NEPS WORKING PAPERS

Aileen Edele, Kristin Schotte, and Petra Stanat

ASSESSMENT OF IMMIGRANT STUDENTS' LISTENING COMPREHENSION IN THEIR FIRST LANGUAGES (L1) RUSSIAN AND TURKISH IN GRADE 9: EXTENDED REPORT OF TEST CONSTRUCTION AND VALIDATION

NEPS Working Paper No. 57
Bamberg, May 2015

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

Working Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) 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
https://www.neps-data.de/projektübersicht/publikationen/nepsworkingpapers

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

Assessment of Immigrant Students' Listening Comprehension in Their First Languages (L1) Russian and Turkish in Grade 9: Extended Report of Test Construction and Validation

Aileen Edele, Humboldt-Universität zu Berlin

Kristin Schotte, Humboldt-Universität zu Berlin

Petra Stanat, Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Humboldt-Universität zu Berlin

E-mail address of lead author:

aileen.edele@iqb.hu-berlin.de

Bibliographic data:

Edele, A., Schotte, K., & Stanat, P. (2015). *Listening Comprehension Tests of Immigrant Students' First Languages (L1) Russian and Turkish in Grade 9: Extended Report of Test Construction and Validation* (NEPS Working Paper No. 57). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Assessment of Immigrant Students' Listening Comprehension in Their First Languages (L1) Russian and Turkish in Grade 9: Extended Report of Test Construction and Validation

Abstract

In large-scale studies, immigrant students' first language (L1) proficiency is typically measured with subjective instruments, such as self-reports, rather than with objective tests. The National Educational Panel Study (NEPS) addresses this methodological limitation by testing the L1 proficiency of the two largest immigrant groups in Germany, namely students whose families have immigrated to Germany from the area of the Former Soviet Union or Turkey. Listening comprehension tests in Russian and Turkish were developed for this purpose. The current paper describes the general framework and requirements for testing first language proficiency within the NEPS and describes the construction of the L1 tests for 9th grade students. Subsequently, the paper reports the item difficulty and reliability of the tests as well as analyses of measurement equivalence indicating that the Russian and Turkish test assess the same construct (configural equivalence). The ability scores and their correlations with other variables are, however, not directly comparable. Analyses of construct validity confirm the unidimensional structure expected for the test. In addition, the L1 test scores correlate with other indicators of L1 proficiency as well as with factors regarded as crucial for L1 acquisition, such as exposure to L1, in the expected way (convergent validity) and they are not substantially related to measures of general cognitive abilities (discriminant validity). We conclude that the listening comprehension tests developed in the NEPS are valid measures of L1 proficiency.

Keywords

first language proficiency, L1 proficiency, listening comprehension, test construction, validity

1. Introduction

The effects immigrant students' first language (L1) proficiency may have on their social integration and educational success are highly disputed.[1] On the one hand, some theoretical perspectives suggest positive effects of L1 proficiency on second language acquisition (e.g., Cummins, 2000; Scheele, Leseman, & Mayo, 2010; Verhoeven, 2007) as well as on third language learning (Hesse, Göbel, & Hartig, 2008; Rauch, Jurecka, & Hesse, 2010) and bilingualism is assumed to promote cognitive development (Adesope, Lavin, Thompson, & Ungerleider, 2010; Bialystok, 2007). On the other hand, neutral or negative effects of proficiency in L1 are proposed (e.g., Dollmann & Kristen, 2010; Esser, 2006; Mouw & Xie, 1999).

The empirical evidence necessary to elucidate this controversy is, however, inconclusive. This is also a result of the methodological constraints of most studies on this issue. In particular, there is a lack of investigations assessing L1 proficiency with objective tests rather than with subjective measures, especially when it comes to analyses with larger sample sizes. Previous large-scale studies typically relied on self-report measures of L1 proficiency. For instance, the National Educational Longitudinal Study (NELS; Mouw & Xie, 1999), the Children of Immigrants Longitudinal Study (CILS; Portes & Rumbaut, 2012), the International Comparative Study of Ethno-cultural Youth (ICSEY; Berry, Phinney, Sam, & Vedder, 2006a), as well as the German Socio-Economic Panel Study (GSOEP; TNS Infratest Sozialforschung, 2012) and the Programme for International Student Assessment (PISA; Frey et al., 2009) examine participants' L1 proficiency based on self-reported questionnaire data. Participants are typically asked to indicate their L1 proficiency on four-point or five-point rating scales with regard to one or several dimensions (understanding, speaking, reading, writing).

A few small-scale studies on the effects of L1 proficiency, however, have implemented objective instruments for assessing L1. For instance, Rauch and colleagues (2010) examined reading comprehension of 8th grade immigrant students' in their L1 (Turkish) with a computer-based adaptive test. Similarly, Leseman, Scheele, Mayo and Messer (2009) assessed pre-schoolers' receptive vocabulary in their L1 Turkish or Tarifit-Berber. Verhoeven (2007) tested various aspects of children's phonological, lexical, morpho-syntactic and textual proficiency in their L1 Turkish. Dollmann and Kristen (2010) used C-tests in order to examine elementary school students' proficiency in their L1 Turkish. These studies, however, are based on small, non-representative samples. It is therefore questionable whether the findings from these studies are generalizable.

The National Educational Panel Study (NEPS; Blossfeld, Roßbach, & von Maurice, 2011) set out to address this research gap by measuring L1 proficiency of students from the two largest immigrant groups in Germany with objective tests. The project assesses the listening comprehension proficiency of children and adolescents whose families immigrated to Germany from the area of the Former Soviet Union or Turkey. As no suitable instruments for

[1] In line with the terminology commonly used in the literature, the term immigrant students refers to the first, second and third immigrant generation in this paper. The term first language (L1) is used here interchangeably with the language in an immigrant family's country of origin, regardless of whether this language was actually acquired prior to the language of the destination country (in our case German), as the label "L1" suggests, or simultaneously.

this purpose were available, tests in Turkish and Russian were developed by the Berlin project team within pillar 4 of the NEPS.[2]

This paper describes the construction of the L1 tests for 9th grade students and reports analyses exploring the tests' validity. The following section delineates the general framework for testing L1 within the NEPS as well as the requirements the instruments had to meet in order to be suitable for implementation in the study. In section 3, we describe the tests developed for students in grade 9. Section 4 reports analyses pertaining to the tests' validity.

2. General Framework and Requirements for Assessing L1 Proficiency within the NEPS

The NEPS assesses L1 proficiency with objective tests at three measurement points. More specifically, the instruments capture basic listening comprehension skills in Russian and Turkish in grade 9 (starting cohort 4 and later on starting cohort 3[3]), in grade 7 (starting cohort 3) and in grade 2 (starting cohort 2). The current paper focuses on the tests developed for 9th grade. At this measurement point, L1 proficiency is assessed shortly before students attending the vocational school tracks transit from school to work life. This makes it possible to explore the extent to which L1 proficiency affects the transition to vocational training and the labor market success. In addition, analyses based on starting cohort 4 are particularly robust due to the particularly large sample size of this starting cohort (Aßmann et al., 2011).

To ensure that analyzed patterns between starting cohorts can be roughly compared, the tests for 7th grade are analogously construed to the tests in 9th grade, but mainly consist of different texts and items. The tests for 2nd grade also resemble the L1 tests for secondary school in general characteristics, but are adapted in some aspects to the test-takers' age.

2.1 Defining and Identifying the Target Population

The NEPS set out to assess the L1 proficiency of the two largest immigrant groups in Germany, namely students from families who have immigrated to Germany from the area of the Former Soviet Union (e.g., Russia, Kazakhstan) or from Turkey. To ensure that all students from these immigrant groups participating in the NEPS are included in the initial sample, the target population is defined as students of the first, second and third generation.[4]

However, not all immigrants maintain their heritage language (Chiswick & Miller, 2001; Esser, 2006; Rumbaut, Massey, & Bean, 2006) and competence testing is, of course, only meaningful if test-takers possess some proficiency in the tested domain. Therefore, we implemented screening tests with very low item difficulty prior to L1 testing. Only students who demonstrated a minimal level of listening comprehension in these tests were asked to participate in the actual L1 assessment (see section 4.2 for further information).

[2] The tests for grade 9 were developed by Aileen Edele and Petra Stanat.

[3] See Blossfeld, von Maurice, and Schneider (2011) for a description of the starting cohorts and project structure in general.

[4] To include all students whose L1 is potentially Russian or Turkish, we initially defined the target population based on country of birth, even though Russian or Turkish is not necessarily the L1 of all families from the Former Soviet Union or Turkey.

2.2 Efficiency

As the NEPS assesses a large number of constructs, testing time available for each competence domain is limited. As a consequence, it was impossible to include multidimensional L1 tests measuring the various components of language proficiency separately, such as vocabulary and grammar. Instead, global and efficient, yet comprehensive measures of L1 proficiency had to be developed. Listening comprehension constitutes a complex process requiring the integration of phonological, syntactic, semantic and pragmatic skills (Anderson, 1995; Flowerdew & Miller, 2005). We therefore decided to assess this aspect as an indicator of general language proficiency.

In order to limit the testing time and financial costs, moreover, the NEPS L1 tests were required to be applicable in group settings and to use a paper-and-pencil format. To avoid costs associated with coding open-response items, moreover, all test items had to be in a multiple-choice format.

2.3 Focus on Listening Comprehension

Models of language proficiency and language testing often distinguish between four basic dimensions of language proficiency: listening, reading, speaking and writing (Harris, 1969; Lado, 1961). Large-scale studies assessing language proficiency focus mainly on reading comprehension. However, children of immigrants typically acquire the L1 in the family context, and the L1 is rarely used in school instruction. A large proportion of immigrant students is therefore unable to read or write in that language. To ensure that students at all levels of L1 proficiency can participate in the assessment, we decided to test the domain of listening comprehension.

Including students who are unable to read and write in their L1 and who may overall have limited L1 skills is important because analyses on most research questions require that a broad spectrum of L1 proficiency is represented in the data. For instance, to identify factors predicting the maintenance or loss of L1, it is crucial to include lower proficiency levels in L1. Similarly, in estimating effects of L1 proficiency on L2 or L3, it may be informative to differentiate between effects of lower and higher levels of L1 proficiency (e.g., Dollmann & Kristen, 2010; Edele & Stanat, submitted).

2.4 Coverage of Proficiency Distribution

In order to ensure that the L1 tests developed for the NEPS would cover a broad range of proficiency levels, we developed listening comprehension texts with varying linguistic difficulty. Based on data from a preliminary study and a larger pilot study, moreover, we also ensured that the difficulty of the items varied substantially (for details see section 4.5; Edele, Schotte, Hecht, & Stanat, 2012). Due to the tests' limited number of items, however, the instruments differentiate most accurately at intermediate proficiency levels, while their capacity to measure precisely very high or low proficiency levels is somewhat restricted.

2.5 Independence of Test Performance from Previous Knowledge

The L1 tests aim at assessing students' ability to understand spoken language in L1. To ensure that the test does, in fact, measure language proficiency rather than prior knowledge, we used texts which either cover topics that should be familiar to all students alike, such as everyday situations in school or family contexts, or topics that are likely to be equally unfamiliar to all participants, such as events in a fictitious narration written specifically for the test.

One aspect that needs to be taken into account in testing immigrant students is the possible impact of cultural knowledge on their performance. There is evidence that text processing is contingent on the fit between cultural knowledge and the content of the text. Steffensen, Joag-Dev and Anderson (1979), for instance, found in their study that participants recalled more information and needed less time when the text they read described content consistent with their cultural knowledge rather than content typical for another culture. Thus, it can be assumed that a text is easier to process when it is in line with the test takers' culturally shaped prior knowledge.

This knowledge is likely to vary considerably in the target population for testing L1 in the NEPS. While some students were born in Germany, others have just recently arrived from their heritage cultures. In addition, immigrants acculturate in different ways: while some immigrant students are strongly oriented towards the heritage culture of their family, others are mostly involved in the culture of the receiving country — or in both or neither of the cultures (e.g., Berry, Phinney, Sam, & Vedder, 2006b; Edele, Stanat, Radmann, & Segeritz, 2013). In order to avoid bias associated with students' culturally shaped knowledge, the stimuli in the L1 tests were chosen such that they should be equally familiar or equally novel to the students with different cultural backgrounds.

2.6 Comparability of the Russian and Turkish Test

Some research questions will focus on one of the two first languages assessed in the NEPS and the respective immigrant group only. For other research questions, however, it may be important or interesting to determine whether the expected pattern of findings generalizes across both L1s and immigrant groups. To ensure that the relationships between L1 proficiency and other constructs can be compared across the two groups, the tests in Russian and Turkish need to capture the same construct. We therefore developed tests in Russian and Turkish which are equivalent with regard to the content of the texts, the questions and the response options. In addition, we tried to keep linguistic features comparable that are likely to affect the difficulty of the texts.

Even the most careful translation process, however, does not necessarily ensure measurement equivalence. Measurement equivalence can be tested with multigroup confirmatory factor analyses (MGCFA). Different forms of equivalence can be distinguished by gradually constraining measurement parameters and subsequently comparing the fit of the resulting models (see Schroeders & Wilhelm, 2011 for a detailed description of testing invariance in categorical data). In short, for categorical data, in the least restrictive model factor loadings, thresholds and residual variance are freely estimated (configural invariance). For strong measurement invariance factor loadings and thresholds are fixed to equality across groups while constraining the residual variance in one group at 1 and estimating it

freely in the other group. In the most restrictive model of strict invariance factor loadings and thresholds need to be equal and residual variances are fixed at 1 in both groups. If configural equivalence is given, it can be assumed that an instrument assesses the same construct in two or more groups or — in our case — that the instruments assess the same construct in the respective target population. Strong invariance is necessary to directly compare the latent means of a test and its correlations with external criteria. In strictly invariant tests, even raw test scores are comparable. Due to the pronounced linguistic differences between the Russian and the Turkish language on the one hand, and the limited time and financial resources for test development on the other hand, it would have been unrealistic to expect strong or even strict invariance of the two tests. We did, however, strive for configural invariance.

3. Test Construction

The L1 tests developed for grade 9 consist of short texts that are orally presented to students from a CD. They are subsequently asked to answer questions about what they have heard. All items have a multiple-choice format.

3.1 Text and Item Construction

The texts and items of the test were initially constructed in German and subsequently translated into Russian and Turkish. The different text units within the test are independent from each other and consist of approximately 100 to 150 words. To assess listening comprehension broadly, the tests include dialogues, expository texts and narrative texts. The text types differ in their linguistic features. While the dialogues have features typical for informal oral language, such as repetitions, redundancies, ellipses, breaks or fragmented language, the expository and narrative texts involve linguistic features typical for written language, such as more explicit vocabulary, correct grammar and a lack of redundancy or repetition (Grotjahn, 2005; Shohamy & Inbar, 1991). A preliminary version of the tests was included in a pilot study (see section 4.2 and 4.4 for more details on the pilot study) for the purpose of item selection.

To cover a range of language proficiency levels, we constructed texts varying in difficulty with regard to grammar and vocabulary. More difficult texts incorporate less frequently used vocabulary, a more complex syntax and a greater variety of tenses, including those that are less frequently used.

For the construction of the texts and questions, we closely collaborated with two L1 experts who have an excellent command of the Russian or Turkish language as well as a very good understanding of the respective culture as they have grown up in Russia or Turkey, respectively, hold university degrees in Slavic studies or Turkology and are experienced teachers of the respective language at the university level.

To ensure that culturally shaped previous knowledge would not threaten the validity of the tests, the L1 experts checked whether the content presented in the texts was not only in line with the German culture, but also consistent with the respective heritage culture. They, for instance, rejected a text describing a typical German birthday party, as a child from a family which is strongly oriented towards its heritage culture may not be familiar with this

situation. We only used texts and items the L1 experts judged to be independent from culturally shaped previous knowledge.

An institute experienced in multilingual translations and the translation of test materials of international large-scale assessment studies, such as PISA, translated the resulting 12 texts and 73 items. In order to keep the tests in Russian and Turkish linguistically as equivalent as possible, we instructed the translators to maintain linguistic features of the original texts, such as text difficulty as well as the features typical for oral versus written language, as far as possible without jeopardizing the linguistic appropriateness in Russian or Turkish.

3.2 Text and Item Selection

Subsequently, the L1 experts submitted the test material to cognitive interviews with four adult native speakers per language. These interviews confirmed that participants answered items based on information retrieved from the text rather than based on their prior knowledge. In addition, the interviews probed once more whether the test material was consistent with the Russian and Turkish culture. Based on the findings, we excluded one text from the tests involving a couple on a hiking tour, as hiking is less popular in Turkey and the area of the Former Soviet Union than in Germany and immigrants from these countries may therefore lack the corresponding background knowledge about navigation (e.g., deploying a map and compass). The resulting version of the tests only contains texts that were judged to be consistent with the Russian and Turkish culture.

To get a first indication of the texts' difficulty, the adult native speakers rated this characteristic on a 3-point scale (1 = easy; 2 = intermediate; 3 = difficult). The mean of the ratings across raters ranged from 1 to 3 for the Russian as well as for the Turkish test with means across all texts of 2.0 for the Russian version and 2.2 for the Turkish version.

After these revisions of the test material had been made, native speakers of Russian or Turkish audio-recorded the remaining 11 texts and items as well as the test instructions. In order to obtain a rough indicator of the item difficulties, 10 adolescents with Russian as L1 and 12 adolescents with Turkish as L1 completed the recorded tests. The item difficulty in terms of the proportion of students who answered the items correctly ranged from $p = .33$ to $p = 1.00$ ($M = .76$) in the Russian test and from $p = .00$ to $p = 1.00$ ($M = .51$) in the Turkish test.

Based on these findings, we chose 9 texts varying in difficulty for the inclusion in a pilot study (see section 4.2 and 4.4 for more details on the study) and excluded a number of extremely easy and difficult items pertaining to the remaining texts. The resulting test versions comprised four dialogues, two narrative texts and three expository texts with five to seven multiple-choice questions each, totaling to 54 items. Every multiple-choice item had four or five response options.

Based on item analyses from the pilot study, we excluded two more texts and 22 questions from the tests. In addition, wrong response options (distractors) which correlated positively with the overall test score were eliminated.

3.3 Final Test Version

The final tests for grade 9 included in the main studies of the NEPS are comprised of seven text units, namely two dialogues, two narrative texts and three expository texts. The audio-recorded texts and questions are presented to the students once before they answer the questions about what they have heard. Every text unit is followed by three to six multiple-choice questions, resulting in a total of 31[5] test items, each with four or five response options. The administration of the test takes 30 minutes (Russian version) and 32 minutes (Turkish version).

4. Validity of the L1 Tests

4.1 Validation Strategy

To investigate the construct validity of the L1 tests, we examined their dimensionality and correlated students' scores with other indicators of L1 proficiency as well as with a nomological net of relevant constructs (Cronbach & Meehl, 1955).

As a first step, we tested whether our L1 tests possess the expected unidimensional structure. To establish the convergent validity (Campbell & Fiske, 1959) of our L1 measures, we then correlated the test scores with another indicator of proficiency in Russian or Turkish, namely the *Bilingual Verbal Ability Test* (Muñoz-Sandoval, Cummins, Alvarado, & Ruef, 1998). As both instruments are objective tests of language proficiency, we expected the correlations to be substantial. However, as the instruments examine different aspects of language proficiency (see below), we did not expect a particularly tight association between them. If the criterion test assessed different aspects of language proficiency than the developed tests, for instance productive instead of receptive language, somewhat lower correlations between the two instruments should occur.

As another indicator of convergent validity, we correlated our L1 test scores with subjective measures of L1 proficiency. Even though subjective assessments, and in particular self-assessments, are susceptible to biases (Edele, Seuring, Kristen, & Stanat, in press), we expected at least a moderately high correlation.

General cognitive abilities served as criteria in our analyses of discriminant validity. They are assumed to influence the efficiency of language acquisition and should therefore relate positively with proficiency (e.g., Esser, 2006; Spolsky, 1989). In addition, text comprehension requires deductive reasoning. However, L1 proficiency depends on a multitude of other factors and should be clearly distinguishable from general cognitive abilities. We therefore expected a significant yet moderate relationship between listening comprehension in L1 and reasoning ability. Speed of perception, by contrast, should be unrelated to performance in our L1 tests as they do not contain a speed component.

To extend the nomological net for the construct validation of the L1 tests, we draw on models of language acquisition from different disciplines. These models suggest a number of conditions that should foster the acquisition and maintenance of L1 proficiency in

[5] Of the 32 items originally included in the final test version, one item was later excluded due to a poor model fit in the main study (see Edele et al., 2012 for further information on the scaling of the tests).

immigrants and their children, such as exposure and motivation for language acquisition and improvement (e.g., Chiswick & Miller, 1994, 2001; Esser, 2006; Spolsky, 1989).

Immigrant students are exposed to their L1 in different contexts. The most important environment for the acquisition and improvement of L1 skills is typically the family. In addition, children and adolescents from immigrant families may have the opportunity to improve their L1 in interactions with co-ethnic peers. The use of media in L1 can also present an important learning opportunity for L1 acquisition. Therefore, students' exposure to L1 in the family, in the peer group and in the media should be positively related to their L1 proficiency (e.g., Duursma et al., 2007; Scheele et al., 2010).

The immigrant generation status can also be assumed to affect exposure to L1 (e.g., Chiswick & Miller, 2001). Generally, a decrease in L1 use and proficiency and an increase in L2 use and proficiency can be observed in immigrants over the time (Rumbaut, 2004; Stanat, Rauch, & Segeritz, 2010). First generation immigrant students thus typically have more opportunities to acquire the L1 in their family context than students who were born in the country of residence. Moreover, first generation immigrants may have acquired the L1 in their heritage country. Accordingly, we expected first generation immigrants to be more proficient in L1 than successive immigrant generations.

Within the first immigrant generation, students who immigrated at an older age were extensively exposed to L1 while they lived in the heritage country – and the quality of the language input can be assumed to be relatively high. Therefore, we expected age at migration to correlate positively with L1 test scores.

In addition to providing learning opportunities, using L1 with family and peers and in media can foster immigrant students' motivation to further improve their L1 skills. As a strong identification with the heritage culture should boost the motivation to improve in L1, moreover, it should be positively related to L1 proficiency.

4.2 Study Design

We draw on data from two studies to examine the validity of our tests. This allows us, on the one hand, to cross-check the findings, as most validation criteria were measured in both studies. On the other hand, the two studies complement each other, as some validation criteria were assessed in only one of the studies. In addition, the second study includes a larger sample (for further details on sampling within the NEPS see Aßmann et al., 2011).

The first investigation (study 1) is a pilot study that was carried out within the NEPS to select items for the final L1 tests. The target population was identified based on information from parents. On the test day, students filled out a questionnaire and subsequently completed a preliminary version of an L1 test in their respective first language. The preliminary test versions implemented in this study included nine texts and 45 corresponding items. The analyses presented in this paper are based on the seven texts and 31 items included in the final test version that was subsequently administered in the second study (see below). The L1 tests in both studies are largely identical, with the only exception that a few false response options (distractors) were excluded from some items in the final tests. This should not substantially affect the patterns of findings relevant for validating the tests. A few

months after the assessment, a sub-sample of the students completed another test measuring proficiency in Russian or Turkish, respectively.

The second study (study 2) on which we draw in the following analyses is the main study of the NEPS for the 9th grade sample of starting cohort 4 (school and vocational training — education pathways of students in 9th grade and higher, doi:/10.5157/NEPS:SC4:4.0.0[6]; see Frahm et al., 2011; von Maurice, Sixt, & Blossfeld, 2011, for further information on this starting cohort). The target population for testing L1 was identified based on student questionnaires implemented in a prior wave. The L1 tests were administered in a group setting on a separate testing day. To ensure that the students had at least a very basic proficiency level in Russian or Turkish, they completed a screening test in the respective language prior to the L1 test. The screening test consists of recordings of eight simple spoken sentences, such as "the dog walks." Participants were asked to relate each sentence to the corresponding picture among five options. Test administrators instantly scored the screening tests using a template. Students who answered a minimum of three items correctly were eligible for participation in the L1 tests.

4.3 Assessment of Validation Criteria

For the validation analyses, we draw on a number of variables measured with student questionnaires, student competence tests and computer-assisted telephone interviews (CATIs) with students' parents.

Objective measure of L1 proficiency

As a concurrent validation criterion, proficiency in Russian and Turkish was tested in individual testing sessions with the Bilingual Verbal Ability Test, BVAT (Muñoz-Sandoval, Cummins, Alvarado, & Ruef, 1998). The goal of the BVAT is to capture bilingual participants' overall language ability — and to thereby avoid underestimating their linguistic capacity — by taking into account language proficiency in L1 and L2. More specifically, the BVAT starts with examining participants' proficiency in L2, typically English. If they fail to solve an item in L2, it is presented in the respective L1. The test results consequently reflect the additive language ability in L1 and L2. The BVAT is available for 17 languages besides English, among them Russian and Turkish. The target population of the test ranges from 5 years to old age. The instrument examines productive language proficiency and includes the four subscales picture vocabulary, synonyms, antonyms and verbal analogies. The test is adaptive by specifying starting items according to participants' age as well as termination criteria after a series of eight (picture vocabulary, verbal analogy) or six (synonyms, antonyms) unsolved items.

As we are specifically interested in students' L1 proficiency, we presented students only with items in Russian or Turkish. In addition, we refrained from adaptive testing and instead presented the same item set to all participants in order to obtain comparable test scores. Due to time constraints, we only employed the subscales picture vocabulary and synonyms. The picture vocabulary scale requests participants to name drawings of objects or activities,

[6] From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

while the synonyms scale requires an active production of synonyms for verbally presented words. We excluded the first eight items of the vocabulary subscale as we considered them too easy for the targeted age group. In total, we administered 51 items from the subscale vocabulary and 20 items from the subscale synonyms of the BVAT.

Trained test administrators, who were native speakers of Russian or Turkish, coded students' responses during the test session. The test sessions were recorded and 61% (Russian sample) or 66% (Turkish sample) of answers were additionally coded by two other native speakers. Inter-rater reliability was very high with Yules Y = .88 in the Russian sample and Y = .87 in the Turkish sample, confirming that the test administrators coded answers adequately.

To deviate as little as possible from the original BVAT, we kept items even if their discriminatory power was lower (.30 > d > 0) than would usually be considered acceptable for psychometric tests (Bortz & Döring, 2002). Only items with a discrimination ≤ 0 in the picture vocabulary scale were excluded from further analyses (five items in the Russian test; eight items in the Turkish test), leaving 66 items in the Russian test and 63 items in the Turkish test in total. The sum of the correctly answered items from both scales served as the validation criterion for our L1 proficiency tests.

The BVAT suffers from a number of conceptual and methodological limitations. Most importantly, the test versions in the various languages are simple translations of the English test. As a consequence, some of the items in the Russian and Turkish versions are not solvable. For instance, for some items in the synonyms scale no equivalent expression exists. The BVAT also contains a number of items which are highly dependent on cultural knowledge. The picture vocabulary scale, for instance, employs illustrations of men canning gold or of an American-style movie theater with the English word "Grand" on top. These items may be of reasonable difficulty for someone familiar with the history, culture and language of the United States of America, but they should be considerably more difficult for someone not possessing this cultural knowledge. What is more, some of the items are outdated, such as the illustration of a printing press or of an antiquated telephone, so that adolescents today are likely to be unfamiliar with the concept or unable to recognize the antiquated illustration. The BVAT also does not provide an exhaustive list of correct responses, such that coding objectivity is jeopardized. Despite its limitations, we decided to use the BVAT as a validation criterion for our L1 tests because no other instruments suitable for testing oral language proficiency in Russian and Turkish exist for the target population of our study.

General cognitive abilities

The NEPS examines *perceptual speed* and *reasoning* as two aspects of cognitive mechanics or indicators of basic information processing (Lang, Kamin, Rohr, Stünkel & Williger, 2014). Aspects of cognitive mechanics are generally assumed to operate relatively independent from content, culture and previous experience and to reflect the neurophysiological basis of the mind (Baltes, Staudinger, & Lindenberger, 1999). The *perceptual speed* test requires students to allocate numbers to symbols according to a provided key. The r*easoning* test consists of matrices similar to those of the RAVEN test (Raven, 1977). In our analyses, we use the sum of correct answers for each scale as ability estimates.

Subjective indicators of L1 proficiency

The validation analyses also draw on several subjective indicators of students' L1 proficiency. First, the student questionnaires measured self-reported L1 proficiency of students with another first language than German ("How good is your command of the other language?" [7]) for the dimensions of understanding, speaking, reading and writing. The 5-point response scale was "very good – rather good – rather poor – very poor – not at all." For the analyses, the arithmetic mean of students' ratings across the four dimensions was computed, resulting in the scale self-estimated global L1 proficiency, which could vary from 0 (not at all) to 4 (very good).

Second, we included two scales reflecting students' self-estimated L1 skills in oral comprehension and written production. These scales require students to estimate their ability with regard to specific examples of linguistic tasks. For instance, the oral comprehension scale assesses students' self-estimated ability to comprehend the temperature predicted for the next day in a radio weather forecast. In the written production scale, students were, for example, asked to estimate their ability to write a description of their activities in the last weekend. The items reflect the proficiency levels A1 to C1 as defined by the Common European Framework of Reference for Languages: Learning, Teaching, Assessment, CEFR (Council of Europe, 2011) and were developed based on a compilation of tasks students are expected to master at each proficiency level (Glaboniat, Müller, Rausch, Schmitz, & Wertenschlag, 2005). With these scales we wanted to check whether students are able to estimate their language proficiency more accurately based on detailed examples compared to the commonly used general dimensions understanding or writing. The oral comprehension scale contains 15 items and the written production scale contains 10 items. The answering scale was the same 5-point response scale measuring the self-estimated global L1 proficiency. We computed the mean of each scale and labeled them self-estimated oral comprehension (specific examples) and self-estimated written production (specific examples).

Students were additionally asked to estimate the number of items they had answered correctly in the L1 test. The self-estimated number of items solved is contingent on the total number of items in the test and could thus range from 0 - 328 (see Lockl, 2013 for further information on the assessment of procedural metacognition in the NEPS). We interpret this scale as another subjective indicator of L1 proficiency.

Parents' estimates of their children's L1 proficiency on the dimensions speaking and writing served as a fourth subjective indicator of students' L1 proficiency. The rating scale was the same as for the students' self-estimated global L1 proficiency, and the two dimensions were averaged.

[7] Before students reached this item, the questionnaire had defined "the other language." Specifically, students were asked to indicate the language other than German they had learned as a child in their family. Afterwards, they were informed that the questionnaire would subsequently refer to this language as "the other language."

[8] Students estimated the number of solved items on the basis of the 32 items originally included in the test of which one was later on eliminated due to poor model fit.

Language use with family and peers

Another set of items in the student questionnaires measured the patterns of language use in the family (with mother, with father, with siblings) and with peers (with best friend and with classmates) for students with another first language than German. An example for these questions is: "What language do you speak with your mother?" The 4-point response scale was "only German – mostly German, sometimes the other language – mostly the other language, sometimes German – only the other language." Students' ratings of language use with the mother and with the father were averaged. Similarly, the ratings of language use with the best friend and with classmates were combined into an indicator of the language used with peers.

Language of media use

The student questionnaires assessed the language of media consumption with seven items capturing the language in which, among other things, students read books, watch television or surf the web. The same 4-point scale as for language use with family and peers was employed. We averaged the seven items to a single indicator of language in media use.

Immigrant generation status and age at immigration

We defined the target persons' immigrant generation status based on the country of birth of the students themselves, their parents and grandparents. The student questionnaires assessed the countries of birth based on a list of countries including Germany, Kazakhstan, the Russian Federation, Ukraine and Turkey, as well as based on an open question allowing students to indicate countries not covered in the list. We classified students who themselves and their parents were born abroad as *first generation*, students who were born in Germany but whose parents were both born abroad as *second generation* and students who were born in Germany, whose parents were born in Germany and whose grandparents (at least two) were born abroad as *third generation*. We further defined students with one parent born abroad and one parent born in Germany as *one parent born abroad*. Students' immigrant generation status was only classified when all data necessary for its univocal identification were available. The student questionnaires also assessed the age which students who were born abroad immigrated to Germany.

Identification with heritage culture

Four items captured students' emotional identification with the heritage culture of their families. One item for instance states: "I feel closely attached to this culture[9]." The 4-point response scale was "does not apply – partially applies – mostly applies – fully applies."

While most validation criteria, particularly the information from the student questionnaires, were measured in both studies, the BVAT and the *self-estimated L1 proficiency based on detailed examples* were only administered in study 1 whereas parents' estimates of

[9] Before students reached the item, the questionnaire requested students to indicate the country other than Germany from which their family originates. Afterwards, it explained that subsequent questions would refer to the culture of this country as students' "heritage culture." An example for this was presented.

students' L1 proficiency, the *self-estimated number of items solved in the L1 test* as well as general cognitive abilities were only included in study 2.

4.4 Sample

Both studies tested L1 proficiency of immigrant students (first, second and third generation) whose families immigrated to Germany from the area of the Former Soviet Union (e.g., Russia, Kazakhstan) or Turkey.[10] More specifically, we defined the target population as students who were either themselves born in one of these countries and students with at least one parent or two grandparents born in the area of the Former Soviet Union or Turkey. Table 1 presents descriptive information on the samples of studies 1 and 2.

Study 1 is based on data from schools located in four federal states (Bavaria, Berlin, Hamburg, North Rhine-Westphalia) attended by high percentages of students speaking Turkish and/or Russian. The Russian L1 test was conducted in 17 schools and the Turkish L1 test in 15 schools.

The Russian sample consists of 224 participants[11] (53% female). On average, students were 16 years old at the time of data assessment. Of these students, 37% were enrolled in a *Hauptschule* (lowest school track of the German secondary education*),* 31% in a *Gesamtschule* (*comprehensive schools*)*,* 17% in a *Schule mit mehreren Bildungsgängen* (schools with several educational tracks) and 15% in a *Gymnasium* (highest track leading to university entrance degree). More than two-thirds (71%) were classified as first generation immigrant students, while only 20% were second generation.

The Turkish sample in study 1 consists of 310 participants[12] (50% female). Participants' mean age was 15.7 years. The majority of students either attended the *Hauptschule* (28%) or the *Gesamtschule* (31%), while a somewhat lower proportion of the students attended a *Schule mit mehreren Bildungsgängen (20%)* or a *Gymnasium* (22%). The majority of students in the Turkish sample (66%) were second generation, while only 7% were first generation immigrants.

A subsample of 79 participants in the Russian sample and of 101 participants in the Turkish sample completed the BVAT.

Study 2 draws on data from schools located in all federal states of Germany. The original Russian sample comprised 743 students and the original Turkish sample 916 students. Not all selected students attended the screening tests, however (see IEA Data Processing and Research Center, 2013): 129 (17%) of the Russian sample and 152 (17%) of the Turkish sample were absent from school at the testing day. In addition, 48 (6%) Russian students and 23 (3%) Turkish students refused to participate in the screening test, due to a complete lack of proficiency in L1. Furthermore, 10 students of the Russian sample (1%) and seven students of the Turkish sample (1%) had left the school in the meantime. The schools

[10] For the sake of brevity, the former group will further be referred to as "the Russian sample" and the latter as "the Turkish sample." These terms do not allude to citizenship or the like, but to the L1 that was tested in the sample.

[11] Overall, 225 students participated in the Russian L1 test. We excluded one student because the number of valid answers provided was too small to scale the test score.

[12] In total, the Turkish sample consisted of 311 students. One student did not provide enough valid answers and therefore had to be excluded from the scaling procedure.

attended by seven students from the Russian sample (1%) and by 29 students (3%) from the Turkish sample did not participate in the NEPS study anymore. The Russian L1 tests were administered in 257 schools and the Turkish L1 tests in 237 schools. 35 students (5%) failed the screening test and were thus excluded from the L1 testing session in the Russian sample. In the Turkish sample, 38 students (4%) failed the screening test. We further excluded 12 (2%) Russian and 4 (0.5%) Turkish students who had participated in the L1 tests from all analyses because the number of valid answers they provided was too small to scale their test scores (see Edele, et al., 2012 for further information on the scaling of the L1 tests).

The resulting Russian sample includes 502 students in total (51% female). On average, they were 15.8 years old at the time of testing. The largest proportion of students attended the *Hauptschule* (41%), followed by students attending the *Realschule* (25%) and the *Gymnasium* (20%). Almost equal proportions of students belonged to the first immigrant generation (47%) and to the second generation (41%). On average, the students born abroad were 5.3 years old when they came to Germany.

The final Turkish sample consists of 662 students (48% female). On average, students in this sample were 15.7 years old. Half of them attended the *Hauptschule* (50%), 20% the *Realschule* and 14% the *Gymnasium*. In this sample, the majority of students (65%) were second generation immigrants; only 9% were first generation.

Table 1: Gender, Age, Attended School Track and Immigrant Generation Status of the Russian and the Turkish Samples in Studies 1 and 2

	Study 1				**Study 2**			
	Russian		**Turkish**		**Russian**		**Turkish**	
	n	***%***	***n***	***%***	***n***	***%***	***n***	***%***
Total	224		310		502		662	
Gender								
Male	106	47.3	155	50.0	248	49.4	342	51.7
Female	118	52.7	155	50.0	254	50.6	320	48.3
School track								
Hauptschule	82	36.6	87	28.1	206	41.0	330	49.8
Realschule	-	-	-	-	124	24.7	130	19.6
Gymnasium	34	15.2	67	21.6	98	19.5	94	14.2
SMB	38	17.0	61	19.7	27	5.4	11	1.7
Gesamtschule	70	31.2	95	30.6	47	9.4	97	14.7

Immigrant generation								
1st generation	159	71.0	23	7.4	234	46.6	62	9.4
2nd generation	44	19.6	205	66.1	206	41.0	430	64.9
3rd generation	-	-	3	1.0	-	-	19	2.9
One parent born abroad	6	2.7	52	16.8	28	5.6	132	19.9
Not determinable	15	6.7	27	8.7	34	6.8	19	2.9

Note. SMB = Schule mit mehreren Bildungsgängen (schools with several educational tracks)

4.5 Results

We analyzed the data from the Russian and the Turkish samples separately, as these groups differ with regard to several important characteristics, such as the proportion of first and second immigrant generation students and the attended school types.

Scaling, item difficulty and reliability

The L1 test scores of both studies were scaled based on a one-parameter logistic model (Rasch model), yielding estimates of students' L1 proficiency as Weighted Maximum Likelihood Estimates (WLEs). Item fit indices are generally satisfying in study 1[13] and in study 2 (see Edele et al., 2012 for further details on scaling).

In study 1, item difficulty on a logit scale of the Russian test ranges from -2.01 to 0.61 (mean item difficulty: b = -.57). The mean of students' proficiency estimates given as WLEs is M_{WLE} = 0.02 ranging from -2.57 to 3.83. In the Turkish test, the mean item difficulty is b = -0.17 ranging from -2.61 to 1.23. Person estimates of Turkish proficiency range between -2.24 and 4.26 with M_{WLE} = 0.01.

In study 2, the mean item difficulty in the Russian L1 test is b = -0.12 ranging from -1.48 to 1.41. The mean of the person estimates is M_{WLE} = .04, ranging from -2.78 to 4.26. The average item difficulty in the Turkish L1 test is b = -0.23, ranging from -1.78 to 1.30. The person estimates range from -3.46 to 3.00 with M_{WLE} = 0.00.

The concurrence of the means in item difficulty and in person estimates suggests that item difficulty and students' L1 proficiency in both studies match well, on average. The range of item difficulties is somewhat limited, however. Graphical analyses of the joint distributions of person estimates and item estimates further indicate that in both studies the majority of items clusters around the center of the scale (medium difficulty), while the boundaries of the scale are covered only by a few items. The tests thus capture very high and very low proficiency levels somewhat less precisely than the proficiency of students at intermediate proficiency levels.

[13] In study 1, one item in each test showed a poor fit (WMNSQ = 1.33 in the Russian test; WMNSQ = 1.21 in the Turkish test), most probably due to attractive distractors. These distractors were eliminated in the final test version.

Yet the tests proved to be highly reliable. In study 1, the WLE-reliability was .86 for the Russian test and .79 for the Turkish test. In study 2, reliability coefficients were .85 for the Russian test and .83 for the Turkish test.

Measurement equivalence

To test measurement equivalence, we conducted a MGCFA on the L1 tests in study 2 (see Table 2). The results show that the fit indices of the model assuming configural invariance are acceptable (see Yu, 2002, for cut-off criteria of fit indices). The more restrictive models assuming strong and strict invariance, however, do not hold, as the model fit indices are not satisfactory and the test of change in model fit is significant.

To check whether the observed lack of invariance is caused by items with low factor loadings in one of the tests (loadings < .40) we excluded seven items from the model. The resulting model assuming configural invariance (CFI = .94; TLI = .96; RMSEA = .04), however, does not fit better than the model including the complete set of items. The more restrictive models, consequently, do not show satisfying fit indices either. Thus, there is no indication for partial strong invariance, either.

These findings confirm configural equivalence of the Russian and the Turkish tests, implying that they measure the same construct. Because more restrictive models of equivalence are not supported, however, neither the ability scores in the Turkish and the Russian tests nor their correlation coefficients with other variables are directly comparable.

Table 2: Tests of Measurement Equivalence of the Russian and the Turkish L1 Tests

	χ2/*df*	***p***	**CFI**	**TLI**	**RMSEA**	**Δχ2/ Δdf**	***p***
Configural equivalence	1315.8/868	.00	.95	.95	.03		
Strong equivalence	1714.0/898	.00	.91	.91	.04	367.72/30	.00
Strict equivalence	1775.3/929	.00	.91	.91	.04	108.39/31	.00

Note. In computing these models with MPlus 6.1 (Muthén & Muthén, 2009), we employed a robust weighted least squares approach (estimator: WLSMV) and estimated model parameters based on Theta parameterization. CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = root mean square error of approximation.

Dimensional structure

To examine whether our L1 tests exhibit the expected unidimensional structure, a 1-dimensional model was tested against an alternative, theoretically plausible 2-dimensional model. The 2-dimensional model assigns items on dialogues to the first dimension and items on expository as well as narrative texts to the second dimension. We computed these models with ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007), using Marginal Maximum Likelihood (MML) estimation with Gauss-Hermite quadrature.

In study 1, the two dimensions correlate very highly, with .97 in the Russian sample and .94 in the Turkish sample. The 1-dimensional model fits the data better than the 2-dimensional model in both language groups, as two indicators of model fit, namely Akaike's information

criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978), indicate (see Table 3).

The two dimensions are also highly correlated in study 2, with .94 in the Russian sample and .98 in the Turkish sample. The model fit indices suggest that the 2-dimensional model fits negligibly better according to the AIC and slightly more poorly according to the BIC in the Russian sample. In the Turkish sample, the 2-dimensional model fits somewhat more poorly as indicated by both indicators (see Table 3). The very high correlation between the two dimensions indicating their near identity and the very similar model fit slightly in favor of the 1-dimensional model suggest that the construct measured with the L1 tests is unidimensional rather than 2-dimensional.

Table 3: Results of the 1-dimensional and 2-dimensional Scaling: AIC and BIC Model Selection Criteria

	Study 1				Study 2			
	Russian Sample		Turkish Sample		Russian Sample		Turkish Sample	
	1-dim	*2-dim*	*1-dim*	*2-dim*	*1-dim*	*2-dim*	*1-dim*	*2-dim*
AIC	7720.9	7724.0	11466.2	11468.7	19201.5	19196.3	25767.9	25772.1
AIC-Diff	3.1		2.5		-5.1		4.2	
BIC	7830.1	7840.0	11585.8	11595.7	19340.7	19344.0	25916.2	25929.4
BIC-Diff	9.9		9.9		3.3		13.2	

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion.

Criterion validity: Other indicators of L1 proficiency and reasoning

To examine the convergent validity of the L1 tests, we correlated in a first step the L1 test scores with students' scores on the BVAT. In the Russian sample, the two objective indicators correlate quite highly, in the Turkish sample the correlation is moderate (see Table 4).

In a second step, we correlated immigrant students' L1 test scores with a number of subjective indicators of L1 proficiency. As expected, students' self-estimated global L1 proficiency is positively related to their L1 test score. Along the same lines, the scales self-estimated oral comprehension (specific examples) and self-estimated written production (specific examples) also substantially relate to tested L1 proficiency in both language groups. These correlations are, however, not higher than those between the L1 tests and the global self-estimates, indicating that an assessment of L1 proficiency based on concrete examples does not enhance the validity of self-estimates. Students' self-estimated number of items solved in the test is also substantially related to their results in the L1 tests; strong correlations emerged in both language groups. Similarly, parents' estimates of their children's L1 proficiency are also strongly (Russian sample) or moderately (Turkish sample) associated with the L1 test scores.

Analyses exploring the discriminant validity of the L1 tests indicate that, as expected, the L1 test scores correlate moderately with reasoning but only weakly and inconsistently with perceptual speed.

Table 4: Pairwise Correlations between Immigrant Students' L1 Test Scores (WLEs) and Validation Criteria

	L1 test score (WLE)			
	Study 1		Study 2	
	Russian	Turkish	Russian	Turkish
BVAT	.72*** (79)	.41*** (101)	-	-
Subjective measures of L1 proficiency				
Self-estimated global L1 proficiency	0.50*** (220)	0.28*** (302)	0.43*** (426)	0.26*** (572)
Self-estimated oral comprehension (specific examples)	0.49*** (223)	0.32*** (306)	-	-
Self-estimated written production (specific examples)	0.43*** (223)	0.18** (305)	-	-
Self-estimated number of solved items	-	-	0.59*** (488)	0.54*** (636)
Parents' estimates of L1 proficiency	-	-	0.57*** (182)	0.40*** (232)
General cognitive abilities				
Perceptual speed	-	-	0.11* (483)	0.03 (632)
Reasoning	-	-	0.16*** (482)	0.28*** (628)
Age at immigration	0.49*** (140)	0.04 (14)	0.34*** (263)	0.30* (48)
Language use				
with parents	0.38*** (224)	0.09 (305)	0.31*** (437)	0.14*** (581)
with siblings	0.27*** (205)	0.07 (293)	0.31*** (399)	0.17*** (545)
with peers	0.34*** (223)	0.18** (304)	0.10* (437)	0.04 (572)
in media use	0.46*** (222)	0.23*** (304)	0.28*** (420)	0.23*** (559)
Identification with heritage culture	0.12 (212)	0.17** (290)	0.17*** (465)	0.15*** (566)

Note. Correlations are given as Pearson's r. Spearman's rank correlation coefficients which we additionally computed because of the non-normal distribution of some validation criteria, yielded almost equal results. Number of cases (N) in parentheses.

*p < .05, **p < .01, ***p < .001

Criterion validity: L1 exposure and motivation

Further analyses of the L1 tests' validity with indicators of exposure to L1 and motivation for L1 acquisition as criteria generally also show the expected pattern of results. In the Russian group, students' L1 test scores in both studies correlate substantially with their age at immigration as well as with the language they use with parents, siblings, peers and in media consumption. In the Turkish group, the age at immigration is also positively related to L1 test scores in study 2 but not in study 1. Yet, the coefficient in study 1 is based on a very small number of students and may therefore not be reliable. The language of media use also shows the expected correlation with L1 test scores in the Turkish group. Language use in the family and with peers overall shows the expected pattern as well, yet the correlation coefficients are somewhat smaller and less consistent in the Turkish group than in the Russian sample.

In a next step, we examined whether the duration of the family's residence in Germany is associated with the L1 test scores. Because the number of third-generation students is very small, the analyses focus on the first and second generations. As expected, the test scores in Russian are higher for first immigrant generation than for second generation students (see Table 5).

Contrary to expectations, in the Turkish sample the first generation does not show higher L1 test scores than the second generation. One possible explanation for this finding is that second generation students with a Turkish background may not encounter significantly fewer learning opportunities for L1 acquisition than first generation students who immigrated from Turkey themselves. To test this idea, we compared first and second generation students' language use with parents, siblings, peers and in media with data from study 2. The results of the t-tests support this explanation to some extent. Students from the first generation of the Turkish group use L1 more often with their parents (M = 3.1) than students from the second generation (M = 2.9, t = 2.38, p <.05). However, there is no evidence for differences between the first and second generation in the usage of Turkish with siblings (M1st = 2.1; M2nd = 1.9, t = 0.88, n.s.), with peers (M1st = 1.7; M2nd = 1.6, t = 0.84, n.s.) or in media consumption (M1st = 1.8; M2nd = 1.7, t = 1.06, n.s.). In the Russian sample, by comparison, the differences in language use between the first and second generation are more consistent and somewhat more pronounced (with parents: M1st = 3.0, M2nd = 2.4, t = 8.03, p <.001; with siblings: M1st = 2.0 ; M2nd = 1.6 , t = 3.81, p <.001; with peers: M1st = 1.5; M2nd = 1.4, t = 1.85, n.s.; in media use: M1st = 1.5; M2nd = 1.4, t = 1.51, n.s).

In addition, we expected that students' identification with the heritage culture should motivate them for L1 acquisition and should consequently correlate with L1 test scores. The expected pattern emerges in both groups, although the coefficients are rather small (see Table 4).

Table 5: L1 Test Scores (WLEs) by Immigrant Generation Status

	Study 1				Study 2			
	First generation		Second generation		First generation		Second generation	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Mean L1 test score Russian	.31	.10	-.81***	.17	.39	.08	-.34***	.07
Mean L1 test score Turkish	.27	.23	.09	.06	-.19	.13	.07	.05

Note. Differences between the first and second immigrant generation groups were tested separately within the Russian and the Turkish samples with Mann-Whitney U-tests.

*** significant difference between first and second immigrant generation ($p < .001$).

5. Discussion

The analyses presented in this article confirm that the L1 listening comprehension tests developed for the NEPS are valid measures of 9th grade students' proficiency in Turkish and Russian. The L1 tests showed convergent validity as evidenced by correlations with another L1 test (BVAT), as well as with a number of subjective estimations of L1 proficiency. As expected, the correlations of our tests with the subjective indicators were somewhat weaker than those with the BVAT, suggesting once more that the subjective proficiency estimates are biased (see also Edele et al., in press; Finnie & Meng, 2005).

The correlations of our tests with the BVAT were more substantial. Yet, given that both instruments are objective tests of proficiency in Russian or Turkish, even higher correlations could have been expected, particularly in the Turkish group for which we observed only a moderate correlation. However, unlike our L1 tests, the BVAT assesses productive language proficiency and examines a linguistic subcomponent (vocabulary). In addition, the BVAT suffers from a number of conceptual and psychometric limitations. These factors most likely limited the correlations with our tests.

Further validation analyses using indicators of exposure to L1 and motivation to acquire L1 as criteria generally also yielded the expected pattern of results, thereby corroborating the tests' validity. In the Turkish group, however, some criterion variables did not show the expected correlations with the L1 test scores. In particular, the first generation did not score higher on the Turkish test than the second generation. This could indicate that the construct validity of the Turkish test is limited. However, the Turkish test was correlated with various other validation criteria in the predicted way, particularly with other subjective as well as objective indicators of Turkish proficiency. The lack of significant relationships with some of the criteria may therefore rather indicate that some of our theoretical assumptions do not fully apply to the Turkish group. Indeed, we found some evidence that second generation students of Turkish origin may not have significantly fewer opportunities for L1 acquisition than first generation students, suggesting that this group was more reluctant to give up the

use of Turkish than assumed by the models of language acquisition (e.g., Chiswick & Miller, 2001).

In general, the correlations of our L1 tests with the criterion variables were higher for the Russian than for the Turkish test. As strongly measurement equivalence could not be confirmed for the two tests, however, these correlations are not directly comparable.

Our L1 tests assess listening comprehension, which allows their administration to students at practically all levels of L1 proficiency, including those with low skills in reading and writing. Moreover, the tests assess students' L1 proficiency efficiently, yet comprehensively. To further extend the analytical potential of the NEPS, it would be interesting to develop tests of students' reading and writing proficiency in Russian and Turkish, which may be particularly relevant for certain outcomes like labor market success. Tests assessing linguistic subcomponents, such as grammar or vocabulary, as well as tests of other L1s frequently spoken by immigrants in Germany, such as Polish, would further complement the data base.

Taken together, our tests present valid measures of an important aspect of L1 proficiency. The instruments will facilitate research on the effects of L1 proficiency on immigrant students' educational development and other outcome variables in a methodologically more sound way than was possible in the past.

References

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*(2), 207-249. doi: 10.3102/0034654310368803

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

Anderson, J. R. (1995). *Cognitive psychology and its implications*. New York: Freeman.

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., … Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaften, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 51-65). Wiesbaden: VS Verlag für Sozialwissenschaften.

Baltes, P. B., Staudinger, U. M., & Lindenberger, U. (1999). Lifespan psychology: theory and application to intellectual functioning. *Annual Review of Psychology, 50*, 471-507.

Berry, J. W., Phinney, J. S., Sam, D. L., & Vedder, P. (2006a). *Immigrant youth in cultural transition: Acculturation, identity, and adaptation across national contexts*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Berry, J. W., Phinney, J. S., Sam, D. L., & Vedder, P. (2006b). Immigrant youth: Acculturation, identity, and adaptation. *Applied Psychology: An international review, 55*(3), 303-332. doi: 10.1111/j.1464-0597.2006.00256.x

Bialystok, E. (2007). Cognitive effects of bilingualism: How linguistic experience leads to cognitive change. *The International Journal of Bilingual Education and Bilingualism, 10*(3), 210-223.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft, 14*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: need, main features, and research potential. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaften, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 5-17). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bortz, J. & Döring, N. (2002). *Forschungsmethoden und Evaluation: für Human- und Sozialwissenschaftler*. Heidelberg: Springer.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the Multitrait-Multimethod-Matrix. *Psychological Bulletin, 56*(2), 81-105.

Chiswick, B. R. & Miller, P. W. (1994). Language choice among immigrants in a multi-lingual destination. *Journal of Population Economics, 7*, 119-131.

Chiswick, B. R. & Miller, P. W. (2001). A model of destination language acquisition: Application to male immigrants in Canada. *Journal of Political Economy, 38*(3), 391-409.

Council of Europe (2011). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Council of Europe.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire* (Books Opinion Papers No. ED446442). Clevedeon: Multilingual Matters Ltd., UTP.

Dollmann, J. & Kristen, C. (2010). Herkunftssprache als Ressource für den Schulerfolg? – Das Beispiel türkischer Grundschulkinder. In C. Allemann-Ghionda, P. Stanat, K. Göbel & C. Röhner (Eds.), *Zeitschrift für Pädagogik, 55. Beiheft. Migration, Identität Sprache und Bildungserfolg* (pp. 123-146). Weinheim: Beltz.

Duursma, E., Romero-Contreras, S., Szuber, A., Proctor, P., Snow, C., August, D., & Calderón, M. (2007). The role of home literacy and language environment on bilinguals' English and Spanish vocabulary development. Applied Psycholinguistics, 28, 171-190. doi: 10.1017/S0142716406070093

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: University of Bamberg, National Educational Panel Study.

Edele, A., Seuring, J., Kristen, C., & Stanat, P. (in press). Why bother with testing? The validity of immigrants' self-assessed language proficiency. *Social Science Research.* doi: http://dx.doi.org/10.1016/j.ssresearch.2014.12.017

Edele, A. & Stanat, P. (submitted). *The Role of First-Language Listening Comprehension in Second-Language Reading Comprehension*. Manuscript submitted for publication.

Edele, A., Stanat, P., Radmann, S., & Segeritz, M. (2013). Kulturelle Identität und Lesekompetenz von Jugendlichen aus eingewanderten Familien. In N. Jude & E. Klieme (Eds.), *Zeitschrift für Pädagogik, 59. Beiheft. PISA 2009 – Impulse für die Schul- und Unterrichtsforschung.* Weinheim: Beltz Juventa.

Esser, H. (2006). *Sprache und Integration: Die sozialen Bedingungen und Folgen des Spracherwerbs von Migranten*. Frankfurt am Main: Campus Verlag.

Finnie, R. & Meng, R. (2005). Literacy and labour market income: self-assessment versus test score measures. *Applied Economics* 37, 1935-1951.

Flowerdew, J. & Miller, L. (2005). *Second language listening: Theory and practice*. Cambridge: Cambridge University Press.

Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, I., ... Kanders, M. (2011). Transition and development from lower secondary to upper secondary school. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaften, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 217-232). Wiesbaden: VS Verlag für Sozialwissenschaften.

Frey, A., Taskinen, P., Schütte, K., Prenzel, M., Artelt, C., Baumert, J., ... Pekrun, R. (Eds.). (2009). *PISA 2006: Skalenhandbuch. Dokumentation der Erhebungsinstrumente*. Münster: Waxmann.

Glaboniat, M., Müller, M., Rausch, P., Schmitz, H., & Wertenschlag, L. (2005). *Profile deutsch: Gemeinsamer europäischer Referenzrahmen: Lernzielbestimmungen, Kannbeschreibungen, Kommunikative Mittel, Niveau A1-A2 B1-B2 C1-C2*. Berlin: Langenscheidt KG.

Grotjahn, R. (2005). Testen und Bewerten des Hörverstehens. In M. Ó. Dúill, R. Zahn & K. D. C. Höppner (Eds.), *Zusammenarbeiten: Eine Festschrift für Bernd Voss* (pp. 115-144). Bochum: AKS-Verlag.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill.

Hesse, H.-G., Göbel, K., & Hartig, J. (2008). Sprachliche Kompetenzen von mehrsprachigen Jugendlichen und Jugendlichen nicht-deutscher Erstsprache. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* (pp. 208-201). Weinheim: Beltz.

IEA Data Processing and Research Center (2013). *Methodenbericht NEPS Startkohorte 4: Haupterhebung – Frühjahr 2011 A47, A68, A84.* Hamburg: IEA Data Processing and Research Center.

Lado, R. (1961). *Language testing: The construction and the use of foreign language tests*. London: Longman.

Lang, F. R., Kamin, S., Rohr, M., Stünkel, C. & Williger, B. (2014). *Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen des Nationalen Bildungspanels: Abschlussbericht zu einer NEPS-Ergänzungsstudie.* (NEPS Working Paper No. 43). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Leseman, P. P. M., Scheele, A. F., Mayo, A. Y., & Messer, M. H. (2009). Bilingual development in early childhood and the language used at home: competition for scarce resources? In I. Gogolin & U. Neumann (Eds.), *Streitfall Zweisprachigkeit - The Bilingualism Controversy* (Vol. 1, pp. 317-331). Wiesbaden: VS Verlag für Sozialwissenschaften.

Lockl, K. (2013). *Assessment of procedural metacognition: Scientific Use File 2013.* Bamberg: University of Bamberg, National Educational Panel Study.

Mouw, T. & Xie, Y. (1999). Bilingualism and the academic achievement of first- and second-generation Asian Americans: Accommodation with or without assimilation? *American Sociological Review, 64*, 232-252.

Muñoz-Sandoval, A. F., Cummins, J., Alvarado, C. G., & Ruef, M. L. (1998). *Bilingual Verbal Ability Tests, Comprehensive Manual.* Chicago: Riverside Publishing.

Muthén, L. K. & Muthén, B. O. (2009). *Mplus (Version 6.1)* [Computer software]. Los Angeles, CA: Muthén & Muthén.

Portes, A. & Rumbaut, R. G. (2012). *Children of Immigrants Longitudinal Study (CILS), 1991-2006*. ICPSR20520-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research (ICPSR) [distributor]. doi: 10.3886/ICPSR20520.v2

Rauch, D., Jurecka, A., & Hesse, H.-G. (2010). Für den Drittspracherwerb zählt auch die Lesekompetenz in der Herkunftssprache. In C. Allemann-Ghionda, P. Stanat, K. Göbel & C. Röhner (Eds.), *Zeitschrift für Pädagogik, 55. Beiheft. Migration, Identität Sprache und Bildungserfolg* (pp. 78-100). Weinheim: Beltz.

Raven, J. C. (1977). Standard Progressive Matrices: Sets A, B, C, D & E. San Antonio, TX: Harcourt.

Rumbaut, A. G. (2004). Ages, life stages, and generational cohorts: Decomposing the immigrant first and second generations in the United States. *International Migration Review, 38*(3), 1160-1205.

Rumbaut, R. G., Massey, D. S., & Bean, F. D. (2006). Linguistic life expectancies: Immigrant language retention in southern Calicornia. *Population and Development Review, 32*(2), 447-460.

Scheele, A. F., Leseman, P. M., & Mayo, A. Y. (2010). The home language environment of monolingual and bilingual children and their language proficiency. *Applied Psycholinguistics, 31*, 117–140.

Schroeders, U. & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71*(5), 849-869. doi: 10.1177/0013164410391468

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

Shohamy, E. & Inbar, O. (1991). Validation of listening comprehension tests: the effect of text and question type. *Language Testing, 8*(1), 23-40.

Spolsky, B. (1989). *Conditions for second language learning: Introduction to a general theory*. Oxford: Oxford University Press.

Stanat, P., Rauch, D., & Segeritz, M. (2010). Schülerinnen und Schüler mit Migrationshintergrund. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, … P. Stanat (Eds.), *PISA 2009: Bilanz nach einem Jahrzehnt* (pp. 200-230). Münster: Waxmann.

Steffensen, M., Joag-Dev, C., & Anderson, R. (1979). A cross-cultural perspective on reading comprehension. *Reading Research Quarterly, 15*, 10-29.

TNS Infratest Sozialforschung (2012). *SOEP 2008 – Erhebungsinstrumente 2008 (Welle 25) des Sozio-ökonomischen Panels*. (SOEP Survey Papers No 109: Series A). Berlin: Deutsches Institut für Wirtschaftsforschung.

Verhoeven, L. T. (2007). Early bilingualism, language transfer and phonological awareness. *Applied Psycholinguistics, 28*, 425-429. doi: 10.1017.S0142716407070233

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: University of Bamberg, National Educational Panel Study.

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.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* [Doctoral dissertation, University of California]. Retrieved from: http://www.statmodel.com/download/Yudissertation.pdf