NEPS
National Educational Panel Study

NEPS Working Papers

Aileen Edele, Kristin Schotte, Martin Hecht & Petra Stanat

Listening comprehension tests of immigrant students' first language (L1) Russian and Turkish in grade 9: Scaling procedure and results

NEPS Working Paper No. 13

Bamberg, September 2012

SPONSORED BY THE
Federal Ministry of Education and Research

Working Papers of the German National Educational Panel Study (NEPS)
at the University of Bamberg

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).
The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS consortium.

The NEPS Working Papers are available at
http://www.uni-bamberg.de/neps/publikationen/neps-working-papers/

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

Listening comprehension tests of immigrant students' first languages (L1) Russian and Turkish in grade 9: Scaling procedure and results

Aileen Edele, Humboldt-Universität zu Berlin
Kristin Schotte, Freie Universität Berlin
Martin Hecht, Humboldt-Universität zu Berlin
Petra Stanat, Humboldt-Universität zu Berlin

E-Mail-Adresse der Erstautorin:

aileen.edele@iqb.hu-berlin.de

Bibliographische Angaben:

Edele, A., Schotte, K., Hecht, M. & Stanat, P. (2012). Listening comprehension tests of immigrant students' first languages (L1) Russian and Turkish in grade 9: Scaling procedure and results (NEPS Working Paper No. 13). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Listening comprehension tests of immigrant students' first languages (L1)[1] Russian and Turkish in grade 9: Scaling procedure and results

Abstract

Immigrant students' proficiency in their first languages Russian and Turkish is examined in three starting cohorts within the National Educational Panel Study (NEPS). This paper reports the scaling procedure and results for the L1-tests in starting cohort 4, grade 9. The tests developed for this purpose are described and the design of the study is delineated, including the criteria participants had to meet in order to participate in the L1-tests and a brief description of the samples. Subsequently, the results of analyses for the Russian and Turkish L1-test are presented. Preliminary analyses for each test address the exclusion of cases and several descriptive statistics such as relative frequencies of correct responses and of missing responses. With regard to scaling, the applied model is described and item parameters as well as item fit indices are presented. In addition, differential item functioning, distractor analyses, the distribution and reliability of person estimates are reported. Tests of unidimensionality and of item dependencies are also described. Overall, the L1-tests fit the Rasch model well and prove suitable for testing the proficiency of the target population in their first languages Russian and Turkish.

Keywords

first language proficiency, L1-proficiency, L1-tests, listening comprehension, scaling

[1]The term first language (L1) is used here interchangeably with the language of the country of origin, regardless of whether the language was actually acquired prior to German, as the labelling L1 suggests, or simultaneously.

1. Testing immigrant students' proficiency in their first languages Russian and Turkish in grade 9

The NEPS assesses a global indicator of immigrant students' proficiency in the languages Russian and Turkish, the L1 of the two largest immigrant groups in Germany. L1 is tested in three starting cohorts within the NEPS, that is, starting cohort 4 (grade 9), starting cohort 3 (grade 7 and expectedly grade 9), and starting cohort 2 (grade 2). To allow for assessing the L1-proficiency of students with low L1-proficiency in written language or a complete lack thereof, we decided to focus on listening comprehension as an indicator of L1-proficiency. As adequate instruments for this purpose were unavailable, tests in Russian and Turkish are newly developed within the Berlin project of pillar 4 of the NEPS. The current paper focuses on the L1-tests in 9th grade[2] (starting cohort 4), which is the starting cohort tested first in L1 within the NEPS.

The L1-tests in Russian and Turkish are construed analogously and consist of 7 independent text units (testlets). The texts include dialogues as well as expository and narrative content and have a length of 98 to 155 words (Russian L1-test) and 97 to 156 words (Turkish L1-test). The text units are followed by 3 to 6 multiple choice questions (items), totaling to 32 items with 4 or 5 response options each. Correct responses (attractors) are scored with 1, incorrect responses (distractors) with 0. The text units, questions and options were recorded with native speakers of Russian or Turkish, respectively, and presented to the students in the test session once. In total, the administration and completion of the tests took approximately 30 (Russian version) and 32 minutes (Turkish version).

2. Study design

The L1-tests were carried out in the 9th grade starting cohort of the NEPS (see von Maurice, Sixt & Blossfeld, 2011 and Frahm et al., 2011 for further details on the sampling, recruitment and administered instruments in this starting cohort) in spring 2011. The target population for testing L1 consists of students of the first, second, and third generation from families who have immigrated to Germany from the Former Soviet Union (e.g. Russia, Kazakhstan) or Turkey. More specifically, students who are either themselves born in one of these countries, and students with at least one parent or two grandparents born in the Former Soviet Union or Turkey were asked to participate in the L1-tests. Students were chosen for participation in the L1-testing based on their answers in a questionnaire administered in autumn 2010.

The L1-tests were administrated on a separate test day. In order to ensure a threshold proficiency in Russian or Turkish, students were asked to answer a screening test with 8 items with low item difficulty. The screening test items consisted in recordings of simple spoken sentences, such as "the dog walks", which students had to relate to the corresponding picture out of five options. Test administrators immediately scored the screening tests by means of a template. Only students who answered a minimum of 3 screening test items correctly attended the L1-tests.

[2] For further information on construction and validity of the L1-tests for grade 9 see Edele, Stanat and Schotte (forthcoming).

The Russian screening test was administered to 549 students of the 13,379 panel students participating in the NEPS study in spring 2011. On average, they answered 7.0 items correctly (*Mdn* = 8), 322 participants solved all 8 screening items. 35 participants answered less than 3 items correctly and therefore did not participate in the L1-testing. The Turkish screening test was completed by 704 students. On average, they answered 7.1 items correctly (*Mdn* = 8). While 451 students solved all 8 screening items, 38 participants solved less than 3 items and did not proceed to the L1-tests.

The Russian and the Turkish L1-tests were analyzed separately and the results are reported successively below. To the extent possible, analyses as well as criteria for the exclusion of cases and items comply with the NEPS standards of scaling competence data (see Pohl & Carstensen, 2012).

3. The Russian L1-test

3.1 Preliminary analyses

3.1.1 Exclusion of cases from the analyses

Of the 514 cases, 12 were excluded from further analyses because less than 3 valid answers were available. Thus, the analyses are based on N = 502 cases.

3.1.2 Sample

The scaling of the Russian L1-test is based on data from 254 (50.6%) female and 248 (49.4%) male students. 206 (41%) of the participants attended the *Hauptschule*, 124 (24.7%) the *Realschule,* 47 (9.4%) the *Gesamtschule* and 98 (19.5%) the *Gymnasium*. In addition, 27 (5.4%) students attended schools with several educational tracks existing in some of the Federal States. On the questions pertaining to students' immigrant generation status, 497 valid responses were available. Almost half of the students (N = 230 or 46.3%) were born in Germany, whereas 267 (53.7%) participants were born abroad.

3.1.3 Descriptive statistics of responses and missing responses

On average, there were 485 valid answers per item, while the mean of omitted items was 3.3% and the mean of invalid answers 0.1% (see table 1). Students on average answered 30.9 items (*Mdn* = 32) or 96.6% of items validly; the percentage of valid answers per person ranged from 25% to 100%. None of the items was extremely easy or difficult. The most difficult item was solved by 24.3% of the students, while the easiest item was solved by 77.1% of the participants. Participants on average answered 52.4% of the Russian L1-Test correctly.

Table 1: Descriptives and missing responses in the Russian L1-Test

Item	Position in the test	Testlet	Valid responses	Omitted responses (in %)	Invalid responses (in %)
nrg90101_c	1	text 1	490	2.39	0.00
nrg90102_c	2	text 1	492	1.99	0.00
nrg90103_c	3	text 1	483	3.78	0.00
nrg90201_c	4	text 2	493	1.79	0.00
nrg90202_c	5	text 2	492	1.99	0.00
nrg90203_c	6	text 2	494	1.39	0.20
nrg90301_c	7	text 3	487	2.99	0.00
nrg90302_c	8	text 3	482	3.98	0.00
nrg90303_c	9	text 3	476	5.18	0.00
nrg90304_c	10	text 3	485	3.39	0.00
nrg90401_c	11	text 4	494	1.59	0.00
nrg90402_c	12	text 4	479	4.58	0.00
nrg90403_c	13	text 4	472	5.98	0.00
nrg90404_c	14	text 4	483	3.78	0.00
nrg90405_c	15	text 4	481	3.98	0.20
nrg90501_c	16	text 5	491	1.99	0.20
nrg90502_c	17	text 5	490	2.19	0.20
nrg90503_c	18	text 5	484	3.59	0.00
nrg90504_c	19	text 5	486	2.99	0.20
nrg90505_c	20	text 5	486	3.19	0.00
nrg90506_c	21	text 5	489	2.59	0.00
nrg90601_c	22	text 6	488	2.79	0.00
nrg90602_c	23	text 6	479	4.58	0.00

nrg90603_c	24	text 6	483	3.59	0.20
nrg90604_c	25	text 6	474	5.58	0.00
nrg90605_c	26	text 6	471	6.18	0.00
nrg90701_c	27	text 7	494	1.39	0.20
nrg90702_c	28	text 7	491	1.99	0.20
nrg90703_c	29	text 7	489	2.59	0.00
nrg90704_c	30	text 7	487	2.79	0.20
nrg90705_c	31	text 7	483	3.78	0.00
nrg90706_c	32	text 7	474	5.58	0.00
Mean			485	3.32	0.06

Note. Valid responses are reported in terms of absolute frequencies; omitted responses as well as invalid responses are provided as relative frequencies (percent).

3.2 Results of scaling

3.2.1 Analyses

Data preparation and subsequent processing of results were executed with Stata/SE 11.0 (Stata Statistical Software: Release 11) and R 2.15.1 (R Core Team, 2012). Correct answers were scored with 1, incorrect answers as well as omitted items and invalid answers were coded as 0. In line with the NEPS standards of competence scaling (Pohl & Carstensen, 2012), a one-parameter logistic model (Rasch model) was applied to the data in ConQuest 2.0 (Wu, Adams, Wilson & Haldane, 2007) using Marginal Maximum Likelihood (MML) estimation with Gauss-Hermite quadrature and 15 nodes. Both convergence criteria (deviance change and parameter change) were set to 0.0001. The person distribution was assumed to be normal with the mean constrained to 0. Standard errors of item parameters were calculated with ConQuest's "quick" method. The syntax of the scaling analyses is provided in the Appendix.

3.2.2 Results

Item parameters and item fit

In line with the NEPS standards for scaling competence data (Pohl & Carstensen, 2012), items with a strong misfit defined as a weighted mean square (WMNSQ) > 1.20 or $t_{WMNSQ} > 8$ were excluded from further analyses. Item nrg90501_c showed such a misfit (WMNSQ = 1.25, t = 6.4). The item fit of the remaining 31 items was .88 ≤ WMNSQ ≤ 1.15 (see table 2). The mean item difficulty for the Russian L1-Test was b = -0.12 indicating that item difficulty and L1-proficiency of participants match well, on average. The range of item difficulty was -1.48 ≤ b ≤ 1.41. The average point biserial correlation of the correct response of items was .46, ranging from $.31 \leq r_{pb} \leq .59$.

Table 2: Item parameters, item fit and differential item functioning of the Russian L1-test

Item	*Difficulty (b)*	SE	WMNSQ	r_{pb}	*DIF* gender	*DIF* school	*DIF* books	*DIF* gen.
nrg90101_c	-1.48	0.12	0.96	0.43	-0.13	0.33	0.08	-0.01
nrg90102_c	0.86	0.11	0.97	0.50	-0.24	0.31	0.24	-0.14
nrg90103_c	-0.29	0.10	1.02	0.44	0.09	0.01	-0.34	-0.01
nrg90201_c	-1.23	0.11	0.95	0.47	0.42	0.88	0.00	0.18
nrg90202_c	-0.51	0.10	1.00	0.46	-0.47	-0.03	0.27	-0.34
nrg90203_c	-0.37	0.10	1.10	0.37	0.41	-0.23	0.17	-0.03
nrg90301_c	0.01	0.10	1.03	0.46	0.47	-0.26	0.05	0.30
nrg90302_c	-0.66	0.10	0.94	0.52	0.36	0.02	-0.08	0.23
nrg90303_c	0.89	0.11	0.92	0.53	0.52	0.14	0.07	0.42
nrg90304_c	-0.13	0.10	1.02	0.45	-0.03	-0.02	0.04	0.33
nrg90401_c	-0.89	0.10	0.91	0.52	0.27	0.18	-0.01	0.28
nrg90402_c	-0.02	0.10	0.93	0.55	-0.25	0.33	0.26	0.31
nrg90403_c	0.76	0.11	1.15	0.31	0.62	0.05	0.30	-0.34
nrg90404_c	0.55	0.10	1.05	0.43	0.20	-0.41	0.04	-0.22
nrg90405_c	-0.10	0.10	0.89	0.58	-0.05	0.38	0.04	0.16
nrg90501_c	-	-	-	-	0.11	-0.36	-0.36	-0.60
nrg90502_c	-0.28	0.10	0.99	0.48	0.19	0.55	0.10	0.09
nrg90503_c	-0.73	0.10	0.99	0.46	-0.22	-0.36	-0.24	0.34
nrg90504_c	-0.16	0.10	1.05	0.43	0.19	0.02	-0.13	-0.35
nrg90505_c	0.79	0.11	1.15	0.34	-0.48	-0.05	0.02	-0.76
nrg90506_c	0.16	0.10	0.99	0.49	-0.26	0.62	0.12	-0.26
nrg90601_c	-0.34	0.10	0.97	0.48	-0.29	0.40	-0.03	-0.16
nrg90602_c	-0.18	0.10	0.93	0.53	-0.53	-0.56	-0.08	0.36

nrg90603_c	-0.36	0.10	1.06	0.41	-0.33	0.00	0.06	-0.28
nrg90604_c	-0.44	0.10	1.12	0.33	-0.09	-0.87	-0.48	0.21
nrg90605_c	0.75	0.11	1.12	0.37	-0.05	-0.10	0.12	-0.76
nrg90701_c	-0.56	0.10	1.06	0.40	-0.71	-0.46	-0.22	-0.01
nrg90702_c	0.14	0.10	0.88	0.59	-0.26	0.10	-0.18	0.44
nrg90703_c	-0.55	0.10	0.98	0.48	-0.35	-0.01	-0.25	0.09
nrg90704_c	-1.17	0.11	1.04	0.39	0.30	-0.15	0.64	-0.20
nrg90705_c	1.41	0.12	0.97	0.45	0.63	-0.56	-0.19	0.38
nrg90706_c	0.26	0.10	0.91	0.56	0.21	0.24	0.06	0.42
Mean	-0.12	0.10	1.00	0.46	0.01	0.00	0.00	-0.01

Note. *b* = item difficulty; *SE = standard error of item difficulty; WMNSQ = weighted mean square;* r_{pb}= point biserial correlation of correct response with total test score; *DIF* gender = differential item functioning according to gender; *DIF* school = differential item functioning according to attended type of secondary school; *DIF* books = differential item functioning according to number of books at home; *DIF* gen. = differential item functioning according to immigrant generation status; *DIF*s are given in absolute differences between groups.

Differential item functioning (DIF)

We examined whether items exhibit DIF with regard to gender, attended type of secondary school (Gymnasium vs. other school types), the number of books in the household (less than 100 vs. more than 100) and the immigrant generation status[3] (first generation vs. other). A negative value on "DIF gender" indicates that the item is easier for female students; a negative value on "DIF school" indicates that the item is easier for students attending the academic track ("Gymnasium"); a negative value on "DIF books" indicates that the item is easier for students with more books in the household; and a negative value on "DIF gen." indicates that the item is easier for students not born in Germany.

Several items showed small (0.4 < |DIF| <= 0.6) or medium (0.6 < |DIF| <= 1) differential item functioning (see table 2). Small DIF-effects were found for 7 items related to gender, for 6 items related to type of school, for 1 item related to number of books in the household, and for 4 items related to immigrant generation status. In addition, 3 items demonstrated medium-size DIF by gender, 3 items by type of school, 1 item by number of books in the household, and 2 items by immigrant generation status. However, none of the items showed a large DIF value (|DIF| > 1). Thus, all items remained in the test (see Pohl & Carstensen, 2012).

[3]The DIF analyses concerning the immigrant generation status does not consider information from the open questions regarding the country of origin in the student questionnaire.

Distractor analyses

Analyses revealed that distractors (incorrect response choices) were on average negatively correlated with participants' overall score in the test (r_{pb} = -.18). The point biserial correlations of distractors with the overall score ranged from -.31 to .02. As none of the distractors correlated positively with the overall test score to a substantial degree, they were all judged to be adequate.

Distribution and reliability of person estimates

In the Rasch model the mean of the person distribution was restricted to M_p= 0. The variance of the person distribution was estimated as σ_p^2 = 1.20; reliability of students' L1-proficiency estimates (WLE) was .85. The mean of the WLE estimates was M_{WLE} = .04 and its variance σ_{WLE}^2 = 1.41, ranging from -2.78 to 4.26. The distribution of WLE estimates was skewed with skewness = 0.72. Graphical analysis of the joint distributions of person estimates and item estimates indicates that the majority of items clusters around the center of the scale (medium difficulty) while the boundaries of the scale are covered by only few items.

Testing unidimensionality

In order to test the scale for unidimensionality, a theoretically plausible, alternative 2-dimensional model was estimated. Items pertaining to texts involving oral features (dialogs) as testlets were assigned to one dimension, while items based on texts representing more literary language (expositions and narrations) as testlets built the other dimension. The two dimensions correlate very highly (*r* = .94). Model fit indices showed that the 2-dimensional model fitted negligibly better on AIC (AIC_{1dim} = 19201.5; AIC_{2dim} = 19196.3; AIC_{Diff} = -5.1) and negligibly worse on BIC (BIC_{1dim} = 19340.7; BIC_{2dim} = 19344.0; BIC_{Diff} = 3.3). The very high correlation indicating near identity of the dimensions and the virtually equal model fit suggest that the construct is unidimensional rather than bidimensional.

Testing for local item dependencies

In order to explore the data for testlet effects, Yen's Q3 statistics (Yen, 1984) were computed. The mean of all bivariate item Q3 values per testlet was used as an indicator of a testlet effect. These Q3 testlet means range from -.02 to .05 (M = .01), indicating that effects due to testlet clustering are negligible. All other bivariate item Q3 values were non-critical as well.

4. Turkish L1-test

4.1 Preliminary analyses

4.1.1 Exclusion of cases from the analyses

Overall, 666 students participated in the Turkish L1-test. Less than 3 valid answers were available for 4 cases and they were therefore excluded from further analyses. Thus, the scaling of the Turkish L1-Tests was based on 662 cases.

4.1.2 Sample

The sample consists of 320 (48.3%) female and 342 (51.7%) male students. Almost half of the participants (N = 330 or 49.9%) attended the *Hauptschule*, 130 (19.6%) the *Realschule*, 97 (14.6%) the *Gesamtschule* and 94 (14.2%) the *Gymnasium.* 11 students (1.7%) attended schools with several educational tracks. The information on the immigrant generation status is based on 657 valid answers. The majority of the participants was born in Germany (N = 584 or 88.9%), while 73 students (11.1%) were foreign-born.

4.1.3 Descriptive statistics of responses and missing responses

There existed 639 valid answers per item on average (see table 3). The mean of omitted responses per item was 3.4% and the mean of invalid responses 0.1%. On average, the students provided valid answers on 30.9 (*Mdn* = 32) or 96.5% of items. Valid answers per person ranged from 9.4% to 100%. As in the Russian test, none of the items was extremely easy or difficult. The most difficult item was solved by 24.8% of the participants, while the item with the lowest difficulty was answered correctly by 82.3% of the students. On average, students answered 54.2% of items correctly in the Turkish L1-test.

Table 3: Descriptives and missing responses in the Turkish L1-Test

Item	Position in the test	Testlet	Valid responses	Omitted responses (in %)	Invalid responses (in %)
ntg90101_c	1	text 1	656	0.91	0.00
ntg90102_c	2	text 1	648	1.96	0.15
ntg90103_c	3	text 1	633	4.38	0.00
ntg90201_c	4	text 2	656	0.91	0.00
ntg90202_c	5	text 2	644	2.57	0.15
ntg90203_c	6	text 2	642	2.87	0.15
ntg90301_c	7	text 3	636	3.78	0.15
ntg90302_c	8	text 3	631	4.68	0.00

ntg90303_c	9	text 3	616	6.65	0.30
ntg90304_c	10	text 3	629	4.98	0.00
ntg90401_c	11	text 4	653	1.21	0.15
ntg90402_c	12	text 4	643	2.87	0.00
ntg90403_c	13	text 4	620	6.34	0.00
ntg90404_c	14	text 4	628	5.14	0.00
ntg90405_c	15	text 4	637	3.78	0.00
ntg90501_c	16	text 5	645	2.57	0.00
ntg90502_c	17	text 5	646	2.27	0.15
ntg90503_c	18	text 5	639	2.87	0.60
ntg90504_c	19	text 5	647	2.11	0.15
ntg90505_c	20	text 5	629	4.83	0.15
ntg90506_c	21	text 5	645	2.57	0.00
ntg90601_c	22	text 6	647	2.27	0.00
ntg90602_c	23	text 6	636	3.78	0.15
ntg90603_c	24	text 6	639	3.32	0.15
ntg90604_c	25	text 6	634	4.23	0.00
ntg90605_c	26	text 6	632	4.53	0.00
ntg90701_c	27	text 7	642	3.02	0.00
ntg90702_c	28	text 7	642	3.02	0.00
ntg90703_c	29	text 7	645	2.57	0.00
ntg90704_c	30	text 7	634	4.23	0.00
ntg90705_c	31	text 7	642	3.02	0.00
ntg90706_c	32	text 7	634	4.08	0.15
Mean			639	3.38	0.08

Note. Valid responses are reported in terms of absolute frequencies; omitted responses as well as invalid responses are provided as relative frequencies (percent).

4.2 Results of scaling

4.2.1 Analyses

The data of the Turkish L1-test were modeled in the same way as the Russian L1-test (see section 1.2.1 and the Appendix for the syntax).

4.2.2 Results

Item parameters and item fit

Even though item ntg90501_c, which was eliminated from the Russian L1-Test, did not exert a misfit in the Turkish L1-test, it was excluded from the scaling in order to keep the Turkish and the Russian test versions analogous. The item fit of the remaining 31 items was .92 ≤ WMNSQ ≤ 1.14 (see table 4). The average item difficulty in the Turkish L1-Test was b = -0.23 with a range of -1.78 ≤ b ≤ 1.30. The average point biserial correlation of the correct response of items with the overall test score was .42, ranging from .24 ≤ r_{pb} ≤ .51.

Table 4: Item parameters, item fit and differential item functioning of the Turkish L1-test

Item	*Difficulty (b)*	SE	WMNSQ	r_{pb}	*DIF* gender	*DIF* school	*DIF* books	*DIF* gen.
ntg90101_c	-1.46	0.10	0.98	0.40	0.17	-0.72	-0.15	0.09
ntg90102_c	-0.90	0.09	0.98	0.41	-0.13	-0.49	-0.04	-0.19
ntg90103_c	-0.53	0.09	0.97	0.46	-0.05	-0.33	0.08	-0.74
ntg90201_c	-1.78	0.11	0.97	0.38	0.46	0.13	-0.33	-0.57
ntg90202_c	-0.28	0.09	0.97	0.46	-0.17	0.35	0.10	-0.23
ntg90203_c	-0.82	0.09	1.01	0.40	0.71	-0.19	0.07	-0.03
ntg90301_c	-0.41	0.09	1.01	0.41	0.52	-0.06	0.45	-0.09
ntg90302_c	-0.01	0.09	0.94	0.50	0.15	0.01	0.03	0.00
ntg90303_c	0.06	0.09	0.98	0.46	0.64	-0.21	0.15	-0.07
ntg90304_c	0.39	0.09	0.98	0.44	0.00	0.05	-0.38	0.29
ntg90401_c	-0.96	0.09	0.96	0.45	-0.09	0.35	-0.11	-0.27
ntg90402_c	0.01	0.09	0.98	0.45	0.11	0.48	-0.19	-0.18
ntg90403_c	0.71	0.09	1.10	0.28	0.14	0.05	-0.11	-0.26
ntg90404_c	0.86	0.09	1.13	0.24	-0.13	-0.68	-0.26	0.06

ntg90405_c	-0.15	0.09	0.97	0.46	0.22	0.11	-0.07	-0.08
ntg90501_c	-	-	-	-	0.08	-0.07	-0.26	0.41
ntg90502_c	-0.37	0.09	0.99	0.43	0.33	-0.02	-0.02	0.18
ntg90503_c	-0.13	0.09	0.99	0.43	0.05	0.19	0.15	-0.06
ntg90504_c	-1.33	0.10	0.92	0.46	-0.05	-0.32	0.11	-0.44
ntg90505_c	0.95	0.09	1.07	0.30	-0.31	0.77	0.11	-0.16
ntg90506_c	-0.31	0.09	0.93	0.51	-0.06	0.05	0.11	0.32
ntg90601_c	-1.02	0.09	0.95	0.46	-0.62	0.19	0.13	0.21
ntg90602_c	0.02	0.09	1.02	0.40	0.23	-0.19	-0.17	-0.18
ntg90603_c	-0.15	0.09	0.99	0.44	-0.39	0.50	0.35	-0.02
ntg90604_c	-0.34	0.09	0.99	0.44	-0.26	-0.04	0.28	0.28
ntg90605_c	1.30	0.10	1.05	0.31	0.33	-0.30	0.14	0.33
ntg90701_c	-0.05	0.09	1.14	0.27	-0.15	-0.70	-0.19	0.30
ntg90702_c	0.02	0.09	0.94	0.51	-0.71	0.30	0.17	0.39
ntg90703_c	-0.44	0.09	0.95	0.48	-0.25	-0.03	-0.11	0.04
ntg90704_c	-0.22	0.09	0.99	0.44	-0.16	-0.10	-0.10	0.21
ntg90705_c	0.29	0.09	1.02	0.41	-0.20	0.36	-0.16	0.40
ntg90706_c	-0.03	0.09	1.06	0.36	-0.25	0.31	0.10	-0.24
Mean	0.06	0.09	1.00	0.42	0.01	-0.01	0.00	-0.01

Note. *b* = item difficulty; *SE = standard error of item difficulty; WMNSQ = weighted mean square;* r_{pb}= point biserial correlation of correct response with total test score; *DIF* gender = differential item functioning according to gender; *DIF* school = differential item functioning according to attended type of secondary school; *DIF* books = differential item functioning according to number of books at home; *DIF* gen. = differential item functioning according to immigrant generation status; *DIF*s are given in absolute differences between groups.

Differential item functioning (DIF)

In the Turkish L1-test, DIF was examined with regard to the same indicators as the Russian L1-test (see section 3.2.2). Again, several items showed small (0.4 < |DIF| <= 0.6) or medium (0.6 < |DIF| <= 1) differential item functioning (see table 4). 2 items had small DIF related to gender as well as 3 items related to type of secondary school, 1 item related to number of books in the household, and 4 items related to immigrant generation status. Medium-size DIF effects were found for 4 items by gender, for 4 items by type of secondary school, and

for 1 item by immigrant generation status. None of the items reached the threshold of |DIF|> 1, so that all items remained in the test.

Distractor analyses

On average, the distractors in the Turkish L1-test were negatively correlated with the overall test score of the test (r_{pb} = -.16), the point biserial correlations ranged from -.33 to .03. Thus, all distractors proved suitable.

Distribution and reliability of person estimates

As in the Russian L1-test, the mean of the person distribution was constrained to M_p= 0. The estimated variance of the person distribution was σ_p^2= 0.86. The reliability of students' L1-profiency estimates (WLEs) amounted to .83; the WLE estimates ranged from -3.46 to 3.00 with M_{WLE} = 0.00 and a variance of σ_{WLE}^2= 1.04. The distribution of WLE estimates seemed normal (skewness = 0.01). Similarly to the Russian L1-test, the graphical analyses displaying the joint distributions of person estimates and item estimates indicated that the majority of items clusters around the center of the scale (medium difficulty), while the boundaries of the scale are covered by fewer items.

Testing unidimensionality

Similarly to the Russian measure, the Turkish L1-items were tested for unidimensionality. The two dimensions specified (comprehension of orally-oriented vs. more literate language) are highly correlated (r = .98). The 2-dimensional model fitted marginally worse as indicated by the AIC (AIC_{1dim} = 25767.9; AIC_{2dim} = 25772.1; AIC_{Diff} = 4.2) and somewhat worse according to the BIC (BIC_{1dim} = 25916.2; BIC_{2dim} = 25929.4; BIC_{Diff} = 13.2). Considering the very high correlation indicating near identity of the dimensions and the almost equal model fit, unidimensionality can be assumed for the Turkish L1-test as well.

Testing for local item dependencies

In the Turkish L1-test, the Q3 item means per testlet range from .02 to .09 (M = .02), indicating that effects due to testlet clustering are negligible. All other bivariate item Q3 values were non-critical as well.

5. Discussion

The paper described the procedure and results of scaling the 9th-grade listening comprehension tests of immigrant students' first languages Turkish and Russian. It results in a final compilation of 31 items in each of the two analogous test forms (Russian and Turkish). Scaling analyses suggest that the instruments are psychometrically sound. All indicators of measurement quality are good or very good. The instrument conforms to the one-parameter logistic model (Rasch model) and tests of dimensionality support the assumption that the developed tests capture a unidimensional construct of listening comprehension in L1. One limitation of the tests is the somewhat sparse occurrence of items in extreme bands of the measurement scale implying a slightly lower measurement precision in these areas. However, the tests precisely measure the proficiency of students at intermediate levels and discriminate well between students with high and low L1-proficiencies. In sum, the developed tests seem to be highly suitable instruments for measuring immigrant students' proficiency in their first languages Russian and Turkish.

Appendix

Syntax of the scaling analyses (ConQuest): The Russian L1-test

```
title = Analysis name: nrg9_scaling;
exportlogfile >> nrg9_scaling.log;
datafile nrg9_scaling.dat;
Format pid 1-10 responses 11-41;
labels  << nrg9_scaling.lab;
set constraints=cases;
set warnings=no,update=yes,n_plausible=5,p_nodes=2000,f_nodes=2000;
model item;
estimate !method=gauss,iter=1000,nodes=15,converge=0.0001,deviancechange=0.0001,
stderr=quick,distribution=normal;
Itanal >> nrg9_scaling.itn;
show cases! estimate=wle >> nrg9_scaling.wle;
show >> nrg9_scaling.shw;
quit;
```

Syntax of the scaling analyses (ConQuest): The Turkish L1-test

```
title = Analysis name: ntg9_scaling;
exportlogfile>> ntg9_scaling.log;
datafile ntg9_scaling.dat;
Format pid 1-10 responses 11-41;
labels<< ntg9_scaling.lab;
set constraints=cases;
set warnings=no,update=yes,n_plausible=5,p_nodes=2000,f_nodes=2000;
model item;
estimate !method=gauss,iter=1000,nodes=15,converge=0.0001,deviancechange=0.0001,
stderr=quick,distribution=normal;
Itanal>> ntg9_scaling.itn;
show cases! estimate=wle>> ntg9_scaling.wle;
show>> ntg9_scaling.shw;
quit;
```

References

Edele, A., Stanat, P., & Schotte, K. (forthcoming). Listening comprehension tests of immigrant students‘ first languages (L1) Russian and Turkish in grade 9: Construction and validation (NEPS Working Paper). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, I., Bos, W., & Kanders, M. (2011). Transition and development from lower secondary to upper secondary school. In H.-P.Blossfeld, H.-G.Rossbach & J. von Maurice (Eds.), Education as a lifelong process. The German National Educational Panel Study (NEPS) [Special Issue]. *Zeitschrift für Erziehungswissenschaft, 14,*217-232.

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.

R Core Team (2012). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* ISBN 3-900051-07-0, URL http://www.R-project.org/

Stata Statistical Software: Release 11 [Computer software]. College Station, TX: StataCorp LP.

von Maurice, J., Sixt, M., & Blossfeld, H.-P. (2011*). The German National Educational Panel Study: Surveying a Cohort of 9th Graders in Germany* (NEPS Working Paper No. 3). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel. Retrieved from: https://www.neps-data.de/de-de/projekt%C3%BCbersicht/publikationen/nepsworkingpapers.aspx

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0 : generalised item response modelling software.* Camberwell, Victoria.

Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 8*(2), 125–145.