RESEARCH ARTICLE

# Linguistic dissimilarity increases age-related decline in adult language learning

Job J. Schepens[1]* , Roeland W. N. M. van Hout[2] and Frans W. P. van der Slik[3]

[1]TU Dortmund University, Dortmund, Germany; [2]Radboud University, Nijmegen, The Netherlands; [3]Radboud University, Nijmegen, The Netherlands and North-West University, South Afrika
*Corresponding author. E-mail: job.schepens@tu-dortmund.de

## Abstract

We investigated age-related decline in adult learning of Dutch as an additional language (L$n$) in speaking, writing, listening, and reading proficiency test scores for 56,024 adult immigrants with 50 L1s who came to the Netherlands for study or work. Performance for all four language skills turned out to decline monotonically after an age of arrival of about 25 years, similar to developmental trajectories observed in earlier aging research on additional language learning and in aging research on cognitive abilities. Also, linguistic dissimilarity increased age-related decline across all four language skills, but speaking in particular. We measured linguistic dissimilarity between first languages (L1s = 50) and Dutch (L$n$) for morphology, vocabulary, and phonology. Our conclusion is that the L1 language background influences the effects of age-related decline in adult language learning, and that the constraints involved reflect both biological (language learning ability) and experience-based (acquired L1 proficiency) cognitive resources.

## Introduction

Age-related decline in learning performance is a pervasive cognitive process that occurs across all sorts of cognitive skills and learning abilities. It typically surfaces when older adults need to process and remember new sorts of information. For example, older adults may continue to learn new, additional languages even at older ages, but the learning ability as well as the ultimate attainment that is achieved in those languages tend to decrease with later starting ages of acquisition. This study focuses on general age effects on additional language (L$n$) learnability over the life span, not on maturational effects that are limited to a specific critical period. For discussions on the critical period, we refer to the large-scale study by Hartshorne, Tenenbaum, and Pinker (2018) and two recent overview studies (Birdsong, 2018; Singleton & Leśniewska, 2021). Interestingly, the L$n$ proficiency data from Hartshorne et al. (2018) show age-related decline of immersion and nonimmersion learners of English. However, to understand how

learning performance reflects age-related decline, it is crucial to compare decline across learning situations in which varying cognitive resources are available. This study compares additional language (L*n*) proficiency measures across adult learners with different starting ages (ages of onset), and a wide range of different first languages (L1s). This enables us to investigate the interaction between the effect of varying L1s and age-related decline.

Age-related decline in cognitive capacities is often subsumed under the general heading of cognitive aging. Cognitive aging directly relates to shrinking biological capacities, such as decline in brain reserve, plasticity, and fluid intelligence (Cabeza et al., 2016; Li et al., 2004; Park et al., 2001; Park & Reuter-Lorenz, 2009; Salthouse, 2012; Stern, 2009). Individual differences in cognitive aging are rather large (Nyberg et al., 2012; Schubert et al., 2020) and specific dimensions cannot be easily teased apart (Cepeda et al., 2013; Deary et al., 2010). Simple cognitive tests suggest that early age-related decline starts at around 20 years. These tests include, for example, associative recall (Shing et al., 2008), operation span (Unsworth et al., 2005), reaction time (Der & Deary, 2006), digit-symbol coding (Hartshorne & Germine, 2015), and numeracy skills (Lipkus et al., 2001). Tasks that require experience-based resources have later starting points of decline (Hartshorne & Germine, 2015), or even no decline at all, for example, vocabulary knowledge (Keuleers et al., 2015).

A distinction between fluid and crystallized intelligence (Horn & Cattell, 1967) seems to be a useful simplification to sharpen the concept of intelligence (see e.g., Kovacs & Conway, 2016; McGrew, 2009). Recent studies have found evidence for more fractionated decompositions (Hampshire et al., 2012; Johnson & Bouchard Jr., 2005; Rhodes et al., 2019). Large-scale testing has revealed a wide variance in age-related peak performances as well as their breaths across tasks that vary in the cognitive resources they require (Hartshorne & Germine, 2015). Moreover, biological and experience-based decline are not easily distinguishable. Ramscar et al. (2014), for example, conclude that older adults' performance on cognitive tests reflect their learning in handling information processing (knowledge based) and not cognitive decline (biological resources).

What can we say about additional language learning in adulthood in relation to experience-based knowledge? Particularly in the domain of pronunciation, previously learned languages are seen as important experience-based knowledge sources or skills that constrain learning success (Best, 1995; Ellis, 2006; Flege, 2018b). The role of previously learned languages might be similar to the way prior knowledge can facilitate or interfere with performance in a new learning task. Just as expectations about a target language based on previously learned languages can facilitate learning, expectations can also impede learning when new input deviates substantially from what would be expected given previous experience (Kleinschmidt & Jaeger, 2015). Available knowledge resources can both harm and help learning performance, depending on its applicability or usefulness (Brod et al., 2013; Umanath & Marsh, 2014). Learning strategies that rely on experience can be relatively effective compared to earlier life stages when less *a priori* knowledge is available (Brod et al., 2013, p. 201; Queen et al., 2012; Umanath & Marsh, 2014).

Age-related decline has strong effects on language processing (Wulff et al., 2019) and language learning (Birdsong, 2014; Bongaerts, 1999; Vanhove, 2013; Hartshorne et al. 2018). The acquisition of an L*n* in adulthood is often regarded as a more demanding and laborious task compared to earlier L*n* acquisition. Explanations range between practical (older adults receiving substantially less helpful exposure [Flege, 2018a]) and

cognitive (adults being less sensitive to new exposure due to previously acquired knowledge [e.g., Ramscar et al., 2014]), but the balance between declining L*n* learning abilities and previously acquired knowledge remains unclear.

L*n* learning outcomes differ more across older adult learners in comparison to younger adults (Marinova-Todd et al., 2000). Adult language learning seems to decline monotonically, ranging over a long period (Hakuta et al., 2003). Furthermore, age-related decline affects both language perception and production (Kemper et al., 2011; Kemtes & Kemper, 1997). However, these effects may vary depending on the specific cognitive demands of the specific language processing skills. For example, language production is generally more cognitively taxing than perception (for review, see Ferreira, 2008; MacDonald, 2013). Also, older L*n* learners experience more problems and stress in expressing grammatical knowledge during speaking and listening compared to writing and reading (McDonald, 2006).

Previously acquired knowledge explains a large part of the differences in L*n* proficiency levels across a wide range of L1s (Schepens et al., 2020), particularly because of similarities between the target language and previously learned languages. One's first language is more important than any additional language background, but additional languages result in similarity effects as well (Schepens et al., 2016). Linguistic dissimilarity or distance can be defined as the sum of linguistic distinctions between a pair of languages. Such dissimilarity measures turned out to be useful in addressing the degree of L*n* learnability with respect to the previously learned languages (Schepens et al., 2020).

Our study adopts a *large-scale* approach that is comparable to the approach taken by Schepens et al. (2020). We rely on language proficiency scores from a state exam on Dutch as a second language (STEX[1] from now on) for adult immigrants who want to study or work in the Netherlands. These are based on a reliable evaluation procedure and comprehensive assessment that includes the four basic language skills (speaking, writing, listening, and reading). Scores are available for more than 50,000 learners from 50 L1 language backgrounds and with an age of arrival between 18 and 50. In contrast to the present study, Schepens et al. (2020) did not investigate age-related decline and focused on testing scores for speaking proficiency only. More generally, our approach can be compared to educational effectiveness studies (Goldstein et al., 2007; Trautwein et al., 2006) which are also based on large-scale (cross-sectional) educational assessment scores (e.g., PISA). Recent studies on L*n* learning have also adopted approaches that analyze large-scale data (Hartshorne et al., 2018; see also van der Slik et al. 2022). These approaches exceed experimental and classroom studies in number of observations, in diversity of the subject population, and (in the present case) the comprehensive measurement of language proficiency.

Importantly, learners could voluntarily fill in a questionnaire when they participated in STEX. We use these accompanying questionnaires in addition to the actual test scores. The two key variables of interest, age of arrival and language background, are based on these questionnaires, as well as a number of other control variables. Age-related decline in L*n* learning is usually studied on the basis of the age of onset or first exposure, which is often operationalized by age of arrival or age at time of testing (e.g., Flege, 2018a; Johnson & Newport, 1989). Schepens et al. (2020) made use of three

---

[1]See https://www.staatsexamensnt2.nl/item/state-exams-dutch-as-a-second for governmental state exam resources.

linguistic similarity measures across three linguistic domains: vocabulary (Schepens et al., 2013b), morphology (Schepens et al., 2013a), and phonology (Schepens et al., 2020). This study also uses these measures to investigate their contribution to age-related decline.

We tested three hypotheses. First, we expect a turning point at around 25 years of age or earlier. We expect a change from an inclining or steady age effect to a monotonically decreasing decline. This expectation is in line with both trajectories of age-related decline in terms of fluid and crystallized intelligence (Li et al., 2004) as well as in terms of more fractionated accounts (Hartshorne & Germine, 2015). The expected turning point is outside of the disputed range of the critical period (cf. Hartshorne et al., 2018; van der Slik et al., 2022). Note that the earliest starting age of acquisition of the participants in our study is 18 years old.

Second, we expect an age-related decline for all four basic language skills with the strongest effect for speaking due to its stronger reliance on cognitive functions and resources typically associated with age-related decline.

Third, we expect that a larger linguistic dissimilarity amplifies aging effects. Specifically, we expected that learning a considerably dissimilar language at an older age should result into a stronger age-related decline compared to learning a more similar language. The extent of biological decline in cognitive functioning may be similar in both situations, but we expect that less helpful cognitive resources in the form of acquired knowledge make learning less efficient. In other words, we expect that acquired knowledge can increase cognitive aging effects. The crucial assumption is that similarity allows more reliance on acquired knowledge and therefore increments L$n$ learnability, while dissimilarity prevents reliance on acquired knowledge and therefore decreases L$n$ learnability.[2]

## Methods

### Data

We made use of a large-scale database of language testing scores gathered in the period 1995–2017. Earlier versions of this data have been used for a number of studies as well (most recently Schepens et al., 2020). This database provides a particular strong testing ground for a number of research questions related to adult language learning, given the large number of available L1s, the many countries of origin, and the available learners' social-demographic and contextual characteristics.

The data comes from the second program of the state examination for Dutch as a Second Language. This second program (STEX II) is targeted specifically at learners who intend to enroll in higher-level education in the Netherlands, or who have a higher-level occupation. Program I (STEX I) is for learners who intend to follow a lower level of (vocational) education, or who have a lower or middle-level occupation. The

---

[2]Our hypotheses emphasize the beneficial effects of similarity. However, similarity can also hamper performance, e.g., through interference (Jarvis, 2015; Jarvis & Pavlenko, 2008), particularly when learning new sounds (Best, 1995). Also, the bilingual advantage effect, if it exists, is not expected to disappear for very distinct languages (Tao et al., 2011). It is possible that interference is more impeding compared to the helpful effects of linguistic similarity at more advanced stages of learning (this study focuses on B2), possibly requiring more flexibility and therefore acting as safeguard against cognitive aging (instead of an amplification effect as we hypothesize here).

requirements for Dutch language proficiency are similar for both levels, but the abstraction (academic) level of Program II is higher. Program I is at the B1 level of the Common European Framework of Reference for Languages (CEFR), while Program II is at the B2 level. Both programs cover four language skills: speaking, listening, writing, and reading. A learner passes an exam when she or he has obtained 500 points or more on each of the four subexams. Learners cannot mix programs.

### Sample

In total, 71,989 learners took at least one of the four subexams in the period 1995–2017. In the case of reexams, we only used the first available test score. Data for age and sex were available for all learners. At the beginning of each exam, learners were invited to fill in a brief questionnaire about various background characteristics, such as year of arrival in the Netherlands, country of birth, L1, sex, and education. The questionnaire was codeveloped with one of the authors of the present study. Learners are informed about the administrative and scientific purposes of the questionnaire. Exclusion of all learners with missing information left 64,353 learners. In addition, lexical, morphological, and phonological distance scores were not available for all L1s. Exclusion of all learners with missing information left 57,603 learners. Exclusion of learners with missing scores for at least one of the four skills left 56,613 learners. Finally, restricting the data to L1s, L2s, and countries of birth containing at least 15 learners left 56,042 learners. The final sample included a diverse selection of 50 L1s,[3] consisting of both very similar languages with many participants (e.g., German) as well very different languages with many speakers (e.g., Arabic, Turkish).

Only adult second language learners who arrived in the Netherlands between 18 and 50 years of age were included in the study. We set the lower bound for age of arrival to 18 years to restrict our study to adult learners only. We set the upper bound for age of arrival to 50 years old because only a few data points were available above the age of 50 (Figure S1).

### Test scores for speaking, writing, listening, and reading

The Dutch proficiency tests were constructed by the Centraal Instituut Toetsontwikkeling (CITO; Central Institute for Test Development) and the Bureau Interculturele

---

[3]Of the L1s, 28 were Indo-European (IE) and 22 were non-Indo-European (non-IE). In the latter group, there were five Afro-Asiatic (Amharic, Arabian, Berber, Somali, Tigre), four Niger-Congo (Igbo, Swahili, Wolof, Yoruba), three Austronesian (Indonesian, Malay, Tagalog), and two Uralic languages (Finnish, Hungarian). There were two Altaic (Mongolic, Turkish), one Kartvelian (Georgian), one Japanese, one Korean, one Dravidian (Tamil), one Austro-Asiatic (Vietnamese), and one Tai-Kadai (Thai) language. The learners reported 117 countries of birth. Learners originated from 40 Western countries (including Australia, Canada, New Zealand, the United States, and former East European countries), and 23 countries from South and Central America. The remaining learners originate from 26 African countries (nine West African, six Nord African, six East African, four Southern African, and three Central African countries) and from 26 Asian countries (13 West Asian, 5 Southeast Asian, 4 Central Asian, 3 East Asian, 2 South Asian countries). Stable estimates of country and language level effects require a sufficient number of observations in the country-level combinations. The minimum amount of observations is open to discussion, however (Bell et al., 2010). We opted for the requirement that countries of origin, L1s, and speaking another L2 if present had to contain a minimum of 15 examinees to be included in this study, as we did in previous studies.

Evaluatie (Bureau ICE; Bureau for Intercultural Evaluation)—two large test battery constructors in the Netherlands. The four tests are administered and taken individually. The degree of difficulty of the examinations was held constant over time, by applying a specific Item Response Theory (IRT) model, namely the One-Parameter Logistic Model—an advanced type of Rasch model. A decisive advantage of IRT models as compared to models based on Classical Test Theory is that the test scores of candidates who took the exam on different occasions are allocated to the same ability distribution, implying that their test results can be analysed together. To achieve this, parts of earlier exams were used in new exams (though the actual design was more complicated). The scores on the exam were standardized. A mark of 500 or higher means that the candidate had passed the exam and indicates that the learner has a proficiency at the B2 level (independent user, vantage level) as defined in the Common European Framework (Council of Europe, 2001), equivalent to IELTS 5.5 (*International English Language Testing System*) (Bechger et al., 2009). The STEX II data includes tests for all four language skills, described next.

### Speaking proficiency test (25 minutes)

The typical speaking test consists of around 15 assignments. Learners are urged to respond orally to prompts like: "*Friends of yours are expecting a baby. They intend to buy a house. They show ads of two houses for sale and ask you for your opinion. You tell your friends which house you like best and why.*" Such prompts are often accompanied by visual aids. These spoken elicitations are recorded individually and digitally. Several independent expert evaluators each evaluate a separate part using both content and correctness criteria. Primary content criteria are the appropriateness of the content related to the task (about 30%) and vocabulary size (around 18%). The most important linguistic criteria are word and sentence formation (about 28%), and pronunciation (about 12%). The remaining 12% refers to fluency, rate of speech, coherence, word choice, and register. Average speaking proficiency was 517.90 (*sd* 36.23).

### Writing proficiency test (100 minutes)

A typical writing test consists of three different tasks: writing seven or eight short responses to prompts, writing two short texts, and one or two longer text of between 150 and 300 words. Several independent expert evaluators evaluate the written production on content and correctness. The primary content criterion is adequacy/comprehensibility (about 40%). The most important linguistic criterion is grammatical correctness (about 40%). The remaining 20% refer to coherence, word choice, spelling, and composition. Average writing proficiency was 521.50 (*sd* 45.51).

### Reading proficiency test (100 minutes)

Learners have to read seven texts varying in length on a variety of subjects (i.e., how to study successfully; protocol for handling complaints) and answer in total around 40 multiple-choice questions. The test evaluates comprehension skills based on instructive, evaluative, descriptive, and persuasive texts in the fields of work and education. Average reading proficiency was 521.50 (*sd* 42.35).

### Listening proficiency test (90 minutes)

Learners have to listen to five recorded interviews using headphones and answer 40 multiple-choice questions in total (on average 8 per interview). The test evaluates global listening skills based on oral reports and opinions. Average listening proficiency was 510.90 (*sd* 39.00). We do not have a clear explanation of the lower average of this test.

### Predictor variables

### Lexical distance

This is a symmetric measure that represents the sum of branch lengths that connect two languages in a phylogenetic language tree of the Indo-European language family (Schepens et al., 2013b). The measure is based on expert cognacy judgments of words in Swadesh lists (Gray & Atkinson, 2003). The branch lengths in the tree represents the degree of evolutionary change over time. We used a maximum distance value for languages that are non-Indo-European because such languages were not part of the tree. This measure is particularly sensitive for distances between Dutch and other Indo-European languages and it assumes that distances between Dutch and non-Indo-European languages are all the same (i.e., maximally distant).

### Morphological distance

This asymmetric measure compares the morphological features between languages according to differences in complexity (Schepens et al., 2013a). We used an existing list with rank orderings for the feature values of 29 morphological features (Lupyan & Dale, 2010). We computed distances for the 49 languages that have at least five available values in WALS (Dryer & Haspelmath, 2011). This measure is particularly sensitive for distances to non-Indo-European languages as it is able to distinguish between the lower morphological complexity of southeast Asian languages and the higher morphological complexity of southwestern Asian languages.

### Phonological distance

This asymmetric measure counts the number of new phonological features in a target language based on complete sound and feature inventories (Schepens et al., 2020). We used the phonological sound and feature inventories from PHOIBLE (Moran & McCloy, 2019). We computed distances for the 62 languages for which PHOIBLE lists a phoneme inventory. The result is a more uniform distribution of distances to Dutch compared to the lexical and morphological measures.

### Age of arrival in the Netherlands

We operationalized age-related changes based on reported age of arrival (AoA). Starting age of exposure or acquisition is a commonly used variable in related studies besides, for example, age at time of testing (AaT). AoA can be computed out of AaT and vice versa using length of residence (LoR, see following text). Only two out of these three variables are enough to carry the same information as all three together because of this redundancy relation. We decided to use AoA and LoR in our models instead of, for

example, AaT and LoR. AoA is more often discussed in the literature, while AaT has a more favorable distribution.

Furthermore, age of arrival is a legitimate substitute for age at first exposure if we assume that learners start to acquire the second language in question from the moment of their arrival in the host country. Van der Slik (2010) argued that this approach would be inaccurate for English as an additional language, given the prominent position of English worldwide in secondary and even primary education. In contrast, Dutch is not part of school curricula across the world, except for Belgium and some schools in the area of Germany bordering the Netherlands. Because Dutch courses in German schools are rare, we decided that we do not need to control for this situation explicitly. Indeed, our findings remain qualitatively the same when we exclude all L1 German speakers from the analysis. The majority of learners will start learning Dutch shortly before or after their arrival. We calculated age at the time of arrival in the Netherlands based on registration data for year of birth and questionnaire answers for year of arrival. The average age of arrival was 31.09 (*sd* 6.29). The average age of arrival was normally distributed across L1s.

### Length of residence

In this study, we are primarily interested in age-related decline and language background, but these effects may be intertwined. Length of residence (LoR) is a measure that can reflect a number of different relevant factors. It is not a direct measure of the degree of exposure to the target language (Flege, 2018b; Higby & Obler, 2016). Numeric measures of language exposure necessarily simplify differences across, for example, social contexts or exposure changes over the years. We control for length of residence in our analyses because of its interrelatedness with age of arrival and age-related decline. The number of years since arrival in the Netherlands was calculated based on the year of the exam and self-reported year of arrival. Average length of residence was 3.92 (*sd* 3.91)

### Length of full-time daily education

From 1995 until 2006, the questionnaire asked about learners' education using a side-by-side matrix question. Learners were asked to mark which type of education they had had (elementary, secondary, or tertiary schooling) by filling in for how many years they had been enrolled, in which country, and whether or not they had graduated. Based on this information, we were able to estimate how many years learners had had daily education from 6 years of age onward. In the present study, we condensed years of education according to the coding scheme used from 2006 onward. The question about learners' education was altered in 2006 and now asks more directly how many years learners have had formal daily education from 6 years of age onward. Possible answering categories are: (1) 0 to 5 years; (8.0%); (2) 6 to 10 years; (6.7%); (3) 11 to 15 years; (45.3%); and (4) 16 years or more. (39.8%). Average category of education was 3.17 (*sd* 0.87). The portion of learners with less than 10 years of education is highest for Armenian (32%) and Somali (31%) speakers, and lowest for Hungarian (5%) and Bosnian (5%) speakers. For all L1s, most learners have a daily education of more than 10 years. The portion of lower educated learners correlates most strongly with phonological distance ($r = .39$, $p < .001$). The variance inflation factor for daily education is unproblematic however (VIF of 1.05).

## Sex

Sex was based on registration data (not self-reported). Sixty-eight percent of learners were female, 32% were male.

## Educational accessibility

Most of the preceding variables vary across individual learners. Only the linguistic similarity measures vary across the L1s of the learners. In addition, some part of the variation in L$n$ proficiency can also be attributed to the country of birth (using a random effect across countries, see following text). Like linguistic similarity, we assume that at least part of this variation is systematic. This is not a central hypothesis of this study but rather a way to control for a possible alternative explanation. Controlling for educational accessibility is indeed a well-balanced way to capture relevant country-specific variability, even though more sophisticated and complex constructions are possible (Schepens et al., 2013b; Van der Slik, 2010; Van Tubergen & Kalmijn, 2009). For example, Van Tubergen and Kalmijn (2005) in a study on language proficiency used a variety of country characteristics such as level of modernity, political suppression, religious origin, and gross domestic product. In a similar way, to control for country differences, we included educational accessibility as a proxy for economic development. The World Bank regularly reports on education data in a wide number of countries around the world.[4] We took the gross enrollment rate in secondary schooling per country in the year the learner has arrived in the Netherlands as an indicator for a country's educational accessibility at the time learners have left their country of origin (as a percentage of the population that has secondary education age). Average enrollment rate was 80.58% (*sd* 27.40) across 117 countries.

## Statistical approach

We applied linear mixed-effects analysis by using the lme4 package (Bates et al., 2015) in R (R Core Team, 2018). Separate analyses were conducted for each of the four language skills (listening, speaking, reading, and writing). The analyses included age of arrival, length of residence, and the three measures of L1–L$n$ similarity as well as all control predictors. Specifically, we included control variables for sex, years of daily education, educational accessibility, and the two-way interactions between educational accessibility with years of daily education and sex (Schepens et al., 2020; van der Slik et al., 2015). We also included squared and cubic terms. Including polynomial terms in regression analysis is common practice to model nonlinear relationships. Visualization is important to interpret resulting models.

The random effects models included crossed random intercepts by country (C), mother tongue (L1), best additional language (L2), and the interaction of first and second languages (L1L2). Together, these random effects aim to account for the multilingual reality of the learners. Migrants from different countries may have the same L1, while migrants from the same country may speak different L1s.

All predictors were centered around their grand mean to reduce multicollinearity in interaction and higher-order terms. Unlike age of arrival and length of residence, the

---

[4]http://databank.worldbank.org/ddp/home.do (accessed April 16, 2019).

three measures of linguistic similarity (lexical, morphological, and phonological) are not intuitively interpretable. To facilitate effect size comparison across these three similarity measures, we standardized them by dividing them through their standard deviation.

## Model selection

Tables S2 and S3 describe five successive models in a stepwise forward selection process by adding additional variables, with the final Model 5 comprising the most variables. We were guided in building up the models by the patterns we observed in the data. We kept effects in our final model that are significant in at least one of the language skills to keep the models comparable. The AIC, BIC, and deviance improvement indices for Models 0 to 5 are given in Table S3 (one table for each skill).

We started with a base model, Model 0, containing only the random effects. After adding more explanatory variables, step by step, we finally arrive at our final model, Model 5. For Model 1, we gave room to nonlinearities in age of arrival effects by including squared and cubic AoA values. The squared and cubic AoA variables are necessary to handle the patterns in the age range between 18 and 27 (see Figures 1 and 2). Higher polynomials were no improvement. For Model 2, we included a linear effect and a quadratic LoR effect that turned out to be sufficient to deal with non-linearities. Another additional relevant effect was the interaction between AoA and LoR (cf. Hilby and Obler, 2016, p. 69). This pattern is visualized in Figure 3. We then included linguistic distances in Model 3 and its interactions are included in Model 4. It turned out that including squared distances in the interaction with the linear AoA
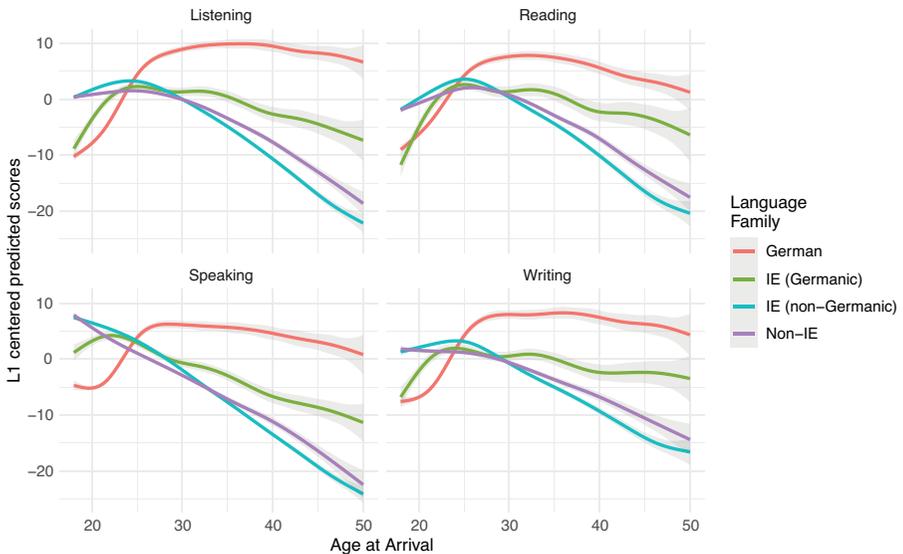


**Figure 1.** Dissimilar language backgrounds show stronger age-related decline for each basic language skill. The predicted scores control for effects of the variables included in Model 5 and they are group-centered across L1s to focus on group differences in development instead of group differences in average performance. Smooths are based on the default setting in the ggplot2 R package (Wickham et al., 2019). IE = Indo-European.
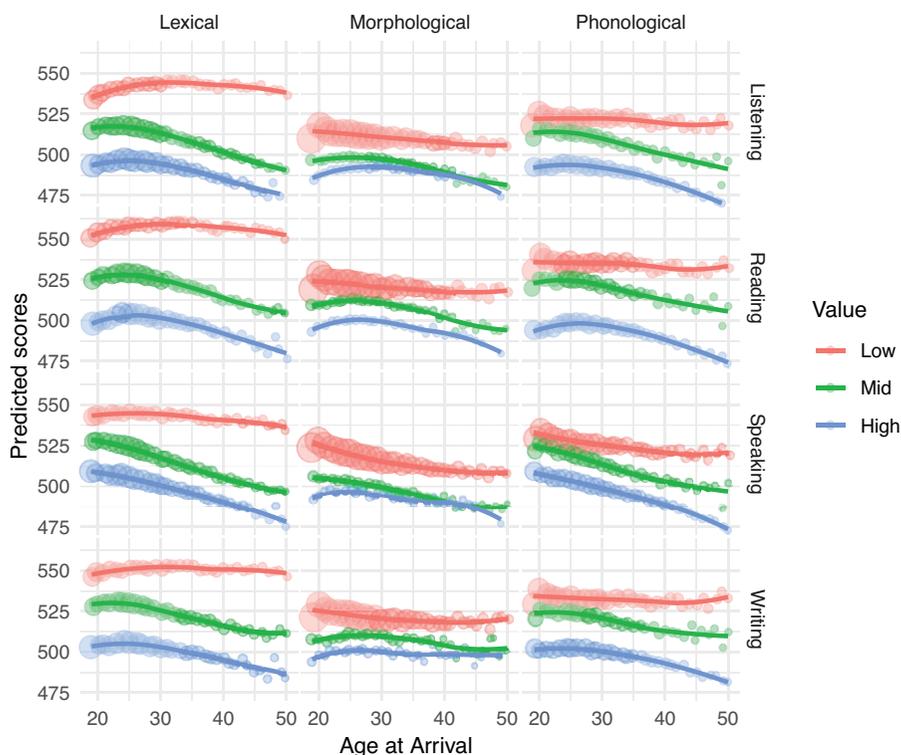
**Figure 2.** Age of arrival shows varying patterns for the three different intervals of lexical, morphological, and phonological distance. The gaps between low and high distances are highest for lexical distance and lowest for morphological distance and increase slightly at older age.

variable gave the best results. These choices are supported by the visualizations of the data patterns (see Figures 2 and 3). We did not include three-way and higher interaction effects. There is no reason to assume them given the existing literature on AoA effects. We tested nevertheless several three-way interactions, without success.

To test if Model 4 might be affected by influential cases, we calculated *dfBetas* for the four random factors, C, L1, L2, and L1L2, using the influence.ME R package (Nieuwenhuis et al., 2012). dfBetas is a measure based on the difference of an estimate with and without a particular case included (Belsley et al., 2005; Fox & Monette, 2002). It appeared that German L1 learners with English as an L2 had average scores that strongly differed from the other groups since they received a *dfBeta* in the range of 6, implying that the parameter estimates of Model 4 could be biased. We loosened the restriction of length of residence of being a fixed factor only, and we added length of residence as a random slope to the random factor L1L2 in Model 5. We chose the L1L2 random factor instead of, for example, L1 to account for as many possible patterns as possible. A recalculation now resulted in a *dfBeta* of only 1.5 for this bilingual group of German speakers. Additional analyses, not presented here, showed that German learners with English as a second language, and who additionally took their Dutch as an L2 exam in the first year of arrival were responsible for the *dfBeta* of 1.5. Excluding this particular group from the analyses resulted in a *dfBeta* of only 0.5. However, the model parameters that we calculated
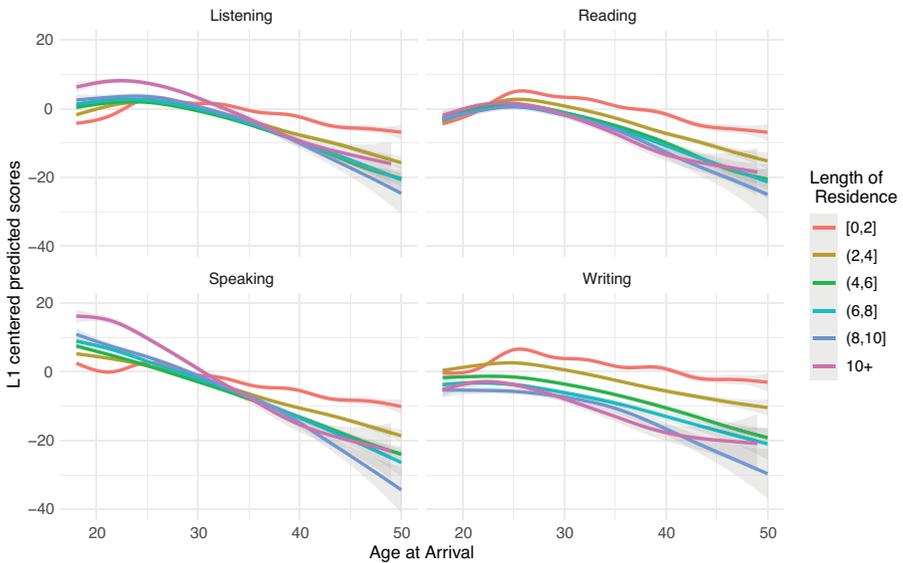
**Figure 3.** Length of residence interacts with age of arrival. Length of residence was cut into six intervals for easier visualization. A very short length of residence (e.g., red line, [0, 2]) has a relatively stable positive effect across all ages of arrival. A longer length of residence only has positive effects for younger ages of arrival. The negative effect of length of residence increases at later ages of arrival.

for the entire sample and the parameters of the model for the sample without this particular group of German language learners were highly similar. None of the $Z$ scores of the differences in parameter estimates was significantly different from 0. Model 5 is presented as our final model in Table S2. Model 4 is not listed because there were only marginal differences in the fixed effects after adding the random slope between length of residence and the L1L2 effect (Model 5). Furthermore, the residuals of Model 5 were normally distributed, except outside the $Z = |2|$ range (see Figure S2). Outside this range, many learners perform better than Model 5's predictions for receptive skills (reading and listening proficiency) and worse for the productive skills (speaking and writing). Model 5 is thus conservative for receptive skills and anticonservative for productive skills, although the differences are larger for receptive skills.

Finally, we calculated Nakagawa's *conditional* and *marginal* $R^2s$ (Nakagawa et al., 2017) using the performanceR package (Lüdecke et al., 2020) for each of the four language skills and each of the five models (see Table S4). Table S4 shows that the three linguistic distance measures explain substantially more variance compared to the other factors. The other factors are also significant, but their explained variance never exceeds 7% while the three linguistic distance measures increase the explained variance with a factor three to four. Most of the linear and nonlinear age and dissimilarities effects in Table S2 (as based on tests using the lmerTest package; Kuznetsova et al., 2017) and all the model comparisons in Table S3 (as based on chi-square tests) are significant. In all, all indices corroborate our choice for Model 5 because all improvements are highly significant.

## Results

We found significant effects for linear, quadratic, and cubic terms for AoA (one $p < .05$, others all $p < .01$), except for the cubic term in both the productive skills ($p > .05$) (see Table S2 in the Online Supplementary Material). To assess whether these aging effects are stronger in online or in productive skills, we compared the coefficients for AoA across the different models using $Z$ tests. We found that the slope for AoA is significantly steeper for Speaking compared to Listening, Reading, or Writing (all $p < .001$). The slope for AoA did not differ significantly across the other skills. The addition of LoR and its interaction with AoA was significant for all language skills ($p < .001$).

The addition of a linear effect for lexical distance was significant across all language skills ($p < .001$). Linear as well as quadratic effects for phonological and morphological distance were significant only for both productive skills (four effects, at least $p < .05$). Linear as well as quadratic interaction effects between AoA and distance were also significant, except for the writing test, in which the linear instead of the quadratic lexical interaction was significant (see Table S2 for parameter estimates and $p$ values).

Figure 1 visualizes the relationship between age of arrival and predicted scores for the four Dutch language proficiency tests. It shows that test scores generally peak before 30 years of age, with additional variation across language background and the four language skills. The recurring successive incline and decline across modalities and language background are in line with a monotonically decreasing effect of age. The pattern for German (red) shows a late start as well as slower decline compared to the other patterns. The start of decline is similar for the other three groups. The decline for non-Indo-European and non-Germanic Indo-European languages is stronger than for Germanic Indo-European languages.

Figure 2 shows how the slope for age of arrival varies according to linguistic distance. We split up the linguistic dissimilarity measures into three equal-sized intervals to help visualization. The aging pattern for similar languages is declining only very modestly across distance measures and skills. The aging pattern for the other two intervals declines more strongly. Accordingly, when distance increases, age-related decline also increases (corroborating our third hypothesis). The interval lines also show additional nonlinear patterns.

Figure 3 visualizes the interactions between length of residence and age of arrival for each language skill. Length of residence is split up in six intervals. The patterns show that a higher length of residence has a positive effect for early ages of arrival and a negative effect for higher ages of arrival. All patterns consistently indicate age-related decline.

Finally, we checked whether language background differences disappear after adding the three linguistic distances and their interactions. Figure 4 shows (for speaking proficiency) that the remaining variance in panels 3 and 4 (representing models including distance measures) is more reduced compared to the remaining variance in panels 1 and 2 (representing models excluding distance measures). Also, the datapoints are ordered less systematically along the y-axis. The residual variance along the *y*-axis in panels 3 and 4 is distributed more randomly across the language families and the explained variance along the *x*-axis is systematic. The reduction in variance indicates the part of variance explained by the linguistic distance explains. Furthermore, the lack of a discernible pattern indicates that remaining by-L1 variance across the *y*-axis results from idiosyncrasies in the data. Figure 4 does not show a clear pattern of larger negative remaining random intercepts for similar language, which indicates that interference effects do not play a large role besides linguistic distance.
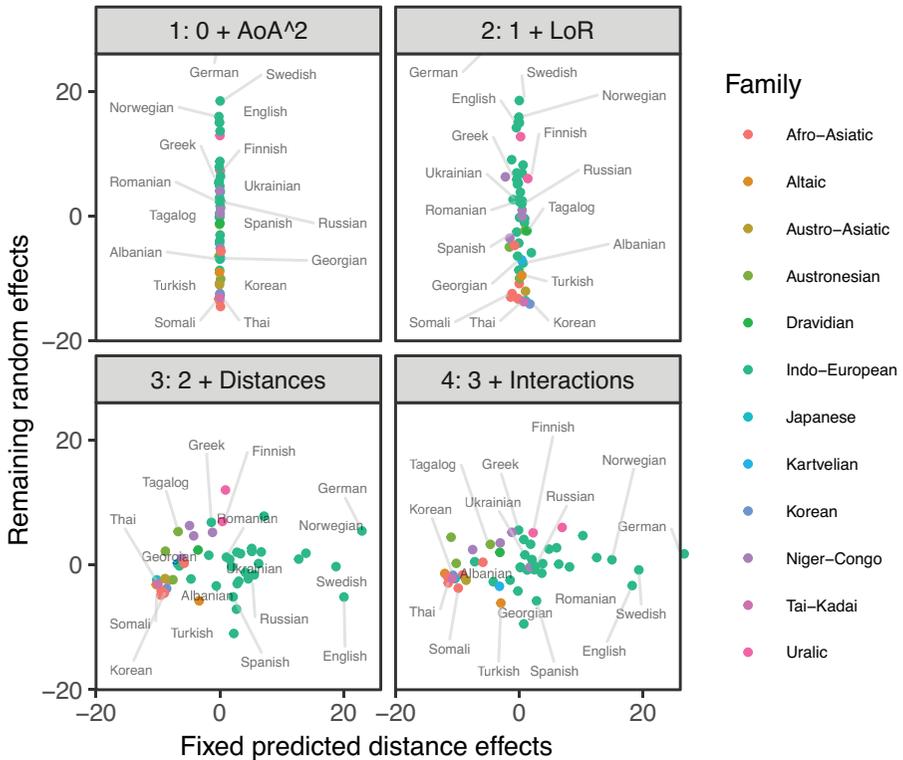
**Figure 4.** Predicted by-L1 differences for speaking proficiency (x-axes) increase with model complexity and remaining by-L1 differences (y-axes) decrease. Less random variance for the remains when more factors are included in the model. Specifically, the remaining unexplained variance of the by-L1 random effect is displayed on the y-axes (BLUPS$_{model\ x}$). The x-axes show the differences between the predicted by-L1 variance of the null model (BLUPS$_{nullmodel}$) and the remaining variance (BLUPS$_{model\ x}$). The value on the x-axes represents the predictions made by the distance effects in terms of reductions in by-L1 BLUPS. The panel numbers correspond to the model numbers in Table S3. Patterns for Models 4 and 5 were visually indistinguishable.

## Discussion

We investigated the effect of starting age or age of onset of learning on adult learners' test performances in large-scale language testing data for Dutch as an additional language for more than 50,000 learners from a broad subject population that includes 50 L1 language backgrounds. The rich and diversified language testing data made it attainable to track age-related decline across many different L1s and a broad range of starting ages of acquisition. We first discuss our findings in relation to our three hypotheses. We also evaluate our approach in general versus experimental and class-room studies and the value of our approach in understanding the role of age in adult language learning. We conclude by pointing out the educational and societal conse-quences of our findings.

First, we found, in line with our first hypothesis, an overall monotonically declining age effect in adulthood. The monotonic decline starts at least before 30 and sometimes at 20 years of age at arrival, complying with the general pattern found for many more

cognitive abilities with a peak around the age of 25 and a linear decline subsequently (Craik & Bialystok, 2006; Li et al., 2004). Growing older can be beneficial until somewhere in the earlier stages of adulthood, after which monotonic decline starts in many biological resources. Notably, peak performance differed only slightly between the four basic language skills. It seems worthwhile to compare those peak performances to those of other abilities that draw on different sorts of cognitive resources because L$n$ learning draws more heavily on higher-level, experience-based comprehension skills compared to, for example, lower-level digit and symbol manipulation skills that are typically associated with fluid cognition.

Second, in line with our second hypothesis, we found a more outspoken negative aging effect for speaking compared to listening, writing, and reading. The stronger negative effect for speaking may reflect a stronger reliance on more cognitive resources because of its online productive properties. Learners in our study had acquired relatively high-level literacy skills already through education because the exam is targeted at learners who intend to enroll in higher-level education in the Netherlands or who have a higher-level occupation. Literacy skills, being firmly established through long-term experiences, might help to compensate for aging effects in offline or receptive skills, flattening its effect, especially when linguistic distances are small (Umanath & Marsh, 2014). These latter patterns seem to shift more to the pattern of available experience-based resources with a peak at middle ages and a decline more moderately than biological resources (Hultsch et al., 1998; Li et al., 2004; Schaie, 2012).

In line with our third hypothesis, we found that a lower L$n$ learnability, as quantified by linguistic distance, shows an increasing age-related decline. This interaction effect was robust across language skills and the three linguistic distance measures. The larger the distance of the L1 to L$n$ Dutch, the more negative the effect of age of arrival on the four language skills. Learners with more distant L1s might show increasing aging effects after the maximum age of arrival of 50 years (which we had to apply given the number of participants in the data). The German as well as the wider group of learners with a Germanic language background only showed a very moderate decline. L1 Germanic learners may have sufficient experience-based resources to compensate for cognitive aging, probably because of their similar language background. This compensation effect has to be investigated further, taking in, if possible, even older learners. Compensation in this sense is based on a comparison to the average decline across all learners. Compensation is, however, also an important neural process in cognitive aging (Cabeza et al., 2016; Park & Reuter-Lorenz, 2009). Furthermore, cognitive reserve, a neuroscientific notion current in the context of Alzheimer's disease, points out that the brain can compensate for losses in brain reserve using alternative functional processes in a similar way (Stern, 2009).

Remarkably, the overall age of arrival effects came out to be stable and robust, also after language background was taken into account. We ruled out that this interaction between age of arrival and linguistic similarity could be due to a bias in prearrival language knowledge of Dutch. Possible reasons for such a bias may include tourism, historical or migration relationships, the size of expat communities, and availability of Dutch education. Although it could be the case that individual learners already speak or have started to learn Dutch as a second language before arriving in the Netherlands, such learners are relatively scarce and not country specific, and their effect would wash out due to the large-scale nature of the study. Furthermore, the baseline as well as the remaining random variance (after the various model parameters are taken into account) did not show significant deviations from normality. If there would be

systematic language or country-specific biases, these deviations (BLUPS) should have shown violations of the normality assumption.

Our approach must be seen as complementary to experimental and classroom studies due to its large-size scale and comprehensive measurement of proficiency. In particular, our approach has the statistical power to detect effects that might otherwise not be detectable (cf. Vanhove, 2013; Hartshorne et al., 2018). The diversity and scale of our sample in combination with professional language testing scores as well as background information helps to answer research questions about fundamental SLA concepts (and their relationships). Although STEX is primarily a language test, research opportunities have been acknowledged as a relevant part of the STEX administration. From the beginning onward, a short questionnaire has been part of the STEX administration procedures. The context of a language test necessitates a short and simple questionnaire that in this case establishes boundaries between L1, L2, and L*n*, which may be more blurred in the multilingual reality of the learners. The necessary compartmentalization (Gullifer & Titone, 2020) of the questionnaire cannot represent degrees of all sorts of language background. Schepens et al. (2016, 2018) have conducted specific studies of the effects of a previously learned additional language besides the L1 on learning Dutch. These studies demonstrated separate distance effects for the L1 and the best other previously learned language (L2) on learning Dutch (L*n*).

The measures of linguistic distance represent indirect measures of the required cognitive resources for learning the target language. These distance measures nevertheless explain an impressive amount of 80% of the variance that mixed-effects models attributed to the differences between the L1s. Linguistic distance measures were defined in a straightforward way, while alternative, more direct cognitive measures are often hard to operationalize. Such measures might include, for example, measures of effort, learning and instruction time, or error analyses.

Our hypotheses did not specifically assume linear effects, so we included quadratic effects in our linear regression approach to arrive at a better fitting model. The resulting model gives an indication that the main effects of age and distance as well as the interaction between age and distance are nonlinear. The nonlinear pattern we found here shows that the benefits from transfer may start to increase almost exponentially at high language similarity levels. Reversely, it perhaps also means that there can be critical limits, and after passing these, language background does not have positive effects any longer. More generally, these nonlinear interaction effects imply that variation in adult L*n* learning may hold valuable information to uncover processes of age-related decline. The age patterns that we exposed seem to show how adult L*n* learning involves a mix of cognitive resources (see e.g., Hartshorne & Germine, 2015). Further research may help to distinguish between language independent-skills and language-dependent skills (Cummins, 1979; Hulstijn et al., 2012).

There are a number of other useful tools to further study these nonlinear patterns. These include general additive modeling (Winter & Wieling, 2016), spline regression, segmented regression analysis (Rutter et al., 2020), exponential learning models (see Hartshorne et al., 2018 for an application of these methods), and cognitive modeling (Greene & Rhodes, 2022). Hartshorne et al. (2018) use large-scale learner data as well, but their proficiency measure is a grammatical judgment test only. Nevertheless, Hartshorne et al. (2018) use these data to argue that in analyzing the critical period, the concept of rate of acquisition or learning is essential. They connect language proficiency levels to rate of learning and learner age by applying a sigmoidal function. This model leads to ceiling effects in the age-related learning curves. Van der Slik et al. (2022) repeat their analysis for separate learner groups to show that the conclusion of

Hartshorne et al. (2018) about the critical period is wrong. Also, the timing of the critical period is too early to be relevant for our study. Crucially, in all models generated for all language learner groups, rate of learning gradually decreases in adulthood for all adult learners to become zero at later ages. All models lead to age-related ceiling effects in proficiency: The later the age of onset of learning, the lower the ultimate proficiency level. That means that the patterns in the Hartshorne et al. (2018) data converge with the results in our study. Learning rate in fact reflects the concept of learning ability, meaning that language learning ability suffers from negative cognitive aging. It shrinks the older the language learner.

We controlled for length of residence because of the many ways that it could influence the role of age of acquisition. Length of residence did not correlate with age of arrival ($r = .05$, *ns*). We found that longer residence had a positive effect at younger ages and a negative effect at older ages. The negative effect at older ages is likely a fossilization effect, indicating that the loss of progress becomes stronger at older ages (Han, 2004). However, the positive effect of length of residence at younger ages suggests that younger learners are likely to immerse in stimulating learning environments, where they can benefit from more exposure time and quality of input. Length of residence is not a direct measure of exposure time or quality (Flege, 2018a; Higby & Obler, 2016). Quantitative measures of language exposure necessarily simplify differences across, for example, social contexts or exposure changes over the years, which average out in comparing groups of learners. The L1 can also lead to L1-specific differences in length of residence, for example due to differences in prearrival knowledge. This is likely the case for target languages such as English or German, which are part of foreign language education in many countries. However, such biases should be less common for languages that are not widely spoken on the international level, such as Dutch.

Length of residence and the three other control variables showed significant effects, but their explained variance never exceeded a modest amount of 7%. This amount was stable across the four different language skills and is in line with previous research (for a review, see Marinova-Todd et al., 2000). The effects of the control variables in the present study are comparable to findings in our previous analyses. For an earlier discussion of the control effects, see van der Slik (2010), and for a specific study of gender and its interaction with educational accessibility, see van der Slik et al. (2015). Furthermore, the model shows that a longer education is more effective in countries with higher educational accessibility but our understanding of the effects of education in combination with linguistic distance is still limited. Other potential sources of individual variation are for instance motivation, language aptitude, living situation, and reasons for migration. Language aptitude might explain part of the wide performance range in additional second language acquisition as well because it addresses the availability of cognitive resources needed in adult language learning (Wen et al., 2019).

We also found that here was no bias any longer toward specific language families in the residual variance of our final model. The three linguistic distance measures and their interactions with age reduced remaining variance across language backgrounds to a random pattern, corroborating the validity of our model. Including lexical, morphological, and phonological distances together increased the explained variance of our models with a factor three to four across all four language skills (see Table S4). Each distance measure also had its unique contribution across all four language skills either as main effect or as an interaction with age of arrival, though in various ways. Lexical distance had comparable effects across skills. Main effects of morphological distance were significant for speaking and writing while interaction effects were stronger for

reading and listening. Phonological distance showed strongest effects for speaking. Although in varying strengths, the separate distance effects remain present across age and language skills. These findings are in line to Schepens et al. (2020), which focused on speaking only.

We conclude that a higher age leads to an increase of linguistic distance effects. Learning a dissimilar language at older age requires significantly more cognitive resources and learning effort than either dissimilarity or high age alone. In other words, a similar language background compensates (partly) for cognitive aging while a dissimilar language background amplifies it. This effect is robust across language skills and linguistic distance measures.

Societally, adult immigrants typically learn an L*n* through a mixture of immersion and instruction. Educational institutions need to understand that learning a new language can be a more demanding task when there are heavier learning difficulties resulting from linguistic distance in combination with higher age. These difficulties make it necessary to invest in professional support to set up L1-tailored educational programs, supplemented by the availability of individual language learning trajectories.

# References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bechger, T. M., Kuijper, H., & Maris, G. (2009). Standard setting in relation to the Common European Framework of Reference for Languages: The case of the state examination of Dutch as a second language. *Language Assessment Quarterly*, *6*, 126–150. https://doi.org/10.1080/15434300802457521

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Loudermilk, B. L., Kromrey, J. D., & Ferron, J. M. (2010). Dancing the sample size limbo with mixed models: How low can you go. *SAS Global Forum*, *4*, 11–14. support.sas.com/resources/papers/proceedings10/197-2010.pdf

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity* (Vol. 571). John Wiley & Sons.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–206). York Press. http://ci.nii.ac.jp/naid/10018033931/

Birdsong, D. (2014). Dominance and age in bilingualism. *Applied Linguistics*, *35*, 374–392.

Birdsong, D. (2018). Plasticity, variability and age in second language acquisition and bilingualism. *Frontiers in Psychology*, *9*, 81.

Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners. In *Second language acquisition and the critical period hypothesis* (pp. 133–159). Erlbaum.

Brod, G., Werkle-Bergner, M., & Shing, Y. L. (2013). The influence of prior knowledge on memory: A developmental cognitive neuroscience perspective. *Frontiers in Behavioral Neuroscience*, *7*, 139.

Cabeza, R., Nyberg, L., & Park, D. C. (2016). *Cognitive neuroscience of aging: Linking cognitive and cerebral aging.* Oxford University Press.

Cepeda, N. J., Blackwell, K. A., & Munakata, Y. (2013). Speed isn't everything: Complex processing speed measures mask individual differences and developmental changes in executive control. *Developmental Science*, *16*, 269–286. https://doi.org/10.1111/desc.12024

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge University Press.

Craik, F. I., & Bialystok, E. (2006). Cognition through the lifespan: Mechanisms of change. *Trends in Cognitive Sciences*, *10*, 131–138.

Cummins, J. (1979). *Cognitive/Academic Language Proficiency, Linguistic Interdependence, the Optimum Age Question and Some Other Matters.* Working Papers on Bilingualism, No. 19.

Deary, I. J., Johnson, W., & Starr, J. M. (2010). Are processing speed tasks biomarkers of cognitive aging? *Psychology and Aging*, *25*, 219–228. https://doi.org/10.1037/a0017750

Der, G., & Deary, I. J. (2006). Age and sex differences in reaction time in adulthood: Results from the United Kingdom Health and Lifestyle Survey. *Psychology and Aging*, *21*, 62.

Dryer, M. S., & Haspelmath, M. (Eds.). (2011). *The world atlas of language structures online.* Max Planck Digital Library. http://wals.info/

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*, 1–24. https://doi.org/10.1093/applin/ami038

Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation*, *49*, 209–246.

Flege, J. E. (2018a). A non-critical period for second-language learning. In *A sound approach to language matters: In honor of Ocke-Schwen Bohn.*

Flege, J. E. (2018b). It's input that matters most, not age. *Bilingualism: Language and Cognition*, *21*, 919–920. https://doi.org/10.1017/S136672891800010X

Fox, J., & Monette, G. (2002). *An R and S-Plus companion to applied regression.* Sage.

Goldstein, H., Burgess, S., & McConnell, B. (2007). Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society: Series A*, *170*, 941–954. https://doi.org/10.1111/j.1467-985X.2007.00491.x

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, *426*, 435–439. https://doi.org/10.1038/nature02029

Greene, N. R., & Rhodes, S. (2022). A tutorial on cognitive modeling for cognitive aging research. *Psychology and Aging*, *37*, 30.

Gullifer, J. W., & Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and Cognition*, *23*, 283–294. https://doi.org/10.1017/S1366728919000026

Hakuta, K., Bialystok, E., & Wiley, E. (2003). Critical evidence: A test of the critical-period hypothesis for second-language acquisition. *Psychological Science*, *14*, 31–38.

Hampshire, A., Highfield, R. R., Parkin, B. L., & Owen, A. M. (2012). Fractionating human intelligence. *Neuron*, *76*, 1225–1237.

Han, Z. (2004). *Fossilization in adult second language acquisition* (Vol. 5). Multilingual Matters.

Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? The Asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science*, *26*, 433–443. https://doi.org/10.1177/0956797614567339

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Higby, E., & Obler, L. K. (2016). Length of residence: Does it make a difference in older bilinguals? *Linguistic Approaches to Bilingualism*, *6*, 43–63.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica*, *26*, 107–129. https://doi.org/10.1016/0001-6918(67)90011-X

Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*, 203–221. https://doi.org/10.1177/0265532211419826

Hultsch, D. F., Hertzog, C., Dixon, R. A., & Small, B. J. (1998). *Memory change in the aged.* Cambridge University Press.

Jarvis, S. (2015). Influences of previously learned languages on the learning and use of additional languages. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual educational environments* (pp. 69–86). Springer. https://doi.org/10.1007/978-3-319-11496-5_5

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. Routledge.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99.

Johnson, W., & Bouchard Jr., T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*, 393–416.

Kemper, S., Hoffman, L., Schmalzried, R., Herman, R., & Kieweg, D. (2011). Tracking talking: Dual task costs of planning and producing speech for young versus older adults. *Aging, Neuropsychology, and Cognition*, *18*, 257–279.

Kemtes, K. A., & Kemper, S. (1997). Younger and older adults' on-line processing of syntactically ambiguous sentences. *Psychology and Aging*, *12*, 362.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, *0*, 1–28. https://doi.org/10.1080/17470218.2015.1022560

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*, 148–203.

Kovacs, K., & Conway, A. R. A. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*, 151–177. https://doi.org/10.1080/1047840X.2016.1153946

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in Linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science*, *15*, 155–163.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*, 37–44.

Lüdecke, D., Makowski, D., & Waggoner, P. (2020). *Performance: Assessment of regression models performance* (0.4.4) [Computer software]. https://CRAN.R-project.org/package=performance

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, *5*, e8559. https://doi.org/10.1371/journal.pone.0008559

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Language Sciences*, *4*, 226. https://doi.org/10.3389/fpsyg.2013.00226

Marinova-Todd, S. H., Marshall, D. B., & Snow, C. E. (2000). Three misconceptions about age and L2 learning. *TESOL Quarterly*, *34*, 9–34.

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, *55*, 381–401.

McGrew, K. S. (2009). *CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research.* Elsevier.

Moran, S., & McCloy, D. (Eds.). (2019). *PHOIBLE 2.0.* Max Planck Institute for the Science of Human History. https://phoible.org/

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*, 20170213.

Nieuwenhuis, R., Te Grotenhuis, H. F., & Pelzer, B. J. (2012). Influence. ME: Tools for detecting influential data in mixed effects models. *The R-Journal*, *4*, 38–47.

Nyberg, L., Lövdén, M., Riklund, K., Lindenberger, U., & Bäckman, L. (2012). Memory aging and brain maintenance. *Trends in Cognitive Sciences*, *16*, 292–305.

Park, D. C., Polk, T. A., Mikels, J. A., Taylor, S. F., & Marshuetz, C. (2001). Cerebral aging: Integration of brain and behavioral models of cognitive function. *Dialogues in Clinical Neuroscience*, *3*, 151.

Park, D. C., & Reuter-Lorenz, P. (2009). The adaptive brain: Aging and neurocognitive scaffolding. *Annual Review Psychology*, *60*, 173–196.

Queen, T. L., Hess, T. M., Ennis, G. E., Dowd, K., & Grühn, D. (2012). Information search and decision making: Effects of age and complexity on strategy use. *Psychology and Aging*, *27*, 817–824. https://doi.org/10.1037/a0028744

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*, 5–42. https://doi.org/10.1111/tops.12078

R Core Team. (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rhodes, S., Jaroslawska, A. J., Doherty, J. M., Belletier, C., Naveh-Benjamin, M., Cowan, N., Camos, V., Barrouillet, P., & Logie, R. H. (2019). Storage and processing in working memory: Assessing dual-task performance and task prioritization across the adult lifespan. *Journal of Experimental Psychology: General*, *148*, 1204–1227. https://doi.org/10.1037/xge0000539

Rutter, L. A., Vahia, I. V., Forester, B. P., Ressler, K. J., & Germine, L. (2020). Heterogeneous indicators of cognitive performance and performance variability across the lifespan. *Frontiers in Aging Neuroscience*, 12. https://doi.org/10.3389/fnagi.2020.00062

Salthouse, T. (2012). Consequences of age-related cognitive declines. *Annual Review of Psychology*, *63*, 201–226. https://doi.org/10.1146/annurev-psych-120710-100328

Schaie, K. W. (2012). *Developmental influences on adult intelligence: The Seattle longitudinal study.* Oxford University Press.

Schepens, J., van der Slik, F., & van Hout, R. (2013a). Learning complex features: A morphological account of l2 learnability. *Language Dynamics and Change*, *3*, 218–244. https://doi.org/10.1163/22105832-13030203

Schepens, J., van der Slik, F., & van Hout, R. (2013b). The effect of linguistic distance across Indo-European mother tongues on learning Dutch as a second language. In L. Borin & A. Saxena (Eds.), *Approaches to measuring linguistic differences* (pp. 199–230). De Gruyter Mouton.

Schepens, J., van der Slik, F., & van Hout, R. (2016). L1 and L2 distance effects in learning L3 Dutch. *Language Learning*, *66*, 224–256. https://doi.org/10.1111/lang.12150

Schepens, J., van Hout, R., & Jaeger, T. F. (2020). Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition*, *194*, 104056. https://doi.org/10.1016/j.cognition.2019.104056

Schubert, A.-L., Hagemann, D., Löffler, C., & Frischkorn, G. T. (2020). Disentangling the effects of processing speed on the association between age differences and fluid intelligence. *Journal of Intelligence*, *8*, 1. https://doi.org/10.3390/jintelligence8010001

Shing, Y. L., Werkle-Bergner, M., Li, S.-C., & Lindenberger, U. (2008). Associative and strategic components of episodic memory: A life-span dissociation. *Journal of Experimental Psychology: General*, *137*, 495–513. https://doi.org/10.1037/0096-3445.137.3.495

Singleton, D., & Leśniewska, J. (2021). The critical period hypothesis for L2 acquisition: An unfalsifiable embarrassment? *Languages*, *6*, 149.

Stern, Y. (2009). Cognitive reserve. *Neuropsychologia*, *47*, 2015–2028.

Tao, L., Marzecová, A., Taft, M., Asanowicz, D., & Wodniecka, Z. (2011). The efficiency of attentional networks in early and late bilinguals: The role of age of acquisition. *Frontiers in Psychology*, *2*, 123.

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, *98*, 788–806. https://doi.org/10.1037/0022-0663.98.4.788

Umanath, S., & Marsh, E. J. (2014). Understanding how prior knowledge influences memory in older adults. *Perspectives on Psychological Science*, *9*, 408–426.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498–505. https://doi.org/10.3758/BF03192720

van der Slik, F. (2010). Acquisition of Dutch as a second language. *Studies in Second Language Acquisition*, *32*, 401–432. https://doi.org/10.1017/S0272263110000021

van der Slik, F., Van Hout, R., & Schepens, J. (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*, e0142056. https://doi.org/10.1371/journal.pone.0142056

van der Slik, F., Schepens, J., Bongaerts, T., & van Hout, R. (2022). Critical Period Claim Revisited: Reanalysis of Hartshorne, Tenenbaum, and Pinker (2018) Suggests Steady Decline and Learner-Type Differences. *Language Learning*, *72*, 87–112.

Van Tubergen, F., & Kalmijn, M. (2005). Destination language proficiency in cross national perspective: A study of immigrant groups in nine western countries. *American Journal of Sociology*, *110*, 1412–1457. https://doi.org/10.1086/428931