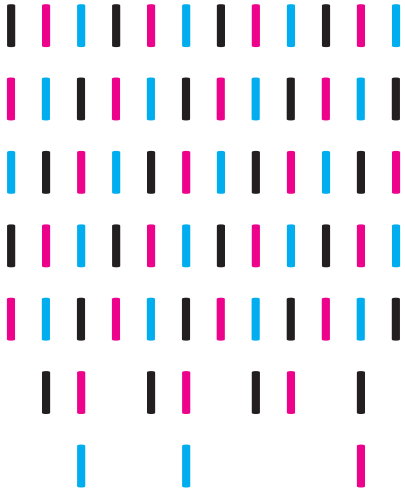

NEPS SURVEY PAPERS

Theresa Rohm, Micha Freund, Timo Gnambs, and Luise Fischer

NEPS TECHNICAL REPORT FOR LISTENING COMPREHENSION: SCALING RESULTS OF STARTING COHORT 3 FOR GRADE 9

NEPS Survey Paper No. 21
Bamberg, May 2017

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 Listening Comprehension: Scaling Results of Starting Cohort 3 for Grade 9

Theresa Rohm, Micha Freund, Timo Gnambs, and Luise Fischer
Leibniz Institute for Educational Trajectories, Bamberg, Germany

E-mail address of lead author:

theresa.rohm@lifbi.de

Bibliographic data:

Rohm, T., Freund, M., Gnambs, T., & Fischer, L. (2017). *NEPS Technical Report for Listening Comprehension: Scaling Results of Starting Cohort 3 for Grade 9* (NEPS Survey Paper No. 21). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Acknowledgements:

We thank Anna Scharl for her assistance in scaling the data.

NEPS Technical Report for Listening Comprehension: Scaling Results of Starting Cohort 3 for Grade 9

Abstract

The National Educational Panel Study (NEPS) investigates the development of competencies from early childhood to late adulthood. Therefore, tests for the assessment of different competence domains are developed. To evaluate the quality of these tests, various analyses based on item response theory (IRT) are performed. This report describes the data and scaling procedures for the listening comprehension test in Starting Cohort 3 (fifth grade) for Grade 9. The listening comprehension test contained 16 items with complex multiple choice response formats that asked respondents about details on two spoken texts. The test was administered to 4,588 students. Their responses were scaled using the partial credit model. Item fit statistics, differential item functioning, Rasch-homogeneity, the tests' 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. Furthermore, test fairness could be confirmed for different subgroups. There was a negligible amount of missing responses; particularly, items that were not reached by the respondents were rare. Challenges of the test included the large number of items targeted toward a lower ability in listening comprehension. Further challenges arose from dimensionality analyses based on different cognitive requirements for the items. Overall, the listening comprehension test had acceptable psychometric properties that supported the estimation of reliable listening comprehension scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the ConQuest syntax for scaling the data.

Keywords

item response theory, scaling, listening comprehension, 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 (2016).

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 listening comprehension in Starting Cohort 3 (fifth grade) in Grade 9. First, the main concepts of the listening comprehension test are introduced. Then, the listening comprehension data of Starting Cohort 3 and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the scientific use file (SUF) is presented.

Please note that the analyses 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 SUF may differ slightly from the data used for the analyses in this paper. However, fundamentally different results are not expected.

2. Testing Listening Comprehension

The framework and test development for the listening comprehension test is described in Berendes, Weinert, Zimmermann, and Artelt (2013) and Hecker, Südkamp, Leser, and Weinert (2015). In the following, specific aspects of the listening comprehension test will be pointed out that are necessary for understanding the presented scaling results presented in this paper.

The administered listening comprehension test included two texts and two sets of eight items each referring to these texts. One text was a non-literary, informal text (e.g., resembling a conversation between people), whereas the other one was a literary, formal text (e.g., resembling a narration; Hecker et al., 2015). Furthermore, the test assessed three cognitive requirements. These were a) literal comprehension of explicit statements, b) text-related reasoning (drawing inferences), and c) comprehension of implicit meanings and statements (reflection and evaluation). The cognitive requirements did not depend on the two texts, but each cognitive requirement was assessed within each text (see Hecker et al., 2015 for a detailed description of the framework). Tables 1 and 2 summarize the number of items per text type and cognitive requirement.

All items of the testlet were prerecorded and presented on a compact disc (CD). Therefore, the time available to respond to the items of the listening comprehension test was set by the duration of the CD tracks. While the listening comprehension texts were presented once to

the test takers, the subsequent items (to be answered with either “true” or “not true”) were each repeated once. Thus, the test takers were able to check and revise their responses. Altogether 28 minutes were scheduled for the administration of the listening comprehension test.

Table 1

Number of Items for the Different Text Types

Text	**Text types**	**Number of items**
Text 1	Non-literary & informal	8
Text 2	Literary & formal	8
Total number of items		16

Table 2

Number of Items for the Different Cognitive Requirements

Cognitive requirements	**Text 1**	**Text 2**
Literal comprehension of explicit statements	4	4
Text-related reasoning (drawing inferences)	2	1
Comprehension of implicit meanings and statements (reflection and evaluation)	2	3
Total number of items	8	8

The listening comprehension test consists only of complex multiple choice (CMC) items. In CMC items, a number of subtasks with two response options are presented. Examples of the different response formats are given in Pohl and Carstensen (2012) and in Hecker and colleagues (2015). The competence test for listening comprehension that was administered in the present study included 16 items. To evaluate the quality of these items, extensive preliminary analyses were conducted. These preliminary analyses identified a poor fit for one subtask of the item lig9016s_c. Therefore, this subtask was not included in the analyses.

3. Data

3.1 The Design of the Study

The study assessed different competence domains including reading competence, declarative meta-cognition, domain general cognitive functioning, and listening

comprehension. The listening comprehension test was always administered as the last test within the test battery. There was no multi-matrix design regarding the order of the items within the test. All subjects received the test items in the same order.

3.2 Sample

A total of 4,599 [1] individuals received the listening competence test. For eleven subjects, less than three valid item responses were available. Because no reliable ability scores can be estimated based on such few valid responses, these cases were excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 4,588 individuals. All respondents of the listening comprehension test were tested in the institutional context (i.e., school).

3.3 Missing Responses

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

Invalid responses occurred, for example, when two response options were selected although only one was required and the chosen response option could, therefore, not be identified. Omitted items occurred when test persons skipped items. As the listening comprehension texts as well as the items were presented on CD a uniform testing time was given. Therefore, items that were not reached by the test takers were expected to be rare. Nevertheless, all missing responses after the last valid response were coded as not-reached. Because all 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. Items were coded as missing if at least one subtask contained a missing response. When just one kind of missing response occurred, the item was coded according to the corresponding missing response. When 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., understanding of instructions). They also need to be accounted for in the estimation of item and person parameters. Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the students were coping with the test. Missing responses per item were examined to evaluate how well each of the items functioned.

3.4 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 items consisted of a set of subtasks that were aggregated to a polytomous variable for each CMC item, indicating the number of correctly responded subtasks within that item. If at least one of the subtasks contained a missing response, the CMC item was scored as missing.

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

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 the items. In these cases, the lower categories were collapsed into one category. As can be seen in Appendix A, categories were collapsed for 15 of the 16 items during the PCM analyses. However, the values of all polytomously scored CMC items in the SUF indicate the number of correctly solved subtasks.

To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats), except for one item (lig9025s_c) for which a scoring of 1 point for each category was used.

Ability estimates for listening comprehension were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989) and will later also be provided in form of plausible values (Mislevy, 1991). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in section 7.

3.5 Checking the Quality of the Test

The listening comprehension 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.

Before aggregating the subtasks of CMC items to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed 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 score, and the item characteristic curves. Only if the subtasks exhibited a satisfactory item fit, they were used to generate polytomous variables that were included in the final scaling model.

After aggregating the subtasks to polytomous variables, their fit 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 total score (equal to the discrimination as computed in ConQuest) 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 listening comprehension test should measure the same construct for all students. If the items favored certain subgroups (e.g., they are easier for male than for female students), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., male and female students) would be biased and, thus, unfair. For the present study, test fairness was investigated for the variables gender, school type, the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012, for a description of these variables). Differential item functioning (DIF) analyses was estimated 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 noteworthy of further investigation, differences between 0.4 and 0.6 as considerable 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 listening comprehension 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 two different subdimensions representing the two texts, and, second, a model with three different subdimensions based on the three cognitive requirements 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.

Since the listening comprehension test consisted of item sets that referred to one of two texts, the assumption of local item dependence (LID) may not necessarily hold. However, the two texts were perfectly confounded with the two text functions. Thus, multidimensionality and local item dependence cannot be evaluated separately with these data.

3.6 Software

The IRT models were estimated in ConQuest version 4.2.5 (Adams, Wu, & Wilson, 2015).

4. Results

4.1 Missing Responses

4.1.1 Missing responses per person

Figure 1 shows the number of invalid responses per person. Overall, there were very few invalid responses. Ninety-six percent of the respondents did not have any invalid response at all; less than one percent had more than one invalid response.

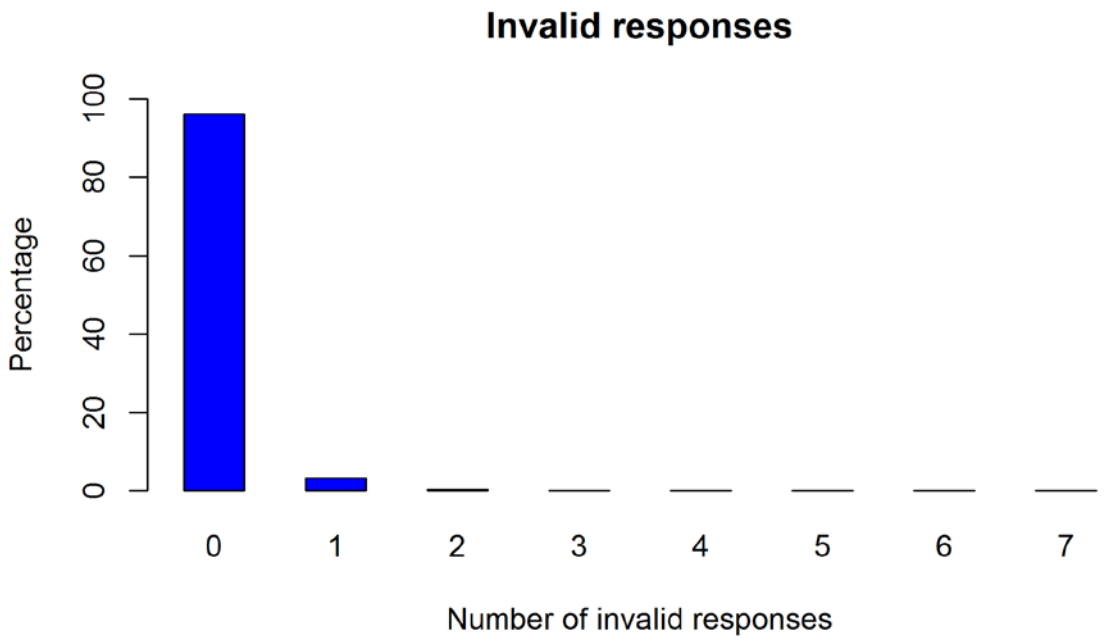

Figure 1. Number of invalid responses.

Missing responses may also occur when respondents omit items. As illustrated in Figure 2 most respondents, 90 percent, did not skip any item and less than one percent omitted more than three items.

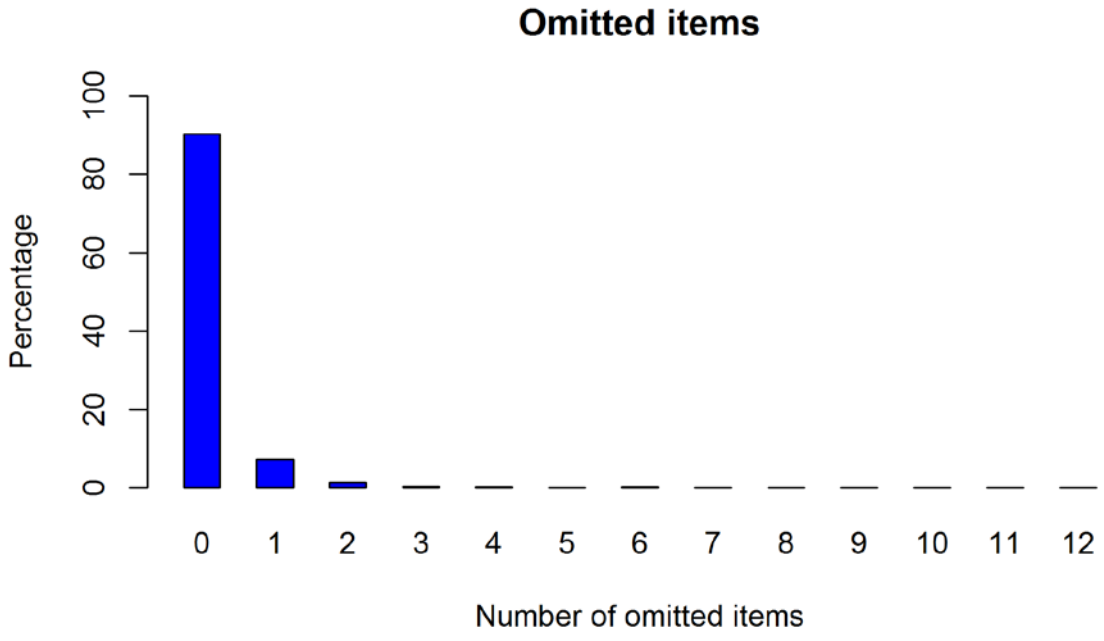

Figure 2. Number of omitted items.

Another source of missing responses were items that were not reached by the respondents; these are all missing responses after the last valid response. The number of not-reached items (Figure 3) was low because the subtasks for each item were repeated once (Hecker et al., 2015). More than 99% of the respondents finished the entire test.

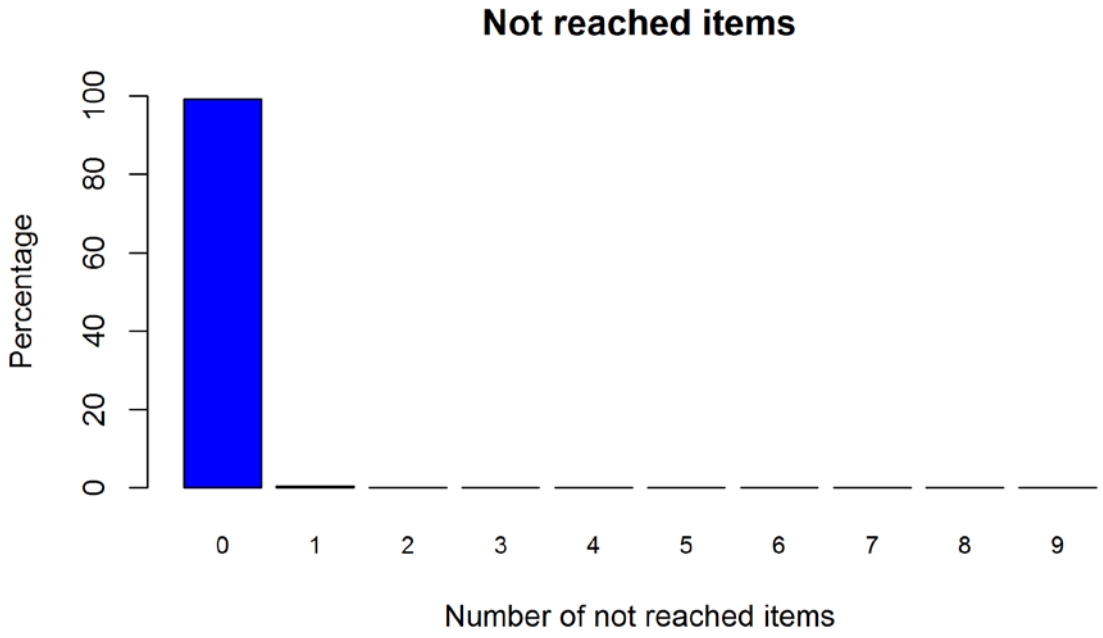

Figure 3. Number of not-reached items.

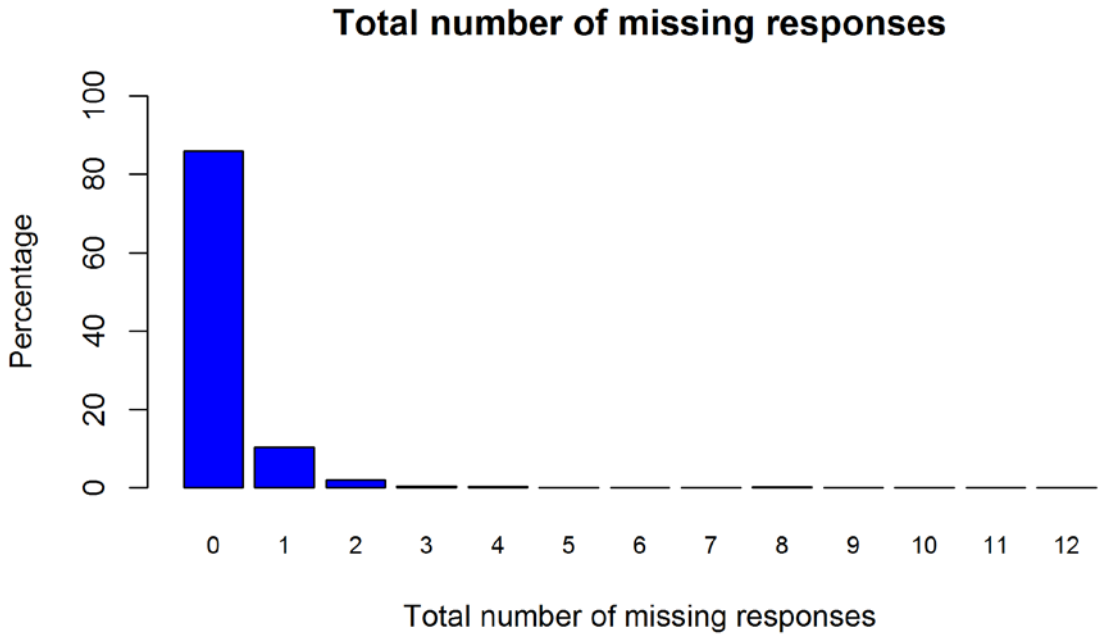

Figure 4. Total number of missing responses.

The aggregated polytomous variables were coded as not-determinable missing response when the subtasks of CMC items contained different kinds of missing responses. However, only a rather small number of not-determinable missing responses occurred. Most respondents, 98.52%, did not have any not-determinable missing response.

The total number of missing responses, aggregated over invalid, omitted, not-reached, and not-determinable missing responses per person, is illustrated in Figure 4. On average, the respondents showed *M* = 0.23 (*SD* = 0.84) missing responses. About 86% of the respondents had no missing response at all and about 1.1% of the participants had four or more missing responses.

In sum, the amount of invalid, not-reached and not-determinable missing responses was very small, whereas a reasonable part of missing responses occurred due to omitted items.

4.1.2 Missing responses per item

Table 3 provides information on the occurrence of different kinds of missing responses per item.

Table 3

Percentage of Missing Values for the Test

Item	Position	*N*	NR	OM	NV
lig9011s_c	1	4,532	0.00	0.92	0.31
lig9012s_c	2	4,539	0.00	0.72	0.35
lig9013s_c	3	4,504	0.00	1.42	0.37
lig9014s_c	4	4,541	0.00	0.92	0.11
lig9015s_c	5	4,511	0.00	1.33	0.35
lig9016s_c	6	4,526	0.00	0.00	0.00
lig9017s_c	7	4,512	0.00	1.31	0.33
lig9018s_c	8	4,520	0.04	0.98	0.41
lig9021s_c	9	4,505	0.09	1.20	0.50
lig9022s_c	10	4,530	0.13	0.76	0.37
lig9023s_c	11	4,526	0.15	0.81	0.39
lig9024s_c	12	4,535	0.17	0.78	0.17
lig9025s_c	13	4,546	0.20	0.59	0.13
lig9026s_c	14	4,506	0.22	1.20	0.35
lig9027s_c	15	4,493	0.26	1.35	0.44
lig9028s_c	16	4,507	0.87	0.61	0.28

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

Overall, the omission rates were low, varying across items between 0.00% and 1.42%. The omission rates correlated with the item difficulties at about *r* = .35. In general, participants were inclined to omit more difficult items. The percentage of invalid responses per item (column 6 in Table 3) was also low with the maximum rate being 0.5 percent. With an item's progressing position in the test, the amount of persons that did not reach the item (column 4 in Table 3) only rose up to less than 1%, which is also depicted in Figure 5.

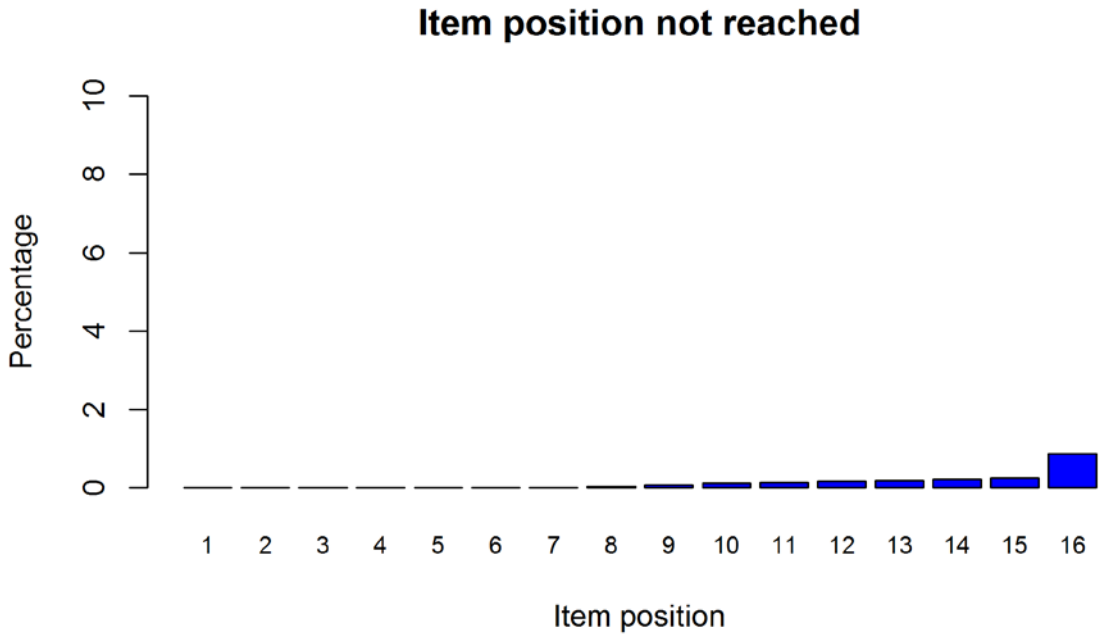

Figure 5. Item position not reached.

4.2 Parameter Estimates

4.2.1 Item parameters

The second column in Table 4 presents the percentage of correct responses in relation to all valid responses for each item. Depicted are the percentages of correct responses on all subtasks of each CMC item. They varied between 26.22% and 83.04%, with an average of 46.40% (*SD* = 15.30%).

The location parameters for the polytomous items are given in Table 4 and the respective step parameters are depicted in Table 5. The item difficulties were estimated by constraining the mean of the ability distribution to be zero. The estimated location parameters for polytomous variables ranged from -2.65 (item lig9028s_c) to -0.19 (item lig9021s_c) with an average difficulty of -1.69 (*SD* = 0.66). Overall, the item difficulties were rather low; there were no items with a high difficulty. Due to the large sample size the standard errors (*SE*) of the estimated item difficulties (column 4 in Table 4) were rather small (all *SE*s ≤ 0.07).

Table 4

Item Parameters

	Item	Percentage correct	Difficulty	*SE*	WMNSQ	*t*	Item-Rest Correlation	Discr.
1.	lig9011s_c	34.33	-0.659	0.050	1.02	1.3	0.34	1.24
2.	lig9012s_c	37.12	-1.451	0.049	0.99	-0.4	0.45	1.50
3.	lig9013s_c	26.22	-1.092	0.044	1.04	2.1	0.48	1.33
4.	lig9014s_c	55.76	-2.207	0.061	0.90	-5.3	0.48	2.24
5.	lig9015s_c	47.26	-2.116	0.057	1.07	3.1	0.34	1.06
6.	lig9016s_c	32.35	-1.997	0.075	0.98	-0.9	0.33	1.51
7.	lig9017s_c	35.26	-1.612	0.053	0.98	-0.9	0.45	1.45
8.	lig9018s_c	45.53	-1.332	0.053	0.96	-2.0	0.43	1.64
9.	lig9021s_c	27.28	-0.194	0.039	1.16	9.0	0.34	0.80
10.	lig9022s_c	38.87	-1.721	0.053	1.00	-0.2	0.42	1.37
11.	lig9023s_c	55.50	-2.620	0.070	1.02	1.0	0.30	1.16
12.	lig9024s_c	47.61	-1.572	0.055	0.97	-2.0	0.41	1.63
13.	lig9025s_c	83.04	-2.113	0.050	0.96	-1.4	0.43	1.64
14.	lig9026s_c	52.44	-1.957	0.052	0.98	-0.8	0.47	1.47
15.	lig9027s_c	55.04	-1.693	0.046	0.98	-0.9	0.50	1.63
16.	lig9028s_c	68.78	-2.654	0.064	0.98	-0.9	0.37	1.52

Note. Difficulty = Item difficulty / location parameter, *SE* = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, *t* = *t*-value for WMNSQ, Discr. = Discrimination parameter of a generalized partial credit model. The Item-rest correlation corresponds to the product-moment correlation between an item and the total-rest score (discrimination value as computed in ConQuest).

Table 5

Step Parameters (with Standard Errors) for Polytomous Items

Item	Step 1 (*SE*)	Step 2 (*SE*)	Step 3 (*SE*)	Step 4 (*SE*)
lig9011s_c	-0.766 (0.031)	0.766		
lig9012s_c	-0.651 (0.052)	-0.269 (0.047)	0.919	
lig9013s_c	-1.053 (0.061)	-0.132 (0.053)	0.163 (0.045)	1.022
lig9014s_c	-0.658 (0.037)	0.658		
lig9015s_c	-0.389 (0.066)	-0.533 (0.058)	0.923	
lig9016s_c	-1.757 (0.040)	1.757		
lig9017s_c	-0.936 (0.055)	-0.180 (0.047)	1.116	
lig9018s_c	-0.601 (0.033)	0.601		
lig9021s_c	-0.480 (0.038)	0.189 (0.044)	0.291	
lig9022s_c	-0.523 (0.059)	-0.559 (0.052)	1.081	
lig9023s_c	-0.948 (0.040)	0.948		
lig9024s_c	-0.662 (0.034)	0.662		
lig9026s_c	-0.669 (0.055)	0.366 (0.052)	0.303	
lig9027s_c	-0.208 (0.052)	0.225 (0.054)	-0.016	
lig9028s_c	-0.198 (0.042)	0.198		

Note. Because item lig9025s_c consists of only two categories, no step parameters are estimated.

4.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person abilities to evaluate the appropriateness of the test for the specific target population. In Figure 6, item difficulties of the listening comprehension items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of item difficulties. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 1.935, which implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = .790, WLE reliability = .764) was good. Although the items covered some range of the ability distribution, the items were too easy. Consequently, person ability in medium- and low-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors of measurement.

Scale (in logits)	Person ability	Item difficulty
4		
	X	
	X	
3	X	
	X	
	XX	
	XX	
	XXX	
	XX	
2	XXXXX	
	XXXX	
	XXXXX	
	XXXXX	
	XXXXXX	
1	XXXXXXXX	
	XXXXXXX	
	XXXXXXXXX	
	XXXXXXXXX	
	XXXXXXXXXX	
	XXXXXXXXX	
0	XXXXXXXX	
	XXXXXXXX	9
	XXXXXXXXX	
	XXXXXXX	1
	XXXXXXX	
-1	XXXXXXX	3
	XXXXXX	
	XXXXX	2 8
	XXXXX	7 12
	XXXXX	10 15
-2	XXXX	6 14
	XXXX	4 5 13
	XXX	
	XX	
	X	11 16
	X	
-3	X	
	X	
-4		

Figure 6. Test targeting. The distribution of person ability in the sample is depicted on the left-hand side of the graph, with each 'X' representing 26.5 cases. The difficulty of the items is depicted on the right-hand side of the graph, with each number representing one item (corresponding to Table 4).

4.3 Quality of the test

4.3.1 Fit of the subtasks of complex multiple choice items

Before the subtasks of CMC items were aggregated and analyzed via a partial credit model, the fit of the subtasks was checked by analyzing the single subtasks in a Rasch model. Counting the subtasks of CMC items separately, there were 61 items. The probability of a correct response ranged from 41.25% to 96.16% across all items (*Mdn* = 82.28%). Thus, the number of correct and incorrect responses was sufficient. Nearly all subtasks showed a satisfactory item fit. Overall, the WMNSQ ranged from 0.87 to 1.12, the respective *t*-value from -7.2 to 11.2, and there were no noticeable deviations of the empirical estimated probabilities from the model-implied item characteristic curves. Due to misfit, item lig90161_c (WMNSQ = 1.12, *t*-value = 9.6) was excluded from CMC item computation and from the subsequent PCM analysis. Due to the good model fit of the remaining subtasks, their aggregation to polytomous variables seemed to be justified.

4.3.2 Item fit

The evaluation of the item fit was performed on the basis of the final scaling model, the partial credit model, using all polytomous CMC items. Altogether, item fit can be considered to be satisfactory (see Table 4). Values of the WMNSQ ranged from 0.90 (lig9014s_c) to 1.16 (lig9021s_c). Only one item exhibited a *t*-value of the WMNSQ greater than 6 (lig9021s_c with a *t*-value of 9). Thus, there is no indication of severe item over- or underfit. Item-rest correlations between each item and the total rest scores ranged from .30 (lig9023s_c) to .50 (lig9027s_c) and had a mean of .41. All item characteristic curves showed a good fit of the items.

4.3.3 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 gender, the number of books at home (as a proxy for socioeconomic status), migration background, and school type (see Pohl & Carstensen, 2012, for a description of these variables). The differences between the estimated item difficulties in the various groups are summarized in Table 6. For example, the column “male vs. female” reports the differences in item difficulties between men and women; a positive value for an item indicates that an item was more difficult for females, whereas a negative value highlights a lower difficulty for females as opposed to males. The negative value for the main effect indicates that men overall performed worse in the test as compared to women. 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 7).

Table 6

Differential Item Functioning

Item	Gender	Books	Migration	School
	male vs. female	< 100 vs. ≥ 100	without vs. with	no sec. vs. sec.
lig9011s_c	-0.140	-0.110	0.158	0.050
	(-0.101)	(-0.085)	(0.115)	(0.042)
lig9012s_c	0.134	0.140	-0.188	0.124
	(0.097)	(0.108)	(-0.136)	(0.103)
lig9013s_c	-0.386	-0.004	0.036	0.164
	(-0.279)	(-0.003)	(0.026)	(0.137)
lig9014s_c	-0.380	0.290	-0.418	0.462
	(-0.275)	(0.225)	(-0.303)	(0.385)
lig9015s_c	-0.144	-0.092	-0.084	-0.078
	(-0.104)	(-0.071)	(-0.061)	(-0.065)
lig9016s_c	-0.598	-0.146	0.314	-0.010
	(-0.433)	(-0.113)	(0.228)	(-0.008)
lig9017s_c	0.340	0.032	-0.112	0.166
	(0.246)	(0.025)	(-0.081)	(0.138)
lig9018s_c	0.032	0.226	0.488	0.266
	(0.023)	(0.175)	(0.354)	(0.222)
lig9021s_c	-0.098	-0.168	0.492	-0.406
	(-0.071)	(-0.130)	(0.357)	(-0.338)
lig9022s_c	0.234	0.144	-0.840	-0.254
	(0.169)	(0.112)	(-0.610)	(-0.212)
lig9023s_c	0.112	-0.280	0.226	-0.358
	(0.081)	(-0.217)	(0.164)	(-0.298)
lig9024s_c	0.090	0.074	-0.100	0.054
	(0.065)	(0.057)	(-0.073)	(0.045)
lig9025s_c	0.004	0.124	0.156	0.040
	(0.003)	(0.096)	(0.113)	(0.033)
lig9026s_c	0.384	-0.118	-0.206	-0.102
	(0.278)	(-0.091)	(-0.150)	(-0.085)

Item	Gender	Books	Migration	School
lig9027s_c	0.056 (0.041)	0.046 (0.036)	0.028 (0.020)	0.216 (0.180)
lig9028s_c	0.348 (0.252)	-0.138 (-0.107)	0.194 (0.141)	-0.242 (-0.202)
Main effect (DIF model)	-0.386 (-0.279)	-1.032 (-0.800)	0.704 (0.511)	-1.420 (-1.183)
Main effect (main effect model)	-0.384 (-0.279)	-1.032 (-0.800)	0.710 (0.516)	-1.416 (-1.182)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. Sec. = Secondary school (German: „Gymnasium").
None of the absolute standardized differences was significantly, $p < .05$, greater than 0.25 (see Fischer, Rohm, Gnambs, & Carstensen, 2016).

Gender: The sample included 2,353 (51%) males and 2,227 (49%) females. Eight respondents that did not indicate their gender were excluded from the analysis. On average, male participants had a lower estimated listening ability than females (main effect = -0.386 logits, Cohen's *d* = -0.279). No item showed DIF greater than 0.6 logits. An overall test for DIF (see Table 7) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). Both the model comparison using Akaike's (1974) information criterion (AIC) and the Bayesian information criterion (BIC; Schwarz, 1978), that takes the number of estimated parameters into account (and, thus, guards against overparameterization of models) favored the model estimating DIF. Nevertheless, regarding the specific differences in item difficulties, no pronounced DIF for gender could be identified.

Table 7

Comparisons of Models with and without DIF

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Gender	main effect	4,580	142,392.42	42	142,476.42	142,746.46
	DIF	4,580	142,231.51	58	142,347.51	142,720.42
Books	main effect	4,471	138,320.83	42	138,404.83	138,673.86
	DIF	4,471	138,276.53	58	138,392.53	138,764.04
Migration	main effect	3,137	96,105.53	42	96,189.53	96,443.67
	DIF	3,137	96,064.71	58	96,180.71	96,531.67
School	main effect	4,588	141,701.63	42	141,785.63	142,055.74
	DIF	4,588	141,586.26	58	141,702.26	142,075.27

Books: The number of books at home was used as a proxy for socioeconomic status. There were 1,747 (38%) test takers with 0 to 100 books at home, 2,724 (59%) test takers with more than 100 books at home, and 117 (3%) 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 1.032 logits (Cohen's *d* = -0.800) lower in listening comprehension than participants with more than 100 books. There was no considerable DIF comparing participants with many or fewer books (highest DIF = 0.290 for item lig9014s_c). The model comparison criterions were ambiguous: the AIC favored the DIF-model while the BIC favored the main effect model (see Table 7).

Migration background: There were 2,961 participants (65%) with no migration background, 1,356 students (30%) with a migration background, and 271 respondents (6%) for whom no information on their migration background was available. In comparison to students with migration background, participants without migration background had on average a higher ability in listening comprehension (main effect = 0.704 logits, Cohen's *d* = 0.511). There was one noteworthy item with DIF due to migration background with a difference in estimated difficulty exceeding 0.6 logits (item lig9022s_c with a DIF of -0.840 logits). Regarding overall model comparison, the AIC favored the DIF-model while the BIC favored the main effects model that does not include item-level DIF.

School type: Overall, 2,123 subjects (46%) who took the listening comprehension test attended secondary school (in German: "Gymnasium") whereas 2,465 (54%) were enrolled in other school types. Students attending secondary school showed, on average, a higher ability in listening comprehension (-1.420 logits, Cohen's *d* = -1.183) compared to students from other school types. There was no noteworthy DIF, as no item exhibited differences in estimated item difficulties greater than 0.6 logits. As before, the AIC indicated a better model fit for the DIF-model while the BIC exhibited a lower value for the main effect model.

4.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item-discrimination parameters are equal. To test this assumption, a generalized partial credit model (GPCM) that estimates different discrimination parameters was fitted to the data. The estimated discriminations differed moderately among items (see Table 4, last column), ranging from 0.8 (lig9021s_c) to 2.24 (lig9014s_c). The average discrimination parameter was 1.45. Model fit indices suggested a slightly better fit of the GPCM (AIC = 142,460.26, BIC = 142,820.41) as compared to the PCM model (AIC = 142,836.28, BIC = 143,099.96). 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 weighting as intended in the theoretical framework.

4.3.5 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, the two texts were used as dimensions, whereas the three different cognitive requirements were subject of the second multidimensional model.

Estimation of the models was carried out in ConQuest using Gauss-Hermite quadrature method.

The estimated variances and correlations between the two dimensions representing the texts are reported in Table 8. Both dimensions had substantial variances, with the highest obtained for the first text. The correlation among the two dimensions was rather high (r = .871). However, it deviated from a perfect correlation (i.e., it was slightly lower than r = .95, see Carstensen, 2013). Moreover, according to model fit indices, the two-dimensional model fitted the data slightly better (AIC = 142,695.308, BIC = 142,971.85, number of parameters = 43) than the unidimensional model (AIC = 142,836.28, BIC = 143,099.96, number of parameters = 41). This could also be a result of the large sample size.

Table 8

Results of Two-Dimensional Scaling

	Dim 1	**Dim 2**
Text 1 (Dim 1)	(2.32)	
(8 items)		
Text 2 (Dim 2)	0.871	(2.03)
(8 items)		

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

Furthermore, as each text function corresponded to only one of the two texts, local item dependence (LID) and the text functions were confounded. As a consequence, the deviation of the correlations from a perfect correlation shown in Table 8 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 conclusion, as the listening comprehension test is constructed to measure a single dimension, the assumption that the two texts measure a common construct is justifiable, although the results indicate not a completely unidimensional construct.

The estimated variances and correlations of the three-dimensional model based on the three different cognitive requirements are presented in Table 9. Correlations between the dimensions varied between r = .91 and r = .94. All correlations deviated from a perfect correlation (i.e., they were lower than r = .95, see Carstensen, 2013). Moreover, the three-dimensional model (AIC = 142,647.61, BIC = 142,943.445, number of parameters = 46) fitted the data slightly better than the unidimensional model (AIC = 142,836.28, BIC = 143,099.96, number of parameters = 41).

Table 9

Results of Three-Dimensional Scaling

	Dim 1	Dim 2	Dim 3
Literal comprehension of explicit statements (Dim 1)	(1.58)		
(8 items)			
Text-related reasoning (Dim 2)	.94	(3.74)	
(3 items)			
Comprehension of implicit meanings and statements (Dim 3)	.92	.91	(2.60)
(5 items)			

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

5. Discussion

The analyses in the previous sections aimed at providing detailed information on the quality of the listening comprehension test in Starting Cohort 3 for Grade 9 and at describing how the listening comprehension score was estimated.

We investigated different kinds of missing responses and examined the item and test parameters. We thoroughly checked item fit statistics for the subtasks of CMC items, as well as for the aggregated polytomous CMC items, and examined the correlations between correct and incorrect responses and the total score. Further quality inspections were conducted by examining differential item functioning, testing Rasch-homogeneity, investigating the tests' dimensionality as well as local item dependence.

Various criteria indicated a good fit of the items and measurement invariance across various subgroups. All types of missing responses were reasonably small, which is an advantage of the item presentation via CD. Furthermore, the test had a high reliability and distinguished well between students. However, the test was mainly targeted at low-performing students and did not accurately measure listening competence of high-performing students. Consequently, ability estimates will be precise for low-performing students but less precise for high performing students.

Some degree of multidimensionality was present for different text functions and cognitive requirements. Thus, the estimation of a single listening comprehension score is slightly challenged. This needs to be addressed in further studies. Nevertheless, Hecker and colleagues (2015) argue that a balanced assessment of listening competence can only be achieved by heterogeneous texts addressing different cognitive requirements. Based on a literacy conception and a functional-integrative perspective on literacy competence they provide arguments for a unidimensional measure of listening competence.

Summarizing these results, the test had acceptable psychometric properties that facilitated the estimation of a unidimensional listening competence score.

6. Data in the Scientific Use File

The data in the SUF contain 16 items. All 16 items were scored as polytomous variables (CMC items). The variable names of CMC items end in 's_c' and the values of the polytomous variables in the SUF correspond to the number of correctly responded subtasks. In the IRT scaling model categories were collapsed (cf. Section 3.4 for a description of the aggregation of CMC items). Except for one item (lig9025s_c), all polytomous CMC variables were scored as 0.5 for each category. The item lig9025s_c was an exception because it had only two possible scores (0 and 1) after categories were collapsed.

In the SUF, a unidimensional listening comprehension score is provided. Manifest listening competence scores are available as WLEs (lig9_sc1u) together with their corresponding standard errors (lig9_sc2u). The ConQuest Syntax for estimating the WLE scores from the items is given in Appendix A. For persons who either did not take part in the listening test or who did not give enough valid responses, no WLEs were estimated. These WLEs and the respective standard errors are denoted as not-determinable missing values.

Plausible values that allow for an investigation of latent relationships of competence scores with other variables will be provided in future data releases. Alternatively, users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values themselves. A description of these approaches can be found in Pohl and Carstensen (2012).

References

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-722. doi:10.1109/TAC.1974.1100705

Berendes, K., Weinert, S., Zimmermann, S., & Artelt, C. (2013). Assessing language indicators across the lifespan within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online*, *5*(2), 15.

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.

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

Hecker, K., Südkamp, A., Leser, C. & Weinert, S. (2015). *Entwicklung eines Tests zur Erfassung von Hörverstehen auf Textebene bei Schülerinnen und Schülern der Klassenstufe 9* (NEPS Working Paper No. 53). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

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

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196. doi:10.1007/BF02294457

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

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

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen & Lydiche.

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450. doi: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. doi:10.1007/s11618-011-0182-7

Appendix

Appendix A: ConQuest-Syntax for estimating WLE estimates in Starting Cohort 3

```
Title Listening Comprehension Test (SC3, Grade 9): Partial Credit Model;

data filename.dat;

format responses 1-16;

/* collapse response categories with less than 200 responses */
recode (0,1,2,3,4)       (0,0,0,1,2)       ! item (1);   /* lig9011s_c */
recode (0,1,2,3,4)       (0,0,1,2,3)       ! item (2);   /* lig9012s_c */
recode (0,1,2,3)         (0,0,1,2)         ! item (4);   /* lig9014s_c */
recode (0,1,2,3,4)       (0,0,1,2,3)       ! item (5);   /* lig9015s_c */
recode (0,1,2,3,4)       (0,0,1,2,2)       ! item (6);   /* lig9016s_c */
recode (0,1,2,3,4)       (0,0,1,2,3)       ! item (7);   /* lig9017s_c */
recode (0,1,2,3,4)       (0,0,0,1,2)       ! item (8);   /* lig9018s_c */
recode (0,1,2,3,4)       (0,0,1,2,3)       ! item (9);   /* lig9021s_c */
recode (0,1,2,3,4)       (0,0,1,2,3)       ! item (10);  /* lig9022s_c */
recode (0,1,2,3,4)       (0,0,0,1,2)       ! item (11);  /* lig9023s_c */
recode (0,1,2,3,4)       (0,0,0,1,2)       ! item (12);  /* lig9024s_c */
recode (0,1,2,3)         (0,0,1,1)         ! item (13);  /* lig9025s_c */
recode (0,1,2,3,4)       (0,0,1,2,3)       ! item (14);  /* lig9026s_c */
recode (0,1,2,3,4)       (0,0,1,2,3)       ! item (15);  /* lig9027s_c */
recode (0,1,2,3)         (0,0,1,2)         ! item (16);  /* lig9028s_c */

/* scoring */
codes 0,1,2,3,4;

score (0,1)         (0,1)              ! items (13);

score (0,1,2)       (0,0.5,1)          ! items (1,4,6,8,11,12,16);

score (0,1,2,3)     (0,0.5,1,1.5)      ! items (2,5,7,9,10,14,15);

score (0,1,2,3,4)   (0,0.5,1,1.5,2)    ! items (3);


set constraint=cases;

model item + item*step;

estimate ! method=gauss, nodes=15, iterations=1000, convergence=0.0001,
stderr=empirical, fit=yes;


show ! estimate=latent    >> show.txt;

itanal                    >> itemanalysis.txt;

show cases ! estimate=wle >> wle.txt;
```