languages

MDPI

Article

Evaluating the Russian Language Proficiency of Bilingual and Second Language Learners of Russian

Tatiana Luchkina [1,*], Tania Ionin [2,*], Natalia Lysenko [3], Anastasia Stoops [4] and Nadezhda Suvorkina [5]

1 Department of Linguistics, Stony Brook University, Stony Brook, NY 11794, USA
2 Department of Linguistics, University of Illinois at Urbana-Champaign, Champaign, IL 61801, USA
3 Department of Linguistics and Humanities, Oryol State Agrarian University named after N.V. Parahin, 302019 Oryol, Russia; n.lysenko@inbox.ru
4 Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA; agusico2@illinois.edu
5 Mezensky Pedagogical College, 302531 Oryol, Russia; vertkar@gmail.com
* Correspondence: tatiana.luchkina@stonybrook.edu (T.L.); tionin@illinois.edu (T.I.)

Abstract: The starting point of most experimental and clinical examinations of bilingual language development is the choice of the measure of participants' proficiency, which affects the interpretation of experimental findings and has pedagogical and clinical implications. Recent work on heritage and L2 acquisition of Russian used varying proficiency assessment tools, including elicited production, vocabulary recognition, and in-house measures. Using such different approaches to proficiency assessment is problematic if one seeks a coherent vision of bilingual speaker competence at different acquisition stages. The aim of the present study is to provide a suite of validated bilingual assessment materials designed to evaluate the language proficiency speakers of Russian as a second or heritage language. The materials include an adaptation of a normed language background questionnaire (Leap-Q), a battery of participant-reported proficiency measures, and a normed cloze deletion test. We offer two response formats in combination with two distinct scoring methods in order to make the testing materials suited for bilingual Russian speakers who self-assess as (semi-) proficient as well as for those whose bilingualism is incipient, or declining due to language attrition. Data from 52 baseline speakers and 503 speakers of Russian who reported dominant proficiency in a different language are analyzed for test validation purposes. Obtained measures of internal and external validity provide evidence that the cloze deletion test reported in this study reliably discriminates between dissimilar target language attainment levels in diverse populations of bilingual and multilingual Russian speakers.

Keywords: Russian; proficiency; cloze test; heritage speakers; L2 learners

Citation: Luchkina, Tatiana, Tania Ionin, Natalia Lysenko, Anastasia Stoops, and Nadezhda Suvorkina. 2021. Evaluating the Russian Language Proficiency of Bilingual and Second Language Learners of Russian. *Languages* 6: 83. https://doi.org/10.3390/languages6020083

Academic Editors: Gita Martohardjono and Jennifer Chard

Received: 23 February 2021
Accepted: 27 April 2021
Published: 11 May 2021

1. Introduction

According to the US Census bureau (U.S. Census Bureau 2019), in the year 2018, an estimated 67.3 million US residents spoke a language other than English at home. This is more than three times the number of the multilingual speakers reported by the US census in the 1980s. As the number of multilingual speakers continues to grow in the exceedingly globalized world, proficiency levels of the modern-day language users deserve careful attention from researchers and educators. Language proficiency is fundamentally important for understanding various aspects of language use, including perception, production and comprehensibility (Lemhöfer and Broersma 2012; Tremblay 2011). Quantitative and qualitative variability in second language (L2) neurocognitive activity (Abutalebi 2008; Kotz 2009), spoken word recognition and phonological processing (Blumenfeld and Marian 2007), sentence level and discourse-level comprehension (Foucart et al. 2016; Van Zeeland and Schmitt 2013) are influenced, to a large extent, by speaker's competence in the target

language (henceforth, TL). Therefore, language proficiency must be taken into consideration when seeking an accurate understanding of language acquisition and bilingualism (Hulstijn 2012).

While the challenge of accounting for bilinguals' and multilinguals' proficiency arises along with the ever-increasing role that these populations play in research and applied contexts, it is further complicated by the fact that unlike in typically developing monolingual speakers, language proficiency of bilinguals and multi-linguals is not only highly variable (Wood Bowden 2016), but also rather dynamic, and may be subject to effects of cross-linguistic influence, language dominance, and language attrition (Kohnert and Bates 2002; Luk and Bialystok 2013; Montrul 2018; Montrul et al. 2008; Montrul and Ionin 2012; White 2003).

Among the non-English languages spoken in the US, the Russian language had the largest proportional increase from 1990 to 2000 (Gildersleeve-Neumann and Wright 2010). The 2017 report by the Census bureau lists Russian as the 9th most spoken language in the US. In promoting the bilingual development of the Russian-English speaking population, both clinicians and researchers must commit to best practices in assessing and educating these bilingual language learners. Despite the recent growth in the population of the Russian speakers, Russian remains a Less Commonly Taught Language (LCTL), which means that its study receives support as one division of foreign language, area, and international studies in US colleges and universities. The need for accurate yet efficient proficiency assessment instruments to be used with Russian learners has become more real ever since Russia reintroduced proficiency requirements for citizenship applicants and international students. Domestically, this has given rise to investigations of assessment centered around the use of comprehensive standardized tests of Russian proficiency (e.g., Belyakova et al. 2013; Basenko-Karmali and Saparova 2020). The Test of Russian as a Foreign Language (TORFL) and the American Council on the Teaching of Foreign Languages (ACTFL) present the two best known commercial standardized tests of Russian proficiency and require specially trained staff to administer testing, and score and evaluate the outcome. At the same time, a lot of research on the acquisition of Russian by bilingual and multilingual populations, past and ongoing, is carried out independently of these recent developments in the formal proficiency assessment arena and demonstrates a shortage of adequately accessible, normed proficiency measures to be used with L2 learners and heritage speakers residing outside of Russia.

To date, to the best of our knowledge, the only peer-reviewed published studies addressing bilingual speakers' proficiency in Russian as a home or second language include Makarova and Terekhova (2017) and Long et al. (2012). Makarova and Terekhova (2017) evaluated the oral proficiency in Russian in a group of 30 5–6-year-old Russian speaking bilinguals and multilinguals all residing in Canada, using an oral elicited production task. Long et al. (2012) presents a comprehensive attempt to deconstruct proficiency levels specified under the Interagency Roundtable Language Scale using data from 68 adult L2 learners and heritage Russian bilinguals tested on over thirty computer-delivered perception and production tasks. The aim of the present study is to contribute to the emergent literature on assessment in Russian and present the validation results of a test designed to evaluate the Russian language proficiency in adult bilingual and multilingual speakers, including those for whom Russian presents a home language or a target language learned via formal instruction. While these populations may differ substantially in terms of the age of exposure to Russian and the context in which the language is acquired, they may be equally subject to incomplete acquisition or attrition of Russian proficiency (Polinsky and Kagan 2007).

The assessment materials presented in this study include an adaptation of a normed language background questionnaire developed by Marian et al. (2007) and a normed cloze deletion test with two distinct response formats (constructed response and multiple choice) and two scoring methods (acceptable answer and exact answer) suited for participants with different attainment levels. Unlike many commercial standardized proficiency tests, the

testing tool we offer is not resource-intensive and affords straightforward and transparent scoring of the test data. In what follows, we report the psychometric properties of the test and explore the relationship between a testing outcome and participant acquisition history, including course level, the age of exposure to the TL, and the context of acquisition. The proposed test achieves satisfactory reliability, internal and external, criterion-based validity, and serves as a robust measure of Russian language skills while successfully meeting the following assessment and scoring criteria:

1. Presents an economical and methodologically simple proficiency assessment solution which researchers and language instructors can use as a unified proficiency gauge for research and/or class placement purposes.
2. Is independent of facts about individual learners' L2 background traditionally used as a proxy of learners' proficiency, including the age of exposure to the TL, and the amount of TL input or semesters of TL instruction.
3. Two normed versions of the cloze deletion test offer flexible testing logistics and provide straightforward scoring guidelines; the multiple-choice version of the test enables automated scoring so as to minimize the impact of the assessor on testing outcomes.

We present analyses of proficiency assessment data from 52 native Russian speakers and 503 Russian speakers and L2 learners with dominant proficiency in a different language and representative of varied learning contexts and language backgrounds. Test takers' data are used to validate the proposed test and demonstrate the application of each scoring method to gauging the Russian language proficiency in bilingual and multilingual speakers ranging from those who self-asses as (semi-)proficient to those whose bilingualism is incipient due to limited input and/or instruction, or declining, due to language attrition. Finally, we provide standardized scores for each test version in order to enable future test administrators to reference their scores to those reported in the present study.

2. Proficiency Assessment in Bilingual and Multilingual Populations

The importance of accurately representing language proficiency is certainly not unique to the Russian language or other LCTLs. The choice of the measure of participants' proficiency presents a starting point of most experimental and clinical examinations of bilingual language development and not only affects the interpretation of experimental findings (see Tremblay 2011 for an extensive discussion of proficiency assessment in L2 French research) but may also have importance for pedagogical and clinical implications. Despite often presenting a methodological challenge, proficiency assessment in research on bilingual and multilingual acquisition is gradually becoming common practice, as it determines participant eligibility for research and helps interpret results of the intervention techniques in a language classroom or tease apart the effects of experimental conditions on participant performance in laboratory research (Kormos 2000; Kotz 2009).

Proficiency and language background measures pulled from large samples of bi- and multi-lingual speakers in the US and beyond (Han 2012; Montanari et al. 2018) reveal important aspects of the social conditions in which bilingual development unfolds in the present days and help educators design optimal educational approaches to supporting life-long bilingualism and multilingualism among the many US-born speakers of minority languages, including Russian. Polinsky and Kagan (2007) list a number of terms used in the language acquisition literature to refer to this population, including "semi-speakers" (Dorian 1981), "incomplete acquirers" (Montrul 2002; Polinsky 2006), and "unbalanced", "dominant", or "early" bilinguals (Baker and Jones 1998). These terms attest to a large degree of variability in the attainment levels of the modern-day US bilingual and multilingual speakers.

Polinsky and Kagan (2007) characterize the Russian language proficiency of the Russian-English early bilinguals in the US as featuring "tremendous variation"; they note that some speakers approach the baseline proficiency characteristics (i.e., appear native-like) in Russian (the minority language), whereas others have significantly reduced

fluency (see Friedman and Kagan 2008 for similar results), perceptible foreign accent, and greatly reduced (over-regularized) morphological paradigms (for similar findings see Polinsky 2007, 2008, 2011; Laleko 2011). Whereas studies investigating heritage Russian converge on the idea that heritage speakers are distinct in terms of their Russian proficiency from the monolingual (baseline) speakers, comparing early Russian-English bilinguals to those with considerable later age of exposure (AOE) to Russian can reveal a selective advantage to early AOE, as well as evidence of L1 transfer. To illustrate, Gor (2019), focusing on the acquisition of Russian morpho-syntax, used a grammaticality judgement task delivered in writing and aurally. Gor's participants were L2ers and heritage Russian bilinguals, matched in terms of their TL proficiency using ACTFL OPI ratings. In Gor's study, heritage speakers outperformed the L2ers across the task modalities. In a similar vein, Ionin et al. (2020, under review) tested adult L1 English L2 Russian learners and English-dominant heritage Russian speakers' sensitivity to the relationship between word order and information structure in Russian; Ionin et al. found that Russian heritage speakers (early/simultaneous bilinguals) reliably detected incongruence between word order, prosody and context, whereas adult Russian L2ers, regardless of the proficiency level, demonstrated invariable preference for default constituent order (SVO) under neutral prosody, seemingly unaware of contextual appropriateness. More target-like performance of the heritage speakers reported by Gor and by Ionin et al. points to an overarching advantage conferred by early exposure to the Russian language and points to qualitative differences in the Russian proficiency of early vs. late Russian/English bilinguals.

While early exposure to a TL in an immigrant family setting may confer an acquisition advantage, it certainly does not guarantee native-like competence across linguistic domains and may be further undermined by subsequent TL attrition. For example, Polinsky (2008) investigated the comprehension of subject and object relative clauses in baseline Russian as well as by English-dominant heritage Russian bilinguals, child and adult. Polinsky established that adult heritage bilinguals differ from Russian monolinguals and child heritage speakers in that they demonstrate non-target-like comprehension of Russian relative clauses. Polinsky's findings support the view that the morphological component of heritage language grammars may be particularly vulnerable/susceptible to attrition (Montrul 2006; Sorace 2004). In a similar vein, Ionin and Luchkina (2019) investigated the effects of word order, prosody and information structure on quantifier scope interpretation by adult Russian English bilinguals. The study found that distinct Russian proficiency levels determine scope interpretation preferences and account for non-target like interpretation of scopally ambiguous sentences in Russian by adult L2 learners and heritage speakers.

Recent work on heritage and L2 acquisition of Russian draws on language proficiency measures which are as distinct as the studies themselves. To illustrate, Gor and Cook (2010); Gor (2019); Ionin et al. (2014) and Polinsky (2005) incorporated heterogenous proficiency estimates varying from elicited production measures to vocabulary recognition and in-house assessment, to standardized test scores. Using such fundamentally different ways of gauging TL proficiency is problematic if we seek a coherent vision of bilingual speakers' competence at different acquisition stages. These considerations further reinforce the need for unified, inclusive proficiency assessment methods developed with L2 learners of Russian as well as bilingual speakers in minority-majority acquisition contexts in mind.

Considerations of Language Background and Acquisition History

Unequivocally, bilinguals' and multilinguals' language proficiency may not be accurately evaluated unless multiple exogenous sources of inter-speaker variance are taken into consideration which may have influence on the TL ultimate attainment levels. A classic study by Johnson and Newport (1989) found significant correlations between adult bilinguals' performance on an aural grammaticality judgment task in English and a number of biographical and attitudinal variables and self-reported proficiency measures. Some of these factors have been routinely introduced as language background variables into proficiency assessment research and include the age of arrival (for immigrant bilinguals),

the age of exposure to the TL, measures of weekly amount of the TL input, self-reported measures of foreign accent, and others.

The importance of biographical data and language history in evaluating bilinguals' performance on language tasks has been firmly established in L2 acquisition and bilingualism research (see, among others, Dunn and Tree 2009; Flege et al. 1998; Gollan et al. 2012; Grosjean 2004; Luk and Bialystok 2013; Sheng et al. 2014). Hyltenstam and Abrahamsson (2003) provided evidence supporting an increasing role of experience-based variables in determining language proficiency in late/adult bilingual learners. These findings call for inclusion of biographical and language history data in assessment procedures. One way to achieve this standard is via combining the primary assessment technique that is independent of the participants' biographical data and language history with a validated instrument for collecting relevant background information. To this end, the present study incorporates an adaptation of the Language Experience and Proficiency Questionnaire (The LEAP-Q) for assessing profiles in bilingual and multilingual speakers created by Marian et al. (2007). Our choice of the LEAP-Q tool as the basis for collecting the self-reported measures of language background and TL input follows the successful implementation of the various adaptations of this questionnaire in research with highly proficient bilinguals (e.g., Conrad et al. 2011; Mercier et al. 2014; Pelham and Abrams 2014; see Kaushanskaya et al. 2020 for more discussion), as well as with emergent bilingual speakers (Antoniou et al. 2015; Nip and Blumenfeld 2015). The LEAP-Q takes into account a number of factors deemed "important contributors to bilingual status" (Marian et al. 2007, p. 943). These factors include language dominance and preference ratings, age and modes of language acquisition, measures of input and usage, as well as duration of stay in the country of residence. Marian and colleagues internally validated the LEAP-Q instrument using data from 52 multilingual speakers and established its criterion-based (external) validity using the LEAP-Q data in conjunction with standardized proficiency scores from 50 adult Spanish-English bilinguals. More specifically, Marian et al. tested a homogenous group of highly proficient bilingual speakers residing in the US who also reported extensive immersion in both their languages. Marian et al. concluded that self-reported language history information was predictive of participants' performance "on specific linguistic tasks" (p. 956).

Because the present study does not seek participants converging in terms of their TL proficiency, it is important to review a language background analysis in the spirit of Tremblay's (2011) norming study of a cloze deletion test created for adult L2 learners of French. Tremblay's sample of 169 L2 learners was characterized as highly heterogenous, based on participants' L1s and self-reported TL exposure and proficiency measures. While Tremblay's study did not use the LEAP-Q instrument, it reported many of the same language background variables, including the age of exposure to the TL, amount of weekly TL input, and history of formal TL instruction. Tremblay concluded that years of instruction in French, followed by self-reported French proficiency, were the best predictors of the cloze test performance in her sample. Results of the external validation of Tremblay's proficiency assessment using self-reported language background variables were superseded by a cluster means analysis based on participants' cloze scores, revealing that proficiency assessment in bilingual populations based solely on biographical data and self-reported proficiency measures may be significantly improved if an independent normed instrument is used in combination with biographical information and language learning history.

3. The Use of Cloze Deletion Tests for Proficiency Assessment

The search for economical yet accurate proficiency estimates is ongoing for many languages under investigation (English: Lemhöfer and Broersma 2012; French: Tremblay and Garrison 2008; Tremblay 2011; Gaillard 2014; Tracy-Ventura et al. 2014; Spanish: Wood Bowden 2016; Mandarin Chinese: Yan et al. 2020). Among the various assessment techniques used in language acquisition and bilingualism research, cloze deletion tests stand out due to their relative ease of implementation, robust ability to discriminate among

the different proficiency levels and well-understood psychometric properties (Brown 2013; Huensch 2014; Hulstijn 2012). Introduced well over half a century ago by Taylor (1953), the cloze deletion format has been widely applied to evaluating language proficiency in individuals beyond the initial acquisition stage (see Watanabe and Koyama 2008 for a meta-analysis of 212 studies; Brown and Grüter 2020). Cloze deletion tests are widely used for holistic language proficiency assessment due to the ease of implementation and robust predictive ability (Brown 1980, 1983, 2013; Kobayashi 2002; Tremblay and Garrison 2008; Tremblay 2011). Cross-linguistically, the cloze test assessment technique may be deployed in language classrooms and beyond, including contexts as diverse as research laboratory and speaker's home. The wide application of cloze tests is supported by the fact that they tap into multiple aspects of the TL proficiency (Storey 1997). These aspects include, but are not limited to, low-level lexical knowledge (Alderson 1979), sentence-level syntactic knowledge (Alderson 1980) and understanding of discourse-level constraints (Chihara et al. 1977).

Traditionally, a cloze deletion test presents an excerpt of a coherent text in the target language in which a portion of the words, content and function, are removed to then be filled in by the test taker. Cloze questions gauge the respondent's ability to supply a word consistent with, and grammatical in, the provided context. Naturally, successful performance requires both high- (discourse) and low- (word) level comprehension processes to ensure that the word chosen for each blank is appropriate from the semantic, syntactic, and inter-sentential standpoint (Van den Broek et al. 2002). Because cloze test design presents options which may affect the internal validity and discriminability of the test instrument, careful decision making is warranted when opting for the deletion method (rational or fixed-ration), the deletion step/interval, and the availability of answer choices associated with each blank provided when the multiple-choice format is adopted as opposed to the "true" fill-in-the-blank format, also known as constructed response (Frey 2018). While each of these options has important repercussions both for test-taker and test developer, they allow to customize the testing tool by making it suitable for respondents falling into diverse proficiency ranges as well as for testing languages which, like Japanese, do not use an alphabetical system (see Douglas 1994 for more discussion) or, like Russian, use a non-roman alphabet. We return to select logistical considerations of cloze test design in Section 7.

In a meta-analysis of published cloze deletion tests, Watanabe and Koyama (2008) reported that in their sample of 212 studies, most used fixed-ration deletion, with the deletion step ranging between every 12th and every 7th word, whereas others used the so called "rational" deletion method. Whereas the fixed-ration deletion method requires that every nth word in the original text is replaced with a blank, the rational deletion method gives the test writer control over the material that is deleted and potentially helps measure a more diverse set of TL proficiency components (Bachman 1985). This may be particularly important for those seeking balance between the numbers of content and function words that the test taker will supply or, when constructing a test for a morphologically rich language, for including reasonably diverse grammatical forms from the same morphological paradigm.

Another critical feature of cloze test design concerns the method for scoring respondent's data. When the fill-in-the-blank format is opted for, each supplied answer must be evaluated for (a) being an exact match to the removed word (known as the exact answer (EX) criterion) or (b) being an acceptable answer and therefore scored as correct (known as the acceptable answer (AC) criterion) (Brown 1980). The latter approach may require that the test writer prepares a list of possible answers for each blank, based on representative responses from native (baseline) speakers who take the cloze first (see Tremblay 2011 for an example of a cloze test offering acceptable answer scoring).

The choice of the scoring method affects the test outcomes: the AC method boosts the score by increasing the probability of supplying a correct answer (Baldauf and Propst 1979; Chapelle and Abraham 1990; Kobayashi 2002) and may affect test discriminability

in one or more parts of the proficiency spectrum. Assessment studies including Brown (1980) and Tremblay (2011) and a meta-study of 144 cloze tests undertaken by Watanabe and Koyama (2008) linked the AC method to greater test reliability. At the same time, Brown and Grüter (2020), using data from 1724 participants who completed Brown's 1980 cloze test for English, reported that the EX method yielded better discriminability among the more proficient respondents. In the same study, Brown & Grüter argued that each discriminability outcome holds, primarily, for the data sample that it is based on.

A different approach to scoring is assumed when a multiple-choice format is opted for. In this case, test takers select one of the provided answers for each blank. The multiple-choice format dramatically simplifies the scoring procedure and, in our own experience, enables test takers from a wider range of TL abilities to complete the test. At the same time, creating incorrect answer choices adds complexity to the test design process, as the test writer must avoid oversimplifying the task (by providing clearly implausible answer choices). Offering distractor answer choices which would require attention from even proficient respondents while being reasonably distinct from the correct answer is necessary to maintain the discriminability of multiple-choice tests (Baldauf and Propst 1979; Mostow and Jang 2012). More design options include the length of the text (the mean value reported in Watanabe & Koyama's report is 374 words, range 125–750), and the number of the blanks (the mean value reported by Watanabe and Koyama 2008 is 34, range 15–80).

While cloze deletion tests are known to offer a methodologically superior solution to the problem of proficiency assessment (Brown 2013; Tremblay 2011), different opinions have been put forward about which aspects of the target language competence these tests tap into. Alderson (1979) and Shanahan et al. (1982) are examples of some early proposals arguing that cloze tests target primarily "low-level" lexical and grammatical skills. Jonz (1990) and Fotos (1991), on the other hand, proposed that cloze tests tap into higher-level discourse competence. In extensive scholarship on cloze test design and application by, e.g., Brown (1983, 1988, 2002, 2013) and in much subsequent work, cloze tests have been characterized as a measure of overall (Brown 1980), general, or global (Huensch 2014) language proficiency. With that being said, one admitted limitation of cloze tests over other methods of proficiency evaluation, including elicited production and elicited imitation tasks (see Gaillard 2014; Wood Bowden 2016 for examples) is that they do not elicit speech data. In her work on the L2 acquisition of stress in L2 French, Tremblay (2009) correlated results from a cloze test and a foreign accent rating task based on read L2 speech and reported a correlation coefficient of 0.43, which translates into a moderate strength correlation. While analysis of production data is beyond the scope of the present study, we proceed by adopting the view endorsed by Abraham and Chapelle (1992); Storey (1997); Hughes (2003); Brown (1980, 2013) and others that cloze test scores present an integrative measure of bilinguals' and multilinguals' proficiency. Consistent with this view is the fact that cloze test scores demonstrate a strong correlation with results of comprehensive standardized tests like the TOEFL (Bachman 1985; Brown 1983; Nunan and Carter 2001) and have been used by major US universities in place of a more comprehensive testing procedure for purposes of ESL placement and foreign language class placement (see Tremblay 2011, p. 344 for specific examples).

4. Constructing the Test

In what follows, we discuss the steps we undertook to construct and validate a cloze deletion test for use with adult populations of bilingual and multilingual Russian speakers. In laying out our approach to test design, we closely followed the steps and recommendations outlined in Brown's original scholarship (Brown 1980, 1988, 2000, 2002, 2003, 2009, 2013). Brown's established guidelines for constructing language proficiency tests and scoring test data, as well as methods of evaluating the validity and reliability of the testing tools, are widely used in the assessment literature (Kim and Rah 2019; Kleijn 2018; Trace 2020). Adhering to Brown's guidelines has yielded successful validation of cloze deletion tests constructed for English (Brown 1980, see Watanabe and Koyama 2008;

Brown and Grüter 2020 for test metanalyses), French (Tremblay 2011), Japanese (Yamashita 2003) and beyond.

The following steps were undertaken to construct a cloze deletion test for Russian and establish its validity, reliability, and discriminability in measuring participant TL proficiency.

4.1. Materials Selection and Deletion Procedure

Russian presents a morphologically rich language with three nominal genders, six morphological cases and three nominal declension types. Morphological case markers are further differentiated based on noun number and gender. Adjectives agree with nouns in number, gender, and morphological case. Verbs require tense and viewpoint aspect marking, and—depending on the tense—person, number, and/or gender markers, and fall into two conjugation types. Function words include pronouns, prepositions and a copula verb used with past and future tense verbs. One consequence of such grammatical organization of the target language which was taken into consideration when constructing the present test is that various grammatical features of Russian can be tested via content word deletion rather than by function word deletion.

Luchkina and Stoops (2013) selected two reading passages published in a Russian newspaper website (https://www.kp.ru/, accessed on 30 April 2021) that were deemed of interest to an average reader of high school age or older. An abridged version of each text was created containing between 300 and 400 words. Working with each text, Luchkina and Stoops applied a rational cloze deletion procedure which resulted in 58 deletions in text 1 and 60 deletions in text 2. Function words accounted for approximately 30% of all deleted material. Distribution of word forms representative of each major grammatical paradigm was examined based on the deleted words. It was determined that deletions in text 1 and in text 2 represented all number and gender forms in verbs and nouns. Additionally, text 1 deletions required that the test taker supplied all three verb tenses, and five case forms in a fairly balanced fashion. In contrast, text 2 deletions contained a smaller number of nominal case forms and verb tense forms compared to text 1.

A focus group of native speakers of Russian included six adult participants, mean age = 36.8. They were instructed to fill in the blanks in each text presented in pen and paper format. All participants resided in the US at the time of participation (average duration of US residence was 2.1 years). Answers were scored using the Appropriate Criterion method (Brown 1980) discussed in Section 5.2. Focus group participants were then asked to rate each text as (a) engaging and (b) culturally specific, i.e., more understandable to a Russian reader as opposed to a foreign reader, on a scale from 1 to 10. The endpoints of the scale were defined as follows: 1 represented "not at all" and 10—"very much so". Results of the focus group analyses are summarized in Tables 1 and 2.

Table 1. Accuracy (means and ranges) of focus group cloze test performance.

Text 1	Text 2
Mean: 98.8% Range: 97.5–100%	Mean: 96.4% Range: 95.6–98.9%

Table 2. Content analysis results based on focus group participant ratings.

Engaging Content	Cultural Specificity
Text 1: 9.3 Text 2: 5.4	Text 1: 3.4 Text 2: 6.8

The focus group participants rated Text 2, about the Baykal lake, as less engaging and more culturally specific. Text 1, reporting a recent scientific discovery, namely, singing abilities in genetically modified mice, was chosen as the basis for the cloze deletion test

due to its more engaging, yet less culturally specific content, as well as greater variability of grammatical forms included among the deleted words.

The chosen reading passage presented an excerpt from the original article abridged to half its original length. The passage contained 307 words; among the 58 words removed using the rational deletion method, 36 were content and 22—function. The rational deletion method was used to ensure that sufficiently diverse grammatical forms are tested. As a result, the deletion step was not fixed and varied between four and six words. The starting point of deletion was the second word in the second sentence of the passage. The grammatical forms included among cloze deletions words are summarized in Table 3.

Table 3. Verb forms and morphological cases of nouns included among the deleted words.

Verb Forms	N Cloze Deletions	Nominal Cases	N Cloze Deletions
Infinitive	2	Nominative	9
Participle	2	Genitive	4
Present	4	Dative	1
Past	4	Accusative	4
Future	1	Instrumental	1
		Prepositional	-
	Total: 13		Total: 19

Lexical frequency of the deleted words was recorded using a frequency dictionary by Lyashevskaya and Sharov (2009), to identify highly infrequent words or wordforms. No such instances were found. Figure 1 shows mean frequencies of occurrence of the deleted words in Russian corpus data. Russian language course syllabi (semesters 1–4) from a public US University and the textbook *Nachalo* (Lubensky et al. 2002), commonly used in courses of Russian as a foreign/heritage language, were examined to gauge the number of semesters of formal instruction which would enable an emergent L2 learner to comprehend the chosen reading passage. It was estimated that satisfactory comprehension requires approximately four-five semesters of classroom instruction in Russian. This supports the view that cloze deletion tests may be inappropriate for learners during the initial acquisition stage, especially those in the first-year courses (Tremblay 2011; Brown and Grüter 2020).

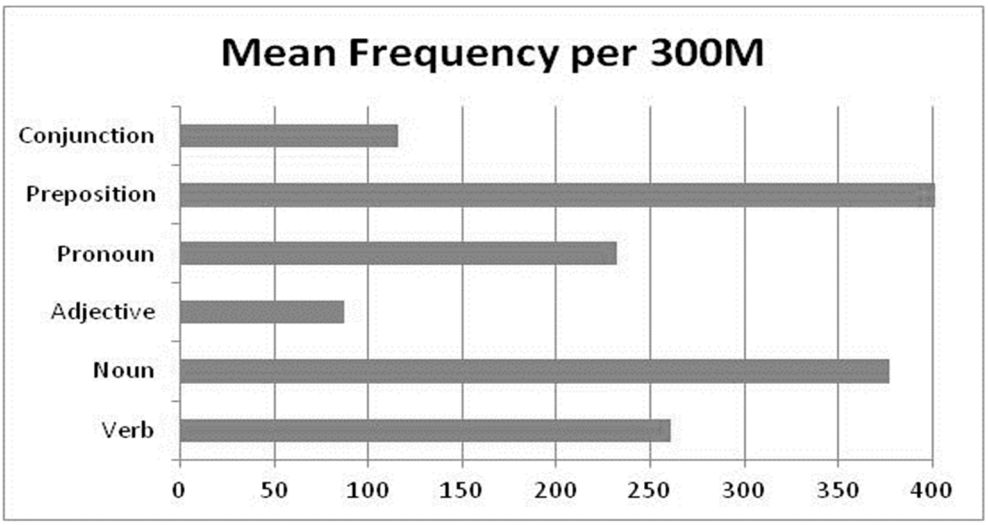


Figure 1. Lexical frequency (means) of deleted words (per 300 million), based on Lyashevskaya and Sharov (2009). X-axis: count; Y-axis: part of speech.

4.2. Response Formats

A Constructed Response version (henceforth, the CR version) of the cloze test was built using Survey Gizmo online testing software. The online test opened with an electronic

consent form in which participants were asked to release their data for research purposes and invited to participate in a follow up production task. The following part of the test was an adaptation of the normed language background questionnaire originally offered in Marian et al. (2007), supplemented with participant-reported proficiency measures. The cloze deletion component of the testing tool was presented next. Participants were instructed to read the text as a whole and answer three broad comprehension questions. Following this initial familiarization phase, participants were presented with the text one paragraph at a time. They were asked to fill in the blanks in Russian, using Cyrillic or Roman scripts. Participants were instructed that their primary task was to fill in each blank with one word such that the chosen word renders the sentence complete and grammatically correct. Instructions were provided in English and in Russian. All responses were submitted online and digitally stored by Survey Gizmo testing software. Participation time lasted between 30 and 60 min and was greater for proficient respondents. Many US-based participants reported no access to a Cyrillic-enabled keyboard, and some—feeling uncertain about supplying romanizations of the Russian words.

A multiple-choice version (henceforth, the MC version) of the same cloze deletion test was constructed to reduce the average testing time and simplify scoring as well as testing logistics. To minimize the amount of cognitive load induced by multiple answer options, we opted for three answer choices to be presented along with each blank. The correct answer choice always matched the deleted word. The distractor choices were created following practices and considerations outlined in Baldauf and Propst (1979) and Mostow and Jang (2012). More specifically, Mostow and Jang (2012) discuss three types of distractors, including ungrammatical, nonsensical, and plausible. In the present study, distractor categories "ungrammatical" and "nonsensical" were conflated for cloze deletions which had to be filled in using function words. For such items, the distractor always matched the target word in terms of segmental length and phonological form. For example, if the target word was a preposition consisting of one consonant or one vowel, such as *k* (Eng. 'to') or *v* (Eng. 'into'), the distractor option was another such preposition, such as *s* (Eng., 'with') or *o* (Eng., 'about'). Similarly, CV prepositions, e.g., *na* (Eng., 'on') were presented along with distractors of the same phonological form, e.g., *po* (Eng., 'along').

Only ungrammatical and plausible distractors were implemented for cloze deletions which had to be completed with open class words. For ungrammatical distractors, the target word was provided featuring a case, number, or gender suffix (on nouns and adjectives), and a tense, person, or number suffix (on verbs) which rendered the wordform contextually ungrammatical. For example, the target genitive plural form *myshey* (Eng., 'mice', genitive) was provided along with distractors featuring the prepositional plural form or the dative plural form of the same noun, *myshah* and *mysham*, respectively. Plausible distractors matched the target word in terms of its part of speech, grammatical form and, where relevant, animacy, while being infelicitous in terms of meaning. For example, the target word *posadili* (Eng., 'seated', plural) was presented along with distractors *polozhili* (Eng., 'laid down', plural) and *postavili* (Eng., 'placed in a standing position', plural). In combination with the noun *myshey* (Eng., 'mice', genitive), each of these distractor choices could seem plausible to an L2 learner or a speaker with incomplete acquisition of Russian, while being clearly anomalous to a native Russian speaker.

The lexical frequency of plausible distractors was examined to identify possible infrequent words or word forms. No such cases were found. The MC test was built using Survey Gizmo and was later implemented in Qualtrics. Additionally, a pen and paper test was created for immersion learners tested in Russia. Each blank in the MC test was presented along with three answer choices. Participants were instructed to select the best answer choice such that it would render the sentence complete *and* grammatically correct. To further decrease the participation time, a shortened version of the language background questionnaire was used with fewer included self-assessed measures related to motivation, language preference, and types of the TL input. These test design modifications reduced the

average participation time to 25–45 min and brought the rate of incomplete participation down to approximately 15% from the 40% recorded for the CR test version.

5. Norming and Validation

5.1. Participants

A control group of native Russian speakers (n = 52, mean age = 32, range 21–52) completed the CR version of the cloze test online. All participants resided in Russia at the time of testing. All reported exposure to Russian at birth and considered Russian to be their only native and primary language.

Five hundred and three speakers of Russian who also reported equal or greater proficiency in at least one other language participated (Additionally, nine learners from a second semester Russian course at a US public university attempted the test during a scheduled testing session but chose to withdraw their data from analysis or reported that the reading passage was too difficult for them to comprehend and declined further participation). Each participant completed one test version, CR or MC, and gave consent to use their background information and test data for research purposes. Participants reported native or native-like proficiency in various languages, including English, Chinese, Turkmen, Kazakh, Turkish, Italian, Korean, French, and Japanese. Most participants who completed the test and the background questionnaire in their entirety, and thereby met the inclusion criteria, were recruited from courses of Russian as a foreign, heritage, or second language at two US universities and two universities in Russia. On average, L2 learners had completed 4.3 (range 2–10) semesters of formal instruction in Russian by the time of participation.

Of the 503 participants, 233 completed the CR test version and 270 participants—the MC test version. We classified each respondent based on the primary type of TL exposure, namely, formal, i.e., mostly classroom-based, and mostly immersion-based. The latter category included participants who resided in Russia at the time of participation and/or reported a history of living in Russia for a period of six months or longer. Additionally, participants who grew up acquiring Russian at home in a language minority setting and eventually developed dominant proficiency in a majority language were classified as heritage Russian bilinguals. Participants' cumulative language background information organized by test version is summarized in Table 4. As Table 4 demonstrates, independent of the test version, each self-reported background measure is broadly distributed (e.g., Length of Residence in a Russian speaking country varies between 0 and 252 months; self-reported foreign accent ratings vary between "very slight" (0) and "very strong" (10), etc.). Based on the language background information and the self-reported measures of TL proficiency, our sample of adult Russian bilingual and multilingual speakers is clearly heterogenous and therefore suitable for test validation purposes.

Table 4. Summary of participants' language background information (based on self-reported measures), for the CR and MC test versions.

Variable	Mean		SD		Range	
	CR	**MC**	**CR**	**MC**	**CR**	**MC**
Age of Exposure (AOE)	16.3	14.06	8.98	7.55	0–44	0–32
Length of residence (LOR, months) in a Russian speaking country	30.2	23.82	49.6	31.86	0–252	0–252
Preference reading in Russian as opposed to other languages known to participant (1–10)	2.1	3.06	2.3	2.4	0–10	0–10
Self-reported accent rating when speaking Russian (1-very slight–10-very strong)	5.79	4.96	3.6	2.34	0–10	0–10

The composition of the sub-samples of participants who completed the CR and the MC test versions is also dissimilar. To illustrate, 56% (n = 132) of the CR version participants were US-based L1 English L2 Russian learners. Among the MC test version participants, approximately 40% were L1 Chinese L2 Russian learners (n = 103), whereas the number of L1 English L2 Russian learners residing in the US was considerably smaller than in the CR sub-sample (n = 28) and accounted for only 10% of the participants. A total of 196 participants resided in Russia at the time of participation or reported an extended period of residence in a Russian-speaking country. One hundred and seventy-one of these participants completed the MC version and 25—the CR version. A total of 116 Russian heritage bilinguals resided in a country where their primary language was spoken but reported an early exposure to Russian through one or both parents. Seventy-three of them completed the CR version and 43—the MC version.

5.2. Data Scoring Methods

Using two distinct response formats, constructed response and multiple choice, requires that the scoring method be adjusted depending on the version of the test administered for assessment. As discussed in Section 3, constructed response tests can be scored using the so-called exact answer (EX) criterion, whereby only those answers are considered accurate (correct) which match the deleted word (Brown 1980). Alternatively, the acceptable answer (AC) criterion renders more than one answer acceptable, as long as the provided word is semantically and otherwise appropriate (e.g., a synonym; see Brown 1980, 1983 for a comparison of these scoring methods). Multiple choice tests, on the other hand, may only be scored using the EX criterion. In the present study, our choice of the AC method was limited to scoring the CR test performance and was based on native speaker participants' data containing more than one appropriate answer option for 24 of the 58 test items. The EX scoring method was used with the MC test version.

The MC test answers were scored automatically in Qualtrics. The CR test data were scored manually. All correct responses were coded as 1 and incorrect—as 0. Native speakers' answers and approximately 40% of the answers supplied by our bilingual and multilingual participants reported in Luchkina and Stoops (2013) were scored by these authors (native Russian speakers) and the rest—by Luchkina. A response was considered correct when it matched one of the answer choices supplied by the native speakers. Numerous spelling- or romanization-induced errors were ignored. Approximately 12% of otherwise acceptable responses were coded as "0" because of agreement, case or tense errors.

5.3. Data Analyses

Results are presented separately for the CR and the MC test versions. We first present the summary statistics and evaluate the distributions of the obtained cloze scores. We use Kuder-Richardson Formula20 estimates to gauge the internal validity of each test. To establish the external validity, we report results of two multivariate analyses of cloze test performance in which the log likelihood of supplying an accurate response is modeled based on self-reported language background and proficiency data. Then, in a series of test item analyses, we gauge individual item difficulty and discriminability, and how these measures vary depending on the response format (CR vs. MC). We conclude our analyses by assigning participants to distinct proficiency levels based on their cloze test performance. Using this approach, we present score ranges for four distinct proficiency levels and probe into the composition of each resulting sub-sample of cloze test participants.

6. Results

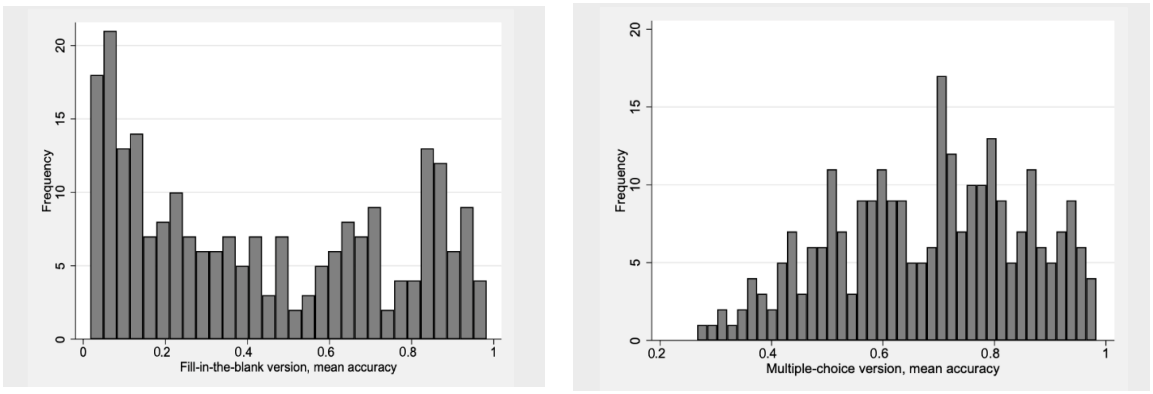
Native speakers' answers were scored using the AC scoring method. This yielded the mean test score of 98.8% (range: 97.5–100%). Based on the responses from this baseline group, a list of acceptable answers was constructed for each test item. Each answer option included in the list was considered grammatical and otherwise appropriate by both raters. The mean cloze probability for the 58 deletions was 2.9, meaning that on average, each

blank in the test could be filled in with approximately 3 words. Function words had the lowest cloze probability ranging between 1 and 2, whereas content words, in particular, adjectives and adverbs, had the highest cloze probability ranging between 1 and 11.

Table 5 summarizes learners'/bilinguals' mean accuracy and score ranges for the CR and MC versions. The corresponding score distributions are provided in Figure 2.

Table 5. Summary statistics of cloze test data.

	CR	MC
Mean accuracy	0.46	0.68
Median	0.41	0.70
SE	0.02	0.01
SD	0.29	0.17
Skewness	0.22	−0.27
Range	5.2–98.9	26.7–98.3



(**a**) (**b**)

Figure 2. Frequency distribution of the cloze scores. (**a**) Fill-in-the-blank test; (**b**) multiple choice test.

Score ranges for the CR and MC tests are widely distributed and indicate that a broad range of TL proficiencies is represented in our participant samples. As shown in Figure 2, the CR test scores form a bimodal distribution with larger portions of scores falling within the lower range (10–40% accuracy) and within the higher range (80–100% accuracy). At the same time, considerably fewer scores fall in between these two ranges. Despite the mean score of 0.46 being fairly well-centered, the distribution is also skewed to the left, indicating that most respondents scored below mean. The MC test scores form a unimodal distribution, skewed to the right. This indicates that fewer MC test participants scored below mean, in addition to a fairly high mean accuracy of 68% obtained in the MC test.

The observed differences in the score patterns obtained in the CR and the MC tests can be attributed to two primary factors. The first is the dissimilar composition of the participant samples: recall that the CR test was administered to a large group of adult L2 learners of Russian and Russian heritage bilinguals residing, primarily, in the US. The MC test, on the other hand, was administered to a large group of bilingual and multilingual speakers whose length of residency in a Russian speaking country (primarily, Russia) exceeded 6 months. At the same time, the number of US college learners and heritage bilinguals in the MC sample was considerably smaller. The second factor entails the dissimilar difficulty of the CR and MC tests, potentially linked to the different response formats. These considerations warrant item-based analyses, which we now turn to, in order to gauge individual test item facility and discriminability, as well as the internal consistency of the CR and MC tests.

6.1. Item-Based Analyses of Cloze Test Performance

Following Brown (2003), the internal consistency of the cloze test was assessed using the Kuder-Richardson Formula20 (K-R20) reliability analysis for individual test items applied to dichotomously scored data. Internal consistency refers to a general agreement among the items that make-up a composite test score and is based on the correlation between individual item scores across participants. The global K-R20 measure is a number between 0 and 1; high internal consistency indicates that the test measure is reliable. Virtually identical K-R20 coefficient values of 0.977 were obtained for the CR test version (K-R20 = 0.9766) and the MC test version (K-R20 = 0.9762) and reveal that independent of the response option, the cloze deletion test is internally consistent.

Individual test item facility and discriminability were evaluated separately for the CR and the MC versions (see Figure 3). Item facility (IF) scores represent mean accuracy rates for individual test items, whereby greater mean accuracy translates into greater item facility. The IF scores obtained in the CR version ranged between 0.14 and 0.77, centered at 0.46, (SD = 0.13); in the MC version, the IF values ranged between 0.32 and 91 and were centered at 0.68 (SD = 14.2).

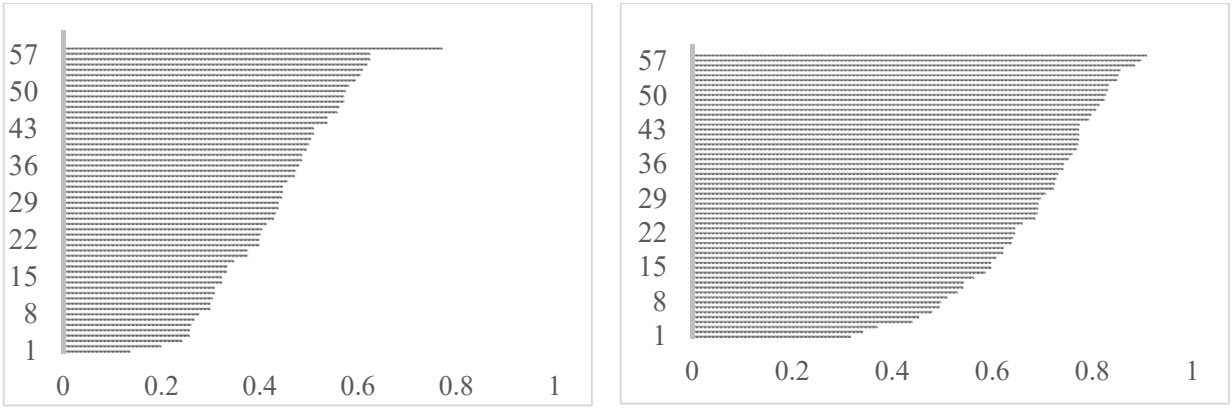


Figure 3. Individual test item facility (means). (**a**) fill-in-the-blanks test; (**b**) multiple-choice test.

The ability of individual test items to discriminate between dissimilar proficiency levels was assessed using the Individual Item Discriminability Analysis adapted from Brown (2003). The analysis yields a measure of discriminability for each test item based on how accurately it discriminates between low-, mid-, and high-scoring participants. A negative discriminability index would serve as an indication of respondents in the bottom third range (the lower tercile of the distribution) being more accurate than those in the top third range (the higher tercile of the distribution). According to Brown (2003), item discriminability index should exceed the cutoff threshold of 0.20, whereas items whose discriminability falls within a 0.30–0.70 range can effectively discriminate between learners of different TL proficiencies.

The mean discriminability indices along with the corresponding score distributions for each test version are summarized in Table 6. In the CR version, the mean item discriminability was computed at 0.72 (SD = 0.14). Good discriminability was obtained for 57 out of the 58 test items. Mean item discriminability in the MC test, centered at 0.39, ranged between 0.13 (poor discriminability, obtained for 4 cloze items) and 0.62. Forty-seven of the 58 MC test items yielded good discriminability (>0.30) and seven items—reached the lower threshold discriminability score of 0.20. Results of the item-based analyses reveal that opting for the MC response option may boost participants' score but is unlikely to compromise the accuracy of the assessment outcome even when most participants in the sample are fairly proficient.

Table 6. CR and MC scores of top and bottom participant terciles and item discrimination indices.

	Mean		SD		Range	
	CR Test	**MC Test**	**CR Test**	**MC Test**	**CR Test**	**MC Test**
accuracy, top third	0.78	0.88	0.19	0.11	0.45–0.99	0.51–0.98
accuracy, bottom third	0.09	0.48	0.11	0.16	0.03–0.32	0.16–0.81
discrimination indices	0.72	0.39	0.14	0.12	0.37–0.96	0.13–0.62

Tremblay (2011) reports that participant accuracy is not only reflective of one's language background and TL exposure, but also hinges on whether any given cloze test item requires that a content vs. a function word is supplied to fill in the blank. Due to their greater cloze probability, function words are associated with greater accuracy and may be supplied by bilinguals whose TL proficiency is merely emergent. Comparing mean accuracy rates for function and content words in the present study yielded the following results. A 5% difference in the mean accuracy for content and function items was obtained in the CR test, in the predicted direction. The difference was 8% in the MC test data, also in the predicted direction. The effects of different parts of speech and lexical frequency, in conjunction with participant self-reported background and proficiency measures were evaluated in two multivariate analyses of the cloze test performance to which we turn next.

6.2. Multivariate Analyses of Cloze Test Performance

The goal of the multivariate analyses was to determine which extra-linguistic factors, in combination with test-specific properties, systematically affected the cloze test performance in our bilingual and multilingual participants. To this end, we modeled the log likelihood of obtaining an accurate response for a single test item, given the following categories of predictors. First, we considered individual test item characteristics, including the lexical frequency and the part of speech of the deleted word (POS). The second predictor category was based on participants' self-reported language background data and included measures of age of exposure to the TL (AOE), age of arrival to a Russian speaking country (AOA), length of residence in a Russian speaking country, in months (LOR), and a number of semesters of formal instruction in Russian. To externally validate the results of the CR and the MC tests, we included participants' self-reported measures of Russian proficiency, including the degree of the foreign accent (evaluated on a 1–10 interval scale) and the preference to read a text in Russian vs. in a different language known to the participant (evaluated on a 1–10 interval scale).

Responses to the CR test (henceforth, the CR model) and the MC test (henceforth, the MC model) were analyzed separately, using two mixed effects logistic regressions implemented in Stata 15. In addition to the fixed effects listed above, each model included random effects (intercepts and slopes) for participant and test item. Main effects (raw logit coefficients and *z* statistics) which reached significance are summarized in Table 7.

Table 7. Significant main effects obtained in multivariate logistic analyses of cloze test performance.

Fixed Effects	**CR Model**		**MC Model**	
	β	*z*	β	*z*
N semesters of formal instruction	0.91	12.64 ***	n.s.	n.s.
AOE	−0.04	−4.45 ***	−0.04	−2.08 *
LOR	0.002	3.27 ***	0.006	2.16 *
Foreign Accent	n.s.	n.s.	−0.06	−2.0 *
POS (conjunction)	0.47	2.55 **	0.9	4.54 ***
POS (pronoun)	0.32	2.1 *	1.76	−10.66 ***
POS (adverb)	n.s.	n.s.	−0.76	−5.61
Lexical frequency of deleted word	0.00002	3.97 ***	0.0008	16.13 ***

Significance levels: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. Same in the following tables.

The multivariate analyses reveal that participant performance accuracy is systematically related to several test item-based characteristics as well as language background and proficiency measures.

Independently of the response option, the likelihood of providing an accurate response was inversely associated with the AOE to Russian but co-moved with the LOR in a Russian speaking country. Additionally, the number of semesters of formal TL instruction was positively associated with the likelihood of supplying an accurate response in the CR test data. Response accuracy was also affected by select test item characteristics, including lexical frequency and the POS of the deleted word. Specifically, in both test versions, we established a positive relationship between accuracy and function word deletions (conjunctions and preposition) further reinforced by the fact that such words are particularly frequent and have naturally greater cloze probability. In summary, the results of the multivariate analyses of the cloze test performance support that the CR and MC versions are externally valid, such that the participant performance is reliably predicted based on the type and length of TL exposure, as well as on individual test item facility. To further tap into the external validity of the present test, we examined the relationship between the number of semesters of TL instruction and the respondent accuracy, computed separately for function and content cloze deletions.

Recall that the relationship between accuracy and the amount of formal TL instruction, often used as a gauge of learners' proficiency, was established as significant for adult L2 learners of Russian who completed the CR version. Figure 4 illustrates a strong positive relationship between the amount of formal instruction and the mean accuracy on function and content word deletions in the CR test data. At the same time, this relationship is not perfectly linear: learners with five and seven semesters of formal TL instruction outperform those with six and eight semesters when compared on function word accuracy. Furthermore, no gains in accuracy on content word accuracy is observed among those with six to eight semesters of instruction. Finally, learners with nine semesters of instruction are equally accurate when supplying content or function words. This pattern, however, does not hold for those with ten semesters of formal TL instruction. The relationship between test performance and the amount of formal TL instruction illustrated in Figure 4 points to variability in individual attainment levels which may stem from various aspects of learners' personal history. We therefore conclude that grouping adult L2 learners based on background information or self-assessment may yield inaccurate outcomes, if only for subgroups of learners. This consideration warrants our final cloze performance analysis in which we assign the CR and MC version participants to distinct Russian proficiency levels based on their cloze test performance and independently of the self-reported measures of TL proficiency or history of TL learning.

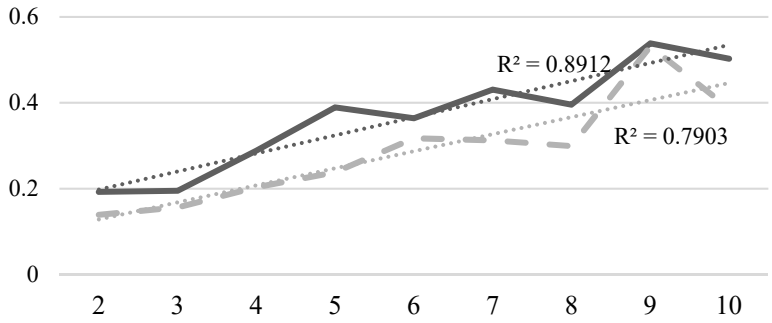


Figure 4. Mean accuracy by number of semesters of TL instruction completed by CR test respondents, Dark line: function words, Light line: content words. Fitted regression lines (dotted) are shown with corresponding R^2 values.

6.3. Using Cloze Performance as a Basis for Proficiency Assessment

A series of K-means cluster analyses were used to identify an optimal number of participant clusters in our CR and MC participant samples. To determine an optimal number of proficiency groups, the analysis was implemented in a stepwise fashion with follow-up analyses of variance performed after each iteration. Using cluster as a between-subject variable, the ANOVAs assessed the fit of models with two, three and four clusters. Three iterations of K-means cluster analysis were implemented using the CR test data. As shown in Table 8, the most dramatic model gain was observed when four clusters were identified. Based on this outcome, participants were divided into four proficiency groups (group one demonstrating the lowest attainment level and group four demonstrating the highest attainment level). To enable comparisons across the CR and MC versions, the same number of clusters was applied to the MC data. Recall that the present study excludes emergent bilingual speakers, i.e., those during the initial stage of TL acquisition. Levels 1–4 are therefore in addition to the initial acquisition level (typically, the first two semesters of formal instruction), not represented in our data.

Table 8. ANOVA parameters for each iteration of the k-means cluster analysis.

k (n Clusters)	df	F	Model Gain
2	1, 230	1024.3 ***	–
3	2, 229	1154.9 ***	130.6
4	3, 228	1639.1 ***	484.2

The cloze test performance of participants assigned to each cluster is summarized in Table 9. The four resulting clusters are differentiated in terms of their cloze scores. To probe into the internal composition of the obtained clusters, we use the following participant category labels (see Section 5.1 for details): adult college learner, adult immersion learner, and heritage Russian bilingual. For illustration purposes, we briefly examine four out of eight clusters in our sample: the two scoring the lowest and the two scoring the highest, on the CR and MC tests.

Table 9. Cluster profiles of CR and MC test participants.

	Mean Accuracy (%)		Accuracy Range (%)		Cluster Size (n) Participants	
Cluster	CR Test	MC Test	CR Test	MC Test	CR Test	MC Test
1	9	41	5–20	27–49	79	44
2	32	57	21–45	50–64	52	70
3	62	73	57–74	65–80	48	87
4	87	89	75–99	81–98	54	68

In the CR sample, the lowest performing cluster (1) included two immersion learners, 67 adult college learners, and 10 heritage Russian bilinguals. The highest performing cluster included 10 immersion learners, 30 heritage Russian bilinguals and 14 college learners. The lowest performing cluster in the MC sample included 17 college learners, 23 immersion learners, and four heritage Russian bilinguals. In contrast, the highest accuracy cluster included only six college learners, 41 immersion learners, and 68 heritage Russian bilinguals. The observed distribution of participants with the lowest vs. highest TL proficiency is in line with the results of the multivariate analyses reported in Section 6.2. Specifically, we reported an advantage of early target language exposure, which is what apparently drives the comparative advantage of the heritage Russian bilinguals in the CR and MC tests. We also established a systematic relationship between the LOR in a Russian speaking country and the likelihood of an accurate response in the MC test data. This is in line with the fact that immersion learners comprise 60% of the best performing MC test participants. Finally, the number of semesters of formal TL instruction presents an

important predictor of response accuracy in the CR test. In line with this result, the lowest performing cluster includes participants with as few as two and as many six semesters of formal instruction, centered at 4.3 semesters. However, in the best performing cluster, this range is between three and ten, and centered at 7.1 semesters of formal instruction.

7. Discussion

This paper reported on building and validating a cloze deletion test designed to assess proficiency in populations of bilingual and multilingual Russian speakers dominant in a different language. The test that we created offers two test item formats, Constructed Response (CR) and Multiple Choice (MC), each linked to a distinct response option and offering more than one way of scoring the responses. The choice of the response format is offered due to logistical and assessment considerations some of which are unique to the target language at hand. We will first review the considerations of testing logistics.

The fact that Russian does not use the Roman script makes open-ended questions which must be answered using the TL problematic for the online test takers without access to a Cyrillic-enabled keyboard. This was the case for most of the L1 English CR test participants in the present study. To enable participation for those without access to the Cyrillic script, we made its use optional and accepted romanizations of Russian words instead. Such decision necessitated extra care when scoring participants' answers, as romanizations may look ambiguous and thereby may mask an otherwise correct response. We committed to the following scoring practices for the CR test data: spelling errors in stems were not considered when scoring test performance regardless of what script was used; if a grammatical affix appeared incorrect due to either a spelling error or a grammar error (e.g., subject-verb agreement error), the answer was considered incorrect and was scored as zero.

A different logistical consideration related to the use of the present test concerns the duration of the testing session: an estimated 15–30 min increase in the duration of the testing session was linked to the CR response format compared to the MC format, and affected proficient test takers more than those with lower TL proficiency. This dissimilar effect can be attributed to proficient participants entertaining a greater number of response options with the CR format than with the MC format. The MC test is therefore optimal for instructors or researchers in need of a time-efficient assessment method which also does not require access to a Cyrillic-enabled keyboard. If implemented using commercial online testing platforms, the MC test allows for automatic scoring of test responses, which makes the scoring process straightforward and time efficient. Cumulatively, these characteristics render the MC test particularly well-suited for assessment practices with inherent time constraints related to participation or scoring of the results.

Turning now to test-specific characteristics, the CR test demonstrated excellent internal consistency, reliability, and discriminability among dissimilar proficiency levels, when evaluated against the standards established in the assessment research (Brown 1980, 2003; Kobayashi 2002; Tremblay 2011). The MC test demonstrated comparable internal consistency and reliability. Less than 10% of the MC test items showed lower than optimal discriminability, which we attribute to the fact that high cloze probability items like conjunctions and prepositions were sufficiently accessible to even the less proficient participants not to warrant the multiple-choice format. While mixed format cloze deletions tests have been argued to be more psychometrically robust (Wang et al. 2016), but see Baghaei and Ravand (2019) on compromised construct validity in mixed-format tests), introducing a number of CR items into the MC test, to maintain comparable difficulty across all items, would defeat the primary motivations for its creation—simplified administration and automated scoring of the results.

The external validity of the CR and MC tests was established by tapping into how the likelihood of providing a correct response is shaped by participants' language background and TL acquisition history. An analysis of participant language background supports the view that the composition of each sample of bilingual or multilingual speakers is

rather unique and may be best characterized by using a subset of background measures relevant for the group whose proficiency is being evaluated. To illustrate, results of the multivariate analyses revealed that the amount of formal TL instruction received by adult L2 learners of Russian, in combination with self-reported measures of L2 exposure and TL proficiency, are among the determinants of successful CR test performance. A closer look at the relationship between the amount of formal TL instruction and test performance showed evidence of individual variability in test scores obtained from the learners in the same semester of instruction. This points to inherent heterogeneity among the learners whose primary source of the TL input is classroom-based. Because Russian is a LCTL, soliciting participants beyond the initial acquisition stage presents an inclusive rather than an exclusive process. The binomial character of the distribution formed by CR test scores reveals that a substantial number of adult US-based L2 learners who completed the test can be identified as emergent Russian-English bilinguals despite the fact that all such participants reported completing two or more semesters of formal classroom instruction in the TL. At the same time, semesters of classroom TL instruction serve as an important predictor of cloze test performance in the CR sample.

Whereas the amount of formal TL instruction was secondary in explaining the performance of the MC test participants, self-reported ratings gauging the strength of the foreign accent were negatively associated with MC test performance. Recall that the MC participant sample included over 70% of bilingual and multilingual speakers who either resided in Russia at the time of participation or reported an extended length of residence in a Russian speaking country. Exposure to the TL in an immersion setting is instrumental to improving proficiency, and often results in an increase in pronunciation accuracy and a concomitant decrease in speech accentedness (e.g., Ingvalson et al. 2011). Although cloze deletion tests do not directly measure oral production or pronunciation skills, test performance may be indirectly related to measures of TL fluency and pronunciation accuracy, in immersion learners. Following Tremblay (2009), we support the importance of externally validating assessment tools which solicit written data only by using spoken (elicited) production data. It is our hope that future uses of the present test in studies involving analyses of production data will afford such additional external validation.

To summarize, analyses of participant language background and TL acquisition history reveal that both formats of the proposed cloze deletion test are externally valid.

Using the CR and MC test scores, we conducted clustering analyses of our participants' scores in order to draw inferences about their TL proficiency independently of self-reported measures of the TL skill level or the amount of formal instruction. Among the four resulting clusters summarized in Table 8, clusters 1 and 2, representing low and low-mid proficiency, have non-overlapping score ranges when compared across the CR and MC test versions. We argue that this result is due to the combined effects of dissimilar composition of participant samples in each test, as well as due to the different response formats offered in these tests. Comparative analysis of test item facility (see Figure 3) reveals that the CR test items, on average, are approximately 20% more difficult for the CR test participants than the MC items are for the MC test participants.

Remarkably, the response format made no difference for participants included in the top performing cluster 4, comprised of highly proficient bilinguals and multilinguals in our sample. The composition of this cluster suggests comparable, high levels of TL proficiency in participants with early age of TL exposure, including a group of heritage Russian bilinguals, and in participants with extended length of residence in a Russian-speaking country. The best-performing cluster also included adult L2 learners with 7–10 semesters of formal classroom instruction in the TL. Based on these results, it is our understanding that the MC response format, while potentially beneficial for speakers with incomplete TL acquisition or for those undergoing TL attrition, yields no participation advantages for speakers with more native-like TL attainment, beyond the practical considerations of test duration or scoring ease. Furthermore, recall that the CR test performance was scored using the acceptable answer criterion. Brown and Grüter (2020) report that in their sample

of over 1700 respondents who completed Brown's (1980) cloze deletion test for English, using the exact answer criterion yielded the most accurate discrimination among the best performing participants. The AC scoring method was chosen in the present study based on the baseline performance of fifty-two native Russian speakers who completed the test and provided more than one, and sometimes numerous, answer options for most content (open class) word deletions. We therefore leave the choice of the scoring method open for future test uses. We provide the language background questionnaire, the cloze deletion test, answer key, and score ranges characterizing each of the proficiency levels obtained based on our data samples in Appendices A–E.

8. Conclusions

Our analyses provide evidence that the Russian cloze deletion test reported in this study reliably discriminates between dissimilar target language attainment levels in diverse populations of bilingual and multilingual Russian speakers. Our participant sample, in all its heterogeneity, is highly representative of the population of the Russian (heritage) bilinguals and learners in the US and beyond, including a growing number of adult L2 learners of Russian who acquire the language for education- and career-related purposes in Russia. As Russian universities are gradually increasing the number of international students, the population of adult learners acquiring the language in an immersion context is expanding, offering SLA researchers more opportunities to conduct empirical work on various questions related to the acquisition of Russian morphological paradigms, variable constituent orders, and other phenomena which await further investigation.

Since its creation, the Russian cloze deletion test has had a number of applied uses in SLA and Bilingualism research. To illustrate, two of the authors of this paper have implemented the CR test in Ionin and Luchkina (2019), investigating the acquisition of quantifier scope in L2 and heritage Russian. The MC test was successfully used in experimental research by Jang et al. (in preparation), with focus on the acquisition of the Russian tense system by adult L1-Chinese learners of Russian. Ionin et al. (2020, under review), and Luchkina et al. (in progress) have all used the MC version of the test as a gauge of Russian proficiency while examining the information structure-word order relationship in heritage and L2 Russian. It is our hope that the proficiency assessment tools provided in this study will enable more investigations of this kind.

In each of the aforementioned studies, participant cloze test performance was instrumental for teasing apart the effects of TL proficiency and other sources of non-target-like performance, such as cross-linguistic influence. Without a doubt, implementing an assessment method that is independent of often ambiguous measures proxying TL proficiency, such as participants' course level or self-reported proficiency estimates, is a first yet critical step towards accounting for the inter-subject variability inherent to SLA and Bilingualism research.

Author Contributions: Data curation, T.L. and A.S.; Methodology, T.L., T.I. and A.S.; Project administration, T.L., N.L. and N.S.; Validation, T.L.; Visualization, T.L.; Writing—original draft, T.L.; Writing—review & editing, T.L. and T.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of University of Illinois at Urbana-Champaign (IRB # 15248, 9/15/14).

Informed Consent Statement: All subjects gave their informed consent for inclusion before they participated in the study.

Data Availability Statement: The data reported in the present study were collected as a part of multiple studies (some ongoing at the time of this article's publication) and are therefore not publicly available.

Acknowledgments: The authors would like to thank coordinators and instructors in the Russian language programs at the University of Indiana, University of Illinois at Urbana-Champaign, and Orel State University for facilitating data collection for this study.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Cloze Deletion Test—The CR Format (English Below)

Мыши умеют петь. Поют, (1)_____правило самцы, в надежде (2)_____ симпатию самок. Это еще (3)_____ 2005 году обнаружили ученые (4)_____ Университета Вашингтона, доказав, что (5)_____ звуки, которые издают влюбленные (6)_____ отнюдь не случайные. Из (7)_____ звуков сотканы узнаваемые музыкальные (8)_____.
К сожалению, мышиные песни находятся (9)_____ той части диапазона, который (10)____ доступен для восприятия людьми. (11)_____ "произведения" можно уловить только (12)_____ помощью специальных приборов.

Мыши (13)_____ очень старательно—не менее (14)_____ и разнообразно, чем певчие (15)_____. Иногда кажется, что они (16)_____ почти осмысленно менять высоту (17)_____ продолжительность своих удивительных (18)_____.
Через 7 лет в результате (19)_____, которое провели в 2012 (20)_____ американские ученые из Университета Дюка (21)_____ Северной Каролине обнаружилось: мыши (22)_____ еще обучаться пению, запоминать (23)_____ воспроизводить новые подслушанные мелодии, (24)_____ не только свои собственные (25)_____. Могут даже петь хором. (26)_____ животных таким талантом обладает (27)_____ человек.

Ученые собрали несколько мышиных самцов в (28) _____ клетку, рядом в другую— (29)_____ самку, и слушали, что (30)_____ ей споют. Через некоторое (31)_____ самцы словно бы спевались (32)_____ начинали распевать хором. При (33)_____ основной мелодией становилась та, (34)_____ предлагал более сильный самец.

(35)_____ раньше японские ученые из (36)_____ экспериментальных технологий Университета Осаки (37)_____ о похожем выдающимся открытии (38) _____их проекте"Эволюция мыши". (39)_____ настолько продвинули группу мышей (40)_____ эволюционной лестнице, что они (41)_____ них запели как птички. Началось (42)_____ с одной мыши, у (43)_____ неожиданно обнаружились вокальные (44)_____ , рассказывает главный руководитель экспериментов. (45)_____ у нас целый хор— (46)____ ста поющих мышей.

Поющая мышь (47)_____ после того, как ее предкам был (48)_____ ген, который отвечает у людей (49)_____ развитие речи. После этого (50)_____ дали этим генетически модифицированным (51)_____ свободно размножаться—то есть, (52)_____ по эволюционной лестнице. И наконец, (53)_____ один прекрасный день, на свет (54)_____ мышь, которая запела. Эксперименты (55)_____ еще одну цель. Ученые (56)_____ пытаться добраться до истоков возникновения речи. И (57)_____ мышах, по сути, моделируют (58)_____ процесс.

English Translation (without blanks):

Mice can sing. As a rule, male mice sing in hopes of being favored by females. This was discovered back in 2005 by researchers from the University of Washington, who proved that the melodious sounds produced by mice are indeed music-like, recognizable compositions. Sadly, the songs are produced at frequencies which cannot be perceived by human ear. The songs of mice can only be registered using special equipment. Mice are elaborate singers. They sing as intricately as birds. In other words, they seem to have control over the pitch and length of the produced sounds.

Seven years later after the original discovery, research conducted in 2012 by American scientists from Duke University in North Carolina showed that mice can also be taught how to sing and have the ability to memorize and reproduce melodies produced by others. They can even sing as a group. In the animal world, such talent is usually considered to be found among humans only.

Scientists placed several male mice in one cage, placed a female in a cage nearby and listened. After some time, males were singing as a group, in unison. At the same time, the melody produced by the strongest male was used as the main tune during the group singing.

A bit earlier, researchers from the University of Osaka in Japan reported a similar remarkable discovery in a project called "The Evolution of Mice". The team was able to advance the singing abilities of mice such that they began singing like birds. "It all began with one mouse that demonstrated signing abilities."-shares the lead researcher. "Now we have a large group of singing mice—about a hundred of them".

A singing mouse was born after researchers gave its ancestors a gene which is responsible for speech abilities in humans. The genetically modified mice were allowed to procreate freely, and therefore, to evolve. Finally, one day, a singing mouse was born. Researchers pursue one other goal by conducting experiments with mice. They search for origins of speech abilities in humans and use mice to model the speech evolution process.

Appendix B. Cloze Deletion Test—The MC Format

Мыши умеют петь. Поют, (1) так/ или/ как правило самцы, в надежде (2) завоевать/обладать/раскрыть симпатию самок. Это еще (3) к/ в/ с 2005 году обнаружили ученые (4) на/из/по Университета Вашингтона, доказав, что (5) мелодичным/мелодичные/мелодичных звуки, которые издают влюбленные (6) мыши/пары/ученые, отнюдь не случайные. Из (7) этих/тех/таких звуков сотканы узнаваемые музыкальные (8) мелодии/композиции/записи. К сожалению, мышиные песни находятся (9) к/в/с той части диапазона, который (10) не/но/ни доступен для восприятия людьми. (11) Наши/их/ваши "произведения" можно уловить только (12) с/без/о помощью специальных приборов.

Мыши (13) пищат/зовут/поют__очень старательно—не менее (14) тихо/плавно/затейливо и разнообразно, чем певчие (15) птицы/звери/животные.
Иногда кажется, что они (16) будут/могут/учатся почти осмысленно менять высоту (17) и/а/о продолжительность издаваемых (18) произведений/сочинений/звуков.
Через 7 лет в результате (19) анализа/исследования/расследования, которое провели в 2012 (20) год/году/годомериканские ученые из Университета Дюка (21) к/ в/ с Северной Каролине обнаружилось: мыши (22) способны/готовы/уверены еще обучаться пению, запоминать (23) и/а/о воспроизводить новые подслушанные мелодии, (24) и/а/о не только свои собственные (25) песни/стихи/рассказы.
Мыши могут даже петь хором. (26) из/вне/среди животных таким талантом обладает (27) иногда/лишь/так же человек.

Ученые собрали несколько мышиных самцов в (28) одну/вторую/дорогую клетку, рядом—в другую— (29) поставили/посадили/положили самку и слушали, что (30) он/она/они ей споют.
Через некоторое (31) место/время/расстояние самцы словно бы спевались (32) и/а/о начинали распевать хором. При (33) этом/этого/этому, основной мелодией становилась та, (34) некоторую/которую/которыми предлагал более сильный самец.

(35) чуть/очень/совсем раньше японские ученые из (36) музея/школы/рынка экспериментальных технологий Университета Осаки (37) сообщали/сообщал/сообщили о похожем выдающимся открытии (38) в/из/к их проекте "Эволюция мыши". (39) иностранцы/ученые/учителя настолько продвинули группу мышей (40) на/вдоль/по эволюционной лестнице, что они (41) и/а/у них запели как птички. Началось (42) всё/весь/вся с одной мыши, у (43) котороя/которой/которую обнаружились вокальные (44) ноты/инструменты/способности—рассказывает главный руководитель экспериментов. (45) теперь/тогда/скоро у нас целый хор— (46) почти/около/наверное ста поющих мышей.

Поющая мышь (47) получился/получилась/получилось после того, как ее предкам был (48) выдан/задан/дан ген, который отвечает у людей (49) от/из/за развитие речи. Потом (50) исследователи/родители/врачи дали этим генетически модифицированным (51) мышам/мышами/мышах свободно размножаться—то есть, (52) двигать/двигался/

двигаться по эволюционной лестнице. И наконец, “(53) в/на/с один прекрасный день” на свет (54) появится/появилась/появляется мышь, которая запела. Эксперименты (55) привлекают/приследуют/предлагают еще одну цель. Ученые (56) были/будут/будущие пытаться добраться до истоков возникновения речи. И (57) в/о/на мышах, по сути, моделируют (58) этот/тот/то процесс.

Appendix C. Answer Key

Table A1. Answer Key.

Test Item	Deletion	Other Acceptable Answers
1	как	как правило
2	завоевать	вызывать, привлечь, заполучить, получить
3	в	
4	из	
5	мелодичные	эти, некоторые, разные, необычные
6	мыши	самцы, грызуны
7	этих	разных, различных, подобных
8	композиции	мелодии, ноты, песни, произведения
9	в	на
10	не	
11	их	мышиные, эти, все, музыкальные, подобные
12	с	
13	поют	
14	затейливо	красиво, продолжительно, богато, старательно, оригинально, выразительно, умело, заливисто, виртуозно, осмысленно, изобретательно
15	птицы	
16	умеют	могут
17	и	
18	звуков	
19	исследования	эксперименты, наблюдения
20	году	
21	в	
22	способны	
23	и	
24	а [1]	причем
25	песни	произведения, творения, мелодии, звуки, трели
26	среди	кроме, из
27	лишь	только
28	одну	
29	посадили	поместили
30	они	самцы
31	время	
32	и	
33	этом	
34	которую	
35	чуть	еще, немного
36	школы	отдела, лабаротории, института, центра, факультета
37	сообщили	написали, доложили, объявили, рассказали, заявили
38	в	об
39	ученые	они, исследователи
40	по	
41	у	
42	всё	это, исследование
43	которой	

Table A1. *Cont.*

Test Item	Deletion	Other Acceptable Answers
44	способности	данные
45	теперь	сейчас, тут
46	около	более
47	получилась	появилась, запела, эволюционировала, стала, родилась
48	дан	внедрен, приобретен, пересажен, передан, привит, введен, вживлен
49	за	
50	исследователи	ученые, они, исследователи, экспериментатор
51	мышам	животным
52	двигаться	продвигаться, двигаться, подниматься, идти, взбираться
53	в	
54	появилась	родилась
55	преследуют	достигли, имели, преследовали, ставили, имеют
56	пытаются	хотят, стремятся, хотели, смогли, сумели, стараются, пытались, решили
57	на	
58	этот	данный, эволюционный, генетический, весь

[1] (The item discriminability analysis reported in Section 6.1 yielded a negative discriminability coefficient for item 24, meaning that in the assessment completed in the present study, this item did not discriminate between different TL attainment levels).

Appendix D. Accuracy Ranges for Four Distinct Proficiency Levels

Native speakers' accuracy: 98–100%.

The CR format:

Level 0: Not tested	(The initial acquisition stage)
Level 1: 5–19% accurate	(Low)
Level 2: 20–44% accurate	(Low-mid)
Level 3: 45–74% accurate	(Mid)
Level 4: 75–100% accurate	(High)

Multiple-choice format:

Level 0: Not tested	(The initial acquisition stage)
Level 1: 5–19% accurate	(Low)
Level 2: 20–44% accurate	(Low-mid)
Level 3: 45–74% accurate	(Mid)
Level 4: 75–100% accurate	(High)

Appendix E. Language Background Questionnaire

The English version:

Language Experience and Proficiency Questionnaire (section 1 of 2)

subject #		list #	today's date:
age		male ☐	female ☐

1 Please list all languages you know in order of dominance and estimate your proficiency in each language.

1. Language A	2. Language B	3. Language C	4. Language D	5. Language E

2 Please list all the languages you know in order of acquisition and indicate how old were you when you started learning each language (your native language first).

1. Language A	2. Language B	3. Language C	4. Language D	5. Language E

3 Please list what percentage of the time you currently use and are exposed to each language. (Your percentages should add up to 100%.)

list lg. here	Lg. A:	Lg. B:	Lg. C:	Lg. D:	Lg. E:
list % here					

4 In choosing to read a text available in all your languages, in what percentage of cases would you choose to read it in each of your languages? (Assume that the original was written in another language, which is unknown to you. Your percentages should add up to 100%.)

list lg. here	Lg. A:	Lg. B:	Lg. C:	Lg. D:	Lg. E:
list % here					

5 When choosing to speak with a person who is equally fluent in all your languages, what percentage of time would you choose to speak each language? Please report percentage of total time. (Your percentages should add up to 100%.)

list lg. here	Lg. A:	Lg. B:	Lg. C:	Lg. D:	Lg. E:
list % here					

6 Please indicate what the native languages of your parents are:

_____

7 How many years of formal education do you have? _____
Please check your highest education level (or the approximate US equivalent to a degree in another country):

☐ less than high school	☐ some college	☐ master's degree
☐ high school	☐ college	☐ PhD/MD/JD
☐ professional training	☐ some graduate school	☐ other:

8 Have you ever had a vision problem, hearing impairment, language disability, or learning disability? (Check if applicable.) If yes, please explain, including any corrections or treatments:_____

Language Experience and Proficiency Questionnaire (section 2 of 2)

Language: Russian
All questions below refer to your knowledge of this language.

1 Age when you...

began acquiring Russian:	became fluent in Russian:	began reading in Russian:	became fluent in reading Russian:

2 Please list the number of years and months you spent in each language environment.

	years	months
A country where Russian is spoken		
A family where Russian is spoken		
A school and/or working environment where Russian is spoken		

3 On a scale from zero to ten, please rate your level of proficiency in speaking, understanding, and reading Russian.

speaking		understand spoken language		reading	

4 On a scale from zero to ten, please rate how much of the following factors contributed to you learning Russian.

interacting with friends		listening to radio/music	
interacting with family		reading	
watching TV/video/web videos		language lab/self-instruction	

5 Please rate to what extent you are currently exposed to Russian in the following contexts:

interacting with friends		listening to radio/music	
interacting with family		reading	
watching TV/video/web videos		language lab/self-instruction	

6 In your perception, how much of a foreign accent do you have in Russian? _____

7 Please rate how frequently others identify you as a non-native speaker based on your accent in Russian?

Bilingual LEAP Questionnaire, adapted from:
Marian et al., 2007. J. Speech, Language, & Hearing Research, v. 50, p. 940–967.

The Russian version:

Языковой опыт иуровень знания: Часть1

#участника		#списка		дата	
возраст		Дата рождения		Муж ☐	Жен ☐

1 Пожалуйста, перечислите, какими языками вы владеете в порядке уменьшения языковых навыков.

1. ЯзыкА	2. Язык Б	3. ЯзыкВ	4. ЯзыкГ	5. Язык Д

2 Пожалуйста, перечислите, какими языками вы владеете в порядке освоения. (родной язык первый).

1. ЯзыкА	2. Язык Б	3. ЯзыкВ	4. ЯзыкГ	5. Язык Д

3 Пожалуйста, укажите в процентномсоотношении, сколько вы пользуетесь каждым языком в настоящее время. (Общаясумма процентов должна быть 100%)

Язык	А:	Б:	В:	Г:	Д:
%					

4 Если текст написан на языке Вам неизвестном, но переводВам доступен на всех известных Вам языках. Укажите процентное соотношение выбора Вами того или иного языка для прочтения такого текста. (Общаясумма процентов должна быть 100%)

Язык	А:	Б:	В:	Г:	Д:
%					

5 Укажите процентное соотношение выбора Вами того или иного языка для общенияс собеседником, который одинаковосвободно владеет всеми известнымиВам языками. (Общаясумма процентов должна быть100%)

Язык	А:	Б:	В:	Г:	Д:
%					

6 С какими культурами(этническими, религиозными) высебяассоциируете. Оцените на шкале от0 до 10 степень принадлежности к каждой группе (Например: Русский, Православный и т.д).

Культура	А:	Б:	В:	Г:	Д:
Шкала (0–10)					

7 Сколько лет вы провели всистеме образования? _____
Укажите уровень вашего образования. Если вы обучались в другойстране укажите наиболее подходящий эквивалент:

☐ не законченное ☐ среднее ☐ среднеесреднее профессиональное	☐ не законченное высшее ☐ высшее ☐ аспирантура	☐ кандидат наук ☐ доктор наук

8 Сколько лет вы жили вСША: _____
Если вы жили в других странах, укажите, в каких исколько лет: _____

9 Были лиу вас когда-либо проблемысо зрением, слухом, илиречью? Если да, тоукажите, какие именно(включая коррекционные процедуры, если были):

Языковой опыт и уровень знания: Часть 2 а

Язык: Русский.
Все последующие вопросы относятся к вашему знанию вышеуказанного языка.

1 В каком возрасте вы:

НачалиучитьРусский:	ВладелисвободноРусским:	Начали читать по-Русски:	Читалисвободно по-Русски:

2 Укажите количество лет и месяцев вы провели в:

	лет	месяцев
Русско-говорящейстране		
Русско-говорящейсемье		
Русско- говорящей школе/работе		

3 Оцените на шкале от0 до 10 ваше умение говорить, понимать, и читать по-Русски:

Речь:		Понимание:		Чтение:	

4 Оцените на шкале от0 до 10 влияние следуюших факторов на ваше изучение Русского:

Общение с друзьями		Радио/Музыка	
Общение всемье		Чтение	
Телевизор		Изучение языка с учителем/Самостоятельное изучение	

5 Оцените на шкале от0 до 10 как часто вы используете Русский вследующих ситуациях:

Общение с друзьями		Радио/Музыка	
Общение всемье		Чтение	
Телевизор		Изучение языка с учителем/Самостоятельное изучение	

6 По-вашему мнению, когда вы говорите по-Русски, насколькосильно присутствуету вас иностранныйакцент?

_____

7 Как часто другие принимают вас за иностранца из-за вашегоакцента?

_____

Языковой опыт и уровень знания: Часть 2 б

Язык: Английский.
Все последующие вопросы относятся к вашему знанию вышеуказанного языка.

1 В каком возрасте вы:

НачалиучитьАнглийский:	ВладелисвободноАнглийским:	Начали читать по-Английски:	Читалисвободно по-Английски:

2 Укажите количество лет и месяцев вы провели в:

	лет	месяцев
Англоговорящейстране		
Англоговорящейсемье		
Англоговорящей школе/работе		

3 Оцените на шкале от 0 до 10 ваше умение говорить, понимать, и читать по-Английски:

Речь:		Понимание:		Чтение:	

4 Оцените на шкале от 0 до 10 влияние следуюших факторов на ваше изучение Английского:

Общение с друзьями		Радио/Музыка	
Общение всемье		Чтение	
Телевизор		Изучение языка с учителем/Самостоятельное изучение	

5 Оцените на шкале от 0 до 10 как часто вы используете Английский вследующих ситуациях:

Общение с друзьями		Радио/Музыка	
Общение всемье		Чтение	
Телевизор		Изучение языка с учителем/Самостоятельное изучение	

6 По-вашему мнению, когда вы говорите по-Английски, насколькосильно присутствуету вас иностранныйакцент?

7 Как часто другие принимают вас за иностранца из-за вашегоакцента?

Языковой опыт и уровень знания: Часть 2 в

Язык:
Все последующие вопросы относятся к вашему знанию вышеуказанного языка.

1 В каком возрасте вы...

Началиучить:	Владелисвободно:	Начали читать по-:	Читалисвободно по-:

2 Укажите количество лет и месяцев вы провели в:

	лет	месяцев
стране		
семье		
школе/работе		

3 Оцените на шкале от 0 до 10 ваше умение говорить, понимать, и читать:

Речь:		Понимание:		Чтение:	

4 Оцените на шкале от 0 до 10 влияние следуюших факторов на ваше изучение:

Общение с друзьями		Радио/Музыка	
Общение всемье		Чтение	
Телевизор		Изучение языка с учителем/Самостоятельное изучение	

5 Оцените на шкале от 0 до 10 как часто вы используете вследующих ситуациях:

Общение с друзьями		Радио/Музыка	
Общение всемье		Чтение	
Телевизор		Изучение языка с учителем/Самостоятельное изучение	

6 По-вашему мнению, когда вы говорите, насколькосильно присутствуету вас иностранныйакцент?

7 Как часто другие принимают вас за иностранца из-за вашегоакцента?

Bilingual LEAP Questionnaire, adapted from:
Marian et al., 2007. J. Speech, Language, & Hearing Research, v. 50, p. 940–67.

References

Abraham, Roberta G., and Carol A. Chapelle. 1992. The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal* 76: 468–79. [CrossRef]

Abutalebi, Jubin. 2008. Neural aspects of second language representation and language control. *Acta Psychologica* 128: 466–78. [CrossRef] [PubMed]

Alderson, J. Charles. 1979. The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly* 13: 219–27. [CrossRef]

Alderson, J. Charles. 1980. Native and nonnative speaker performance on cloze tests. *Language Learning* 30: 59–76. [CrossRef]
Antoniou, Mark, Eric Liang, Marc Ettlinger, and Patrick C. M. Wong. 2015. The bilingual advantage in phonetic learning. *Bilingualism* 18: 683. [CrossRef]
Bachman, Lyle F. 1985. Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly* 19: 535–56. [CrossRef]
Baghaei, Purya, and Hamdollah Ravand. 2019. Method bias in cloze tests as reading comprehension measures. *SAGE Open* 9: 2158244019832706. [CrossRef]
Baker, Colin, and Sylvia Prys Jones, eds. 1998. *Encyclopedia of Bilingualism and Bilingual Education*. Bristol: Multilingual Matters.
Baldauf, Richard B., Jr., and Ivan K. Propst Jr. 1979. Matching and multiple-choice cloze tests. *The Journal of Educational Research* 72: 321–26. [CrossRef]
Basenko-Karmali, N. A., and N. A. Saparova. 2020. *Testirovanije detej po russkomu jazyku: Opyt organizacii i podgotovki uchenikov. [Russian Language Testing for Child Learners: Organizational Issues and Student Preparation]. Uchitelju russkoj zarubezhnoj shkoly*. St. Petersburg: Zlatoust, vol. 290.
Belyakova, L. F., N. A. Veryanova, and A. V. Dikareva. 2013. *Problemy podgotovki inostrannyh studentov k sertificirovaniyu na tretij uroven' vladeniya russkim yazykom kak inostrannym (TRKI-3). [Preparing International Students for Certification in Third Level of Proficiency in Russian as a Foreign Language (TRKI-3)]*. Izvestiya Volgogradskogo gosudarstvennogo tehnicheskogo universiteta. Series: Novye obrazovatel'nye sistemy i tehnologii obucheniya v vuze. Volgograd: Volgograd State Technical University, pp. 27–29. (In Russian)
Blumenfeld, Henrike K., and Veronica Marian. 2007. Constraints on parallel activation in bilingual spoken language processing: Examining proficiency and lexical status using eye-tracking. *Language and Cognitive Processes* 22: 633–60. [CrossRef]
Brown, James Dean. 1980. Relative merits of four methods for scoring cloze tests. *The Modern Language Journal* 64: 311–17. [CrossRef]
Brown, James D. 1983. A closer look at cloze validity and reliability. In *Issues in Language Testing Research*. Edited by John W. Oller. Rowley: Newbury House Publishers, Inc.
Brown, James Dean. 1988. Tailored cloze: Improved with classical item analysis techniques. *Language Testing* 5: 19–31. [CrossRef]
Brown, James Dean. 2002. *Do Cloze Tests Work? Or Is It Just an Illusion?* University of Hawai'i Second Language Studies Paper 21. Honolulu: University of Hawai'i.
Brown, James Dean. 2003. Norm-referenced item analysis (item facility and item discrimination). *Shiken: JALT Testing and Evaluation SIG Newsletter* 7: 16–19.
Brown, James Dean. 2009. Statistics Corner. Questions and answers about language testing statistics: Choosing the right number of components or factors in PCA and EFA. *Shiken: JALT Testing & Evaluation SIG Newsletter* 13: 19–23.
Brown, James Dean. 2013. My twenty-five years of cloze testing research: So what. *International Journal of Language Studies* 7: 1–32.
Brown, James Dean, and Theres Grüter. 2020. The same cloze for all occasions? Using the Brown (1980) cloze test for measuring proficiency in SLA research. *International Review of Applied Linguistics in Language Teaching*. [CrossRef]
Chapelle, Carol A., and Roberta G. Abraham. 1990. Cloze method: What difference does it make? *Language Testing* 7: 121–46. [CrossRef]
Chihara, Tetsuro, John Oller, Kelley Weaver, and Mary Anne Chavez-Oller. 1977. Are cloze items sensitive to constraints across sentences? *Language Learning* 27: 63–70. [CrossRef]
Conrad, Markus, Guillermo Recio, and Arthur M. Jacobs. 2011. The time course of emotion effects in first and second language processing: A cross cultural ERP study with German–Spanish bilinguals. *Frontiers in Psychology* 2: 351. [CrossRef]
Douglas, Masako O. 1994. Japanese cloze tests: Toward their construction. *Japanese Language Education Around the Globe* 4: 117–31.
Dorian, Nancy. 1981. *Language Death: The Life Cycle of a Scottish Gaelic Dialect*. Philadelphia: University of Pennsylvania Press.
Dunn, Alexandra L., and Jean E. Fox Tree. 2009. A quick, gradient bilingual dominance scale. *Bilingualism* 12: 273. [CrossRef]
Flege, James E., Elaina M. Frieda, Amanda C. Walley, and Lauren A. Randazza. 1998. Lexical factors and segmental accuracy in second language speech production. *Studies in Second Language Acquisition*, 155–87. [CrossRef]
Fotos, Sandra S. 1991. The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations? *Language Learning* 41: 313–36. [CrossRef]
Foucart, Alice, Carlos Romero-Rivas, Bernharda Lottie Gort, and Albert Costa. 2016. Discourse comprehension in L2: Making sense of what is not explicitly said. *Brain and Language* 163: 32–41. [CrossRef] [PubMed]
Frey, Bruce B., ed. 2018. *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. Thousand Oaks: Sage Publications.
Friedman, Debra, and Olga Kagan. 2008. Academic writing proficiency of Russian heritage speakers: A comparative study. In *Heritage Language Education: A New Field Emerging*. London: Routledge, pp. 181–98.
Gaillard, Stéphanie. 2014. Implementing an Elicited Imitation Task as a Component of a Language Placement Test in French at the University Level. Ph.D. thesis, University of Illinois Urbana-Champaign, Champaign, IL, USA.
Gildersleeve-Neumann, Christina E., and Kira L. Wright. 2010. English speech acquisition in 3-to 5-year-old children learning Russian and English. *Language, Speech, and Hearing Services in Schools* 41: 429–44. [CrossRef]
Gollan, Tamar H., Gali H. Weissberger, Elin Runnqvist, Rosa I. Montoya, and Cynthia M. Cera. 2012. Self-ratings of spoken language dominance: A multi-lingual naming test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. *Bilingualism (Cambridge, England)* 15: 594. [CrossRef] [PubMed]
Gor, Kira, and Svetlana Cook. 2010. Nonnative processing of verbal morphology: In search of regularity. *Language Learning* 60: 88–126. [CrossRef]
Gor, Kira. 2019. Morphosyntactic knowledge in late second language learners and heritage speakers of Russian. *Heritage Language Journal* 16: 124–50. [CrossRef]

Grosjean, François. 2004. Studying bilinguals: Methodological and conceptual issues. In *The Handbook of Bilingualism*. Hoboken: Wiley Online Library, pp. 32–63.
Han, Wen-Jui. 2012. Bilingualism and academic achievement. *Child Development* 83: 300–21. [CrossRef]
Huensch, Amanda. 2014. The Perception and Production of Palatal Codas by Korean L2 Learners of English. Ph.D. dissertation, University of Illinois at Urbana-Champaign, Champaign, IL, USA.
Hughes, Arthur. 2003. *Testing for Language Teachers*. Stuttgart: Ernst Klett Sprachen.
Hulstijn, Jan H. 2012. Incidental learning in second language acquisition. In *The Encyclopedia of Applied Linguistics*. Hoboken: John Wiley and Sons, Inc.
Hyltenstam, Kenneth, and Niclas Abrahamsson. 2003. Maturational constraints in second language acquisition. In *Handbook of Second Language Acquisition*. Edited by Catherine Doughty and Michael Long. Malden: Blackwell, pp. 539–88.
Ingvalson, Erin M., James L. McClelland, and Lori L. Holt. 2011. Predicting native English-like performance by native Japanese speakers. *Journal of Phonetics* 39: 571–84. [CrossRef]
Ionin, Tania, and Tatiana Luchkina. 2019. Scope, syntax and prosody in Russian as a second or heritage language. *Exploring Interfaces: Lexicon, Syntax, Semantics and Sound*, 141–70.
Ionin, Tania, Maria Goldshtein, Tatiana Luchkina, and Sofya Styrina. 2020. Word Order and Information Structure in Russian as a Heritage or Second Language. Paper presented at 44th Boston University Conference on Language Development, Boston, MA, USA, November 7–10. Edited by Megan M. Brown and Alexandra Kohut. Cascadilla Press: Somerville.
Ionin, Tania, Tatiana Luchkina, and Anastasia Stoops. 2014. Quantifier scope and scrambling in the second language acquisition of Russian. Paper presented at 5th Conference on Generative Approaches to Language Acquisition–North America, Lawrence, KS, USA, October 11–13, pp. 169–80.
Ionin, Tania, Maria Goldshtein, Luchkina Tatiana, and Sofya. Styrina. Under review. Who did what to whom, and what did we already know? Word order and information structure in heritage and L2 Russian.
Jang, C., A. Myachikov, and V. Shtyrov. In preparation. Acquisition of Russian tense by Chinese-speaking learners: A self-paced reading investigation. (in preparation)
Johnson, Jacqueline S., and Elissa L. Newport. 1989. Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* 21: 60–99. [CrossRef]
Jonz, John. 1990. Another turn in the conversation: What does cloze measure? *TESOL Quarterly* 24: 61–83. [CrossRef]
Kaushanskaya, Margarita, Henrike K. Blumenfeld, and Viorica Marian. 2020. The language experience and proficiency questionnaire (leap-q): Ten years later. *Bilingualism: Language and Cognition* 23: 945–50. [CrossRef]
Kim, Hyunwoo, and Yangon Rah. 2019. Constructional Processing in a Second Language: The Role of Constructional Knowledge in Verb-Construction Integration. *Language Learning* 69: 1022–56. [CrossRef]
Kleijn, Suzanne. 2018. *Clozing in on Readability: How Linguistic Features Affect and Predict Text Comprehension and On-Line Processing*. Utrecht: LOT.
Kobayashi, Miyoko. 2002. Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *The Modern Language Journal* 86: 571–86. [CrossRef]
Kohnert, Kathryn J., and Elizabeth Bates. 2002. Balancing bilinguals II: Lexical comprehension and cognitive processing in children learning Spanish and English. *Journal of Speech, Language & Hearing Research* 45: 347–59.
Kormos, Judit. 2000. The role of attention in monitoring second language speech production. *Language Learning* 50: 343–84. [CrossRef]
Kotz, Sonja A. 2009. A critical review of ERP and fMRI evidence on L2 syntactic processing. *Brain and Language* 109: 68–74. [CrossRef] [PubMed]
Laleko, Oksana. 2011. Restructuring of verbal aspect in heritage Russian: Beyond lexicalization. *International Journal of Language Studies* 5: 13–26.
Lemhöfer, Kristin, and Mirjam Broersma. 2012. Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods* 44: 325–43. [CrossRef]
Long, Michael H., Kira Gor, and Scott Jackson. 2012. Linguistic correlates of second language proficiency: Proof of concept with ILR 2–3 in Russian. *Studies in Second Language Acquisition* 34: 99–126. [CrossRef]
Lubensky, Sophia, Gerard L. Ervin, Larry McLellan, and Donald K. Jarvis. 2002. *Nachalo*, 2nd ed. Books 1 and 2. New York: McGraw Hill.
Luchkina, Tatiana, and Anastasia Stoops. 2013. Cloze test as a measure of L2 Russian proficiency. Paper presented at SLRF (Second Language Research Forum), Provo, UT, USA, November 2.
Luk, Gigi, and Ellen Bialystok. 2013. Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology* 25: 605–21. [CrossRef]
Lyashevskaya, O., and S. Sharov. 2009. *Chastotnyj slovar' sovremennogo russkogo jazyka (na materialakh Nacional'nogo korpusa russkogo jazyka) [The Frequency Dictionary of Modern Russian Language Based on Russian National Corpus]*. Moscow: Azbukovnik. (In Russian)
Makarova, Veronika, and Natalia Terekhova. 2017. Russian language proficiency of monolingual and Russian–English bi/multilingual children. *OLBI Journal* 8: 53–69. [CrossRef]
Marian, Viorica, Henrike K. Blumenfeld, and Margarita Kaushanskaya. 2007. The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research* 23: 945–50. [CrossRef]

Mercier, Julie, Irina Pivneva, and Debra Titone. 2014. Individual differences in inhibitory control relate to bilingual spoken word processing. *Bilingualism: Language and Cognition* 17: 89–117. [CrossRef]
Montanari, Simona, Robert Mayr, and Kaveri Subrahmanyam. 2018. Bilingual speech sound development during the preschool years: The role of language proficiency and cross-linguistic relatedness. *Journal of Speech, Language, and Hearing Research* 61: 2467–86. [CrossRef]
Montrul, Silvina. 2002. Incomplete acquisition and attrition of Spanish tense/aspect distinctions in adult bilinguals. *Bilingualism* 5: 39. [CrossRef]
Montrul, Silvina. 2006. On the bilingual competence of Spanish heritage speakers: Syntax, lexical-semantics and processing. *International Journal of Bilingualism* 10: 37–69. [CrossRef]
Montrul, Silvina. 2018. Heritage language development: Connecting the dots. *International Journal of Bilingualism* 22: 530–46. [CrossRef]
Montrul, Silvina, and Tania Ionin. 2012. Dominant language transfer in Spanish heritage speakers and second language learners in the interpretation of definite articles. *The Modern Language Journal* 96: 70–94. [CrossRef]
Montrul, Silvina, Rebecca Foote, and Silvia Perpiñán. 2008. Gender agreement in adult second language learners and Spanish heritage speakers: The effects of age and context of acquisition. *Language Learning* 58: 503–53. [CrossRef]
Mostow, Jack, and Hyeju Jang. 2012. Generating diagnostic multiple choice comprehension cloze questions. Paper presented at Seventh Workshop on Building Educational Applications Using NLP, Montreal, QC, Canada, June 7; pp. 136–46.
Nip, Ignatius S. B., and Henrike K. Blumenfeld. 2015. Proficiency and linguistic complexity influence speech motor control and performance in Spanish language learners. *Journal of Speech, Language, and Hearing Research* 58: 653–68. [CrossRef]
Nunan, David, and Ronald Carter, eds. 2001. *The Cambridge Guide to Teaching English to Speakers of Other Languages*. Stuttgart: Ernst Klett Sprachen.
Pelham, Sabra D., and Lise Abrams. 2014. Cognitive advantages and disadvantages in early and late bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 40: 313. [CrossRef] [PubMed]
Polinsky, Maria, and Olga Kagan. 2007. Heritage languages: In the 'wild' and in the classroom. *Language and Linguistics Compass* 1: 368–95. [CrossRef]
Polinsky, Maria. 2005. Word class distinctions in an incomplete grammar. In *Perspectives on Language and Language Development*. Boston: Springer, pp. 419–34.
Polinsky, Maria. 2006. Incomplete acquisition: American Russian. *Journal of Slavic Linguistics* 1: 191–262.
Polinsky, Maria. 2007. Reaching the end point and stopping midway: Different scenarios in the acquisition of Russian. *Russian Linguistics* 31: 157–99. [CrossRef]
Polinsky, Maria. 2008. Relative clauses in heritage Russian: Fossilization or divergent grammar. In *Formal Approaches to Slavic Linguistics*. Ann Arbor: Michigan Slavic, vol. 16, pp. 333–58.
Polinsky, Maria. 2011. Reanalysis in adult heritage language: New evidence in support of attrition. *Studies in Second Language Acquisition* 1: 305–28. [CrossRef]
Shanahan, Timothy, Michael L. Kamil, and Aileen Webb Tobin. 1982. Cloze as a measure of intersentential comprehension. *Reading Research Quarterly* 16: 229–55. [CrossRef]
Sheng, Li, Ying Lu, and Tamar H. Gollan. 2014. Assessing language dominance in Mandarin-English bilinguals: Convergence and divergence between subjective and objective measures. *Bilingualism (Cambridge, England)* 17: 364. [CrossRef] [PubMed]
Sorace, Antonella. 2004. Native language attrition and developmental instability at the syntax-discourse interface: Data, interpretations and methods. *Bilingualism Language and Cognition* 7: 143–45. [CrossRef]
Storey, Peter. 1997. Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing* 14: 214–31. [CrossRef]
Taylor, Wilson L. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly* 30: 415–33. [CrossRef]
Trace, Jonathan. 2020. Clozing the gap: How far do cloze items measure? *Language Testing* 37: 235–53. [CrossRef]
Tracy-Ventura, Nicole, Kevin McManus, John M. Norris, Lourdes Ortega, and P. Leclercq. 2014. Repeat as much as you can': Elicited imitation as a measure of oral proficiency in L2 French. In *Measuring L2 Proficiency: Perspectives from SLA*. Multilingual Matters. Bristol: CCSD, pp. 143–66.
Tremblay, Annie. 2009. Phonetic variability and the variable perception of L2 word stress by French Canadian listeners. *International Journal of Bilingualism* 13: 35–62. [CrossRef]
Tremblay, Annie. 2011. Proficiency assessment standards in second language acquisition research: "Clozing" the gap. *Studies in Second Language Acquisition* 33: 339–72. [CrossRef]
Tremblay, Annie, and Meryl D. Garrison. 2008. Cloze tests: A tool for proficiency assessment in research on L2 French. In *Selected Proceedings of the Second Language Research Forum*. Cascadilla Proceedings Project. Somerville: Iowa State University, pp. 73–88.
U.S. Census Bureau. 2019. *Languages Spoken at Home by Language: 2017*; Language Use in the United States. Suitland-Silver Hill: U.S. Census Bureau. Available online: https://www.census.gov (accessed on 19 February 2021).
Van den Broek, Paul, Sandra Virtue, Michelle Gaddy Everson, Yuhtsuen Tzeng, and Yung-chi Sung. 2002. Comprehension and memory of science texts: Inferential processes and the construction of a mental representation. *The Psychology of Science Text Comprehension* 1: 131–54.
Van Zeeland, Hilde, and Norbert Schmitt. 2013. Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics* 34: 457–79. [CrossRef]

Wang, Wei, Fritz Drasgow, and Liwen Liu. 2016. Classification accuracy of mixed format tests: A bi-factor item response theory approach. *Frontiers in Psychology* 7: 270. [CrossRef]
Watanabe, Yukiko, and Dennis Koyama. 2008. *A Meta-Analysis of Second Language Cloze Testing Research*. University of Hawai'i Second Langauge Studies Paper 26. Honolulu: University of Hawai'i.
White, Lydia. 2003. Fossilization in steady state L2 grammars: Persistent problems with inflectional morphology. *Bilingualism* 6: 128–41. [CrossRef]
Wood Bowden, Harriet. 2016. Assessing second-language oral proficiency for research. *Studies in Second Language Acquisition* 38: 647. [CrossRef]
Yamashita, Junko. 2003. Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing* 20: 267–93. [CrossRef]
Yan, Xun, Yuyun Lei, and Chilin Shih. 2020. A corpus-driven, curriculum-based Chinese elicited imitation test in US universities. *Foreign Language Annals* 53: 704–32. [CrossRef]