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

Evidence from *Duolingo*

Joop Adema †

aap

July 15, 2024

Abstract

Alongside the rise of the internet and smartphones, language learning applications have become hugely popular. This technology encourages and facilitates language learning, at low cost. As language knowledge is a crucial determinant of migrants' earnings, the availability of low-cost language learning may have noteworthy consequences for prospective migrants prior to moving as well as migrants in the host country. In this paper, I study this using the staggered introduction of 84 language courses on the widely used platform *Duolingo*. As these courses enable learning a specific target language from a specific source language, they generate rich variation in the availability of low-cost language learning across migration corridors. The roll-out of these courses was plausibly supply-constrained and targeted the largest language-learning pairs first. Furthermore, past levels and changes in migration intentions and flows do not predict module introduction, reassuring that courses were not rolled in expectation of increasing migration flows. First, I establish that course rollout increases online search interest in the target language and that courses with English as a target language improve *TOEFL* test scores in reading and listening, but not writing and speaking. Second, using the gravity model of migration I find that the introduction of a typical module increases the stock of individuals aspiring to migrate by about 6%. In addition, actual migration flows to *OECD* countries start increasing several years after the introduction of a relevant course. Third, I find that the pre-migration availability of a relevant Duolingo course increases the probability that a migrant in the *EU* speaks the destination country language at least at beginner level by about 11 percentage points, and increases the proportion of economic migrants. In similar vein, I find that Duolingo courses to English increases migrants' probability to speak English at least well by 2 percentage points in the *US*. In the near future, I plan to study the effects on employment outcomes, wages and occupational skill-content of migrants in both the US and Europe.

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*Preliminary and incomplete - please do not cite/circulate.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

Contents

1 Introduction

The rapid rollout of mobile internet and the adoption of smartphones has transformed international migration. Not only does it enable more than 4 billion people (GSMA, 2019) to search for information about opportunities in potential destination countries and continuously stay in contact with people far away through social media, it also gives people access to modern educational technologies. One of the most popular educational technologies are (gamified) language learning applications, being actively used by more than 60 million people worldwide in 2022, of whom 56.5 million on the platform *Duolingo*.[1] The introduction of these applications lowered the financial and convenience cost of language learning, which may increase language learning in the extensive and intensive margin as well as the languages studied.

Although large productivity differences across the world imply large potential benefits from labor mobility between countries, a lack of language-related country-specific skills is one impediment to reaping the benefits. Knowledge of a local language is an important determinant of foreign workers' earnings (Chiswick and Miller, 2015) and arguably important in other aspects of life as well. Unsurprisingly, migration flows between countries sharing a language and with smaller linguistic distance between its main languages are larger (Belot and Ederveen, 2012; Adsera and Pytlikova, 2015). Not only are there benefits of foreign language knowledge abroad, foreign language skills have also been associated with higher earnings at home (Isphording, 2013; Di Paolo and Tansel, 2015; Stöhr, 2015). These motives are in line with the fact that foreign language learning is widespread across the globe. Important macro-level determinants of foreign language learning are international trade flows, language size and linguistic proximity (Ginsburgh, Melitz and Toubal, 2016).

Throughout history, language learning has been rigid and costly, which was acknowledged by Luis von Ahn, a professor of Computing Science at Carnegie Mellon University. Von Ahn was born in Guatemala and noted that learning English was prohibitively expensive for many Guatemalans[2]. With the subsequent introduction of the application *Duolingo*, co-founded by von Ahn in 2012, low cost language learning became available to a wide public – it is available on smart phones and desktops for free.[3] *Duolingo* provides interactive language learning in courses enabling one to learn a target language from a specific source language. Starting with the courses English→Spanish, Spanish→English and English→German in 2012, many additional courses were rolled out in subsequent years. As of 1st of January 2023, 84 courses to living languages are have reached the final phase of development. Each course consists of a series of topical lessons lasting a

[1] See https://www.businessofapps.com/data/language-learning-app-market/ for other statistics regarding the language learning market.

[2] Von Ahn mentioned that this would be beneficial for income at the origin country as well: "And knowledge of English in a non-English speaking country can usually mean that your income potential is doubled. I mean, you literally make twice as much money if you know English. So that's kind of where the idea came from to have a free way to learn languages, and that was Duolingo." from https://www.bbc.com/news/business-51208154

[3] *Duolingo* also offers an ad-free platform with some extras for a modest fee. By 2022, *Duolingo* has almost 3 million paid users.

few minutes consisting of several items which consist of matching through listening, translating, speaking, multiple choice items and stories for some courses. New users can do a placement test to start at a lesson at a higher level. A course typically provides approximately several thousands of lessons, including 10s of thousands unique sentences and 1000s of unique words, and is under constant development.[4] 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.

As of the third quarter of 2022, the application has been downloaded over 600 million times and has 56.5 million monthly active users worldwide. Figure A3 shows the most learned language by country on the platform in 2021. In most of the world, English prevails as the most learned language, which is partially driven by the fact that there are many (22) courses to English available. A few notable exceptions appear: Korean is the most learned language in Mongolia (following the flow of many seasonal workers to Korea) and Swedish is the most learned language in Sweden, which is driven by Arabic-speaking refugees learning the host country's language.

That the availability of language learning affects international migration has been shown recently by Huber and Uebelmesser (2019) and Jaschke and Keita (2021). Both studies use the opening of German language learning centers (so-called *Goethe* institutes) on international migration. Huber and Uebelmesser (2019) find that migrant flows to Germany increase after opening of a German language learning institute in an origin country and Jaschke and Keita (2021) find that the migrant pool has better German skills upon arrival and becomes positively self-selected in education. However, contrary to foreign language institutes such as Goethe, *Duolingo* courses do not provide generally accepted certification certification.[5] Moreover, such online private alternatives to language learning may have large effects on the uptake of language learning, especially when in-class alternatives are not provided in immigrants' proximity (Foged and van der Werf, 2023).

In this paper, I study whether the availability of low-cost language learning through mobile-assisted language learning (MALL) affected international migration between 2007 and 2021. We model language learning and migration as a simple two-period setting, where agents decide on a language learning investment at a convex cost in period 1 and observe a migration cost shock and decide to migrate or not in period 2. By linking information on the roll-out of *Duolingo* courses with the share of speakers of the source and target languages in origin and destination country respective, I construct a time-dependent measure that captures the probability that a random person in the origin country can attain the language of a random

[4] Data on the size of the language courses is obtained from https://ardslot.com/duolingodata.html

[5] *Duolingo* provides certification for English by means of a adaptive language test costing 49 US dollar. Although introduced in 2016, the Duolingo English Test became popular only during the COVID-19 pandemic, see https://learningenglish.voanews.com/a/duolingo-english-test-gains-support-questions-remain/6357539.html

person in the destination country through a specific *Duolingo* course. This measure proxies the usefulness of the available language course for potential emigrants to learn the language of the destination country. As multiple courses can bridge two countries through different languages which are spoken widely, I cap the scale of the measure at the largest value of the measure for any of the courses or at unity. Importantly, I also include non-native languages spoken to the extent that the information is available. We obtain information on language requirement policies from MIPEX to study whether low-cost uncertified language learning has implications for migration to countries that require additional certification for permanent residence.

We combine this measure to survey data from the Gallup World Polls (GWP) and aggregate data on yearly bilateral migration from the OECD. The GWP elicits individuals' desire to emigrate and preferred destination country. As the GWP is a representative survey targeting 1,000 individuals every country every year, I can construct a measure of the stock of people desiring to emigrate between a country surveyed in GWP and any other country in the world. The OECD bilateral migration data records migration flows from all countries in the world to OECD destination countries (33 in our main estimation sample).

Although actual migration is a more tangible outcome than migration aspirations based on a limited sample, the latter still provides interesting evidence for a variety of reasons. First of all, changes to migration aspirations are predictive of subsequent migration (Tjaden, Auer and Laczko, 2019). Secondly, studying migration aspirations offers advantages to studying migration responses when the treatment of interest likely has only effects on actual migration flows and longer-run flow data is not (yet) available. Third, the GWP data covers all destination countries, which reduces concerns about destinations missing not at random. Nevertheless, I provide additional evidence by considering emigration to OECD countries. Despite the limited geographic coverage of the OECD data, many large migrant-receiving countries are included.

Using a gravity model with staggered introduction of *Duolingo* courses with continuous treatment intensity, I find that the introduction of a language course spoken by the full population in the origin and destination country increases the desire to emigrate to that specific country by 24%. As a typical course has an exposure score of 0.32, such a course increases bilateral flows by about 7% for. Event study estimators robust to heterogeneous and dynamic treatment effects confirm this result and moreover show that there are no pre-trends before the roll-out of a *Duolingo* course. Furthermore, by using data on international migration to OECD countries, I find evidence that migration flows start increasing 4 years after the roll-out of a language course.

The remainder of the paper is structured as follows. Section 2 reviews related literature and discusses this paper its contributions to it. Section 4.1 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. Section 3 introduces the language learning app *Duolingo* and describes the roll-out of courses and section 4.2 discusses

the main empirical strategy and its identifying assumptions. Sections 5, 6 and 7 present the results of this paper in three parts. First, 5 shows how *Duolingo* spurred interest in languages and test scores. Second, section 6 reveals its effects on international migration to OECD countries Third, section 7 displays the effects on migrants' language skills and economic integration in the United States. Section 8 concludes and explores future steps.

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 on the drivers of 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, many scholars have studied the role of language and culture in migrants' earnings. 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%.[6] 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 avoidance and perceived gender roles are the most important cultural impediments to international migration. Adsera and Pytlikova (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

[6] 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 (2022) conclude that language learning policies for refugees have positive effects on earnings in the short run.

factors correlated to language. We 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 language spoken in the destination country through available technology. Contrary to studies using time-independent measures of language and culture, I can do so controlling for unobserved dyadic fixed effects.

The second strand of literature concerns (digital) technologies that affect international migration. The internet changed the speed and way information spreads across the globe, which is likely to have large consequences for international migration. 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 origin countries. 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. We contribute to this literature by showing how a very specific type of technology, namely 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. Dustmann (1999) has shown that temporary migrants with a longer horizon in Germany have a larger incentive to attain a host country language and have better languages skills. Along similar lines, Wong (2022) exploits random allocation of refugees in Switzerland, finding that linguistic proximity is related to better labor market outcomes. 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. The latter two studies are suggestive of the fact that labor market potential increases when language learning is easier. This motivated policy makers to introduce (obligatory) language courses for some immigrant groups. In a recent survey of the literature, Foged, Hasager and Peri (2022) conclude that language learning policies for refugees have small positive effects on earnings in the short run. However, 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. However, immigrants able to speak English are moving to places with lower English skills, suggesting higher returns when English skills are scarce (Fenoll and Kuehn, 2019). 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 and decreased fertility. Access to language learning has been studied in several settings. First, quasi-experimental variation in integration policies with language as a key component has been shown to be an effective tool for migrant and particularly refugee integration (Dahlberg et al., 2024). Only few studies have examined the role of language learning in isolation. Foged and Van der Werf (2023) study the availability of language training 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.[7] Huber and Uebelmesser (2019) 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 (2019) 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, also those who would be liquidity constrained. Furthermore, compared to the yearly attendance of the Goethe Institutes, the number of Monthly Active Users on *Duolingo* is 500 times 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 language learning application with gamified features, consisting of bilateral and directional courses: a course enables one to learn a specific target using a source language. Although *Duolingo* was not the first online language learning platform, it gained considerably more traction than its competitors (Shortt et al., 2023). This allowed Duolingo to amass a market share of 64% in 2022, more than six times that of its closest competitor.[8] Compared to its competitors, Duolingo's entry threshold is very low. First of all, most of Duolingo's content is available for free[9] Secondly, it allows learning from scratch, as the language of instruction is the source language. This feature also gives rise to rich variation in availability of low-cost

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

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

[9] 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.

language learning across speakers of different languages, which is absent for its competitors.

Figure A1 shows a series of typical tasks on Duolingo: 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.[10] The rightmost screenshot of Figure A1 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.[11] Nevertheless, courses were gradually extended over time, and not all courses include lessons up to B2 to date.[12] *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; Ersoy, 2021; Rachels and Rockinson-Szapkiw, 2018). 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.

The first courses where rolled out in 2012, with English to Spanish, Spanish to English and English to German.[13] 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*.[14] Nevertheless, many *Duolingo* language course 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.

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

[11] 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 (CEFR, 2023).

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

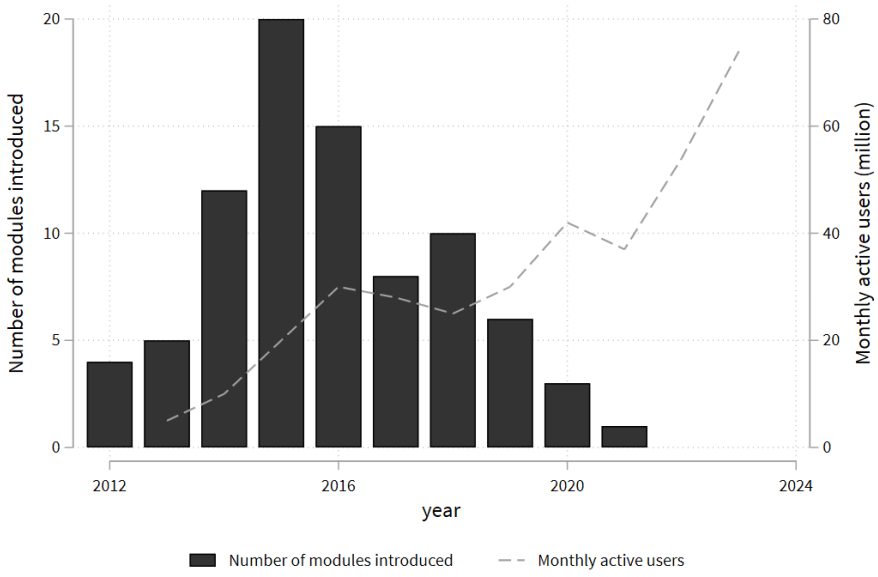
[13] More than half a million people signed up for the *beta* versions in June 2012, mobile applications on iOS and Android were released in November 2012 and June 2013, respectively.

[14] 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.

I obtained the available courses and the date of rollout from a fan-based website.[15] 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.[16]

We discard Esperanto, Klingon, High Valyrian and Latin, as they are not widely used and therefore irrelevant to international migration. Until 2022, 110 courses involving two existing languages have been developed, of which 87 have reached the final phase. We 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 A2 shows all courses available since 2021 by connecting all 23 unique source and 30 unique target languages in a Sankey diagram. 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. We further study the determinants of course rollout and uptake in section 4.2.

Figure 1: Introduction of *Duolingo* courses



[15] These dates can be found on https://duolingo.fandom.com/wiki/Course_list. The dates were verified using the available languages on *Duolingo* through the Wayback Machine.

[16] 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, we 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.

In 2013, Duolingo had 5 million monthly active users, which gradually rose to 74.1 million in 2023, of whom 21.4 million active on a daily basis. 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.[17] 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-relatee reasons, whereas those in non-English speaking countries and younger generations are more likely to use it for school.[18] Although I use the rollout of Duolingo on the language level, it is still insightful to see where Duolingo's users are located. Table A2 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.

4 Model and Empirical Strategy

4.1 A Model of Language Learning and Migration

Duolingo courses are available for free. This stands in stark contrast to traditional in-class language courses. For example, in the Goethe institutes studied by Huber and Uebelmesser (2019), a course at CEFR A2 level costs 484 euro in Colombia and 260 euro in Bangladesh in 2022. Motivated by this, I model the effects of Duolingo introduction as a decrease in the cost of language learning on language skill investments and migration decisions in a modified random utility model (RUM) of migration.

We model costly language learning and migration from origin country o to a destination country d as a two-step process. In the first step, individuals in o choose the optimal acquisition of a foreign language skill $s_{oT} \in (0, 1)$. As individuals from country o are ex-ante indistinguishable, optimal language skill decisions are homogeneous and s_{oT} is not indexed by i. The cost of acquiring a target language skill T is convex and proportional to c_{oT}:

$$c_{oT} s_{oT}^2 = \frac{\kappa_{oT}}{2(1 + \eta_{oT} \alpha_{oS} Duolingo_{ST})} s_{oT}^2 \tag{1}$$

Here, we parametrize c_{oT} to depend on a parameter κ_{oT}, which is lowered by the availability of a Duolingo course to T for people in o with effectiveness η_{oT}. A Duolingo module is available in o for the share of population that speaks the source language of the course, α_{oS}.

In the second step, an individual observes the idiosyncratic benefits of migration to all alternatives d, ϵ_{iod}, and chooses to migrate to the destination which offers highest utility. Following the literature, we model

[17] 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. https://blog.duolingo.com/dear-duolingo-how-does-language-learning-differ-between-generations/

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

the idiosyncratic term as an iid 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 s_{oT} from country o when moving to d is:

$$U_{iod} = \ln w_{od} + \epsilon_{iod} = \mu_{od} + s_{oT} b_{odT} + \epsilon_{iod} \tag{2}$$

Here, μ_{od} are earnings for someone without relevant language skills in country d, s_{oT} are language skills of individuals from o in language T and b_{odT} denotes the return to language skill for individuals from o in country d, which we assume to be always finite, positive, and strictly positive for at least one d. For simplicity, we do not explicitly include migration costs, but these can be thought of as absorbed in μ_{od} and in b_{odT}, if language skills reduce migration costs. Using the properties of the EVT-1 shock, utility maximization gives the following migration probabilities:

$$\mathbb{P}_{od} = \frac{e^{\mu_{od} + s_{oT} b_{odT}}}{\sum_{d'} e^{\mu_{d'} + s_{oT} b_{od'T}}} \tag{3}$$

The denominator sums over all potential destinations d', which also includes the origin country itself. 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 migration to d over staying in o (Bertoli and Moraga, 2013; Beine, Bertoli and Fernández-Huertas Moraga, 2016):

$$\ln\left(\frac{\mathbb{P}_{od}}{\mathbb{P}_{oo}}\right) = \mu_{od} - \mu_o + (b_{odT} - b_{ooT}) s_{oT} \tag{4}$$

This expression is independent of the deterministic part of utility in alternative destinations. It shows that the ratio of migrants to d over stayers is an increasing function of s_{oT} if returns to the language skill is higher in the destination country than at home ($b_{odT} > b_{ooT}$). However, this result does not imply that total migration to d is increasing in s_{oT}. Appendix Section A.1 shows that this is only the case when b_{odT} exceeds the migration probability-weighted in all alternative destinations (including the origin). Likewise, Equation 21 shows that total emigration from o increases only when the weighted foreign return exceeds domestic return 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. We assume that the discount factor is one. The expected utility for someone from country o in period one is given by the expected utility in period two minus the cost of

language learning:

$$U_o^1 = U_o^* - c_{oT} s_{oT}^2 = \sum_d U_{od}^* \mathbb{P}_{od} - c_{oT} s_{oT}^2 \tag{5}$$

Using the envelope theorem, $\frac{\partial U_{od}^*}{\partial \mathbb{P}_{od}} = 0$, I obtain the following first order condition:

$$2 c_{oT} s_{oT} = \sum_d \mathbb{P}_{od} b_{odT} \tag{6}$$

The left hand side represents the marginal cost of one unit of language skills, whereas the right hand side represents the marginal benefit. Importantly, the migration probabilities $\mathbb{P}_{od}$ depend on s_{oT}. An equilibrium pinning down s_{oT}^* exists and is unique.[19] For most countries, migration probabilities are small compared to the probability of staying. In that case, the right hand side becomes linear in s_{oT} and an expression for s_{oT}^* can be derived:

$$s_{oT}^* \approx \frac{\mathbb{P}_{oo} b_{ooT} + \sum_{d \neq o} \mathbb{P}_{od}(0) b_{odT}}{2 c_{oT} - \sum_{d \neq o} \mathbb{P}_{od}(0) b_{odT}^2} \approx \left(\mathbb{P}_{oo} b_{ooT} + \sum_{d \neq o} \mathbb{P}_{od}(0) b_{odT} \right) \frac{1 + \eta_{oT} \alpha_{oS} Duolingo_{ST}}{\kappa_{oT}} \tag{7}$$

Here, $\mathbb{P}_{od}(0)$ denotes the migration probability in absence of the language skill. Different motives for language learning can be isolated. First, higher earnings on the domestic labor market can motivate language skill acquisition. Second, a language skill can increase earnings abroad. Hence, the expected benefit is a migration-probability weighted average of foreign returns to skill. Third, a larger language skill shifts the migration probabilities towards destinations wherethe language skill is more valued, deflating the denominator.[20] However, the latter can be assumed small compared to c_{oT}.[21] After invoking this and plugging in the parameterization of the cost function yields the term on the right hand side.

Plugging s_{oT}^* from equation 7 into equation 4 yields the following expression for the log odds of migration:

$$\ln\left(\frac{\mathbb{P}_{od}}{\mathbb{P}_{oo}}\right) = \mu_{od} - \mu_o + (b_{odT} - b_{ooT}) \left(\mathbb{P}_{oo} b_{ooT} + \sum_{d \neq o} \mathbb{P}_{od}(0) b_{odT} \right) \frac{1 + \eta_{oT} \alpha_{oS} Duolingo_{ST}}{\kappa_{oT}} \tag{8}$$

As with language skills, the log odds ratio of migration depends on the returns to language skills at home and abroad. Moreover, it depends on the *difference* in earnings abroad and at home. Hence, both the

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

[20] The second term in the denominator is quadratic in b_{od} as higher returns increase returns as well as the probability proportional to b.

[21] To see why this is likely, we estimate c_{oT} in a case it is relatively low and compare the second in the denominator. Returns for English in European countries, where 30-70% of the population learns English, vary between 10-50%. Equating marginal costs and benefits in absence of migration (if anything this gives a lower bound to returns and underestimates costs) using the middle of the ranges for b and l gives $c_{oT} = b/2s = 0.3/(2 * 0.5) = 0.3$. As migration probabilities are very low, and returns to foreign languages abroad and at home are not much higher than 50% (see 2 as well as below), this term is much smaller than 0.3.

size and strength of the effect on Duolingo availability depends on domestic and foreign returns to language skills.

Our model makes several simplifying assumptions. First of all, the model assumes that the EVT-1 shock realizes only after decisions on language skills and thus language learning is origin country- but not individual specific. However, in reality individuals have varying migration expectations and are more likely to invest in language skills relevant to destination countries with higher expected probabilities. Hence, effects are 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 expected migration probabilities. Moreover, we assume that the EVT-1 shock is i.i.d distributed, which is a very strong assumption. In reality, preferences on the individual level are driven by preferences for (unobserved) country characteristics, which plausibly are not randomly distributed. Relaxing this assumption introduces an error term that may be correlated with the regressors in Equation 8. We come back to potential violations of the iid 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, 2019). 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 courses for Japanese for English speakers, 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.

Ultimately, our model does not touch upon several facets that are relevant to language learning and international migration. First, as the skill-specific distribution of migrants has profound impacts on sending and receiving countries, self-selection of migrants has gained a prominent place in the migration literature (Borjas, 1987). Second, the model does not consider post-migration language learning. As Duolingo courses enable low-cost language learning both before and after migration, this poses an interesting dynamic decision problem of prospective migrants deciding on the relative timing of language learning and migration. However, in absence of longitudinal data on prospective migrants' language learning to estimate such a model, we do not pursue this.

In the following, we discuss how the model can be brought to the data. The coefficient of a regression of language skills and migration odds on a measure of Duolingo availability consists of parameters of the cost function as well as returns to language skills. As these strongly vary across origin countries dyads, the

rollout of a Duolingo module has strongly divergent impacts across dyads. In the following, we discuss how to proxy for returns to language skills, and to bring Equation 7 and 8 to the data.

4.2 From Model to Empirical Strategy

Proxying returns to language skills Returns to language skills play an essential role in the model outlined above. However, these returns are not comprehensively measured across countries and languages. Nevertheless, the literature has provided estimates in several salient case studies. 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.[22] Returns to English has been found to be related to 10-50% higher earnings across countries.[23] 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.[24] An additional caveat to many of these studies of returns to skills is that they are correlational and if they credibly estimate causal effects, they identify local treatment effect, rather than average treatment effects across the population. The latter is the return that best proxies the average returns for in the uniform-population setting of our model.

In absence of comprehensive estimates of returns across dyads and languages, we proxy the return by estimating the probability a Duolingo course between S and T enables communication between two randomly drawn individuals in o and d, DL_{od}^{ST}.[25] Using the distribution of speakers of languages across countries and the (strong) assumption that languages within countries are independently distributed, we 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.[26] This approach can be implemented within ($o = d$) and between ($o \neq d$) countries, as well as for spoken and official target languages. In the special case that o and d do not share any languages before the introduction of a course, this measure is simple equal to the share of target language speakers in the destination α_{dT}. The larger the overlap between

[22] 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.

[23] 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).

[24] 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).

[25] It seems plausible that the ability to communicate with more individuals in the society one works in. Nevertheless, earnings may not be proportional to the number of people one can communicate in the society of residence. We study the non-parametric relation in robustness tests.

[26] This approach is also followed by Melitz and Toubal (2014), who 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.

languages between o and d, the lower the potential gains from a given Duolingo course. Moreover, 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. In Appendix A.3 I discuss how to calculate this object in the general case. Contrary to returns to language skills, this does depend on the source language. The presence of multiple Duolingo courses with relevant source and target languages poses a challenge when studying language learning across language pairs or migration along country dyads, which we discuss below.

We obtain the share of speakers 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.

We implement this approach for spoken source languages, as speaking a language simply allows communication in the language as well as using it to learn other languages. For target languages, we calculate it for spoken and official target languages. The reason for the latter is that official languages may better reflect returns to language skills b_{odT}. For example, spoken languages include minority languages and foreign languages, which have limited use on the country's labor markets. For target languages that are not widely spoken outside of native-speaking countries, in other countries domestic returns are likely close to 0, whereas destination-country returns are large if T is a widely spoken language in d. $b_{odT} > b_{ooT}$. Therefore, it is plausible that $b_{ooT} > b_{odT}$ for some dyad-years, $b_{ooT} < b_{odT}$ and both zero for others. Nevertheless, not every language is created equal. As mentioned in Footnote 23, returns to some languages may be rather large irregardless of the number of speakers in the country of residence. This particularly pertains to English as a lingua franca, but may also be the case for German, French, Russian and Spanish. Hence, these may have non-zero b_{ooT} and b_{odT} for a wide set of dyads. Nevertheless, b_{odT} is probably smaller than b_{ooT} when the target language is a lingua franca that is not a national language, although one could reap benefits from knowing English without knowing the national language in many countries.

Language learning and language skills Equation 7 suggests that equilibrium language learning and skills depend on domestic and foreign motives for language learning. To test the existence and (relative) strength of these channels, we study the effects on interest in language learning and language skills. We rely on data at the language pair level over time as well as on the origin by target language level. The latter requires aggregation, as multiple Duolingo courses using different source languages may enable learning to a specific target language. Using the proxies for returns to language skills, $b_{ooSTt} = k_o DL_{oo}^{ST}$ and $b_{odSTt} = k_f DL_{od}^{ST}$,

we can bring equation 7 to the data. We construct the Duolingo exposure as such:

$$DL^{h}_{oTt} = \max_{S} DL^{ST}_{oo} \alpha_{oS} Duolingo_{STt} \tag{9}$$

$$DL^{a}_{oTt} = \max_{S} \sum_{d} \mathbb{P}_{od} DL^{ST}_{od} \alpha_{oS} Duolingo_{STt} \tag{10}$$

In the presence of multiple Duolingo courses enabling target language learning from different source languages, we take the source language with the largest exposure. Although this can underestimate the probability an individual from o can use Duolingo to learn T, this is often a good approximation of total exposure. Appendix Figure X shows that in most cases the largest exposure exceeds the second largest exposure. Using these measures we estimate the following regression equations:

$$s_{oTt} = \beta_1 DL^{h}_{oTt} + \beta_2 DL^{a}_{oTt} + \phi_{oT} + \theta_{Tt} + (\psi_{ot}) + \epsilon_{oTt} \tag{11}$$

s_{oTt} denotes measures of language learning or skills in language T for individuals from country o to country d at time t. DL^{h}_{oTt} and DL^{a}_{oTt} denote the time-varying exposure to Duolingo, which measure the probability a Duolingo module enables communication to T at home and abroad, respectively. ψ_{oT} captures unobserved origin country-by-target language level unobserved factors. ϕ_{ot} and θ_{dt} indicate origin-year and target language-year fixed effects that capture unobserved heterogeneity at those levels. Following the model, we expect both domestic and foreign motives for language learning to increase language learning and thus skills, $\beta_i > 0$. The ratio between β_1 and β_2 identifies the relative strength of domestic and foreign motives for language learning. Nevertheless, as we will deal with English as one destination language, our measure of returns to skills may have limited power to capture returns to skill. Hence, when possible, we do analysis for English separately. We bring the language-skill decision to the data in section **??**, using country-specific online search interest in Duolingo, specific target languages and language test scores for English, the target language with most Duolingo courses.

Migration In a similar fashion, I bring Equation 8 to the data. Here, we have four distinct terms related to the availability of Duolingo modules, which we operationalize in the following way:

$$DL^1_{odt} = \max_{S,T} DL^{ST}_{od} DL^{ST}_{oo} \alpha_{oS} Duolingo_{STt}$$

$$DL^2_{odt} = \max_{S,T} DL^{ST}_{od} \left(\sum_d \mathbb{P}_{od} DL^{ST}_{od} \right) \alpha_{oS} Duolingo_{STt}$$

$$DL^3_{oot} = \max_{S,T} DL^{ST}_{oo} DL^{ST}_{oo} \alpha_{oS} Duolingo_{STt}$$

$$DL^4_{oot} = \max_{S,T} DL^{ST}_{oo} \left(\sum_d \mathbb{P}_{od} DL^{ST}_{od} \right) \alpha_{oS} Duolingo_{STt}$$

Here, we aggregate over multiple source languages, as in the previous section, as well as over multiple target languages. Appendix Figure X shows that in most cases the largest exposure is considerably larger than the second largest exposure. We estimate Equation 8 by stacking all origin countries o, exponentiating both sides 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 staggered variation with continuous treatment intensity:

$$\frac{M_{odt}}{M_{oot}} = \exp\left[\beta_1 DL^1_{odt} + \beta_2 DL^2_{odt} + \beta_3 DL^3_{oot} + \beta_4 DL^4_{oot} + \gamma' \mathbf{X_{odt}} + (\phi_{ot}) + \theta_{dt} + \psi_{od}\right] \eta_{odt} \tag{12}$$

M_{odt} is the number of GWP respondents desiring to emigrate from country o to country d at time t, M_{oot} is the number of GWP respondents not desiring to emigrate from country o. Here, $Duolingo_{odt-1}$ is a the exposure to *Duolingo* courses (see above) the year before. We take the first lag of the *Duolingo* exposure because a course made available during year $t-1$ is surely available to respondents of the GWP at first at t. As the mass of GWP interviews is in the middle of the $\mathbf{X}_{odt-1}$ include a dummy for joint EU membership, and WTO trade agreements. 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[27] and to account for bias due to the inward and outward multilateral resistances.[28] As some of the terms following from the model only vary at the origin-time level, we estimate models with and without ϕ_{ot}. η_{odt} is an error term with unit mean.

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

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

4.3 Identification

The identification strategy underlying the estimation equations 11 and 12 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), we need the following identifying assumptions. First, we assume that there are no anticipation effects. It seems plausible that language learning and migration intentions are not affected by the future availability of Duolingo courses. Second, we assume parallel trends. For the additive model in equation 11, we assume that origin-target language pairs follow parallel trends in levels and for the multiplicative model of 12 of origin-destination pairs follow parallel trends, conditional on the covariates and fixed effects. In other words, language skills in treated origin country-target language pairs would have followed parallel trends to untreated (including not yet treated) pairs in absence of the treatment. Likewise, the *growth* in the odds ratio of migration would have been the same in treat and untreated the absence of any treatment.[29]

The rollout of language learning courses across dyads may not be exogenous to trends in language learning and the formation of migration aspirations and actual migration plans. This potentially invalidates the parallel trends assumption underlying my empirical design. However, there are three reasons to believe that this is likely not the case for *Duolingo* courses. First of all, course rollout likely did not depend on trends in leaning nor in migration intentions, but rather on levels. Second, trends before roll-out in all outcomes are parallel. Third, our results on migration intention are robust to omission of the largest countries by source and target languages. First, as the *Duolingo* platform was released in 2012, the rollout of courses was initially supply-constrained as courses had to be developed from scratch. Courses are thus plausibly developed based on the current demand for language learning, rather than on the growth rate. In the next section, we assess what predicts course rollout on the country pair level and the level of exposure on the country pair level. Even if *Duolingo* anticipates future trends in the demand for language learning, it would require the foresight that migration intentions are on the rise or factors affecting migration aspirations. Another feature of *Duolingo* supports this argument. As Duolingo, contrary to other competitors, only makes courses available to the public once there is a considerable amount of content available, the development stage of the average course takes several 100s of days. The primary motive of language learning on *Duolingo* is often different for 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

[29] Note that this differs from the parallel trends assumption in the triple differences case (Gruber, 1994; Olden and Møen, 2022). This assumption works on the dyad level where treatment is not only assigned in the two cross-sectional dimension, instead of the case where one dimension is used as a pure control dimension.

tourism-related reasons) and 8.4% because of other reasons (which could include migration-related reasons). Those numbers are comparable among Spanish- and French-learners. Hence, it seems unlikely that the reverse causal path of language courses being developed because *Duolingo* anticipates increases in migration along a course's migration corridors plays a substantial role. Second, there are no pre-trends in event study estimator: In an event study around substantial course rollouts (defined as a change in $Duolingo_{odt}$ of 0.7 or larger) there are no significant pre-trends, but I do find a strong upward sloping post-trend response. The absence of significant pre-trends in migration aspirations indicates that migration aspirations developed similarly between treated and control units before course roll-out. Furthermore, I use event study estimators that are robust to the issue of negative weights in staggered two-way fixed effects regressions. As our identification strategy relies on staggered introduction, this is in principle susceptible to the issue described in de Chaisemartin and D'Haultfœuille (2019), among others. The event study estimators suggest that this is not driving our results. Third, all results are robust to removing the contribution of the country with the most speakers for each language, for both the source and target language. This alleviates concerns that specific language courses are developed to enable language learning for prospective emigrants along a specific migration corridor.

4.3.1 What predicts course development?

To study what determines course rollout and subsequently Duolingo exposure, we analyze the timing of rollout on the language-pair and country-pair level. For the language-level analysis, we regress (1) a binary indicator of whether a course has been rolled out by the end of our sample period, 31st of December 2023, and conditional on rollout, (2) the year of rollout on (bilateral) characteristics of languages. Second, we perform a similar analysis using the dyadic exposure measure on the country pair level and the year the Duolingo exceeds 0.2. 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 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 and later rollout. Importantly, 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 does not exclude dynamic selection into treatment. We partially study this by assessing pre-trends in migration trends in subsequent sections.

Table 1: The Determinants of the Rollout of *Duolingo* Cou

	(1) $Duolingo_{23}^{ST}$	(2)	(3) Year of rollout	(4)	(5) $Duolingo_{od23}$	(6
	(1) Course available	(2) Course available	(3) Year of rollout	(4) Year of rollout	(5) Exposure in 2022	(6) Exposure
Log source speakers	0.004**		-0.051			
	(0.002)		(0.265)			
Log target speakers	0.003**		-0.447**			
	(0.002)		(0.175)			
Log source speakers × Log target speakers		0.002**		0.460		
		(0.001)		(0.267)		
Sharing an official language					-0.198***	-0.20
					(0.026)	(0.0
Log population-weighted distance					0.053***	-0.02
					(0.009)	(0.0
Log GDP pc PPP in origin					0.038***	
					(0.009)	
Log GDP pc PPP in destination					0.026***	
					(0.008)	
Log bilateral migrant stock + 1 (2005)					0.009***	-0.0
					(0.002)	(0.0
Observations	13225	13225	84	52	22005	220
Source and Target FE		✓		✓		
Origin and Destination FE						✓

OLS regressions of language-pair and country-pair level exposure and year of rollout on language- and country characteristics. We include all N source a (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 t all source and target languages appearing only once. Similarly, column 5 and 6 regress the measure of Duolingo exposure in 2023 on country and dy regress the year in which exposure to Duolingo first exceeds 0.2 on country and dyadic characteristics, and column 8 adds origin- and destination fixed e 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 bor 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 obta on country characteristics, except linguistic proximity, is obtained from CEPII. Linguistic proximity is obtained from (Adsera and Pytlikova, 2015).

4.4 Estimation

We rely on two distinct but complementary estimation strategies for both 12 and 11. First, we use use all available variation using a standard three-way Second, we acknowledge that N-way fixed effects regressions of staggered difference-in-differences settings as ours can give biased results in the presence of heterogeneous effects, and we employ robust event study estimators around large increases in Duolingo exposure. This additionally provides canonical test of the presence of pre-trends before roll-outs.

First, we estimate equation 11 by OLS with three-way fixed effects regressions. To estimate treatment effects on dyadic data such as trade and migration flows, scholars typically estimate gravity models by Pseudo-Poisson Maximum Likelihood (PPML) rather than log-like transformed outcomes for two main reasons.[30] 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, 2023). Second, heteroskedasticity can introduce bias in non-linear models (Silva and Tenreyro, 2006). An often overlooked but important difference of the Poisson model is that the estimand it targets is different from that of OLS on a log-like outcome, as

[30]I borrow the term log-like from Chen and Roth (2023), 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.

noted by Tyazhelnikov and Zhou (2021) and Chen and Roth (2023).[31] A drawback of this estimand is that it weights a unit increase similar irregardless of where it happens on the distribution. Hence, we show effects OLS on log+1 transformed outcomes too. The (panel) Poisson estimator is the only non-linear estimator that is not biased because of the Incidental Parameter Problem (IPP). Nevertheless, Weidner and Zylkin (2021) have shown that the three-way FE does face IPP to some extent and develop a correction to the prevailing bias. Although this bias is limited in most cases, we report

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, 2018; de Chaisemartin and D'Haultfœuille, 2019, 2020; Sun and Abraham, 2021; Borusyak, Jaravel and Spiess, 2021; Callaway and Sant'Anna, 2018; Wooldridge, 2021). Hence, in extreme cases, the presence of heterogeneous and dynamic 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., 2022). Strezhnev (2023) and Nagengast and Yotov (2023) have pointed out that this problem naturally extends to the three-way fixed effects setting.[32] Nagengast and Yotov (2023) provide a solution which resembles the approach of Wooldridge (2023).[33]

[also note on what we do for language learning] Hence, I will show 3WFE effect estimates as well as event study estimators around sharp increases in Duolingo exposures instead. Although not all increases in Duolingo are very sharp (as the case of Germany → Chile in Table **??**), the vast majority of large increases is large and does not increase much after the first treatment.

This approach estimates a Poisson regression with dummy variables for every treated cohort-time cell rather than a single indicator for treated units.[34]

[mention that log odds does not have simple MR term]

An additional estimation concern for studying migration-related outcomes arises from the strong assumptions underlying our model. We make the strong assumption that the discrete choice problem fulfills the Independence of Irrelevant Alternatives (IIA) assumption. However, in reality, individual-level preference

[31] Chen and Roth (2023) calls $e^{\hat{\beta}} - 1$ the average proportional treatment effect on the treated. OLS and PPML target the same estimand if treatment effects are ex-ante homogeneous, but not when treatment effects are heterogeneous.

[32] To assess the degree to which negative weights occur, we can estimate the weights with the Stata-command TWOWAYFEWEIGHTS (de Chaisemartin and D'Haultfœuille, 2019)

[33] Harmon (2022) has shown that, although the alternatives to two-way fixed effects estimators are unbiased, the efficiency of the proposed estimators may differ based on the nature of the data generating process. Regression imputation-based estimators such as that by (Wooldridge, 2023) are inefficient in cases where innovations in outcomes are persistent. Other estimators only use the last pre-treatment period to calculate long differences to the respective time horizon of interest. Although similar concerns are present in our setting, there is no long-difference based Poisson estimator available.

[34] The approach is slightly more complicated when including time-varying covariates. However, due to the dimensionality of the variation and the included fixed effects, there are arguably few dyadic time-varying factors that are not potential bad controls. Hence, I do not discuss the case with covariates here.

shocks for different destinations are not independent. Bertoli and Moraga (2013) show that this generates additional terms in the error term, giving rise to endogeneity and generate complicated spatial correlations in the residual. Bertoli and Moraga (2013) conceptualize the multilateral resistance term, by assuming that shocks are correlated in (potentially overlapping) nests m of destination countries. This can reflect that two or more countries in the same nest may be close substitutes.[35] In the general case, the multilateral resistance term takes the form:

$$r_{odt} = \ln \left(\sum_{m} \left(\alpha_{odt}\right)^{1/\tau} \left(\sum_{d' \in b_m} \left(\alpha_{od't} e^{V_{od't}}\right)^{1/\tau} \right)^{\tau-1} \right) \tag{13}$$

In limiting cases, this term does not vary by dyad over time. One such case, as (Ortega and Peri, 2013) and (Bertoli and Moraga, 2013) discuss, assumes that there are only two nests: one for staying, and one for all destinations. Then $r_{odt} = r_{ot}$, as the value is identical for all destination countries, which can be subsumed by including origin-time Fixed Effects. Alternatively, assuming that the nesting structure and the deterministic component of utility for destinations and time is identical acros origin countries, $r_{odt} = r_{dt}$, which can be subsumed by destination-time fixed effects. In our main estimation equation, we include both origin-time and destination-time fixed effects, but this does not dispel all concerns. A potential solution could be to include origin-time-destination fixed effects, where nests are chosen based on relevant observable characteristics of destination (as in (Beine, Bierlaire and Docquier, 2021) for individual-level data). However, this may absorb a large part of the variation, especially when nests overlap with destination country languages. Furthermore, the spatial correlation in errors that multilateral resistance introduces, which I further discuss in section 4.5.

To see how this can effect our estimates, consider the following example. Duolingo availability to Spanish increases the attractiveness of all Spanish speaking destinations. However, the Spanish speaking countries are also close substitutes and preference draws of individual i for these destinations is correlated. Hence, the effect size is underestimated because the pulling effect all these destinations exert on each other makes the migration rate increase less than if only one destination could have been treated. This would bias our estimates towards 0. In similar vein, if a Duolingo course becomes available to people in Spanish speaking countries, this may increase the willingness to emigrate. However, it may be that because of visa restriction in destination countries, the different destination countries compete for the same places and exert a downward effect on each other.

Hence, multilateral resistance likely introduces terms that are negatively correlated to, leading to a downward bias in our estimates. To nevertheless reduce concerns, we follow (**?**) and introduce origin-destination nest-year fixed effects across various dimensions and test whether cross-sectionsl dependence is

[35] This can be driven by several factors. For example, different groups of people desiring to different countries in the origin country may be subject to similar shocks.

still present.

4.5 Inference

To account for the possibility that not all observations are independent, heteroskedasticity-robust standard errors may overestimate the precision of our estimates. As our treatment is assigned at a more aggregate level than the unit of analysis, conventional knowledge is to cluster at the level of treatment assignment (Abadie et al., 2023). However, in dyadic data such as ours all flows from o to d can be correlated to all dyads where either o or d are represented. Cameron and Miller (2014) suggest a *dyadic-robust* variance estimator. In practice clustering two-way at both the sending and receiving unit gives comparable standard errors. In the following, we cluster at sending and receiving unit because of the unavailability of dyadic-robust variance estimators in conventional statistical packages.

Geographically, economically or culturally close countries often speak the same language or languages from the same language family. Due to their proximity these countries could be subject to common shocks, and observations may not be fully independent. To account for this, I cluster standard errors on the language family pair level. (Adsera and Pytlikova, 2015) collect detailed information on language families of the main language by country from *Ethnologue*. We assign languages to the same language family if they belong to the same third level of the linguistic tree of *Ethnologue* (such as English and German being West Germanic languages) as used by Adsera and Pytlikova (2015). We extend their data with data from *Glottolog* to complete the language family classification for countries absent in (Adsera and Pytlikova, 2015). This gives 17 distinct language families on our baseline sample.

5 Language Learning

In this section, I study the impact of Duolingo course rollout on take-up, language learning effort and language skills. First, I study the determinants of course take-up. 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. These can test the aggregate predictions of the model. These study something different from studies examining the efficacy of Duolingo and other online language learning technologies. Such studies consider language learning outcomes in controlled environments and look at short run outcomes. Instead, I study whether online language courses affect aggregate real-world outcomes when low-cost language learning becomes available.

5.1 Course take-up

To study the determinants of take-up, we rely on the number of learners by language course, which is continuously provided by Duolingo. Figure A5 and A6 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 A5 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 English is the most learned second language and takes a special role in the global economy as a *lingua franca*. Figure A6 shows that among 15 widely spoken languages, the number of Duolingo learners by origin country exceeds 5% of speakers. Several rarely languages have attracted many learners relative to the number of speakers, such as Welsh, Norwegian and Hebrew. Moreover, several more widely spoken languages have garnered a relatively large number of speakers for tourism and cultural reasons, such as Greek, Italian, Japanese, and Korean. Even among more commonly spoken language, such as English, Spanish and French, the number of learners exceeds 10% of the number of speakers.

Table A1 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 0.8 for every 100 source language speakers as well as target language speakers.[36] This shows that not only the total pool of potential learners is relevant in the decision to learn a language, but also the applicability of the target language. The positive interaction effect between the number of source and target language speakers in column 2 also suggest that demand for courses is particularly high between languages with many speakers, such as English and Spanish.

It is a priori unclear how low-cost language learning interacts with traditional (in-class) language learning. In Appendix D.1 I find that the rollout of Duolingo decreased course participation, but has not affected exam participation. This is in line with online language learning substituting for learning, but not for certification, which may be needed for visa or employment.

5.2 Interest in *Duolingo* and available languages

The availability of a Duolingo course induces interest in the platform Duolingo in countries where the source language is spoken. Moreover, it may generate interest towards the source languages offered. As online search behavior is a useful proxy of the interest in a topic, I collect information about relative search intensity on the widely used search engine *Google*. The Google Trends API enables one to query time series of normalized

[36] This is the slope of users to speakers. The percentage of learners in July 2024 of source language speakers is 2%. However, because of existing users quitting Duolingo and users starting courses, the total number of ever learners likely exceeds this figure.

search intensity of a search term or topic over time, referred to as the Google Trends Index (GTI).[37]

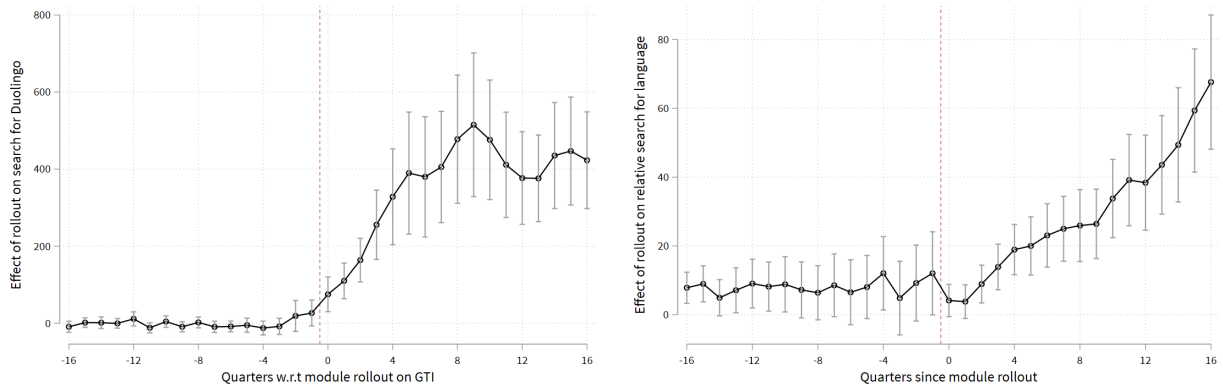
We collect GTIs from all countries for the search term Duolingo and common transliterations,[38] as well as for the *topics* of the 50 most spoken languages (plus 6 languages which are available as a target language in Duolingo but are not among the 50 most spoken languages.). To interpret absolute changes in the GTIs as equal increases in relative search intensities across countries, I always scale the GTI with the GTI across regions. We denote the resulting scaled GTI for term T in country o at time t by (GTI_{ot}^{T}) and the GTI scaled across countries but not across terms as $GTI_{ot}^{\widetilde{T}}$. In the following, I use the scaled Google Trends Index to study whether course introduction spurred interest in Duolingo as well as languages.

Interest in Duolingo As Duolingo was founded in 2011, relative global interest went from practically zero to large values (see Appendix Figure A11a). Across the whole period interest for Duolingo is nonzero in 75 countries and highest in Guatemala. 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 0.5 or larger in Duolingo exposure on the origin country-level. Figure 2a shows the development of interest in Duolingo around 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. As in some cases Duolingo was available as a beta version before the course was released, there was some interest the two 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.

[37] For a detailed description about how the time series are constructed, Google Trends is queried for search terms and topics, see Appendix F. Importantly, the Google Trends Index is a measure of relative search intensity relative to the highest search intensity in the queried time series. We obtain a normalized measure of relative search intensity comparable across geographic regions and search terms.

[38] I add the transliterated counterpart of "Duolingo" in the non-latin languages with Duolingo source courses: Arabic, Chinese, Japanese, Korean, and Ukrainian and Russian. Duolingo is not transliterated in Greek, Hindi, Thai and Vietnamese.

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



(a) Duolingo

(b) Languages

Estimates from staggered introduction of Duolingo modules on (a) search interest in Duolingo and its transliterations across countries and (b) search interest in target languages across countries using the Borusyak, Jaravel and Spiess (2021) estimator on a quarterly level. The two-way fixed equivalent of the event study corresponds to a regression in (a) of the GTI in Duolingo on origin country and quarter fixed effects and an aggregation of the Duolingo exposure similar to that in Equation 12 $GTI_{ot}^{Duolingo} = \beta \mathbb{1}\left((\max_{S,T,d} \alpha_o^S \alpha_d^T DL_t^{ST}) > 0.5\right) + \psi_o + \phi_t + \epsilon_{ot}$ and in (b) of the GTI in target languages on origin-language, origin-quarter and target-quarter fixed effects: $GTI_{ot}^{T} = \beta \mathbb{1}\left((\max_{S} \alpha_o^S DL_t^{ST}) > 0.5\right) + \psi_{oT} + \phi_{tot} + \theta_{Tt} + \epsilon_{oTt}$.

Interest in target languages As Duolingo courses enable to learn a particular language, it could not only increase the interest in Duolingo, but could also spur the interest in the specific language. Average relative search interest in foreign languages was stable since 2009, but started increasing in 2016 (see Appendix Figure A11b). 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. We 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. Compared to Figure 2a, this has the benefit that I can account for three-way fixed effects that partial out all unobserved heterogeneity on the origin-year, destination-year and country pair-level. We find that interest in the language starts increasing two quarters 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, as shown in Figure 1. These two exercises also validate that the rollout dates, based on the date courses enter the final phase, well capture the relevant timing of course introduction. Moreover, the latter results provide evidence that Duolingo impacts search behavior towards available languages. This implies that course rollout increases either the number

of people interested in the target language, or that learners search more intensively for information related to the target language. This suggests that Duolingo courses impact language learning beyond the Duolingo platform.

5.3 Language skills

The previous sections have shown that many people took up Duolingo modules, and spurred interest in Duolingo and avalable target languages. A pressing question is whether access to low-cost language learning also improved language skills among the general population. However, internationally comparable data on foreign language skills among the general population is scarce.[39]

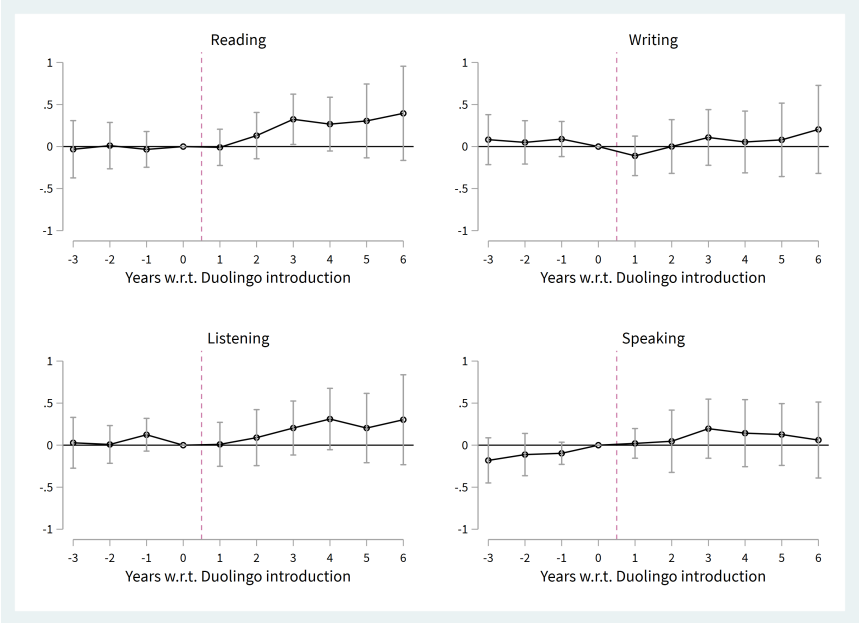
The best available alternative to representative data are scores from widely-used tests. The Test of English as a Foreign Language (TOEFL) test organized by ETS, which is an English language test taken annually by more than 2 million individuals across the world, mostly for (foreign) university enrollment. 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 have worse skills at the time of the test than the rest of the populaion. This concern is somewhat alleviated with the high-stakes TOEFL test participants, 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, so comparing changes in passive language skills among test takers and active changes can provide a partial test of the influence of low-cost language learning on language skills.

ETS compiles yearly reports of average section scores by language and by origin country. We use the former as it naturally maps to the structure of Duolingo courses. Figure 3 shows the results of event studies of test scores around the introduction of a Duolingo course (using the estimator of de Chaisemartin and D'Haultfœuille (2019)). In the first two years before course rollout there are no discernible pretrends. In the first five years after course rollout, scores have increased by 0.29 for reading, for writing by 0.07, for listening by 0.22 and for speaking by 0.08. The long term effect after 5 years for reading raises to 0.5, due to the gradual increase. To put this coefficient into perspective, using a linear model with native language fixed effects, the average reading score has improved by 0.13 points per year between 2007 and 2022. 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.

[39]The EU Adult Education Survey, a representative survey across the EU, includes questions about language knowledge and skills among. In a next version of this paper, I will use the 2011, 2016 and 2022 waves to study the effect of online language learning availability on language skills among the general population

This exercise does not identify the effect of Duolingo on the average population for two important reasons. First, as TOEFL takers are more likely to have studied English, they are much more likely to have used Duolingo. Second, Duolingo availability may affect who takes the test. The latter is unlikely to have an upward effect on test scores. As Duolingo enables the study of English language at beginner and intermediate levels, it seems unlikely only more proficient individuals take the test.

Figure 3: The Effect of Duolingo Rollout on Component Scores of English language (TOEFL) test (2007-2021)



Notes: Results from a de Chaisemartin and D'Haultfœuille (2019) event study estimator of TOEFL test scores around Duolingo course rollout. Standard errors are clustered at the language level. We use TOEFL data from 2007 up to and including 2022. N = 1,134.

6 Migration Aspirations and Flows

In this section, I study whether the staggered introduction of low-cost language learning has impacted migration flows. As bilateral migration flow data is only available on a yearly level for a limited number of countries, this section will mostly rely on migration intention as elicited in the Gallup World Poll (GWP). Bilateral migration intentions as elicited in the GWP are strongly correlated to migration flows across countries and are predictive of subsequent migration (Tjaden, Auer and Laczko, 2019). Hence, bilateral migration intentions are an important probe of individuals' future plans.

6.1 Data

We use the 2007-2021 vintages of the Gallup World Poll (GWP), which is a representative survey of about 1000 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).[40] When individuals mention that they desire to emigrate, they also indicate where they would like to migrate.[41] Using this question and data on the population of these origin countries, I construct the estimated stock of desiring to emigrate from country o to country d in year t.[42] Hereafter, I call this the stock of aspiring migrants. We use the GWP between 2007 to 2023, comprising of more than 1 million interviews across 156 countries, which are visited on average 11.5 times across the 17-year period.

The OECD bilateral 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 from many origin countries (in some countries administered by nationality, in others by country of birth). We focus on the time period from 2007 to 2019. We exclude later data because of the large influence of the Covid-19 pandemic on international mobility.

In order to estimate gravity models of international migration, I use the database by Conte, Cotterlaz and Mayer (n.d.). This dataset includes all important variables to estimate gravity models up to and including 2020: trade flows, trade agreements, geographical distances, macroeconomic indicators, from a variety of original sources.

6.2 Migration Aspirations

Table 4 shows the main results from the model introduction in Equation XXX.

Column 1 includes country pair and destination-year fixed effects, whereas Column 2 additionally include origin-year fixed effects. Column 3 destination- and origin-year fixed effects, and Column 4 includes all 3-way fixed effects. The point estimate in our preferred specification in Column 4 suggests that a language course from a language spoken by the full population in the origin to a language spoken by the full population in the destination increases bilateral migration desire by 21 percent. The exposure for the median dyad with nonzero exposure is 0.32, suggesting that a typical language course increases the desire to emigration by 7

[40] 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**. In the latter two cases I discard the observations.

[41] 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.

[42] We 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 respondents N_{ot} and the share of respondents aspiring to emigrate from o to d, N_{odt}: $M_{odt} = pop_{ot} N_{odt}/N_{ot}$.

Table 2: The Effect of *Duolingo* Courses on Bilateral Migration Aspirations

	(1)	(2)
DL_odt^{abroad}	0.273***	0.400***
	(0.070)	(0.083)
DL_$odt^{domestic}$	-0.427**	
	(0.200)	
Observations	123484	123484
Unique origin countries	153	153
Unique destination countries	193	193
Unique dyads	10663	10663
Origin-destination FE	✓	✓
Origin-year FE		✓
Destination-year FE	✓	✓

Estimated by PPML. Standard errors are clustered at origin-destination level. 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.

percentage points. We can correct the estimates of column (4) for bias in three-way fixed effects regressions caused by the incidental parameter problem Weidner and Zylkin (2021) and the results remain very similar. Table A5 and A6 show duplicates of Table 4, but with standard errors clustered two-way, at the origin and destination country level, and at the linguistic pair-level (as discussed in section 4.5). In both cases, the results in Column 4 are still highly significant.

At first glance, these may sound as incredibly large effect sizes. However, one has to take into account that these are average partial effects on the odds ratio. The average share of individuals desiring to emigrate in the GWP data lies around 0.20, and as there are about 200 alternative options,

At an average odds ratio of 0.0016, an increase of

Table 3: The Effect of *Duolingo* Courses on Bilateral Migration Aspirations

	(1)	(2)	(3)	(4)
DL_$odt^{S=EN,abroad}$	0.772***	0.899***		
	(0.205)	(0.208)		
DL_$odt^{S!=EN,abroad}$	0.194**	0.295***		
	(0.082)	(0.100)		
DL_$odt^{S=EN,domestic}$	-0.248			
	(0.241)			
DL_$odt^{S!=EN,domestic}$	-0.434*			
	(0.256)			
DL_$odt^{T=EN,abroad}$			0.186**	0.581***
			(0.092)	(0.092)
DL_$odt^{T!=EN,abroad}$			0.239**	0.106
			(0.106)	(0.092)
DL_$odt^{T=EN,domestic}$			-0.507*	
			(0.285)	
DL_$odt^{T!=EN,domestic}$			-0.074	
			(0.221)	
Observations	124878	123484	124878	123484

Estimated by PPML. Standard errors are clustered at origin-destination level. 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.

6.2.1 Event Study around Introduction of Influential Courses

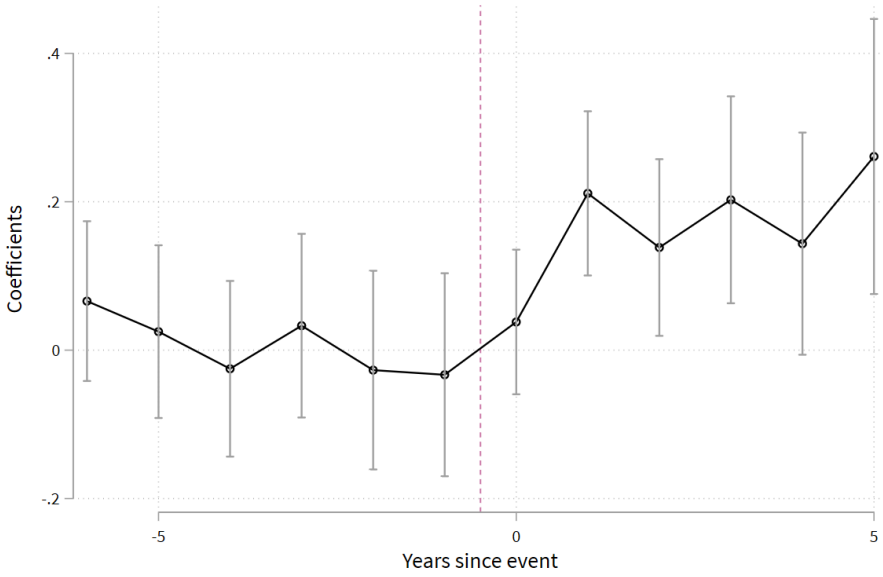
As I would like to infer how the effects of *Duolingo* courses play out over time, to assess the plausibility of the parallel trends assumption by considering pre-trends, and to alleviate the concerns that our results are driven by negative weights in staggered difference-in-differences settings (Goodman-Bacon, 2018), I use the estimator by Borusyak, Jaravel and Spiess (2021). Contrary to other recently developed estimators developed to deal with forbidden comparisons in staggered difference-in-differences (de Chaisemartin and D'Haultfœuille, 2019; Callaway and Sant'Anna, 2018; Abraham and Sun, 2018), the imputation-based estimator by Borusyak, Jaravel and Spiess (2021) can account for the whole set of 3-way fixed effects. As this estimator requires a binary treatment, I define an event as a dyad receiving an increase of 70 percentage points in the *Duolingo* exposure measure, and all others as a control group (note that there are many regions that are treated between 20 and 70pp, so this is not the cleanest control group).

The lth pre-trend coefficient of the Borusyak, Jaravel and Spiess (2021) estimator essentially indicates the regression-adjusted difference in levels between treatment units (receiving treatment exactly l periods

later) compared to control units (never or not yet receiving treatment l periods later). This differs from regular two-way fixed effects event studies, which always identify effects with respect to some pre-treatment period. As there is no readily available solution to the negative weighting problem for staggered introduction in nonlinear diff-in-diff settings with many fixed effects,[43] I use $\ln M_{odt} + 1$ as our explanatory variable of interest in a linear model.

Figure 4 shows the Borusyak, Jaravel and Spiess (2021) estimates of the analysis. We find no effect before the introduction of such an influential course, but an increase one period after the introduction, that increases thereafter. This is in line with (i) course adoption needing some time, (ii) the fact *Duolingo* that *Duolingo* became only and (iii) a heterogeneous effect where earlier treated units may have higher treatment effects.

Figure 4: Event study around 70pp increases in Duolingo$_{odt}$ on Desire to Emigrate (2007-2021)



Note: Standard errors clustered at origin-destination level.

6.3 The Role of English

English is both the language with most courses as a target language (22), most all-time learners (around 150 million) and is arguably the language with the largest labor market benefits as a foreign language. Hence, results may be driven by English as a foreign language and English-native speaking destinations. To assess this possibility, I show the results omitting the six high-income countries as a destination country in Figure

[43] In a future version of the paper, I want to estimate the model of equation 12 using the ETWFE estimator developed by Wooldridge (2022), which enables to estimate staggered difference-in-differences in non-linear models.

?? at a time does not render the results insignificant, but that omitting all of them at the same time does. Nevertheless, the point estimate is still significant.

6.4 Heterogeneity

Table ?? shows the interaction of our main treatment with several other variables. We find that the effect is stronger for destinations further away (this may be strongly correlated by the availability of other languages) and the effect is somewhat stronger within EU- compared to non-EU dyads. Column 3 and 4 show that the effect disappears when both countries share an official language, but not when both countries only share a non-official language share by more than 9% of population. In Column 5 and 6 I find that larger relative stocks in the destination country also reduce the effect: if there is a diaspora in the destination, the need of learning another in the destination becomes lower.

Table 4: The Effect of *Duolingo* Courses on Bilateral Migration Aspirations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
DL_odt^{abroad}	0.347***	0.206**	0.487***	0.400***	0.350***	0.054	0.645***
	(0.079)	(0.086)	(0.101)	(0.083)	(0.097)	(0.233)	(0.137)
DL_odt^{abroad} × GTI for Duolingo (2006-2023)	0.000**						
	(0.000)						
DL_odt^{abroad} × $3G_{oy}$ (Collins)		0.327**					
		(0.153)					
DL_odt^{abroad} × Broadband_{oy} (ITU)		-0.002					
		(0.005)					
DL_odt^{abroad} × Linguistic proximity (AP15)			-0.527**				
			(0.243)				
DL_odt^{abroad} × Shared official language				-0.406**			
				(0.198)			
DL_odt^{abroad} × GDP PPP_oy					0.003		
					(0.003)		
DL_odt^{abroad} × GDP PPP_dy						0.007	
						(0.005)	
DL_odt^{abroad} × Log of migrant stock (2010)							-0.027*
							(0.016)
Observations	121385	89933	114404	122461	118102	115367	122161

Estimated by PPML. Standard errors are clustered at origin-destination level. 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.

Different treatment definitions Table A4 shows how changing the definition of *Duolingo* exposure affects the results. Column 1 sets all treatment exposure from Arabic to other countries in Equation ?? to 0, as Arabic to English, German, French and Swedish was rolled out rapidly during the 2015/6 refugee crisis. The results barely differ from that of Table 4. Column 2 shows the deletion of the single target

language with the most speakers in the construction of the exposure measure: English. We performed the same exercise using all origin- and destination languages and results are robust to that. The point estimate becomes considerable lower, which may not be surprising as English is the most learned language of Duolingo. Column 3 removes the contribution to the *Duolingo* exposure from the destination country with the largest number of speakers for each of the target languages, Column 4 performs a similar exercise by removing the contribution to the *Duolingo* exposure from the origin country with the largest number of speakers for each of the source languages, and Column 5 removes both sets of contributions to the exposure. These results partially alleviate concerns that courses are developed for country pairs that become more integrated in unobserved ways. The point estimate is somewhat smaller, but remains significant.

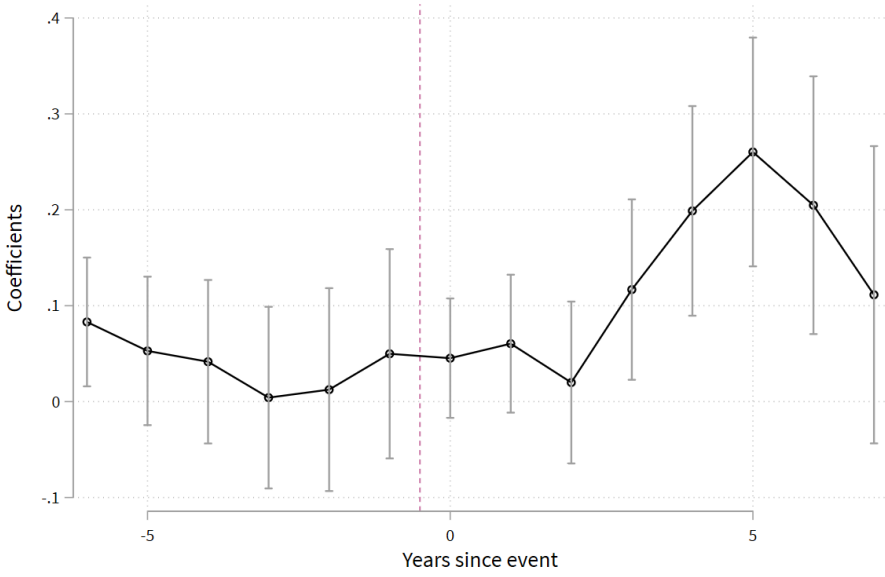
Column 6 normalizes the *Duolingo* exposure by dividing by the sum of *Duolingo* exposure on the origin-time level. When many different courses that enable emigration to many destinations are available, the effect of a given language course is likely to be lower. The results remain highly significant, but the coefficient becomes hard to assign meaning to. Table A4 shows the results for slightly different treatment definitions. All results are qualitatively similar to the main results, except for column 5. This suggests that the results are insignificant if two countries share a common language, which renders language learning irrelevant.

6.5 Migration Flows to OECD countries

To answer the question whether the availability of low-cost language learning impacted *actual* migration patterns, we use the bilateral migration data from the OECD.

We can also perform an event study similar to that in Figure 4 for migration flows to OECD countries. We set a dyad to treated when they face a sharp (70pp) increase in *Duolingo* exposure, using dyads with less than a 30pp increase or no increase as control units. Unsurprisingly, I find that actual migration responds slower to the rollout of courses than migration aspirations: in Figure 5 I show 4 years after introduction of a course actual migration starts increasing. This is roughly in line with Huber and Uebelmesser (2019), who find that introduction of German language learning institutes leads to higher migration after 6 years. Our estimates suggest larger effects than that of Goethe institutes in Huber and Uebelmesser (2019): whereas a single Goethe institutes increases migration flows about 10 percent, a salient Duolingo course increases bilateral flows by more than 20% 5 year after course introduction.

Figure 5: Event study around 70pp increases in Duolingo$_{odt}$ on Migration to OECD countries (2007-2020)



Standard errors clustered at origin-destination level.

7 Migrants' Language Skills, Sorting and integration

[Being aware of my previous results, there is likely a delayed response on actual flows. Nevertheless, this could hide a compositional shift in the first 3 years after treatment. A event study is crucial.]

Not only could accessibility of language learning affect the extent of migrant flows, it could also affect who is migrating and the language skills of those migrating. Moreover, migrants could arrive with better language skills or even learn the host country language faster after migration. In turn, differential selection and higher language skills could influence the pattern of linguistic, economic and social integration. To study these, we turn to two complementary data sets: the EU Labor Force Survey (EU LFS) and American Community Survey (ACS). English is the most learned second language on Duolingo and across the world and the language with the most courses as target language on Duolingo (22). Furthermore, the US is a popular destination country for immigrants originating from many language areas.[44] The ACS surveys 1% of the US population every year, which includes questions on self-assessed language skills every year. However, it concerns only a single target language and destination country, which limits the variation and renders the identifying assumptions weaker. The EU LFS does not face this problem, as it (including the

[44]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/

United Kingdom, Switzerland, and Norway) hosts many languages that are target of Duolingo courses (62), as well as several countries that are not. Unfortunately, the EU LFS does not distinguish the exact origin country of immigrants, but rather a broad origin country group. This allows for an identification strategy relying on variation at the origin group by destination country over time. A further drawback of the EU LFS is that it only incorporates questions on language skills and reasons for migrating in 2014 and 2021. An advantage of the 2021 module is that it also includes questions of self-assessed language skills upon arrival as well as concurrently.

In the following, we will first discuss the EU LFS and ACS and the construction of Duolingo exposure for both data sets. Thereafter, I will present empirical strategies and results concerning the effect of Duolingo availability on upon-arrival migrants' language skills and selection and sorting as well as on after-arrival language skills and integration.

An important consideration in these analyses is that I use a Duolingo exposure purely based on availability, rather than the actual likelihood of use. However, because of the strongly increasing popularity of Duolingo over time, it is plausible that later cohorts are more likely to have adopted it. Also the quality of courses may have improved. As in the previous section, we turn to event study estimators to study the dynamic effect. An important caveat of dynamic estimators of a recent phenomenon is that it does not enable to well disentangle cohorts effects as discussed as well as dynamic effects of the treatment [To continue after results]

7.1 Data and Duolingo exposure

EU LFS 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, labor market participation of households. 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 an of one's citizenship, as well as the years of residence. These regions are discussed in Table **??**. 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,[45] whether one participates in a language course,[46] and one's self assessed contemporaneous language skills[47] in both years, as well as language skills upon arrival in 2021 on the same scale. This has the advantage that I can cleanly separate the effect of pre-migration learning from post-migration learning.

As we are interested in language skills in the receiving country's national language, we only consider

[45] 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

[46] 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

[47] beginner, intermediate, advanced, mother tongue

Duolingo exposure arising from courses to one of the national languages T_d [how do we deal with BE and CH??]. As we are interested in not (yet) naturalized as well as naturalized immigrants, we construct the measure based on an immigrant's country of birth. To overcome limitation of having country groups for birth g, we weight the exposure on the origin country by target language level with the distribution of immigrants in a prior period:

$$DL_{gc}^{T=T_d} = \frac{\sum_{o \in g} m_{odt} DL_{oc}^{T_d}}{\sum_{o \in g} M_{oc}^{T_d}} \tag{14}$$

Here m_{odc} is the migrant inflow in year c from country o to country d.[48] In similar vein to section **??**, $DL_{oc}^{T_d} = \max_S \alpha_o^S \times DL_c^{ST_d}$ represents the origin country-by-destination national language level exposure. 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.

In line with the main analysis, we start the analysis in 2007 and hence omit all individuals who have migrated before. [other sample restrictions]

ACS 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.[49][50] Although the language spoken at home would allow me to construct an exposure measure, I choose to use the country of birth as the language spoken at home could be endogenous. We construct the exposure to Duolingo to English as:

$$DL_{ot}^{T=T_d=EN} = \max_S \alpha_o^S \times DL_t^{S,T=EN} \tag{15}$$

We restrict the sample to those aged between 12 to 64 who immigrated to the US in the past 5 years. We restrict the sample to those who immigrated to the US since 2009, to prevent large differences in the

[48] Although contemporaneous this could generate a spurious correlation between the outcome and this exposure measure when the scale of the migrant flow is correlated to this outcome. Nevertheless, as we have seen in Section 6 above, the size of migration flows to a very similar set of countries (OECD countries) only increase several years after Duolingo rollout. Hence, in a robustness test I also use the migrant stock M_{od2005} in 2005 born in country o living in country d.

[49] Only English, very well, well, not well, does not speak English

[50] Self-reported language skills in the ACS have been shown to strongly correlate to actual language proficiency 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-ability-question.html. However, some studies have found that active learners under-assess their learning gains, e.g. https://onlinelibrary.wiley.com/doi/abs/10.1111/flan.12379

composition of immigrant flows before the global financial crisis. In a later version of the paper I assess robustness of different definitions.

[Descriptives about language levels, Mincer of language levels]

7.2 Reasons to Migrate, Language Skills, Sorting and Selection

7.2.1 Empirical strategy

To study the effects of Duolingo availability on migrants' characteristics and language upon arrival, I compare individuals who have migrated in presence of low-cost language learning opportunities to those who don't. As the availability of Duolingo for a dyad may be more likely in dyads with different levels, we always include dyad fixed effects. We estimate varieties of the following equation:

$$y_{iodct} = \beta DL_{oc}^{T=T_d} + \phi_{od} + (\psi_{oc}) + \theta_{dc} + \xi_t + \epsilon_{iodct} \tag{16}$$

For the US, there is only one destination country d, and Duolingo exposure varies at the origin-cohort level. For the EU, the origin indicator o is replaced by the origin group indicator g. 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 the ideal setting, this outcome is realized at year c and faithfully reported in t. An example is the language skills upon arrival or the degree one held upon migrating (e.g. the EU LFS asks for the year of highest qualification). In other settings, we only observe the contemporaneous value of y. In those cases we only use those who are interviewed shortly after arrival at the expense of statistical power. This is especially relevant to isolate the effect of pre-migration availability from post-migration availability.

ϕ_{od} captures dyadic fixed effects, ψ_{oc} and θ_{dc} capture origin-cohort and destination-cohort fixed effects. The term ψ_{oc} captures all cohort-specific factors. ξ_t captures year-specific factors unrelated

Following the discussions in section 4.4 and 4.5, we estimate all models using a dynamic treatment estimator and cluster standard errors twoway at the origin- by destination- level.

7.2.2 Empirical strategy

[to copy below when done] To study the effects of Duolingo availability on migrants' integration, we pursue a different identification strategy that emphasises the *path* of integration after arrival

$$y_{iodct} = \beta DL_{oc}^{T=T_d} + \phi_{od} + (\psi_{oc}) + \theta_{dc} + \xi_t + \epsilon_{iodct} \tag{17}$$

This is especially relevant to isolate the effect of pre-migration availability from post-migration availability.

7.2.3 Results

Reasons to migrate (EU) Although the analysis of section 6 studies how migration aspirations change, the EU LFS allows us to study whether the composition of realised migration flows changes. One important aspect of migration is why people migrated, which may shed light on who benefits most from migration as well as the (intended) activities of migrants in the host country. One appealing aspect of the EU LFS is that it asks on the reason for migrating.

[Enter table]

Language Skills upon arrival and selection Our sample is restricted to those migrants that arrived after the year 2000 who originate from a country where the destination-country language is not spoken as a mother tongue by more than 20% of the migrants across the sample. This sample includes 33,774 migrants originating from 17 distinct origin countries and residing in 28 distinct destination countries. First, I estimate equation **??** to study pre-learning effects on migrants' language skills. Figure **??** shows that exposure to a Duolingo course enabling you to learn the host country language in the 3 years before migration increases the probability to speak the host country language at least at beginner level by 14 percentage points, but there is no discernible effect on higher levels of languages skills. Figure shows the estimation results of equation **??**. The results show that exposure to Duolingo after arrival does not significantly increase language skills, although the estimate on at least beginner levels is just not statistically significant (point estimate = 0.5, p=0.078). The latter two results are in line with Duolingo enabling users to use basic language skills, but once in the destination does not offer additional beneifts.

[INPUT LANDSCAPE TABLE]

To additionally study whether the type of migration is changed by the introduction of Duolingo, I rely on a question in the same add-hoc course on the reason for migration. We find that more people arrive for employment reasons, without having found a job before (Figure **??**). In a future version of the paper, I extend this analysis to the 2008 and 2014 add-hoc courses as well.

7.3 Integration

Not only could accessibility to language learning affect the extent of migrant flows (as in Huber and Uebelmesser (2019)), it could also affect language skills upon arrival as well as selection (as in Jaschke and Keita (2021)). After arrival, Duolingo courses enable migrants to further attain the destination country language. As English is the most learned second language across the world and the language with the most target languages in Duolingo courses (22), an English-speaking country that is also a popular destination

country for immigrants speaking many different mother tongues is the ideal setting to study the effect of the rollout of Duolingo on immigrant skills and charcteristics. Therefore, I turn to the United States to study how the staggered introduction of Duolingo courses has changed migrants' characteristics upon arrival and thereafter. 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.[51]

7.3.1 Empirical strategy

To assess the effect of Duolingo on the language skills of migrants, I estimate the following model for various outcomes y_{itco}, which vary at the individual i, time of interview t, time of arrival in the US (cohort) c and country of origin o (measured as country of birth):

$$y_{itco} = \alpha DL_{co}^{pre} + \beta DLyears_{tco}^{post} + \gamma^{'}\mathbf{X_{itc}} + \delta^{'}\mathbf{Z_{c-1o}} + \phi_{c(t-c)} + \psi_{o(t-c)} + \epsilon_{itco} \tag{18}$$

DL_{co}^{pre} and $DLyears_{tco}^{post}$ denote the pre-arrival and post-arrival Duolingo exposure of immigrants interviewed in year t who arrived in the US in year c, expressed in number of years of availability. We construct the pre-migration exposure as a binary indicator for whether a relevant Duolingo course was available in the year prior to moving and the post-migration exposure as the number of years one was able to study English from a language widely spoken in their country of birth (by at least 25% of the population) at origin before and after arrival, respectively:

$$DL_{co}^{pre} = \mathbb{1}\left(c > \tau_{introduction,l}\right) \tag{19}$$

$$DLyears_{cot}^{post} = max(min(t - \tau_{introduction,l}, t - c), 0) \tag{20}$$

Where $\tau_{introduction,l}$ denotes the year the first relevant Duolingo course from a source country l spoken by at least 25% of the population of country o was introduced. To control for language knowledge at arrival and learning patterns specific to immigrants from specific countries I control for country-specific effects at each number of years since arrival through origin country-by-time-since-arrival fixed effects $\psi_{o(t-c)}$. To control for different assimilation paths by cohort of arrival, I include cohort-by-time-since-arrival fixed effects $\phi_{c(t-c)}$, which also capture unobserved year-specific factors. Individual-level controls $\mathbf{X}_{itc}$ capture the effect of age at arrival, its square and sex, and cohort controls at the year before immigration at the country level

[51] https://blog.duolingo.com/the-united-states-of-languages-an-analysis-of-duolingo-usage-state-by-state/

$\mathbf{Z}_{c-1,o}$ capture differential selection of immigrants: those include unemployment rates, GDP per capita and a measure for conflict at origin. ϵ_{itco} denotes the individual-level error term. We always cluster standard errors at the country-of-origin level.

The main identifying assumption to identify the causal effect of availability of language learning through Duolingo on migrant characteristics is a parallel trend assumption: the path of outcomes of immigrants originating from a country for which a relevant Duolingo course towards English becomes available would have been the same as that of an immigrant from a country for which no such course is available (yet), in the counterfactual case no relevant Duolingo course would have rolled out.

7.3.2 Results

In Table 5 I show the effect of pre-arrival and post-arrival Duolingo exposure on English language knowledge, as well as age at immigration, sex, and educational attainment. First of all, I find that the availability of a Duolingo course from the prior to arrival affects language skills of immigrants considerably: the share of immigrants speaking at least some English increases by 0.9 percentage points. This effects is sizable as the baseline share of individuals speaking some English is already relatively high at 88.8%. Similarly, I find a similar-sized effect on the share of individuals speaking English at least well. Contrary to the case of certified language learning (Jaschke and Keita, 2021), I do not find evidence that the skill composition changes (column 5 and 6 - on the subsample of those). The length of Duolingo availability after arrival has a positive impact on self-assessed language skills. One additional year of Duolingo availability increases the probability to report to speak English well and very well by 0.8 percentage point. Table 6 shows the effects on labor market outcomes. We find no significant effects on the probability to be employed of both pre-arrival and post-arrival exposure. However, there is no effect on earnings. As total earnings may be a noise measure of income and occupational quality (as confirmed by the standard error on that estimate of almost 4 log point), I also include the occupational (ranging from 0 (unemployed) to 80) as a dependent variable. In the future, I aim to study whether the increase in language skills has contributed to differences in occupational composition.

As I use data from immigrants who arrived between 0 and 5 years prior to interview, I measure the effect of pre-departure and post-departure Duolingo exposure on language skills after arrival on average. To better assess the effect of Duolingo availability on language skills as closely as possible after arrival, I restrict the sample to those who have been in the US for strictly less than a year.[52] This answers a different question than the analysis in the previous section, as immigrants without any available Duolingo course could have

[52] In other words: the year of interview and the year of arrival should be the same. Assuming a constant distribution of immigrant arrivals and interviews across the year, the average time since arrival is likely to be only several months on average. The ACS does not provide more detailed information on the time of arrival and interview.

Table 5: The effect of *Duolingo* availability before and after arrival on language skills and demographics

	Speaks EN	Speaks EN at least well	Speaks EN at least very well	Age at immigration	Female	At least 9th grade (at least 18)	At least some tertiary education (at least 25)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$DL^{pre}{}_{oc}$	0.008**	0.009	0.007	0.167	0.004	0.009*	-0.008
	(0.003)	(0.007)	(0.006)	(0.132)	(0.004)	(0.005)	(0.005)
$DLyears^{post}{}_{oc}$	0.003**	0.009***	0.007**	0.116*	-0.002	0.003	0.010*
	(0.001)	(0.002)	(0.003)	(0.066)	(0.002)	(0.002)	(0.005)
Observations	400217	400217	400217	400217	400217	343283	245411
R^2	0.23	0.33	0.30	0.07	0.03	0.21	0.32
Average dependent variable	0.889	0.716	0.477	29.628	0.506	0.902	0.639
Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Controls for age, age2 and sex	✓	✓	✓				

Estimated by OLS, Standard errors clustered at the country of origin.

Table 6: The effect of *Duolingo* availability before and after arrival on integration outcomes

	Employed	Total income earned	Occupational score
	(1)	(2)	(3)
$DL^{pre}{}_{oc}$	-0.003	-0.055	-0.108
	(0.005)	(0.052)	(0.163)
$DLyears^{post}{}_{oc}$	0.001	0.016	0.166
	(0.003)	(0.039)	(0.106)
Observations	400217	379120	400217
R^2	0.25	0.22	0.29
Average dependent variable	0.531	6.026	19.014
Fixed Effects	✓	✓	✓
Controls for age, age2 and sex	✓	✓	✓

Estimated by OLS, Standard errors clustered at the country of origin.

attained English through other means after arrival and catch up with those that had such courses available. Table 7 shows the results of the analysis. We find that closely upon arrival, self-assessed English skills are better. Access to a relevant Duolingo course for at least one year prior to arrival improves the likelihood an immigrants speaks English well or better by 2.0 percentage points.

Table 7: The effect of *Duolingo* availability before arrival on language skills and demographics of those who arrived less than 1 year ago

	Speaks EN	Speaks EN at least well	Speaks EN at least very well	Age at immigration	Female	At least 9th grade (at least 18)	At least some tertiary education (at least 25)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$DL^{pre}{}_{oc}$	0.009	0.020*	0.009	-0.342	-0.011	0.011*	-0.002
	(0.010)	(0.012)	(0.011)	(0.264)	(0.008)	(0.006)	(0.011)
Observations	53671	53671	53671	67423	67423	61429	42189
R^2	0.32	0.38	0.34	0.12	0.08	0.25	0.34
Average dependent variable	0.836	0.656	0.417	31.411	0.487	0.894	0.607
Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Controls for age, age2 and sex	✓	✓	✓				

EXPLANATION

EFFECT SIZE INTERPRETATION IS CRUCIAL: also Fabian's comment

[Compare effect sizes to that known in the literature]

8 Conclusion

Language is paramount to the success of immigrants in the host country, which motivates learning. The ability to learn a host country's language may enable prospective immigrants to obtain country-specific skills that (1) increase labor market earnings and (2) lower migration costs but also (3) foster interest in a country's language and culture by lowering the perceived linguistic distance. Language learning apps could 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. In this paper, I study the effect of the availability of language courses on the popular mobile and desktop application *Duolingo* on migration aspirations to a particular destination, actual migration to OECD countries and the language skills and integration of immigrants in the United States.

The availability of a typical language learning course increases the bilateral desire to emigrate by 7%. Furthermore, by using data on international migration to OECD countries, I find evidence that migration flows start increasing 4 years after the roll-out of a language course. As these estimates are obtained using a three-way fixed regression those results imply that dyadic flows increase *relative* to untreated dyads, but not whether or not total migration flows are affected as well. Importantly, I find that the effects are driven by English as a target language and by countries that do not have language requirements. The latter informs policy makers that such requirements are relevant tools to steer immigration in a world where digital technologies can bridge skills, rather than certification.

Surveys among immigrants in migrant-receiving countries can provide information on whether immigrant cohorts that had access to a relevant *Duolingo* course prior to immigration differ from those who had not in terms of education, language knowledge and other characteristics. In contrast to the findings of Jaschke and Keita (2021) that the availability of language learning institutes providing certification made migrants to Germany more positively selected, the availability of *low-cost* language learning may have different implications for immigrant self-selection. By studying the characteristics of immigrants in the USA, I find that the rollout of relevant Duolingo courses made immigrants assess their language skills to be better upon arrival, but also to enable learning after arrival. Our results suggest that it indeed does, which may inform policy makers about the relative effectiveness of provided language courses and private alternatives.

Several interesting questions concerning low-cost mobile language learning among immigrants are left for future research. Furthermore, as many immigrant-receiving countries invest heavily in (compulsory) language courses for recent immigrants, this raises the question whether language courses and low-cost mobile language learning are complementing each other or not.

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.

Abraham, Sarah, and Liyang Sun. 2018. "Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects." *SSRN Electronic Journal*.

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."

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

Adsera, Alicia, and Mariola Pytlikova. 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." *The Palgrave handbook of economics and language*, 342–372.

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 AR, 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." *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. 1987. "Self-Selection and the Earnings of Immigrants." *The American Economic Review*, 77(4): 531–553.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2021. "Revisiting event study designs: Robust and efficient estimation." *arXiv preprint arXiv:2108.12419*.

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

Callaway, Brantly, and Pedro H C Sant'Anna. 2018. "Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment." DETU Working Papers 1804.

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. 2023. "Logs with zeros? Some problems and solutions." *The Quarterly Journal of Economics*, qjad054.

Chiswick, Barry R, and Paul W Miller. 1995. "The endogeneity between language and earnings: International analyses." *Journal of labor economics*, 13(2): 246–288.

Chiswick, Barry R, and Paul W Miller. 2015. "International migration and the economics of language." In *Handbook of the economics of international migration*. Vol. 1, 211–269. Elsevier.

Conte, m, P Cotterlaz, and T Mayer. n.d..

Dahlberg, Matz, Johan Egebark, Ulrika Vikman, and Gülay Özcan. 2024. "Labor market integration of refugees: RCT evidence from an early intervention program in Sweden." *Journal of Economic*

Behavior & Organization, 217: 614–630.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2019. "Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects." NBER Working Papers 25904.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." *arXiv preprint arXiv:2007.04267*.

de Chaisemartin, Clément, Xavier d'Haultfoeuille, Félix Pasquier, and Gonzalo Vazquez-Bare. 2022. "Difference-in-differences estimators for treatments continuously distributed at every period." *arXiv preprint arXiv: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, 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.

Fenoll, Ainoa Aparicio, and Zoë Kuehn. 2019. "Immigrants move where their skills are scarce: Evidence from English proficiency." *Labour Economics*, 61: 101748.

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. 2022. "Comparing the Effects of Policies for the Labor Market Integration of Refugees."

Freeman, Cassie, Audrey Kittredge, Hope Wilson, and Bozena Pajak. 2023. "The duolingo method for app-based teaching and learning." *Duolingo Research Report*.

Ginsburgh, Victor A, and Juan Prieto-Rodriguez. 2011. "Returns to foreign languages of native workers in the European Union." *ILR Review*, 64(3): 599–618.

Ginsburgh, Victor, Jacques Melitz, and Farid Toubal. 2016. "Foreign Language Learning and Trade." *Review of International Economics*, 25(2): 320–361.

Ginsburgh, Victor, Jacques Melitz, and Farid Toubal. 2017. "Foreign language learning and trade." *Review of International Economics*, 25(2): 320–361.

Goodman-Bacon, Andrew. 2018. "Difference-in-differences with variation in treatment timing."

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.

Gruber, Jonathan. 1994. "The incidence of mandated maternity benefits." *The American economic review*, 622–641.

GSMA. 2019. "Mobile Internet Connectivity 2019 ."

Hahm, Sabrina, and Michele Gazzola. 2022. "The value of foreign language skills in the German labor market." *Labour Economics*, 76: 102150.

Harmon, Nikolaj A. 2022. "Difference-in-Differences and Efficient Estimation of Treatment Effects." Working paper.

Huber, Matthias, and Silke Uebelmesser. 2019. "Presence of language-learning opportunities and migration."

Isphording, Ingo E. 2013. "Returns to Foreign Language Skills of Immigrants in S pain." *Labour*, 27(4): 443–461.

Isphording, Ingo E, and Sebastian Otten. 2011. "Linguistic distance and the language fluency of immigrants." *Ruhr Economic 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 Plon-

sky, 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.

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."

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

Olden, Andreas, and Jarle Møen. 2022. "The triple difference estimator." *The Econometrics Journal*, 25(3): 531–553.

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

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." *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, JMC Santos, and Silvana Tenreyro. 2006. "The log of gravity." *The Review of Economics and statistics*, 88(4): 641–658.

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." 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.

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. 2022. "The effect of linguistic proximity on the labour market outcomes of the asylum population." *Journal of Population Economics*, 1–44.

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. 2022. "Simple Approaches to Nonlinear Difference-in-Differences with Panel Data." *Available at SSRN 4183726*.

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

A Model details and extentions

A.1 Total migration

The expression for the probability to migrate between o and d does depend on the relative attractiveness of all alternative destinations:

$$\frac{\partial \mathbb{P}_{od}}{\partial s_{oT}} = \frac{e^{\mu_{od}+s_{oT}b_{odT}}\left[\left(\sum_{d'} e^{\mu_{d'}+s_{oT}b_{od'T}}\right)b_{odT} - \left(\sum_{d'} b_{od'T}e^{\mu_{d'}+s_{oT}b_{od'T}}\right)\right]}{\left(\sum_{d'} e^{\mu_{d'}+s_{oT}b_{od'T}}\right)^2} = \mathbb{P}_{od}\left(b_{odT} - \sum_{d'}\mathbb{P}_{od'}b_{od'T}\right) \lessgtr 0 \tag{21}$$

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 $d \neq o$), this term (and thus the domestic return to language skills) plays a prominent role in Equation 21. 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 22 shows the condition for which this is the case. Total emigration increases if the probability-weighted sum of returns to a destination exceeds domestic returns. Hence, if migration probabilities are low enough, and a language skills is moderately values 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 costs).

$$\frac{\partial \sum_{d\neq o}\mathbb{P}_{od}}{\partial s_{oT}} = \mathbb{P}_{oo}\sum_{d\neq 0}\mathbb{P}_{od}b_{odT} - (1-\mathbb{P}_{oo})\mathbb{P}_{oo}b_{ooT} > 0 \implies \frac{1}{\sum_{d\neq 0}\mathbb{P}_{od}}\sum_{d\neq 0}\mathbb{P}_{od}b_{odT} > b_{ooT} \tag{22}$$

A.2 Derivation of Equation 23 in the low migration limit

Equation 6 can be written as:

$$2c_T s_{oT} \approx b_{ooT} + \sum_{d'\neq o}\mathbb{P}_{od}(0)b_{odT}(1+b_{odT}s_{oT}) \tag{23}$$

A.3 Calculation of the communication probabilities

α_{cL} denotes the number of language L in country c. If we 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), we can calculate the probability that two randomly chosen individuals can communicate. We 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. Using the law of total probability, we can write this as a function of α_{cL}'s, showing the first $k = 4$ out of $k = N$ terms:

$$\mathbb{P}(comm_{od}) = \sum_{l}\alpha_l^2 - \sum_{ll'}\alpha_l^2\alpha_{l'}^2 + \sum_{ll'l''}\alpha_l^2\alpha_{o'}^2\alpha_{l''}^2 - \sum_{ll'l''l'''}\alpha_l^2\alpha_{o'}^2\alpha_{l''}^2\alpha_{l'''}^2 \ldots \tag{24}$$

For a large parameter space of common languages, this implies that there are many terms. For N languages, the k$^{\text{th}}$ term there are $\binom{N}{k}$ elements. However, as language knowledge is fairly sparse and $\alpha_{cL} \leq 1$, the largest k=4$^{\text{th}}$ order term is only 0.02. Hence, we truncate the calculation at N = 4. Using this result, and setting $\alpha_{oS} = 1$, we obtain the probability conditional on the individual in o speaking S, $\mathbb{P}(comm_{od}|S)$. Additionally, setting $\alpha_{oT} = 1$, we 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 by:

$$\mathbb{P}(comm_{od}|DL_{S\to T}, S) = \mathbb{P}(comm_{od}|S \wedge T) - \mathbb{P}(comm_{od}|S) \tag{25}$$

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:

$$\mathbb{P}(comm_{od}|DL_{S\to T}) = \mathbb{P}(comm_{od}|DL_{S\to T}, S)\mathbb{P}(S) = \mathbb{P}(comm_{od}|DL_{S\to T}, S)\alpha_{oS} \tag{26}$$

In the paper, we use shorthand notation for the increased probability of communicating between two countries, $\mathbb{P}(comm_{od}|DL_{S\to T}, S) = DL_{od}^{ST}$.

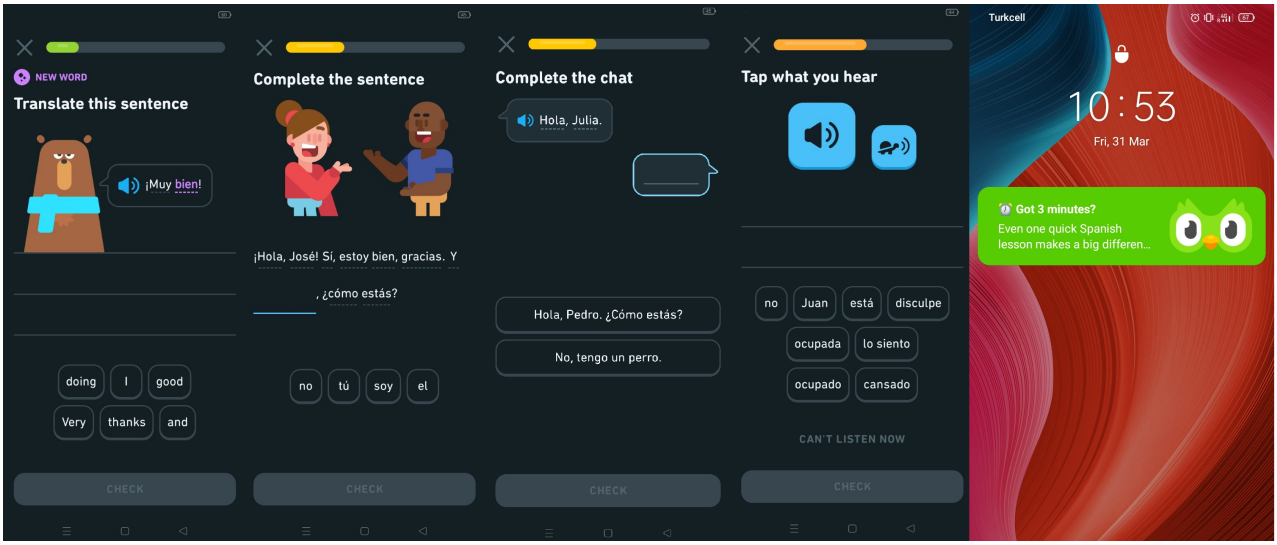
B Detailed Description of Data

C Descriptives

C.1 Duolingo

C.1.1 Course content

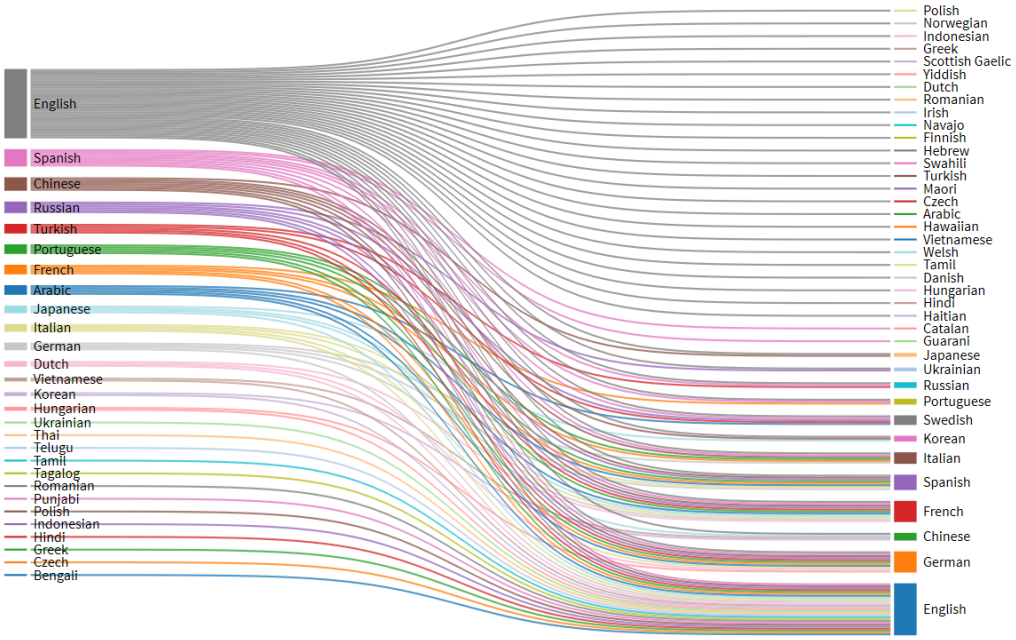
Figure A1: Characteristic tasks on *Duolingo*



Note: Example of a *Duolingo* course: English to Spanish.

C.1.2 Courses and Users

Figure A2: Available Courses as of 2022

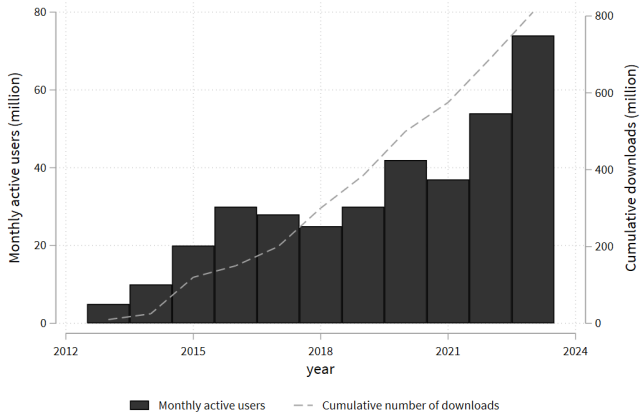


Note: Sankey diagram of all available courses on Duolingo. In total, this comprises 84 modules over 23 source languages and 30 target languages. Information on courses is obtained from the Duolingo website.

Figure A3: Most learnt language by country on *Duolingo* in 2021

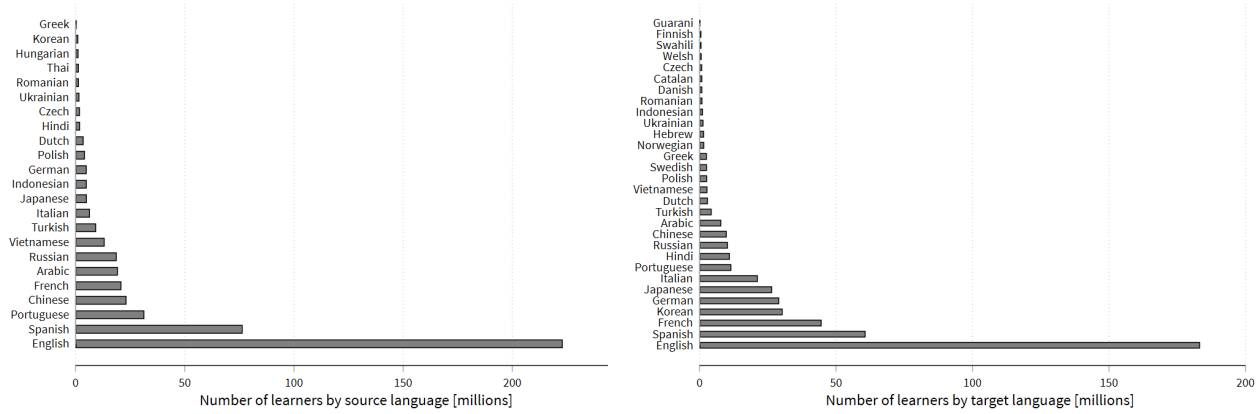
Data from the 2021 Duolingo Language Report.

Figure A4: Growth in Monthly Active Users on Duolingo between 2012 and 2023



Numbers on the Monthly Active Users and cumulative downloads are obtained from https://www.businessofapps.com/data/duolingo-statistics/.

Figure A5: Number of *Duolingo* learners by source and target language

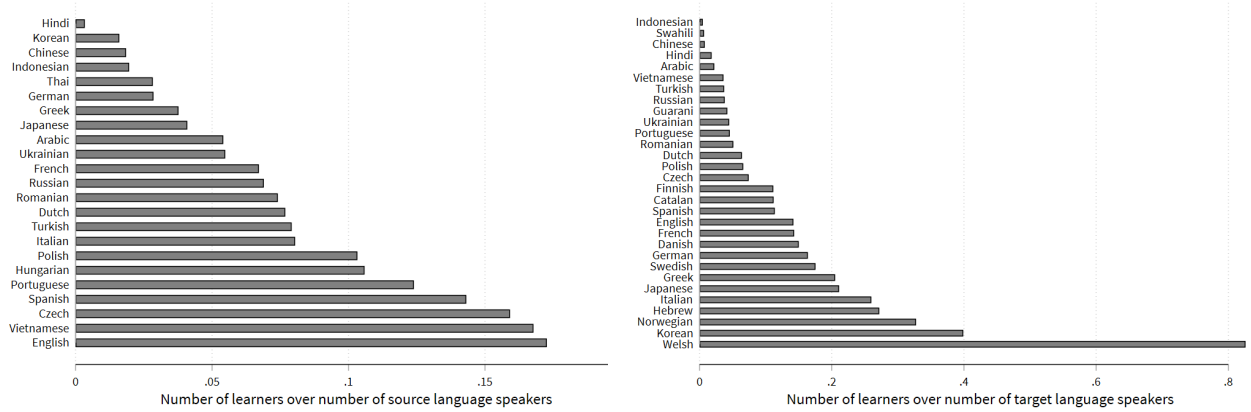


(a) Source language

(b) Target language

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, we sum over all courses. This thus 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 A6: Intensity of *Duolingo* language learning by source and target language



(a) Source language

(b) Target language

See notes to Figure A5 for a description of the calculation of the total number of learners. To calculate the number of learners relative to speakers, we divide the numbers from A5 by the total number of speakers by language from Ginsburgh, Melitz and Toubal (2017)

Table A1: The Determinants of the Number of Learners of *Duolingo* Courses

	(1) Learners	(2) Learners
Source language speakers	0.008*** (0.001)	0.004*** (0.000)
Target language speakers	0.008*** (0.001)	0.004*** (0.000)
Source language speakers × Target language speakers		0.000*** (0.000)
Observations	84	84
Source and Target FE		✓

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. Data on learners is obtained from the Duolingo platform in July 2024. Standard errors are clustered twoway on the source and destination language. The total number of learners is 478 million.

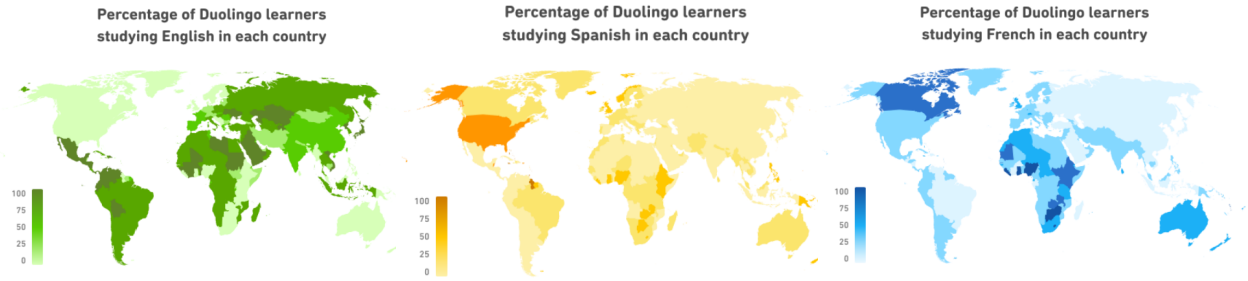
C.1.3 Learners of prominent languages by country

Table A2: 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.

Figure A7: Percentage of learners learning English, Spanish, or French across the world in 2020



(a) English (b) Spanish (c) French

C.1.4 Exposure

Figure A8: Duolingo Exposure by directed dyad over time

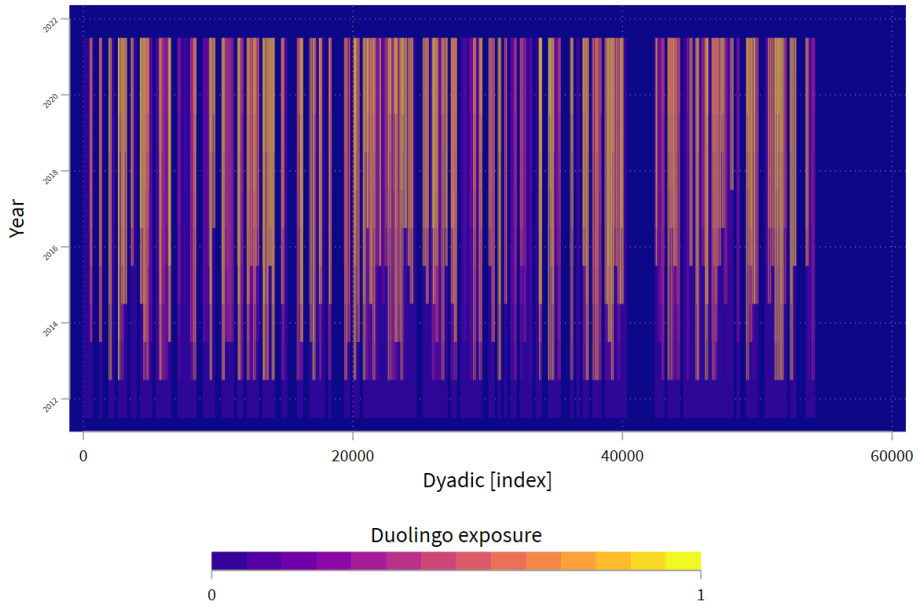


Figure A9: Average *Duolingo* Exposure by origin country

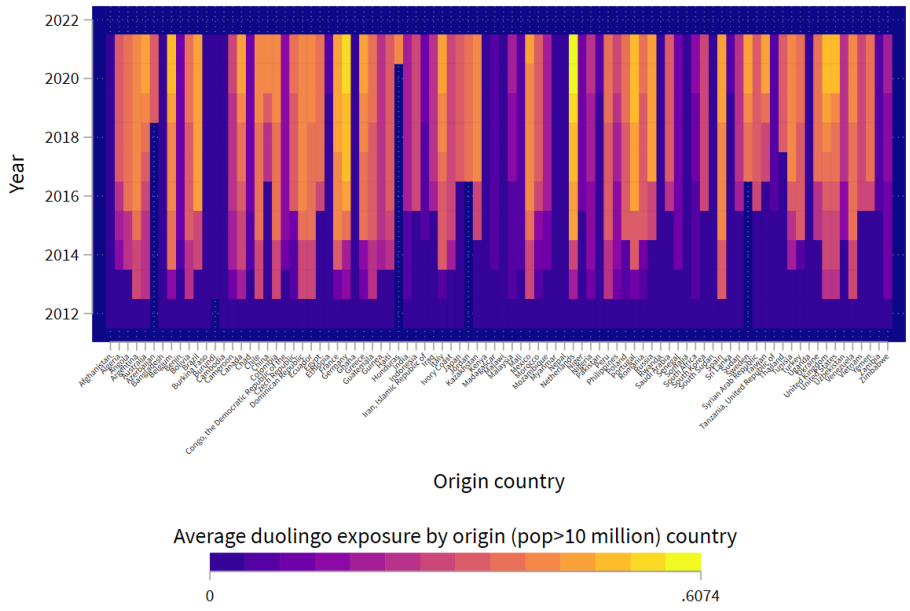
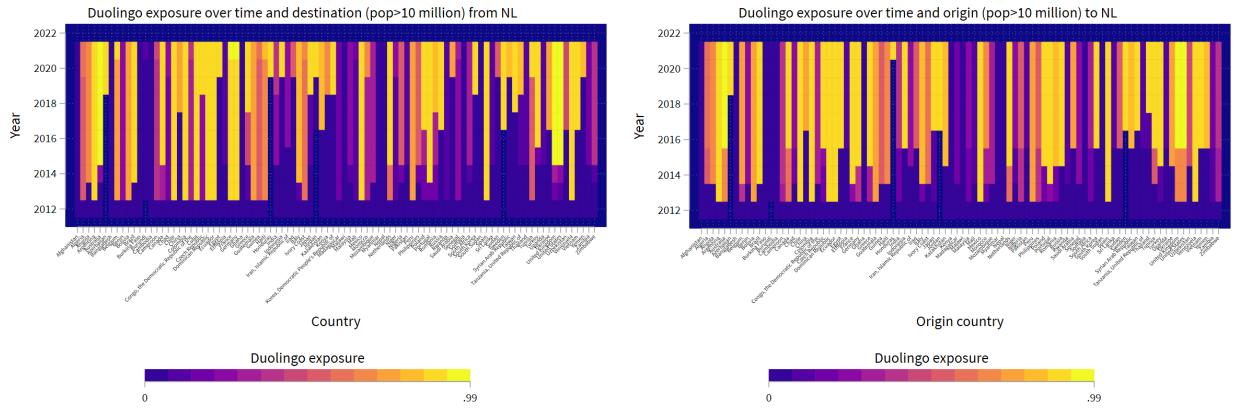


Figure A10: Variation in *Duolingo* exposure to and from the Netherlands



(a) From the Netherlands

(b) To the Netherlands

D Additional results

D.1 Does Duolingo crowd out traditional 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, which could increase in-class course participation, particularly at higher proficiency levels. To provide some evidence on this, we turn to data on the number of German language course and exam takers at German language learning institutes (*Goethe* institutes) outside of Germany. By combining the dataset of Uebelmesser, Sommerfeld and Weingarten (2022) with recent yearly reports of the Goethe Institute, we can study how the introduction of Duolingo affected German language course and exam participation.

We 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. Data for 2015 is missing. There are 12 global regions, who have about 12 Goethe institutes each on average. Using the sum of registrations and exam takers by country in the year 2014, we 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, we construct a weighted exposure to Duolingo courses on the global region level. We estimate a two-way fixed effects Poisson model regressing the number of exam or course takers on the Duolingo exposure and fixed effects for region and year. The results are reported in Table A3.

Although somewhat imprecise, the results in Appendix Figure A3 are telling. We find that introduction of a relevant Duolingo course decreases the number of registrations by about 11% (p=0.054), but the number of exams is hardly affected. 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 language skills for some types of residence, and German higher education institutes for admission.

Table A3: The Effect of *Duolingo* Courses on Institutional German Learning

	(1) Number of exams	(2) Number of course participants	(3) Exams per course participant
Duolingo exposure	-0.019	-0.112*	-0.196
	(0.132)	(0.058)	(0.601)
Observations	180	180	180

Columns 1 and 2 are estimated by PPML. Standard errors are clustered at the origin group level.

E Additional Robustness

E.1 Migration Aspirations

Alternative Exposures

Table A4: Using Different Definitions of *Duolingo* Exposure

	(1) M_{odt}	(2) M_{odt}	(3) M_{odt}	(4) M_{odt}	(5) M_{odt}	(6) M_{odt}	(7) M_{odt}	(8) M_{odt}
$DL_{\text{setting less than 0.2 to 0}}$	0.182***							
	(0.048)							
$DL_{\text{summed over modules and capped at 1}}$		0.196***						
		(0.052)						
$DL_{\text{share in origin if spoken by at least 70\% in d}}$			0.146***					
			(0.039)					
$DL_{\text{subtract the sum of shares of CSLs}}$				0.282***				
				(0.051)				
$DL_{\text{only official languages in destination}}$					0.131***			
					(0.041)			
$DL_{\text{omit d-t pair with most speakers per lang.}}$						0.180***		
						(0.054)		
$DL_{\text{omit o-s pair with most speakers per lang.}}$							0.301***	
							(0.058)	
$DL_{\text{omit d-t and o-s pairs with most speakers per lang.}}$								0.254***
								(0.066)
Observations	100507	100507	100507	100507	100507	100507	100507	100507
Origin-destination FE	✓	✓	✓	✓	✓	✓	✓	✓
Destination-year FE	✓	✓	✓	✓	✓	✓	✓	✓
Origin-year FE	✓	✓	✓	✓	✓	✓	✓	✓
Origin-destination clustered SEs	✓	✓	✓	✓	✓	✓	✓	✓

Estimated by PPML, Standard errors clustered two-way at the origin-destination level.

E.2 Migration Flows

Table A5: Main results

	(1) M_{odt}	(2) M_{odt}	(3) M_{odt}	(4) M_{odt}
$Duolingo_{odt}$	0.242**	0.122	0.202***	0.190***
	(0.112)	(0.098)	(0.070)	(0.068)
Observations	96762	100507	100507	100507
Unique origin countries	158	158	158	158
Unique destination countries	188	188	188	188
Unique dyads	9747	9747	9747	9747
Origin, destination and year FE	✓	✓	✓	✓
Shared official- and common language- dummy	✓			
Origin-destination FE		✓	✓	✓
Origin-year FE			✓	✓
Destination-year FE			✓	✓
Trade controls				✓

See notes to Table 12. Standard errors clustered two-way at the origin and at the destination level.

E.3 Heterogeneity

Table A6: Main results

	(1) M_{odt}	(2) M_{odt}	(3) M_{odt}	(4) M_{odt}
$Duolingo_{odt}$	0.241**	0.129	0.204***	0.191***
	(0.110)	(0.110)	(0.046)	(0.045)
Observations	90652	93429	93429	93429
Unique origin countries	151	151	151	151
Unique destination countries	178	178	178	178
Unique dyads	9032	9032	9032	9032
Origin, destination and year FE	✓	✓	✓	✓
Shared official- and common language- dummy	✓			
Origin-destination FE		✓	✓	✓
Origin-year FE			✓	✓
Destination-year FE			✓	✓
Trade controls				✓

See notes to Table 12. Standard errors clustered two-way at the linguistic pair level. Number of observations is slightly lower than in Table 12 due to missing linguistic family information for few countries.

E.4 Merging Duolingo exposure on the language spoken at home

Table A7: Heterogeneity in the effect of *Duolingo* availability before and after arrival on speaking English well

	25 and younger (1)	Between 25 and 50 (2)	50 and older (3)	Male (4)	Female (5)	Not married upon arrival (6)	Married upon arrival (7)
$DL^{pre}{}_{oc}$	0.020**	0.007	0.013	0.019**	0.004	0.022**	-0.002
	(0.010)	(0.007)	(0.011)	(0.008)	(0.009)	(0.010)	(0.007)
$DLyears^{post}{}_{oc}$	0.005	0.012***	0.005	0.005**	0.010***	0.004	0.013***
	(0.003)	(0.005)	(0.007)	(0.003)	(0.003)	(0.003)	(0.005)
Observations	201277	242100	37324	236022	244721	284898	195843
R^2	0.31	0.32	0.35	0.32	0.32	0.32	0.31
Average dependent variable	0.765	0.689	0.479	0.715	0.695	0.758	0.627
Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Controls for age, age2 and sex	✓	✓	✓	✓	✓	✓	✓

Estimated by OLS, Standard errors clustered at the country of origin.

Table A8: The effect of *Duolingo* availability before and after arrival on language skills and demographics

	Speaks EN	Speaks EN at least well	Speaks EN at least very well	Age at immigration	Female	At least 9th grade (at least 18)	At least some tertiary education (at least 25)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$DL^{pre}{}_{lc}$	0.008**	0.011**	0.004	0.226***	0.006	0.010**	-0.006
	(0.003)	(0.005)	(0.006)	(0.076)	(0.004)	(0.004)	(0.004)
$DLyears^{post}{}_{lc}$	0.004**	0.009***	0.008***	0.121**	-0.002	0.002	0.012**
	(0.002)	(0.003)	(0.003)	(0.055)	(0.002)	(0.002)	(0.005)
Observations	397749	397749	397749	397749	397749	340960	243463
R^2	0.23	0.36	0.38	0.07	0.03	0.20	0.33
Average dependent variable	0.889	0.715	0.476	29.625	0.506	0.902	0.639
Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Controls for age, age2 and sex	✓	✓	✓				

Estimated by OLS, Standard errors clustered at the country of origin.

Table A9: The effect of *Duolingo* availability before and after arrival on integration outcomes

	Employed	Total income earned	Occupational score
	(1)	(2)	(3)
(4)			
$DL^{pre}{}_{lc}$	-0.003	-0.068	-0.003
	(0.004)	(0.051)	(0.186)
$DLyears^{post}{}_{lc}$	0.001	0.022	0.190*
	(0.003)	(0.031)	(0.107)
Observations	397749	376577	397749
R^2	0.25	0.24	0.29
Average dependent variable	0.531	6.308	18.998
Fixed Effects	✓	✓	✓
Controls for age, age2 and sex	✓	✓	✓

Estimated by OLS, Standard errors clustered at the country of origin.

Table A10: The effect of *Duolingo* availability before arrival on language skills and demographics of those who arrived less than 1 year ago

	Speaks EN	Speaks EN at least well	Speaks EN at least very well	Age at immigration	Female	At least 9th grade (at least 18)	At least some tertiary education (at least 25)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$DL^{pre}{}_{lc}$	0.014	0.015	0.001	0.035	-0.015	0.011	-0.000
	(0.009)	(0.014)	(0.013)	(0.282)	(0.009)	(0.008)	(0.008)
Observations	53729	53729	53729	53729	53729	49060	33933
R^2	0.29	0.38	0.41	0.06	0.04	0.19	0.31
Average dependent variable	0.835	0.655	0.416	31.467	0.481	0.892	0.616
Fixed Effects	✓	✓	✓	✓	✓	✓	✓
Controls for age, age2 and sex	✓	✓	✓				

F Google Trends: obtaining panel data of relative search intensity

GTI series across time and space

Google Trends allows one to query the relative search interest (relative to all search activity) of a search term (1) by region for a given time period or (2) over time for a given region.[53] Importantly, it is not possible to directly query the relative search interest of a panel of interest in a search term over time across countries. Moreover, Google Trends does not output the relative search intensity. The output is simply scaled to 100 for the highest relative intensity within a query, 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 two limitations, a panel dataset measuring relative search intensity can be calculated. The first object is the interest over time for a given geographic region (the whole world, a country, or a subnational region), which I denote by $GTI^{\widetilde{T}}_{\widetilde{o}t}$. o is the geographic region of interest, t is the time period of interest and T is the term or topic of normalized. The variables with a tilde indicate that search interest is not normalized *across* that dimension, whereas the absence of a tilde indicates that it is normalized across that dimension. 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 are interested in longer periods, for our purpose the monthly frequency suffices and I always query time series between 2006 and 2022.

The second object is the interest by region for a given time period, which I denote by $GTI^{\widetilde{T}}_{o(2006-2022)}$. Here, the available regions are 240 countries and territories. For many of these regions, Google has a large market share and is thus fairly representative of the online population. For an analysis of Google's market share on the search market across the world for out time period of interest, see above. This approach is still limited by the rounding of the index on whole 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 inaccurate results for low interest regions.

Using the interest by regions and across time for the same T, geographic area and time, 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 over time and across regions. Please note that it is not possible to make direct statements about the share of searches concerning a given topic, due to the scaling of GTI. However, to further interpret the normalized index one can in principle calculate the search intensity relative to an undoubtedly popular topic, such as the News.

Anchorbanking across terms and topics

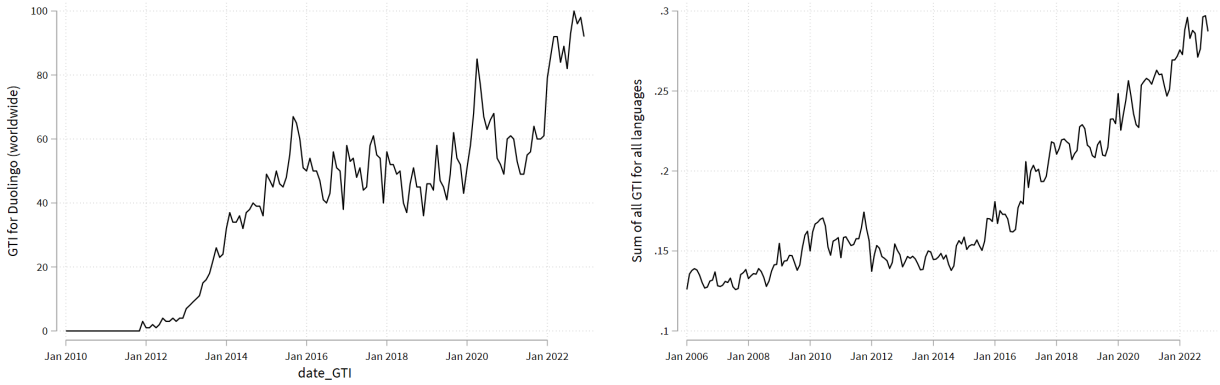
As one can query up to five terms or topics, one can identify the normalized relative search intensity across these terms or topics.[54] In case one has a larger set of terms or topics $\mathcal{T}$, one has to query the search terms in overlapping folds of 5 (the first fold includes the first five terms, 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

[53] Search terms can be simply a set of words, or a Topic. Topics are defined by Google Trends, which enables searching for the popularity of a coherently defined topic across languages, taking account of synonyms.

[54] This can be done either across regions for a given time period or over time for a given geographic region.

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. 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 our purpose, repeating this process two or three times suffices to obtain a distribution of Google Trend Indices that does not change upon another repetition. Through this so-called anchorbanking procedure the relative interest in keywords and topics for a given geographic are can be determined, $GTI^{T}_{world,t}$. As I are interested in the worldwide interest across topics, I take the time-average to obtain $GTI^{T}_{world,(2006-2022)}$.

Figure A11: Worldwide Google Trends Index for *Duolingo* and Languages over time



(a) Duolingo

(b) Languages

Relative search intensity for (a) *Duolingo* and its transliterations and (b) relative search intensity for all languages on *Google*.