Preprint of paper published in Poznań Studies in Contemporary Linguistics 58(3), 2022
https://doi.org/10.1515/psicl-2022-0020

PREDICTING FOREIGN LANGUAGE SKILLS BASED ON FIRST LANGUAGES:

THE ROLE OF LEXICAL DISTANCE AND RELATIVE MORPHOLOGICAL COMPLEXITY

RAPHAEL BERTHELÉ, PETER LENZ, ELISABETH PEYER
University of Fribourg
raphael.berthele@unifr.ch

supplementary material: https://osf.io/9tqmx/?view_only=e7f1883585f948e9ae1c71c8b6e8cb51

ABSTRACT

Research on adult second language learning shows the importance of the linguistic proximity of acquired languages to the target language as a predictor of learning. Not much research has been done on the impact of linguistic distance in foreign language learning settings. We analysed the speaking, writing, listening, and reading skills of multilingual 14-year-old French learners (N = 409) in a German-speaking context who indicated at least one language other than German as (a) L1. Using mixed-effect models, we tested associations between linguistic contrasts between French and the individuals' first languages and their success on tests of French as a foreign language. The models also controlled for motivation, curriculum-related variables as well as social, economic and educational background information. Results show a small effect for lexical distance with all four skills, while relative morphological complexity seems negligible. The study therefore shows that the impact of lexical distance is measurable not only in immersive second language learning settings, but also in foreign language instruction settings with very limited exposure to the target language.

KEYWORDS: linguistic distance; foreign language learning; French; multilingualism; transfer.

1. Introduction

Studies in immersive second language learning settings have shown a robust impact of linguistic distance on second language speaking skills in adult immigrants (Schepens 2014; Schepens et al. 2013 and 2016). Whereas linguistic distance seems to be robustly associated with the pace of learning in these immersion contexts, instructional guidance in the language learning process

could arguably cancel out the distance effects found in (largely) untutored settings. This study therefore investigates if an impact of linguistic distance is also measurable in foreign language instruction settings, i.e. in settings with very limited hours of contact with the target language and with learning that is strongly shaped and mediated by teachers and teaching materials. More specifically, this article examines if linguistic distance between first languages and French may have an effect on skills in French as a foreign language. We discuss data collected in a large-scale assessment study of French in the German-speaking part of Switzerland (Peyer et al. 2016). Although French is an official language and spoken by the largest minority in the Western part of multilingual Switzerland, the country's territorial regime imposes local institutional monolingualism in most areas except in specific zones close to the traditional linguistic borders. Therefore, we use the term 'foreign language' for French despite its official status in Switzerland. The study focuses on the performance of 14-year-old students learning French as a compulsory second foreign language after English. They were tested in French listening, reading, speaking and writing. We use the four measures of French foreign language skills to study associations between these skills and certain properties of languages or language varieties in the repertoires of the multilingual students in the sample.

2. Background: the role of linguistic distance in language learning

2.1. Contrastive approaches in the past

The study of learning has had a long-standing interest in the transfer of skills and knowledge to a new domain of application (Bransford et al. 2000, chapter 3). A robust finding regarding such transfer is that its likelihood rises with increasing degree of overlap between the source and the target domains (a topic already investigated by Thorndike and Woodworth 1901). With regard to the learning and use of closely related languages, the question arises as to what degree transfer or cross-linguistic influence can be predicted by structural and lexical overlap between a source and a target language. Past and current thinking about such transfer processes in the domain of language learning draw on three different but complementary perspectives: Focus on structural similarity or difference, focus on language skills, and focus on curricula.

Our study examines the effects of two measures of linguistic distance, thus building on the first perspective. Most commonly, scholars refer to Lado's

Contrastive Analysis Hypothesis (CAH, Lado 1957) as an early approach to predicting difficulty in language learning based on linguistic comparison of source and target systems. However, linguistic proximity or distance as a factor to be taken into account in language learning has been subject to considerations for centuries. According to the Dominican Robert Holcot, it was recommended after the Norman conquest of England to teach Latin through French due to the similarity of the two target languages ("et per gallicum latinum", Kristol 1990). In 20th-century theorizing on foreign language learning, there is a continued interest in the particular ease that learners have in developing receptive skills in languages that are closely related to languages previously learned or acquired (Ringbom 1978; Kürschner et al. 2008). Kloss (1929) predicted closely related languages to be easy in reception but more difficult in production. Similar differential effects of comprehension versus production are also discussed in more recent studies (for an overview see Puig-Mayenco et al., 2018). In addition to the objective and measurable genealogical relatedness of language systems, scholars proposed that the perceived proximity, termed psychotypology (Kellerman 1986), was also an important factor determining the amount of transfer from source to target language.

2.2. Structure-oriented research: current findings

In recent years, scholars have investigated the impact of overlap in the domains of vocabulary and grammar. In the lexical domain, the spontaneous comprehension of cognate words in unlearnt but genealogically close languages has been investigated in different language families (Otwinowska and Szewczyk 2019; Gooskens et al., 2018; Vanhove and Berthele 2015). These studies show that, rather unsurprisingly, the degree of overlap between an acquired lexical form (in a first or second/foreign language) and a target word is a robust predictor of the probability of correct recognition of the item. In the domain of syntax and morphosyntax, recent works within formal approaches to multilingual language learning (e.g., Puig-Mayenco et al. 2020; Westergaard et al. 2017) emphasize the importance of linguistic proximity of pre-acquired languages in the acquisition of additional languages. On the other hand, there are scholars who warn against overly simplistic predictions of unconstrained transfer from acquired languages onto new target languages (Dimroth 2018). Overall, analyses of cross-linguistic influence of multilingual language learning consistently yield linguistic proximity as a predictor of target-language

production and reception. Assuming the robustness of this overall finding, linguistic proximity should have effects not only on discrete-point tests of specific items and structures but also on skills such as reading, writing, listening and speaking.

2.3. Recent findings in skills-oriented research

Recent investigations of the impact of linguistic distance on second language skills in adult immigrants in the Netherlands (Schepens 2014; Schepens et al. 2013 and 2016) show a robust association of linguistic distance, measured as lexical distance and as relative morphological complexity of source and target systems, on second language proficiency (cf. van der Slik et al. 2017; for details on the distance measures see the methods section below). These studies draw on data sets that are untypically large for our field. They focus on predictors for speaking skills specifically, and they are set in a second-language learning environment (migrants learning the local language as opposed to students learning a foreign language in a language classroom). It remains unclear, however, whether similar effects of linguistic distances on other language skills than speaking can be found or, in other words, whether all skills are similarly affected by linguistic distance or whether there are effects on specific skills but not others (e.g. on reception vs. production, or on oral vs. written skills). Also, the Dutch studies relate to a second language that is contextually strongly supported by the environment the learners migrated into. It is an open question whether similar effects can be found in foreign-language learning, i.e. the learning of a language that is mostly confined to the classroom and not spoken in the learner's immediate living environment.

As regards literacy specifically, Cummins'(e.g. 1979) framework suggests that in particular in the academic-educational domain, reading and writing skills in the languages of bi- and multilinguals are interdependent. There are mixed results on the mediating role of linguistic proximity on literacy transfer. In some cases, proximity of the languages involved shows no or very weak effects (e.g. Dressler and Kamil 2006; Berthele and Vanhove 2020). Other evidence, however, suggests the influence of linguistic proximity on skills (Proctor et al. 2010).

Multilingual approaches to language learning assume that learners of a third or additional language have a relative advantage due to enhanced language learning skills and metalinguistic awareness, among other things (Jessner 1999). Metalinguistic awareness includes the perception of linguistic dif-

ferences and similarities across languages, as suggested by Kellerman (1986), and the assumption is that this perception can be beneficial for additional language learning. There are a few investigations into bi- or multilingual schoolchildren's learning of an additional language as a compulsory foreign language in the curriculum. These studies show mixed results: in some cases, an advantage for migrant children was indeed detected (Klieme and DESI-Konsortium 2008), while other studies found no effects (Sanders and Meijers 1995; Lorenz et al. 2021) or even disadvantages for bi- and multilingual learners of additional languages (Elsner 2007; Engel et al. 2009). Other studies revealed advantages in foreign language learning for bilingual children whose bilingualism is not a consequence of international migration but of the local language situation (Cenoz and Valencia 1994; Safont Jordà 2005). At least in European contexts, large genealogical distances between immigrant and local languages can coincide with large differences in socioeconomic and educational status between immigrant and resident populations (e.g. disproportional presence of Non-Europeans and thus also of speakers of Non-Indo-European languages in the lowest strata as reported in Bartelheimer and Wolter (2016) for Germany). Therefore, it is important in such studies to control for socioeconomic status and educational background. Otherwise, these background variables are confounded with individual multilingualism so that a potential advantage of multilinguals in further language learning is masked. Moreover, different types of bi- and multilingualism involving typologically markedly different languages are often treated indifferently. Since the studies cited above do not systematically factor in linguistic distance, we do not know to what extent there is a straightforward effect of being bilingual/multilingual on additional language learning and to what extent such an effect is mediated by the linguistic distance between a first/previously learned language and the target language (see Schepens et al. 2016).

2.4. The present study

In the light of the findings laid out in the literature review above, this study starts from the assumption that linguistic distance at the level of grammar and vocabulary has a measurable impact on the acquisition of specific linguistic structures. Furthermore, in relation to the studies on typological effects on grammar learning discussed above, we assume that effects on the acquisition of words and structures will also affect the acquisition of linguistic skills that

rely on the production and/or comprehension of words and grammatical constructions. It seems reasonable to assume that linguistic skills are likely to benefit from the cumulative effects of the structural similarities or differences that can facilitate or hinder the learning of words and constructions. Thus, the question to be investigated here is to what extent linguistic distances and relative differences in grammatical complexity between languages in the individual repertoires predict learning of an additional language. More specifically, we intend to estimate the impact of linguistic differences between languages found in the repertoires of multilingual learners and the French language on the results on tests of French reading, writing, listening and speaking.

Looking at both productive and receptive skills allows, on the one hand, to test Kloss' (1929) hypothesis of a more positive effect of overlap in reception than in production. On the other hand, it allows to investigate the claim made in the interdependence framework (Cummins 1979) whether there are more transfer effects in literacy-related language activities (writing, reading) than in oral language activities.

As discussed in Section 3.4, we use two metrics that compare the first languages and the target language French, a metric for lexical overlap of language pairs and a metric for relative morphological complexity in language pairs. Our research question is as follows: are lexical distance and relative morphological complexity predictors of target language learning? From this, the following hypotheses are derived:

H0: lexical distance of previously learnt languages to French and relative morphological complexity compared to French do not predict French foreign language skills;

H1a: the larger the lexical distance between a student's L1[1] and French, the lower the French scores;

H1b: if a student's most complex L1 is less complex than French, the French score is lower.

3. Method

The hypotheses spelled out above are tested drawing on data collected for an evaluation study of French as a foreign language, mandated by the cantonal

[1] If a student's multilingual repertoire involves several previously acquired languages, the language closest to French will be considered in the analysis.

ministers of education of Central Switzerland. The impact of two linguistic metrics (lexical distance and relative morphological complexity) on French reading, writing, listening, and speaking skill scores is tested while accounting for a set of control variables that are also deemed relevant for foreign language learning outcomes.

3.1. Participants

The participants in this study stem from representative samples of pupils in 6th and 8th grade. In the following, we focus on the group of those 8th-graders who have at least one non-local language in their multilingual repertoire, normally because at least one parent speaks that language at home. We also analysed data from 6th-graders in a strictly parallel fashion, and the results are largely identical. For the sake of brevity, we focus on the group of the older students.

French is the second compulsory foreign language in this curricular re- gion. While it is compulsory in lower grades, in 8th grade, about 10% of children are exempt from the French lessons or have opted out of the subject (see Peyer et al. 2016: 9 for details). All participants studied English from 3rd grade on.

The great majority of the participants started studying French in 5th grade, also as a compulsory school subject. A small portion of the sample started learning French only in 7th grade. However, they had more teaching hours per week and used a different textbook. Total exposure of our students to the target language in the classroom varied depending on the local curricula (Swiss federalism allows for regional variation in curricula). 8th-graders had either 9, 10, 12 or 14 cumulative weekly lessons (CWL) of French since they started learning that language.[2] A total of 2212 8th-graders (aged around 14 years) took all or some of the skill-specific tests. However, the subsample used in the present study includes only those pupils who indicated at least one language other than German as a "mother tongue" (provided this additional language is not French, the target language) and who also filled out the questionnaire where they supplied the background data we needed. From this subsample, 44 cases had to be excluded because the information available on the morphological complexity of some of the first languages (L1s) involved was too sparse, or (in two cases) no lexical distance measure could be computed. The remaining pupils had a

[2] The principal purpose of the original evaluation study was to investigate the effect of this curricular factor.

total of 40 different "mother tongues". Twelve languages were mentioned by ten or more pupils. Tables A3 and A4 in the supplementary material give the lists of the languages for which linguistic measures were calculated (see details below). Our final subsample comprises 409 8th-graders overall, 390 students for listening, 337 for reading, 294 for writing, and 196 for speaking. A total of 167 pupils completed tasks in all four skills. There are various reasons why the numbers vary across the four skills. As regards speaking, practicality and cost were the main reasons that lead to a smaller subsample while the samples for reading and writing are reduced (compared to listening) because testing time was needed for additional elements of the survey. Particularly in the case of writing, the subsample was further reduced due to performances that were unusable (e.g. the pupils wrote texts in English or German instead of French, see supplementary material, section 1 for more details on the sample and data preparation).

3.2. Instruments

The tasks used to test writing, listening and reading were taken (and partly adapted) from the *lingualevel* pool of assessment tasks (Lenz and Studer 2007). These tasks are related to *real-world tasks* (Nunan 1989) and constructed to test the skills of young learners. They are based on descriptors ("can do" statements) that were developed for *lingualevel* by adapting Common European Framework of Reference (CEFR) scales and descriptors (Council of Europe 2001) to better suit the needs of adolescent learners.[3] To test a sufficient number of curricular objectives, a total of 11 writing tasks, 14 reading comprehension and 12 listening comprehension tasks was used in the test. The team of evaluators compiled them into eight different, overlapping test booklets per skill (reading, listening, and writing) so that each individual student only worked on a selection of the complete set of tasks (principle of task rotation, typical of large-scale assessments with a focus on the population rather than the individuals; e.g. OECD 2014: 30). The individual test booklets for listening and reading each contained four tasks (comprising several items), the test booklets for writing contained two to three tasks. Four test booklets were

[3] These "self-assessment descriptors for learners in lower secondary education" are available on the Council of Europe website (https://www.coe.int/en/web/portfolio/overview-of-cefr-related-scales).

compiled specifically for pupils with lower educational requirements (comprising test tasks for the CEFR levels A1 to A2) and four booklets for pupils with higher educational requirements (with test tasks covering the levels A2 to B1+). The two series of booklets were linked through common tasks.

The writing test covered a variety of writing purposes and text types. It comprised tasks such as writing a packing list, a message to a friend or a film review. The reading test included tasks such as reading a travel blog, short film reviews or an invitation to a party. The listening tasks included recordings of public announcements, short travel reports or a text about the history of the Eiffel tower. The reading and the listening comprehension tests covered a variety of text types and reading/listening styles. Comprehension was assessed by either multiple-choice, multiple-matching, true/false or short-answer questions. For assessing speaking, the evaluators developed two versions of a computer-based test: a simpler version for students in classes with lower educational requirements and a more demanding version for pupils with higher educational requirements. About one third of the items of the two versions were overlapping to ensure comparability between the two groups of learners. Both versions were related to curricular objectives and consisted of an interactive and a productive part. In the first part, students were to interact with Julien, a French-speaking boy whose picture appeared on the screen and whose voice they heard. This helped to minimize the influence different test administrators might have as all pupils heard the same recorded voice. Part of the test was concerned with personal topics. To improve comparability between the test takers, the pupils were assigned a new identity of either a girl (if the test-taker was a girl) or a boy (if the test-taker was a boy). A personal profile of these identities, consisting of drawings and some text in the language of schooling (cf. the example profile Figure A1 in the supplementary material), provided the information necessary to answer questions on the test. During the speaking test, the pupils had to put themselves in the assigned role and answer Julien's questions about "their" family and "their" hobbies. Either three or five times, depending on the version of the test, a pop-up note on the screen prompted them to ask Julien a question (e.g. "ask Julien where he lives").

After the interactive part, the learners saw two pictures of people in different situations (e.g. a market stand with fruit and vegetables or a birthday party). In a first step, the test takers were to answer the question "What is happening in the picture?". A second question per picture was used to check specific vocabulary and/or structures (e.g. "What time of day is it?", "What do people sell at the market?").

In addition to the language tests, a student questionnaire was used for further data collection. It covered various topics, such as the social and economic background, language learning motivation and anxiety. The questionnaire items on these topics were to a large extent inspired by previous research carried out in primary or secondary schools, on the one hand, the three large-scale assessment studies PISA (OECD 2012), PIRLS (BIFIE 2013) and DESI (Wagner et al. 2009), on the other hand, by a monograph on motivational aspects in young learners' language learning (Heinzmann 2013). Most items were adapted to suit the specific context and age group (see supplementary material, Section 2 for details).

3.3. Procedure

Data collection. The listening, reading and writing tests were administered as paper and pencil tests. The maximum time allowed to complete the three tests was 95 minutes in the case of 8th-graders with lower educational requirements and 120 minutes for 8th-graders with higher educational requirements. The speaking test usually took 10–15 minutes to finish.

Marking and rating. The listening and reading tests were marked according to the *lingualevel* guidelines, which include a key for all item types, including short-answer questions. The written texts were rated using a holistic writing scale that was based on the *lingualevel* writing assessment grid (Lenz and Studer 2007). In accordance with the curricular objectives, the scale focused on lexical, grammatical and textual aspects; spelling was not included. To rate the spoken performances, an assessment grid was developed which focused on lexical range and, to a lesser degree, on lexical and grammatical correctness (see Table A2 in the supplementary material).

The written and spoken texts were rated by a team of five language-teaching students who were familiar with the educational context of the study. Before the actual rating process, the raters took part in several activities to familiarize themselves with the rating scales, the tasks and a series of illustrative performance samples from this age group. In a next step, the team rated several texts individually and then discussed the ratings, in order to gain a common understanding of the rating scales. Such standardization sessions were repeated regularly during the rating process. To control rater severity/leniency and to assess inter-rater reliability, approximately one third of the written and spoken texts underwent double-blind rating.

Data preparation. Data preparation involved many steps including data entry and data cleaning, scale-building and imputation. The four language test measures were constructed by IRT-scaling (for language testing terms see ALTE 1998) the item responses (listening and reading; Masters 1982) and the ratings (writing and speaking; Linacre 1994) using the R package "TAM" (Kiefer et al. 2015). The person measures for each of these scales were output as WLEs (Warm's Weighted Likelihood Estimates; Warm 1989). The WLE reliability (indicating measurement precision; a value of 1 would express that person parameters are estimated error-free; Adams 2005) of the four language test scales is as follows: listening 0.77, reading 0.72, speaking 0.84, writing 0.80. The inter-rater reliabilities (indicating the degree of agreement between different raters or assessors; ALTE 1998) of the raw ratings amount to $\alpha = 0.93$ in the case of speaking and $\alpha = 0.92$ in the case of writing (Krippendorff's alpha coefficient for ordinal data; Krippendorff 2011; Gamer et al. 2012, see section 1 in the supplementary material for details on data preparation). All available student variables were merged into a dataset. Based on this dataset, multiple imputation was performed to replace missing values and scale values with known measurement error by plausible values (see section 1 in the supplementary material for details). The resulting ten imputed datasets formed the basis for further analyses.

3.4. Linguistic distance measures

We calculated two types of comparative measures between the first languages of our student sample and the target language French. The first is a measure of lexical distance, and the second is a measure of comparative morphological complexity.

If participants reported to have several first languages ("mother tongues"), we used the lexical distance value of the language that is closest to French. For the value of relative morphological complexity, we chose the language with the smallest increase in complexity between a student's L1 and French. If a student did not mention German, the local language of schooling, among his or her L1s, we added German because students with seriously limited knowledge of that language were excluded from the assessment.

Lexical distances. We used the ASJP62 Software (version 2.2, Wichmann et al. 2016; Bakker et al. 2009) to calculate the "Levenshtein Distance Normalized Divided" (LDND) between the first languages and French. The LDND is

based on Levenshtein distance matrices based on 40 words in the ASJP database. The distances are normalized by word length and average chance similarity. The lexical distances of the L1s in the original sample vary between 0 (for French, not used in the present study) and 103.12 (for Japanese,-see Figure 1 for details).

Relative morphological complexity. To describe the morphological complexity of the languages of the world, Lupyan and Dale (2010) initially suggested a list of 28 features from the World Atlas of Linguistic Structures (Haspelmath 2005). Based on this list, Schepens et al. (2013) calculated the morphological complexity of the first languages in their sample relative to Dutch and showed that this measure is a valid predictor of Dutch speaking skills. In a second study, van der Slik et al. (2017) reduced the list of morphological features to eight features that seem to be equally valid predictors of Dutch speaking skills.

We base our measure of relative morphological complexity on the reduced list of morphological features. There are obviously linguistic structural differences between Dutch and French as target languages, but six of the eight features are identical for French and Dutch. The differences between the two languages pertain to Features 22 (Inflectional Synthesis of the Verb) and 92 (Position of Polar Question Particles). French is more complex with respect to feature 22 (the WALS value is 3, as compared to 2 for Dutch). The other feature difference concerns polar question particles which are non-existent in Dutch but do exist (in first position) in French (according to van der Slik et al. 2019: 69−70, the Dutch feature is more complex than the French). For each first language in the student sample, we compared the eight features listed in Table 1 to the corresponding values for French. Whenever the feature of an L1 was of lower complexity than in French, we counted this as an increase in complexity (towards French). When the feature was equally or more complex in that L1 than in French, we counted it as equal. As suggested in van der Slik et al. (2017), we then calculated the relative complexity score by dividing the number of equally or more complex features by the total number of features available for a specific language so that a language that is morphologically at least equally complex as French received a score of 1. The features were not weighted for importance. At least four out of the eight features needed to be available for the language (and the corresponding students) to be included in further analyses.

Table 1. Features used for morphological distance calculation. Sample application to Vietnamese and French. Seven out of eight features are more complex in French if the hierarchy in van der Slik et al. (2019: 69) is applied. The relative complexity score (number of equally or more complex features in L1 divided by total number of features) is 0.13.

WALS Feature	French (FL)	Vietnamese (L1)	Complexity change
Inflectional Synthesis of the Verb (WALS 22)	3 (4–5 categories per word)	1 (0–1 category per word)	L1 less complex than FL
Prefixing vs. Suffixing in Inflectional Morphology (WALS 26)	2 (predominantly suffixing)	1 (little or no inflectional morphology)	L1 less complex than FL
Syncretism in Verbal Person/ Number Marking (WALS 29)	2 (Subject person/ number marking is syncretic)	1 (No subject person/ number marking)	L1 less complex than FL
Past Tense (WALS 66)	1 (Past/non-past distinction marked; no remoteness distinction)	4 (No grammatical marking of past/non-past distinction)	L1 less complex than FL
Position of Polar Question Particles (WALS 92)	1 (Question particle at beginning of sentence)	2 (Question particle at end of sentence)	L1 equally complex as FL
Alignment of Verbal Person Marking (WALS 100)	2 (accusative alignment)	1 (neutral alignment, no verbal person marking)	L1 less complex than FL
Verbal Person Marking (WALS 102)	2 (Person marking of only the A argument)	1 (No person marking of any argument)	L1 less complex than FL
Negative Morphemes (WALS 112)	2 (negative particle)	4 (Negative word, unclear if verb or particle)	L1 less complex than FL

As can be seen in Figure 1, a large majority of the first languages in the sample are either of equal complexity as French or of somewhat lower complexity (0.88). A count of the data points revealed that 96.7% of all participants

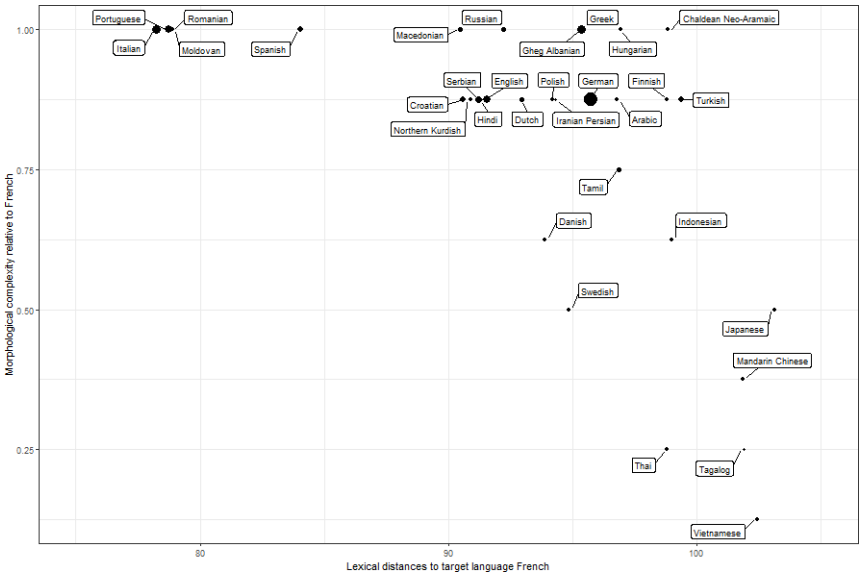

Figure 1. Lexical distances to French and morphological complexity relative to French of the languages in the sample. The size of the data points indicates the number of subjects speaking the language in question (see Tables A3 and A4 in in the online supplementary material for more details).

are distributed among these two levels of relative complexity. Assuming that all students in regular 8th-grade classes had extensive contact with German, the language of schooling, we decided to consider German as an L1-like language in their repertoire and to assign all students a complexity measure of at least 0.88, the value for German. Thus, the relative complexity variable is a binary variable coding morphological complexity equal to French as 1 and a relative complexity score of 0.88 as 0. For the regression models below, we selected those students from the full sample of the assessment study for whom we had complete distance and complexity measures, whose first language repertoire does not include French, and whose family languages involve at least one language other than German. (See Tables A1 and A2 in the supplementary material for more detail on the distance and relative complexity measures of the L1s considered in the sample.)

3.5. Regressors used

We added the two linguistic measures – the (standardized) lexical distance to French, labelled *Lex.Distance French*, relative morphological complexity (as a binary variable), labelled *Morph.Complexity French* – and a total of 14 control variables to the regression models. These latter variables pertain to curricular, motivational or socioeconomic aspects.

A first series of variables (dummy variables labeled **9/12/14 Lessons**) refers to the exposure to French in the classroom**.** As shown in many studies (e.g. Jaekel et al. 2017), exposure to the target language in the classroom is a robust predictor for proficiency and thus needs to be accounted for here. The pupils tested either had 9, 10 (here the reference category), 12 or 14 cumulative weekly lessons (CWL) of French before they were tested. For example, the 14 CWL group had 3 weekly lessons in 5th grade, 3 in 6th, 4 in 7th and 4 in 8th grade.

At the end of primary school, Swiss pupils are selected into secondary school tracks with lower or higher requirements. Children in the higher track are expected to perform better in general as well as in the target foreign language French. The binary variable **Higher Track** identifies pupils in the more/most demanding track present in the sample. Related to this variable, children who have skipped or repeated a grade are identified by the variables **Grade Skipped/Grade Repeated.**

As argued in Saito et al., different motivational dimensions explain variance in language learning (Saito et al. 2018)**.** We used two standardized scores based on 6 questionnaire items operationalizing pupils' interest and pleasure in learning French (**Intrinsic Motivation**) and another score based on 13 questionnaire items (**Extrinsic Motivation)** operationalizing pupils' motivation to learn French to receive good grades, or to use the language in domains such as occupation or the internet. Both intrinsic and extrinsic motivation have been shown to be relevant for foreign language achievement in past studies (see Saito et al. 2018 for a literature review). Moreover, we used a standardized score based on 7 questionnaire items covering various aspects of foreign language learning anxiety and feelings of overburdening (**FL Anxiety).**

A final set of control variables are social variables, many of them related to the well-documented influence of socioeconomic background on linguistic development and language skills, both productive and receptive (see Avineri and Johnson 2015 for an overview). We include a standardized estimate of the number of books at home, based on a 6-point scale (min. "0–10"; max. "more than 500", **Books Home**). Furthermore, we use a standardized score for the

highest educational level completed by either parent (from "no educational level completed" to "higher education degree", **Education Parents**). We also include a standardized variable accounting for possessions such as a room of one's own or the number of cars, computers or works of art in the household **(Home Possessions)**. Simplifying the PISA (Programme for International Student Assessment) four-category IMMIG index (OECD 2014: 307), we considered a pupil having a migrant background if both parents are born abroad, independently of his or her own country of birth (binary variable, **Parents Migrants**). Finally, we include a variable indicating that a student is male, since many studies have shown a female advantage in foreign and second language learning (binary variable, **Gender Male**; see van der Slik et al. 2015 for a discussion).

3.6. Modelling procedure

To test the effects of the background variables and the two linguistic measures on the skills in French listening, reading, writing and speaking, we used 10 imputed datasets (for details see section 1 in the supplementary material) concurrently to fit linear mixed models using the R packages "lme4" (Bates et al. 2015) and "mice" (van Buuren and Groothuis-Oudshoorn 2011). The clustering of the data (students nested in classes) was accounted for by random intercepts.[4] We fitted models that included one or both linguistic distance measures in addition to the curricular, motivational and socioeconomic control variables. In the following results section, we focus on the full model containing both distance measures.

4. Results

We first present a visual inspection of the relationship between the two distance measures and language test scores.[5] Subsequently, we report the results of the inferential statistical modeling, thereby focusing on the association between the linguistic distance measures and the language test results. The data for all four language skills were analysed in a strictly parallel fashion.

[4] Random slope models are not supported by the data, as there are a number of classes with only one student who meets the selection criteria.

[5] For this descriptive part, data from one of the ten imputed datasets is used.

4.1. Visual inspection of the data

Figures 2 and 3 visualize the association between the two linguistic distance measures and the results of the listening comprehension test (all skills show similar patterns, for full detail see Figures A2 and A3 in the supplementary material).

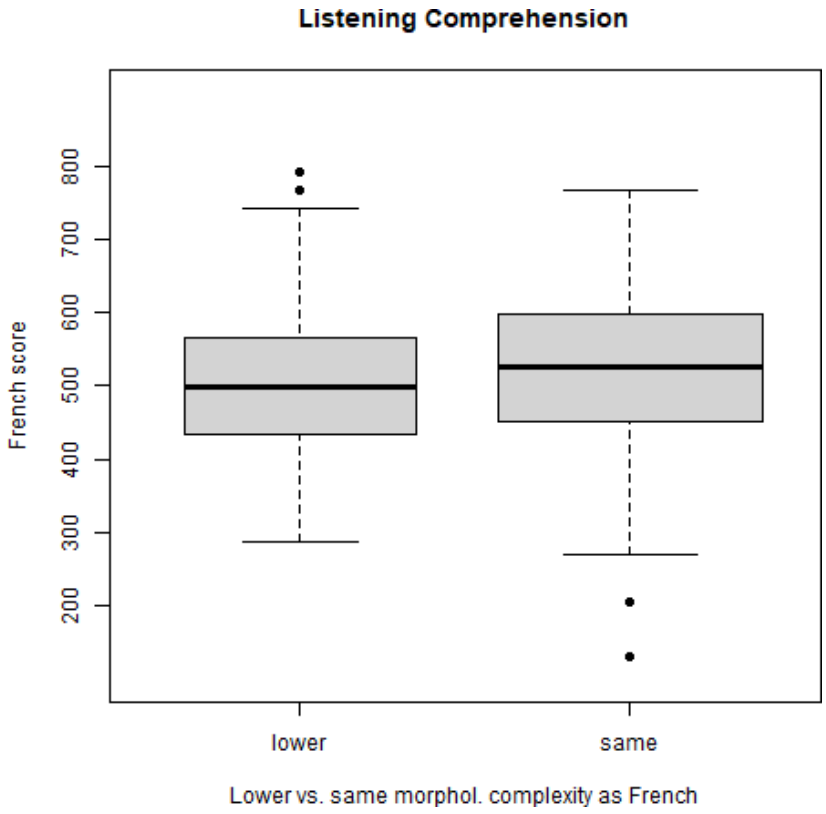

Figure 2. Relative morphological complexity and French listening comprehension scores.

Figure 2 shows the association between listening comprehension and the level of relative morphological complexity. Independent of the L1 background, all

distributions of language test results span a broad scale range. However, the observed median test results are slightly lower for students whose repertoire of L1s comprises only languages of lower morphological complexity than French. This observation is true for all four skills tested, see Figure A2 in the supplementary material. This finding suggests that lower complexity in L1 is associated, as expected, with lower French proficiency. This association, however, seems to be relatively weak given the large overlap of the data ranges for the two groups.

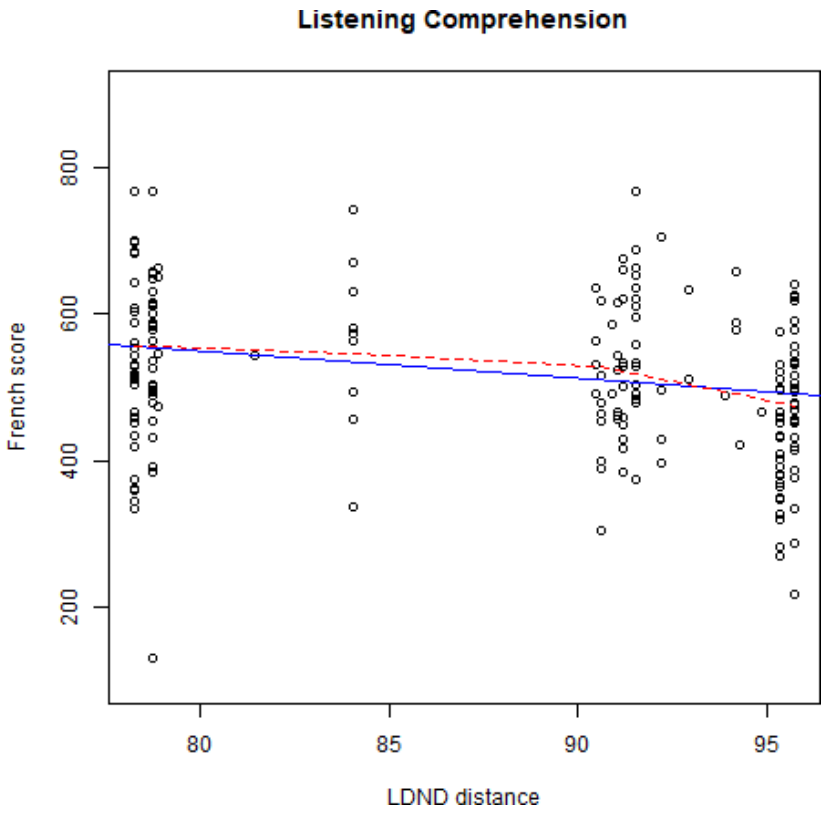

Figure 3. Lexical distance and French listening comprehension scores

Figure 3 shows the unmodelled association of lexical distance and French listening comprehension. Again, as predicted, the test results on the listening comprehension test are lower as the lexical distances between French and the learners' L1s increase. This applies to all four skills, as Figure A3 in the supplementary material illustrates. In view of the relatively flat slope of the regression line, this effect does not appear to be dramatic in size. The non-linear (lowess) smoother line for listening comprehension shows a similar tendency with a dip at a lexical distance value just above 90. The languages in question (see Figure 1) are languages such as Serbian, Croatian or Hindi. In the group of speakers of these first languages, the scores seem to be particularly low. There is a wide range of scores at a distance of around 95. At this distance to French, there are languages that have a very different role and relationship with respect to the main language of instruction (e.g. German, Swedish, Albanian).

4.2. Inferential statistics

To test whether lexical distance and relative morphological complexity are statistically significant predictors of performance on the French tests, we fitted linear mixed models that include a series of control variables in addition to these linguistic predictors.

If both linguistic predictors (complexity and distance) are concurrently included in the models as fixed effects, only lexical distance is a significant predictor. If relative morphological complexity is considered without the lexical distance measure, it is a significant predictor in the case of two language skills, listening and writing. In the online supplementary material, tables A12 and A13 provide the fixed-effect estimates and the random effect parts of the model containing morphological complexity only, tables A9 and A10 provide the results of the full model containing both linguistic measures. An effect size measure for the predictors can be derived directly from the fixed effect coefficients (Baayen et al. 2008). As morphological complexity is a binary variable, the fact that a learner's most complex L1 is morphologically equally complex as French instead of less complex, improves his/her listening score by (only) 5/100 standard deviations (SD; 100 being the SD of the listening scale) according to the full model with both linguistic measures. This effect is not significant. Also, if the lexical distance between French and a learner's closest L1 is one SD larger, the result on the listening test decreases by 20/100 SD on average. Alternative methods of quantifying effect size are based on R^2, the

proportion of the total variance that is explained by a model. Cohen's f^2 effect size statistic (Cohen 1988) provides a standardized measure for R^2 differences induced by adding predictors to multiple regression models. As actual R^2 estimates are not available for mixed-effect models, we used the marginal pseudo-R^2 as defined in Nakagawa and Schielzeth (2013) and implemented in the R package *r2glmm* (Jaeger 2017) instead. With regard to listening comprehension, the inclusion of relative morphological complexity alone (in addition to the control variables) shows an effect of $f^2 = 0.015$, 95% CI [0.014, 0.016]. The corresponding f^2 for lexical distance alone is 0.035, 95% CI [0.032, 0.038]. For the combination of both predictors f^2 amounts to 0.0363, 95% CI [0.033, 0.039]. According to Cohen's standard interpretation, f^2 values from 0.02 to 0.15 stand for a small effect. Therefore, while we observe a small effect for lexical distance, morphological complexity seems negligible. Effect sizes are very similar in the case of all four language tests (please refer to Table A10 for full detail for all four skills).

The coefficients of the fixed effect predictors of the full model for listening comprehension are plotted in Figure 4. Figures A5 to A8 in the online supplementary material show these same estimates for all four skills.

At a purely descriptive level, the two distance measures on which we focus in our analyses show similar patterns across all four skills, i.e. effects in the expected direction can be observed with regard to all four language test scores. The lexical distance of L1s to French is also a statistically significant predictor of performance on the French tests. The larger the lexical distance of an L1 to French, the lower the test scores. Unlike lexical distance, relative morphological complexity as measured by the comparison of eight features is not significantly associated with French language proficiency if both linguistic predictors (complexity and distance) are added to the model. The two predictors are negatively correlated ($r = -0.54$; t-value: -12.87; $p < 0.001$).

The French scores on all four skills are (negatively) associated with the level of foreign language anxiety as measured by our questionnaire items. A higher track level coincides with better performance in French in three out of the four language skills tested. A factor affecting both receptive skills is the students' estimate of the number of books at home – a proxy for the relative importance of family literacy practices. As expected, the higher the number of books indicated the better the listening and reading skills. We do not know why the productive skills are not affected by this predictor to a similar degree.

No significant association with any of the four skills is found for socioeconomic constructs such as parents' education, and possessions at home. Parents being migrants is not associated with the test results in three out of the

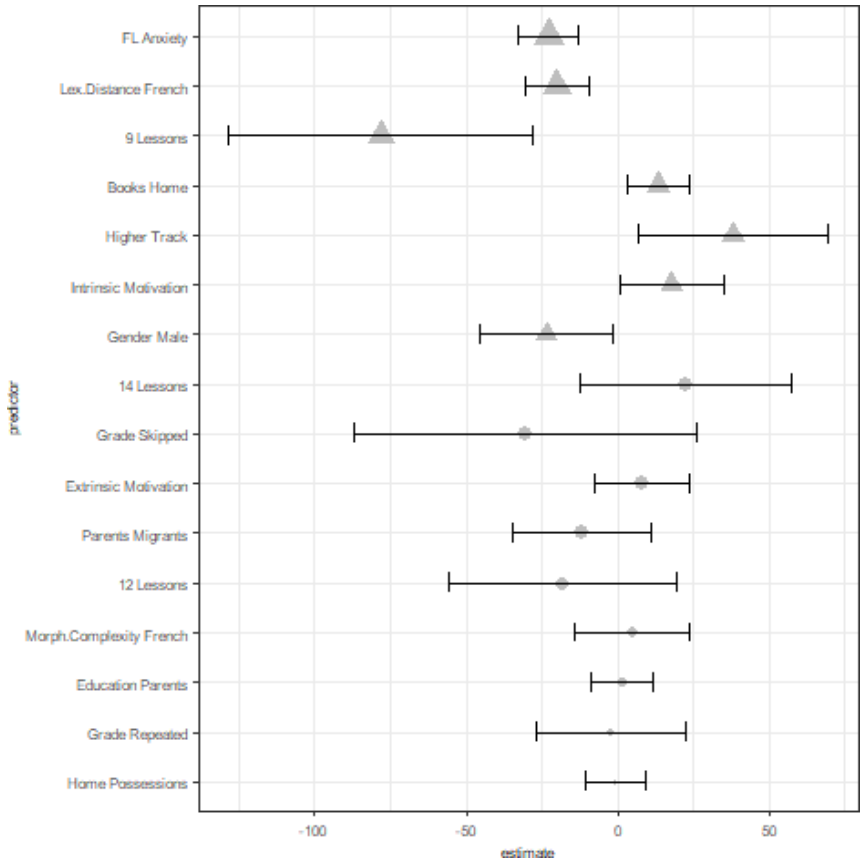

Figure 4. Estimates and 95% confidence intervals for the predictors of French listening comprehension (full model). Triangles represent estimates that are significant at the $p < 0.05$ level. The size of the symbols represents the absolute value of the t-statistic (see Table A5 for full detail).

four skills (there is a negative association with the speaking skill). Furthermore, the fact that a student previously skipped or repeated a school year does not show any noteworthy association with any of the four skills.

The curriculum-related variables overall yield the expected patterns of association with the skills measures: The speaking and writing scores are positively associated with the longest total exposure time (14 CWL) while the listening scores are negatively associated with the shortest total exposure time (9

CWL). Reading is not significantly associated with the number of weekly lessons.

Extrinsic motivation to learn French is positively associated with reading and writing, while intrinsic motivation is positively associated with all skills except reading comprehension. Our expectations were that they be associated with all four skills in a similar fashion and we have no specific explanation why the effect on reading is too small to reach statistical significance.

Gender, to conclude, affects, three out of the four skills, namely listening, reading, and writing. Male students perform worse than female students on these skills. We have no founded explanation why success on the speaking test does not seem to be associated with gender.

The results regarding the hypotheses H1a and H1b are as follows: A similar pattern emerges for the two variables operationalizing linguistic contrasts between languages. Regarding all four skills, lexical distance between L1 (or German) and the target language French shows a significant negative association with foreign language skills. Morphological complexity is associated as expected with the outcome variables; however, the predictive power of this second variable is not statistically significant according to our models.

None of the variables is associated with French skills in a counter-intuitive way. Some patterns are relatively unexpected, such as the occasional absence of associations across the board (e.g. parents' education) or the absence of an effect of a variable in relation to only one skill, e.g. the lack of a gender effect on speaking.

5. Discussion

Linguistic distance between first languages (or German) and the target language French is negatively associated with foreign language skills when controlling for a considerable list of factors known to be associated with (language) learning in school. This finding, based on a dataset gathered in the context of a large-scale assessment study of foreign-language proficiency, ties in with large-scale data from rather different settings of language learning, as the one discussed in Schepens (2014). Our study thus confirms and extends previous findings in several respects: The impact of linguistic distance seems to be measurable not only in immersive second language learning settings, but also in foreign language instruction settings with very limited exposure to the target language. Moreover, whereas the second language learning studies

(where the target language is the strong environmental language of the immigration context) investigating the impact of linguistic distance have focused on the speaking skill exclusively, our study covers the four skills of listening, reading, speaking and writing. The analyses show that all four skills are – descriptively – affected in the same direction by the lexical and morphological measures and that lexical distance between L1s (or the language of schooling) and French is consistently positively associated with all skills. In our study, skills were focused rather than constructional, grammatical competence as in studies from a formal or learner-variety perspective discussed in the state of the art section above. Our results are nevertheless in line with some of these previously reported results (e.g. in Puig-Mayenco et al. 2020; and Westergaard et al. 2017). Our study, however, does not confirm differential effects of comprehension versus production. Moreover, no fundamental difference between the two oral and the two written skills appears in our data.

The two measures applied are not measures of perceived distance (psychotypology, Kellerman 1986) but objective measures relating to the language systems. Our distance measures have clear limitations, most obviously so with regard to languages that are genealogically unrelated to the target language: The lexical distance from French to Italian is smaller than from French to Spanish, which is due to more and less overlap in the list of words used to calculate this distance. In contrast, the difference between the lexical distances of, say, Hungarian or Bahasa Indonesia to French is most certainly not meaningful given the absence of a shared inherited (cognate) vocabulary which this distance measure explores. What really matters is the greater distance between French and these languages compared to the distance between French and, in particular, Romance languages. This weakness of lexical measures is what motivates the inclusion of other linguistic measures, as e.g. the morphological complexity measure used in our study. The dip in the smoother line in Figure 3 shows that the association of the lexical distance measure with foreign language skills is not strictly linear with regard to listening comprehension. Although languages such as Dutch or Swedish are relatively different from French, the scores obtained by the group of speakers of these first languages do not follow a strictly linear pattern. This might be an effect of the fact that their first languages are closer to the language of schooling (German), and it is possible that rapid assimilation to the language of schooling due to this similarity also has a positive impact on the learning of French as a foreign language in a German-speaking context. Additionally, in Switzerland, the sociology of families with a Dutch or Swedish background, on the one hand, and families with for instance a Serbian, Croatian or Hindi background, on the

other, is expected to be quite different, which in turn is known to have an impact on educational success (e.g. Becker et al. 2013). Therefore, it is important to assess the impact of linguistic distance by taking into account other, socioeconomic variables, as we did. The regression models (especially the effect size measures) show that the effect of linguistic distance as measured in the lexical distance variable remains small but robust, while the relative morphological complexity is not relevant if the effect of lexical distance is included in the model.

The importance of linguistic distance as a predictor of additional language learning ties in not only with the linguistic but also with the general literature on transfer discussed in the literature review above. The effect of the degree of overlap of source and target knowledge also predicts learning in the foreign language domain.

The measure of relative morphological complexity applied was based on previous work on immigrant languages and Dutch as a target language. The choice of drawing on this previous work might not be an optimal fit for the languages under investigation here. A different measure drawing on other linguistic features could and maybe should be developed, however this would require extensive pre-testing and gathering of large amounts of typological features that are difficult to obtain for certain languages in our sample.

The other variables we included in our models showing linear associations with French skills represent effects scholars would expect. Intrinsic and extrinsic motivational constructs are positively associated with learning French overall while anxiety is negatively associated. Varying length of exposure (due to different curricula) is also associated in the expected direction, indicating that, quite unsurprisingly, more exposure leads to better skills. A positive association with the number of books variable as a proxy for family literacy practices is also to be expected. A noteworthy finding is that the other socioeconomic variables that are generally expected to be related to school performance do not play out in the expected way if we take into account all the predictors we added to our models.

French is minimally the fourth language of the participants in our study. This count depends on how Alemannic Swiss German, which is the native language of virtually all German-speaking Swiss and also of many Swiss with one migrant parent, is treated: If it is not counted as a separate language, then the pupils in our sample typically start with German as one of their first or as a second language, English as a first foreign language followed by French as second foreign language. For children with two migrant parents in our sample, German (and Swiss German) would typically be the second language, English

third, and French fourth. It is thus difficult to draw conclusions on the impact of a supposed positive bilingualism effect on third language acquisition. What the results do show, and this might be relevant for curriculum planning, is that there is no clear evidence for a (negative or positive) impact of the status of being a child with a migrant background on the learning of French as an additional language. Although the estimates of the variable "Parents Migrants" are negative, they do not reach significance in three out of four skills (see Table A10 in the supplementary material for details). This finding seems important from the point of view of educational policy: Opponents of the teaching of two foreign languages at primary school level often argue that migrant children whose first language is not the local language of schooling are more likely to be overburdened by two foreign languages (in addition to their second language, i.e. German). The multilingual advantage theory for its part would predict multilingual (migrant) children to outperform monolinguals simply by being multilingual. Since we exclusively analyse multilingual children's scores, we cannot put the latter theory to the test. After controlling for the effects of linguistic distance and the other variables, our results show no negative effect of the migration status. The data thus do not lend strong support to any one of the two opposing positions outlined above. Moreover, we should keep in mind that in all schools a certain percentage of children are exempt from the French lessons or have opted out of the subject. Given this risk of a "study in winners" due to this potential bias in our data, we refrain from both, i.e. the positive ("multilingualism advantage") and the negative ("overburdening") conclusions on the effects of individual bi- or multilingual repertoires – notwithstanding the languages actually concerned – on additional language learning. What emerges from our data, however, is that the degree of overlap between the multilinguals' linguistic repertoires indeed predicts foreign language skills.

6. Conclusion

Our contribution adds to the research on linguistic contrasts between mastered languages and target languages that are being acquired. Our study is different from previous work (e.g. Schepens et al. 2016) in that we analyse data from classroom-based foreign language learning where exposure is necessarily rather limited. Our analyses show that lexical distance is consistently associated with proficiency in the foreign language French. This result ties in with long-standing claims on transfer (Thorndike and Woodworth 1901) that predict transfer effects to be positively associated with similarity or overlap between

existing knowledge and the object of learning. As discussed in our contribution, further development of linguistic distance measures seems a promising avenue for future research. Such measures that could take into account morphological and syntactic but also semantic similarities and contrasts might contribute to a better understanding of the role of previous linguistic knowledge in the learning of additional languages.

7. Acknowledgements

We wish to thank all our colleagues who contributed to this project, especially Mirjam ANDEXLINGER, Karolina KOFLER, and Christopher HUGUENIN

REFERENCES

Adams, R. J. 2005. "Reliability as a measurement design effect". *Studies in Educational Evaluation* 31(2–3). 162–172.

ALTE. 1998. *Multilingual glossary of language testing terms*. Cambridge: Cambridge University Press.

Avineri, N. and E. J. Johnson. 2015. "Invited forum: Bridging the 'Language Gap'". *Journal of Linguistic Anthropology* 25(1). 66–86.

Baayen, R. H., D. J. Davidson and D. M. Bates. 2008. "Mixed-effects modeling with crossed random effects for subjects and items". *Journal of Memory and Language* 59(4). 390–412.

Bakker, D., A. Müller, V. Velupillai, S. Wichmann, C. H. Brown, P. Brown, … E. W. Holman. 2009. "Adding typology to lexicostatistics: A combined approach to language classification". *Linguistic Typology* 13(1). 169–181.

Bartelheimer, P. and M. I. Wolter. 2016. "Demografischer Wandel – ändert Zuwanderung den Trend? Die neue Migration nach Deutschland – sozioökonomisch betrachtet". *SOFI-Mitteilungen* 10. 7–12.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. "Fitting linear mixed-effects models using lme4". *Journal of Statistical Software* 67(1). 1–48.

Becker, R., F. Jäpel and M. Beck. 2013. "Diskriminierung durch Lehrpersonen oder herkunftsbedingte Nachteile von Migranten im Deutschschweizer Schulsystem"? *Swiss Journal of Sociology* 39(3). 517–549.

Berthele, R. and J. Vanhove. 2020. "What would disprove interdependence? Lessons learned from a study on biliteracy in Portuguese heritage language speakers in Switzerland". *International Journal of Bilingual Education and Bilingualism* 23(5). 550–566.

BIFIE, Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens. 2013. *Schülerfragebogen Standardüberprüfung 8. Schulstufe 2013.* Retrieved from https://www.bifie.at/system/files/dl/BIST-UE_E8_2013_Schuelerfragebogen_2013-01-24_0.pdf.

Bransford, J. D., A. L. Brown and R. R. Cocking. 2000. *How people learn: Brain, mind, experience, and school.* (Expanded, subsequent edition). Washington, D.C: National Academies Press.

Cenoz, J. and F. Valencia. 1994. "Additive trilingualism: Evidence from the Basque country". *Applied Psycholinguistics* 15(2). 195–207.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences* (2nd edn.). Hillsdale, N.J: L. Erlbaum Associates.

Council of Europe (ed.). 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press. Retrieved from https://rm.coe.int/1680459f97.

Cummins, J. 1979. "Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters". *Working Papers on Bilingualism* 19. 197–205.

Découvertes. 2004. Stuttgart: Ernst Klett Verlag.

Dimroth, C. 2018. "Beyond statistical learning: Communication principles and language internal factors shape grammar in child and adult beginners learning polish through controlled exposure". *Language Learning* 68(4). 863–905.

Dressler, C. and M. L. Kamil. 2006. "First- and Second-Language Literacy". In: D. August and T. Shanahan (eds.), *Developing literacy in second language learners: Report of the national literacy panel on language-minority children and youth*. Washington D.C.: Lawrence Erlbaum Associates Publishers. 197–238.

Elsner, D. 2007. *Hörverstehen im Englischunterricht der Grundschule: Ein Leistungsvergleich zwischen Kindern mit Deutsch als Muttersprache und Deutsch als Zweitsprache.* Frankfurt am Main u.a.: Peter Lang.

Engel, G., B. Groot-Wilken and E. Thürmann. 2009. *Englisch in der Primarstufe – Chancen und Herausforderungen: Evaluation und Erfahrungen aus der Praxis.* Berlin: Cornelsen.

Envol. 2000. (6th edition). Zürich: Lehrmittelverlag Zürich.

Gamer, M., J. Lemon, I. Fellows and P. Singh. 2012. "irr: Various coefficients of interrater reliability and agreement". Retrieved from https://CRAN.R-project.org/package=irr.

Gooskens, C., V. J. van Heuven, J. Golubović, A. Schüppert, F. Swarte and S. Voigt. 2018. "Mutual intelligibility between closely related languages in Europe". *International Journal of Multilingualism* 15(2). 169–193.

Haspelmath, M. 2005. *The world atlas of language structures.* Oxford: Oxford University Press.

Heinzmann, S. 2013. *Young language learners' motivation and attitudes: Longitudinal, comparative and explanatory perspectives.* London/New York: Bloomsbury Academic.

Jaeger, B. 2017. *r2glmm: Computes R Squared for mixed (multilevel) models*. Retrieved from https://CRAN.R-project.org/package=r2glmm.

Jaekel, N., M. Schurig, M. Florian and M. Ritter. 2017. "From Early Starters to Late Finishers? A Longitudinal Study of Early Foreign Language Learning in School". *Language Learning* 67(3). 631–664.

Jessner, U. 1999. "Metalinguistic awareness in multilinguals: Cognitive aspects of third language learning". *Language Awareness* 8(3-4). 201–209.

Kellerman, E. 1986. "An eye for an eye: Crosslinguistic constraints on the development of the L2 lexicon". In: E. Kellerman and M. Sharwood Smith (eds.), *Cross-linguistic influence in second language acquisition*. New York et al.: Pergamon Institute of English. 35–48.

Kiefer, T., A. Robitzsch and M. Wu. 2015. *TAM: Test Analysis Modules*. http://CRAN.R-project.org/package=TAM.

Klieme, E. and DESI-Konsortium (eds.). 2008. *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie.* Weinheim et al.: Beltz. Retrieved from https://www.pedocs.de/volltexte/2010/3149/pdf/978_3_407_25491_7_1A_D_A.pdf.

Kloss, H. 1929. *Nebensprachen. Eine sprachpolitische Studie über die Beziehungen eng verwandter Sprachgemeinschaften.* Wien, Leipzig: Braumüller.

Krippendorff, K. 2011. *Computing Krippendorff's alpha-reliability.* University of Pennsylvania. Retrieved from http://repository.upenn.edu/asc_papers/43.

Kristol, A. M. 1990. "L'enseignement du français en Angleterre (XIIe-XVe siècles). Les sources manuscrites". *Romania* 111(443–444). 289–330.

Kürschner, S., C. Gooskens and R. V. Bezooijen. 2008. "Linguistic determinants of the intelligibility of Swedish words among Danes". *International Journal of Humanities and Arts Computing* 2(1–2). 83–100.

Lado, R. 1957. *Linguistics across cultures: Applied linguistics for language teachers.* Ann Arbor: University of Michigan.

Lenz, P. and T. Studer. 2007. *Lingualevel. Instrumente zur Evaluation von Fremdsprachenkompetenzen.* Bern: Schulverlag.

Linacre, J. 1994. *Many-facet Rasch measurement* (2nd edn.). Chicago: MESA Press.

Lupyan, G. and R. Dale. 2010. "Language structure is partly determined by social structure". *PLoS ONE* 5(1). e8559.

Lorenz, E., T. E. Toprak, and P. Siemund (2021) "English L3 acquisition in heritage contexts: Modelling a path through the bilingualism controversy". *Poznań Studies in Contemporary Linguistics* 57(2). 273–298.

Masters, G. N. 1982. "A Rasch model for partial credit scoring". *Psychometrika* 47. 149–174.

Nakagawa, S. and H. Schielzeth. 2013. "A general and simple method for obtaining R^2 from generalized linear mixed-effects models". *Methods in Ecology and Evolution* 4(2). 133–142.

Nunan, D. 1989. *Designing tasks for the communicative classroom*. Cambridge: University Press.

OECD (ed.). 2012. *Programme for International Student Assessment 2012: Internationaler Schülerfragebogen PISA 2012.* Retrieved from https://www.bifie.at/system/files/dl/pisa12_internationaler_nationaler_schuelerfragebogen.pdf.

OECD (ed.). 2014. *PISA 2012 technical report.* Retrieved from http://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm

Otwinowska, A. and J. M. Szewczyk. 2019. "The more similar the better? Factors in learning cognates, false cognates and non-cognate words". *International Journal of Bilingual Education and Bilingualism* 22(8). 974–991.

Peyer, E., Andexlinger, M., Kofler, K., & Lenz, P. (2016). *Projekt Fremdsprachenevaluation BKZ. Schlussbericht zu den Sprachkompetenztests*. Kompetenzzentrum Mehrsprachigkeit. https://folia.unifr.ch/unifr/documents/304950

Proctor, C. P., D. August, C. Snow and C. D. Barr. 2010. "The interdependence continuum: A perspective on the nature of Spanish–English bilingual reading comprehension". *Bilingual Research Journal* 33(1). 5–20.

Puig-Mayenco, E., J. González Alonso and J. Rothman. 2020. "A systematic review of transfer studies in third language acquisition". *Second Language Research* 36(1). 31–64.

Ringbom, H. 1978. "On learning related and unrelated languages". *Moderna Språk* 72. 21–25.

Rubin, D. B. 1987. *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Sanders, M. and G. Meijers 1995. "English as L3 in the elementary school". *ITL Review of Applied Linguistics* 107. 59–78.

Safont Jordà, M. P. 2005. *Third language learners: Pragmatic production and awareness.* Clevedon; Buffalo: Multilingual Matters.

Saito, K., J.-M. Dewaele, M. Abe and Y. In'nami. 2018. "Motivation, emotion, learning experience, and second language comprehensibility development in classroom settings: A cross-sectional and longitudinal study". *Language Learning* 68(3). 709–743.

Schepens, J. 2014. *Bridging linguistic gaps: The effects of linguistic distance on the adult learnability of Dutch as an additional language.* Utrecht: Utrecht University.

Schepens, J., F. Van der Slik and R. Van Hout. 2013. "Learning complex features: A morphological account of L2 learnability". *Language Dynamics and Change* 3(2). 218–244.

Schepens, J. J., F. Van der Slik and R. Van Hout. 2016. "L1 and L2 distance effects in learning L3 Dutch". *Language Learning* 66(1). 224–256.

Thorndike, E. L. and R. S. Woodworth. 1901. "The influence of improvement in one mental function upon the efficiency of other functions: III. Functions involving attention, observation and discrimination". *Psychological Review* 8(6). 553–564.

Van Buuren, S. and K. Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R". *Journal of Statistical Software* 45(3). 1–67.

Van der Slik, F. W. P., R. W. N. M. van Hout, and J. J. Schepens. 2015. "The gender gap in second language acquisition: Gender differences in the acquisition of Dutch among immigrants from 88 countries with 49 mother tongues". *PLoS ONE* 10(11).

Van der Slik, F., R. van Hout and J. Schepens. 2019. "The role of morphological complexity in predicting the learnability of an additional language: The case of La (additional language) Dutch". *Second Language Research* 35(1). 47−70.

Vanhove, J. and R. Berthele. 2015. "Item-related determinants of cognate guessing in multilinguals". In: G. de Angelis, U. Jessner and M. Kresić (eds.), *Crosslinguistic Influence and Crosslinguistic Interaction in Multilingual Language Learning*. New York: Bloomsbury Academic. 95–118.

Wagner, W., A. Helmke and E. Rösner. 2009. *Deutsch-Englisch-Schülerleistungen-International (DESI). Dokumentation der Erhebungsinstrumente für Schülerinnen und Schüler, Eltern und Lehrkräfte.* Frankfurt am Main: Gesellschaft zur Förderung Pädagogischer Forschung / Deutsches Institut für Internationale Pädagogische Forschung. Retrieved from https://www.pedocs.de/volltexte/2010/3252/pdf/MatBild_Bd25_1_D_A.pdf

Warm, T. 1989. “Weighted likelihood estimation of ability in item response theory”. *Psychometrika* 54. 427–450.
Westergaard, M., N. Mitrofanova, R. Mykhaylyk and Y. Rodina. 2017. “Crosslinguistic influence in the acquisition of a third language: The Linguistic Proximity Model”. *International Journal of Bilingualism* 21(6). 666–682.
Wichmann, S., C. H. Brown and E. W. Holman (eds.). 2016. *The ASJP Database*. Jena: Max Planck Institute for the Science of Human History. Retrieved from http://asjp.clld.org/

Address correspondence to:
Raphael Berthelé

Institute of Multilingualism Université de
Fribourg Rue de Rome 1
CH-1700 Fribourg/Freiburg
raphael.berthele@unifr.ch

Predicting foreign language ability in large-scale assessment

Supplementary material

Raphael BERTHELE/Peter LENZ/Elisabeth PEYER

24.9.2022

1 Data preparation

The data we use for our study was collected for a large-scale evaluation study whose purpose it was to verify the attainment of the curricular objectives for the foreign-languages in Central Switzerland. Most of the data preparation was carried out on the occasion of that large-scale assessment. Data cleaning, the scaling of item-based data and finally t h e i m putation o f m issing d ata a nd d ata w ith k nown m easurement e rror were crucial steps in this process.

Overall, the dataset for French consisted of 3687 cases, i.e. pupils in either 6th or 8th grade. As each student completed only a subsample of all tasks forming the test, a large proportion of the test data was missing "by design". All available test data per language skill was linked through common items. In such designs, working with test scores (counts of the correct answers) makes little sense. Instead, item response theory (IRT) models are used to bring all items (or criteria) and test-takers together on a single test scale. IRT is also used to identify problematic items, raters or test-takers.

The listening and reading data was scaled using the Rasch Partial Credit Model (Masters, 1982) as implemented in the R package "TAM" (Kiefer, Robitzsch & Wu, 2015). The PCM can deal with polytomous items (items with scores beyond 0 or 1) if needed. The same method was applied to item-based questionnaire data relating to motivational constructs. The writing and speaking data differ from objectively scorable items in that they do not represent direct learner responses but ratings of learner performances (i.e. spoken and written texts). Roughly one third of the performances were double-rated. Thanks to this rater overlap, it was possible to account for rater severity by using the Many-Facet Rasch Model (Linacre, 1994). Based on the double ratings, this model corrects each person (pupil) measure by the severity of the raters involved. A rater's severity is the difference between that rater's average severity, i.e. the tendency to rate performances more or less strictly than the others who rate the same performances, and all raters' average severity.

Finally, for each of the Rasch scales, person measures were produced in the shape of Warm's Weighted Likelihood Estimates (WLEs, Warm, 1989). Although these measures are not scores (counts of correct answers) in a strict sense, we use the word "score" to relate to them in our article, as they are a proxy to actual scores in a more complicated setting. The reliabilities of the WLE scales are given in Table A1.

Table A1: Reliabilities of WLE-based Scales

Construct	Scale reliability
Listening	0.77
Reading	0.72
Speaking	0.80
Writing	0.84
Intrinsic Motivation	0.86
Extrinsic Motivation	0.90
Language Learning Anxiety	0.79

Test scale reliabilities are not very high[1] because, as usual in large-scale assessments, which focus on the population (or the educational system) rather than the individuals, the tests used were rather short. Data imputation (see below) helps to avoid potentially distorting effects of less than perfect reliabilities, i.e. the measurement errors afflicting each person measure, on the results of subsequent analyses.

The WLE-based scales were combined with all variables potentially relevant for data analysis, including the linguistic distance measures. The resulting (original) dataset consisting of all students who provided usable test data, had a considerable portion of missing data. The main cause for this was a return rate of 78.9% for the student questionnaire. In order to be able to work with complete datasets in the analyses, data imputation was performed. The approach to data imputation that was chosen dealt with missing data and measurement error at the same time (cf. Blackwell, Honaker & King, 2015). The imputation was performed using the R packages "mice" (multivariate imputation by chained equations, van Buuren & Groothuis-Oudshoorn, 2011) and "miceadds" (Robitzsch, Grund & Henke, 2019). Put simply, chained-equation functions take one variable containing missing values at a time as the dependent variable and set up multivariate regression equations by identifying correlated variables in the dataset and using these as well as interactions between some of these variables as covariates. Based on the ad-hoc regression model, plausible data values are drawn from the prediction interval to fill in missing values in the momentarily dependent variable. In the case of WLE estimates, such regression equations are used to narrow down the scale ranges for the person estimates initially defined by a WLE point estimate and a measurement error interval. Based on the regression equation, random values are drawn from the narrowed WLE error intervals to define an "error-free" and complete variable. For the purpose of the original evaluation study, 100 imputed datasets were produced in this manner. In order to have independent draws for the missing data, only every tenth set was used for data analysis. Subsequently, all statistical analyses based on the slightly differing sets (due to the random draws) had to be carried out ten times independently. The results were combined using the appropriate combination rules (Rubin, 1987). Cases (pupils) that had merely imputed (i.e. originally completely missing) data in the criterion variable (e.g. the speaking test score in the analysis of the speaking test) were excluded from an analysis. Therefore, the number of cases that could be used, varied depending on the language skill in question.
In the present study which deals with the influence of typological differences between a student's L1 and the learning of French, the proportion of imputed, previously missing data in the independent variables is very small because all students who did not return their questionnaire had to be excluded in advance due to the lack of information on the personal language repertoire. In the subset of the finally selected 8th graders (N=409), only two variables included previously missing imputed data – "Parents Migrants" (1.2%) and "Education Parents" (7.8%).

References

Adams, R. J. (2005). Reliability as a measurement design effect. Studies in Educational Evaluation, 31(2–3), 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008
Blackwell, M., Honaker, J. & King, G. (2015). A unified approach to measurement error and missing data: overview and applications. Sociological Methods & Research. https://doi.org/10.1177/0049124115585360
Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). irr: Various coefficients of interrater reliability and agreement. Retrieved from https://CRAN.R-project.org/package=irr
Kiefer, T., Robitzsch, A., & Wu, M. (2015). TAM: Test Analysis Modules. http://CRAN.R-project.org/package=TAM
Linacre, J. (1994). Many-facet Rasch measurement (2nd ed.). Chicago: MESA Press. Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149–174. https://doi.org/10.1007/BF02296272
Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests, expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.
Robitzsch, A., Grund, S. & Henke, T. (2019). miceadds: Some additional multiple imputation functions, especially for mice. Retrieved from https://CRAN.R-project.org/package=miceadds
Wang, S., & Wang, T. (2001). Precision of Warm's Weighted Likelihood Estimates for a polytomous

[1]Reliability measures from Classical Test Theory (CTT), such as Cronbach's coefficient alpha, make limited sense in the Item Response Theory (IRT) framework. WLE (and competing) reliability indices replace traditional indices in the IRT context. The magnitude of a WLE reliability coefficient has a similar practical interpretation as e.g. alpha. Its substantive interpretation, however, is different: instead of the internal consistency of a test, it gauges the size of the overall measurement error for the persons on the test: higher WLE reliability indicates more measurement precision (Adams, 2005; Wang & Wang, 2001:320).

model in computerized adaptive testing. Applied Psychological Measurement, 25(4), 317–331. https://doi.org/10.1177/01466210122032163
Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(3), 427-450. https://doi.org/10.1007/BF02294627

2 Examples from Tasks and Scales

Table A2: Scoring scheme for the speaking test. Each answer/question is rated individually.

Number of Points	Description: lexical range/correctness
3 Points	Both vocabulary and structure (especially syntax, verbal morphology) of the utterance are in correct French. Mistakes of grammatical gender in the article or possessive pronoun (la pull, ma père) as well as in the adjective (des chaussures blancs) are not considered incorrect.
2 Points	The student is able to express him/herself in French in an understandable way. This means that the content words are in French, the structure may be incorrect (e.g. verb wrongly/not conjugated).
1 Point	The student's vocabulary is limited. This means that some but not all content words/phrases are in French.
0 Points	No answer; no content words in French (or only a cognate like "banana")

Extrinsic-instrumental motivation scale

- I will be able to use French even after my school days. (Adapted from Bifie 2013)
- I know exactly why I learn French. (Self-developed)
- It is important for me to achieve good grades in French. (Adapted from DESI 2009)
- With French you can communicate in many places. (Adapted from Bifie 2013)
- It is important to know French in order to get a good job later. (Adapted from DESI 2009)
- I am determined to make full use of my French lessons. (Adapted from DESI 2009)
- I learn French because I will need it later in my professional life. (Self-developed)
- I learn French because many people in the world speak French. (Adapted from Heinzmann 2013)
- I learn French to be able to communicate during holidays in a French speaking country. (Self-developed)
- I learn French to read French books and magazines. (Adapted from Heinzmann 2013)
- I learn French so that I can understand information on the Internet. (Adapted from Heinzmann 2013)
- I learn French to be able to talk to French-speaking friends or relatives. (Adapted from Heinzmann 2013)
- I learn French to be able to understand French-speaking music and films. (Adapted from Heinzmann 2013)

Intrinsic motivation scale

- I like French. (Adapted from Bifie 2013)
- I think French is an interesting subject. (Adapted from Bifie 2013)
- I like learning French. (Adapted from Bifie 2013)
- French is a beautiful language. (Adapted from Bifie 2013)
- I learn French because I like the language. (Self-developed)
- I learn French because I enjoy speaking it. (Adapted from Heinzmann 2013)

Language learning anxiety scale

- I am afraid of getting bad marks in French. (Adapted from Heinzmann 2013)
- I am afraid of written exams in French. (Adapted from DESI 2009)
- I am often stressed in French class because everything is so difficult. (Adapted from Heinzmann 2013)
- I am always glad if I do not have to say anything during French class. (Adapted from Heinzmann 2013)
- I am afraid of making mistakes when speaking in French class. (Adapted from Heinzmann 2013)
- It is too much for me to learn two foreign languages at school. (Self-developed)
- I find it hard to learn French in addition to English. (Adapted from Heinzmann 2013)

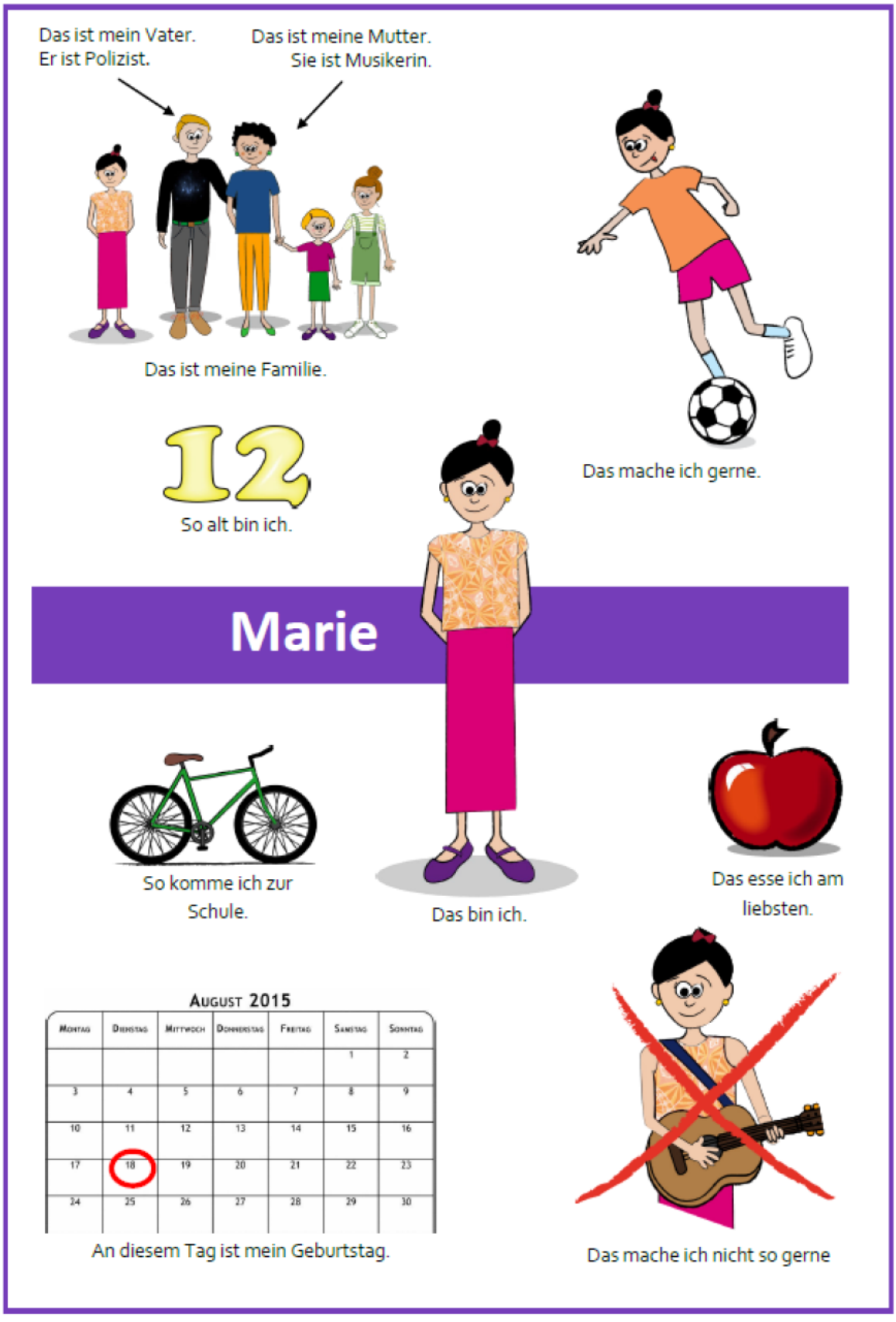

Figure A1: Example profile for the speaking test

3 Lexical distances and relative morphological complexity

Table A3: Lexical Distance of Closest L1 or German to French; Number of Students per Language

Lexical Distance to French	No. of Pupils Any Skill	No. of Pupils Listening	No. of Pupils Reading	No. of Pupils Speaking	No. of Pupils Writing	L1 (ISO639.3)
78.24	66	63	58	31	45	Italian (ita)
78.74	50	46	41	33	42	Portuguese (por)
78.89	4	4	4	4	3	Romanian (ron)
79.86	1	1	1	0	1	Galician (glg)
81.42	2	2	2	1	1	Romansh (roh)
84.03	20	19	18	9	15	Spanish (spa)
90.46	7	7	2	4	4	Macedonian (mkd)
90.6	23	20	18	9	15	Croatian (hrv)
90.9	6	6	5	2	5	Northern Kurdish (kmr)
91.03	10	10	7	6	7	Bosnian (bos)
91.21	34	32	26	12	21	Serbian (srp)
91.32	1	1	0	0	0	Hindi (hin)
91.55	39	37	34	17	25	English (eng)
92.21	8	8	7	4	6	Russian (rus)
92.27	1	1	0	0	0	Czech (ces)
92.96	8	8	7	2	6	Dutch (nld)
93.88	2	2	1	1	1	Danish (dan)
94.2	5	4	4	3	3	Polish (pol)
94.31	1	1	1	1	1	Iranian Persian (pes)
94.85	1	1	1	1	1	Swedish (swe)
95.35	61	61	50	25	47	Gheg Albanian (aln)
95.36	2	1	1	0	1	Modern Greek (ell)
95.73	57	55	49	31	44	German (deu)

Table A4: Relative Morphological Complexity of Most Complex L1 (or German) in Repertoire; Number of Students per Complexity Group

Morphol. Complexity Compared to French	No. of Pupils Any Skill	No. of Pupils Listening	No. of Pupils Reading	No. of Pupils Speaking	No. of Pupils Writing	L1 (ISO639.3)
Lower than French (0)	179	169	145	78	122	bos, ces, dan, deu, eng, glg, hin, hrv, kmr, nld, pes, pol, roh, srp, swe
Equal to French (1)	230	221	192	118	172	aln, ara (Arabic), bul (Bulgarian), cld (Chaldean Neo-Aramaic), ell, heb (Hebrew), hun (Hungarian), ita, mkd, por, ron, rus, spa

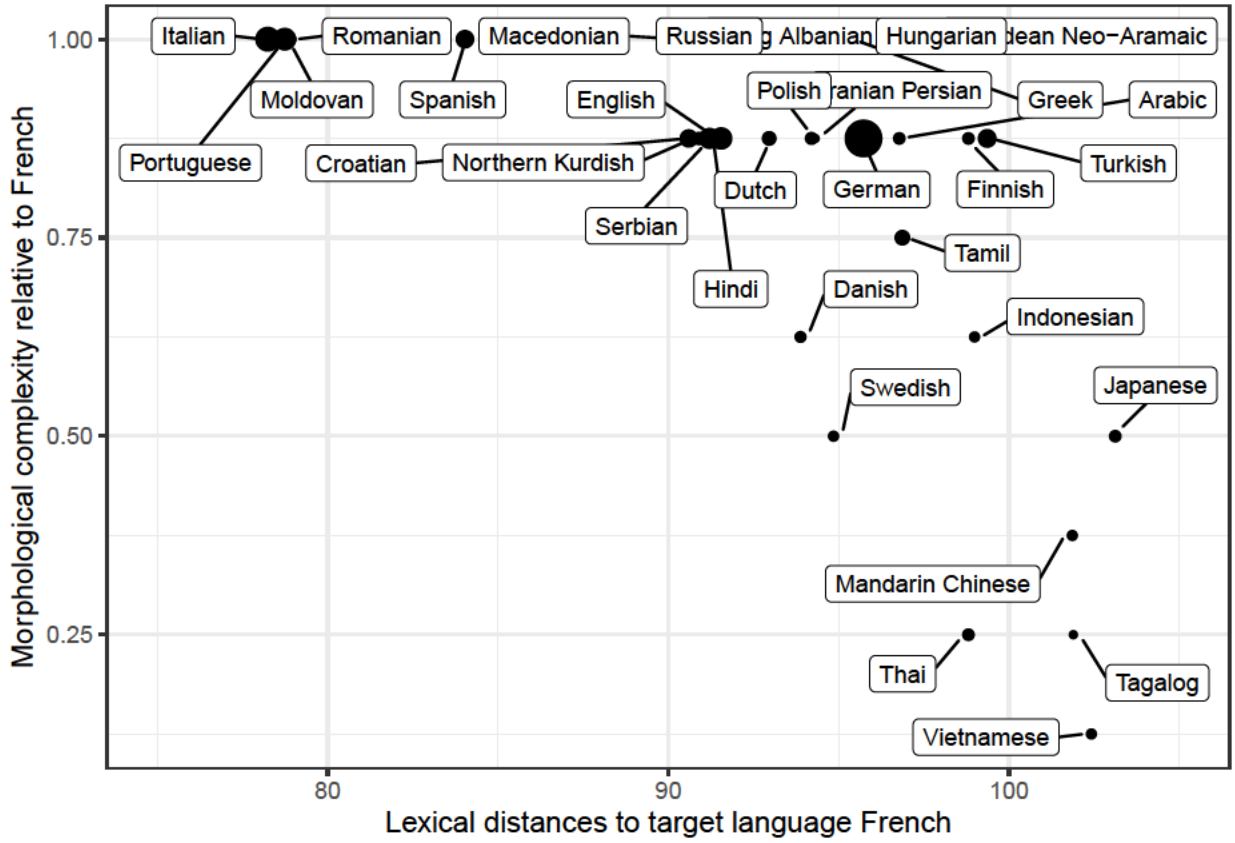

Figure A2: Lexical distances to French and morphological complexity relative to French of the languages in the sample. The size of the data points indicates the number of subjects speaking the language in question

4 Association of linguistic measures and the skills

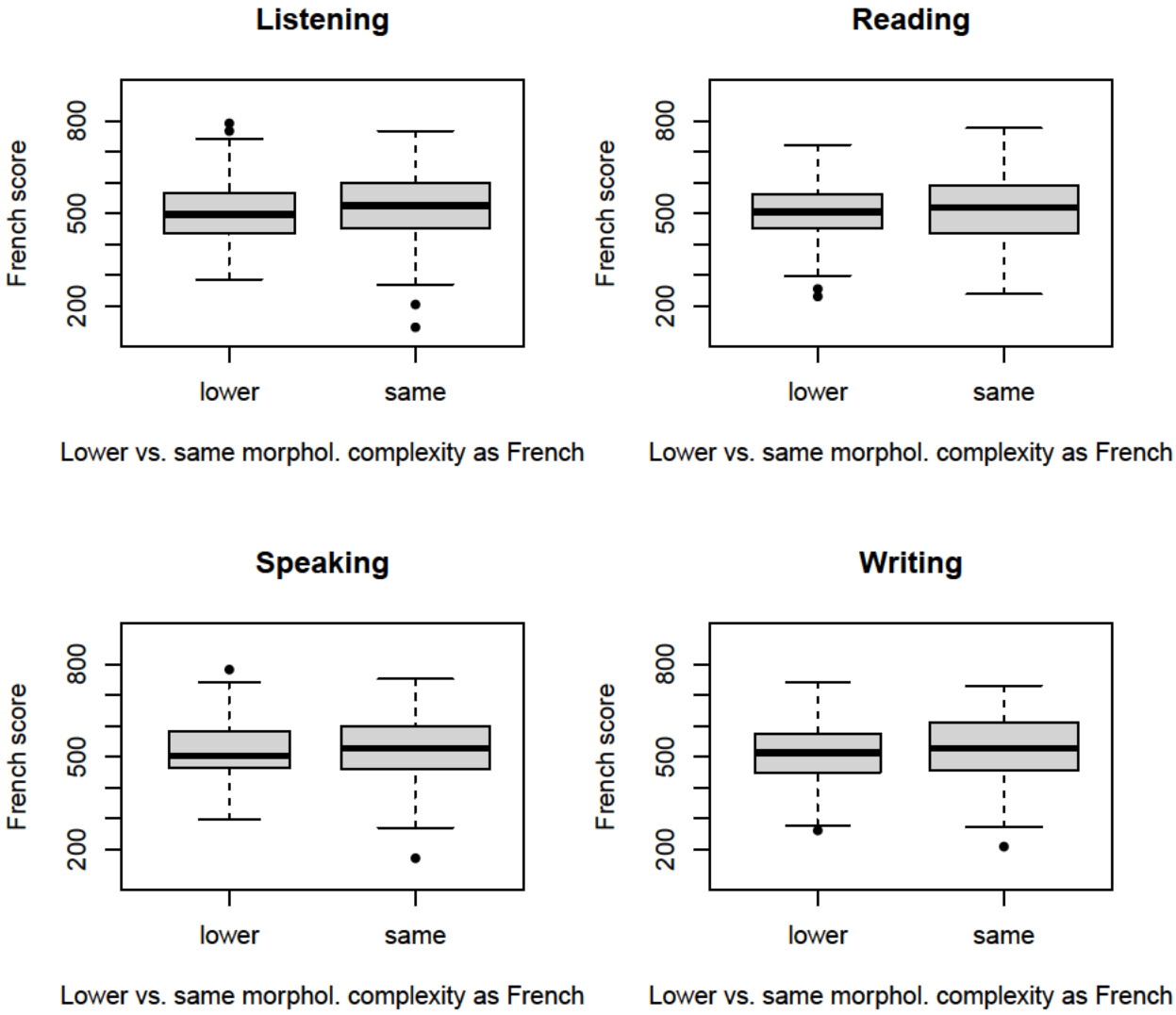

Figure A3: Relative morphological complexity and scores in the four skills

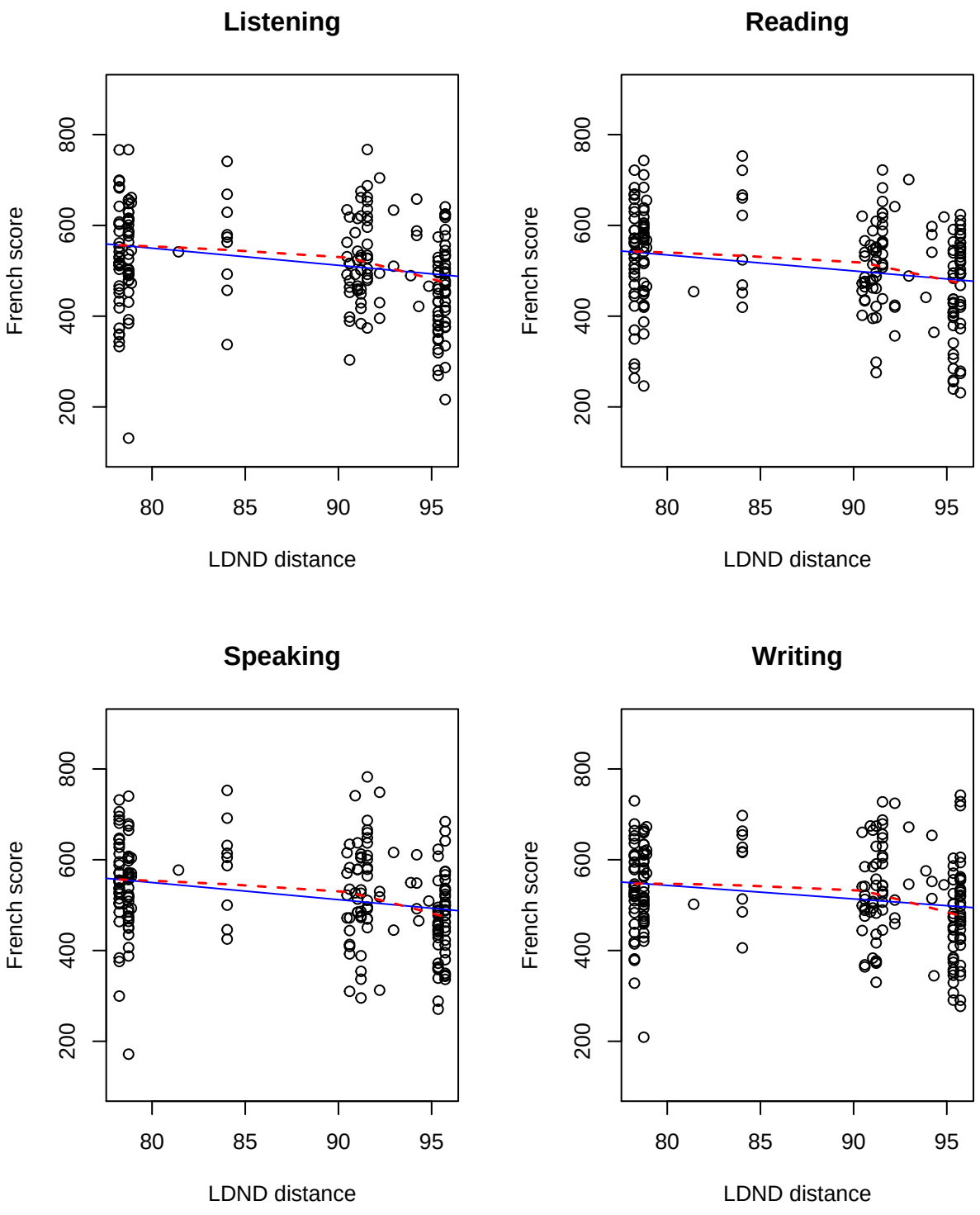

Figure A4: Lexical distance and scores in the four skills

5 Regression models

Table A5: Regression estimates listening skill

term	estimate	std.error	statistic	df	p.value
(Intercept)	511.57	15.94	32.09	193.83	0.00
Morph.Complexity French	4.88	9.70	0.50	192.76	0.62
Lex.Distance French	-20.02	5.37	-3.73	118.74	0.00
9 Lessons	-78.10	25.68	-3.04	141.53	0.00
12 Lessons	-18.08	19.27	-0.94	182.64	0.35
14 Lessons	22.77	17.86	1.27	223.44	0.20
Higher Track	38.41	15.92	2.41	81.48	0.02
Gender Male	-23.33	11.19	-2.08	34.24	0.04
Parents Migrants	-11.64	11.74	-0.99	33.75	0.33
Grade Repeated	-2.10	12.63	-0.17	56.49	0.87
Grade Skipped	-30.35	28.89	-1.05	143.65	0.30
Education Parents	1.54	5.09	0.30	139.09	0.76
Books Home	13.60	5.18	2.63	122.37	0.01
Home Possessions	-0.66	5.12	-0.13	63.37	0.90
FL Anxiety	-22.83	5.05	-4.52	55.08	0.00
Extrinsic Motivation	7.95	7.94	1.00	66.02	0.32
Intrinsic Motivation	18.03	8.64	2.09	44.78	0.04

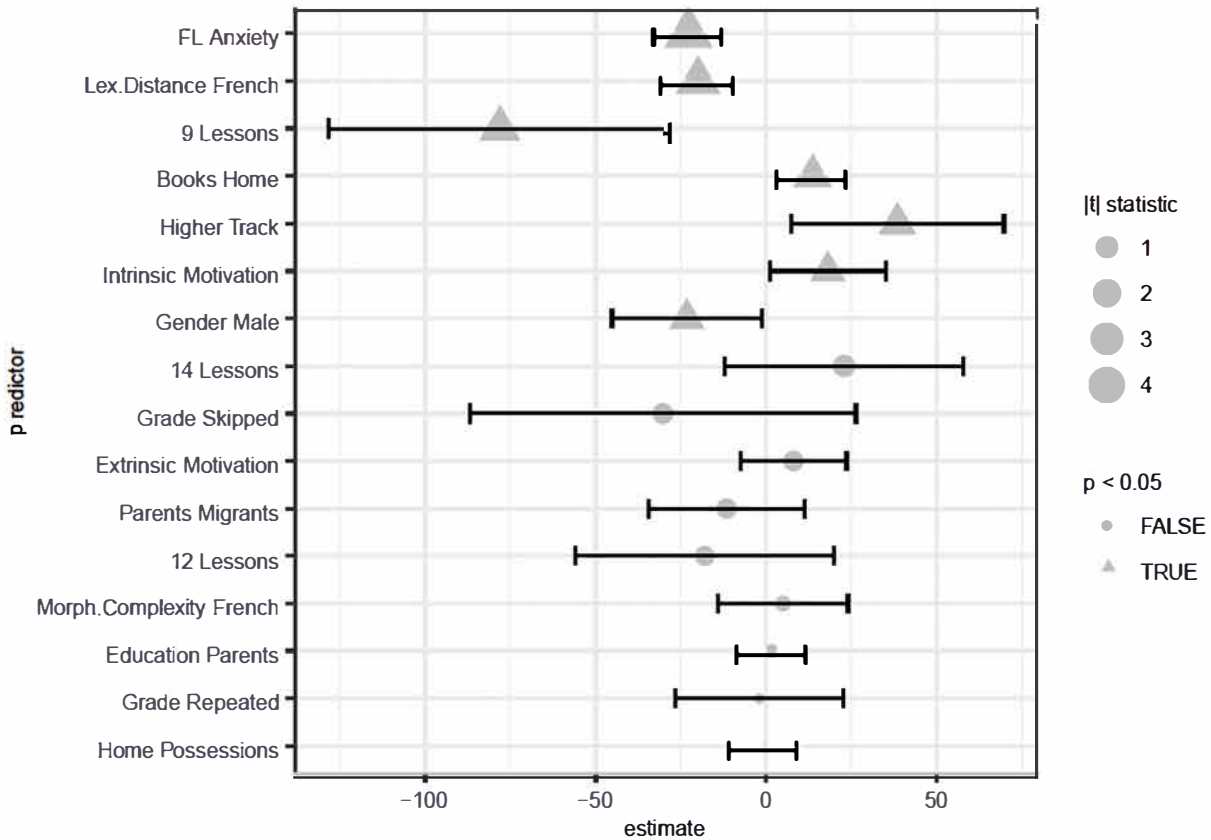

Figure A5: Listening skill

Table A6: Regression estimates reading skill

term	estimate	std.error	statistic	df	p.value
(Intercept)	502.36	19.02	26.41	78.08	0.00
Morph.Complexity French	-6.38	13.60	-0.47	41.88	0.64
Lex.Distance French	-20.87	6.68	-3.12	58.68	0.00
9 Lessons	-57.30	30.15	-1.90	106.66	0.06
12 Lessons	-0.32	22.23	-0.01	69.62	0.99
14 Lessons	13.58	18.21	0.75	185.69	0.46
Higher Track	30.99	15.48	2.00	126.42	0.05
Gender Male	-24.17	11.57	-2.09	59.01	0.04
Parents Migrants	-5.66	14.08	-0.40	28.24	0.69
Grade Repeated	18.53	14.19	1.31	78.91	0.20
Grade Skipped	-10.73	31.73	-0.34	101.90	0.74
Education Parents	4.85	6.13	0.79	80.95	0.43
Books Home	19.99	6.64	3.01	50.65	0.00
Home Possessions	-3.81	5.60	-0.68	88.59	0.50
FL Anxiety	-27.15	6.45	-4.21	36.34	0.00
Extrinsic Motivation	19.40	8.33	2.33	132.24	0.02
Intrinsic Motivation	11.63	9.37	1.24	53.38	0.22

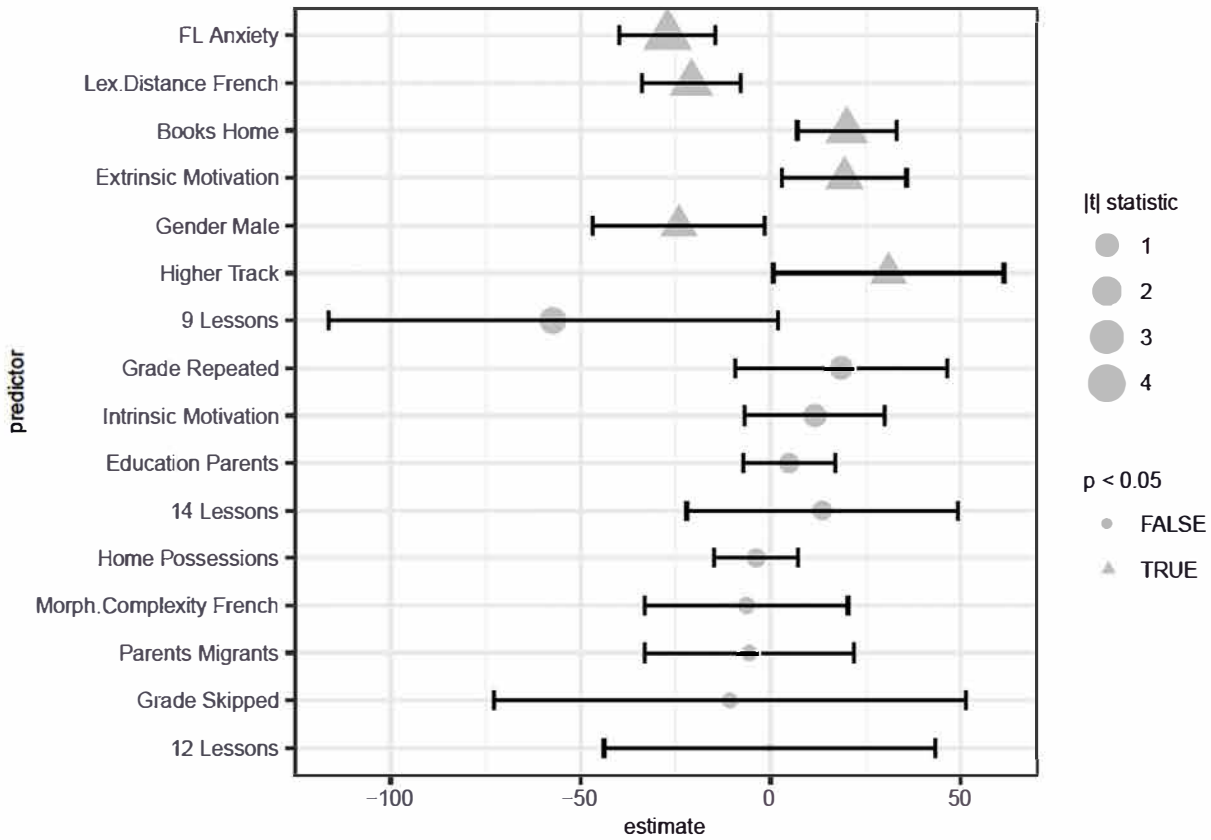

Figure A6: Reading skill

Table A7: Regression estimates writing skill

term	estimate	std.error	statistic	df	p.value
(Intercept)	487.79	16.35	29.84	153.25	0.00
Morph.Complexity French	10.13	11.48	0.88	35.34	0.38
Lex.Distance French	-14.71	6.32	-2.33	29.36	0.03
9 Lessons	-11.42	26.13	-0.44	165.70	0.66
12 Lessons	-10.45	19.22	-0.54	174.61	0.59
14 Lessons	41.61	18.77	2.22	154.31	0.03
Higher Track	57.47	14.16	4.06	236.35	0.00
Gender Male	-22.87	8.67	-2.64	111.39	0.01
Parents Migrants	-8.59	9.31	-0.92	106.80	0.36
Grade Repeated	0.86	11.51	0.07	81.99	0.94
Grade Skipped	-31.72	23.22	-1.37	100.54	0.17
Education Parents	5.02	5.38	0.93	63.78	0.35
Books Home	8.26	5.15	1.60	68.63	0.11
Home Possessions	-2.70	4.52	-0.60	133.39	0.55
FL Anxiety	-27.62	5.25	-5.26	37.08	0.00
Extrinsic Motivation	18.68	6.73	2.78	85.02	0.01
Intrinsic Motivation	18.52	6.45	2.87	152.96	0.00

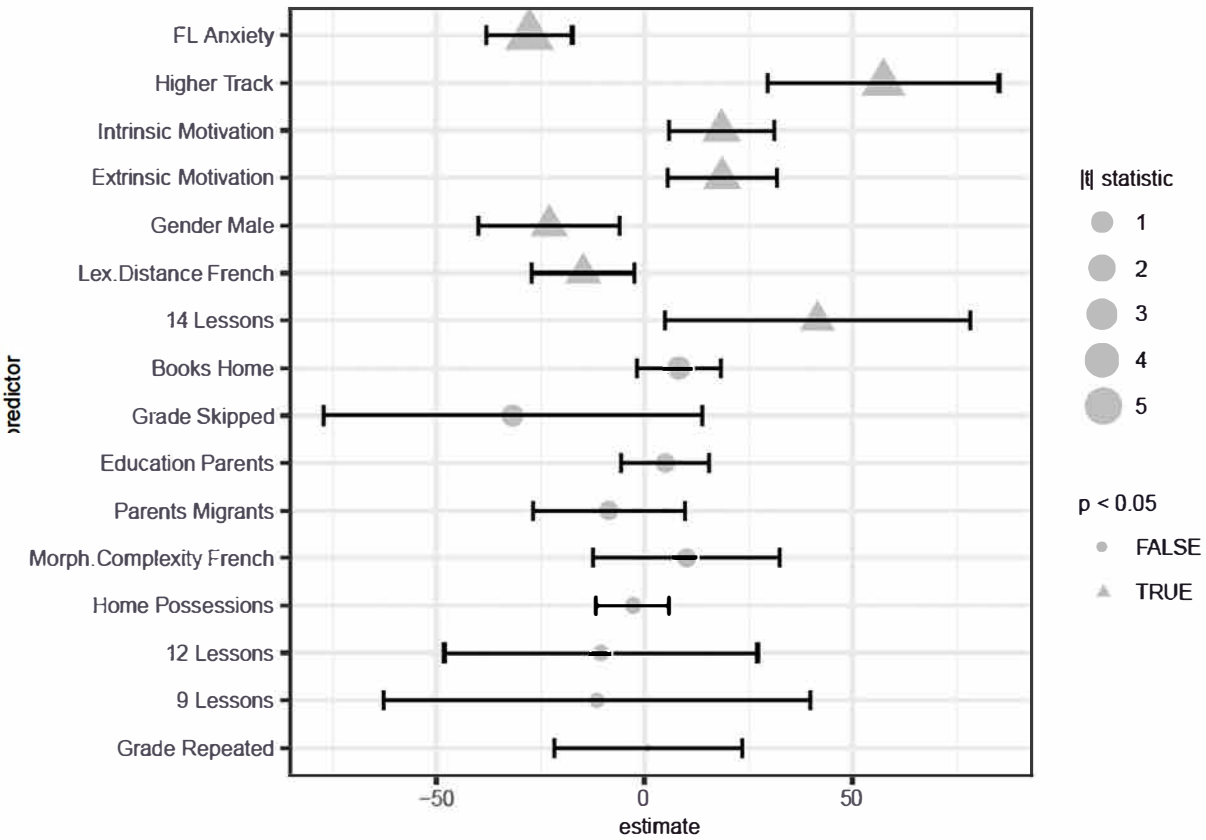

Figure A7: Writing skill

Table A8: Regression estimates speaking skill

term	estimate	std.error	statistic	df	p.value
(Intercept)	512.02	20.07	25.51	111.47	0.00
Morph.Complexity French	-8.25	13.93	-0.59	145.82	0.55
Lex.Distance French	-20.85	6.77	-3.08	138.95	0.00
9 Lessons	-36.46	27.66	-1.32	90.40	0.19
12 Lessons	12.64	20.62	0.61	102.82	0.54
14 Lessons	40.69	20.27	2.01	75.16	0.05
Higher Track	42.63	15.13	2.82	142.51	0.01
Gender Male	0.83	11.96	0.07	91.62	0.94
Parents Migrants	-26.70	12.93	-2.07	123.67	0.04
Grade Repeated	-17.15	15.79	-1.09	78.95	0.28
Grade Skipped	-55.44	31.32	-1.77	125.36	0.08
Education Parents	-1.17	6.86	-0.17	84.74	0.87
Books Home	4.36	7.09	0.62	108.86	0.54
Home Possessions	8.19	6.31	1.30	106.14	0.20
FL Anxiety	-27.35	6.65	-4.11	53.60	0.00
Extrinsic Motivation	13.00	9.80	1.33	65.04	0.19
Intrinsic Motivation	25.44	9.76	2.61	63.78	0.01

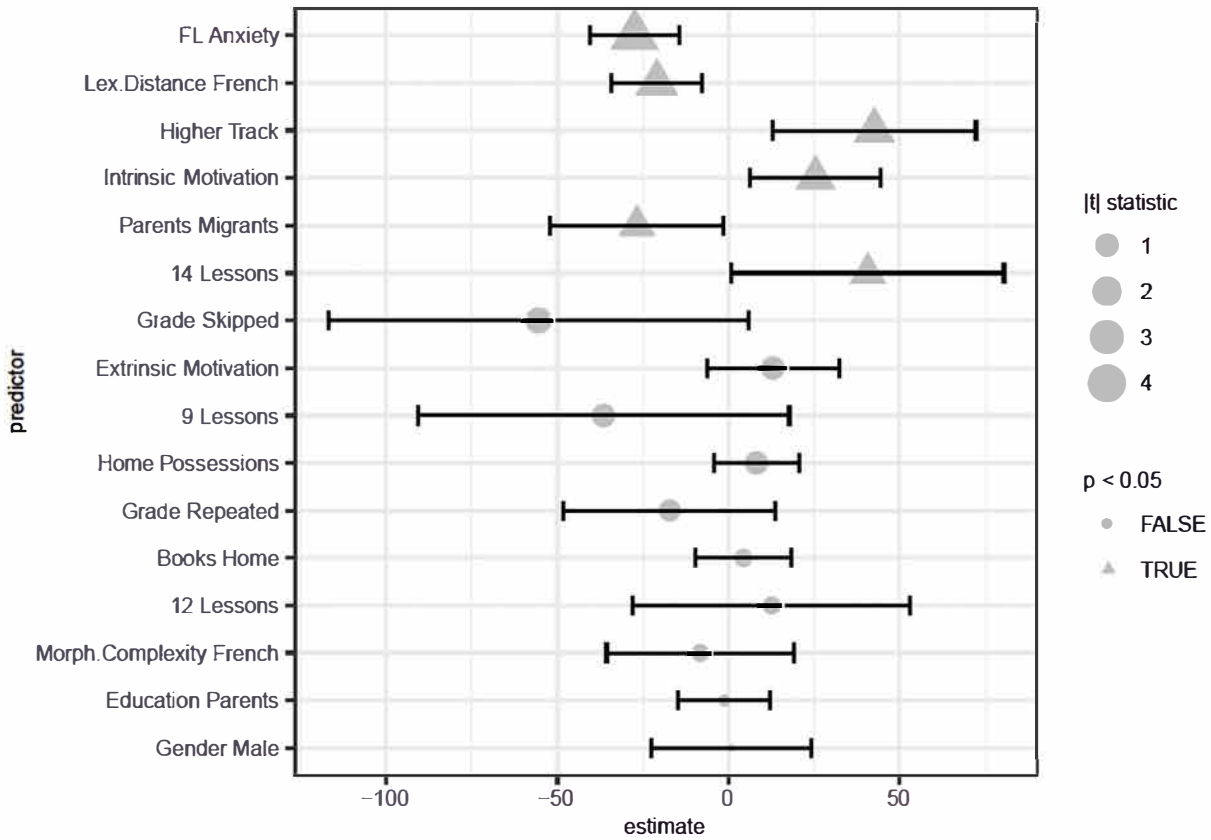

Figure A8: Speaking skill

Table A9: Random effect estimates and standard errors (se) for the four regression models that include relative morphological complexity and lexical distances

term	group	listening estimate	listening se	reading estimate	reading se	writing estimate	writing se	speaking estimate	speaking se
sd intercept	class_F	54.53	3.62	50.8	2.83	58.61	2.40	45.42	7.49
sd observation	Residual	68.71	2.46	74.4	2.17	54.58	1.34	65.62	2.83

Table A10: Fixed effect estimates and standard errors for the mixed models that include relative morphological complexity and lexical distances ($p<0.01$ flagged as ** and $p<0.05$ as *)

term	listening est	listening t	reading est	reading t	writing est	writing t	speaking est	speaking t
(Intercept)	511.57	32.09*	502.36	26.41*	487.79	29.84*	512.02	25.51*
Morph Compl French	4.88	0.5 (n.s.)	-6.38	-0.47 (n.s.)	10.13	0.88 (n.s.)	-8.25	-0.59 (n.s.)
Lex.Distance French	-20.02	-3.73*	-20.87	-3.12*	-14.71	-2.33**	-20.85	-3.08*
9 Lessons	-78.10	-3.04*	-57.30	-1.9 (n.s.)	-11.42	-0.44 (n.s.)	-36.46	-1.32 (n.s.)
12 Lessons	-18.08	-0.94 (n.s.)	-0.32	-0.01 (n.s.)	-10.45	-0.54 (n.s.)	12.64	0.61 (n.s.)
14 Lessons	22.77	1.27 (n.s.)	13.58	0.75 (n.s.)	41.61	2.22**	40.69	2.01**
Higher Track	38.41	2.41**	30.99	2**	57.47	4.06*	42.63	2.82*
Gender Male	-23.33	-2.08**	-24.17	-2.09**	-22.87	-2.64*	0.83	0.07 (n.s.)
Parents Migrants	-11.64	-0.99 (n.s.)	-5.66	-0.4 (n.s.)	-8.59	-0.92 (n.s.)	-26.70	-2.07**
Grade Repeated	-2.10	-0.17 (n.s.)	18.53	1.31 (n.s.)	0.86	0.07 (n.s.)	-17.15	-1.09 (n.s.)
Grade Skipped	-30.35	-1.05 (n.s.)	-10.73	-0.34 (n.s.)	-31.72	-1.37 (n.s.)	-55.44	-1.77 (n.s.)
Education Parents	1.54	0.3 (n.s.)	4.85	0.79 (n.s.)	5.02	0.93 (n.s.)	-1.17	-0.17 (n.s.)
Books Home	13.60	2.63*	19.99	3.01*	8.26	1.6 (n.s.)	4.36	0.62 (n.s.)
Home Possessions	-0.66	-0.13 (n.s.)	-3.81	-0.68 (n.s.)	-2.70	-0.6 (n.s.)	8.19	1.3 (n.s.)
FL Anxiety	-22.83	-4.52*	-27.15	-4.21*	-27.62	-5.26*	-27.35	-4.11*
Extrinsic Motivation	7.95	1 (n.s.)	19.40	2.33**	18.68	2.78*	13.00	1.33 (n.s.)
Intrinsic Motivation	18.03	2.09**	11.63	1.24 (n.s.)	18.52	2.87*	25.44	2.61**

Table A11: Pseudo R2 and confidence intervals of two partial effects in the multiple regression models

	Reading			Listening			Writing			Speaking		
	r^2	lower	upper	r^2	lower	upper	r^2	lower	upper	r^2	lower	upper
full model	0.322	0.248	0.397	0.330	0.261	0.399	0.443	0.370	0.517	0.449	0.361	0.537
no linguistic measures	0.294	0.219	0.369	0.305	0.236	0.375	0.420	0.345	0.496	0.424	0.333	0.514
morphology only	0.300	0.225	0.374	0.315	0.246	0.385	0.434	0.359	0.508	0.427	0.337	0.517
lexical only	0.321	0.247	0.396	0.329	0.260	0.398	0.441	0.367	0.514	0.447	0.359	0.536
full vs. no ling	0.043	0.039	0.047	0.036	0.033	0.039	0.041	0.039	0.044	0.045	0.043	0.049
morphology vs. no ling	0.008	0.008	0.009	0.015	0.014	0.016	0.023	0.022	0.025	0.005	0.005	0.006
lexical vs. no ling	0.041	0.038	0.045	0.035	0.032	0.038	0.036	0.034	0.038	0.043	0.040	0.047
full vs. morphology	0.034	0.031	0.037	0.021	0.019	0.023	0.017	0.017	0.018	0.040	0.037	0.043
full vs. lexical	0.002	0.002	0.001	0.001	0.001	0.001	0.005	0.005	0.006	0.002	0.003	0.002

Table A12: Random Effect Estimates and Standard Errors (se) for the Four Regression Models that Include Relative Morphological Complexity but Not Lexical Distances

term	group	listening estimate	listening se	reading estimate	reading se	writing estimate	writing se	speaking estimate	speaking se
sd intercept	class_F	51.71	3.25	49.90	3.16	57.55	2.43	43.75	7.47
sd observation	Residual	71.25	2.53	76.49	2.69	56.20	1.71	68.29	2.72

Table A13: Fixed Effect Estimates and Standard Errors for the Mixed Models for All Skills; Model with Relative Morphological Complexity Only (p<0.01 flagged as ** and p<0.05 as *)

term	listening est	listening t	reading est	reading t	writing est	writing t	speaking est	speaking t
(Intercept)	505.50	31.53*	494.23	26.35*	481.55	30.45*	502.67	24.73*
Morph Compl French	23.35	2.56**	14.40	1.33 (n.s.)	24.27	2.68*	15.55	1.29 (n.s.)
9 Lessons	-69.72	-2.77*	-44.06	-1.46 (n.s.)	-3.39	-0.13 (n.s.)	-26.27	-0.95 (n.s.)
12 Lessons	-16.57	-0.88 (n.s.)	1.29	0.06 (n.s.)	-9.17	-0.48 (n.s.)	15.48	0.75 (n.s.)
14 Lessons	27.54	1.58 (n.s.)	18.17	1 (n.s.)	45.45	2.43**	45.66	2.26**
Higher Track	36.70	2.35**	30.43	1.96 (n.s.)	57.53	4.07*	42.69	2.81*
Gender Male	-25.00	-2.18**	-25.28	-2.15**	-24.11	-2.74*	-1.69	-0.14 (n.s.)
Parents Migrants	-21.52	-1.93 (n.s.)	-15.66	-1.17 (n.s.)	-15.29	-1.57 (n.s.)	-35.76	-2.77*
Grade Repeated	-2.24	-0.17 (n.s.)	19.58	1.36 (n.s.)	2.51	0.21 (n.s.)	-18.16	-1.13 (n.s.)
Grade Skipped	-20.70	-0.69 (n.s.)	-6.43	-0.2 (n.s.)	-26.15	-1.11 (n.s.)	-47.56	-1.49 (n.s.)
Education Parents	-0.41	-0.08 (n.s.)	3.37	0.55 (n.s.)	3.87	0.71 (n.s.)	-2.00	-0.29 (n.s.)
Books Home	15.60	2.96*	20.81	3.09*	9.44	1.83 (n.s.)	4.35	0.6 (n.s.)
Home Possessions	0.02	0 (n.s.)	-2.47	-0.43 (n.s.)	-2.01	-0.44 (n.s.)	9.55	1.47 (n.s.)
FL Anxiety	-24.04	-4.68*	-28.96	-4.53*	-28.23	-5.32*	-28.60	-4.22*
Extrinsic Motivation	9.42	1.17 (n.s.)	21.25	2.5**	19.98	2.95*	14.28	1.43 (n.s.)
Intrinsic Motivation	18.52	2.11**	12.37	1.3 (n.s.)	19.24	2.9*	27.70	2.82*