Low-cost Language Learning: a Boost to Move?*

Evidence from Duolingo

Joop Adema†

February 26, 2025

Abstract

With the rise of the internet and smartphones, language-learning applications have become increasingly popular. Since foreign language skills enhance migrants' earning potential, the accessibility of such tools can have significant implications for international migration. This study examines the impact of low-cost language learning on (i) language acquisition, (ii) migration patterns, and (iii) migrants' language skills and integration, leveraging the staggered introduction of 84 language courses on the widely used Duolingo platform. Each course targets a directed language pair, giving rise to rich variation across country pairs over time. First, the analysis shows that course availability improved language skills among the general population. Second, a course bridging two countries strongly increases migration intentions across that corridor. Evidence on actual migration flows to OECD countries paints a less clear picture: effects are smaller and insignificant. Third, the availability of relevant language courses before migration boosts the proportion of migrants arriving with basic language skills and those securing jobs upon arrival in the European Union. Post-arrival access to relevant language courses further enhances migrants' employment outcomes, further highlighting that access to host-specific skills improves migrant outcomes.

Keywords: International Migration; Mobile Internet; Language Learning; Digital Infrastructure; Educational Technologies

JEL Codes: F22; I20; L86

*Preliminary and incomplete - please do not cite/circulate. I thank Simone Bertoli, Claudia Steinwender, Davide Cantoni, Mathias Bühler, Panu Poutvaara, Leander Andres and Lena Specht, as well as participants at the LMU Applied Micro seminar, the Verein Für Socialpolitik meetings and the Economics of Migration Junior Workshop, for their helpful comments and suggestions. I thank Lara Bieske and Padma Kadambi for excellent research assistance.

†Adema is a PhD student at the University of Munich and ifo Institute, adema@ifo.de

1 Introduction

The rapid rollout of mobile internet and the subsequent adoption of smartphones has transformed international migration. Not only does this enable more than 4.5 billion people (GSMA, 2023) to search for information about opportunities abroad and stay in contact with family and friends, it also unlocks access to modern educational technologies.[1] One of the most popular educational technologies is language learning platforms, which strongly reduced the financial and convenience cost of foreign language learning. As foreign language skills are crucial to integrating into a host society (Adserà and Pytliková, 2016), this technology is of special interest to migrants. It is especially relevant as many (prospective) immigrants lack good language skills, which has generated interest in policies that try to remedy that (Foged, Hasager and Peri, 2024). Improving their language skills improves welfare of immigrants and natives through better integration and enables to reap the large gains from immigration from less to more productivity countries (Clemens, 2011).

A decrease of the cost of foreign language learning is expected to increase learning and improve skills, and increases the attractiveness of migration destinations with attainable languages. However, it is an empirical question to what extent it does. Moreover, it is a priori unclear how the introduction of a low-cost technology impacts the selection of migrants in terms of general and language skills and whether it improves the preparedness of migrants for host-county labor markets. Hence, examining how the introduction of *low-cost language learning* platforms has affected migration patterns and migrant integration provide valuable information to academics understanding human capital decisions of cross-border workers and policy makers to design sound immigration and integration policies.

In this paper, I study the effect of online low-cost language learning on migration and integration patterns using the staggered rollout of language courses on the platform Duolingo. Duolingo provides freely available language courses, includes gamified elements[2] and targets learners at low initial levels of proficiency as it offers learning a target language using instructions in a particular source language. Each course consists of a series of topical lessons lasting a few minutes consisting of several items which consist of matching through listening, translating, speaking, multiple choice items and stories for some courses. Importantly, Duolingo was found to be an effective way for English speakers to study Spanish at a beginner level, indistinguishable from in-class instructions (Rachels and Rockinson-Szapkiw, 2018; Ersoy, 2021). Partially due to these features, Duolingo soon became the market leader on the language learning app market and has 74 million monthly active users

[1] In accordance, the share of population engaging in online courses has risen rapidly in recent years. Across the EU, the percentage of population doing any online course has risen from 5% in 2015 to 15% in 2023. Figure 1 shows the share of population doing an online course across the EU over time, showing considerable divergence across countries.

[2] To motivate and engage learners, Duolingo provides several *gamified* elements, such as competing on a leaderbord, virtual badges for additional content and regular reminders to self-set target. A course typically provides approximately several thousands of lessons, including 10s of thousands unique sentences and 1,000s of unique words, and is under constant development.

in 2023.[3] Duolingo first rolled out courses in 2012, starting with three courses in 2012, and gradually rolled out more courses in subsequnt years. As of 1st of January 2023, 84 courses to living languages are have been introduced.

To study the impact of low-cost language learning, I exploit the feature that Duolingo courses are *dyadic* in nature by enabling learning from source to target language. Combining this with information on languages spoken by country gives rise to rich variation in low-cost language learning availability on the country-language pair level and country pair level. This is best illustrated with an example. After introduction of a course between French and English, learning English for those in French-speaking countries becomes easier. I study (among others) whether interest in English increased in French-speaking countries, whether English test scores improved among native French speakers, whether migration intentions and flows increased between French-speaking and English speaking countries, and whether migrants from French-speaking countries hosted have better language skills upon arrival and integrate faster economically. Beyond this stylized example, reality is less dichotomized as many languages are spoken by a part of population in a given country. I proxy the intensity of exposure to Duolingo between an origin country and target language by the probability that a given course enables a randomly picked person in the origin to be able to newly learn the target language. Likewise, I calculate the probability a course between two languages enables two randomly picked individuals in two countries to newly communicate.[4] Combining these communication probabilities with course rollout dates, I construct a measure of exposure to low-cost language learning.

All analyses in this paper compare individuals or geographic units more exposed to those less exposed to Duolingo before and after exposure, which requires the assumption of counterfactual parallel trends. This is jeopardized if courses are rolled out in anticipation of stronger future demand for language learning (which could be related to trends in language skills and migrant integration). I argue that this is unlikely the case as Duolingo was supply-constraint for most of its history. As Duolingo was established in 2011, it had to rapidly keep up with demand and rapidly developed courses for the largest languages. Hence, Duolingo likely anticipated *levels* in demand for language learning across two languages rather than *trends*. In addition, several courses were developed by the user-community, who had no other motive than making their language available for others to learn. When it comes to migration-related outcomes, it is particularly unlikely that future trends are anticipated. Duolingo courses are not primarily targeted towards migrants; most users learn for work- or education-related reasons. Across applications, I nevertheless test for the presence of pre-trends. I proceed in three steps. First, I study how course availability has shaped language-related outcomes, such as

[3] See https://www.businessofapps.com/data/language-learning-app-market/. The market share of Duolingo in 2023 is 60%.

[4] In absence of information of the joint distribution of language skills within countries, these require the simplyifying assumption that language knowledge is randomly among a country's population.

search interest on Google Trends, test scores in the TOEFL and GRE scores and the impact on traditional language learning. Second, I estimate its effect on migration aspirations in the Gallup World Polls (GWP) and flows through a theory-guided gravity model. Third, I study its effect on the language and general skills of migrants upon arrival on subsequent integration in the European Union and the United states.

First, I study how the rollout of Duolingo courses affected online search behavior for Duolingo and available languages. I find that the first relevant course rollout in a country strongly increases search for Duolingo courses on Google Trends. Moreover, I find that online search interest in available languages increases as well, which suggest that the availability of a language course also spurs off-platform learning. To study the impact of low-cost language learning on language skills, I turn to the English-languaged TOEFL and GRE tests. These tests are typically taken by young individuals, among many whom take the test to apply for an educational program of job abroad. Using the TOEFL test, I find that exposure to Duolingo increases passive elements of language skills (reading and listen) but not active elements of language skills (writing and speaking), which is in line with the skills online platforms such as Duolingo predominantly practice. Using the GRE, which test both language and quantitative skills, I find that Duolingo exposure increases language skills, but does not change the number of test takers or changes the scores on quantitative scores, suggesting that the improved English skills are not driven by selection, but by learning.

To understand how the availability of new language learning technologies interacts with traditional language learning, I study how the participation in adult and school-based foreign language learning develops when Duolingo becomes available. Using information on German learning in *Goethe* language institutes across the world, I find that the availability of Duolingo courses to German seem to reduce the number of course participants, but not the number of exams taken. Using information on foreign languages taught in schools across the EU, I find that full Duolingo exposure increases the share of pupils learning available languages by about two percentage points.

Second, I study the effect on migration intentions and flows. For the main analysis, I use the Gallup World Polls (GWP), which is a representative survey in more than 150 countries globally and includes a question on migration aspirations and where people would like to migrate. A large advantage of the migration intention data over the flow data is that information is available for all destination countries globally. Reassuringly, migration aspirations have been shown to be predictive of subsequent migration flows (Tjaden, Auer and Laczko, 2019) and are thus a meaningful income. Motivated by a random utility model of migration, I estimate a gravity model of migration including exposure to Duolingo. I can control for pair, origin-year and destination-year fixed effects, which can control for multilateral resistance terms as well as general push and pull shocks. I find that upon large increases of Duolingo exposure, migration intentions increase gradually in the first three years. On average, I find that migration intentions increase by 45% if a course newly enabling communication between origin and destination country becomes available. I find that less than half of this effect is driven by individuals who otherwise would not have desired to emigrate, and the rest by diversion of migration intentions towards destinations with learnable languages. To examine whether this has also led to larger migration flows, I turn to the OECD yearly bilateral migration migration flow data, which includes all origin countries, but only 37 OECD countries as destinations. I find no conclusive evidence that migration flows have increased, but estimates are relatively imprecise. As an alternative, I use a recent database of scholarly migration flows developed by (Akbaritabar, Theile and Zagheni, 2024) between all countries in the world. I find that exposure to Duolingo increases these flows by 4 percentage points, which is driven by courses from English to other languages.

Third and last, I study the integration of migrants across the EU using the EU Labor Force Survey (EU LFS) from 2007–2021. Exploiting variation in timing of the rollout of relevant Duolingo courses across country pairs, relative to the arrival cohort of immigrants and survey interview timing, the effects of pre-arrival and post-arrival exposure to Duolingo can be disentangled. An advantage of this data over data from a single country is that it includes many destination countries, enabling me to control for origin-year fixed effects which capture time-varying selection into migration. Hence, I exploit variation in the availability of low-cost language learning across origin group-destination pairs over time. I find that the probability that migrants speak a language at least at beginner level increases by 20 percentage points. These effects are driven by migrants who found a job already before arrival, family migrants and those coming for education. Moreover, I find that the availability of Duolingo increases the share of migrants coming with a job upon arrival and the share of male migrants. The initial gain of employment upon arrival of about 10 percentage points decreases over time in the destination. Nevertheless, exposure to Duolingo after arrival also has a strong positive effect on employment rates.

I complement this with a study of migrant integration in the US, using the American Community Survey (ACS). Although I can not control for origin-year fixed effects, there are no trends in migrant characteristics before strong increases in Duolingo exposure. Contrary to the EU, language skills within the first year after arrival do not improve. Exposure to Duolingo after arrival does increase language skills and increased employment rates by four percentage points. In both the EU and US results seem to suggest that the availability of Duolingo before arrival shifts the educational composition of migrants towards low-skilled individuals. In addition, the English language intensity of immigrants' occupations in the US decreases. This suggests that the basic skills that Duolingo enables users to attain could be sufficient for low-skilled workers to find suitable employment that requires some language skills but less than average, which reduces the average language requirements of immigrant jobs.

The remainder of the paper is structured as follows. Section 2 reviews related literature and discusses this paper its contributions to it. Section 3 introduces the language learning app Duolingo and describes the roll-out over time and languages. Section 4 sets up a simple model of investments in language learning to understand how a decrease in the cost of language learning can impact migration patterns, and introduces and discusses the main empirical strategy. Section 5 encompasses an analysis how Duolingo course availability has impacted language learning and proficiency, section 6 studies its effects on migration aspirations and flows to OECD countries and among scholars and section 7 studies the effects on migrants' language skills and economic integration in the European Union and the United States. Section 8 concludes and discusses implications of this study.

2 Literature

This paper relates to three strands of literature: the literature on the economic, cultural and linguistic determinants of international migration, the literature on the role of new technologies in international migration, and – most closely related to this work – the literature on the role of language learning in international migration.

The first strand of literature is that aims to understand the drivers of (international) migration. Many authors have assessed economic determinants of international migration, focusing on the extent of migration flows as well as the selection and sorting of migrants Borjas (1987); Grogger and Hanson (2011). Furthermore, the role of language and culture in migrants' earnings has been thoroughly studies. Micro-level evidence has shown that relevant language skills contribute to higher labor market earnings (Chiswick and Miller, 1995; Dustmann and Fabbri, 2003). Apart from financial costs, it may also reduce the burden of applying for visas in a foreign language and many other important frictions (Jaschke and Keita, 2021). Linguistic distance between languages is a key determinant, as it enables individuals attain languages easier. Isphording and Otten (2011) show that language attainment among migrants in Germany strongly correlated to the distance between languages based on a measure of lexical distance between languages. Adserà and Pytliková (2016) survey the literature, finding that the host-country language premium ranges between 5 and 35%.[5] As these studies show that labor markets reward foreign language skills, countries sharing a language or speaking a similar language should be positively related to the size of bilateral migration flows. Belot and Ederveen (2012) have shown that between OECD countries cultural and linguistic distance is associated with lower migration flows. White and Buehler (2018) have shown that differences in individualism, uncertainty

[5]These large differences supported motivated policy makers to introduce (obligatory) language courses for some immigrant groups. In a recent survey of the literature, Foged, Hasager and Peri (2024) conclude that language learning policies for refugees have positive effects on earnings in the short run.

avoidance and perceived gender roles are the most important cultural impediments to international migration. Adserà and Pytliková (2015) have advanced the study of language by showing that lower linguistic distance is associated with larger international migration flows, after controlling for sharing a language. A fundamental limitation of such studies is that one can not control for all pair-level unobserved heterogeneity, such as unobserved cultural factors correlated to language. I contribute to this literature by showing that not only sharing a language or linguistic proximity between languages affects bilateral migration, but also the ease of learning a destination country language through available technology. Contrary to studies using time-independent measures of language and culture, I can do so controlling for unobserved dyadic factors.

The second strand of literature concerns (digital) technologies that affect international migration. The internet changed the velocity and way information spreads across the globe, which is likely to have large consequences for (prospective) international migrants. (Adema, Aksoy and Poutvaara, 2022) show that the worldwide rollout of 3G mobile technologies increased the desire and intentions to emigrate, using data from 120 countries of origin. In addition, their analysis suggests that preferred destinations change: as the internet lowers the cost of acquiring information about previously lesser known destinations, preferred destinations become more diverse. Böhme, Gröger and Stöhr (2020) show that online search behavior predicts migration flows, suggesting online search is important to finding information about potential destinations. Diving into one important element of the modern-day world wide web, Dekker and Engbersen (2014) conceptualize how social media has transformed international migration by interviewing individuals from three origin countries in the Netherlands, showing that social media reduced perceived distance to family and friends at home and enabling migrants to leverage weak social networks to organize migration and integration, thereby facilitating migration. I contribute to this literature by showing how a very specific type of internet technology, the availability of language learning apps, shape migration aspirations.

The third and final strand of literature is that of the role of language learning in international migration. The seminal work of Bleakley and Chin (2004, 2010) has documented that immigrants age at arrival is crucial for attaining host-country language skills. Using this relationship, they found that immigrants' language skills increases earnings and intermarriage with natives. Dustmann (1999) has shown that temporary migrants with a longer horizon in Germany have a larger incentive to attain the host country language and have better languages skills. Along similar lines, Wong (2023) exploits random allocation of refugees across the linguistic regions of Switzerland, finding that linguistic proximity is related to better labor market outcomes through quicker language learning. Adserà and Ferrer (2021) find that linguistic distance to English reduced earnings upon arrival more for college educated than for non-college educated migrants in Canada. Furthermore, they find that labor market earnings of men from linguistically distant countries increased substantially over time. These two studies are suggestive of the fact that labor market potential increases when language learning is easier. In line with this, policy makers have introduced (mandatory) language courses for some immigrant groups, particularly refugees. Using quasi-experimental variation in the availability such integration policies, their efficacy has been studied in several countries. In a recent survey of the literature, Foged, Hasager and Peri (2024) conclude that integration policies for refugees, including language training, have small positive effects on earnings in the short run. Only few studies have examined the role of language learning in isolation. Foged and Van der Werf (2023) study the availability of language learning in Denmark for refugees through variation in the proximity to language training centers, finding strong impacts to stay in that locality and some positive results on labor market outcomes.[6]

Contrary to language learning after arrival, less is known about language learning prior to migrating. Nocito (2021) show that English as a language of instruction in Master's degree strongly increase graduate migration from Italy. Huber and Uebelmesser (2023) and Jaschke and Keita (2021) study the closing and opening of German language institutes (Goethe Institutes; GI) where up to 100,000 individuals study German each year. Huber and Uebelmesser (2023) show that six years after opening a GI international migration flows from the country where the institute located sends more migrants to Germany. Jaschke and Keita (2021) find in the same setting that GIs affect the self-selection of migrants: upon arrival they have better language skills and are higher educated. Nevertheless, course participants still pay a considerable fee for a language course in the Goethe institutes. Freely accessible online language courses provide an opportunity to many people across the world, including to those who would not be able to afford institutional language courses. Furthermore, compared to the yearly attendance of the Goethe Institutes, the number of Monthly Active Users on Duolingo is order of magnitudes larger. Therefore, studying how low cost language learning affects international migration is complementary to studying the availability of certified language courses and provides a setting for studying the availability of language learning not in an isolated setting of migration to Germany, but practically the whole world.

3 Duolingo: an Educational Technology

Duolingo is a freely available online language learning platform with gamified features, consisting of bilateral and directional courses: a course enables one to learn a specific target using a source language. Hence, it particularly targets language learning at low initial levels of target-language proficiency. Although Duolingo was not the first online language learning platform, it gained considerably more traction than its competitors for two reasons: Duolingo is available for free and it contains many features that increase user engagement (Shortt et al., 2023). Users can set targets and Duolingo reminds users to meet the target and users can

[6] Di Paolo and Mallén (2023) use a similar design in Barcelona, finding that distance to a language center improves Catalan language skills, but not labor market outcomes.

compete against each other on a leaderboard. This allowed Duolingo to amass a market share of 64% in 2022, more than six times that of its closest competitor.[7] Duolingo's entry threshold is very low. First of all, most of Duolingo's content is available for free[8] Secondly, it allows learning from scratch, as the language of instruction is the source language the user has command over. This feature gives rise to rich variation in availability of low-cost language learning across speakers of different languages.

Users can access the platform through a desktop browser or a mobile application. Figure 2 shows a series of typical tasks of a course: it includes translation, sentence completion, written conversations and dictation. These elements are mostly very helpful to attain passive skills such as reading and listening, but potentially lack active elements of languages, such as writing and speaking. Moreover, Duolingo's learning philosophy is based on learning by doing, and does not include explicit grammar exercises (Freeman et al., 2023). By forcing users to set learning targets and reminding users of these regularly, Duolingo keeps users engaged, which could aid in overcoming commitment problems.[9] The rightmost screenshot of Figure 2 shows how Duolingo encourages users to fulfull their targets. In addition, Duolingo fosters engagement through several gamification elements, which allow users to collect points through learning and to compete against others on the platform.

As Duolingo provides language learning of a target language by instruction in a source language, it naturally targets language learning at low levels of proficiency. To also target more advanced users, a placement test is offered, so that users can start at an appropriate level. Many courses are extensive: they comprise 1,000s of practice lessons; several courses reach up to and including CEFR level B2.[10] Nevertheless, courses were gradually extended over time, and not all courses include lessons up to B2 to date.[11] Duolingo was found to be an effective way for English speakers to study Spanish at a beginner level, indistinguishable from in-class instructions (Vesselinov and Grego, 2012; Rachels and Rockinson-Szapkiw, 2018; Ersoy, 2021). Moreover, Duolingo's research department has extensively studied the efficacy of its platform on reading and listening skills, finding outcomes on par with several semesters of university courses (Jiang et al., 2021*a*,*b*). Nevertheless, there is less independent evidence on its efficacy at more advanced stages of language learning and on speaking and writing skills. However, Duolingo should not be seen as a pure substitute to traditional language learning. After obtaining a basic level of a language on Duolingo, an individual can continue learning a language at an institution that offers certification. As the individual possesses some initial language knowledge, the individual can start institutional language learning at a higher level, reducing the total cost

[7] https://seekingalpha.com/article/4570169-duolingo-stock-gamified-learning-great-growth-potential

[8] Duolingo is free to use with advertisements. An ad-free premium version is available too. In 2024, Duolingo had more than 5 million paying subscribers for the premium version.

[9] The recent work of Brade et al. (2024) shows that overcoming commitment problems improves academic performance.

[10] A B2 user can understand main ideas from complex text, interact with native speakers without strain and produce detailed test on a wide range of subjects.

[11] For an overview of the extent of specific courses, see https://duolingodata.com/.

spent on language certificates.

The first courses where made available in 2012, with English to Spanish, Spanish to English and English to German.[12] An important element of course development is that subsequent courses were built using strong support of volunteers during most of the history of Duolingo, who suggested courses in the *Duolingo Incubator*.[13] Nevertheless, many Duolingo language courses are introduced with commercial motives to attract more users to the platform and increase engagement. I discuss the determinants of rollout of Duolingo courses in section 4.2, and the uptake of courses in section 5.1.

I obtained information on the roll-out dates from an online source.[14] Courses go through three stages of development. Although courses may be available to a smaller audience before the final phase, I identify the rollout date as the day the course entered the final phase.[15] To validate the relevance of the roll-out dates, I study the impact of on online search behavior for Duolingo and available target languages in section 5.

I discard Esperanto, Klingon, High Valyrian and Latin, as they are not widely used and therefore irrelevant to international migrants. Until 2022, 110 courses involving two existing languages have been developed, of which 87 have reached the final phase. I do not include Irish, Hawaiian and Scottish Gaelic, because of the low number and lack of reliable information on the current day speakers, leaving 84 courses. Figure 1 shows the number of courses rolled out by year of introduction, as well as the total number of monthly active users across all courses. The first four courses were rolled out in 2012 and the most recent introduction took place in 2021. Figure 3 shows all available courses in 2021 in a Sankey diagram by connecting source and target languages. In total, there are 23 unique source- and 30 unique target languages. The diagram highlights that English is the most prevalent source (27 courses) and target language (22 courses), but that there is considerable variation across other languages.

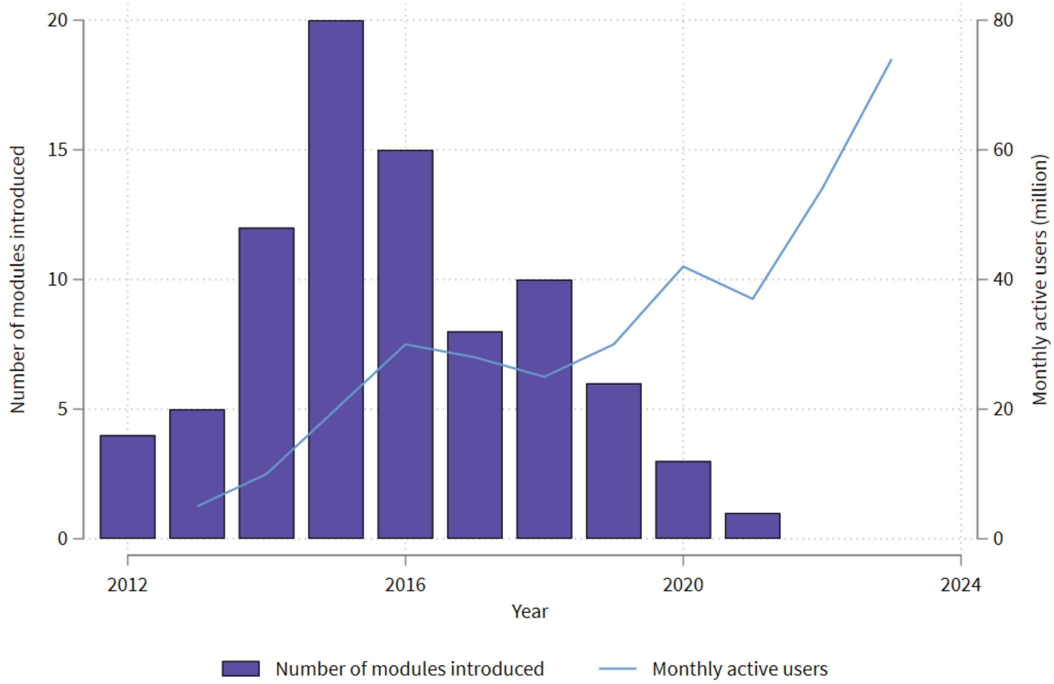
[12] More than half a million people signed up for the *beta* versions in June 2012, mobile applications on iOS and Android were released in November 2012 and June 2013, respectively.

[13] This option ended in 2021, see `https://blog.duolingo.com/ending-honoring-our-volunteer-contributor-program-2/` and `https://duolingo.fandom.com/wiki/Incubator`. As an example, the English to Russian course was fully developed by volunteers.

[14] These dates can be found on `https://duolingo.fandom.com/wiki/Course_list`. I verified these dates using the available languages on Duolingo through the Wayback Machine.

[15] In the first phase, courses are being developed, but can't be used by the general public. In the second (beta) phase the course is in testing and can be accessed by users, although it is typically not widely used. In the third and final phase the course is operational and widely used. As the number of users in the beta phase is small, I use the final phase as the rollout date. As it takes time for a course to achieve a wide audience, especially when no other course is available from the same source language or in early years when Duolingo had limited visibility, the introduction date of a course may be an imperfect proxy of when a course becomes widely adopted. Section 5 shows that online search interest in Duolingo is elevated but still very low in the quarters before a country experiences the first rollout of a relevant language course, increasing rapidly in the first 1.5 years after introduction.

Figure 1: Introduction of Duolingo Courses and the User Base over Time



Notes: Total number of courses by year of introduction and the number of Monthly Active Users (MAUs). A course is counted within a year if it reached the final stage of development within that year. MAUs are defined as unique users who engage with our Duolingo App or the learning section of the Duolingo website each month.

In 2013, Duolingo had 5 million monthly active users, which gradually rose to 74.1 million in 2023, of whom 21.4 million are active on a daily basis. In 2023, 1.5 billion hours were spent on learning, suggesting an average of about 20 hours per user. Duolingo's user base consist for 50% of females, and is relatively young: 29% of users is aged between 18 and 24, 26% is between 25 and 34, 18% between 35 and 44, 13% between 45 and 54, 9% between 55 and 64, and just 6% over 65.[16] Users' main reason for learning on Duolingo vary across country-language pairs and age categories. English learners in the U.S. are most likely to use the platform for work-related reasons, whereas those in non-English speaking countries and younger generations are more likely to use it for school.[17] Although I use the rollout of Duolingo on the language level, it is still insightful to see where Duolingo's users are located. Table 2 shows internet traffic data to Duolingo by global region. Although 27% of traffic comes from North America, considerable traffic originates from other regions across the world.

[16] Data obtained from https://www.similarweb.com/website/duolingo.com/#demographics, which only lists information on users 18 and over. Data on the numbers of users in the US suggests that about 20% of users are below 18.

[17] See https://blog.duolingo.com/english-learner-motivations/.

4 Model and Empirical Strategy

4.1 A Model of Language Learning and Migration

Duolingo is available to anyone in possession of a desktop or smartphone with internet connection. The use of the ad-based version of Duolingo is free of cost regardless of the number of courses and intensity of learning. This stands in stark contrast to traditional institutional language courses. For example, in 2024, a German course at CEFR A2 level costs 340 euro in Colombia and 180 euro in Bangladesh (Goethe Institute, 2024*a*,*b*). Motivated by this, I model the availability of a relevant Duolingo course as a decrease in the cost of language learning in a modified random utility model (RUM) of migration, in which agents choose how much to invest in language learning before observing their utility from each destination.

I model costly foreign language acquisition and migration from origin country o speaking language S to a destination country d where language T is spoken as a two-step process. I focus on one language skill as a Duolingo course enables to learn one specific *target* language, but discuss the extension to multiple language skills below. I assume that the discount factor is one. In the first step, individuals choose the optimal level of a foreign language skill $l_{oS}^T \in (0,1)$. As individuals are ex-ante indistinguishable, optimal language skill decisions are homogeneous and l_{oS}^T is not indexed by i. The cost of acquiring a target language skill T is convex in l_{oT}:

$$c_{oST}(l_{oS}^T)^2 = \frac{\kappa_{oST}}{2(1+\eta_{oST}Duolingo_{ST})}(l_{oS}^T)^2 \tag{1}$$

Here, c_{oST} depends positively on parameter κ_{oST} and negatively on the availability of a Duolingo course with effectiveness η_{oST}. In the second step, an individual observes the idiosyncratic benefits of migration to all alternatives d, ϵ_{iod}, and migrates to the destination which offers highest utility. Following the literature, I model the idiosyncratic term as an i.i.d. EVT-1 shock (Beine, Bertoli and Fernández-Huertas Moraga, 2016), which gives a convenient closed-form solution for destination choice probabilities. The utility for individual i with language skills l_{oS}^T from country o when moving to d is:

$$U_{ioSd} = \ln w_{od} + \epsilon_{iod} = \mu_{od} + l_{oS}^T b_{oSd}^T + \epsilon_{iod} \tag{2}$$

Here, μ_{od} are earnings in absence of relevant language skills in country d, l_{oS}^T are language skills of individuals from o in language T and b_{oSd}^T denotes the return to language skill for individuals from o speaking language S in country d, which I assume to be always finite, positive, and strictly positive for at least one d. For sake of simplicity, I do not explicitly include migration costs, but these can be thought of as absorbed in μ_{od} and in b_{oSd}^T, if language skills reduce migration costs. Using the properties of the EVT-1 shock, utility maximization gives the following migration probabilities:

$$\mathbb{P}_{od} = \frac{e^{\mu_{od}+l^T_{oS}b^T_{oSd}}}{\sum_{d'} e^{\mu_{od'}+l^T_{oS}b_{od'T}}} \tag{3}$$

The denominator is the sum of the deterministic part of utility in all potential destinations d', which also includes the origin. The number of migrants from o to d is given by $M_{od} = \mathbb{P}_{od}P_o$, where P_o is the initial population of origin o. Dividing the share of individuals migration $\mathbb{P}_{od}$ by the share staying $\mathbb{P}_{oo}$ and taking the natural logarithm gives a convenient expression for the log odds of migrating to d over staying in o, which is independent of the utility of alternative destinations (Bertoli and Fernández-Huertas Moraga, 2013; Beine, Bertoli and Fernández-Huertas Moraga, 2016):

$$\ln\left(\frac{\mathbb{P}_{od}}{\mathbb{P}_{oo}}\right) = \mu_{od} - \mu_o + (b^T_{oSd} - b^T_{oSo})l^T_{oS} \tag{4}$$

The log odds of migrating increase in l^T_{oS} if returns to the language skill are higher in the destination country than at home ($b^T_{oSd} > b^T_{oSo}$). However, this result does *not* imply that total migration to d is increasing in l^T_{oS}. Section A.1 shows that this is only the case when b^T_{oSd} exceeds the migration probability-weighted in all alternative destinations (including the origin). equation 1 shows that total emigration from o increases only when the weighted foreign return to language skills exceed domestic returns to the language skill.

Returning to the first step, individuals from o decide how much to invest in the language skill, given their expected utility from language skills. The expected indirect utility for someone from country o in period one is given by the expected utility in period two minus the cost of language learning:

$$U^*_{oS} = \mathbb{E}\left(U_{oSd}\right) = \sum_d U_{oSd}\mathbb{P}_{od} - c_{oST}(l^T_{oS})^2 \tag{5}$$

Language skills affect indirect utility in two ways. First, language skills increase the utility of destinations where the skill is valuable (keeping migration probabilities constant). Second, language skills increase migration probabilities towards destinations where the skill is valuable (keeping the destination-specific utility constant). As the indirect utility function is optimized w.r.t. migration probabilities, I can use the Envelope Theorem: $\frac{\partial U^*_{oS}}{\partial \mathbb{P}_{od}} = 0$. Hence, the net change in indirect utility from the second channel is zero. I obtain the following first order condition w.r.t. l:

$$2c_{oST}l^T_{oS} \approx \sum_d \mathbb{P}_{od}b^T_{oSd} \tag{6}$$

The left hand side represents the marginal cost of one unit of language skills, whereas the right hand side represents the expected marginal benefit. Importantly, the migration probabilities $\mathbb{P}_{od}$ depend on l_{oS}^T. An equilibrium pinning down l_{oT}^* exists and is unique.[18] For most countries of origin, migration probabilities are small compared to the probability of staying. In this low migration case, the right hand side only weakly depends on l_{oS}^T and an expression for optimal language skill acquisition l_{oS}^{T*} can be derived:

$$l_{oS}^{T*} \approx \left(\mathbb{P}_{oo} b_{oSo}^T + \sum_{d \neq o} \mathbb{P}_{od}(0) b_{oSd}^T \right) \frac{1 + \eta_{oST} Duolingo_{ST}}{\kappa_{oST}} \tag{7}$$

Here, $\mathbb{P}_{od}(0)$ denotes the migration probability in absence of the language skill. Two motives for language learning can be isolated. First, higher earnings on the domestic labor market increase incentives for language skill acquisition. Second, as language skills increase earnings abroad, the expected benefit is a migration-probability weighted average of returns to skill abroad. Plugging l_{oT}^* from equation 7 into equation 4 yields the following expression for the log odds of migration:

$$\ln \left(\frac{\mathbb{P}_{od}}{\mathbb{P}_{oo}} \right) = \mu_{od} - \mu_o + (b_{oSd}^T - b_{oSo}^T) \left(\mathbb{P}_{oo}(0) b_{oSo}^T + \sum_{d \neq o} \mathbb{P}_{od}(0) b_{oSd}^T \right) \frac{1 + \eta_{oST} Duolingo_{ST}}{\kappa_{oST}} \tag{8}$$

The log odds of migration depend on the distribution of returns to language skills across countries. Importantly, it depends linearly on the *difference* in earnings abroad and at home. Hence, both the size and magnitude of the effect of low-cost language learning (an increase in $Duolingo_{ST}$) depends on domestic (in o) and foreign (in d) returns to language skills. Moreover, the term in large brackets represents the degree to which language skills respond to changes in costs: it depends on the migration probability-weighted return across all countries. As most people never migrate, migration probabilities are small. Thus, this term is particularly large for foreign languages which are rewarded on domestic labor markets, such as English. Although equation 8 pertains to the low migration limit, l^* is an increasing function of $Duolingo_{ST}$ in more general cases. However, the effect is generally not linear and its precise functional form depends on the distribution of mean wages and return to language skills across countries.

Limitations. The model makes several simplifying assumptions. First of all, the model assumes that the EVT-1 shock realizes only after decisions on language skills, which renders language learning origin country-

[18] To see this, note that the right hand side is already strictly positive when l_{oS}^T is 0 ($\sum_d \mathbb{P}_{od}(l=0) b_{oSd}^T > 0$), finite when l_{oS}^T is large and everyone migrates to the country with largest returns to skill ($\lim_{l \to \infty} \sum_d \mathbb{P}_{od}(s) b_{oSd}^T = max_d b_{oSd}^T$), and that its derivative w.r.t. l_{oS}^T is always positive, as with increasing l_{oS}^T the migration probability weights on destinations with larger b_{oSd}^T become larger.

specific but not individual-specific. However, in reality individuals have heterogeneous skills and migration preferences and are more likely to invest in language skills relevant to preferred destination countries. Hence, effects are plausibly heterogeneous across individuals in the origin country. Nevertheless, the main mechanism remains the same: given expectations about migration probabilities, decreasing the cost of language learning increases language learning, which increases foreign earnings conditional on migrating and increases migration probabilities. Moreover, I assume that the EVT-1 shock is i.i.d. distributed, which is a strong assumption. In reality, preferences on the individual level are driven by preferences for (unobserved) country characteristics, which are likely to be correlated across countries. Relaxing this assumption introduces a complex error structure which depends on the characteristics of alternative destinations (Beine, Bertoli and Fernández-Huertas Moraga, 2016). I come back to the consequences of potential violations of the i.i.d. assumption in section 4.2. Furthermore, the model excludes non-earnings related explanations for language learning. As an example, language learning may have a consumption value as well (Huber and Uebelmesser, 2023). In this case, akin to the love-of-variety argument underpinning many trade models (Krugman, 1980), language learning may actually be stronger among more (linguistically) distant languages. This could explain the popularity of the Duolingo courses for Japanese, despite the limited returns to Japanese on domestic labor markets and limited number of migrants in Japan. Furthermore, it also excludes mechanisms affecting migration (intentions) through other mechanisms than language learning. Duolingo may spark interest may make target-language speaking countries more salient and induce interest in a country's culture. Such channels may ultimately affect migration.

The model does not incorporate heterogeneity across the foreign and general (i.e. education levels) skill distribution, which could have profound effects on migrant selection. First, as Duolingo targets language learning at low levels of proficiency, it could be particularly beneficial to those without or with limited initial proficiency. This could exert a downward pressure on the average foreign language skills of migrants. Second, the effects on selection into migration in terms of general skills is ex-ante ambiguous. On the one hand, as Duolingo is freely available, it could enable low-skilled individuals to whom language learning was previously too costly to acquire foreign language skills. On the other hand, foreign language learning could be complementary to general skills. As higher-skilled individuals have higher propensities to emigrate, they face stronger incentives to learn foreign languages, which in turn could improve their language skills upon migrating. Moreover, high-skilled individuals could use foreign languages as source languages (e.g. English) to learn other foreign languages. I empirically study the net effects of these forces by considering the impact of Duolingo availability on migrants' characteristics and language skills upon arrival in section 7.4. Secondly, the model only considers pre-migration learning, but not post-migration language learning. As Duolingo courses enable language learning both before and after migration, this poses a dynamic decision problem of prospective migrants deciding on the relative timing of language learning and migration. For example, migrants could delay language learning, knowing that they can utilize low-cost language learning technologies after arrival. Moreover, unanticipated access to language learning after arrival could further improve migrants' language skills. I empirically study the role of post-migration in section 7.4.

4.2 From Model to Empirical Strategy

To bring equation 8 to the data, one needs to proxy for returns to foreign language skills specific to the directed country pair *od*. In the following, I discuss how to approximate the returns using information on the distribution of languages spoken across countries.

Proxying returns to language skills. Returns to language skills play a crucial role in the model outlined above. However, these returns are not extensively measured across countries and languages. Nevertheless, the literature has provided estimates in several salient settings. Adserà and Pytliková (2016) survey the literature on immigrants' returns to destination country language skills, finding returns between 5 and 35% across contexts. Additional language skills can also increase earnings on domestic labor markets.[19] Returns to English has been found to be related to 10-50% higher earnings across countries.[20] This also extends to other widely spoken languages, such as French, German, Russian and Spanish, although the size of the estimated returns have been found to be smaller than for English.[21] An additional caveat to many of these studies of returns to skills is that they are correlational or identify local treatment effects among a specific sub-population.

In absence of such estimates of returns across dyads and languages, I proxy the foreign return of a Duolingo course between S and T by approximating the probability the course enables communication between two randomly drawn individuals in o and d, $\mathbb{P}(comm_{od}|DL_{S\to T})$. Using the distribution of speakers of languages across countries and the (strong) assumption that languages within countries are independently distributed, I can calculate the likelihood that a random individual speaking S in o can communicate with a random person in d, and how much this probability increases when one would also speak T.[22] This approach can

[19] These include both foreign and non-foreign languages. In multilingual countries, which have multiple official or widely spoken languages, returns to non-foreign languages which are not one's native language may yield considerable returns. This concerns for example German for native-French Swiss and English in India.

[20] English has large returns in India 34% (Azam, Chin and Prakash, 2013), Turkey 40% (Di Paolo and Tansel, 2015), Poland 50-60% (Adamchik et al., 2019), China 10% (Wang, Smyth and Cheng, 2017), Germany 13% (Hahm and Gazzola, 2022; Stöhr, 2015), Spain (Isphording, 2013) and across Europe (Ginsburgh and Prieto-Rodriguez, 2011).

[21] In the US foreign language skills yield small positive returns (Saiz and Zoido, 2005), in Turkey, Russian, French and German (Di Paolo and Tansel, 2015), in Poland, French, German and Spanish (Liwiński, 2019), and across Europe for French, German and Spanish (Ginsburgh and Prieto-Rodriguez, 2011).

[22] This approach is borrowed from the trade literature: Melitz and Toubal (2014) study the effect of common languages on trade flows between countries. They use the probability two individuals speak the same spoken, native and official languages to explain the role of language in international trade.

be implemented both for spoken and official languages. In the special case that o and d do not share any languages, this measure is simply equal to product of the share of source- and target language speakers in the origin and destination, respectively: $\alpha_{oS}\alpha_{dT}$. The higher the share of source language speakers in the origin and the share of target language speakers in the destination, the larger the potential gains. However, the larger the existing overlap of languages between o and d, the lower the potential gains from a given Duolingo course. In section A.3 I discuss how to calculate this object in the general case. However, this approach does not work for domestic returns to foreign languages. As most countries have one dominant language, the probability to newly communicate with someone in your own country from a language course is negligible. Instead, I assume that the domestic return to learning T from S is simply the product between the share of speakers of S and T, $\alpha_{oS}\alpha_{oT}$. The reasoning behind this is that non-native languages T that are valued on domestic labor markets amass a considerable number of speakers.

I obtain the share of speakers across countries and languages, α_{cL} and official languages by country from Melitz and Toubal (2014) and Ginsburgh, Melitz and Toubal (2017), who collected information about all languages spoken (natively) by at least 4% of population in most countries worldwide, as well as all official languages. For several missing observations, I complete the data using the most recent CIA World Factbook. I implement this approach for spoken source languages, as command of a non-native source language may facilitate the learning of other target languages. For target languages, I calculate it using spoken languages as well, but also for official languages as a robustness test. The reason for the latter is that official languages may better reflect returns to language skills b^T_{oSd}. For example, spoken languages include minority languages and foreign languages, which have limited use on the country's labor markets.

Constructing foreign and domestic exposures. To construct a time-varying measure of foreign and domestic Duolingo exposure, I interact the proxy for returns with a binary indicator that takes value one the first year a Duolingo course between S and T has been fully available.[23] In practice, multiple courses ST may "bridge" a country pair od. For sake of simplicity, I take the largest exposure at any point in time. This implies that the treatment can increase more than once. For example, exposure from Netherlands to France increased for the first time when the English to French module became available, for the second time when German to French became available and for the third time when Dutch to English became available (see Figure 12). By construction, these measures are bounded between 0 and 1.

[23] For example, this switches from zero in 2012 to one in 2013 for the course English to Spanish, which has been introduced during 2012.

$$DL_{odt} = \max_{S,T} \mathbb{P}(comm_{od}|DL_{S\to T}) Duolingo_{STt}$$

$$DL_{oot} = \max_{S,T} \alpha_{oS}\alpha_{oT} Duolingo_{STt}$$

Figure 10 shows a heatmap of the intensity of exposure DL_{odt} across country dyads, showing considerably variation within and across dyads over time. Although by construction the largest exposure is one, plenty of dyads have treatment intensities less than one because a course only newly enables communication between part of the origin and destination country. Figure 11 shows the average exposure by origin countries show that there is plenty of variation across origin countries over time, except for several linguistically isolated countries, or countries that only speak languages not covered by Duolingo. Figure 13 shows the domestic exposure across countries, showing that about half of the countries are unexposed.

Throughout the remainder of this paper, I also estimate treatment effects of Duolingo exposure on outcomes varying at different levels than the country pair level. In such cases, I construct the exposure analogously. For example, for analyses on the country-language level I construct the exposure as the probability a Duolingo course enables communication to a person speaking target language T, which is an increasing function of α_{cS} and a decreasing function of α_{cT}.

Estimation equation. I estimate equation 8 by stacking all origin countries o, exponentiating both sides, plugging in the exposure measures and adding a well-defined error term with mean 1 and adding a time component t. This boils down to the following gravity model of migration with two continuously varying treatments:

$$\frac{M_{odt}}{M_{oot}} = \exp\left[\beta_1 DL_{odt} + \beta_2 DL_{oot} + \gamma^{'}\mathbf{X_{odt}} + (\phi_{ot}) + \theta_{dt} + \psi_{od}\right]\eta_{odt} \tag{9}$$

$\frac{M_{odt}}{M_{oot}}$ are the migration odds, comprised of M_{odt}, the number of (aspiring) migrants from country o to country d at time t, and M_{oot}, the number of (aspiring) stayers from country o. The regression coefficients β_1 identifies the effect of the availability of a Duolingo course enabling communication between the full populace of o in d on the value of living in d. Importantly, this is not the same as the effect of a typical Duolingo course on migration odds, due to the multilateral resistance exercised because alternative same-language destinations receive treatment at the same time, which could be close substitutes to the focal destination. One can think of β_1 of the counterfactual effect on migration odds if a single destination would have received treatment. β_2 identifies the effect of being able to learn another domestically spoken language.

The covariate vector $\mathbf{X}_{odt}$ includes a dummy for joint EU membership, WTO trade agreements and bilateral trade flows. ψ_{od} captures unobserved pair-level unobserved factors. ϕ_{ot} and θ_{dt} indicate a set of origin-year and destination-year fixed effects that capture unobserved heterogeneity at those levels. Without origin-year level fixed effects and destination-year level fixed effects, language area-specific shocks temporally coinciding with the rollout of Duolingo courses could generate spurious effects[24] and to account for bias due to the inward and outward multilateral resistances.[25] As some of the terms following from the model only vary at the origin-time level, I estimate models with and without ϕ_{ot}. η_{odt} is an error term with unit mean.

4.3 Identification

The identification strategy underlying equation 9 is a generalized differences-differences strategy. Hence, to interpret the estimates of β's in prior sections as Average Treatment effects on the Treated (ATT), the following identifying assumptions need to be satisfied. First, there should be no anticipatory effects of Duolingo availability. It seems plausible that language learning and migration intentions are not affected by the future availability of Duolingo courses. Second, trends between exposed and unexposed units in absence of treatment should be common, for all levels of treatment intensity. For the multiplicative model of 9, this requires migration odds of origin-destination pairs to follow parallel trends in proportions in absence of treatment, conditional on the (origin-year) and destination-year fixed effects. This implies that growth rates in migration odds would have developed similarly in units with more or less Duolingo exposure, if the relevant Duolingo courses would not have been made available.

The parallel trends assumption could be violated if the rollout of courses anticipates trends in demand for language learning that are correlated to trends in migration outcomes. However, there are two reasons why this is likely not the case for Duolingo courses: courses did not target migrants and course rollout was plausibly supply-constrained and did not depend on trends in demand for learning, but rather on levels. The primary motive of language learning on Duolingo is often not preparation for migration. Duolingo asks users for the main motivation for studying foreign languages. In 2020, 33.8% indicated to learn English for school, 15.8% for work, 13.2% for brain training, 9% for family and 7.3% for cultural reasons. Only 12.6% learns English because of travel (which may be partially for tourism-related reasons) and 8.4% because of other reasons (which could include migration-related reasons). In fact, Duolingo's founder, Luis von Ahn, was motivated to increase domestic wages of Guatemalans on their home-country labor market. In addition, as Duolingo was released in 2012, the rollout of courses was initially supply-constrained as courses had to be developed from scratch. The development stage of the average course took several 100s of days. Courses

[24]For example, this could happen when Hispanophone countries experience higher unemployment rates, or Anglophone countries introducing stricter immigration laws

[25]Section 4.4 discusses the challenges multilateral resistance poses to my estimates and how I deal with it.

are thus plausibly developed based on the current demand for language learning across a language pair, rather than on the growth rate. If Duolingo nevertheless aimed to anticipate future trends in the demand for language learning driven by migration trends, it is unlikely to happen for migration aspirations, which is the first step in the migration process.

To alleviate remaining concerns about parallel trends violations, I perform several diagnostic and robustness exercises. First, I study pre-trends in migration and language learning outcomes between strongly exposed and other country dyads before the strongly exposed received treatment. Using event study estimators that are robust to the issue of negative weights in staggered effects regressions de Chaisemartin and d'Haultfoeuille (2020); Nagengast and Yotov (2023) (as further discussed in section 4.4), I find no evidence of differential pre-trends in migration intentions and flows. Moreover, there are no pre-trends in levels of the following language-related outcomes: online search interest in languages, language instruction among language skills among English test takers, and language skills conditional on migration in the EU. Second, courses may be developed to tailor to the origin and destination countries with the most speakers by language. To exclude that my main results are driven by these countries, I set the contribution for language-country pairs to zero if the country is the one where the language is spoken by the most people globally. For example, if a course is developed between French and English, it could be driven by anticipation of trends for individuals in France. However, also all other French-speaking countries to English speaking are affected by the Duolingo course. This exercises removes the former variation, but keeps the latter. I perform this exercise at the source language-origin and the target language-destination pair, as well as both at the same time. Third and last, I construct two Instrumental Variables approaches, one based on the notion that Duolingo courses are rolled out between languages with many speakers and the notion that course development is easier if there are more courses using the same source- or target languages. I re-estimate the model of equation 9 using a control function approach (Wooldridge, 2015). The results are robust to these two exercises.

4.3.1 What predicts course development?

To study what determines course rollout and hence Duolingo exposure, I analyze the timing of rollout on the language-pair and country-pair level. For the language-level analysis, I regress (1) a binary indicator of whether a course has been rolled out by the end of my sample period, and conditional on rollout, (2) the year of rollout on (bilateral) characteristics of languages. Second, I perform a similar analysis using the country pair-level foreign Duolingo exposure country pair level and the year the Duolingo exceeds 0.5. Table 1 shows the main results. I find unsurprisingly that courses are rolled out between larger languages, and courses to languages with less target speakers are rolled out later. Moreover, the probability a module is rolled out increases strongly if both languages have many speakers. Turning to the country-level analysis, I find that the Duolingo exposure in 2023 is almost 20 percentage points smaller for countries sharing a language and increasing in GDP per capita of both the origin and target languages, suggesting that courses are rolled out to languages spoken by rich countries. After inclusion of origin and destination fixed effects, distance between countries is related to a lower Duolingo exposure. Importantly, the log of the bilateral stock of migrants does not explain Duolingo exposure in 2023, nor does it explain rollout timing once unobservable country-level characteristics are accounted for. However, this analysis does not exclude dynamic selection into treatment. I assess this by estimating pre-trends in various outcomes in subsequent sections.

Table 1: The Determinants of the Rollout of Duolingo Courses

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$Duolingo^{ST}_{2022}$		Year of rollout		$DL_{od,2023}$		Year of large rollout	
Log source speakers	0.004**		-0.051					
	(0.002)		(0.265)					
Log target speakers	0.003**		-0.447**					
	(0.002)		(0.175)					
Log source × Log target speakers		0.002**		0.460				
		(0.001)		(.)				
Sharing an official language					-0.191***	-0.202***	1.848***	0.410
					(0.027)	(0.028)	(0.419)	(0.330)
Log population-weighted distance					0.054***	-0.021***	-0.031	0.048
					(0.009)	(0.006)	(0.215)	(0.046)
Log GDP pc PPP in origin					0.036***		0.239**	
					(0.008)		(0.095)	
Log GDP pc PPP in destination					0.031***		0.373*	
					(0.009)		(0.195)	
Log bilateral migrant stock + 1 (2005)					0.010***	-0.000	-0.213***	-0.004
					(0.002)	(0.002)	(0.048)	(0.012)
Observations	13225	13225	84	52	22217	22217	4098	4098
Source and Target FE		✓		✓				
Origin and Destination FE						✓		✓

Notes: OLS regressions of language-pair and country-pair level exposure and year of rollout on language- and country characteristics. I include all N source and target languages Column 1 and 2 regress a binary module for the presence of a module (by the end of 2023) on the log of source- and target language speakers and its interaction, where column 2 adds fixed effects. Columns 3 and 4 regress the year of rollout on the same characteristics, where column 4 add fixed effects, dropping all source and target languages appearing only once. Similarly, column 5 and 6 regress the measure of Duolingo exposure in 2023 on country and dyadic characteristics, and column 6 adds origin- and destination fixed effects. Column 7 regress the year in which exposure to Duolingo first exceeds 0.5 on country and dyadic characteristics, and column 8 adds origin- and destination fixed effects. Hence, columns 7 and 8 only include dyads for which exposure exceeds 0.5 by the end of 2023. Columns 5-8 also control for log of origin country population, log of destination country population, a dummy for countries sharing a border, linguistic proximity between both country's main languages (Adserà and Pytliková, 2015), log of trade value in 2005, origin country in EU, destination country in EU and both countries in the EU. Data on speakers by language is obtained using Ginsburgh, Melitz and Toubal (2017) and World Bank population data, data on country characteristics, except linguistic proximity, is obtained from CEPII. Linguistic proximity is obtained from Adserà and Pytliková (2015). Standard errors reported in parentheses are two-way clustered: on the source language and target language level (1-4) or on the country of origin and country of destination level (5-8). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.4 Estimation

To estimate treatment effects from multiplicative models of dyadic trade or migration flows, scholars typically estimate gravity models by Pseudo-Poisson Maximum Likelihood (PPML) rather than log-like transformed outcomes for two main reasons.[26] First, the presence of zero flows make the estimates effect size dependent on the unit in which the outcome is stated when there is an effect on the extensive margin (Chen and Roth, 2024). Second, heteroskedasticity can introduce bias in OLS estimates of non-linear models (Silva and Tenreyro, 2006).[27] Moreover, non-linear models are biased due to the incidental parameter problem (IPP). The (panel) Poisson estimator is the only non-linear estimator that does not face this issue. However, this is not the case for three-way fixed effects Poisson estimation. Weidner and Zylkin (2021) develop a correction to the prevailing IPP bias. Although this bias is limited in most cases, I report bias-corrected estimates.

Negative weights. A wealth of recent literature has shown that two-way fixed effects regressions of staggered treatments do not identify an estimand of positively-weighted treatment effects (Goodman-Bacon, 2021; de Chaisemartin and d'Haultfoeuille, 2020; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2021; Borusyak, Jaravel and Spiess, 2024; Callaway and Sant'Anna, 2021; Wooldridge, 2021). In extreme cases, the negative weights on some treatment effects may lead the researcher to find results that take opposite signs. Alternative estimators have been proposed for the staggered setting with binary absorbing treatment, as well as with staggered adoption of multi-valued and continuous treatments Callaway, Goodman-Bacon and Sant'Anna (2021) as well as fully continuously distributed treatments (de Chaisemartin et al., 2024). Wooldridge (2023) pointed out that the same issue arises in non-linear models and Strezhnev (2023) and Nagengast and Yotov (2023) have shown that this problem naturally extends to the three-way fixed effects setting. In this paper, I estimate event study specifications using the regression-based estimator developed by Wooldridge (2023) and implemented by (Yotov, Nagengast and Rios-Avila, 2024). This estimator prevents "forbidden comparisons" by estimating treatment effects heterogeneity by separately estimating treatment effects by cohort and time period. These treatment effects can be aggregated by event time, enabling the study of pre-trends and dynamic effects. A major advantage of this estimator is that it can be implemented for non-linear models and that it flexibly allows for the inclusion of additional fixed effects. A limitation of this approach is that its usage is limited to binary and absorbing treatments. For implementation, I study large increases in Duolingo exposure. As a large share of variation is driven by a single course rollout per dyad, this allows me to estimate event studies using most identifying variation.

[26] I borrow the term log-like from Chen and Roth (2024), who define it as "functions $m(y)$ that are well-defined at zero but behave like $log(y)$ for large values of y, in the sense that $m(y)/log(y) \to 1$". This includes the often-used $log(y + 1)$ and inverse hyperbolic sine $log(y + \sqrt{1 + y^2})$ transformations.

[27] An often overlooked but important difference between the Poisson and OLS of a log-transformed outcome model is that the estimand it targets is different, as noted by Tyazhelnikov and Zhou (2021) and Chen and Roth (2024). Chen and Roth (2024) calls $e^{\hat{\beta}} - 1$ the average proportional treatment effect on the treated. The Poisson estimator targets the relative average effect, whereas the log-like estimator targets the average relative effect. Hence, a given treatment effect has the same effect on the Poisson estimate when it happens to units with low- or high levels of untreated potential outcomes.

Multilateral resistance to migration. An additional estimation concern for studying migration-related outcomes arises from the strong assumptions underlying the random utility of migration, including that the discrete choice problem fulfills the Independence of Irrelevant Alternatives (IIA) assumption. However, in reality, individual-level preference shocks for different destinations are not independent. Bertoli and Fernández-Huertas Moraga (2013) show that this generates additional terms in the error term, giving rise to an endogeneity problem in equation 9 and could bias estimates if the independent variable of interest is correlated across destinations. In limiting cases, this term does not vary by dyad over time. One such case, as Ortega and Peri (2013) and Bertoli and Fernández-Huertas Moraga (2013) discuss, assumes that correlations of the EVT-1 shock only happens within two nests: one for the origin country, and one for all foreign destinations. The resulting multilateral resistance term only varies on the origin country by year, which can be accounted for by the inclusion of origin-time fixed effects.

More generally, for a given origin country, the correlation structure of the EVT-shock across destinations may be complicated, giving rise to multilateral resistance terms varying at the origin-destination-year level. Relevant to my setting, destination-specific shocks may be correlated across destinations sharing languages. To see how this can effect my estimates, consider the following example. Duolingo availability from a given language to Spanish increases the attractiveness of all Spanish speaking destinations. However, the Spanish speaking countries are also close substitutes and preference draws for these destinations are correlated. Hence, the effect size is underestimated because the competing effect all these destinations exert on each other makes the migration rate increase less than if only one destination would have been treated in isolation. This would generate a bias towards zero in my estimates. As origin-time fixed effects may not account for all resistance terms, an additional remedy is to include origin-time-destination nest fixed effects, where nests are chosen based on sharing relevant observable characteristics of destinations that determine the attractiveness to migrants (Bertoli and Fernández-Huertas Moraga, 2013; Beine, Bierlaire and Docquier, 2021). I address this in section 6.2.3.

Inference. Standard heteroskedasticity-robust standard errors may overestimate the precision of regression estimates because of correlation of the error term across observations. Conventional knowledge is to cluster at the level of treatment assignment (Abadie et al., 2023). In dyadic data, however, all flows from o to d can be correlated to all dyads where either o or d are represented. In practice, clustering two-way at both the sending and receiving unit gives approximately correct standard errors Cameron and Miller (2014). However, geographically, economically or culturally close countries often speak the same language or languages from the same language family. Due to this, observations across origin and destination countries sharing languages may not be fully independent. As the treatment is correlated across origins sharing languages to a given destination and vice versa, this could lead to an underestimate of the variance of the estimates. To account for this, I consider alternative two-way clustered standard errors: on the main spoken language in the origin country and on the main spoken language in the destination country.

5 Language Learning

In this section, I study the impact of Duolingo course rollout on language learning and learning outcomes. First, I study the determinants of take-up of Duolingo courses and its effect on traditional language learning. Second, I examine to what extent the introduction of a course induces online search behavior in Duolingo and towards the target language of the rolled out course by origin country. Third, I study whether the language skills of English test takers is impacted by Duolingo modules to English. Fourth and last, I study how adult and in-school language learning intensity is affected by Duolingo course availability.

5.1 Language learning

To study the determinants of take-up, I rely on the number of learners by language course, which I obtained from the Duolingo website in October 2022. Because of users quitting Duolingo, the total number of ever learners likely exceeds the number of learners at a given point in time. Figure 8 and 9 show the total number of learners by language and as a share of the number of speakers by language, aggregated by source and target language. Figure 8 shows that English is by far the most used source language as well as the most learnt target language. The former suggests that English is also used by many non-native English speakers to learn third languages. The latter reflects that returns to English proficiency are high due to its status as a *lingua franca*. Even among other widely spoken languages, such as Spanish and French, the number of learners exceeds 10% of the number of speakers. Moreover, several more widely spoken languages have garnered a relatively large number of speakers likely for tourism and cultural reasons, such as Greek, Italian, Japanese, and Korean. Table 1 shows the results from a regression of the number of users by course on the number of speakers of the source and target languages. and its interaction. The results show that the number of users increases by about 8 for every 1000 source language speakers as well as for every 1000 target language speakers.[28] This shows that not only the total pool of potential learners is relevant in the decision to learn a language, but also the extent of applicability of the target language. The positive interaction effect between the number of source and target language speakers in column 2 and 3 also suggest that demand for courses is particularly high between languages with many speakers, such as English and Spanish.

[28]This is the slope of users to speakers. The percentage of learners of total source language speakers is about 2%. As individual can speak multiple languages, this number is lower than the total share of world population that have used Duolingo.

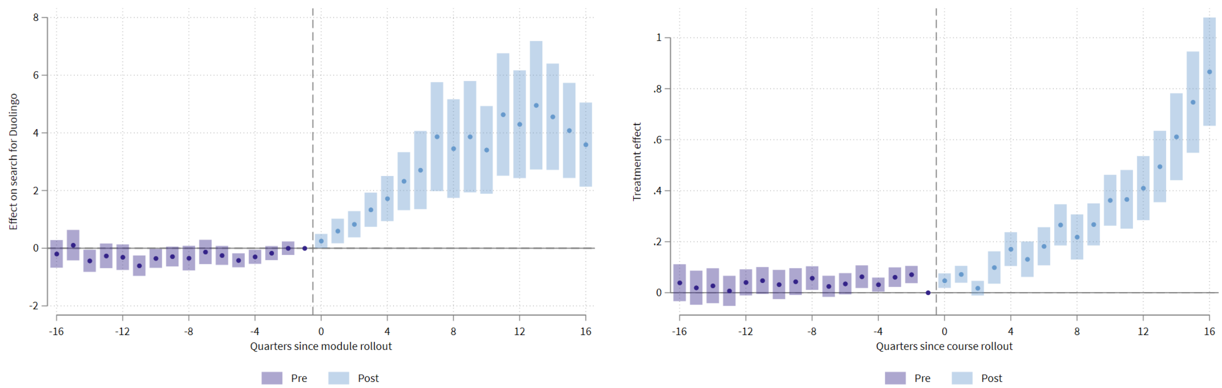
5.2 Interest in Duolingo and available languages

Online search behavior is a useful proxy of the interest in a topic. To validate that the course roll-out dates and to study how courses affect demand for language-related information, I turn to online search interest for Duolingo and available languages. I collect information about relative search intensity on the widely used search engine *Google* through the Google Trends API. This API enables one to query time series of normalized search intensity of a search term or topic over time, referred to as the Google Trends Index (GTI). To ensure that the GTI represent the same relative search intensities across countries, I always scale the GTI across regions. For a detailed description about how consistent time series are constructed, Google Trends is queried for search terms and topics, see C.

Interest in Duolingo. As Duolingo was founded in 2011, global search interest went from practically zero to large values (see Figure 1a). To study the effect of course rollout on the search interest for Duolingo, I study the scaled GTI four years before and after first increases of 50 percentage points or larger in foreign Duolingo exposure on the origin country-level.

Figure 2a shows the development of interest in Duolingo four years before and after the introduction of the first relevant Duolingo course on the country of origin level, controlling for country of origin and year fixed effects. After course introduction the country-specific interest starts increasing gradually, which is in line with the increasing popularity of Duolingo over time documented in Figure 1. As in some cases Duolingo was available as a beta version before the course was released, there was some interest several quarters before courses became available. This could also be driven by foreign language speakers (migrants or visitors) that search for Duolingo as a course may have already been available to them.

Figure 2: The effect of influential course introductions on interest in Duolingo and Languages



(a) Duolingo

(b) Languages

Notes: Results from Wooldridge (2023) event study estimators around increases in Duolingo exposure. (A) shows estimates of the effect of increases in Duolingo exposure on the relative search interest in Duolingo and its transliterations across countries proxied be the scaled Google Trends Index. A country is counted as treated if the share of inhabitants speaking a source language with a Duolingo course is at least 50 percentage points. The two-way fixed effect counterpart of the Nagengast and Yotov (2023) event study corresponds to a regression of (a) the GTI in Duolingo on origin country and quarter fixed effects and an aggregation of the Duolingo exposure: $GTI_{ot}^{Duolingo} = \beta \mathbb{1}\left((\max_{S,T} \alpha_o^S DL_t^{ST}) > 0.5\right) + \psi_o + \phi_t + \epsilon_{ot}$. N = 8,058 from 158 countries, of which 112 treated. Shaded blue bars indicate 95% confidence intervals based on cluster-robust standard errors at the country level. (B) shows estimates of the effect of increases in Duolingo exposure on relative search interest in available target languages proxied be the Google Trends Index scaled across countries but not languages. The regression counterpart is a three-way fixed of the GTI on origin-language, origin-quarter and target-quarter fixed effects: $GTI_{ot}^{\tilde{T}} = \beta \mathbb{1}\left((\max_S \alpha_o^S DL_t^{ST}) > 0.5\right) + \psi_{oT} + \phi_{tot} + \theta_{Tt} + \epsilon_{oTt}$. In the latter case, I restrict the time period to 2011-2022, as Google Trends Indices for less frequently spoken languages are noisy due to the limited search interest. N = 545,220, 15,145 pairs of which 1,279 treated from 233 origins and 65 queried target languages. Shaded blue bars indicate 95% confidence intervals based on two-way cluster-robust standard errors at the country and language level. Data obtained by repeatedly querying Google Trends. For a further discussion on the construction of the Google Trends Indices, see section C.

Interest in target languages. As Duolingo courses enable to learn a particular language, it could also spur the interest in the specific language or increase the intensity of learning. Relative search interest in languages was stable since 2009, but started increasing in 2016 (see Figure 1b). As for Duolingo, I study the interest in languages four years before and after course rollout. Contrary to the previous section, as I obtain dyadic search interest from origin countries to target languages, I can partial out all origin-time and source language-time variation with fixed effects. I study the scaled GTI around first increases of 0.5 or larger in dyadic Duolingo exposure on the origin country-source language level.

Figure 2b shows the event study results around the introduction of a salient course on the *bilateral* interest between the origin country and the target language. I find that prior to course roll-out, interest in languages is not trending before course rollout and that interest in the language slowly starts increasing after the introduction of the course and increases steadily thereafter. The continuing gradual increase over time reflects that Duolingo has become considerable more popular over time, but also reflect that course availability increases online search to further study languages by e.g. searching for translations or further study material. These two exercises also validate that the rollout dates, based on the date courses enter the final phase, reasonably capture the relevant timing of course introduction.

5.3 Language skills

The previous section has documented the take-up of Duolingo courses, and that course roll-out spurred interest in Duolingo and available target languages. A pressing question is whether access to low-cost language learning also improved language skills among the general population, as predicted by the model in section 4.1. However, internationally comparable data on foreign language skills among the general population is scarce. In the following, I study whether foreign language skills in the EU have increased. To study the effect on likely prospective migrants and specific language skills, I consider origin-specific test scores from widely-used language proficiency tests. To study how low-cost language learning affect other modes of learning, I examine the effect on the number of language learners in schools in the EU and German language-learning institutes globally.

5.3.1 Language skills across Europe

The EU Adult Education (AES) survey is an irregularly fielded representative survey of skills and training among the adult population of the European Union. I use responses from 677,876 responses from four cross-sections of the AES: 2005-2008, 2011-2012, 2016-2018, and 2022-2023. The AES collects information on up to two mother tongues, up to seven spoken languages and levels of proficiency for the two best spoken foreign languages. [29]

I construct a dataset where the unit of observation is the respondent-by-target language, for each of the 30 target languages available on Duolingo. I exclude other languages, as there is no variation in the treatment variable. I consider an observation treated if a course from either of the mother tongues to the target language has become available in the year before the interview or earlier. 9% of all respondents indicate having a second mother tongue. The main outcome variables are binary indicators for whether a respondent indicates to speak a language, or whether it speak the language at an intermediate or advanced level. I omit observations for which the target language is a native language for the country of interview, as

[29] The question on mother tongues is worded as: *Which language(s) is (are) your mother tongue(s)? (Mother tongue refers to language spoken at home as a child.* The question on languages spoken is: *Which other languages do you understand or speak, except your mother tongue(s).* The question on language skills among the two best spoken languages is: *How would you rate your skills in ..... ? Which of the following statements fits best?*, with the following three answer options: ***Basic:*** *I can understand and use the most common everyday expressions. I use the language in relation to familiar things and situations.* ***Intermediate:*** *I can understand the essential of clear language and produce simple text. I can describe experiences and events and communicate fairly fluently.* ***Advanced:*** *I can understand a wide range of demanding texts and use the language flexibly. I master the language most completely.*

in some survey waves in some countries, individuals could additionally include their own native languages among the spoken languages. 1 shows descriptive statistics of the estimation sample. Individuals speak on average 1 non-native language, 0.51 of which are spoken at at least intermediate level and 0.23 spoken at advanced level. The number of languages spoken is similar for natives and immigrants. On average individuals have access to Duolingo courses to 4.63 out of 30 target languages in 2023.

To estimate the effect of the availability of low-cost language learning on language skills, I estimate the following three-way fixed effects model:

$$y_{iT} = \beta DL^{m}_{Tt} + \gamma X_i + \phi_{m_1 T} + \psi_{m_1 t} + \phi_{Tt} + \epsilon_{iT} \tag{10}$$

y_{iT} is a binary indicator for whether individual i interviewed at t speaks language T. DL^{m}_{Tt} is a binary indicator for whether a Duolingo course is available using one of respondent's mother tongues m (either the first m_1 or second m_2 mother tongue) and target language T at time t. X_i includes age, age squared, a dummy for being male, a dummy for being born abroad and dummies for secondary and tertiary educational attainment. $\phi_{m_1 T}$, $\psi_{m_1 t}$, and ϕ_{Tt} represent first mother tongue-target language pair, first mother tongue-year and target language-year fixed effects. The first capture the time-invariant propensity among speakers of every mother tongue to speak a given target language, the second capture the time-varying propensity to speak foreign languages by first mother tongue and the third captures the time-varying propensity to speak as a specific foreign languages for every mother tongue. I weight the regressions using demographic weights provided by EU AES to make the results representative of the EU population. To account for arbitrary correlation in the error term across language pairs, I cluster standard errors on the first mother tongue-by-target language pair. If Duolingo rolled out courses in anticipation of increased demand for learning across a language corridor, this could be potentially violate the main assumption of parallel trends. To examine its plausibility, we study pre-trends prior to rollout.

Results Table 2 shows the main results. The availability of low-cost language learning increases the probability to speak a language by 1.5 percentage points. This effect is large, given the low average share of people speaking is just 3.5% across the 30 target languages. The effects on higher levels of language skills are lower and the effect on advanced language skills is small and insignificant. As the treatment is staggered, we can examine the dynamic effect of course rollout in an event study. Figure 3 shows the results. I find that there are no pre-trends in language skills before, which increases confidence in the parallel trends assumption. As countries are often surveyed in only one year in a given AES wave, the panel is unbalanced and hence the placebo and dynamic effect estimates may suffer from compositional changes if treatment effects are heterogeneous. Nevertheless, the results suggest that the effects appeared already in the first year after rollout.

These results are robust to a set of robustness tests reported in Table 2. Importantly, the results are similar when additionally including fixed effects for the second mother tongue-year and second mother tongue-target language pair, respondent fixed effects and country-cohort-target language fixed effects. The first ensures that the effects are not driven by changing compositions of second mother tongue speakers, the second controls for all observed unobserved individual-level compositional differences across survey waves and the third ensures that the found effects happen within birth cohorts. The latter estimate is only slightly smaller if including these fixed effects, suggesting that the effects are driven by increases in target-specific language skills among individuals born in the same over time.

Table 3 shows that the effects are equally strong if we construct our measure based on only the first mother tongue, and that effects are somewhat smaller, but still present if we base our measure also on other languages spoken at intermediate level. As the average number of target languages with Duolingo courses available in 2023 is 4.93, the rollout of Duolingo led to an average increase of 0.07 languages spoken per individual and 0.03 languages spoken at least at intermediate level.

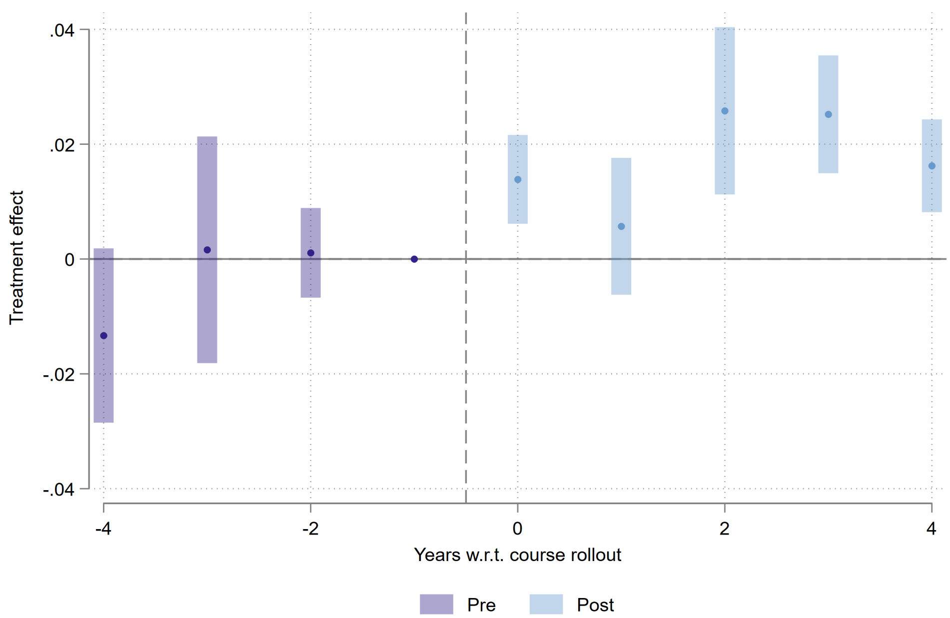
Table 4 reports heterogeneity among individual characteristics. The effect is strongest for young people, but still present among older individuals, and slightly stronger for women than for men. Remarkably, the effect does not depend on educational attainment. The effects are considerably stronger among immigrants than natives. Table 5 shows that the results are particularly strong for European languages, suggesting that skill acquisition is particularly strong for languages spoken in nearby countries. Unsurprisingly, the effect is strong for English particularly among immigrants, but is also present for other languages. Table 6 shows that the effect for immigrants is particularly strong for low proficiency levels of host country languages, but remains present for immigrants for other languages. These results suggest that the introduction of online language courses increase the probability to speak learnable languages. I explore the effect of course availability on immigrants' language skills upon and after arrival in more detail in section 7.3.

Table 2: The Effect of Duolingo Courses on Language Skills across the EU

	(1) Spoken	(2) Intermediate	(3) Advanced
DL^m_{Tt}	0.015***	0.009***	0.001
	(0.004)	(0.002)	(0.002)
Observations	21,553,104	21,553,104	21,553,104
R^2	0.388	0.259	0.129
Average dependent variable	0.036	0.018	0.008
Three-way fixed effects	✓	✓	✓

Notes: OLS estimates of language proficiency on a Duolingo course availability, controls and three-way fixed effects. The dependent variable is a binary indicator for whether individual i with mother tongues m speaks language T or speaks language T at at least intermediate or advanced level. Controls include age, age squared, a dummy for being male, a dummy for being born abroad and dummies for secondary and tertiary educational attainment. Data on language skills originates from the EU Adult Education Survey and the rollout dates of Duolingo courses. All estimations are weighted using weights provided by EU AES to render the results representative of the EU populations. Regressions are weighted using demographic weights provided in the EU AES. Standard errors reported in parentheses are clustered on the language pair level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 3: The Dynamic Effect of Duolingo Rollout on Language Skills in the EU



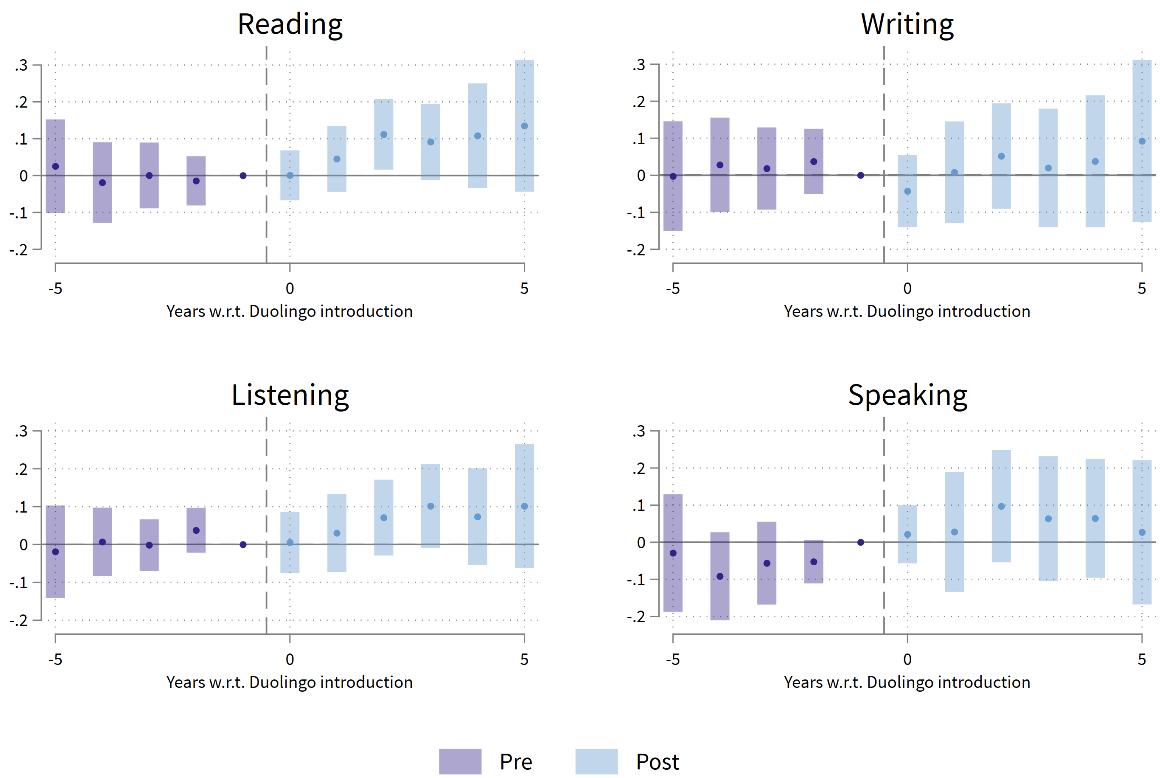
Notes: Results from linear Wooldridge (2023) event study estimators around introduction of a Duolingo course. The specification follows that of column 1 of Table 2, but estimates every treatment effect separately by treatment cohort and time period. Shaded blue bars indicate 95% confidence intervals based on standard errors clustered at the first mother tongue-target language level. For details on the data, see notes to Table 2.

5.3.2 Language tests

TOEFL. The Test of English as a Foreign Language (TOEFL) test organized by ETS, which is an English language test taken annually by more than 2 million individuals across the world, mostly for (foreign) university enrollment. Hence, the base of test takers likely includes many prospective migrants. The TOEFL test is scored on a scale of 0 to 120 points, where each of the four sections (Reading, Listening, Speaking, and Writing) receives a score from 0 to 30. However, test participants may be differentially selected after introduction of low cost language technologies. For example, individuals without prior access to language learning may take up low-cost language learning, but still have worse skills at the time of the test than the rest of the population. Moreover, individuals could defer the test to improve their test scores as they can continue learning. The latter concern is somewhat alleviated as many test takers take it at the end of high school or a degree program, which leaves little margin to defer the test. Furthermore, differential selection is unlikely to affect the relative performance across elements. In particular, as Duolingo predominantly enables the learning of passive language skills, I can test whether scores on passive (reading and listening) rather than active elements (writing and speaking), providing a partial test of the influence of low-cost language learning on language skills.

ETS compiles yearly reports of average section scores by test takers' native language. Figure 4 shows the results of event studies of test scores around the introduction of 22 Duolingo courses to English. In the first years before course rollout there are no discernible pretrends. In the first five years after course rollout, scores have increased by 0.12 s.d. for reading and 0.08 s.d. for listening, but no effects for writing and speaking. Altogether, the results suggests that passive skills have improved considerably, without an effect on active skills. Based on the nature of Duolingo vis-a-vis in-person language courses with more active components, this is not surprising. This exercise does not identify the effect of Duolingo on the average population for two important reasons. First, TOEFL takers are much more likely to have used Duolingo. Second, Duolingo availability may affect selection into taking the test. As Duolingo enables the study of English language at beginner and intermediate levels, it seems unlikely only initally more proficient individuals take the test. Nevertheless, test scores have become more positive on average, suggesting this effect is not dominating. As Duolingo may be more useful to those with low levels of initial skills, I study whether the effects are stronger by native languages with low pre-Duolingo test scores. Figure 1 shows effects separately for the subsample of languages with below-median scores in 2010. Effects are considerable larger, and the results also hint at small positive effects for active language skills.

Figure 4: The Effect of Duolingo Rollout on Component Scores of English language (TOEFL) test (2007-2022)



Notes: Results from linear Wooldridge (2023) event study estimators around introduction of a Duolingo course to English. The panels report results for four different outcomes: standardized TOEFL scores reading (upper left), writing (upper right), listening (lower left) and speaking (lower right) by native language of the test takers. As the unit of observation is the native language level, the Duolingo exposure is binary. N = 1,808 from 121 languages, of which 22 are treated. Shaded blue bars indicate 95% confidence intervals based on standard errors clustered at native language level. Data is obtained from the yearly TOEFL iBT Test and Score Data Summary between 2007 and 2022.

GRE. To further address the potential risk posed by differential selection into test taking, I turn to the GRE test. The GRE is a English-languages general ability test taken by approximately 300,000 individuals every year. Most test takers use the GRE test for university admission in a foreign country (not necessarily English-speaking). Contrary to the TOEFL test, for the GRE the number of GRE test takers are reported, which enables testing whether Duolingo availability has changed the size of the pool of test takers. Moreover, the GRE includes separate linguistic and quantitative elements, which allows us to test whether selection in terms of general skills has changed. As Duolingo exposure should not have a causal effect on quantitative skills, this provides a test of changes in selection of test takers. As the GRE scores are published by origin country, I construct a measure of exposure on the origin country-level for Duolingo courses to English. Figure

2 reports results from an event study specification using the Wooldridge (2023) estimator. The number of test takers has not changed in a statistically significant way. Verbal and Analytical writing skills, however, have increased by up to 0.2 S.D. after three years. Reassuringly, I find null effects on quantitative scores, which improve confidence that the availability of Duolingo did not affect selection into the GRE based on other skills.

5.3.3 Traditional Language Learning

It is a priori unclear how low-cost language learning interacts with traditional (in-class) language learning. On the one hand, potential learners may use Duolingo instead of traditional in-class language courses. On the other hand, Duolingo may spur language learning at basic levels and generate interest in destination language countries and culture, which could increase in-class learning, particularly at higher proficiency levels. This could increase the total number of learners, and increase average proficiency, as suggested by the model of section A. I test which effect prevails for two different types of language learning: adult language learning in German language learning institutes and in-school foreign language instruction across the EU.

Goethe institutes offer German language learning in more than 90 countries worldwide. These institutes both offer languages courses, as well as exams providing generally accepted certification. In D.3 I present suggestive evidence that the rollout of 9 Duolingo courses to German decreased course participation in traditional German language courses, but had no effect or even a slight positive effect on the number of exams taken. This is in line with online language learning substituting for costly learning, but not for certification, which may be needed for visa or employment.

In most countries in the Eurppean Union, schools instruct pupils in one or more foreign languages. Data on the share of pupils learning any given foreign language by education level is available through *Eurostat*. In D.4, I study how Duolingo course availability has impacted in-school language learning. Reassuringly, before Duolingo course availability, there are no discernible pre-trends in the share of students learning available foreign languages. Moreover, I find that across school levels, low-cost language learning increased the share of students by 1–2 percentage points up to five years after course introduction. The effect size is increasing over time, which is in line with gradual adoption of Duolingo courses over time. Excluding the most learnt foreign language across the EU, English, the effects are still positive but smaller in magnitude, and marginally significant in most cases. These results also imply that low-cost language learning likely increases the total online and offline learning intensity of available languages, improving foreign language skills.

6 Migration Aspirations and Flows

In this section, I study whether the staggered introduction of low-cost language learning has impacted migration patterns in accordance with the model predictions of section 4.1. As comprehensive bilateral migration flow data is only available on a yearly level for a limited number of countries, this section will mostly rely on migration intentions as elicited in the Gallup World Poll (GWP). To complement evidence the with evidence on migration flows, I additionally study migration flows to OECD countries and global scholarly migration flows.

6.1 Data

I use the 2007–2022 vintages of the Gallup World Poll (GWP), which is a representative survey of about 1,000 individuals per year in more than 150 countries. Besides many questions concerning demographic, economic and social issues, it includes a question on whether one would like to emigrate if one had the opportunity, as in Adema, Aksoy and Poutvaara (2022). The question's wording is *Ideally, if you had the opportunity, would you like to move permanently to another country, or would you prefer to continue living in this country?*, and the answer options are **yes**,**no**, **don't know** and **prefer not to answer**.[30] In the latter two cases I discard the observations. If individuals mention that they desire to emigrate, they are also asked where they would like to migrate.[31] Importantly, bilateral migration intentions as elicited in the GWP are strongly correlated to migration flows across migration corridors and are predictive of subsequent migration flows (Tjaden, Auer and Laczko, 2019). Based on the answers to these questions, I construct the stock of people in country o aspiring to emigrate to country d from the origin country's population at t, the total number of GWP respondents N_{ot} and the share of respondents aspiring to emigrate from o to d, N_{odt}: $M_{odt} = pop_{ot} N_{odt}/N_{ot}$.[32] To construct the odds ratio of migrating to d over staying in o, I calculate $\frac{M_{odt}}{M_{oot}} = N_{odt}/(N_{ot} - \sum_{d'} N_{odt})$. I weight this procedure using the individual-level sampling weights provided by Gallup. The GWP comprises 2,003,036 interviews from 166 jurisdictions of whom 22.8% desire to emigrate.[33] Jurisdictions are visited on average 11.5 times across the 16-year period. 5% of those aspiring to migrate do not indicate a preferred country or indicate a region (e.g. "African country") or jurisdictions without information on languages spoken. I regard those individuals as stayers, to ensure that the sum of

[30] Figure 5 shows that the share of world population that aspires to migrate has increased since 2010. During the Covid-19 pandemic migration aspirations by 3–4 percentage points, but reverted to previous trends soon after. Adema, Aksoy and Poutvaara (2022) find that at least 2–3 percentage points of this increase can be attributed to the rollout of mobile internet networks.

[31] The question's working is: *To which country would you like to move?* to which respondents can give an answer which is codified to a country by the interviewer if possible.

[32] I obtain yearly data on population pop_{ot} from the World Bank.

[33] Several jurisdictions in the GWP are absent in the dataset of Ginsburgh, Melitz and Toubal (2017) nor is there reliable information on languages spoken from the CIA World Factbook. These are omitted from analysis.

migrants and stayers used for calculation in the equals unity.

I complement the data on migration intentions with actual migrant flow data from two different sources. The OECD International Migration database records yearly bilateral migration from virtually all countries in the world to 37 OECD countries. This data consists of collected national statistics about inflows of migrants, in some countries administered by nationality, in others by country of birth. I construct the odds of migration by using the origin-country population from the World Bank. I focus on the time period from 2007 to 2019, due to the large influence of the Covid-19 pandemic on international mobility barriers. A particularly mobile population are academic scholars. Akbaritabar, Theile and Zagheni (2024) constructed a dataset of yearly scholar migration between all countries in the world over time from the OpenAlex (version 2024_v1) database of published scholarly articles (Priem, Piwowar and Orr, 2022). In this dataset, a scholar is counted as a migration from o to d in t if her main affiliation on a paper published in t is an institute in d, and the last available main affiliation prior is from an institution in o. I construct the odds of scholarly migration using the total number of scholars publishing in o at t, also provided by Akbaritabar, Theile and Zagheni (2024). I use the 2007 – 2019 data to study the effect of exposure to Duolingo on scholar migration.

To complete the dataset with additional information on country- and country pair-level, I use the database by Conte, Cotterlaz and Mayer (2022). This dataset includes all important variables to estimate gravity models up to and including 2022: trade flows, trade agreements, geographical distances, macroeconomic indicators, from a variety of original sources.

6.2 Aspirations

Three-way Fixed Effects. Table 3 shows the main estimation results of the model in Equation 9. As hypothesized by the model, foreign exposure to Duolingo increases migration intentions strongly across specifications. Inclusion of origin-year fixed effects increases the point estimate, which could be driven by the downward bias exercised by multilateral resistance as discussed in section 4.4. In line with the model, domestic exposure to Duolingo is negatively correlated to the log odds of migration, although the estimates are on the brink of significance. The inclusion of controls on the pair level over time barely influences the estimates. Because of the unavailability of data for one origin and eight destinations, I subsequently discuss results without the control variables.

One unit of foreign Duolingo exposure, corresponding to a language course enabling communication between everyone in the origin country to everyone in the destination county, increases the log odds of migration by 45% (column 2). In comparison, one unit of domestic Duolingo exposure *decreases* the log odds of migration by 25% (column 1). In line with the model, this suggests that the introduction of low-cost

Table 3: The Effect of Duolingo Courses on Bilateral Migration Aspirations (2007 – 2022)

	(1) $\frac{M_{odt}}{M_{oot}}$	(2) $\frac{M_{odt}}{M_{oot}}$	(3) $\frac{M_{odt}}{M_{oot}}$	(4) $\frac{M_{odt}}{M_{oot}}$
DL_{odt}	0.267***	0.374***	0.306***	0.352***
	(0.065)	(0.080)	(0.062)	(0.068)
DL_{oot}	-0.225		-0.269*	
	(0.160)		(0.140)	
Observations	123263	123263	98019	98019
Unique origin countries	153	153	152	152
Unique destination countries	196	196	188	188
Unique dyads	9439	9439	9439	9439
Origin-destination FE	✓	✓	✓	✓
Origin-year FE		✓		✓
Destination-year FE	✓	✓	✓	✓
Controls			✓	✓
Weidner-Zylkin correction		0.386***		0.360***
		(0.060)		(0.058)

Notes: PPML estimation of a Gravity model of migration without (odd columns) and with origin-year fixed effects (even columns). The dependent variable is the ratio of the total number of people desiring to emigrate from origin country o to destination country d in year t over the total number of people not desiring to emigrate from country o in year t. Trade controls include a dummy for joint EU membership, a dummy for a WTO trade agreement between two origin and destination country, as well as the log of trade flows from the origin to the destination country. Because data on trade flows is not available for all destinations, columns 3 and 4 have less observation than columns 1 and 2. The bottom row reports results from the bias correction of Weidner and Zylkin (2021) for three-way fixed effects Poisson regression. Data on migration intentions originates from the Gallup World Poll and the Duolingo exposure is constructed using data from (Ginsburgh, Melitz and Toubal, 2017) and the rollout dates of Duolingo courses. Standard errors reported in parentheses are two-way clustered: on the country of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

language learning courses to languages rewarded on foreign labor markets increase migration intentions, whereas the introduction of courses to languages that are rewarded on domestic labor markets decrease intentions. The net effect of the availability of low-cost language learning depends on the relative return to language skills at home and abroad.

One may be wondering whether courses with higher returns have stronger effects. To study this, I subdivide treatment intensity in four bins of exposure and include indicators for these bins in equation 9 instead of a continuous variable for both foreign and domestic exposure. Table 10 presents the results. I find that the effect for foreign exposure is close to zero for small levels of exposure, but increases monotonically with higher treatment intensity. The effect for domestic exposure is driven by those with a very large exposure. These are countries where two languages are spoken by a large proportion of the population. In such cases, knowledge of both languages may enable one to earn a large premium on the labor market.

Table 11 shows the main results using the Duolingo exposure calculated using official languages at the destination. The effects are statistically significant, but smaller. This could be driven by the strong effects for non-official but widely spoken languages among prospective migrants, such as English.

At first glance, the effect sizes may seem very large. However, one has to consider that these are relative average effects on the odds ratio of migration. About 20% of individuals in GWP desire to emigrate, and as there are about 200 alternative destination countries, the bilateral stock of aspiring migrants is about 0.1% on average. A 45% increase implies that the stock of aspiring migrants across a dyad receiving full treatment increases by about 0.045 percentage points. As average foreign Duolingo exposure is 0.26 in 2022 and there are 196 potential destinations in the main estimation sample, this suggest a persuasion rate of about 2.3% of the population.[34] This simple back-of-the-envelope calculation is an upper bound of the share of people that changed their migration intention due to Duolingo course availability, for two reasons. First, as most exposure intensities are considerably lower than unity, the average per-unit effect is closer to 37.4%. Second, it disregards multilateral resistance effects. For example, if a language course to English becomes available, this increases migration aspirations to all English-speaking destinations. However, it increases migration aspirations with less than the point estimate to a given English-speaking destination as other English-speaking countries exert a downward influence, as language-sharing destinations likely are close substitutes.[35]

Importantly, these results do not imply that total emigration intentions on the origin-level would increase by 45%. A positive point estimate in a gravity model could imply (a combination of) two effects. First, the availability of language learning may have shifted the preferred destination of aspiring migrants, or induced migration intentions among those who would not have desired to migrate in absence of a course. To study which effect prevails, I collapse the data at the origin-year level and regress the emigration rate on the domestic and average foreign Duolingo exposure and country and year fixed effects. Table 9 reports the results. Although statistically insignificant, the effects suggests that moving from no to full foreign Duolingo exposure increases the emigration rate by about 4 percentage points. Columns 3 and 4 report results using a migration stock-weighted measure of average foreign Duolingo exposure. Latter results are similarly sized, but statistically significant, indicating that courses in more attractive destinations increase the emigration intention rate. If the effect of 45% would be fully driven by newly aspiring migrants, this would explain 10 percentage points. Hence, less than half of the increase in migration aspiration odds can be explained by new attraction, and the rest by diversion effects.

Event study. To assess the plausibility of the parallel trends assumption by considering pre-trends, to study the dynamic effect and to alleviate the concerns that my results are driven by negative weights in

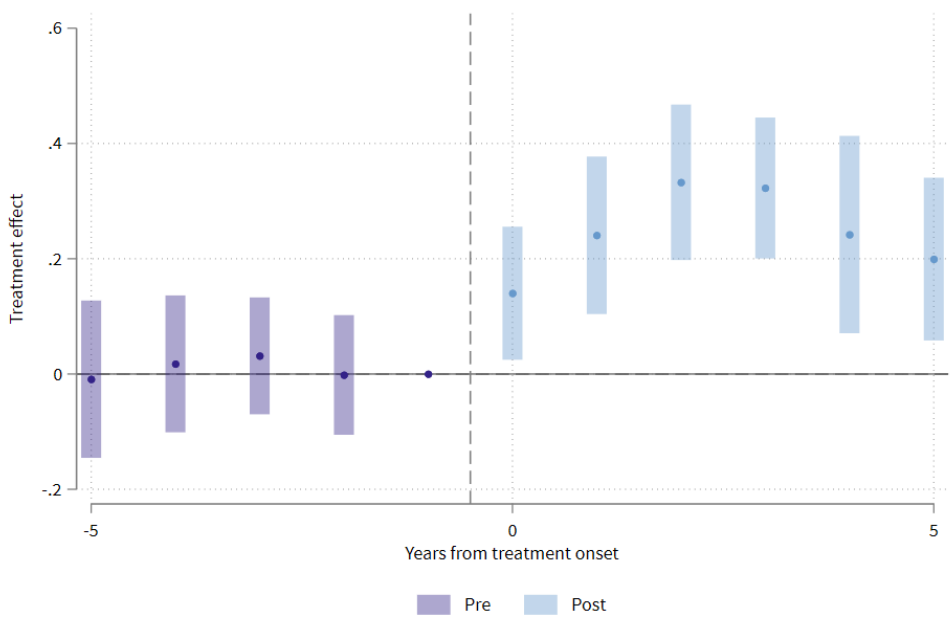
[34] This is a considerable share of the population that used Duolingo. By the end of 2022, more than 500 million people are registered on Duolingo (see e.g. https://blog.duolingo.com/2022-duolingo-language-report/), which is about 6.2% of world population. As the GWP does not survey the full world population, the propensity to take up Duolingo

[35] Based on gravity model estimates, one can simulate the effects of a language course on the distribution of migrants taking multilateral resistance effects into account. see e.g. Guichard and Machado (2024).

staggered difference-in-differences settings (Goodman-Bacon, 2021), I estimate an event study estimator for large increases in foreign Duolingo exposure across a dyad. As the Nagengast and Yotov (2023) estimator requires a binary treatment, I define an event as a dyad receiving an increase of 50 percentage points in foreign Duolingo exposure measure, using all other dyads as a control group. Figure 6 shows that the foreign Duolingo exposure increases strongly, and that pre-trends, driven by treatment and control units receiving small amounts of exposure, are negligible.

Figure 5 shows that the main event study results. I find no evidence of pre-trends before strong increases in exposure, increasing confidence in the conditional parallel trends assumption. I find a gradual increase in treatment effect in the first three years after treatment, which dampens thereafter.[36] The slow onset of treatment effects is in line with the gradual adoption of courses.

Figure 5: Event Study around Large Increases of DL_{odt} on Bilateral Migration Aspirations (2007-2022)



Notes: PPML regressions of the heterogeneity-robust event study estimator by Nagengast and Yotov (2023) of migration aspiration odds on a binary indicator for whether a country pair has experienced increase in foreign Duolingo exposure exceeding 50 percentage points, including origin-destination pair, origin-year and destination year fixed effects. An event is defined as an increase in Duolingo exposure of more than 50 percentage points. 6 shows that treated units on average experience a stable increase of about 60 percentage relative to the control group. Estimates are shown for the 5 years before and after an event. Shaded blue bars indicate 95% confidence intervals based on standard errors clustered at native language level. See notes to Table 3 for information on the data and sample.

[36] I also perform a similar event study using large increases in domestic Duolingo exposure, controlling for foreign Duolingo exposure. The results, presented in Figure 7, suggest that domestic exposure decreases migration aspirations gradually in the first three years after exposure.

6.2.1 The Role of English

The dominant role of English is clear from the model: the equilibrium level of language learning $l^T_{oS}*$ is an increasing function of the migration probability-weighted returns to the language skill. In the case of English, the domestic contribution to this is large across countries. As shown in Figure 8, this is reflected in learner numbers. Moreover, English takes a special role as a source language as many non-native speakers can use English as a language of instruction.

To study the special role of English, I split out the contribution of English and all other languages, calculating two exposure measures for both foreign and domestic exposure. I do this separately for English as a source language and a target language. Table 4 shows the results. Columns 1 and 2 show that the effect of foreign Duolingo exposure is significant for both English and other languages as source languages, but the former is considerably stronger. This could be driven by the fact that many English speakers are non-native speakers themselves, who are higher educated and at the same time have a larger propensity to migrate. Columns 3 and 4 shows that the effect is also strongest for English as a target language. However, it is considerable smaller and insignificant for other languages. This is in line with the model, as can be seen from equation 8: expected returns to English are considerably larger than for other languages, not in the last place because of returns on the domestic labor market. Hence, uptake of language learning is stronger for English, which may lead to a larger increase in bilateral migration, given the same level of foreign Duolingo exposure.

As the effect for foreign Duolingo exposure is driven by English as a target language, one could be concerned that the effect is driven by one or few English-speaking countries. In Table 15 I exclude five high-income native English-speaking destination country at a time and all at the same time. Although the point estimates are somewhat smaller when excluding the US and few destination countries, which is not the case.

The results for domestic language exposure in 4 show a similar pattern: the effects are strongly negative for English as a source- and target-language. This suggests that countries where English is spoken by many people, the opportunity to learn English, or to learn another widely used language using English, strongly decreases migration intentions.

Table 4: The Role of English as a Source and Target language

	(1) $\frac{M_{odt}}{M_{oot}}$	(2) $\frac{M_{odt}}{M_{oot}}$	(3) $\frac{M_{odt}}{M_{oot}}$	(4) $\frac{M_{odt}}{M_{oot}}$
$DL_{odt}^{S=EN}$	0.729***	0.827***		
	(0.155)	(0.186)		
$DL_{odt}^{S\neq EN}$	0.207***	0.251***		
	(0.073)	(0.097)		
$DL_{oot}^{S=EN}$	-0.623***			
	(0.112)			
$DL_{oot}^{S\neq EN}$	0.038			
	(0.172)			
$DL_{odt}^{T=EN}$			0.219***	0.543***
			(0.084)	(0.093)
$DL_{odt}^{T\neq EN}$			0.257***	0.109
			(0.095)	(0.087)
$DL_{oot}^{T=EN}$			-0.705***	
			(0.216)	
$DL_{oot}^{T\neq EN}$			0.129	
			(0.175)	
Observations	123263	123263	123263	123263
Origin-destination	✓	✓	✓	✓
Origin-year FE		✓		✓
Destination-year FE	✓	✓	✓	✓

PPML regressions based on the sample and specification of column 1 and 2 of Table 3. See notes to Table 3 for the estimation strategy, data and sample. Column 1 and 2 report results from models where the Duolingo exposures are calculated separately for courses with- and without English as source language. Column 3 and 4 report results from models where the Duolingo exposures are calculated separately for courses with- and without English as target language. Standard errors reported in parentheses are two-way clustered: on the country of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Notes: See notes to Table 3

6.2.2 Heterogeneity

Table 5 shows the interaction of foreign Duolingo exposure with several other variables. As a validation, I confirm that the effect is stronger for those countries searching more intensively for Duolingo. As most users use Duolingo on mobile devices, mobile internet access should increase the availability of Duolingo at the extensive (number of users) intensive (intensity of learning) margin. I interact the foreign Duolingo exposure with the country-year level share of individuals covered by mobile networks, finding that access to mobile networks more than doubles the effect size. However, I find that the country-level number of broadband users is irrelevant to the effect size. Column 3 and 4 show that the effect size is smaller for linguistically closer countries and countries sharing a language. If countries' languages are linguistically close, learning costs were already plausibly lower, so the introduction of online language learning did not decrease costs that much. If countries already share a language, benefits to learning another destination-country language may be very low and other language learning opportunities to learn the shared language may be widespread. Although insignificant, column 5 and 6 shows that the effect is slightly stronger for higher-income origins and considerably stronger for higher-income destination countries, which is in line with larger gains from migration in high-income destination countries. In Column 7 I find that larger pre-existing migration networks reduce the effect. This could be driven by the reduced need for language learning if there is a diaspora in the destination. This implies that the availability of low-cost language learning expands the choice set of aspiring migrants, raising interest to destinations with low pre-existing migrant networks.As Duolingo does not offer certification, learning on Duolingo may be less of a relative cost reduction in language learning costs in countries with language requirements for immigrants. Table 8 test whether the presence of language requirements for residency moderate the effect of foreign Duolingo exposure. Although the effect is 5% lower to destinations with language requirements, it is not statistically significant.

Table 5: Effect Heterogeneity of Foreign Duolingo Exposure

	(1) $\frac{M_{odt}}{M_{oot}}$	(2) $\frac{M_{odt}}{M_{oot}}$	(3) $\frac{M_{odt}}{M_{oot}}$	(4) $\frac{M_{odt}}{M_{oot}}$	(5) $\frac{M_{odt}}{M_{oot}}$	(6) $\frac{M_{odt}}{M_{oot}}$	(7) $\frac{M_{odt}}{M_{oot}}$
DL_{odt}	0.318***	0.154*	0.451***	0.373***	0.378***	0.212	0.604***
	(0.078)	(0.092)	(0.096)	(0.079)	(0.079)	(0.137)	(0.108)
$DL_{odt} \times GTI_o^{Duolingo}$ (2006-2022) [0,1]	0.729**						
	(0.306)						
$DL_{odt} \times 3G_{oy}$ [0,1]		0.332**					
		(0.152)					
$DL_{odt} \times \text{Broadband}_{\text{oy}}$ [0,100]		-0.000					
		(0.005)					
$DL_{odt} \times$ AP15 Linguistic $\text{proximity}_{\text{od}}$ [0,1]			-0.505**				
			(0.218)				
$DL_{odt} \times$ Shared official $\text{language}_{\text{od}}$ {0,1}				-0.416**			
				(0.170)			
$DL_{odt} \times GDPpc_{oy}$ (1 s.d.)					0.062		
					(0.065)		
$DL_{odt} \times GDPpc_{dy}$ (1 s.d.)						0.153	
						(0.102)	
$DL_{odt} \times$ Above-median migrant $\text{stock}_{\text{od,2005}}$							-0.255**
							(0.123)
Observations	121180	89759	114137	122242	117908	115121	122279

Notes: PPML regressions based on the sample and specification of column 2 of Table 3. See notes to Table 3 for the estimation strategy, data and sample. Every column introduces an interaction effect between foreign Duolingo exposure and a moderator. Due to limited availability of data on the moderator the sample size varies. GTI is obtained from Google Trends, 3G data is obtained from Collins Bartholomew, Broadband subscription data from ITU, Linguistic Proximity from Adserà and Ferrer (2021), a dummy for sharing an official language from Conte, Cotterlaz and Mayer (2022), GDP per capita (PPP) from the World Bank and the stock of migrants in 2005 from the UN International Migrant Stock database. For ease of interpretation and to prevent a spurious correlation with time trends, I standardized the measure of GDP for every year of the data. Using the migration stock in 2005, I calculate the median number of migrants by origin country. Using the median, I construct a binary indicator taking value one if a particular destination housed an above-median amount of migrants. Standard errors reported in parentheses are two-way clustered: on the country of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6.2.3 Robustness

I perform a series of additional robustness tests on the main results of Table 3. First, I examine four potential risks to identification and thereafter the sensitivity of the main results to changes in the sample.

Identification. Although I argue that course enrollment is plausibly exogenous to trends in migration intentions across language areas, one can use features of Duolingo rollout to construct exogenous instruments for Duolingo exposure. Section D.5.2 discusses two such instruments: one based on Duolingo's tendency to develop courses for language dyads with many speakers of the source and target languages and one based on the reduced cost of course development when there are other courses developed for both the source and target language. Table 7 reports the results. Both instruments predict foreign Duolingo exposure well, although the latter is weaker, which increases estimated standard errors considerably. In both cases the effect size is larger than the baseline PPML results. In part, this could be driven by measurement error in the Duolingo exposure. In particular, as Duolingo courses have become more extensive over time and as more people have access to a desktop or smartphone, the actual exposure to low-cost learning may have been gradually increasing over time, which is captured by these instruments linearly increasing over time.

As mentioned in 4.4, residual multilateral resistance could be present in the model with three-way fixed effects. If preference shocks for destinations are correlated within a nest or group of countries, multilateral resistance terms are common to that nest (by year and origin country). Hence, Table 14 shows the results for specifications with origin-year-destination fixed effects for three types of nests: World Bank's 7 global regions, World Bank's 4 country groups, and membership of the EU. As these specifications are demanding as they introduce many fixed effects to the model, I report the total number of FEs estimated in the table. Although the fixed effects based on the World Bank regions reduce the point estimate somewhat, the estimate remains and large across all specifications.

As discussed in section 4.4, the Duolingo exposure measures are strongly correlated across countries speaking the same languages, which could lead to an underestimation of the standard error if there is intra-cluster correlation in migration intentions. To allow for arbitrary correlations within origins and within destinations with the same native language, I cluster standard errors on different levels in Table 12. In particular, in columns 5 and 6 I cluster standard errors by most spoken language in the origin (63 clusters) and most spoken language in the destination (70 clusters). The results are only slightly less significant than the baseline results clustered by origin- and destination country. However, clustering on the most spoken language may be too granular. For example, many former colonies may house many speakers of one of the world languages for which many Duolingo courses are rolled out, although the most spoken language is a native language. As migration patterns in such countries could be correlated to each other and their colonial hosts, this could be a concern. Hence, I modify the condition for most spoken language in the following way: I count an origin (destination) country into the language group of the most spoken language that is a source (target) language in any Duolingo course. This yields considerably fewer clusters: 22 on the origin level and 30 on the destination level. Nevertheless, columns 7 and 8 of Table 12 show that the standard errors only become slightly larger, confirming prior results.

As mentioned in 4.2, if Duolingo modules are developed due to increased demand for language learning, this is most likely driven by the countries with the most source- and target speakers. To test whether the countries with most speakers by language are not driving the results, I perform three exercises. First, I remove the contribution of the source language in the origin country with most speakers, for every source language, from the Duolingo exposure. Second, I omit the contribution of the target language in the destination country with most speakers, for every target language. Third, I do both at the same time. Table 13 shows the results, which have changed little compared to Table 3. Hence, it shows that the effects are not (just) driven by the countries for whose markets Duolingo courses are most likely developed.

Sample. The first year in the data is 2007, which means that 6 years of data before the first course introduction is included in the main estimation sample. Although this choice of starting year is driven by the starting date of the GWP interviews, it is still somewhat arbitrary, and one may be concerned that such a long pre-period, including the 2007-2008 financial crisis and its aftermath, could strongly affect the diff-in-diff estimates. Table 16 shows that it is not the case; using later time periods decreases the point estimate on foreign Duolingo exposure. Moreover, the last column of Table 16 shows that omission of the Covid-19 period and subsequent years has only a limited effect on the estimate.

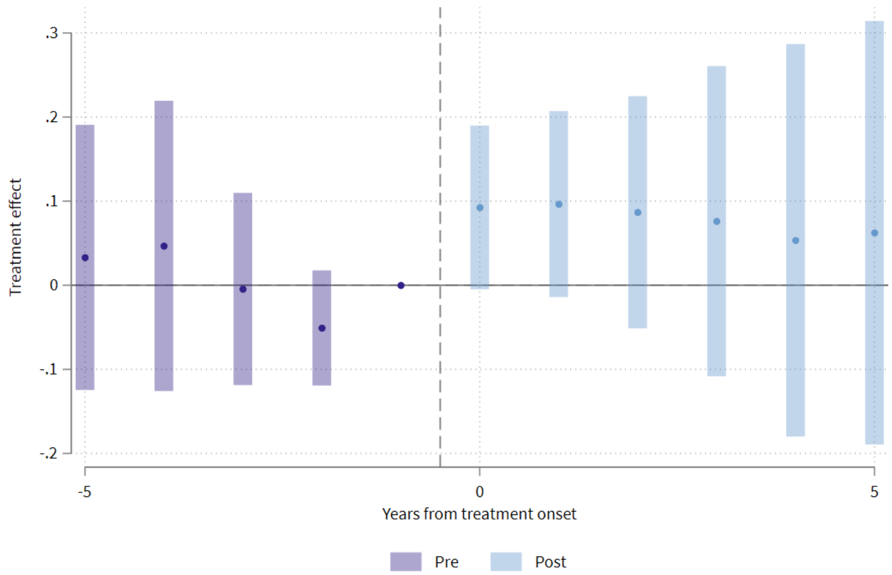
To study how sensitive the results are to single Duolingo courses, I recalculate the main measure of foreign Duolingo exposure omitting a single course at a time. Figure 8 report both the point estimates and p-values from this exercise. Results are mostly insensitive, with the exception of one course: Spanish to English. The coefficient on foreign Duolingo exposure decreases to 0.23, but remains significant (p=0.006). This is not completely surprising, as it is the most popular course on Duolingo with more than 50 million learners (see Figure 7). To ensure that the estimates are not driven by single origin- or destination countries, I omit one country at a time. The results, shown in Figure 9 and 10, indeed confirm that the effects are not driven by any single country.

6.3 Flows to OECD Countries

Previous section has established that low-cost language learning has a strong effect on bilateral migration intentions. However, did these translate into increases in migration flows? Unfortunatally, there is no dataset containing information about yearly gross migration flows between countries. To nevertheless provide a partial answer to this question, I study the impact on yearly bilateral migration flows to OECD countries. In similar vein to 5, I present event study estimates in Figure 6 around increases in foreign Duolingo exposure of 50 percentage points or more. The results are not conclusive. Although the instantaneous and first post-treatment effect are positive and close to significant, the first pre-treatment estimator is negative, suggesting that migration flows increased between the second and first year before treatment. If anything, these results would suggest a considerable effect directly upon treatment, which seems implausible given the delayed response for migration intentions in 6 and the time it take for language learning and migration preparations.

I interpret these results with caution for two additional reasons. First, the uncertainty of the estimates is large in comparison. The standard error of the average effect if 0.12, which is too large to detect a reasonably sized effect. Second, an important limitation of the OECD migration dataset is that it is incomplete on the destination level. This is problematic in a setting where treatment is correlated across same-language destinations which are partially inside and partially outside the dataset. For example, if a course becomes available to the language all same-language destinations are affected and exercise a downward spillover effect on each other. However, because these observations are outside the dataset it is impossible to account for these multilateral resistance effects.

Figure 6: Event study of Migration Odds to OECD Countries around Large Increases in Duolingo$_{odt}$ (2007-2019)



Notes: PPML regressions of the heterogeneity-robust event study estimator by Nagengast and Yotov (2023) of migration flow odds on a binary indicator for whether a country pair has experienced an increase in foreign Duolingo exposure exceeding 50 percentage points, including origin-destination pair, origin-year and destination year fixed effects. Estimates are shown for 5 years before and after an event. Shaded blue bars indicate 95% confidence intervals based on standard errors clustered at native language level. Data on migration flows originates from the OECD and the Duolingo exposure is constructed using data from (Ginsburgh, Melitz and Toubal, 2017) and the rollout dates of Duolingo courses.

6.4 Global Flows of Scholars

To nevertheless estimate the effect on global migration flows, I turn to data on a specific type of migration flows: academic scholars moving between institutions. Table 6 shows the results from the estimation of a gravity model of scholarly migration odds, net of three-way fixed effects. I find that foreign Duolingo exposure increases bilateral flows by slightly more than 4%, which is just significant at the 5% level. As most scholars in English-speaking countries are publishing in English or are immigrants less proficient in the local language who are more mobile than natives, courses from English may be particularly fruitful. Effects are indeed stronger for courses from English: full foreign Duolingo exposure increases flows with more than 7.5%. Basic levels of language skills could aid high-skilled workers to take up jobs abroad, even though they do not need the local language for their job: For example, it could facilitate other aspects of life including finding housing or interacting with authorities. Partitioning the exposure measure over English as a target language, I do not find any significant effect.

Table 6: The Effect of Duolingo on Scholarly migration flows (2007-2019)

	(1) $\frac{M_{odt}^{scholar}}{M_{oot}}$	(2) $\frac{M_{odt}^{scholar}}{M_{oot}^{scholar}}$	(3) $\frac{M_{odt}^{scholar}}{M_{oot}^{scholar}}$
DL_{odt}	0.042**		
	(0.021)		
$DL_{odt}^{S=EN}$		0.074**	
		(0.032)	
$DL_{odt}^{S\neq EN}$		0.023	
		(0.025)	
$DL_{odt}^{T=EN}$			0.031
			(0.033)
$DL_{odt}^{T\neq EN}$			0.039
			(0.025)
Observations	167670	168024	168024

PPML regressions of the odds of scholar migration based on the sample and specification of column 2 of Table 3. See notes to Table 3 for the estimation strategy, data and sample. Column 2 and 3 use the same procedure as Table 4, splitting the Duolingo exposure by English and other languages for source- and target languages, respectively. Data on scholarly migration flows originate from Akbaritabar, Theile and Zagheni (2024) and the Duolingo exposure is constructed using data from (Ginsburgh, Melitz and Toubal, 2017) and the rollout dates of Duolingo courses. The estimation sample concerns 189 unique origin and 194 unique destination countries. Standard errors reported in parentheses are two-way clustered: on the country of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

7 Migrants' Language Skills, Selection and integration

One group that has large incentives to take up low-cost language learning are prospective migrants. Greater pre-arrival learning in expectation of a future move most likely improves the language skills of migrants upon arrival. However, the access to low-cost language learning may also particularly facilitate migration for low-skilled individuals who only need basic skills for employment and who had less access to language leaening before arrival. Hence, a priori the expected effect of pre-arrival language learning opportunities is ambiguous. In addition, better language skills before arrival can be used to search more effectively for employment opportunities in the destination. After arrival, migrants can take up low-cost language learning while residing in the host country. Unless this strongly affects selection into return migration, this imaginably improves their language skills. All of these may influence subsequent migrant integration.

To study these effects, I turn to two large-scale survey data sets for two distinct geographic settings between 2007 and 2022: the EU Labor Force Survey (EU LFS) in the EU and American Community Survey (ACS) in the US. Both data sets include information on respondents background characteristics, including country of birth and year of immigration, as well as questions on economic integration. The ACS includes questions on self-assessed language proficiency at the time of interview throughout, but has no question about language skills upon arrival, which the 2021 EU LFS does have.

Both settings have considerable variation in the availability of Duolingo courses over immigrant origin (and destination for the EU) countries over time- English is the most learned second language on Duolingo and the language with the most courses as target language on Duolingo (22).[37] However, it concerns only a single target language for a single destination country, which limits the variation and renders the identifying assumptions weaker. The EU LFS does not face this problem, as it (including the United Kingdom, Switzerland, and Norway) hosts many languages that are target of Duolingo courses (62), as well as several countries that are not. A drawback of the EU LFS is that it does not distinguish the exact origin country of birth of immigrants, but rather one of 24 broad origin country groups.

In the following, I present the main strategies to identify the effect of pre- and post-arrival exposure to Duolingo on migrants' language skills, selection and integration. Because this identification strategy is arguably stronger for the EU, I present the results using the EU LFS in sections 7.3 and 7.4. Nevertheless, in section 7.5 I discuss results for the US using the ACS, which are further presented in detail in section D.7, and compare the findings.

7.1 Empirical strategies

Upon arrival. To study the effect of low-cost language learning on migrants' characteristics and language skills upon arrival, I estimating the following specification, which uses the staggered introduction of Duolingo courses:

$$y_{iodct} = \beta DL_{oc}^{T_d} + \phi_{od} + (\psi_{oc}) + \theta_{dc} + \xi_t + \epsilon_{iodct} \tag{11}$$

Here, y_{iodct} denotes the outcome of individual i who migrated from origin country o to destination country d in year c interviewed at year t. In some cases, this outcome is realized at year c and faithfully reported in t. An example is the language skills upon arrival. In other cases, I only observe the contemporaneous value of y. For these, I rely only on those interviewed within the first year after arrival, to ensure that outcomes are as close to their values upon arrival as possible. The main independent variable of interest is the Duolingo exposure $DL_{oc}^{T_d}$ to the native language of d. ϕ_{od} are dyadic fixed effects, which capture origin-destination pair specific immigrant characteristics. For example, this controls for the linguistic distance between main languages in countries, which are an important determinant of migrant integration. ψ_{oc} and θ_{dc} are origin-cohort and destination-cohort fixed effects, capturing unobserved heterogeneity in origin-

[37] As English is the second most studied language in the USA on Duolingo and in 2017 between 2 and 6% of inhabitants across the 50 states used Duolingo, it is plausible that many immigrants use it to improve their English skills after arrival. https://blog.duolingo.com/the-united-states-of-languages-an-analysis-of-duolingo-usage-state-by-state/

specific and destination-specific characteristics of migration cohorts, such as selection into emigration and destination-country immigration policies. ξ_t captures year-specific factors.

This constitutes a triple differences in cohort time with a fully continuous treatment. For the US, there is only one destination country d, which removes the possibility of including origin-cohort FE (and I include origin FE instead) and destination-cohort FE simply become cohort FE. In this case, it becomes a simple two-way fixed effects specification. To nevertheless account for origin-specific cohort quality, I include a comprehensive vector of time-varying controls to eliminate differences in cohort quality explained by observable factors. To identify β as the causal effect of the availability of low-cost language learning, the outcomes need to fulfill parallel trends in levels for all levels of treatment intensity. The plausibility of this assumption can be examined by considering whether there are pre-trends in outcomes between strongly treated and untreated units.

As in section 6, I use the foreign exposure to Duolingo, proxying for the returns of a language skills. However, as the surveys ask about language skills in the main native question, here I construct the Duolingo exposure $DL_{oc}^{T_d}$ as the probability that a Duolingo course enables communication between a random person in the origin country and a random person in the destination country who speaks the native language. For the EU, the origin indicator o is replaced by the origin group indicator o_g and I calculate the aggregated exposure at the origin group level, weighting with origin-specific cohort sizes using the yearly bilateral flow data introduced in section 6.3:

$$DL_{o_g c}^{T_d} = \frac{1}{\sum_{o \in o_g} N_{odc}} \sum_{o \in o_g} N_{odc} DL_{oc}^{T_d} \tag{12}$$

Here, $N_{o_d c}$ are the gross flow of immigrants from o to d arriving in c. For some origin regions the weighting is not restrictive because a single language dominates (such as North Africa or Latin America) or because migrants mostly originate from one country in a global region (e.g. Vietnamese in Czech Republic). However, for others this introduces measurement error. Section D.6.2 discusses this further and provides an estimate of the degree of attenuation bias introduced due to the aggregation: it is about 25%.

After arrival. After an immigrant has arrived in the host country she can continue learning on Duolingo or start a Duolingo course, if a relevant course is available. To isolate this post-arrival effect from that of pre-arrival exposure to Duolingo, I sketches the identification problem in Figure 11. For sake of simplicity, I consider a binary treatment, but the logic also extends to settings with continuous treatment intensity. Each panel of Figure 11 shows the availability of a relevant low-cost language course before and after arrival based on the time of arrival and time of interview relative to course rollout. I illustrate the exposure regimes someone falls into (upper) and the resulting pre- and post-migration exposure (lower panel). I do this in two cases: one where the migrant is interviewed within the first year upon arrival (left) and in the second year after arrival (right). In the former case, those arriving before rollout had no access to the course and those arriving after rollout had access to the course course before moving, but as they arrived recently had no time yet to study after arrival. In the latter case, those arriving at least two years before course rollout had no opportunity to use Duolingo, whereas those arriving closer to the rollout date have a gradually longer time window to have used the course after arrival. This provides variation in pre-arrival To capture this, I calculate the post-arrival exposure as the average exposure to Duolingo since arrival:

$$DL_{otc}^{T_d,post} = \frac{1}{t-c}\sum_{\tau=1}^{t-c} DL_{o(t+1-\tau)}^{T_d} \tag{13}$$

Here, I denote time since arrival in the destination by $t-c$. $DL_{o(t-\tau)}^{T_d}$ is the Duolingo exposure from o to d in year $(t-\tau)$. For those interviewed during the year of arrival $(t=c)$, I set $DL_{otc}^{T_d,post}$ to 0, as they had limited time to take up the course after arrival. To estimate the effect of pre- and post-arrival exposure jointly, I estimate the following model:

$$y_{iodct} = \sum_i \beta_i DL_{oc}^{T_d} \times \mathbb{1}(t-c=i) + \gamma DL_{otc}^{T_d,post} + \phi_{od(t-c)} + (\psi_{oc}) + \theta_{dc} + \xi_t + \epsilon_{iodct} \tag{14}$$

Here, notation follows that of equation 11. Compared to equation 11, I flexibly estimate the effect of pre-arrival exposure on outcomes for each number of completed years since arrival. This is important, as an initial effect of pre-arrival learning may diminish over time relative to the unexposed immigrants, who may catch-up. As pre-arrival and post-arrival exposure are correlated, flexibly capturing dynamic effects of pre-arrival exposure is also important to ensure that the post-arrival exposure does not spuriously capture pre-arrival effects. Additionally, I include pair-by-time since arrival fixed effects $\phi_{od(t-c)}$. These control for general and origin-by-destination-specific integration patterns over time, such as cultural and linguistic distance, and selection into return migration. To make the sample representative of the EU and US, I weight all regressions using the representative yearly weights provided by the EU LFS and ACS, respectively.

To interpret γ as the causal effect of post-arrival exposure on immigrant characteristics and outcomes, stronger exposed immigrants should have had similar integration patterns than weaker or unexposed immigrants in absence of post-arrival Duolingo exposure, conditional on pre-arrival exposure. Furthermore, to pin down the causal effect of learning on outcomes I also need to assume that post-arrival exposure does not predict. This could be violated if selection into migration changes due to immigrants anticipating future increases in Duolingo exposure (i.e. due to the future rollout of a course). The latter is particularly unlikely for those migrants who had no access to Duolingo before arrival. To isolate the effect of post-treatment exposure in absence of pre-arrival exposure, I can estimate equation 14 on the subsample with $DL_{oc}^{T_a} = 0$ (i.e. among those who were completely unexposed upon arrival). However, the effect of post-arrival exposure may be different than the additional effect of post-arrival learning when pre-arrival learning was present. If migrants already were exposed before leaving, the effect of post-arrival learning is plausibly weaker as low-hanging learning gains have already been exploited.

An important caveat is that both data sets are repeated cross-sections of individual interviews. Hence, differential selection into answering the survey, or differential selection into return migration may affect the estimates of γ. This is potentially important as levels of linguistic and economic integration may affect return migration decisions (Dustmann, 2003). Differential selection on observables can be partially evaluated using the effect of arrival measures on age, sex and educational attainment (among older individuals).

7.2 Data

The EU Labor Force Survey (LFS) are harmonized surveys conducted by the national statistical agencies of EU countries as well as some non-EU countries. The surveys include many questions on demographic characteristics and labor market participation. For an individual's main job, it includes a variable on the monthly income decile. For migrants it includes a variable on the global region of birth, as well as the years of residence. I use the surveys between 2008 and 2021, as prior to 2008 information on the years of residence was unavailable. The EU LFS fielded add-on surveys in 2014 and 2021 covering part of the sample. The add-on modules asked the reason for migration,[38] whether one participated in a language course,[39] and language skills upon arrival in 2021, with the following answer options: hardly or none, beginner, intermediate, advanced, or mother tongue.

I restrict the sample to those who arrived at age 18 or older, to capture those who have opted to migrate themselves. Moreover, as I am interested in labor market outcomes, I restrict the sample to those who are aged 60 or below at the time of interview. Moreover, I focus on immigrants who are interviewed up to and including 9 years of arrival as longer times. In line with previous analysis, I consider immigrants who have arrived in 2007 and thereafter. To construct a measure of Duolingo exposure at the origin group-destination-year level, I use the OECD bilateral migration data as discussed in section 7.1. Unfortunately, migration data for Cyprus, Greece, Ireland, Malta and Romania is sparse, and I drop these destinations. The full sample includes 668737 individuals from 24 origin regions in 23 destination countries.

I construct several different datasets: those answering the question on language skills upon arrival in

[38] Employment – job found before migrating, Employment – no job found before migrating, no job found before migrating, Family reasons, Education or training, Retirement (2021), International protection or asylum, Other

[39] Yes – general language course (2021), Yes – work-specific language course (2021), Yes (2014), No – because language courses were not available or affordable (2021), No – because language skills were sufficient (2021), No – was not necessary (2014), No – for other reasons

2021, those answering questions on reasons for migration in 2014 and 2021, all of those interviewed within one year of migration, and a comprehensive dataset of all individuals up to and including the fifth of arrival. Descriptive statistics of the three samples are shown in Table 17.

7.3 Language skills and integration upon arrival

Table 7: The Effect of Duolingo Exposure on Language Skills upon Arrival, Migration Reasons and Characteristics

	(1)	(2)	(3)	(4)	(5)
Panel A: Language skills upon arrival (2021)					
	At least beginner	At least intermediate	At least advanced	Mother tongue	Did a language course
$DL_{oc}^{T_d}$	0.198***	0.151***	0.047	-0.031	-0.006
	(0.044)	(0.037)	(0.033)	(0.022)	(0.053)
Observations	19254	19254	19254	19254	18803
R^2	0.32	0.43	0.50	0.74	0.31
Mean dep. var.	0.506	0.343	0.253	0.170	0.405
Panel B: Reason for migration (2014 & 2021)					
	Employment, job on arrival	Employment, no job on arrival	Family	Education	Refugee
$DL_{oc}^{T_d}$	0.062***	-0.012	-0.008	-0.004	-0.045
	(0.024)	(0.028)	(0.040)	(0.019)	(0.028)
Observations	64648	64648	64648	64648	64648
R^2	0.13	0.17	0.10	0.13	0.42
Mean dep. var.	0.203	0.218	0.373	0.067	0.080
Panel C: Observable Characteristics in first year after arrival (2008–2021)					
	Primary educated	Secondary educated	At least tertiary education	Female	Age
$DL_{oc}^{T_d}$	0.039	-0.006	-0.033	-0.090**	-0.211
	(0.030)	(0.035)	(0.044)	(0.037)	(0.617)
Observations	37088	37088	37088	50444	50444
R^2	0.18	0.11	0.17	0.05	0.09
Mean dep. var.	0.216	0.311	0.473	0.542	32.675
Origin group-year FE	✓	✓	✓	✓	✓
Destination-year FE	✓	✓	✓	✓	✓
Origin group-Destination FE	✓	✓	✓	✓	✓

Notes: OLS estimations of the model of equation 11. Panel A is from the 2021 add on sample, Panel B is from the 2014 and 2021 add-on samples, Panel C is from the full 2008-2021 LFS. "At least a beginner" in column 1 of panel A is 1 if a respondent indicates to have had at least beginner level language skills and 0 if she answers to have had hardly any or no language skills. The subsequent levels in column 2–4 indicate binary indicators for higher minimum levels of language skills. "did a language course" is a binary indicator for whether an immigrant did a language course after arrival. Panel B reports the levels of migration reasons, where the omitted category is "other". Panel C reports current educational attainment for respondents at least 25 years of age and a binary indicator for female and an integer variable for age. The measure of Duolingo exposure is constructed using data from (Ginsburgh, Melitz and Toubal, 2017) and the rollout dates of Duolingo courses. Standard errors reported in parentheses are two-way clustered: on the country group of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

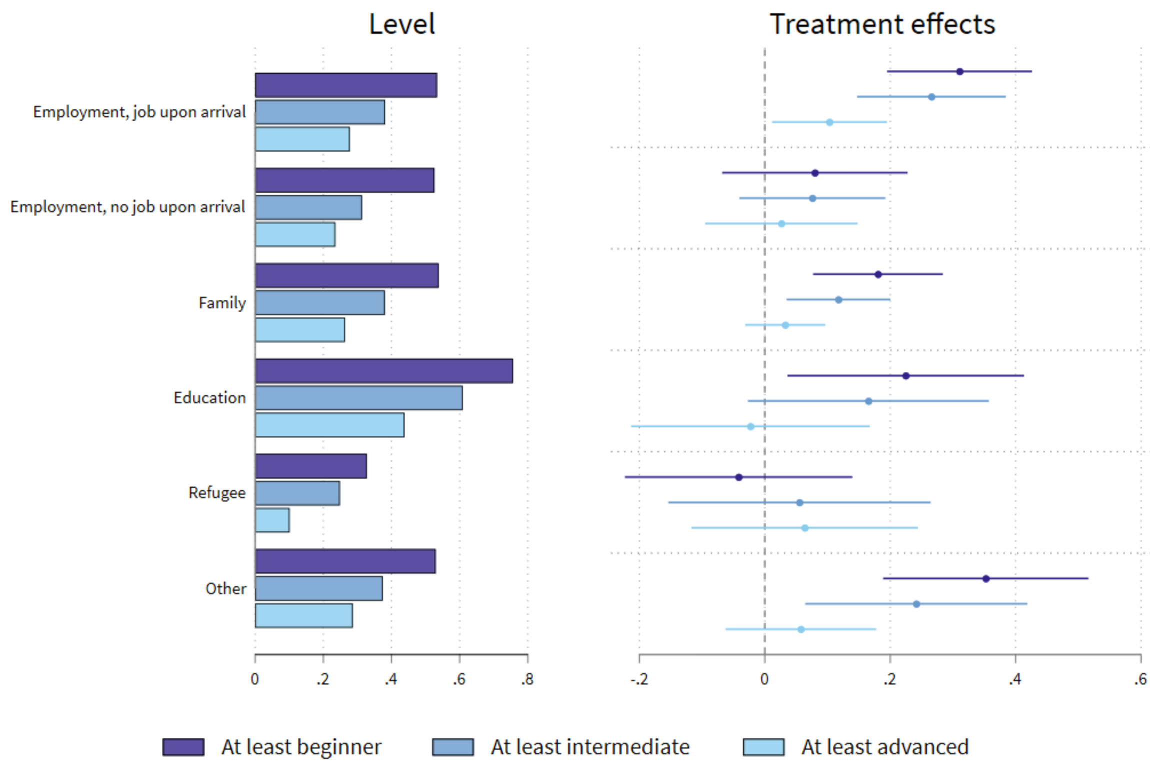
Figure 13 shows the results from an event study of three levels of language skills around large increases in Duolingo exposure. Due to the relatively low number of observations I bin the cohorts relative to arrival. I find no evidence of pre-trends in language skills upon arrival, but a considerable increase of language skills after arrival. Panel A of Table 7 reports estimation results from the model of equation 11. On a mean of about 51%, I find that full Duolingo exposure increased the language skills of migrants by 20 percentage points. As Duolingo exposure in 2021 is about 40 percentage points across the EU migrant pool, this implies that the share of respondents with at least beginner skills, or an increase of around 15%. Moreover, the share of individuals reporting at least intermediate language skills has also improved considerable, but there is no effect on advanced language skills or speaking the native language as your mother tongue. The latter is reassuring, as the availability of language learning should not impact those speaking the language natively. Moreover, as Duolingo enables acquisition of basic language skills, it probably has no large effects on advanced skills. Ultimately, there is no effect on the propensity to have done a language course in the origin country, suggesting that Duolingo neither crowds out taking language courses in the host country, but also does not boost it. Panel C of 8 explores the effect on reasons for taking or not taking a language course. Respondents are much more likely to answer that they did not take a course because they possessed sufficient language skills upon arrival.

Does this large increase also affect the reasons for immigration? It turns out that it does. Panel B of Table 7 shows that migrants are 6 percentage points more likely to have arrived for employment with having a found a job before arriving. This suggest that better language skills enable migrants to find employment on a distance. Panel C studies the selection of immigrants. I find no significant change in the skill composition in terms of formal education among those migrants aged 25 or above, although results suggest that the pool of migrants became lower skilled. Moreover, I do find that the proportion of men increases strongly.

Figure 7 examines the heterogeneity of the effect by reason for migration. Immigrants arriving for education have the strongest language skills: almost 80% have some skills upon arrival. Language skills are particularly low for refugees: about 30% have basic language skills. I find that the treatment effects are largest for immigrants who arrive for employment with a job upon arrival and "other" migrants. Moreover, the effect at beginner levels of skills are also significant for family and educational migrants. Skills among migrants who are less prepared upon arrival, those arriving for employment without a job and refugees, are unaffected by exposure to low-cost language learning.

Panel A and B of Table 8 reports results for economic integration. In the first year after arrival, Duolingo exposure increases the probability to work by 10 percentage points, which goes at the expense of all other categories. In term of job characteristics, I find no difference in the propensity to be self employed or to work on a temporary contract. In line with the increase in employment, working hours increase by 15% on average. However, using the information on income deciles based on monthly wage income, it seems that migrants do not earn more. In fact, these estimates suggest that hourly wages declined. This could be partially driven by the worsening skill composition of the immigrant pool as suggested by the results in Table 7.

Figure 7: Migration Reason-specific Language Skills and the Effect of Duolingo



Notes: Levels of language skills upon arrival by migration reason among those who arrived before any Duolingo exposure. (left) and the treatment effect of Duolingo exposure by migrant group (right). OLS estimations of the model of equation 11 including indicators for the levels of reasons for migration (not shown) and an interaction between the pre-arrival Duolingo exposure and all levels of reasons for migration. The estimates on the interaction terms are shown with 95% confidence intervals based on two-way clustered standard errors: on the country group of origin and country of destination level. For information on the data and outcome variables and migration reasons, see notes to Table 7.

Table 8: The Effect of Duolingo Exposure on Additional Outcomes

	(1)	(2)	(3)	(4)	(5)
Panel A: Main activity in first year after arrival					
	Employed	Unemployed	Retired	Student	Stay-at-home
$DL_{oc}^{T_d}$	0.104**	-0.012	-0.016*	-0.037	-0.023
	(0.049)	(0.044)	(0.009)	(0.036)	(0.038)
Observations	45216	45216	45216	45216	45216
R^2	0.23	0.11	0.04	0.15	0.14
Mean dep. var.	0.496	0.141	0.040	0.146	0.155
Panel B: Employment in first year after arrival					
	Self-employed	Temporary work	log of hours usually worked	Income decile	Lowest income decile
$DL_{oc}^{T_d}$	-0.018	0.025	0.143**	-0.305	-0.055
	(0.031)	(0.041)	(0.064)	(0.373)	(0.046)
Observations	26699	24028	25989	16000	16000
R^2	0.12	0.17	0.11	0.24	0.14
Mean dep. var.	0.093	0.324	3.559	4.937	0.134
Panel C: Language courses (2021)					
	Yes, General	Yes, Work-specific	Not available of affordable	Not because skilled enough	Not for other reasons
$DL_{oc}^{T_d}$	-0.008	-0.003	-0.039	0.085**	-0.035
	(0.054)	(0.016)	(0.027)	(0.036)	(0.039)
Observations	19128	19128	19128	19128	19128
R^2	0.28	0.13	0.15	0.41	0.16
Mean dep. var.	0.356	0.053	0.107	0.345	0.140
Origin group-year FE	✓	✓	✓	✓	✓
Destination-year FE	✓	✓	✓	✓	✓
Origin group-Destination FE	✓	✓	✓	✓	✓

Notes: OLS estimations of the model of equation 11. Panel A and B consider those interviewed in the first year after arrival in the full 2008-2021 LFS, Panel C is from the 2021 add-on sample. The outcomes in Panel A are the mutually exclusive categories of main activity within the first year after arrival, Panel B shows other job characteristics. The EU LFS does not include labor income but reports the within country-year income decile. In Column 4 I report OLS regressions using the integer income decile between 1 and 10 as outcome variable, and column 5 uses a binary indicator for the lowest income decile. The outcomes in Panel C are the levels of the categorical question on whether someone has taken a traditional language course after arrival and the reasons why if yes and not. The measure of Duolingo exposure is constructed using data from (Ginsburgh, Melitz and Toubal, 2017) and the rollout dates of Duolingo courses. Standard errors reported in parentheses are two-way clustered: on the country group of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

7.4 Integration after arrival

To study integration in the first year after arrival, I restrict the sample to those in the first five years after arrival, for two reasons. First, as after five years many immigrants have considerable language skills and it seems implausible that the introduction of low-cost language learning affects is helpful. Second, the longer after arrival, the stronger return migration impacts the estimates. Unfortunately, questions on language skills are only included in 2014 and 2021, which makes a detailed analysis of the impact on language skills over time infeasible. Instead, Table 9 shows the result of pre- and post-arrival exposure on several employment-related outcomes. Column 1 shows the effect upon arrival as shown in previous tables. Column 2-4 report results from the full sample of immigrants within the first 5 years of arrival. To study whether the initial gains in outcomes fade out over time, I interact the Duolingo exposure with the log of years since arrival plus one in column 2. Column 3 introduces the post-arrival treatment intensity and Column 4 includes an interaction of pre-arrival intensity with dummies of years since arrival (not shown). The last column limits the sample to those who had no exposure to Duolingo before arrival.

Panel A of Table 9 shows that employment gains from pre-arrival exposure slowly fade out. Exposure to Duolingo after arrival positive affects employment. This effect is stronger among those who migrated when Duolingo was not yet available: the probability to be employed increased by almost 11 percentage points. This estimate is remarkably similar to the effect from pre-arrival exposure in Column 1.

Table 9: The Effect of Duolingo Exposure on Language Skills after arrival

	(1) Upon arrival	(2)	(3) Full	(4)	(5) Arrival before any Duolingo Exposure
Panel A: Main activity: employment					
$DL_{oc}^{T_d}$	0.104**	0.103***	0.090**	0.070*	
	(0.049)	(0.034)	(0.035)	(0.039)	
$DL_{oc}^{T_d} \times \log(t-c+1)$		-0.037	-0.063**		
		(0.026)	(0.026)		
$DL_{otc}^{T_d,post}$			0.062***	0.048*	0.114***
			(0.023)	(0.026)	(0.041)
Observations	45216	398697	398697	398697	234743
R^2	0.23	0.15	0.15	0.15	0.14
Mean dep. var.	0.496	0.558	0.558	0.558	0.550
$DL_{oc}^{T_d} \times (t-c)$ FE				✓	

Notes: OLS estimations of the model of equation 14, with the following outcomes: Panel A shows results for a binary indicator for having employment as one's main activity. Column 1 shows the effect upon arrival as shown in previous tables. Column 2-4 report results from the full sample of immigrants within the first 5 years of arrival. To study whether the initial gains in outcomes fade out over time, I interact the Duolingo exposure with the log of years since arrival plus one in column 2. Column 3 introduces the post-arrival treatment intensity and Column 4 includes an interaction of pre-arrival intensity with dummies of years since arrival (not shown). The last column limits the sample to those who had no exposure to Duolingo before arrival. The measure of Duolingo exposure is constructed using data from (Ginsburgh, Melitz and Toubal, 2017) and the rollout dates of Duolingo courses. Standard errors reported in parentheses are two-way clustered: on the country group of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

7.5 Language skills in the US

Section D.7 shows and discusses a similar analysis for the US as presented above for the EU. Contrary to the EU case, I find that pre-exposure Duolingo availability does not increase language skills in the US. Moreover, I find suggestive evidence that selection into migration becomes more negative in terms of education, although statistically insignificant. However, full exposure to Duolingo after arrival increases language skills of immigrants, and improves employment rates by 4 percentage points. The stark differences between the EU and US results for pre-arrival exposure could be explained by several reasons. First, language skills upon arrival are arguably better in the case of the US than for the EU as English is more widely spoken abroad than any of the other European languages. However, as the EU LFS and ACS ask questions with different answer options, it is not possible to test difference in the levels of immigrants' language skills directly. Second, English courses were often the first available courses, when Duolingo was not yet popular in a given linguistic region of origin. On the contrary, the courses to European languages were often rolled out after English and in later years when Duolingo was already established and amassed many users. This could have led to a quicker and stronger learning response in the EU. Third, European countries have stricter language requirements than the US, which may provide stronger motives for learning before arrival.

Altogether, I find employment effects of 10 percentage points for full pre-arrival exposure and post-arrival exposure in the EU, and no effect for pre-arrival exposure and a 4 percentage point effect for post-arrival exposure in the US. The magnitude of some of these effects is surprisingly large, especially given the potential attenuation bias due to aggregation of exposure measures as discussed above for the EU. Comparing my estimates to those from the literature, these estimates seem large. However, it is hard to compare these estimates directly, as I do not have reliable information about the share of migrants taking up Duolingo. Foged, Hasager and Peri (2024) finds that extensive language training programs (200 additional hours of instruction) in Denmark improve short-run employment rates of refugees by 3–4 percentage for women and 8 percentage points for men.[40] Lochmann, Rapoport and Speciale (2019) finds similarly large effect sizes among immigrants in France, which also includes non-refugees. However, these are effect sizes of actual uptake, and my estimates are intention-to-treat estimates with an unknown take-up rate.

The calculated Duolingo exposures based on communication probabilities are imperfect measures of whether a Duolingo course is relevant or not across a migration corridor, which could lead to overestimation of the estimate of the effects of course availability. For example, as many courses to European languages are using English as a source language, an underestimation of the share of English speakers among the pool of prospective migrants could lead to an exaggeration of the effect size. This can be driven by two different factors: first, the migrant pool may have considerably larger English skills than the whole native population and second, the data I use on the share of speakers from (Ginsburgh, Melitz and Toubal, 2017) relies predominantly on data from 2012 and before.

8 Conclusion

Language proficiency is paramount to the success of immigrants. The ability to learn a host country's language enable the acquisition of skills that increase labor market earnings and lower migration costs. Moreover, the ability to learn a language may also foster interest in a country's language and culture,

[40] In line with this paper, I find that initial employment gains relative to unexposed migrants become smaller after several years.

which subsequently. Low-cost language learning offered on online platforms facilitate language acquisition among individuals without prior access to language learning, deem investing in a language course too risky, or perceive the cultural distance as large. Using the staggered introduction of language courses on Duolingo and the global distribution of spoken languages, I provide causal evidence of the availability of low-cost language learning on migration intentions and flows, selection into migration, the language skills of immigrants and subsequent economic integration. As a first step, I aim to understand how the availability of low-cost language availability affected language learning and skills. I find that availability of Duolingo improved scores on (predominantly) passive components of English-language tests and increased school-based instruction of learnable languages.

To study the effect of language course availability on bilateral migration intentions and flows, I construct a measure of how beneficial taking up a language course is for a prospective migrant across a corridor. This foreign Duolingo exposure captures the probability course take-up of newly enables two randomly picked individuals in two countries to communicate to eachother. Moreover, I additionally construct a measure of domestic language exposure capturing the usability of a language course at home. Using these, I find that foreign Duolingo exposure strongly increases migration intentions within the first 3 years after course introduction, by on average 45%. Additional analysis shows that more than half of this increase is driven by diversion of migration intentions between destinations. However, I do not find strong evidence that this translates to increases in the extent of migration flows to OECD countries. Using a recently published dataset on global scholarly migration flows, I find that courses from English to other languages increase by about 7% suggesting that the availability of low-cost language learning at least can change the composition of the migrant pool.

To further study the effects conditional on migration, I turn to the European Union Labor Force Survey (EU LFS). As many migrants lack basic language skills upon arrival, they have large economic and social incentives to engage in language learning. I find that availability of an appropriate language course before arrival strongly increases the probability to possess beginner-level language skills by 20 percentage points, and to possess intermediate-level skills by 15 percentage points, but find no effects at the highest level of skills or on the share of mother tongue speakers. Moreover, I find that the proportion of migrants arriving for employment reasons with a job on arrival increases, which is suggestive of the fact that improved language proficiency increases the ability to search and find a job before arrival. Considering heterogeneity in the effect sizes by migration motive, I find that the effect is driven by those arriving with a job upon arrival, family migrants and those moving for education, which are groups who often had time to prepare before migration. I find that migrants who are unlikely to (be able to) prepare before arrival, economic migrants without pre-arrival jobs and refugees, to be unaffected. More migrants are in work due to more jobs being found before arrival, but this advantages diminishes after 3 years as unexposed migrants catch up. Despite the lowered threshold to language learning, I do not find evidence that low-cost language learning considerably worsened the educational composition of the migrant pool.

Although the identification strategy is weaker for a similar analysis among U.S. migrants, I find that no discernible pre-trends in migrant characteristics exist before exposure to Duolingo. Contrary to the EU, I do not find any effect on language skills upon arrival in the US. This could be driven by an offsetting effect through changing selection: the estimates suggest that the share of immigrants with a tertiary education has decreased by several percentage points. Exposure to Duolingo after arrival to the US improves language skills and integration outcomes. The differences in results between the EU and US could be explained by various factors, including the lack of basic skills among many EU migrants, the absence of language requirements in the US for many visa types and the relative popularity of Duolingo in the US compared to typical origin countries.

The findings of this paper suggest that availability of low-cost language learning increases the pool of potential migrants considerably and that it increases language skills at the low end of the distribution upon arrival. Policy makers concerned with addressing worker shortages or with improving integration could take this information to the heart by facilitating language learning opportunities abroad. However, contrary to the availability of costly traditional language learning as in (Jaschke and Keita, 2021), low-cost language learning does not increase the average educational attainment of migrants. As many refugee hosting countries spend vast resources on integration courses due to low initial language knowledge, the positive impact of digital learning methods on language skills can motivate the targeted development of digital language courses for major refugee origin languages.

Altogether, this paper shows how one aspect of the internet can improve the integration of migrants. However, less remains known about the overall effect of the internet on integration. Through improved information provision, immigrants may become better informed and make better destination choices, improving integration (Porcher, 2020). However, much in line with the results of this paper, migrants' skills upon arrival have improved during the internet era. Yet, due to the availability of internet in migrants' origin country, migrants spend more time online and less time with natives, worsening linguistic integration after arrival (Yarkin, 2024). Fruitful avenues for further exploration in this literature are the role of remote work on migrant selection and the role of migration experiences transmitted through social media on migration decisions.

Ultimately, the rich variation in low-cost language learning explored in this paper can be used by other scholars trying to understand how language learning opportunities affect other aspects of integration, such as social integration through time use with natives or intermarriage rates. Moreover, it can be employed to study how it affects other bilateral ties between countries, including interest in foreign cultures, tourist visits and trade patterns.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2023. "When should you adjust standard errors for clustering?" *The Quarterly Journal of Economics*, 138(1): 1–35.

Adamchik, Vera A., Thomas J. Hyclak, Piotr Sedlak, and Larry W. Taylor. 2019. "Wage returns to English proficiency in Poland." *Journal of Labor Research*, 40: 276–295.

Adema, Joop, Cevat Giray Aksoy, and Panu Poutvaara. 2022. "Mobile internet access and the desire to emigrate." CESifo Working Paper 9758.

Adserà, Alícia, and Ana Ferrer. 2021. "Linguistic proximity and the labour market performance of immigrant men in Canada." *Labour*, 35(1): 1–23.

Adserà, Alícia, and Mariola Pytliková. 2015. "The role of language in shaping international migration." *The Economic Journal*, 125(586): F49–F81.

Adserà, Alícia, and Mariola Pytliková. 2016. "Language and migration." In *The Palgrave Handbook of Economics and Language*. 342–372. London:Palgrave Macmillan UK.

Akbaritabar, Aliakbar, Tom Theile, and Emilio Zagheni. 2024. "Bilateral flows and rates of international migration of scholars for 210 countries for the period 1998-2020." *Scientific Data*, 11(1): 816.

Allcott, H., J.C. Castillo, M. Gentzkow, L. Musolff, and T. Salz. 2024. "Sources of Market Power in Web Search: Evidence from a Field Experiment."

Azam, Mehtabul, Aimee Chin, and Nishith Prakash. 2013. "The returns to English-language skills in India." *Economic Development and Cultural Change*, 61(2): 335–367.

Beine, Michel, Michel Bierlaire, and Frédéric Docquier. 2021. "New York, Abu Dhabi, London or stay at home? Using a cross-nested logit model to identify complex substitution patterns in Migration." *Using a Cross-Nested Logit Model to Identify Complex Substitution Patterns in Migration*.

Beine, Michel, Simone Bertoli, and Jesús Fernández-Huertas Moraga. 2016. "A practitioners' guide to gravity models of international migration." *The World Economy*, 39(4): 496–512.

Belot, Michèle, and Sjef Ederveen. 2012. "Cultural barriers in migration between OECD countries." *Journal of Population Economics*, 25(3): 1077–1105.

Bertoli, Simone, and Jesús Fernández-Huertas Moraga. 2013. "Multilateral resistance to migration." *Journal of Development Economics*, 102: 79–100.

Bleakley, Hoyt, and Aimee Chin. 2004. "Language skills and earnings: Evidence from childhood immigrants." *The Review of Economics and Statistics*, 86(2): 481–496.

Bleakley, Hoyt, and Aimee Chin. 2010. "Age at arrival, English proficiency, and social assimilation among US immigrants." *American Economic Journal: Applied Economics*, 2(1): 165–192.

Böhme, Marcus H., André Gröger, and Tobias Stöhr. 2020. "Searching for a better life: Predicting international migration with online search keywords." *Journal of Development Economics*, 142: 102347.

Borjas, George J. 1985. "Assimilation, changes in cohort quality, and the earnings of immigrants." *Journal of labor Economics*, 3(4): 463–489.

Borjas, George J. 1987. "Self-selection and the earnings of immigrants." *The American Economic Review*, 77(4): 531–553.

Borjas, George J. 2015. "The slowdown in the economic assimilation of immigrants: Aging and cohort effects revisited again." *Journal of Human Capital*, 9(4): 483–517.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting event-study designs: Robust and efficient estimation." *Review of Economic Studies*, forthcoming.

Brade, Raphael, Oliver Himmler, Robert Jaeckle, and Philipp Weinschenk. 2024. "Helping Students to Succeed–The Long-Term Effects of Soft Commitments and Reminders." CESifo Working Paper 11001.

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics*, 225(2): 200–230.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with a continuous treatment." *arXiv preprint arXiv:2107.02637*.

Cameron, A Colin, and Douglas L Miller. 2014. "Robust inference for dyadic data." *Unpublished manuscript, University of California-Davis*.

Chen, Jiafeng, and Jonathan Roth. 2024. "Logs with zeros? Some problems and solutions." *The Quarterly Journal of Economics*, forthcoming.

Chiswick, Barry R., and Paul W. Miller. 1995. "The endogeneity between language and earnings: International analyses." *Journal of Labor Economics*, 13(2): 246–288.
Clemens, Michael A. 2011. "Economics and emigration: Trillion-dollar bills on the sidewalk?" *Journal of Economic Perspectives*, 25(3): 83–106.
Conte, Marc N., Pierre Cotterlaz, and Thierry Mayer. 2022. "The CEPII Gravity Database." https://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=8 (last accessed March 2, 2024).
de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." *arXiv preprint arXiv:2007.04267.*
de Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*, 110(9): 2964–2996.
de Chaisemartin, Clément, Xavier d'Haultfoeuille, Félix Pasquier, and Gonzalo Vazquez-Bare. 2024. "Difference-in-differences estimators for treatments continuously distributed at every period." arXiv.org Papers 2201.06898.
Dekker, Rianne, and Godfried Engbersen. 2014. "How social media transform migrant networks and facilitate migration." *Global Networks*, 14(4): 401–418.
Di Paolo, Antonio, and Aysit Tansel. 2015. "Returns to foreign language skills in a developing country: The case of Turkey." *The Journal of Development Studies*, 51(4): 407–421.
Di Paolo, Antonio, and Bernat Mallén. 2023. "Does geographical exposure to language learning centres affect language skills and labour market outcomes in a bilingual city?" *Economic Analysis and Policy*, 80: 429–449.
Dustmann, Christian. 1999. "Temporary migration, human capital, and language fluency of migrants." *Scandinavian Journal of Economics*, 101(2): 297–314.
Dustmann, Christian. 2003. "Return migration, wage differentials, and the optimal migration duration." *European Economic Review*, 47(2): 353–369.
Dustmann, Christian, and Francesca Fabbri. 2003. "Language proficiency and labour market performance of immigrants in the UK." *The Economic Journal*, 113(489): 695–717.
Ersoy, Fulya. 2021. "Returns to effort: experimental evidence from an online language platform." *Experimental Economics*, 24(3): 1047–1073.
Foged, Mette, and Cynthia Van der Werf. 2023. "Access to language training and the local integration of refugees." *Labour Economics*, 102366.
Foged, Mette, Linea Hasager, and Giovanni Peri. 2024. "Comparing the effects of policies for the labor market integration of refugees." *Journal of Labor Economics*, 42(S1): S335–S377.
Freeman, Cassie, Audrey Kittredge, Hope Wilson, and Bozena Pajak. 2023. "The Duolingo method for app-based teaching and learning." https://duolingo-papers.s3.amazonaws.com/reports/Duolingo_whitepaper_duolingo_method_2023.pdf (last accessed Sep 13, 2024).
Ginsburgh, Victor A., and Juan Prieto-Rodriguez. 2011. "Returns to foreign languages of native workers in the European Union." *Industrial and Labor Relations Review*, 64(3): 599–618.
Ginsburgh, Victor, Jacques Melitz, and Farid Toubal. 2017. "Foreign language learning and trade." *Review of International Economics*, 25(2): 320–361.
Goethe Institute. 2024*a*. "German Courses in Bangladesh: Dates and prices." https://www.goethe.de/ins/bd/en/spr/kur/tup.cfm (last accessed Sep 13, 2024).
Goethe Institute. 2024*b*. "German Courses in Colombia: Dates and prices." https://www.goethe.de/ins/co/de/spr/kur.html (last accessed Sep 13, 2024).
Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*, 225(2): 254–277.
Grogger, Jeffrey, and Gordon H. Hanson. 2011. "Income maximization and the selection and sorting of international migrants." *Journal of Development Economics*, 95(1): 42–57.
GSMA. 2023. "Mobile internet connectivity 2023." https://www.gsma.com/r/wp-content/uploads/2023/10/The-State-of-Mobile-Internet-Connectivity-Report-2023.pdf?utm_source=website&utm_medium=button&utm_campaign=somic23 (last accessed Sep 13, 2024).
Guichard, Lucas, and Joël Machado. 2024. "The externalities of immigration policies on migration flows: the case of an asylum policy." *Journal of Economic Geography*, lbae016.
Hahm, Sabrina, and Michele Gazzola. 2022. "The value of foreign language skills in the German labor market." *Labour Economics*, 76: 102150.

Huber, Matthias, and Silke Uebelmesser. 2023. "Presence of language-learning opportunities and migration." *Labour Economics*, 84: 102409.

Huber, Matthias, Ann-Marie Sommerfeld, and Silke Uebelmesser. 2022. "Language learning: human capital investment or consumption?" *Empirica*, 49(4): 897–948.

Isphording, Ingo E. 2013. "Returns to foreign language skills of immigrants in Spain." *Labour*, 27(4): 443–461.

Isphording, Ingo E., and Sebastian Otten. 2011. "Linguistic distance and the language fluency of immigrants." Ruhr Economic Papers Working Paper 274.

Jaschke, Philipp, and Sekou Keita. 2021. "Say it like Goethe: Language learning facilities abroad and the self-selection of immigrants." *Journal of Development Economics*, 149: 102597.

Jiang, Xiangying, Haoyu Chen, Lucy Portnoff, Erin Gustafson, Joseph Rollinson, Luke Plonsky, and Bozena Pajak. 2021*a*. "Seven units of Duolingo courses comparable to 5 university semesters in reading and listening." *Duolingo Research Report DRR-21-03*.

Jiang, Xiangying, Joseph Rollinson, Luke Plonsky, Erin Gustafson, and Bozena Pajak. 2021*b*. "Evaluating the reading and listening outcomes of beginning-level Duolingo courses." *Foreign Language Annals*, 54(4): 974–1002.

Krugman, Paul. 1980. "Scale economies, product differentiation, and the pattern of trade." *The American Economic Review*, 70(5): 950–959.

Liwiński, Jacek. 2019. "The wage premium from foreign language skills." *Empirica*, 46(4): 691–711.

Lochmann, Alexia, Hillel Rapoport, and Biagio Speciale. 2019. "The effect of language training on immigrants' economic integration: Empirical evidence from France." *European Economic Review*, 113: 265–296.

Ma, Wenyue, and Paula Winke. 2019. "Self-assessment: How reliable is it in assessing oral proficiency over time?" *Foreign Language Annals*, 52(1): 66–86.

Melitz, Jacques, and Farid Toubal. 2014. "Native language, spoken language, translation and trade." *Journal of International Economics*, 93(2): 351–363.

Nagengast, Arne, and Yoto V Yotov. 2023. "Staggered difference-in-differences in gravity settings: Revisiting the effects of trade agreements." CESifo Working Paper No. 10782.

Nocito, Samuel. 2021. "The effect of a university degree in english on international labor mobility." *Labour Economics*, 68: 101943.

Ortega, Francesc, and Giovanni Peri. 2013. "The effect of income and immigration policies on international migration." *Migration Studies*, 1(1): 47–74.

Pischke, Steve. 2007. "Lecture notes on measurement error." *London School of Economics, London.*

Porcher, Charly. 2020. "Migration with Costly Information." Princeton University Unpublished working paper.

Priem, Jason, Heather Piwowar, and Richard Orr. 2022. "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts." *arXiv preprint arXiv:2205.01833*.

Rachels, Jason R, and Amanda J Rockinson-Szapkiw. 2018. "The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy." *Computer Assisted Language Learning*, 31(1-2): 72–89.

Saiz, Albert, and Elena Zoido. 2005. "Listening to what the world says: Bilingualism and earnings in the United States." *The Review of Economics and Statistics*, 87(3): 523–538.

Shortt, Mitchell, Shantanu Tilak, Irina Kuznetcova, Bethany Martens, and Babatunde Akinkuolie. 2023. "Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from public release of 2012 to early 2020." *Computer Assisted Language Learning*, 36(3): 517–554.

Silva, J. M. C. Santos, and Silvana Tenreyro. 2006. "The log of gravity." *The Review of Economics and Statistics*, 88(4): 641–658.

Statcounter. 2024. "Search Engine Market Share Worldwide." https://gs.statcounter.com/search-engine-market-share (last accessed Sep 13, 2024).

Stöhr, Tobias. 2015. "The returns to occupational foreign language use: Evidence from Germany." *Labour Economics*, 32: 86–98.

Strezhnev, Anton. 2023. "Decomposing Triple-Differences Regression under Staggered Adoption." *arXiv preprint arXiv:2307.02735*.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, 225(2): 175–199.

Tjaden, Jasper, Daniel Auer, and Frank Laczko. 2019. "Linking migration intentions with flows: Evidence and potential use." *International Migration*, 57(1): 36–57.

Tyazhelnikov, Vladimir, and Xinbei Zhou. 2021. "PPML, Gravity, and Heterogeneous Trade Elasticities." Unpublished Working Paper.

Uebelmesser, Silke, Ann-Marie Sommerfeld, and Severin Weingarten. 2022. "A macro-level analysis of language learning and migration." *German Economic Review*, 23(2): 181–232.

US Census Bureau. 2015. "How Well Do You Speak English? Assessing the Validity of the American Community Survey English-Ability Question." https://www.census.gov/newsroom/blogs/research-matters/2015/10/how-well-do-you-speak-english-assessing-the-validity-of-the-american-community-survey-english-abilit html (last accessed Oct 17, 2023).

Vesselinov, Roumen, and John Grego. 2012. "Duolingo effectiveness study." *City University of New York, USA*, 28(1-25).

Wang, Haining, Russell Smyth, and Zhiming Cheng. 2017. "The economic returns to proficiency in English in China." *China Economic Review*, 43: 91–104.

Weidner, Martin, and Thomas Zylkin. 2021. "Bias and consistency in three-way gravity models." *Journal of International Economics*, 132: 103513.

White, Roger, and David Buehler. 2018. "A closer look at the determinants of international migration: Decomposing cultural distance." *Applied Economics*, 50(33): 3575–3595.

Wong, Lorraine. 2023. "The effect of linguistic proximity on the labour market outcomes of the asylum population." *Journal of Population Economics*, 36: 609–652.

Wooldridge, Jeffrey M. 2015. "Control function methods in applied econometrics." *Journal of Human Resources*, 50(2): 420–445.

Wooldridge, Jeffrey M. 2021. "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators." *Available at SSRN 3906345.*

Wooldridge, Jeffrey M. 2023. "Simple approaches to nonlinear difference-in-differences with panel data." *The Econometrics Journal*, 26(3): C31–C66.

Yarkin, Alexander. 2024. "Does the 'melting pot' still melt? Internet and immigrants' integration." Unpublished Working Paper.

Yotov, Yoto V, Arne Nagengast, and Fernando Rios-Avila. 2024. "The European Single Market and Intra-Eu Trade: An Assessment with Heterogeneity-Robust Difference-In-Differences Methods." Drexel University School of Economics Working Paper Series 2024-5.

A Model details and extentions

A.1 Total migration

The derivative of the probability to migrate to language skills is:

$$\frac{\partial \mathbb{P}_{od}}{\partial l_{oS}^T} = \frac{e^{\mu_{od}+l_{oS}^T b_{oSd}^T}\left[\left(\sum_{d'} e^{\mu_{d'}+l_{oS}^T b_{od'T}}\right) b_{oSd}^T - \left(\sum_{d'} b_{od'T} e^{\mu_{d'}+l_{oS}^T b_{od'T}}\right)\right]}{\left(\sum_{d'} e^{\mu_{d'}+l_{oS}^T b_{od'T}}\right)^2} = \mathbb{P}_{od}\left(b_{oSd}^T - \sum_{d'} \mathbb{P}_{od'} b_{od'T}\right) \lessgtr 0 \quad (1)$$

A change in language skills increases the probability to migrate to destination d if the return to the language skill in the destination is larger than the average return across potential locations, weighted with migration probabilities. As most people do not migrate ($\mathbb{P}_{oo}$ is large w.r.t. $\mathbb{P}_{od}$, where $d \neq o$), the domestic return to language skills plays a prominent role in equation 1. If language skills are only rewarded in one destination, larger language skills increase the probability to migrate to that destination. However, if the language skill is rewarded more in other destinations or at home, this may decrease migration flows.

Moreover, larger language skills do not need to imply larger *total* emigration. equation 2 shows the condition for which this is the case. Total emigration increases if the migration probability-weighted foreign returns exceed domestic returns. Hence, if migration probabilities are low enough, and a language skills is moderately valued on the domestic labor markets, total emigration decreases. This is likely to be the case for many countries where English is rewarded on domestic labor markets but migration links to English-speaking destinations are weak (e.g. due to high moving costs).

$$\frac{\partial \sum_{d\neq o} \mathbb{P}_{od}}{\partial l_{oS}^T} = \mathbb{P}_{oo} \sum_{d\neq 0} \mathbb{P}_{od} b_{oSd}^T - (1-\mathbb{P}_{oo})\mathbb{P}_{oo} b_{oSo}^T > 0 \implies \frac{1}{\sum_{d\neq 0} \mathbb{P}_{od}} \sum_{d\neq 0} \mathbb{P}_{od} b_{oSd}^T > b_{oSo}^T \quad (2)$$

A.2 Derivation of equation 7

The marginal benefit of language skills generally is a function of the level of language skills through the migration probabilities. In the low migration limit, the deterministic part of utility of staying exceeds the utility of migrating. This is a reasonable assumption, as most people never migrate across borders during their lifetime. Using (1) the low migration limit to eliminate the skill-dependence of the denominator of the migration probability and (2) the property that $e^x \approx (1+x)$ if x is small, equation 6 can be written as:

$$\sum_d \mathbb{P}_{od} b_{oSd}^T \approx \sum_d b_{oSd}^T \frac{e^{\mu_{od}}}{e^{\mu_{oo}}} e^{l_{oS}^T b_{oSd}^T} = b_{oSd}^T(1 + b_{oSd}^T l_{oS}^T)\mathbb{P}_{od}(0) \quad (3)$$

$\mathbb{P}_{od}(0)$ denotes the approximate migration probability if $l = 0$. Due to the exponential form of the migration probabilities, the benefits from larger language skills are increasing in language skills returns in the low migration limit. Plugging this expression into equation 6 gives the following:

$$l_{oS}^{T}\left(2c_T - \sum_d \mathbb{P}_{od}(0)(b_{oSd}^{T})^2\right) = b_{oSd}^{T}\mathbb{P}_{od}(0) \tag{4}$$

Rearranging gives the following expression for equilibrium levels of language skills:

$$l_{oS}^{T*} \approx \frac{\mathbb{P}_{oo}b_{oSo}^{T} + \sum_{d\neq o}\mathbb{P}_{od}(0)b_{oSd}^{T}}{2c_{oST} - \sum_d \mathbb{P}_{od}(0)(b_{oSd}^{T})^2} \tag{5}$$

The second term in the denominator can be assumed small compared to $2c_{oST}$. To see why this is the case, I estimate c_{oST} and compare it to the second term in the denominator. Returns for English in European countries, where 30-70% (l) of the population learns English, vary between 10-50% (b, see section 4.1 for a discussion on these). Equating marginal costs and benefits in absence of migration using the mid-point of the ranges for b and l gives an estimated cost of $c_{oST} = b/2s = 0.3/(2*0.5) = 0.3$. As migration probabilities are low, and returns to foreign languages abroad do not exceed 50% (see section 2), this term is much smaller than $2c_{oST}$ in the low migration limit. Using this approximation, I arrive at the right hand side of equation 7:

$$l_{oS}^{T*} \approx \left(\mathbb{P}_{oo}b_{oSo}^{T} + \sum_{d\neq o}\mathbb{P}_{od}(0)b_{oSd}^{T}\right)\frac{1+\eta_{oST}Duolingo_{ST}}{\kappa_{oST}} \tag{6}$$

A.3 Calculation of the proxy for returns to skills

α_{cL} denotes the number of language L in country c. If I assume that all languages are equally and randomly distributed among a country's population (e.g. if $\alpha_{cL} = 0.5$ and $\alpha_{cL'} = 0.5$, 25% of people speak none of L and L'. 25% speak only L, 25% speak only L' and 25% speak both), I can calculate the probability that two randomly chosen individuals can communicate. I denote the product of the number of speakers of a language in the origin and destination as $\alpha_l^2 = \alpha_{ol}\alpha_{dl}$ and the number of languages spoken in either country by N. For convenience, l is an ordered index of languages. Using the law of total probability, I can write the probability two randomly picked individuals, one in the origin and one in the destination, as a function of α_{cL}'s. Here, I show the first $k = 4$ out of $k = N$ terms:

$$\mathbb{P}(comm_{od}) = \sum_l \alpha_l^2 - \sum_{l>l'}\alpha_l^2\alpha_{l'}^2 + \sum_{l>l'>l''}\alpha_l^2\alpha_{o'}^2\alpha_{l''}^2 - \sum_{l>l'>l''>l'''}\alpha_l^2\alpha_{o'}^2\alpha_{l''}^2\alpha_{l'''}^2 \dots \tag{7}$$

For a large number of common languages, this implies that there are many terms. For N languages, the k$^{\text{th}}$ term contains are $\binom{N}{k}$ elements. As the largest number of shared languages between any two countries in the data is 6, I calculate the terms up to and including k = 6. Using this result, and setting $\alpha_{oS} = 1$, I obtain the probability conditional on the individual in o speaking S, $\mathbb{P}(comm_{od}|S)$. Additionally, setting $\alpha_{oT} = 1$, I can calculate $\mathbb{P}(comm_{od}|S \wedge T)$. Using those, the availability of learning T from S (sloppily denoted by $DL_{S\to T}$) expands the set of people to communicate with by:

$$\mathbb{P}(comm_{od}|DL_{S\to T}, S) = \mathbb{P}(comm_{od}|S \wedge T) - \mathbb{P}(comm_{od}|S) \tag{8}$$

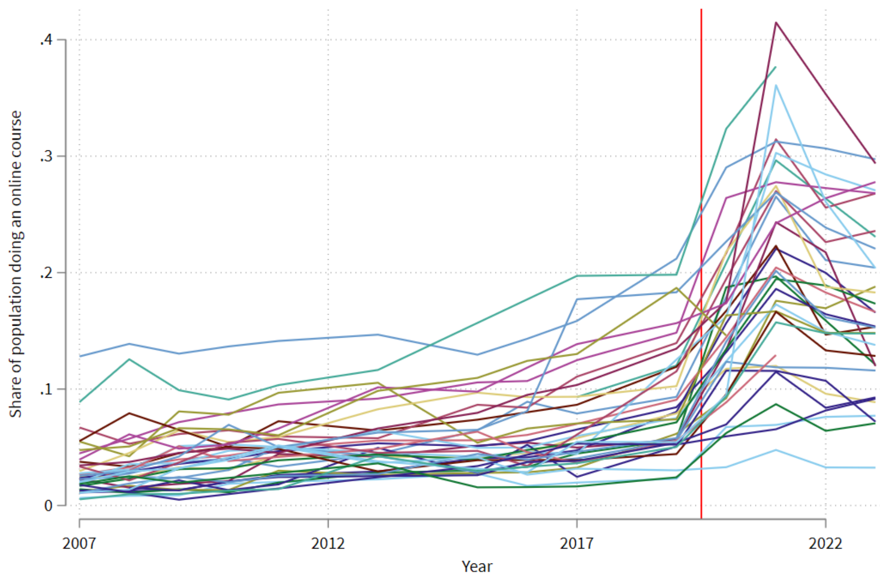
The total probability of a Duolingo course from S to T facilitating communication between two randomly chosen individuals in o and d is then simple found by:

$$\mathbb{P}(comm_{od}|DL_{S\to T}) = \mathbb{P}(comm_{od}|DL_{S\to T}, S)\mathbb{P}(S) = \mathbb{P}(comm_{od}|DL_{S\to T}, S)\alpha_{oS} \tag{9}$$

B Descriptives on Language Learning and Duolingo

B.1 Online courses

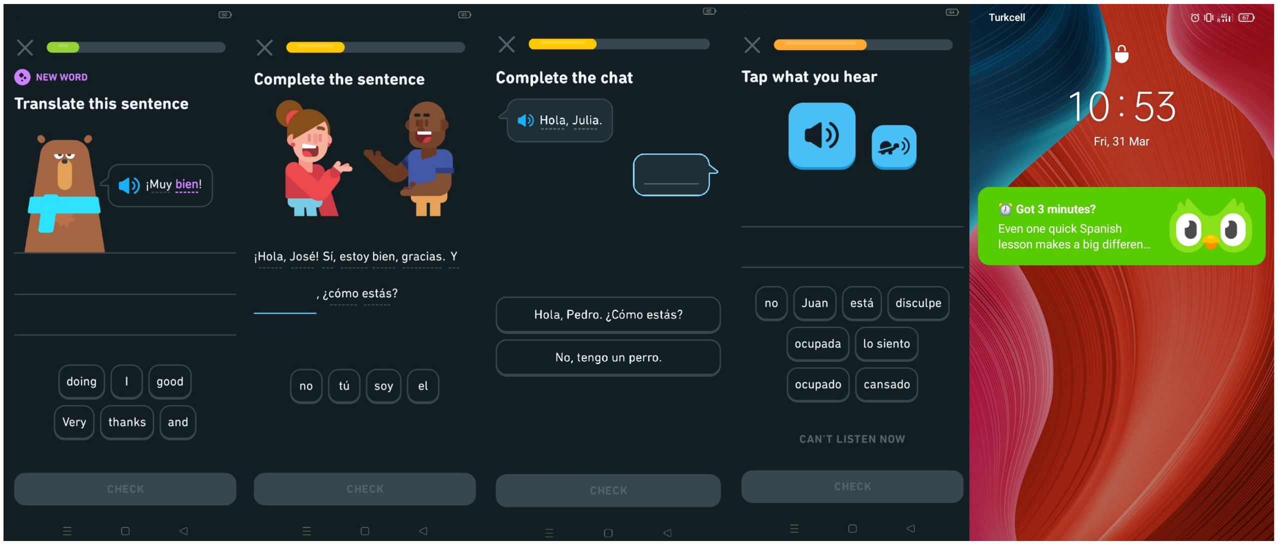
Figure 1: Share of Population Doing any Online Course across the EU



Notes: The share of population doing any online course between 2007 and 2023, by EU member state. The vertical line denotes the onset of the Covid-19 pandemic. Data stems from the EU survey on the use of Information and Communication Technologies (ICT) in households and by individuals and is available in eurostat table isoc_ci_ac_i.

B.2 Course content

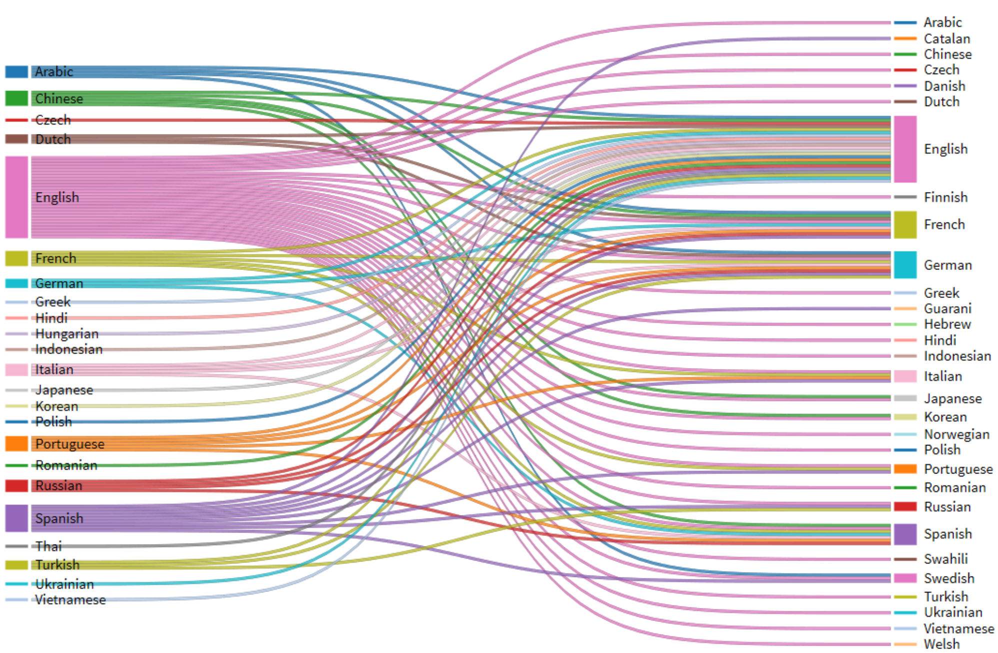
Figure 2: Tasks on Duolingo



Notes: Example of typical tasks on Duolingo for the English to Spanish course.

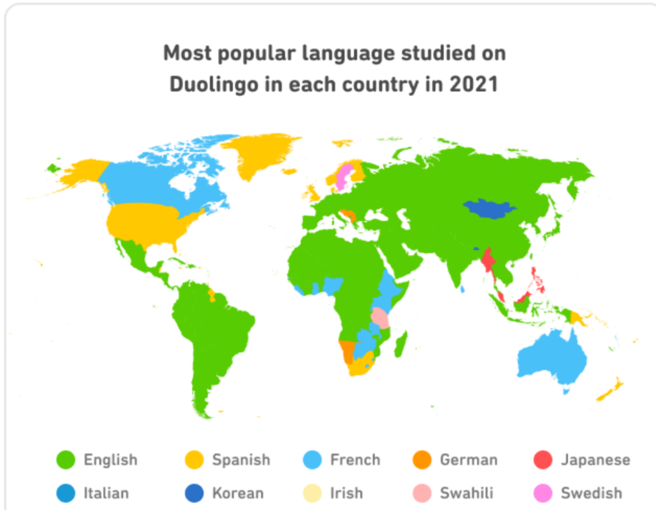
B.3 Duolingo Courses

Figure 3: Available Courses as of 2022



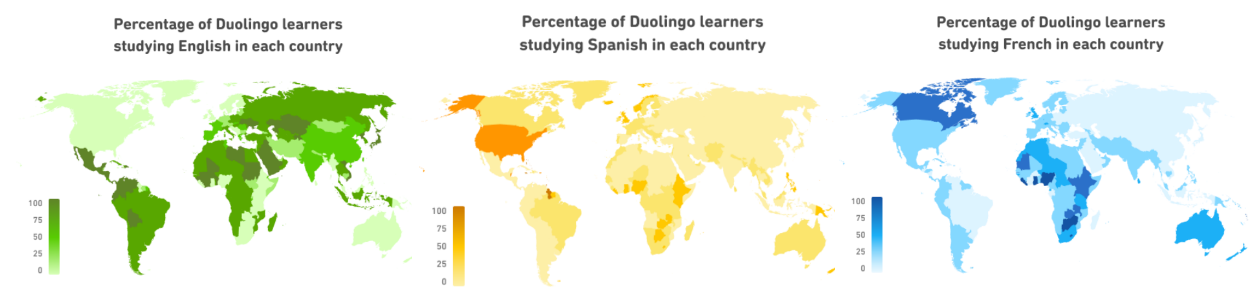
Notes: Sankey diagram of all available courses on Duolingo. In total, this comprises 84 modules over 23 source languages and 30 target languages. Information on courses and rollout dates is obtained from Fandom and verified using the Duolingo website.

Figure 4: Most Studies Language by Country on Duolingo in 2021



Notes: Most studied language by country in 2021. Data from the 2021 Duolingo Language Report.

Figure 5: Percentage of Learners Learning English, Spanish, or French across the World in 2020



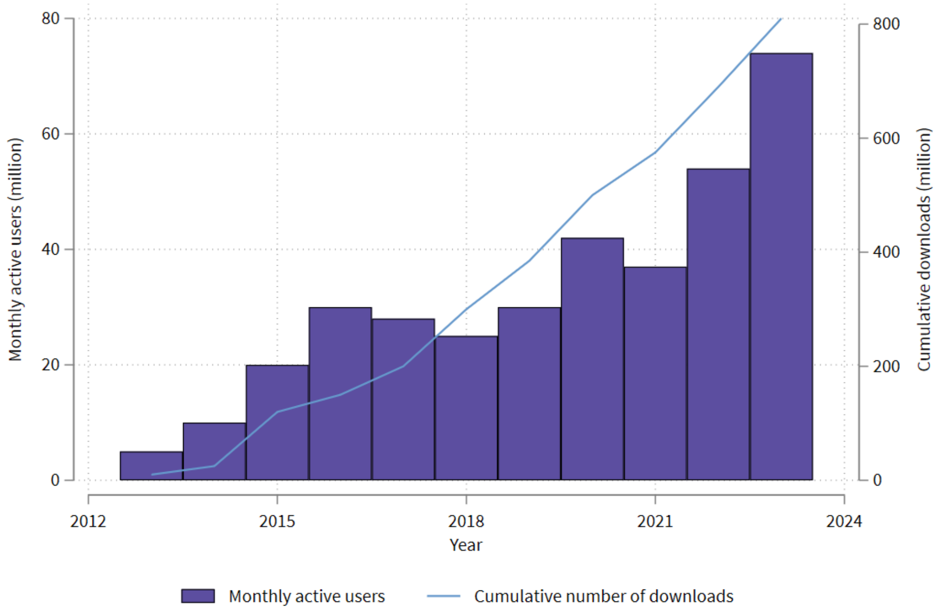
(a) English

(b) Spanish

(c) French

Notes: Most studied language by country in 2021. Data from the 2020 Duolingo Language Report.

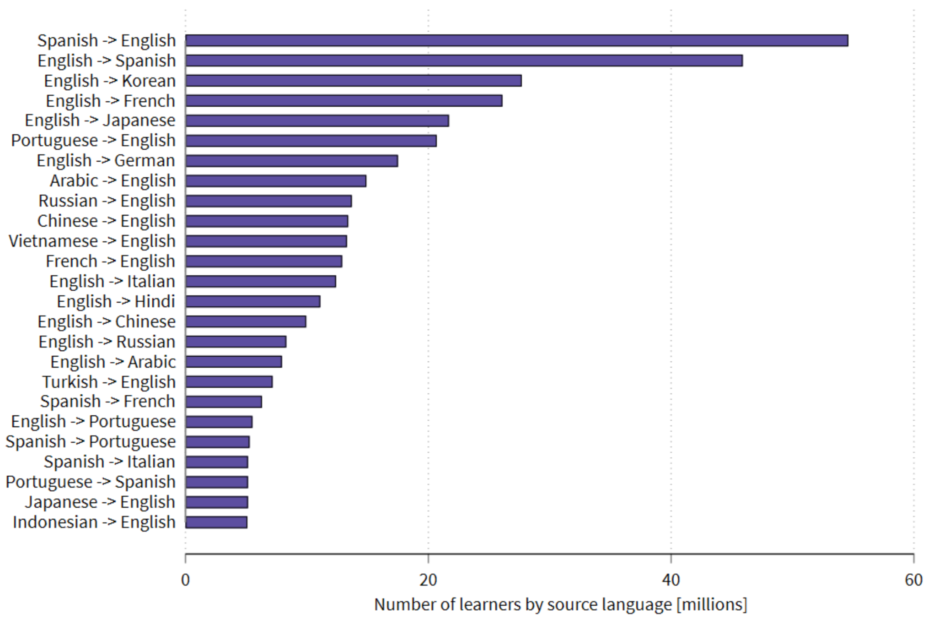
Figure 6: Monthly Active Users on Duolingo between 2012 and 2023



Notes: Monthly active users and the number of downloads of the Duolingo mobile application. For the definition of a monthly active users, see notes to Figure 1. Numbers on the Monthly Active Users and cumulative downloads are obtained from https://www.businessofapps.com/data/duolingo-statistics/.

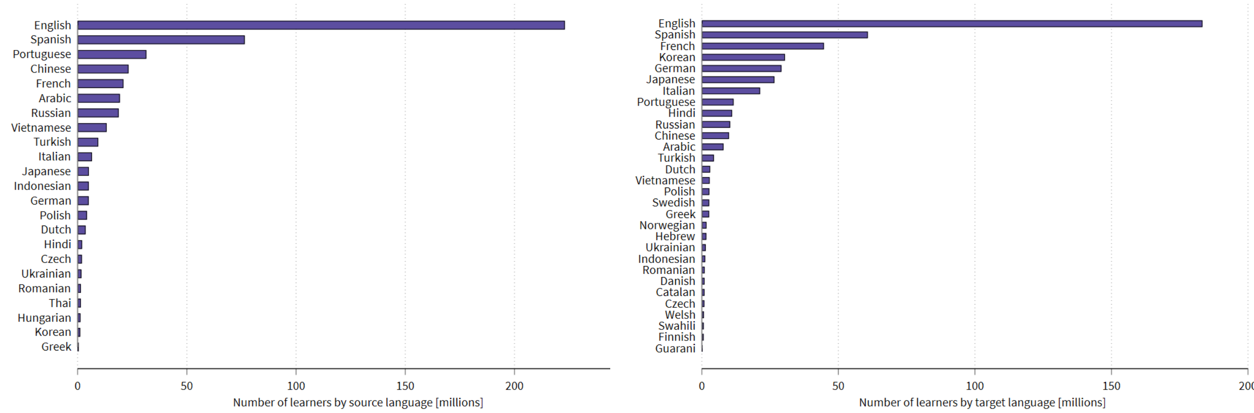
B.4 Uptake of Courses

Figure 7: Number of Learners by Course for 25 Most Popular Courses



Notes: Total numbers of learners by course, as indicated on the Duolingo website on October 6, 2022.

Figure 8: Number of Duolingo Learners by Source and Target Language

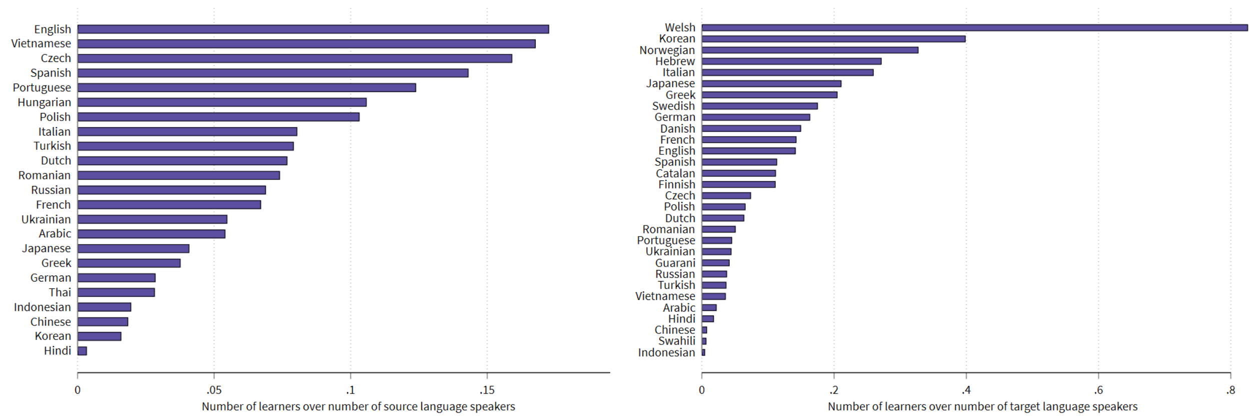


(a) Source language

(b) Target language

Notes: Total numbers of learners by (a) source and (b) destination language indicated on Duolingo on October 6, 2022. To calculate the total number of learners by source and target language, I sum over all courses. This sum represents the total number of instances someone started learning a specific target language from a specific source language on Duolingo. In practice, users may initiate multiple courses from the same source language, so the numbers in (a) are higher than the total unique individuals using a specific source language. Likewise, the numbers in (b) represent the total number of learner-language attempts for a specific target language.

Figure 9: Relative Number of Duolingo Learners by Source and Target Language



(a) Source language (b) Target language

Notes: See notes to Figure 8 for a description of the calculation of the total number of learners. To calculate the number of learners relative to speakers, I divide the numbers reported in Figure 8 by the total number of speakers by language from Ginsburgh, Melitz and Toubal (2017).

Table 1: Determinants of the Number of Learners of Duolingo Courses

	(1)	(2) Users	(3)
Source language speakers	0.008*** (0.001)	0.004*** (0.000)	
Target language speakers	0.008*** (0.001)	0.004*** (0.000)	
Source speakers × Target speakers (100 million)		0.001*** (0.000)	0.002* (0.001)
Observations	84	84	52
Source and Target FE			✓

Notes: OLS regressions of the number of learners on Duolingo, as measured of the number of learners by language course, on the number of speakers of the source and the target language. Column (3) introduces fixed effects on the source- and target language level, which drops 32 courses where either the source or target language is a singleton. Standard errors are clustered two-way on the source and destination language. Data on learners is obtained from the Duolingo platform in October 2022.

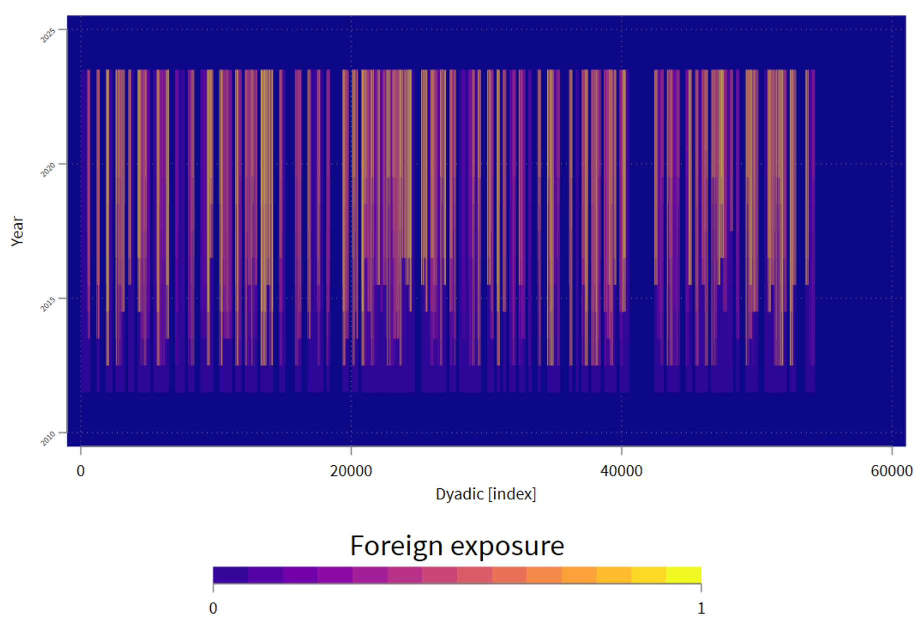
Table 2: Internet Traffic to Duolingo by Global Region

Global region	Share of traffic
North America	27%
South America	12%
Western Europe	11%
Eastern Europe	8%
Northern Europe	7%
East Asia	6%
Southern Europe	5%
South East Asia	5%
Central America	4%
Other	15%

Notes: Data has been obtained from Semrush (https://de.semrush.com/website/duolingo.com/overview/) in June 2024. Other includes Africa, Middle East and Turkey, and Oceania. Traffic from these regions is too low to analyze in isolation, but together accounts for about 15% of all traffic.

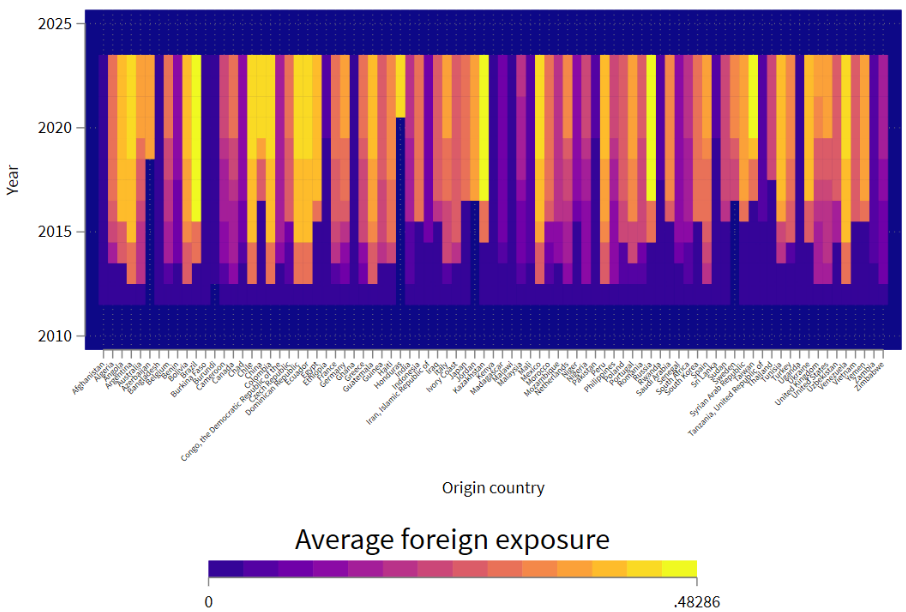
B.5 Visualizing Duolingo Exposure

Figure 10: Foreign Duolingo Exposure by Directed Country Pair (2012-2023)



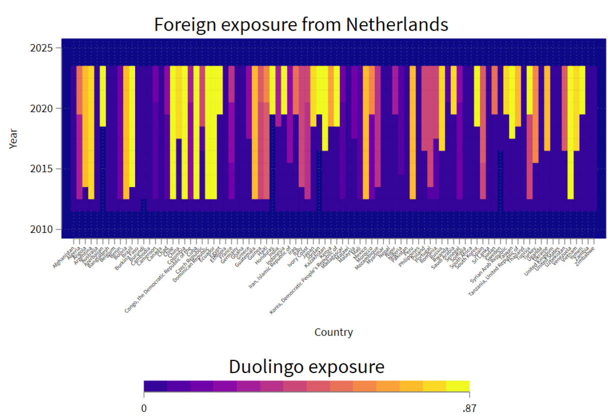
Notes: Dyadic foreign exposure to Duolingo DL_{odt} between 2012 and 2023. Brighter colors indicate a larger exposure.

Figure 11: Average Foreign Duolingo Exposure by Origin Country (2012-2023)

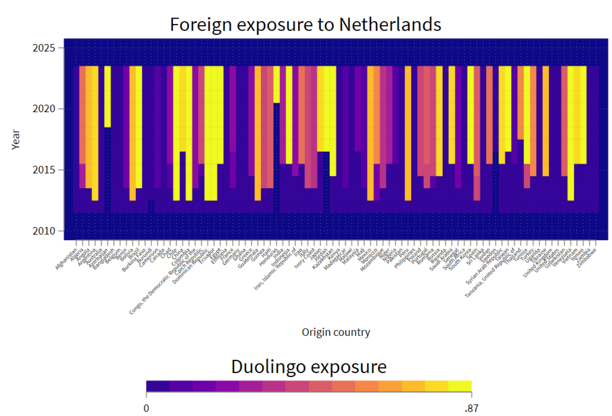


Notes: Simple-weighted foreign exposure to Duolingo $\frac{1}{N}\sum_d DL_{odt}$ between 2012 and 2023 across origin countries. Brighter colors indicate a larger exposure.

Figure 12: Variation in Foreign Duolingo Exposure to and from the Netherlands (2012-2023)



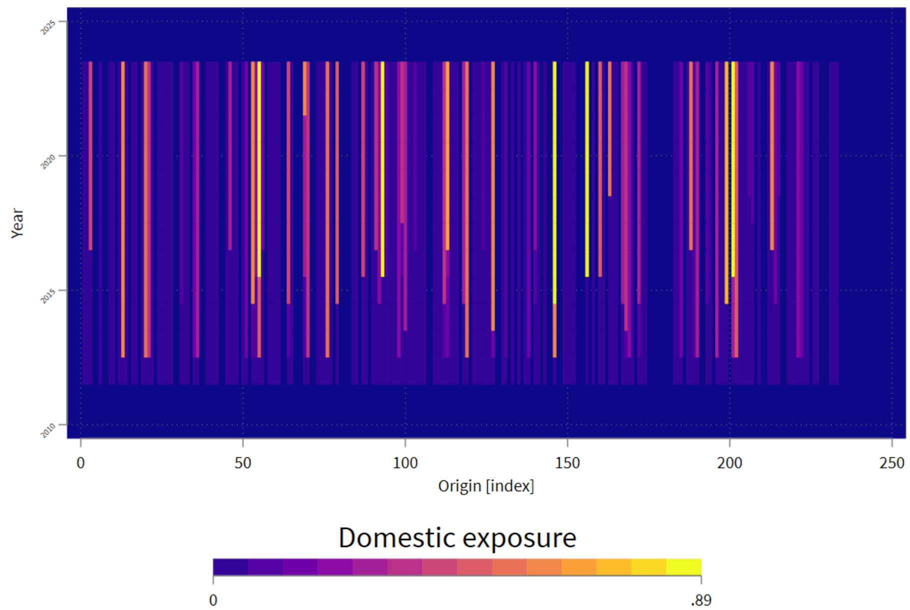
(a) From the Netherlands



(b) To the Netherlands

Notes: Foreign exposure to Duolingo within an (a) origin and (b) destination country (the Netherlands) to/from all other countries with more than 10 million inhabitants (for legibility). The Netherlands is used as an illustration as it features multiple spoken language to and from which courses are rolled out at different times. Hence, there is variation in treatment timing across (a) destinations and (b) origins and in several instances treatment changes more than once. Brighter colors indicate a larger exposure.

Figure 13: Domestic Duolingo Exposure by Origin Country over Time (2012-2023)



Notes: Dyadic domestic exposure to Duolingo DL_{oot} between 2012 and 2023. Brighter colors indicate a larger exposure.

C Google Trends

Google is the most-used search engine, with a global market share of about 90% between 2009 and 2024 (Allcott et al., 2024; Statcounter, 2024). Google provides *Google Trends*, a platform which enables users to query the relative search interest of a search term relative to all search activity on *Google Search*. Users can query the Google Trends Index (GTI), which is a measure of relative search intensity for a search term (i) by region for a given time period or (i) over time for a given region.[41][42] Importantly, it is not possible to directly query the relative search interest in a search term over time across countries. The relative search interest is normalized to 100 for the highest relative intensity within a query of type (i) or (ii), and all other data points get an integer score 0-100 relative to the highest relative intensity.

In the following, I discuss how, despite these limitations, a panel dataset measuring relative search intensity can be constructed. I denote (i) the interest by region as $GTI^{\widetilde{T}}_{o(2006-2022)}$ and (ii) the interest over time for a given geographic region (e.g. the whole world, a country, or a subnational region) as $GTI^{\widetilde{T}}_{\tilde{o}t}$. Here, the available regions are 240 countries and territories. T is the term or topic, o is the geographic region of interest, and t is the time period of interest. The variables with a tilde indicate that search interest is not scaled *across* that dimension, whereas the absence of a tilde indicates that it is not scaled. For example, $GTI^{\widetilde{T}}_{\tilde{o}t}$ is obtained through querying the GTI for every combination of origin region and search term. Hence, it is normalized to 100 for every origin-search term combination, and is uninformative about the relative search

[41]Search terms can be simply a set of words, or a Topic. Google Trends *topics* are language-agnostic and include synonyms and common misspellings. This has the large advantage that it captures search behavior of without the need of translating and accounting for different grammatical forms.

[42]The temporal frequency available on Google Trends depends on the period of interest. If the period of interest is more than 5 years, the frequency is monthly. If it is between 8 months and 5 years, it is weekly. If it is shorter than 8 months, the frequency is daily. As I am interested in longer periods, for my purpose the monthly frequency suffices and I always query time series between 2006 and 2022.

interest across origins and terms. Using the interest by regions and across time for the same search terms, geographic areas and time period, one can construct and index that is normalized across geographic regions and time: $GTI^{\widetilde{T}}_{ot} = \frac{1}{100} * GTI^{\widetilde{T}}_{\widetilde{o}t} \times GTI^{\widetilde{T}}_{o(2006-2022)}$. This enables us to compare relative search intensity across regions over time.[43] To proxy relative search interest for the search term Duolingo and its commonly used transliterations (in Arabic, Cyrillic, Japanese, Korean, and Mandarin scripts), I use $GTI^{Duolingo}_{ot}$. To proxy relative search interest for languages, I use $GTI^{\widetilde{T}}_{ot}$, which is *not* scaled across languages. The reason for this is that if one would scale across languages, the full variation in GTI^{T}_{ot} would be driven by languages with large absolute levels of search interest, limiting variation to mostly English and Spanish. I query search interest for all Topics that are used in Duolingo courses, all other languages with more than 50 million speakers according to the 2022 Ethnologue, and the following smaller European languages: Bulgarian, Lithuanian, Albanian, Latvian, Estonian, Slovak, Slovenian, Serbian, and Croatian. This gives a sample of 65 languages. Although I do not use it for the event studies in section 5, in the following section I explain how to obtain the interest scaled across search terms GTI^{T}_{ot} using a process called *Anchorbanking*.

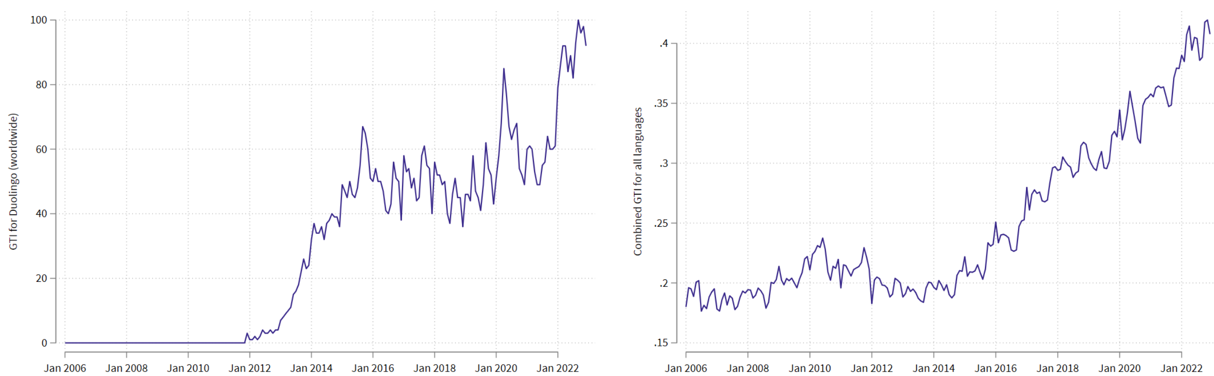
Anchorbanking across terms and topics

As one can query up to five terms or topics, one can identify the normalized relative search intensity across these terms or topics. In case one has a set of terms or topics $\mathcal{T}$ exceeding five, one has to query the search terms in overlapping folds of 5 (the first fold includes the first five terms, the second fold includes the fifth to the ninth term). After querying, one rescales the GTI of the terms from the second fold onwards using the ratio of the time-averaged Indices of the overlapping term in the first over that in the second fold, and repeats this procedure for all subsequent folds. After this procedure all the GTIs are normalized to the highest value in the first fold. Furthermore, as long as one has topics and terms all across the distribution of relative search intensity, this also allows circumventing the rounding problem between terms. For this it is not sufficient to just query with overlap, but also to re-order after rescaling and repeating the procedure of querying with overlap and rescaling. For my purpose, repeating this process two or three times suffices to obtain a distribution of GTI^{T}_{ot} that barely changes upon another repetition.

Figure 1a shows the global interest in Duolingo over time $GTI^{Duolingo}_{world,t}$. Before its existence, search interest was virtually zero, increasing rapidly between 2013 and 2015. Figure 1b shows the aggregated interest in languages using the sum of the fully scaled search interest $GTI^{languages}_{ot} = \sum_{T} GTI^{T}_{world,t}$, showing a strong increase from 2016 onwards.

[43]This approach is still limited by the rounding of GTI on integers. Hence, regions with less interest of around two orders of magnitudes smaller than the most interested region are strongly subject to rounding errors. This may lead to noisy results for low interest regions.

Figure 1: Worldwide Google Trends Index for Duolingo and Languages over time



(a) Duolingo

(b) Languages

Notes: Global relative search intensity for (a) Duolingo and its transliterations and (b) relative search intensity for all 65 queried languages. Data obtained by repeatedly querying Google Trends.

D Additional results

D.1 EU Adult Education Survey

Table 1: Descriptive Statistics of AES

	All		Natives		Immigrants	
	mean	s.d.	mean	s.d.	mean	s.d.
Age	43.95	12.96	44.20	13.10	42.24	11.85
Male	0.50	0.50	0.50	0.50	0.49	0.50
Secondary educated	0.43	0.50	0.45	0.50	0.34	0.47
Tertiary educated	0.34	0.47	0.34	0.47	0.32	0.47
Number of non-native languages spoken	1.00	0.99	1.01	1.00	0.93	0.95
Number of non-native languages spoken (intermediate level)	0.50	0.75	0.51	0.77	0.47	0.64
Number of non-native languages spoken (advanced level)	0.23	0.57	0.22	0.57	0.28	0.54
$\frac{1}{30}\sum_T DL^m_{T,2023}$	4.93	6.75	5.02	6.81	4.38	6.35
Observations	675,726		604,648		71,078	

Notes: Descriptive statistics of the full, native and immigrant sample of the EU AES.

Table 2: Robustness tests

	(1) Unweighted	(2) Mother tongues with source language	(3) Country-year FEs	(4) Respondent FEs	(5) Country-cohort-target language FEs
DL^{m}_{Tt}	0.020***	0.013***	0.020***	0.024***	0.012***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.003)
Observations	19,664,985	14,382,749	19,599,001	19,664,985	19,664,926
R^2	0.412	0.372	0.399	0.421	0.432
Mean dep. var.	0.036	0.032	0.035	0.035	0.035

Notes: OLS estimates of language proficiency on a Duolingo course availability and three-way fixed effects. The specification follows column 1 of Table 2 with several adjustments. For details on the specification and data, see notes to Table 2. Column 1 reports results without regression weights, column 2 drops all respondents speaking languages not featured as a source language on Duolingo, column 3 includes country-year fixed effects, column 4 includes respondent fixed effects and column 5 includes country-year of birth-target language fixed effects. Standard errors reported in parentheses are clustered on the language pair level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Alternative ways of treatment assignment

	(1)	(2)	(3)	(4)
DL^{m1}_{Tt}	0.018***			
	(0.005)			
$DL^{b,adv}_{Tt}$		0.016***		
		(0.002)		
$DL^{b,int}_{Tt}$			0.011***	
			(0.002)	
Observations	21,553,104	21,553,104	21,553,104	
R^2	0.388	0.389	0.388	
Average dependent variable	0.036	0.036	0.036	

Notes: OLS estimates of language proficiency on a Duolingo course availability and three-way fixed effects. The specification follows column 1 of Table 2 with different treatments. For details on the specification and data, see notes to Table 2. Column 1 reports results using a treatment that is one if a Duolingo course is available from the first mother tongue only, column 2 for both mother tongues (our baseline estimate), column 3 for the mother tongues and other languages spoken at advanced level and column 4 for the mother tongues and other languages spoken at least at intermediate level.

Table 4: Heterogeneity across individual characteristics

	(1) 44 and below	(2) 45-64	(3) 65 and above	(4) Male	(5) Female
DL^m_{Tt}	0.016***	0.016***	0.006*	0.013***	0.017***
	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)
Observations	9,996,531	9,443,217	1,616,956	10,221,485	11,329,986
R^2	0.448	0.362	0.361	0.390	0.389
Average dependent variable	0.041	0.031	0.029	0.035	0.036
	Primary	Secondary	Tertiary	Immigrant	Native
DL^m_{Tt}	0.012***	0.012***	0.014***	0.022**	0.014***
	(0.004)	(0.004)	(0.004)	(0.009)	(0.003)
Observations	5,351,456	9,493,540	6,690,067	2,134,806	18,207,434
R^2	0.255	0.388	0.522	0.302	0.409
Average dependent variable	0.017	0.034	0.052	0.034	0.036

Notes: OLS estimates of language proficiency on a Duolingo course availability and three-way fixed effects for subsamples of respondents. The specification follows column 1 of Table 2. For details on the specification and data, see notes to Table 2. Panel A reports results for three age bins and for men and women separately, Panel B report the results for three levels of educational attainment and for those born abroad and in the country of interview. Standard errors reported in parentheses are clustered on the language pair level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Heterogeneity over target languages

	(1) Natives	(2) Immigrants
DL^m_{Tt}	0.011***	0.012**
	(0.003)	(0.005)
DL^m_{Tt} × English	0.026	0.112*
	(0.022)	(0.058)
DL^m_{Tt} × French	-0.010	0.012
	(0.020)	(0.013)
DL^m_{Tt} × German	0.028*	0.007
	(0.016)	(0.019)
DL^m_{Tt} × Spanish	0.015*	0.014
	(0.008)	(0.019)
DL^m_{Tt} × Non-EU	-0.007**	0.001
	(0.003)	(0.005)
Observations	18,207,434	2,134,806
R^2	0.409	0.303
Average dependent variable	0.036	0.034

Notes: OLS estimates of language proficiency on a Duolingo course availability and three-way fixed effects. The specification follows column 1 of Table 2, interacting the treatment with indicators for the target language (group). For details on the specification and data, see notes to Table 2. Column 1 reports results for natives only and column 2 for immigrants only. Standard errors reported in parentheses are clustered on the language pair level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

D.2 TOEFL and GRE Test Scores

Figure 1: The Effect of Duolingo Rollout on Component Scores of the English language (TOEFL) test (2007-2021)

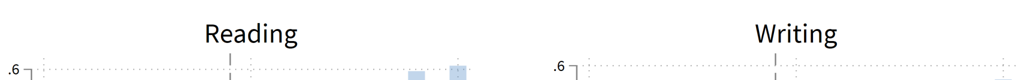
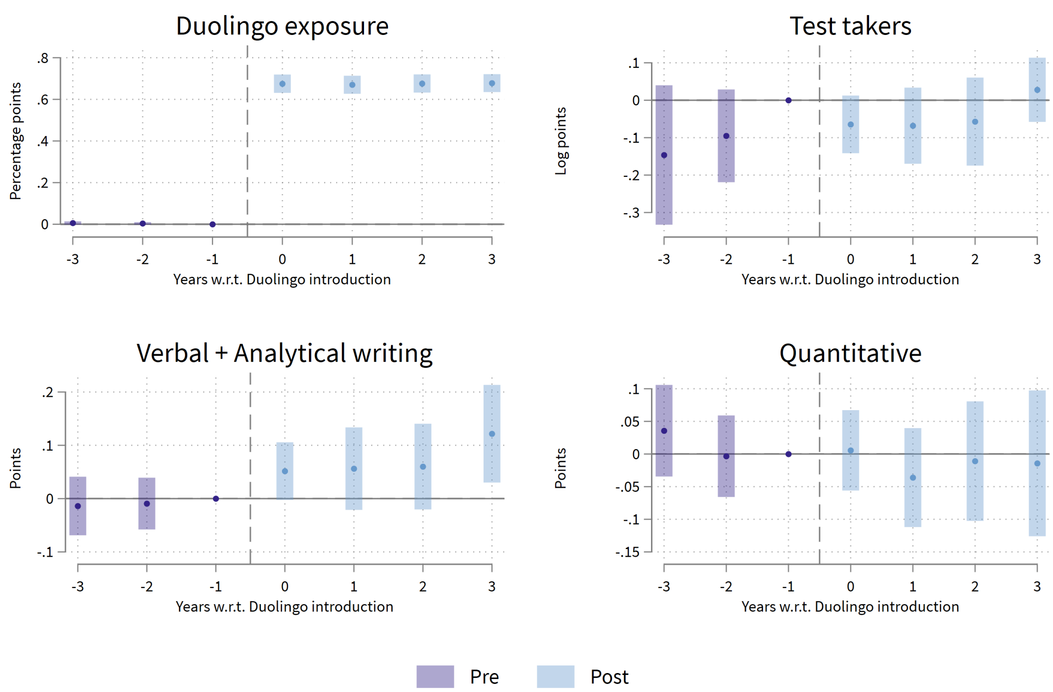


Table 6: The effect on language skills of immigrants

	(1)	(2)	(3)	(4)	(5)	(6)
	Host country language			Other languages		
	Spoken	Intermediate	Advanced	Spoken	Intermediate	Advanced
DL^m_{Tt}	0.076***	0.011	0.011	0.022**	0.009***	0.004**
	(0.016)	(0.011)	(0.010)	(0.009)	(0.003)	(0.002)
Observations	79,077	79,077	79,077	1,892,002	1,892,002	1,892,002
R^2	0.632	0.458	0.290	0.299	0.246	0.168
Mean dep. var.	0.487	0.281	0.181	0.035	0.018	0.011

Notes: OLS estimates of language proficiency on a Duolingo course availability and three-way fixed effects. The specification follows column 1-3 of Table 2 with several adjustments. For details on the specification and data, see notes to Table 2. Columns 1-3 of this table report results only for host country native languages and columns 4-6 report results for all other languages. Standard errors reported in parentheses are clustered on the language pair level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 2: The Effect of Duolingo Rollout on GRE Test Takers and Scores (2011-2022)



Notes: Results from Wooldridge (2023) event study estimators around increases in Duolingo exposure exceeding 20 percentage points. The panels report results using the following specifications and outcomes: a linear model of the Duolingo exposure as an outcome (upper left), (upper right) a Poisson model of the number of test takers, (lower left) a linear model of average scores for the verbal and analytical writing part of the GRE test and (lower right) a linear model of average scores for the quantitative parts. Effect sizes in the bottom two figures are standardized. The number of test takers and scores are reported as averages by country of citizenship. N = 1,510 from 148 distinct origin countries, of which 89 are treated. Shaded blue bars indicate 95% confidence intervals based on cluster-robust standard errors at the country of citizenship level. Data is obtained from the GRE annual reports between 2011 and 2022.

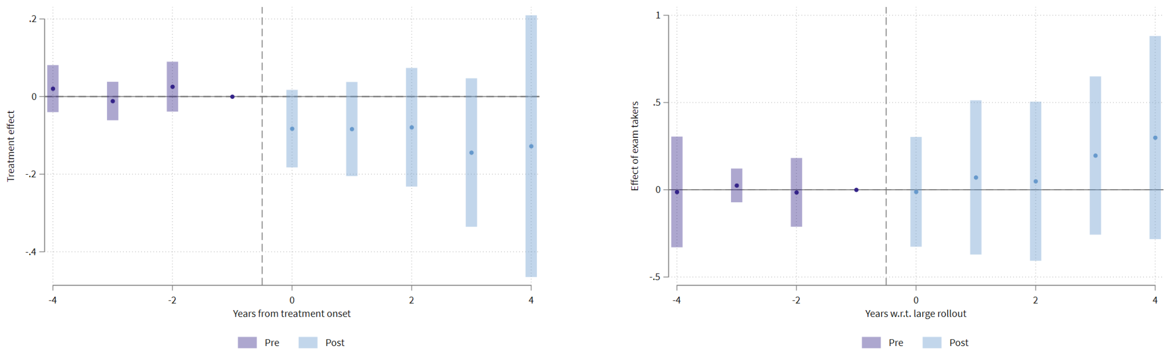
D.3 Does Duolingo crowd out traditional adult language learning?

It is a priori unclear how the introduction of Duolingo courses affect traditional language learning. On the one hand, potential learners may use Duolingo instead of traditional in-class language. On the other hand, Duolingo may spur language learning at basic levels and generate interest in destination language countries and culture and increase in-class course participation, particularly at higher proficiency levels. To provide some evidence on this, I turn to data on the number of German language course and exam takers at German language learning institutes (*Goethe* institutes) across more than 90 countries outside of Germany. By combining information from these language learning institutes with the staggered rollout of 9 Duolingo courses between 2014 and 2020, I can study whether the availability of low-cost language learning affected German language course and exam participation.

Participants of courses at Goethe institutes are predominantly high-skilled young adults (77% are aged 35 or below), almost 60% are female, and almost half are still in education. Education and cultural interest are the predominant reasons for learning, and only a minority share for migration-related reasons (Huber, Sommerfeld and Uebelmesser, 2022). I obtain the number of exams and course takers by country from Uebelmesser, Sommerfeld and Weingarten (2022) from 2007-2014 and collect the same information from Goethe's yearly reports between 2016 and 2022 on the global region level. In the latter, data is aggregated for 12 global regions in the latter, who house about 12 Goethe institutes each on average. Using the sum of registrations and exam takers by country in the year 2014, I construct a matrix X_{rc} of weights of every country c in every region r. Using this matrix and the time-varying Duolingo exposure by origin country for German, I construct a weighted exposure to Duolingo courses on the global region level. To study the impact of Duolingo availability on the number of course participants and test takers, I estimate a Poisson model event study using the Wooldridge (2023) estimator using increases in the Duolingo exposure of more than 20 percentage points. The results are reported in Table 3.

Although imprecisely estimated, the results in Figure 3 indicate that introduction of a relevant Duolingo course decreases the number of registrations, but the number of exams is even increasing if anything. This is consistent with language learners substituting Goethe courses with learning on Duolingo, but exams to a lesser extent. This is not surprising, given that Germany requires proof of language skills for some types of residence, and German higher education institutes for admission.

Figure 3: The Effect of Duolingo on Institutional German Learning



(a) Course participants

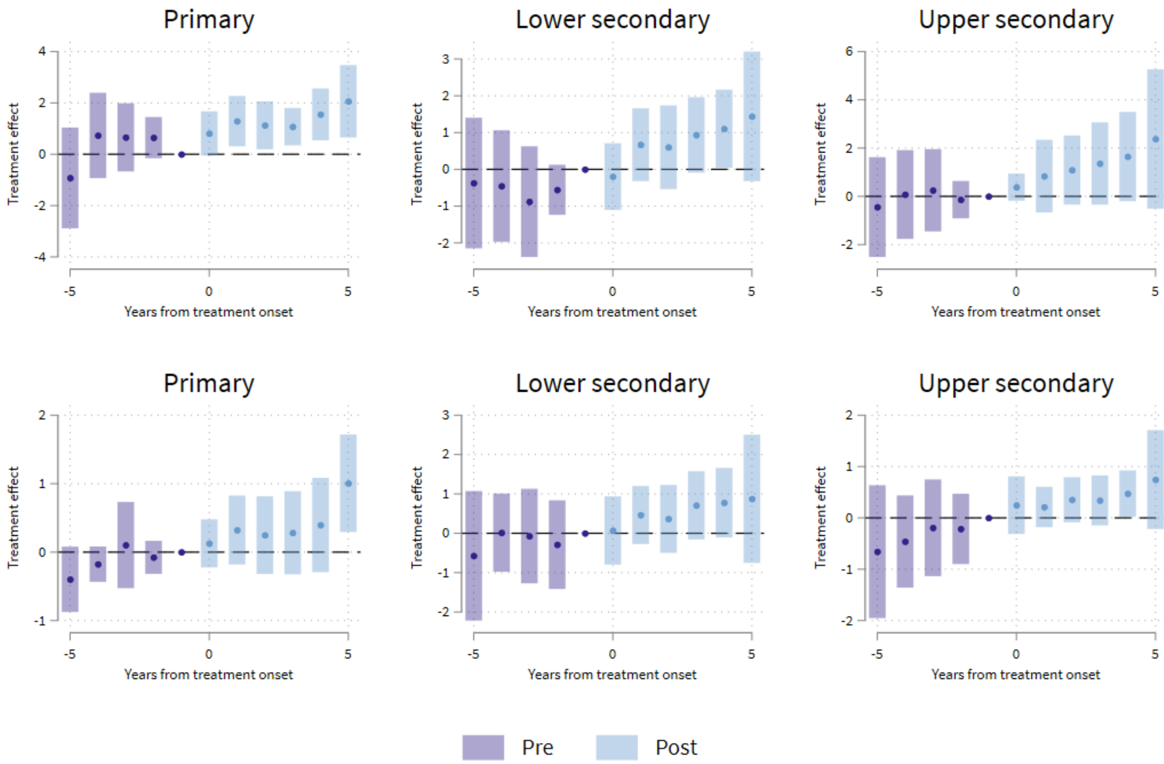
(b) Exam takers

Notes: Results from Wooldridge (2023) event study PPML estimators around increases in average region-level Duolingo exposure exceeding 20 percentage points. The text provides a discussion about how Duolingo exposure is constructed on the regional level. Standard errors are clustered at the origin region level. N = 192 with 12 unique origin regions, of which 6 are treated. Data on participants and test takers originate from Uebelmesser, Sommerfeld and Weingarten (2022) and the yearly reports (*Jahrbuch*) of the Goethe Institute.

D.4 Does low-cost language learning affect in-school instruction in the EU?

The availability of low-cost language learning may affect in-class instruction in various ways. First, the availability of a language course may foster Second, it may induce interest in the particular language among pupils, who continue to study the language further in class. To study whether these effects are at play, I resort to data on foreign language learning across the EU. The EU consistently reports numbers on the share of pupils learning specific foreign languages for three stages of education. English is by far the most learned foreign language across the EU: 73% of pupils learn it in primary education, 92% in lower secondary and 82% in upper secondary education. Hence, I show results with and without English included. Figure 4 shows the results. The upper panel suggests that across levels Duolingo exposure increases the number of pupils. Without English, the results are smaller in magnitude, but results still seem to be positive.

Figure 4: The Effect of Duolingo Courses on In-school Language Learning

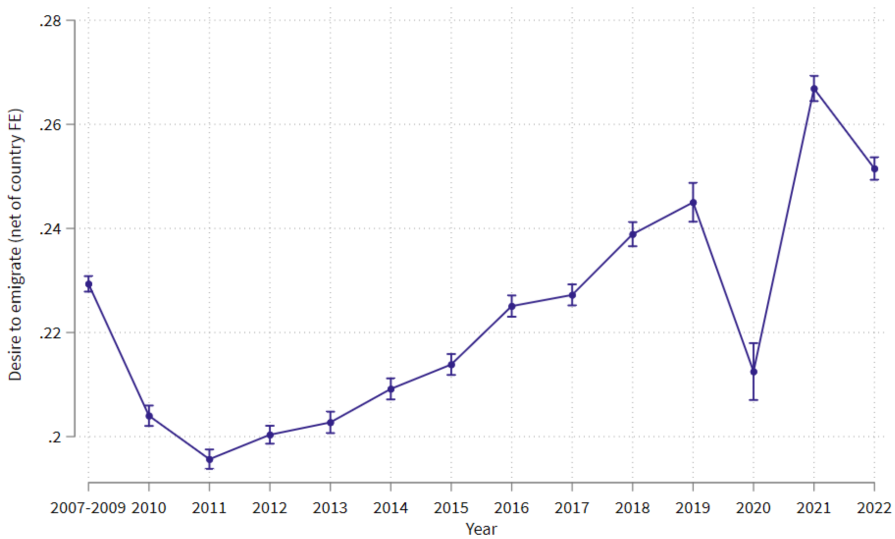


Notes: Results from linear Wooldridge (2023) event study estimators around increases in Duolingo exposure exceeding 50 percentage points. The Duolingo exposure is constructed by the maximum of the share of native-language speakers of the source language of a Duolingo course. I use the share of native language speakers in this case as school-aged children are unlikely to speak foreign spoken languages already. The dependent variable is the share of pupils studying a specific language. The columns report results for primary, lower secondary and upper secondary students separately. The upper row presents results including English, the bottom row excluding English. N = 9,890 from 34 countries and 28 languages, 879 country-language pairs of which 100 treated in the upper panel, N = 9,458 from 34 countries and 27 languages, 846 country-language pairs of which 84 treated in the bottom panel. Shaded blue bars indicate 95% confidence intervals based on two-way cluster-robust standard errors at the country and language level. Data is obtained from Eurostat table educ_enrl1tl for 2007–2012 and educ_enrllng1 for 2012–2022. In case a country-language pair has an observation in both in 2012, I take the simple unweighted average between both.

D.5 Migration Intentions

D.5.1 Pattern over time

Figure 5: Share of GWP Respondents Desiring to Emigrate



Notes: Share of individuals answering positively to the question on desiring to emigrate over time, net of country fixed effects. The graph reports coefficients of a regression of a dummy for desiring to emigrate on country fixed effects and year fixed effects. I report the year fixed effects, adding back the constant. Because of the limited number of countries visited by GWP in the first years, I aggregate the years 2007–2009 in a single category. 95% confidence intervals are reported based on standard errors are clustered at the country level. Data are from the Gallup World Polls between 2007 and 2022. N = 2,014,359.

D.5.2 Control Function Approach

Although it is plausible that Duolingo courses were not rolled out in anticipation of trends in international migration, one may still be concerned about the presence of unobserved confounding between the availability of DDuolingo courses and bilateral migration (intentions) over time. To further mitigate such concerns, I construct two distinct instruments for foreign exposure to Duolingo and re-estimate the main results using a control function approach.

As demand for language learning requires source language learners being interested in target languages, Duolingo's incentives to roll out languages courses are increasing in the product of the source- and target language size. To see this, assume there are two distinct source languages S_1 and S_2 with 2 and 1 million speakers respectively and two distinct target languages T_1 and T_2 with 2 and 1 million speakers respectively. I assume that the probability that a source language speaker is interested in a target language is proportional to its applicability, e.g. the number of speakers. Consequently, the potential market size for a $S_1 \rightarrow T_1$ course is four times as large as a $S_2 \rightarrow T_2$ course. Hence, the potential market size of a course varies on the bilateral level and is an increasing function of the product of the number of speakers. As witnessed in Table 1, Duolingo indeed prioritized courses between languages with many speakers, after controlling for the number of source- and target language speakers. Moreover, the slope of the number of learners by course per source language speaker is also increasing in the number of target language speakers, as witnessed in Table

1. Hence, I instrument the rollout of Duolingo courses by the interaction between the log number of speakers of the source and target language, interacted with a linear time trend starting in 2012 (the introduction year of the first courses). I opt to construct the instrument using the log of speakers rather than the the absolute number, as otherwise the vast amount of variation is driven by country-pairs speaking the largest two languages: one speaking English and the other speaking Mandarin.

$$Z_{STt} = \begin{cases} Z_{STt} = log(N^S)log(N^T)t & \text{if } t \geq 2012 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The second instrument is based on Duolingo's propensity to roll out courses to the same source and target languages as existing courses. For example, if Duolingo rolls out courses from a specific source languages, they have language-specific knowledge that facilitates the development of further courses using that language as a source language. The same logic applies to target languages. To construct a bilateral instrument for course roll-out, I interact the source language and target language propensity, omitting the focal source- and target language. The latter would generate a mechanical correlation between the instrument and the Duolingo exposure.

$$Z_{STt} = \begin{cases} \left(\sum_{S' \neq S} Duolingo_{ST}\right) \left(\sum_{T' \neq T} Duolingo_{ST}\right) t & \text{if } t \geq 2012 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Here, $Duolingo_{ST}$ is a binary indicator for whether there is a Duolingo course from language S to language T by 2022. In both cases, I aggregate the instrument in similar vein as the main exposure to Duolingo:

$$Z_{odt} = \max_{S,T} \alpha_{oS} \alpha_{dT} Z_{STt} \tag{12}$$

Using either instrument, I estimate the following linear first stage:

$$DL_{odT} = Z_{odt} + \phi_{ot} + \theta_{dt} + \psi_{od} + \epsilon_{odt} \tag{13}$$

As the gravity model is a non-linear model, naive 2SLS estimation is invalid. Instead, one should resort to a control function procedure (Wooldridge, 2015). To implement this, I estimate equation 13, obtain the residuals $\hat{\epsilon}_{odt}$, and include these residuals in the baseline gravity model. Any variation in the potentially endogenous regressor unexplained by the instrument and the fixed effects is absorbed by this residual. Hence, inclusion of the residual in the second stage controls for the endogenous variation.

The remaining identification assumption behind the first control function strategy is that trends in migration aspirations between countries both speaking widely spoken languages have followed similar trends than between countries where one of the languages spoken is smaller, after controlling for three-way fixed effects. This implies that differential trends for countries speaking source- and target languages of varying sizes are absorbed by the origin-year and destination-year fixed effects. The remaining identification assumption behind the second control function approach is that outcome trends would have followed similar paths between country pairs both speaking languages that have other Duolingo courses than country pairs that do not, in absence of strong Duolingo exposure between the treated country pair.

Results. Table 7 shows the results of the control function approach. Columns 1 and 3 shows that both instruments strongly predict Duolingo exposure, but that the language size instrument is particularly strong.

Moreover, I find that the IV results are in line with the OLS results, only somewhat larger. The results suggest that, if anything, the endogeneity bias is negative, and that the effect size exceeds 80%.

Table 7: Control Function Estimates of the Effect of Duolingo of Migration Aspirations

	(1)	(2)	(3)	(4)
	Language size		Course propensity	
	DL_{odt}	$\frac{M_{odt}}{M_{oot}}$	DL_{odt}	$\frac{M_{odt}}{M_{oot}}$
Z_{odt}	0.001***		3.609***	
	(0.000)		(0.776)	
DL_{odt}		0.602***		1.141***
		(0.189)		(0.366)
Control function		-0.351*		-0.855**
		(0.191)		(0.367)
Observations	121698	121477	121698	121477
Unique origin countries	153	153	153	153
Unique destination countries	196	196	196	196
Unique dyads	10,527	10,527	10,527	10,527
Kleibergen-Paap F-statistic	444.3		21.7	
Origin-destination FE	✓	✓	✓	✓
Origin-year FE	✓	✓	✓	✓
Destination-year FE	✓	✓	✓	✓
Estimator	OLS	PPML	OLS	PPML

Notes: Control function estimation of the effect on Duolingo exposure on migration odds for two different instruments. Odd columns report estimates from a linear first stage and even columns report the results of the three-way gravity model estimated by PPML. Columns (1) and (2) estimate this procedure using the instrument based on language size as described in equation 10 and columns (3) and (4) estimate this procedure using the instrument based on the propensity to roll out languages to particular courses as described in equation 11. For notes on the data and estimation sample, see notes to Table 3. Standard errors reported in parentheses are two-way clustered: on the country of origin and country of destination level.

D.5.3 The Role of Language Requirements

As Duolingo effectively enables low-cost language learning without certification, this could reduce migration barriers more for destination countries without language requirements. (which typically require a costly certificate). I therefore draw on the MIPEX database, which provides information on migration policy for 56 destination countries. I retrieve the MIPEX indicator for language requirements for permanent residence (originally taking values 0, 50, or 100) and convert it to a 0-2 scale. Figure 8 shows the effect of the interaction of Duolingo exposure with the MIPEX indicator. I find that the effect is strongest for those countries without requirements, although the interaction is not statistically significant.

Table 8: Heterogeneity of Results by Destination-country Language Requirements

	(1) $\frac{M_{odt}}{M_{oot}}$	(2) $\frac{M_{odt}}{M_{oot}}$	(3) $\frac{M_{odt}}{M_{oot}}$
DL_{odt}	0.374***	0.237***	0.286***
	(0.080)	(0.080)	(0.097)
DL_{odt} × Permanent residence language requirements (0-2, MIPEX)			-0.056
			(0.042)
Observations	123263	41699	41699

PPML regressions based on the sample and specification of column 2 of Table 3. See notes to Table 3 for the estimation strategy, data and sample. Columns 2–4 restrict the sample to those 56 destination countries where MIPEX is available. Column 3 introduces an interaction between an indicator for language requirements for permanent residency, taking values 0 (no language requirements), 1 (some language requirements) or 2 (strict language requirements). The number of observations are slightly different from Table 3 because of 3 countries not speaking any language present on Duolingo required for Column 7 and 8. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

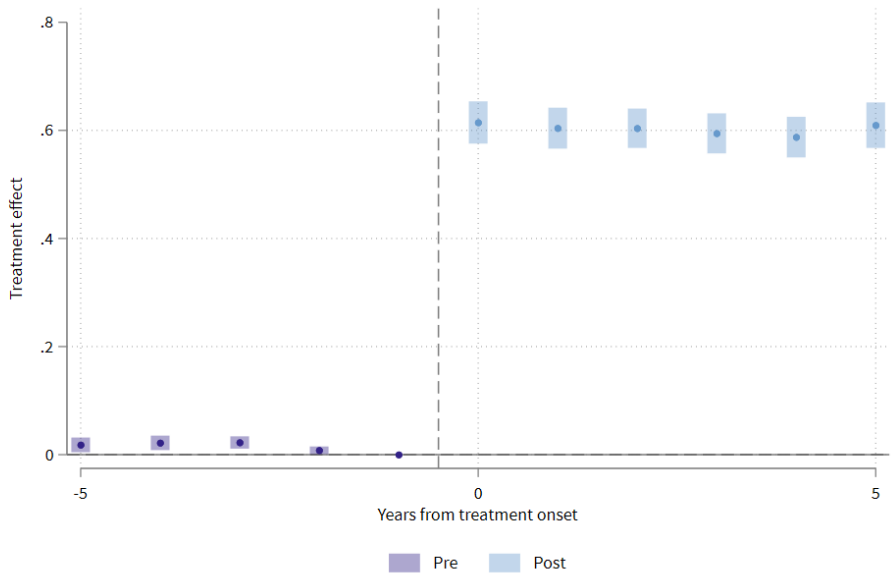
Table 9: The Effect of Duolingo on Total Emigration

	(1) $\frac{M_{ot}}{P_{ot}}$	(2) $\frac{M_{ot}}{P_{ot}}$	(3) $\frac{M_{ot}}{P_{ot}}$	(4) $\frac{M_{ot}}{P_{ot}}$
$DL_{ot}^{foreign}$	0.036	0.042		
	(0.037)	(0.037)		
$DL_{ot}^{domestic}$		-0.035*		
		(0.019)		
$DL_{ot}^{foreign}$ (weighted)			0.048**	0.047**
			(0.020)	(0.020)
$DL_{ot}^{domestic}$				-0.029
				(0.018)
Observations	1757	1757	1757	1757
Average dependent variable	0.219	0.219	0.219	0.219

Notes: OLS regression with country- and year fixed effects of migration rates on foreign- and domestic Duolingo exposure. Columns 1 and 2 use unweighted averages of Duolingo exposure, whereas columns 3 and 4 uses a weighted average measure of foreign exposure to Duolingo, using the bilateral stock of migrants in 2005 as weights. See notes to Table 3 for the sample and data. Standard errors reported in parentheses are clustered on the country of origin level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

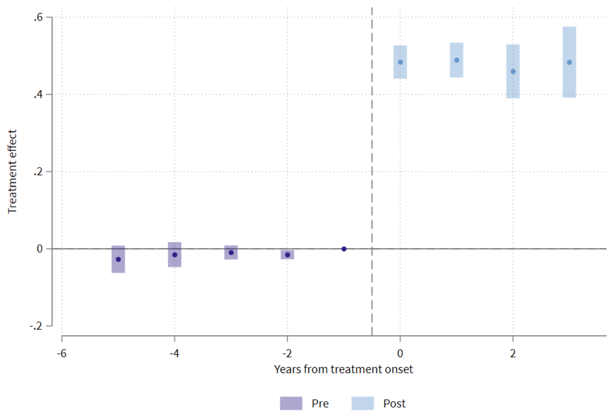
D.5.4 Event study

Figure 6: Change in Duolingo Exposure around Increases in Exposure Exceeding 50 pp

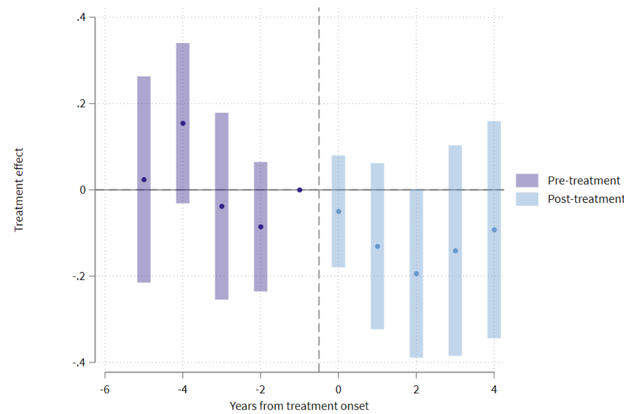


Notes: OLS regression of the heterogeneity-robust Wooldridge (2023)-estimator with three-way fixed effects of the treatment exposure on a binary indicator for whether a origin-destination pair has experienced an increase in foreign Duolingo exposure exceeding 50 percentage points. See notes to Table 3 for information on the data and sample.

Figure 7: Event Study around Large Increases in Domestic Duolingo Exposure



(a) Domestic Duolingo Exposure



(b) Migration Aspiration Odds

Notes: (a) OLS regression of the heterogeneity-robust Nagengast and Yotov (2023)-estimator with three-way fixed effects of the treatment exposure on a binary indicator for whether a country pair has experienced increase in domestic Duolingo exposure exceeding 50 percentage points. See notes to Table 3 for information on the data and sample. Standard errors reported in parentheses are two-way clustered: on the country of origin and country of destination level. (b) PPML regressions of the heterogeneity-robust event study estimator by Nagengast and Yotov (2023) of migration aspiration odds on a binary indicator for whether an origin country pair has experienced increase in domestic Duolingo exposure exceeding 50 percentage points, including origin-destination pair and destination year fixed effects. An event is defined as an increase in Duolingo exposure of more than 50 percentage points. Estimates are shown for the 5 years before and after an event. Shaded blue bars indicate 95% confidence intervals based on standard errors clustered at native language level. See notes to Table 3 for information on the data and sample.

D.5.5 Alternative Treatment Definitions

Table 10: Testing Monotonicity of the Effect

	(1)	(2)
$0 < DL_{odt} \leq 0.25$	-0.066	0.004
	(0.048)	(0.054)
$0.25 < DL_{odt} \leq 0.5$	0.024	0.158***
	(0.058)	(0.058)
$0.5 < DL_{odt} \leq 0.75$	0.132**	0.180**
	(0.055)	(0.070)
$0.75 < DL_{odt} \leq 1$	0.170***	0.297***
	(0.061)	(0.080)
$0 < DL_{oot} \leq 0.25$	0.011	
	(0.068)	
$0.25 < DL_{oot} \leq 0.5$	-0.139	
	(0.103)	
$0.5 < DL_{oot} \leq 0.75$	0.077	
	(0.117)	
$0.75 < DL_{oot} \leq 1$	-0.369***	
	(0.105)	
Observations	123263	123263
Origin-year fixed effects		✓

Notes: Gravity model estimated by PPML without (column 1) and with (column 2) origin-year fixed effects. See notes to Table 3 for information on the data and sample. Instead of numerical values for both treatments, these results include binary indicators for four bins of treatment intensity per exposure variable and omits the regressors reported in Table 3. Standard errors reported in parentheses are two-way clustered: on the country of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 11: Using Official Target Languages in the Destination

	(1)	(2)
DL_{odt}^{off}	0.167***	0.198***
	(0.056)	(0.061)
DL_{oot}^{off}	-0.161	
	(0.376)	
Observations	123655	123655
Origin-destination FE	✓	✓
Origin-year FE		✓
Destination-year FE	✓	✓

PPML regressions based on the sample and specification of column 2 of Table 3. See notes to Table 3 for the estimation strategy, data and sample. Exposure measures are calculated by only taking into account target languages that are official languages in the destination country. Standard errors reported in parentheses are two-way clustered: on the country of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

D.5.6 Robustness on Empirical Approach

Table 12: Different ways of clustering standard errors

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Level of clustering:	Pair		Origin & Destination		Main origin & destination language		Main Duolingo origin & destination language	
DL_{odt}	0.267***	0.373***	0.267***	0.373***	0.267***	0.373***	0.267***	0.373***
	(0.065)	(0.052)	(0.065)	(0.080)	(0.060)	(0.094)	(0.058)	(0.109)
DL_{oot}	-0.225***		-0.225		-0.225		-0.225	
	(0.075)		(0.160)		(0.157)		(0.169)	
Observations	123180	123180	123180	123180	123180	123180	123180	123180
Number of clusters	10641	10641	153	153	63	63	22	22
Number of clusters (2)			194	194	70	70	30	30
Origin-destination FE	✓	✓	✓	✓	✓	✓	✓	✓
Origin-year FE		✓		✓		✓		✓
Destination-year FE	✓	✓	✓	✓	✓	✓	✓	✓

PPML regressions based on the sample and specification of column 1 and 2 of Table 3. See notes to Table 3 for the estimation strategy, data and sample. Standard errors, reported in parentheses, are clustered on different levels. Columns 1 and 2 cluster on the country pair level, 3 and 4 on the origin- and destination level (as Table 3), 5 and 6 on the main origin- and main destination language and 7 and 8 on the main origin- and main destination language that is available in any course on Duolingo. The number of observations are slightly different from Table 3 because of 3 countries not speaking any language present on Duolingo required for Column 7 and 8. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 13: Omission of Exposure Contribution in Countries with Most Speakers by Language

	(1)	(2)	(3)	(4)	(5)	(6)
	Omission of contribution to exposure in:					
	Origins with most source speakers		Destinations with most target speakers		Both	
DL_{odt}	0.276***	0.421***	0.326***	0.346***	0.388***	0.408***
	(0.059)	(0.083)	(0.100)	(0.123)	(0.080)	(0.123)
DL_{oot}	-0.209		-0.254		-0.240	
	(0.155)		(0.169)		(0.162)	
Observations	123180	123180	123180	123180	123180	123180
Origin-destination FE	✓	✓	✓	✓	✓	✓
Origin-year FE		✓		✓		✓
Destination-year FE	✓	✓	✓	✓	✓	✓

PPML regressions based on the sample and specification of column 1 and 2 of Table 3. See notes to Table 3 for the estimation strategy, data and sample. Columns 1 and 2 includes alternative measures of Duolingo exposure, excluding the contribution of the source language in the origin country with most speakers, for every language. Columns 3 and 4 includes alternative measures of Duolingo exposure, excluding the contribution of the target language in the destination country with most speakers, for every language. Columns 5 and 6 includes alternative measures of Duolingo exposure, excluding both types of contributions. Standard errors reported in parentheses are two-way clustered: on the country of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 14: Controlling for Origin-destination-nest-year Fixed Effects

Nest:	(1) none	(2) WB region	(3) WB income	(4) EU members	(5) All three
DL_{odt}	0.365***	0.264***	0.346***	0.363***	0.231***
	(0.080)	(0.090)	(0.098)	(0.065)	(0.085)
Observations	111251	101005	103788	111198	95942
Number of Fixed Effects	14120	21500	18389	15884	30484
Number of groups		7	4	2	7+4+2

PPML regressions based on the sample and specification of column 2 of Table 3. See notes to Table 3 for the estimation strategy, data and sample. The sample sample is somewhat smaller than the baseline sample of Table 3 due to missing information of WB income groups for some jurisdictions. Column 1 shows that Every column additionally includes origin-nest-year fixed effects where the nests are given by the column header Column 4 uses the three. Standard errors reported in parentheses are two-way clustered: on the country of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

D.5.7 Robustness on Sample

Table 15: Omitting a High-income Native-English Destination Country at a Time

Omission of destination:	(1) AU	(2) CA	(3) UK	(4) US	(5) IE	(6) All
DL_{odt}	0.349***	0.349***	0.359***	0.327***	0.349***	0.295***
	(0.069)	(0.071)	(0.071)	(0.067)	(0.068)	(0.082)
Observations	96448	96442	96434	96435	96746	90429

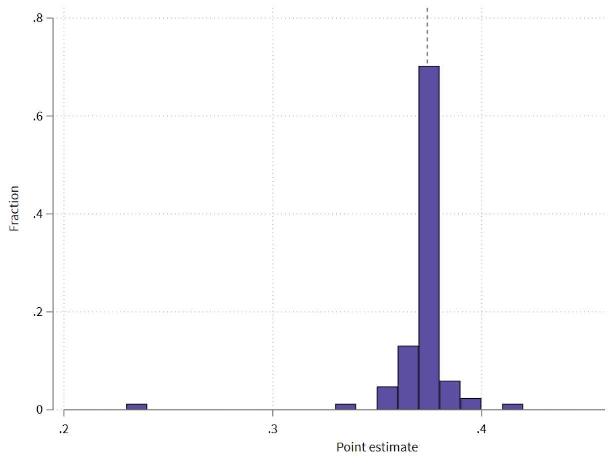
PPML regressions based on the sample and specification of column 2 of Table 3. See notes to Table 3 for the estimation strategy, data and sample. Columns 1-5 each remove a destination country, column 6 removes all countries at once. Standard errors reported in parentheses are two-way clustered: on the country of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 16: Different Sample Periods

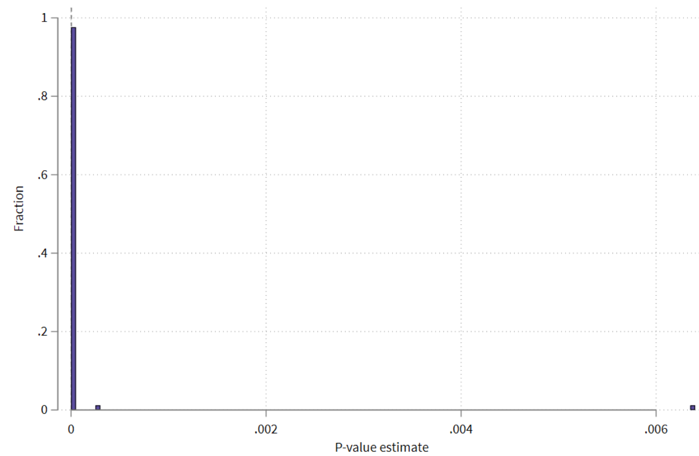
Time period:	(1) 2008–2022	(2) 2009–2022	(3) 2010–2022	(4) 2011–2022	(5) 2007–2019
DL_{odt}	0.349***	0.328***	0.311***	0.277***	0.332***
	(0.077)	(0.077)	(0.076)	(0.077)	(0.061)
Observations	117919	111894	102948	93437	93611

PPML regressions based on the sample and specification of column 2 of Table 3. See notes to Table 3 for the estimation strategy, data and sample. Every column restricts the sample to the years in the column header. Standard errors reported in parentheses are two-way clustered: on the country of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 8: Omission of a Duolingo course at a time



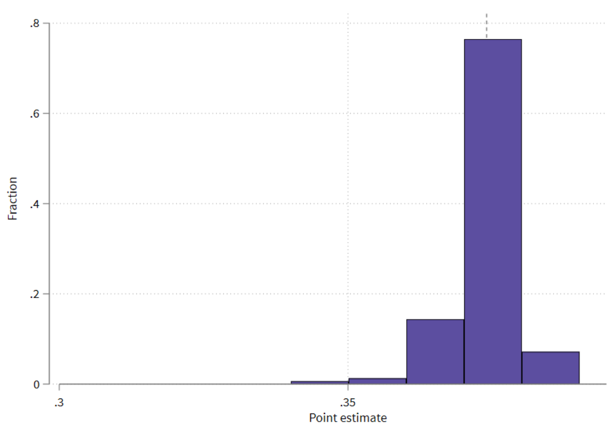
(a) Point estimates



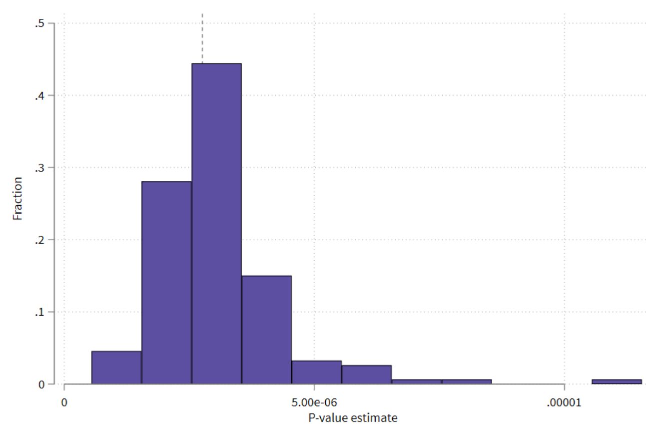
(b) p-values

Notes: Re-estimation of the results reported in column 2 of Table 3, omitting a language course at a time in the construction of DL_{odt}. (A) shows the point estimates on foreign Duolingo exposure, (B) shows the p-values of the coefficient.

Figure 9: Omission of An Origin Country at a Time



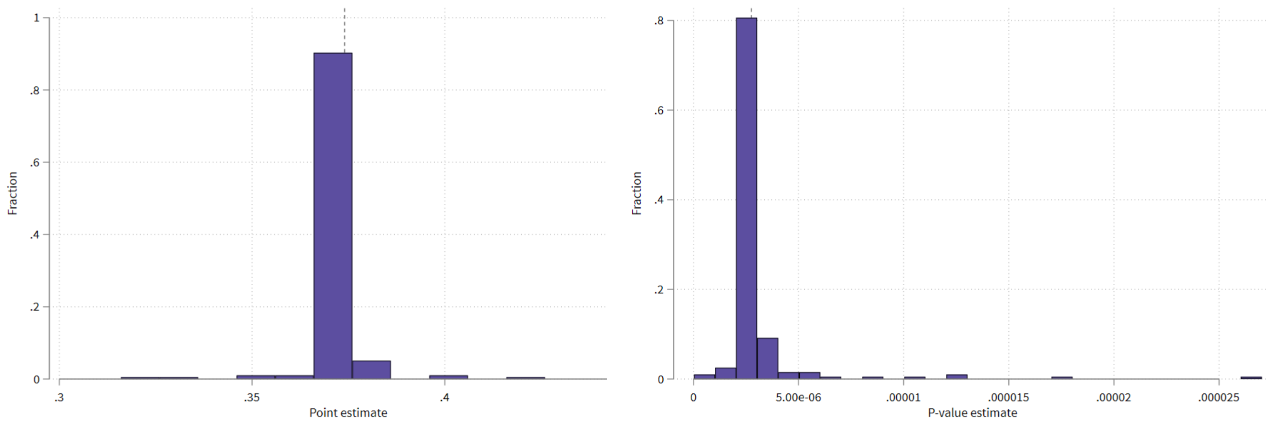
(a) Point estimates



(b) p-values

Notes: Re-estimation of the results reported in column 2 of Table 3, omitting an origin country at a time. (A) shows the point estimates on foreign Duolingo exposure, (B) shows the p-values of the coefficient.

Figure 10: Omission of A Destination Country at a Time



(a) Point estimates

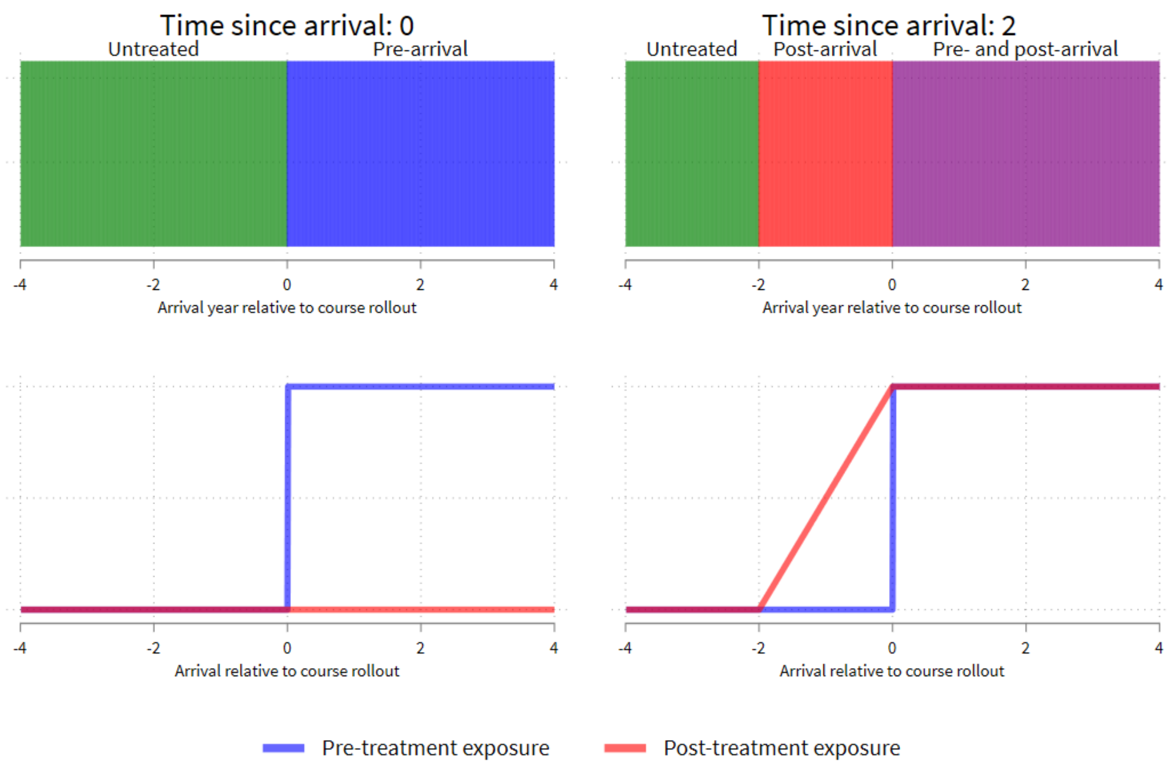
(b) p-values

Notes: Re-estimation of the results reported in column 2 of Table 3, omitting a destination country at a time. (A) shows the point estimates on foreign Duolingo exposure, (B) shows the p-values of the coefficient.

D.6 Migrant Integration in the EU

D.6.1 Identification of pre- and post-arrival effects

Figure 11: Identification of Pre- and Post-treatment Exposure



Notes: This figure illustrates the identification of the pre-arrival and post-arrival effects of Duolingo availability on migrant outcomes. The x-axis in all 4 panels represents the time of arrival of the migrant relative to the roll-out of a language course. The left column considers a migrant interviewed within the first year of arrival and the right column a migrant interviewed two years after arrival. The upper row identifies three regimes: untreated, only post-arrival learning and both pre- and post-arrival learning. The bottom row shows the pre- and post-arrival treatment intensity.

D.6.2 Attenuation bias due to aggregation

As the group-level exposure is aggregated from country-level exposure, it is a noisy measure of individual-level availability of low-cost language learning. To quantify the degree to which this noise attenuates point estimates, I examine the extent of Figure 12 shows the scatterplot of the Duolingo exposure $DL_{o_g dt}$ and the migration-flow weighted within-group variance. In absence of measurement error all circles lie on the x-axis, if classical measurement error was maximal all lines would lay on the hump-shaped line. The figure shows that across average exposure levels there is considerably less intra-group variance than maximal. About three quarters of the variance in the Duolingo exposure by origin country, destination and arrival year is driven by origin country *group*, destination and arrival year. This suggest that attenuation bias plays only a small but not negligible role. Using the formula for attenuation bias under classical measurement error from

Pischke (2007) I estimate that I underestimate the true effect by about 25%:

$$\beta \approx \frac{\sigma^2_{\Delta x} + \sigma^2_{\Delta u}}{\sigma^2_{\Delta x}} \hat{\beta} \approx \frac{5}{4} \hat{\beta} \tag{14}$$

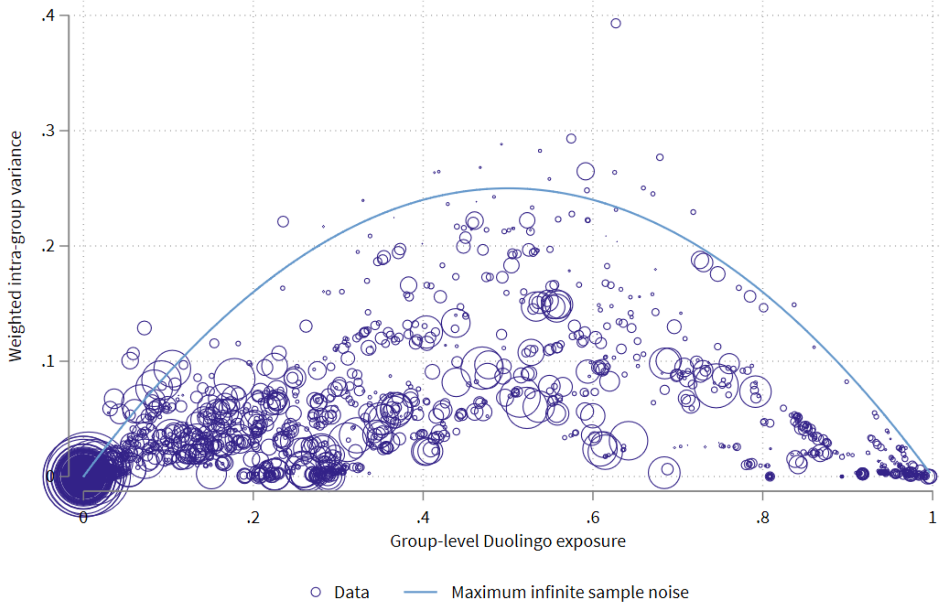
D.6.3 Descriptives and additional results

Table 17: Descriptive Statistics of Main LFS Samples

	Language skills upon arrival (2021)			Upon arrival ($t-c=0$)			Full sample		
	mean	s.d.	N	mean	s.d.	N	mean	s.d.	N
Female	0.53	0.50	19341	0.53	0.50	51464	0.53	0.50	668737
Age	37.47	9.03	19341	31.34	9.30	51464	35.23	9.07	668737
Primary educated	0.40	0.49	19250	0.46	0.50	49598	0.37	0.48	655011
Secondary educated	0.29	0.45	19250	0.20	0.40	49598	0.29	0.45	655011
Tertiary educated	0.31	0.46	19250	0.34	0.47	49598	0.33	0.47	655011
Time since arrival	6.71	4.08	19341	0.00	0.00	51464	4.65	3.55	668737
Main activity: employment	0.62	0.49	19289	0.46	0.50	46127	0.59	0.49	631765
Pre-treatment Duolingo exposure	0.17	0.23	19279	0.20	0.26	50464	0.10	0.20	664591
Language upon arrival: at least advanced	0.31	0.46	19341	0.23	0.42	607	0.31	0.46	19341
Language upon arrival: at least intermediate	0.41	0.49	19341	0.32	0.47	607	0.41	0.49	19341
Language upon arrival: at least beginner	0.56	0.50	19341	0.51	0.50	607	0.56	0.50	19341
Reason: employment, job before arrival	0.16	0.37	18911	0.23	0.42	2801	0.18	0.38	64780
Reason: employment, no job before arrival	0.21	0.41	18911	0.13	0.34	2801	0.22	0.41	64780
Reason: family	0.37	0.48	18911	0.36	0.48	2801	0.34	0.47	64780
Reason: education	0.10	0.30	18911	0.15	0.36	2801	0.10	0.30	64780
Reason: refugee	0.10	0.30	18911	0.03	0.18	2801	0.10	0.30	64780

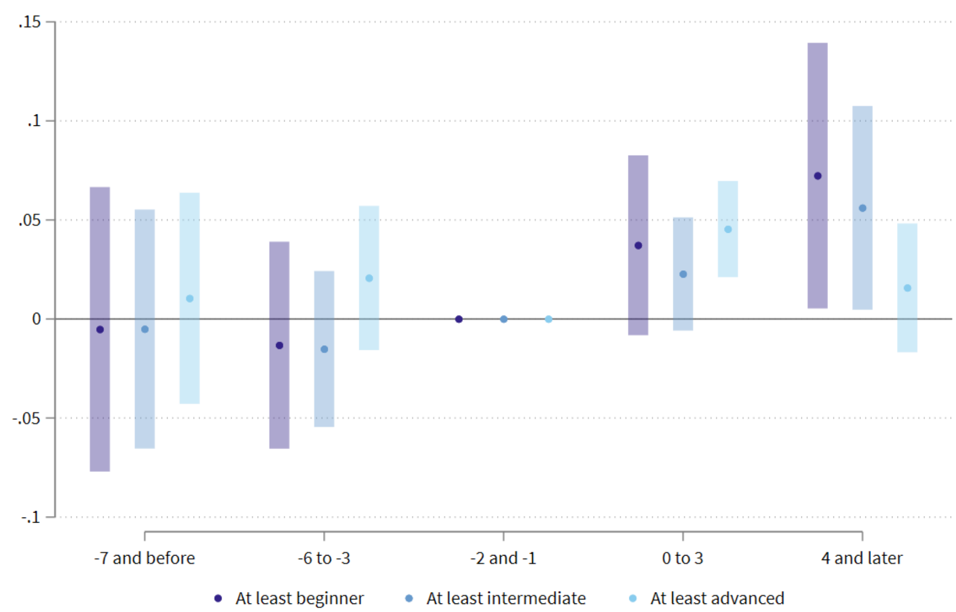
Notes: Descriptive statistics of the three samples used for the EU LFS.

Figure 12: Measurement Error due to Group-level Duolingo exposure



Notes: Scatter plot between the intra-origin group variation and the mean level Duolingo exposure on the origin group by destination by year of arrival level. N = 3,643. Empty circles' surface area is proportional to the weighted number of observations in the full estimation sample.

Figure 13: Event study of Language Skills upon Arrival around the Large Increases in Duolingo Exposure



Notes: OLS event study of language skills around large increases (at least 50 percentage points) in Duolingo exposure. Shaded blue bars indicate 95% confidence intervals based on two-way cluster-robust standard errors at the origin group and destination country level.

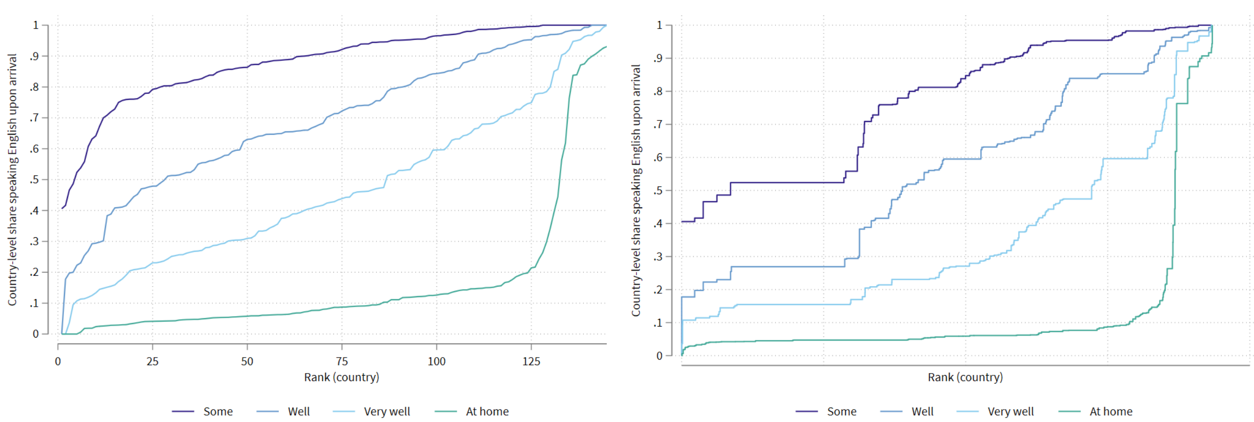
D.7 Migrant Language Skills and Integration in the US

I follow the same empirical strategy as section 7. However, because there is no variation in target languages across the US, the main equation can not be estimated with origin-by-year fixed effects for the US. Variation in origin-specific "cohort quality" may be large (Borjas, 1985), and entry wages of US migrants in more recent cohorts is decreasing (Borjas, 2015). Hence, these estimates would be particularly sensitive to changes in cohort quality among migrants from specific native languages areas. Careful assessment of differential exposure between treated and untreated countries of origin before the availability of relevant Duolingo courses is crucial.

D.7.1 Data

The American Community Survey (ACS) is a large yearly household survey fielded by the US Census Bureau among more than 3 million people each year. Respondents are randomly selected each year and are legally obliged to answer, providing information about themselves and other members of their household. The ACS collects information of a range of relevant demographic characteristics and economic outcomes, as well as information on an individuals' migration history and country of birth, the language spoken at home and a self-assessment of the contemporaneous language skills of all household members, with the following answer options: Only English, very well, well, not well, does not speak English.[44] Although the language spoken at home could be used to construct an exposure measure, I choose to use the country of birth as the language spoken at home could be endogenous. I restrict the sample to those who arrived on ar after age 18 and those currently 59 or younger who immigrated to the US in the past 10 years. In line with the other datasets, I start the analysis in 2007 until the last available year, 2022.

Figure 14: Distribution of English Language Skills Upon Arrival in the US



(a) Unweighted

(b) Weighted

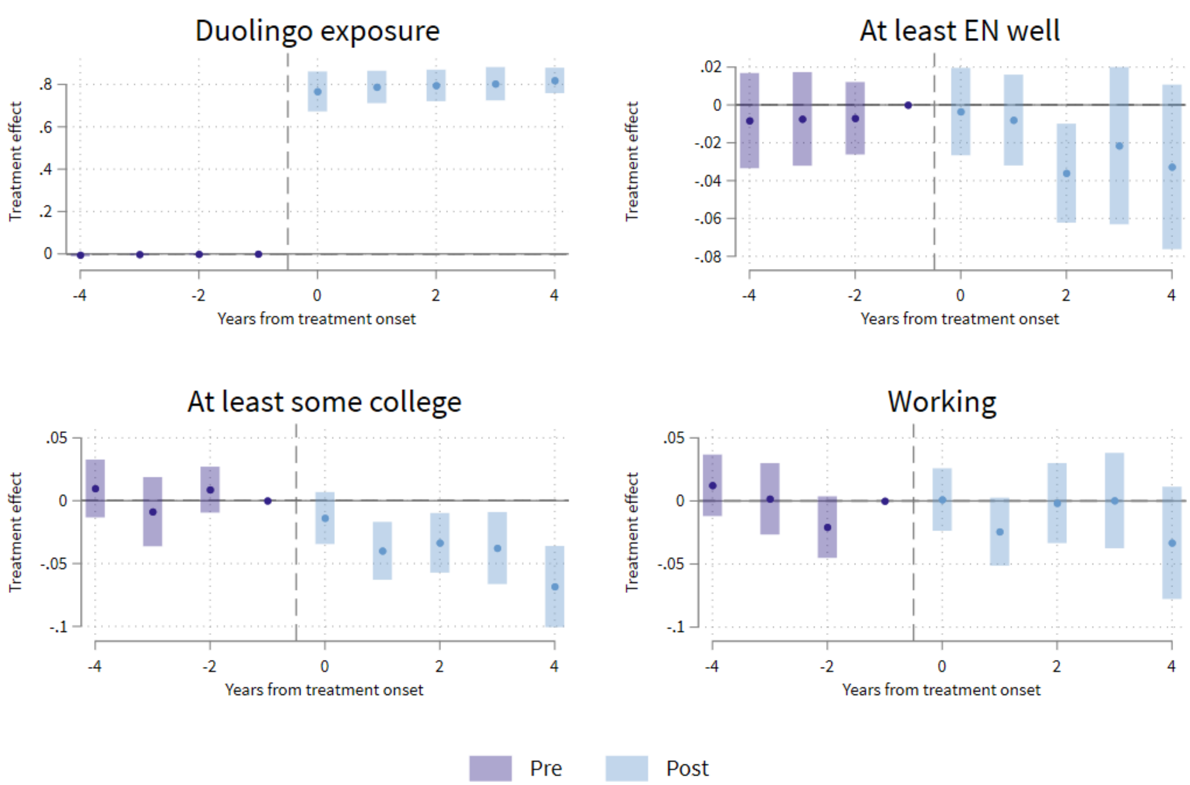
Notes: Distribution of the level of language skills upon arrival in the US between 2007 and 2022, by origin country (left) and weighted by cohort size (right). The strong difference between both graphs shows that immigrants from several large immigrant origin countries have poort English skills (such as Mexico).

Figure 14 shows the cumulative distribution of levels of language skills among those within the first year

[44]Self-reported language skills in the ACS have been shown to strongly correlate to actual language proficiency (US Census Bureau, 2015). However, some studies have found that active learners under-assess their learning gains, e.g. Ma and Winke (2019)

of arrival. This shows that many respondents lack good English language skills before arriving to the US. As discussed in section 7, the identification strategy in the US assumes parallel trends in outcomes between treated and untreated origin countries. A potential risk to this identification assumption could be that language skills for English skills have trended differently for example for countries that speak more widely spoken languages (and who are more likely to have received a Duolingo course) than for less widely spoken languages. Figure 15 shows that pre-trends before large increases in Duolingo exposure are small. However, the results suggest that the share of respondents with at least some college decreased after introduction of Duolingo.

Figure 15: The Effect of Duolingo Exposure on Migrant Outcomes upon arrival



Notes: OLS results from a Nagengast and Yotov (2023) event study estimator around large increases in Duolingo exposure on outcomes among those, including origin and year fixed effects. The panels report results for four different outcomes: the exposure itself (upper left), speaking English at least well (upper right), having at least some college education (lower left) and currently working (lower right) by native language of the test takers. As the unit of observation is the native language level, the Duolingo exposure is binary. Shaded blue areas indicate 95% confidence intervals based on standard errors that are clustered at the origin country level.

D.7.2 Results

Table 14 examines the language skills, characteristics and employment outcomes upon arrival, with and without country-year controls. The results suggest that the probability to speak English very well in the first year since arrival has decreased, although it is not statistically significant. Moreover, as suggested in Figure 15, the probability to have college educations has decreases by 2 percentage points on average, but in the two-way fixed effects regression the effect is not significant. Panel C find some suggestive evidence that workers' earnings increase, but that they perform jobs in which English skills are less important. This is suggestive of the fact that the availability of low-cost language learning facilitates immigrants to find jobs in which they do not need to be very proficient in English.

Turning to integration of immigrants beyond the first year after arrival, I analyze the effects of pre- and post-arrival exposure on language skills in Table 19 and on economic integration in Table 20. The estimates suggest that the lower share of individuals with very good English language skills upon arrival is only temporary as exposed individuals catch-up. Moreover, the estimates for post-arrival exposure suggest that being able to learn languages after migration increases the probability to speak English at least well with 2 percentage points. Moreover, Table 20 shows that the introduction of a Duolingo course after arrival increases the probability to be working by 4 percentage points and increases incomes by 7%. The initial lower English-language intensity of immigrants is rapidly catching up to that of unexposed immigrants. In addition, I do not find that exposure to Duolingo after arrival decreases the English intensity of immigrants' jobs. This suggest that the effect of language knowledge does not decrease the English intensity of jobs. This further suggests that selection effects are driving the decreased English intensity of jobs upon arrival.

Table 18: The Effect of Duolingo Exposure on Language Skills upon Arrival in the U.S.

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Language skills						
	At least some		At least well		At least very well	
$DL_{oc}^{T_d}$	0.006	-0.005	-0.006	-0.013	-0.014	-0.024
	(0.011)	(0.011)	(0.013)	(0.014)	(0.009)	(0.015)
Observations	44611	34824	44611	34824	44611	34824
R^2	0.32	0.32	0.35	0.35	0.22	0.22
Mean dep. var.	0.820	0.821	0.619	0.624	0.350	0.352
Panel B: Selection						
	Female		At least 9th grade		At least some college	
$DL_{oc}^{T_d}$	-0.016	-0.022	-0.002	0.005	-0.009	-0.020
	(0.011)	(0.013)	(0.011)	(0.011)	(0.012)	(0.012)
Observations	44611	34824	44611	34824	44611	34824
R^2	0.03	0.03	0.17	0.18	0.24	0.25
Mean dep. var.	0.482	0.476	0.886	0.886	0.451	0.464
Panel C: Integration						
	Working		Log earnings		Occupation: importance English	
$DL_{oc}^{T_d}$	0.029	0.028	0.034	0.109**	-0.942*	-1.010*
	(0.019)	(0.019)	(0.062)	(0.046)	(0.566)	(0.538)
Observations	44611	34824	18650	14673	21157	16202
R^2	0.15	0.17			0.37	0.40
Mean dep. var.	0.418	0.421	31,921	32,512	40.947	40.880
Country-year controls		✓		✓		✓

Notes: PPML (Panel C column 3 and 4) and OLS (all others) estimations of the model of equation 11. Panel A, B and C consider those interviewed in the first year after arrival in the full 2007-2022 ACS. Panel A includes three binary indicators for minimum levels of language skills, Panel B includes binary indicator for being female, having at least completed 9th grade and a binary indicator for having at least some college educated. Panel C includes a binary indicator for being in work, yearly labor income in US Dollar and the occupation-level importance of English from ONET. This is a score between 0 and 100 indicating how important a skill is for the job. The skill description for use of English Language: "Knowledge of the structure and content of the English language including the meaning and spelling of words, rules of composition, and grammar".
All even columns include extensive origin-year level controls: log of GDP per capita, unemployment rates, the share of population with tertiary education, the number of conflict deaths from the Global Burden of Disease dataset, the GINI coefficient from the World Inequality Database, the median income from the World Bank, and the share of admissions of legal permanent residents by visa type from Yearbook of Immigration Statistics of the DHS. The measure of Duolingo exposure is constructed using data from (Ginsburgh, Melitz and Toubal, 2017) and the rollout dates of Duolingo courses. Standard errors reported in parentheses are two-way clustered: on the country group of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 19: The Effect of Duolingo Exposure on Language Skills after arrival in the USA

	(1) Upon arrival	(2)	(3) Full	(4)	(5) Interview before Duolingo Exposure
Panel A: Speaks at least some English					
$DL_{oc}^{T_d}$	0.006	0.004	0.001		
	(0.011)	(0.009)	(0.010)		
$DL_{oc}^{T_d} \times \log(t-c+1)$		-0.005	-0.004		
		(0.005)	(0.005)		
$DL_{otc}^{T_d,post}$			0.006	0.008	0.002
			(0.006)	(0.007)	(0.008)
Observations	44611	376541	373819	373819	209922
R^2	0.32	0.26	0.26	0.26	0.26
Mean dep. var.	0.820	0.882	0.882	0.882	0.910
Panel B: Speaks English at least well					
$DL_{oc}^{T_d}$	-0.006	-0.004	-0.011		
	(0.013)	(0.010)	(0.009)		
$DL_{oc}^{T_d} \times \log(t-c+1)$		-0.002	0.002		
		(0.008)	(0.008)		
$DL_{otc}^{T_d,post}$			0.019**	0.020**	0.020**
			(0.009)	(0.010)	(0.010)
Observations	44611	376541	373819	373819	209922
R^2	0.35	0.37	0.37	0.37	0.38
Mean dep. var.	0.619	0.679	0.678	0.678	0.715
Panel C: Speaks English at least very well					
$DL_{oc}^{T_d}$	-0.014	-0.020***	-0.023**		
	(0.009)	(0.008)	(0.009)		
$DL_{oc}^{T_d} \times \log(t-c+1)$		0.009*	0.010*		
		(0.004)	(0.005)		
$DL_{otc}^{T_d,post}$			0.006	0.008	0.011
			(0.009)	(0.010)	(0.009)
Observations	44611	376541	373819	373819	209922
R^2	0.22	0.27	0.27	0.27	0.28
Mean dep. var.	0.350	0.416	0.416	0.416	0.437
$DL_{oc}^{T_d} \times (t-c)$ FE				✓	

Notes: OLS estimations of the model of equation 14, with the following outcomes: Speaking at least some English (A), speaking English at least well (B) and speaking English at least very well (C). Column 1 shows the effect upon arrival as shown in previous tables. Column 2-4 report results from the full sample of immigrants within the first 5 years of arrival. To study whether the initial gains in outcomes fade out over time, I interact the Duolingo exposure with the log of years since arrival plus one in column 2. Column 3 introduces the post-arrival treatment intensity and Column 4 includes an interaction of pre-arrival intensity with dummies of years since arrival (not shown). The last column limits the sample to those who had no exposure to Duolingo before arrival. The measure of Duolingo exposure is constructed using data from (Ginsburgh, Melitz and Toubal, 2017) and the rollout dates of Duolingo courses. Standard errors reported in parentheses are two-way clustered: on the country group of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 20: The Effect of Duolingo Exposure on Migrant Outcomes after arrival in the USA

	(1) Upon arrival	(2)	(3) Full	(4)	(5) Interview before Duolingo Exposure
Panel A: Working					
$DL_{oc}^{T_d}$	0.029	0.051***	0.037***		
	(0.019)	(0.013)	(0.012)		
$DL_{oc}^{T_d} \times \log(t-c+1)$		-0.032***	-0.027***		
		(0.008)	(0.007)		
$DL_{otc}^{T_d,post}$			0.035***	0.032**	0.039***
			(0.011)	(0.012)	(0.010)
Observations	44611	376541	373819	373819	209922
R^2	0.15	0.19	0.19	0.19	0.21
Mean dep. var.	0.418	0.627	0.626	0.626	0.623
Panel B: Log yearly labor income					
$DL_{oc}^{T_d}$	0.034	0.006	-0.029		
	(0.062)	(0.045)	(0.034)		
$DL_{oc}^{T_d} \times \log(t-c+1)$		-0.019	-0.001		
		(0.029)	(0.022)		
$DL_{otc}^{T_d,post}$			0.072	0.084	0.069**
			(0.049)	(0.053)	(0.031)
Observations	18650	236043	234102	234102	130799
R^2					
Mean dep. var.	31920.838	45468.648	45446.903	45446.903	48654.692
Panel C: Occupation-level importance of English language					
$DL_{oc}^{T_d}$	-0.942*	-1.150***	-0.919***		
	(0.566)	(0.424)	(0.308)		
$DL_{oc}^{T_d} \times \log(t-c+1)$		0.669**	0.565**		
		(0.284)	(0.235)		
$DL_{otc}^{T_d,post}$			-0.522	-0.516	-0.114
			(0.435)	(0.456)	(0.406)
Observations	21157	238209	236135	236135	121743
R^2	0.37	0.32	0.33	0.33	0.33
Mean dep. var.	40.947	41.014	41.017	41.017	41.793
$DL_{oc}^{T_d} \times (t-c)$ FE				✓	

Notes: PPML (Panel B) and OLS (Panel A and C) estimations of the model of equation 14, with the following outcomes: Being in work, yearly wage income in US Dollar and the occupation-level importance of English from ONET. This is a score between 0 and 100 indicating how important The skill description for use of English Language: "Knowledge of the structure and content of the English language including the meaning and spelling of words, rules of composition, and grammar". Column 1 shows the effect upon arrival as shown in previous tables. Column 2-4 report results from the full sample of immigrants within the first 5 years of arrival. To study whether the initial gains in outcomes fade out over time, I interact the Duolingo exposure with the log of years since arrival plus one in column 2. Column 3 introduces the post-arrival treatment intensity and Column 4 includes an interaction of pre-arrival intensity with dummies of years since arrival (not shown). The last column limits the sample to those who had no exposure to Duolingo before arrival. The measure of Duolingo exposure is constructed using data from (Ginsburgh, Melitz and Toubal, 2017) and the rollout dates of Duolingo courses. Standard errors reported in parentheses are two-way clustered: on the country group of origin and country of destination level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.