 1				wave 12			
	Item	**Δσ**	***SE*$_{Δσ}$**	***F***	**Item**	**Δσ**	***SE*$_{Δσ}$**	***F***
5.	res10160_c	-0.11	0.10	1.14	rea903010_c	-0.11	0.10	1.21
6.	res10170_c	-0.01	0.12	0.01	rea903020_c	0.01	0.10	0.01
7.	res10180_c	-0.15	0.18	0.69	rea903030_c	-0.08	0.11	0.52
8.	res10190_c	0.16	0.10	2.55	rea903040_c	0.19	0.11	3.00
9.	res1021s_c	-0.10	0.09	1.11	rea90305s_c	-0.11	0.05	4.31
10.	res1022s_c	0.16	0.10	2.70	res1203080_c	0.04	0.11	0.13
11.	res10230_c	-0.13	0.11	1.36	rea904010_c	0.22	0.11	3.85
12.	res1024s_c	0.09	0.08	1.28	rea90402s_c	-0.15	0.05	8.82
13.	res10250_c	-0.09	0.11	0.64	res1204050_c	-0.01	0.12	0.02
14.	res10260_c	0.16	0.14	1.19	rea906010_c	0.05	0.11	0.20
15.	res10270_c	-0.02	0.12	0.02	rea906020_c	0.12	0.10	1.46
16.	res10310_c	-0.22	0.17	1.70	rea906030_c	0.29	0.12	5.93
17.	res1032s_c	0.11	0.06	3.49	rea90604s_c	0.09	0.06	2.29
18.	res10330_c	-0.27	0.15	3.22	rea907010_c	-0.05	0.15	0.14
19.	res10340_c	0.14	0.12	1.52	rea907020_c	-0.21	0.15	2.11
20.	res10350_c	0.11	0.12	0.94	rea90704s_c	-0.09	0.08	1.20
21.	res10360_c	0.12	0.13	0.88				
22.	res10370_c	-0.17	0.13	1.77				
23.	res10380_c	0.01	0.18	0.00				
24.	res10410_c	0.08	0.18	0.22				
25.	res10420_c	0.22	0.24	0.82				
26.	res1043s_c	-0.17	0.06	6.76				
27.	res10440_c	-0.25	0.17	2.13				
28.	res10450_c	-0.35	0.21	2.87				

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in wave 1 or 12 and the link sample (positive values indicate easier items in the link sample); *SE*$_{Δσ}$ = 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} (2, 2,811) = 34.89. A non-significant test indicates measurement invariance.

Analyses of differential item functioning between the link sample and Starting Cohort 5 identified neither for wave 1 (difference in logits: *Min* = 0.01, *Max* = 0.47) nor for wave 12 (difference in logits: *Min* = 0.01, *Max* = 0.29) items with significant (α = .05) DIF. The reading

competence tests administered in the two waves were linked using the "mean/mean" method for the anchor-group design (see Fischer et al., 2016).

The correction term was calculated as $c = 0.226$. This correction term was subsequently added to each difficulty parameter estimated in wave 12 (see Table 9) 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 0.046 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 Linking in the SC 6

The sample of Starting Cohort 6 had to be splitted into two parts, because one part of the respondents took the first reading test in the main study 2010/11 (see Hardt, Pohl, Haberkorn & Wiegand, 2013) and the other part of the respondents took the same test in 2012 (refreshment sample; see Koller, Haberkorn & Rohm, 2014). Each part was linked with its respective previous sample by using the above mentioned anchor-group design.

7.2.3.1 Sample

The first part of Starting Cohort 6 participated in the main study 2010/11, as well as in the current measurement occasion, yielding a subsample of 3,314 adults (52% women). These respondents were used to link the two tests across both wave 3 and 9 (see Fischer et al., 2016). A subsample of 2,021 adults (49% women) participated in the main study 2012 of Starting Cohort 6, as well as in the current measurement occasion. These respondents were used to link the two tests across wave 5 and 9 (see Fischer et al., 2016.). The independent link sample of *N* =1,294 adults (53% women) was used to link both parts of Starting Cohort 6. Respondents of the link sample received both tests within a single measurement occasion.

7.2.3.2 The design of the link study

The test administered to adults in wave 3 and 5 included 32 items (see Hardt, Pohl, Haberkorn & Wiegand, 2013; Koller, Haberkorn & Rohm, 2014), whereas the test administered to adults in wave 9 included 23 items (see above). A random link sample of 670 adults received the reading test before working on a mathematics test, whereas the remaining 624 adults received the mathematics test before the reading test. Moreover, 393 of the randomly selected participants received the easy test version of the wave 9 reading test and 307 persons were administered the difficult version. No multi-matrix design regarding the selection and order of the items within a test was established. Thus, all test takers were given the reading items in the same order.

7.2.3.3 Results

To examine whether the two tests administered in the link sample measured a common scale, we compared a one-dimensional model that specified a single latent factor for all items to a two-dimensional model that specified separate latent factors for the two tests. The information criteria slightly favored the two-dimensional model, AIC = 36,282, BIC = 36,803, over the one-dimensional model, AIC = 36,299, BIC = 36,811. However, an examination of the residual correlations for the one-dimensional model using the corrected Q_3 statistic (Yen, 1984) indicated a largely unidimensional scale—the average absolute residual correlation was $M = .06$ ($SD = .05$, $Max = .32$). This indicates that the reading competence tests administered in waves 3, 5 and 9 were essentially unidimensional.

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 6, main study 2010/2011 (wave 3). The differences in item difficulties between the link sample and Starting Cohort 6, main study 2010/2011 (wave 3) and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 18.

Table 18

Differential Item Functioning Analyses between Wave 3, Wave 9, and the Link Sample.

	Wave 3				Wave 9			
	Item	**Δσ**	***SE*$_{Δσ}$**	***F***	**Item**	**Δσ**	***SE*$_{Δσ}$**	***F***
1.	rea30110_c	-0.07	0.23	0.09	rea90101s_c	-0.17	0.07	6.14
2.	rea3012s_c	-0.15	0.09	2.36	rea90102s_c	0.03	0.09	0.13
3.	rea30130_c	0.06	0.20	0.08	rea901030_c	-0.22	0.13	3.09
4.	rea30140_c	-0.16	0.14	1.26	rea90104s_c	-0.14	0.09	2.31
5.	rea3015s_c	0.06	0.07	0.80	rea90105s_c	-0.24	0.05	19.77
6.	rea30210_c	0.30	0.20	2.14	rea90201s_c	-0.17	0.05	12.55
7.	rea30220_c	-0.11	0.16	0.42	rea902020_c	0.25	0.10	6.46
8.	rea30230_c	0.34	0.17	4.10	rea902030_c	0.08	0.10	0.63
9.	rea30240_c	-0.19	0.16	1.41	rea902040_c	0.20	0.10	4.31
10.	rea30250_c*	-0.52	0.16	10.60	rea90206s_c	-0.09	0.09	1.04
11.	rea3028s_c	-0.08	0.04	4.81	rea903010_c	0.03	0.10	0.11
12.	rea30310_c	0.04	0.11	0.11	rea903020_c	0.07	0.10	0.49
13.	rea30320_c	-0.23	0.14	2.51	rea903030_c	-0.05	0.11	0.19
14.	rea30330_c	-0.17	0.13	1.78	rea903040_c	0.10	0.10	1.13
15.	rea30340_c	0.10	0.11	0.84	rea90305s_c	0.02	0.05	0.10
16.	rea30350_c	0.11	0.14	0.56	rea903060_c	0.09	0.10	0.80
17.	rea30360_c	0.17	0.14	1.62	rea90307s_c	-0.23	0.08	7.47
18.	rea30370_c	0.24	0.12	4.13	rea904010_c	0.11	0.11	1.04
19.	rea3038s_c	-0.08	0.08	1.15	rea90402s_c	-0.05	0.05	1.30
20.	rea30410_c	-0.14	0.12	1.38	rea90403s_c	-0.01	0.05	0.02
21.	rea3042s_c	-0.07	0.07	1.08	rea904040_c	0.09	0.11	0.57
22.	rea30430_c	-0.39	0.15	6.84	rea905010_c	-0.33	0.17	3.60
23.	rea30440_c	0.38	0.18	4.40	rea905020_c	-0.31	0.20	2.37

	Wave 3				Wave 9			
	Item	**Δσ**	***SE*$_{\Delta\sigma}$**	***F***	**Item**	**Δσ**	***SE*$_{\Delta\sigma}$**	***F***
24.	rea30450_c	0.02	0.16	0.01	rea905030_c	0.22	0.16	1.88
25.	rea30460_c	-0.02	0.12	0.02	rea905040_c	-0.47	0.17	7.55
26.	rea30510_c	0.32	0.22	2.13	rea905050_c	-0.09	0.20	0.18
27.	rea3052s_c	0.02	0.18	0.02	rea905060_c	-0.23	0.21	1.30
28.	rea30530_c	0.29	0.18	2.73	rea906010_c	0.27	0.14	3.75
29.	rea3054s_c	0.08	0.11	0.58	rea906020_c	0.19	0.14	1.89
30.	rea30550_c	-0.16	0.19	0.71	rea906030_c	0.08	0.17	0.22
31.					rea90604s_c	0.18	0.08	5.65
32.					rea907010_c*	-0.54	0.23	5.34
33.					rea907020_c*	0.52	0.21	5.97
34.					rea907030_c*	0.51	0.22	5.31
35.					rea90704s_c	0.31	0.16	3.75

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in wave 3 or wave 9 and the link sample (positive values indicate easier 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} (2, 4,606) = 51.98. A non-significant test indicates measurement invariance. *item excluded from the linking procedure.

Analyses of differential item functioning between the link sample and SC 6 identified one item with indicated DIF (difference in logits: -0.52) for wave 3 and three items with incidated DIF for wave 9 (differences in logits: 0.51, 0.52, and -0.54). The relevant items are marked with an asterisk in Table 18 and were excluded prior to linking the reading competence tests using the “mean/mean” method for the anchor-group design (see Fischer et al., 2016).

The correction term was calculated as $c = -0.040$. This correction term was subsequently added to each difficulty parameter estimated in wave 9 (see Table 9) 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 0.050 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.

The differences in item difficulties between the link sample and longitudinal subsample from the Starting Cohort 6, main study 2012 (wave 5) and the respective tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 19.

Table 19

Differential Item Functioning Analyses between Wave 5, Wave 9, and the Link Sample.

	Wave 5				Wave 9			
	Item	**Δσ**	$SE_{\Delta\sigma}$	***F***	**Item**	**Δσ**	$SE_{\Delta\sigma}$	***F***
1.	rea30110_c	-0.08	0.24	0.12	rea90101s_c	-0.22	0.07	8.89
2.	rea3012s_c	-0.10	0.10	1.06	rea90102s_c	0.02	0.09	0.04
3.	rea30130_c	-0.17	0.20	0.74	rea901030_c	-0.17	0.13	1.53
4.	rea30140_c*	-0.55	0.14	14.87	rea90104s_c	-0.13	0.10	1.84
5.	rea3015s_c	0.20	0.07	7.32	rea90105s_c	-0.32	0.06	30.04
6.	rea30210_c	0.19	0.21	0.86	rea90201s_c	-0.05	0.05	0.88
7.	rea30220_c	-0.23	0.17	1.95	rea902020_c	0.12	0.10	1.52
8.	rea30230_c	0.21	0.17	1.57	rea902030_c	-0.04	0.11	0.11
9.	rea30240_c	-0.08	0.16	0.26	rea902040_c	0.12	0.10	1.27
10.	rea30250_c	-0.41	0.16	6.20	rea90206s_c	-0.08	0.09	0.72
11.	rea3028s_c	0.09	0.04	5.76	rea903010_c	0.02	0.11	0.05
12.	rea30310_c	0.17	0.12	2.07	rea903020_c	0.12	0.10	1.49
13.	rea30320_c	-0.16	0.15	1.19	rea903030_c	-0.05	0.11	0.20
14.	rea30330_c	-0.01	0.14	0.01	rea903040_c	0.06	0.10	0.33
15.	rea30340_c	0.02	0.11	0.04	rea90305s_c	0.10	0.06	3.52
16.	rea30350_c	0.26	0.15	2.87	rea903060_c	0.10	0.10	0.94
17.	rea30360_c	-0.01	0.14	0.01	rea90307s_c	-0.14	0.09	2.83
18.	rea30370_c	-0.01	0.12	0.01	rea904010_c	0.07	0.11	0.37
19.	rea3038s_c	-0.03	0.08	0.19	rea90402s_c	-0.04	0.05	0.72
20.	rea30410_c	-0.13	0.12	1.09	rea90403s_c	-0.01	0.06	0.01
21.	rea3042s_c	0.08	0.07	1.23	rea904040_c	0.09	0.12	0.61
22.	rea30430_c	-0.42	0.15	7.45	rea905010_c	-0.27	0.18	2.35
23.	rea30440_c	-0.10	0.18	0.31	rea905020_c	-0.45	0.21	4.85
24.	rea30450_c	0.06	0.17	0.12	rea905030_c	0.18	0.17	1.20
25.	rea30460_c	0.07	0.13	0.28	rea905040_c*	-0.52	0.18	8.38
26.	rea30510_c	0.43	0.23	3.59	rea905050_c	-0.23	0.20	1.28
27.	rea3052s_c	0.10	0.18	0.31	rea905060_c	-0.09	0.22	0.17
28.	rea30530_c	0.41	0.19	4.77	rea906010_c	0.23	0.15	2.32
29.	rea3054s_c	0.23	0.12	3.89	rea906020_c	0.03	0.15	0.03

	Wave 5				Wave 9			
	Item	**Δσ**	***SE*$_{\Delta\sigma}$**	***F***	**Item**	**Δσ**	***SE*$_{\Delta\sigma}$**	***F***
30.	rea30550_c	0.00	0.20	0.00	rea906030_c	0.05	0.18	0.09
31.					rea90604s_c	0.18	0.08	4.72
32.					rea907010_c	-0.24	0.25	0.98
33.					rea907020_c*	0.53	0.23	5.43
34.					rea907030_c*	0.68	0.24	8.25
35.					rea90704s_c	0.34	0.17	3.99

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in Wave 5 or Wave 9 and the link sample (positive values indicate easier 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} (2, 3,313) = 39.75. A non-significant test indicates measurement invariance. *item excluded from the linking procedure.

Analyses of differential item functioning between the link sample and SC 6 identified one item with DIF (difference in logits: -0.55) for wave 5 and three items with DIF for wave 9 (differences in logits: -0.52, 0.53, and 0.68). The relevant items are marked with an asterisk in Table 19 and were excluded prior to linking the reading competence tests using the "mean/mean" method for the anchor-group design (see Fischer et al., 2016).

The correction term was calculated as $c = 0.102$. This correction term was subsequently added to each difficulty parameter estimated in wave 9 (see Table 9) 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 0.048 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.

7.3 Reading competence scores

In the SUF, manifest reading competence scores are provided in the form of two different WLEs. In Starting Cohort 4, these are called "rea10_sc1" and "rea10_sc1u", including their respective standard errors, "rea10_sc2" and "rea10_sc2u". The corresponding names in Starting Cohort 5 are "res12_sc1" and "res12_sc1u", with the respective standard errors "res12_sc2" and "res12_sc2u". In the Starting Cohort 6, the WLEs are labelled as "rea9_sc1" and "rea9_sc1u", while the standard errors are labelled "rea9_sc2" and "rea9_sc2u". For the variables ending with "u" (e.g. "res12_sc1u"), person abilities were estimated using the linked item difficulty parameters. Subsequently, the estimated WLE scores were corrected for differences in the test position. In some studies, the reading test was either presented as the first or the second test within the test battery (see page 5). To correct for differences in the test position, we added the main effect related to the test position (see Table 11) to the WLE scores of respondents that received the reading test after working on another test. As a result, the WLE scores ending with "u" (e.g. "res12_sc1u" for Starting Cohort 5) can be used for longitudinal comparisons between the two waves (e.g. waves 1 and 12 in Starting Cohort 5). The resulting differences in WLE scores can be interpreted as development trajectories

across measurement points. In contrast, the WLE scores without a "u" at the end (e.g. "res12_sc1") are not linked to the underlying reference scale of the respective previous wave. However, they are corrected for the position of the reading test within the booklet. As a consequence, they cannot be used for longitudinal purposes but only for cross-sectional research questions. Because no substantial DIF was found for the proctored CBT and the unproctored WBT conditions in Starting Cohort 5, WLEs for respondents receiving the WBT were estimated using the fixed item parameters from the CBT scaling[4]. The R Syntax for estimating the WLE is provided in Appendix B. For persons who either did not take part in the reading test or who did not give enough valid responses, no WLE is estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values.

4 The test taking behavior in unproctored testing cannot be properly supervised and, thus, might not be comparable to proctored settings (see Kröhne, Gnambs, & Goldhammer, 2019). Therefore, we inspected the response times for respondents in the WBT condition. For 86 respondents exhibiting breaks, with no test interaction of more than five minutes during the test, no WLEs were estimated because they were suspected to adopt different test taking strategies.

References

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

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.

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, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Gehrer, K., & Artelt, C. (2013). Literalität und Bildungslaufbahn: Das Bildungspanel NEPS. In A. Bertschi-Kaufmann, & C. Rosebrock (Eds.), *Literalität erfassen: bildungspolitisch, kulturell, individuell* (pp. 168-187). Weinheim, Germany: Juventa.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). *The assessment of reading competence (including sample items for grade 5 and 9)*. Scientific Use File 2012, Version 1.0.0. Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online, 5*, 50-79.

Gnambs, T., Fischer, L., & Rohm, T. (2017). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 4 for Grade 12* (NEPS Survey Paper No. 13). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

Hardt, K., Pohl, S., Haberkorn, K., & Wiegand, E. (2013). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 6 for adults in main study 2010/11* (NEPS Working

Paper No. 25). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

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

Kröhne, U., Gnambs, T., & Goldhammer, F. (2019). Disentangling setting and mode effects for online competence assessment. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), *Education as a lifelong process* (2nd ed., pp. 171-193). Wiesbaden, Germany: Springer VS. https://doi.org/10.1007/978-3-658-23162-0_10

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. https://doi.org/10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Applied Psychological Measurement, 16*, 159-176. https://doi.org/10.1177/014662169201600206

Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement, 50*, 447-468. https://doi.org/10.1111/jedm.12028

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

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

Pohl, S., Haberkorn, K., & Hardt, K. (2014). *NEPS Technical Report for Reading – Scaling results of Starting Cohort 5 for first-year students* (NEPS Working Paper No. 34). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

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

Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules*. R package version 2.12-18. https://CRAN.R-project.org/package=TAM

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

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

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

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0*. Camberwell, Australia: Acer Press.

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

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

Appendix

Appendix A: Different Text Types and Cognitive Requirements

Item	Seq. Position	Text Types	Cognitive Requirements
rea90101s_sc4a10_c	1	Instruction text	Drawing text-related conclusions
rea90102s_sc4a10_c	2	Instruction text	Drawing text-related conclusions
rea901030_sc4a10_c	3	Instruction text	Drawing text-related conclusions
rea90104s_sc4a10_c	4	Instruction text	Drawing text-related conclusions
rea90105s_sc4a10_c	5	Instruction text	Reflecting and assessing
rea90201s_sc4a10_c	6	Commenting text	Drawing text-related conclusions
rea902020_sc4a10_c	7	Commenting text	Drawing text-related conclusions
rea902030_sc4a10_c	8	Commenting text	Reflecting and assessing
rea902040_sc4a10_c	9	Commenting text	Reflecting and assessing
rea90205s_sc4a10_c	10	Commenting text	Drawing text-related conclusions
rea90206s_sc4a10_c	11	Commenting text	Reflecting and assessing
rea903010_sc4a10_c	12	Information text	Reflecting and assessing
rea903020_sc4a10_c	13	Information text	Reflecting and assessing
rea903030_sc4a10_c	14	Information text	Finding information
rea903040_sc4a10_c	15	Information text	Reflecting and assessing
rea90305s_sc4a10_c	16	Information text	Reflecting and assessing
rea903060_sc4a10_c	17	Information text	Finding information
rea90307s_sc4a10_c	18	Information text	Finding information
res1203080_c	19	Information text	Finding information
rea904010_sc4a10_c	20	Advertising text	Drawing text-related conclusions
rea90402s_sc4a10_c	21	Advertising text	Reflecting and assessing
rea90403s_sc4a10_c	22	Advertising text	Drawing text-related conclusions
rea904040_sc4a10_c	23	Advertising text	Reflecting and assessing
res1204050_c	24	Advertising text	Drawing text-related conclusions
rea905010_sc4a10_c	25	Literary text	Finding information
rea905020_sc4a10_c	26	Literary text	Finding information
rea905030_sc4a10_c	27	Literary text	Drawing text-related conclusions
rea905040_sc4a10_c	28	Literary text	Reflecting and assessing
rea905050_sc4a10_c	29	Literary text	Reflecting and assessing
rea905060_sc4a10_c	30	Literary text	Drawing text-related conclusions
rea906010_sc4a10_c	31	Instruction text	Drawing text-related conclusions
rea906020_sc4a10_c	32	Instruction text	Drawing text-related conclusions
rea906030_sc4a10_c	33	Instruction text	Drawing text-related conclusions

rea90604s_sc4a10_c	34	Instruction text	Reflecting and assessing
rea907010_sc4a10_c	35	Literary text	Reflecting and assessing
rea907020_sc4a10_c	36	Literary text	Drawing text-related conclusions
rea907030_sc4a10_c	37	Literary text	Reflecting and assessing
rea90704s_sc4a10_c	38	Literary text	Reflecting and assessing

Note. Seq. Position = Sequential Position, item position within tests for SC4 and SC6 with item positions 19 and 24 added for additional SC5 items.

Appendix B: R-Syntax for estimating WLEs in Starting Cohorts 4, 5, & 6

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


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

# 38 items of the reading competence test
items <- c("rea90101s c", "rea90102s c",
           "rea901030 c", "rea90104s c",
           ...)

# identify polytomous items
f <- c("rea90307s_c", "rea90305s_c", "rea90101s_c",
       "rea90102s_c", "rea90104s_c", "rea90105s_c",
       "rea90201s_c", "rea90205s_c", "rea90206s_c",
       "rea90704s_c", "rea90402s_c", "rea90403s_c",
       "rea90604s_c")
f <- items %in% f

# collapse response categories
dat$rea90101s_c <- recodeVar(dat$rea90101s_c,
                             c(0, 1, 2, 3, 4),
                             c(0, 0, 1, 2, 3))
dat$rea90206s_c <- recodeVar(dat$rea90206s_c,
                             c(0, 1, 2, 3, 4),
                             c(0, 1, 2, 3, 3))
dat$rea90403s_c <- recodeVar(dat$rea90403s_c,
                             c(0, 1, 2, 3, 4, 5),
                             c(0, 0, 1, 2, 3, 4))

# define Q-matrix for 0.5 scoring of PCM
Q <- matrix(1, nrow = length(items), ncol = 1)
Q[f, 1] <- 0.5    # score of 0.5

# estimate partial credit model
mod <- tam.mml(resp = dat[, items], Q = Q, irtmodel = "PCM2",
               pid = dat$ID_t)
summary(mod)

# item fit
tam.fit(mod)

# WLE
tam.wle(mod)
```

Appendix C: Variable Names in Different Starting Cohorts

Seq. Position	SC 4	SC 5	SC 6
1	rea90101s_sc4a10_c		rea90101s_c
2	rea90102s_sc4a10_c		rea90102s_c
3	rea901030_sc4a10_c		rea901030_c
4	rea90104s_sc4a10_c		rea90104s_c
5	rea90105s_sc4a10_c		rea90105s_c
6	rea90201s_sc4a10_c	rea90201s_sc5s12_c	rea90201s_c
7	rea902020_sc4a10_c	rea902020_sc5s12_c	rea902020_c
8	rea902030_sc4a10_c	rea902030_sc5s12_c	rea902030_c
9	rea902040_sc4a10_c		rea902040_c
10	rea90205s_sc4a10_c	rea90205s_sc5s12_c	rea90205s_c
11	rea90206s_sc4a10_c	rea90206s_sc5s12_c	rea90206s_c
12	rea903010_sc4a10_c	rea903010_sc5s12_c	rea903010_c
13	rea903020_sc4a10_c	rea903020_sc5s12_c	rea903020_c
14	rea903030_sc4a10_c	rea903030_sc5s12_c	rea903030_c
15	rea903040_sc4a10_c	rea903040_sc5s12_c	rea903040_c
16	rea90305s_sc4a10_c	rea90305s_sc5s12_c	rea90305s_c
17	rea903060_sc4a10_c		rea903060_c
18	rea90307s_sc4a10_c		rea90307s_c
19		res1203080_c	
20	rea904010_sc4a10_c	rea904010_sc5s12_c	rea904010_c
21	rea90402s_sc4a10_c	rea90402s_sc5s12_c	rea90402s_c
22	rea90403s_sc4a10_c		rea90403s_c
23	rea904040_sc4a10_c		rea904040_c
24		res1204050_c	
25	rea905010_sc4a10_c		rea905010_c
26	rea905020_sc4a10_c		rea905020_c
27	rea905030_sc4a10_c		rea905030_c
28	rea905040_sc4a10_c		rea905040_c
29	rea905050_sc4a10_c		rea905050_c
30	rea905060_sc4a10_c		rea905060_c
31	rea906010_sc4a10_c	rea906010_sc5s12_c	rea906010_c
32	rea906020_sc4a10_c	rea906020_sc5s12_c	rea906020_c
33	rea906030_sc4a10_c	rea906030_sc5s12_c	rea906030_c
34	rea90604s_sc4a10_c	rea90604s_sc5s12_c	rea90604s_c

Seq. Position	SC 4	SC 5	SC 6
35	rea907010_sc4a10_c	rea907010_sc5s12_c	rea907010_c
36	rea907020_sc4a10_c	rea907020_sc5s12_c	rea907020_c
37	rea907030_sc4a10_c		rea907030_c
38	rea90704s_sc4a10_c	rea90704s_sc5s12_c	rea90704s_c

Note. Seq. Position = Sequential Position. The position of items in the three different test booklets is depicted in Table 8.