Language, Identity and Ethnic Cohesion in Schools

Preliminary Draft Please Do Not Circulate

Ari Anisfeld *
University of Chicago
Job Market Paper

Pepi Pandiloski
University of Chicago

October 2, 2024
Click For Latest Draft

Abstract

Building social cohesion in diverse societies presents a significant challenge, one that is further compounded by linguistic diversity. This study investigates the impact of language and identity on cohesion built through inter-group contact in segregated high schools in Macedonia, where students are segregated by language of instruction. Using a pair-level contact intervention, we randomly assign students to complete cooperative tasks with either co-ethnic or cross-ethnic partners. Randomization also creates variation in the language skills that partners share. Participants in cross-ethnic teams communicate less and rate the experience lower than those in co-ethnic teams. Nonetheless, inter-ethnic contact enhances social cohesion without a statistically significant effect on productivity. The cohesion, experience, and communication effects are strongly correlated with the level of mutual language skill (a combination of predetermined covariates). Inspired by exposure mapping, we use propensity weighting to disentangle *match effects*– the benefit of having shared language skills – versus *selection effects* – the association between unobservables and language skills. The existence of match effects suggests room for policy to improve cohesion at a relatively low cost. Finally, we ask: Does the particular language for which students have the highest communication potential matter? Student experience and cohesion outcomes increase when they match on their home language (i.e. they have a cross-ethnic partner who can communicate in the focal player's home language). Contrary to theory, there does not appear to be a "backlash" effect,

*Corresponding author: anisfeld@uchicago.edu. We would like to thank Steven Durlauf, Anjali Adukia, and Lesley Turner for their thoughtful advice and gracious support, and to Oendrila Dube, James Robinson, Leo Bursztyn and Raul Sanchez de la Sierra for their invaluable guidance throughout the project. We also thank Violeta Petroska Beshka, Dan Black, Arun Chandrasekhar, Terrence Chau, Josh Dean, Michael Dinerstein, Ben Feigenberg, Max Foxman, Rachel Glennester, Jens Ludwig, Alicia Menendez, Max Tabord-Meehan, Molly Offer-Westort, and Linh Tô as well as seminar participants at the Harris School of Public Policy PhD Workshop, the University of Chicago Committee on Education Workshop and to the participants at the many brown bag lunches who provide feedback and moral support over the years. This research is funded by the JPAL European Social Inclusion Initiative, The Pearson Institute, the Development Economics Center and the Becker Friedman Institute at the University of Chicago. We'd like to thank Luan Imeri, Safet Ballazhi, and Zana Jusufi-Osmani at the Center for Human Rights and Conflict Resolution for excellent project implementation in the field, Jasmina Mihailovska and Tea Todorova at M-Prospect for coordinating data collection, and to 20 enumerators and 19 session facilitators whose diligent work made this study possible. We had excellent RA support from Alex Chen. Ari, specifically, expresses gratitude to Nikola and Nvenka Pandiloski, along with Pepi's extended family and friends, for a warm and welcoming introduction to Macedonia.

and if anything speaking a neutral language is worse than using the other player's home language. The results have implications for integration and language policy.

1 Introduction

The cohesion of a community reflects ongoing effort to build trust and a spirit of cooperation among its members. In diverse societies, communication frictions can increase the cost of this effort and shift who bears the costs, leading to increased tensions and lower investment in public goods (Alesina and Ferrara, 2005). To overcome these frictions, schools function as social capital builders where students learn shared identities and social skills in integrated classrooms (Alan et al., 2021; Chyn, 2018). However, in linguistically diverse societies, policymakers face a tradeoff between cohesion and learning. Learning in one's mother tongue can be more efficient, particularly early in life, but this typically demands segregation by language–reducing the development of mutual communication skills and common identities.[1]

Many communities face tradeoffs between cohesion building and efficient learning in home languages.[2] In the United States, for example, one in five households use a language other than English, and state-level policies have oscillated between banning and providing well-resourced bilingual education.[3] India, with over 100 local languages, has struggled to implement a language policy that carves out more resources for home language instruction (Jain, 2017; Kumar et al., 2022). African governments, from South Africa to Uganda to Morocco, face similar challenges (Heugh, 2013; Miguel, 2004; Angrist and Lavy, 1997; Laitin and Ramachandran, 2022).[4] In lieu of integrated classrooms, a common policy prescription is for local institutions to organize activities where students from different groups can cooperate toward a common goal (Allport et al., 1954).[5] But, language skills developed under segregation may limit the effectiveness of such contact interventions. In this paper, we ask: How does language mediate the development of social cohesion? If two students have a better ability to communicate between themselves, do they have a better experience? Do they perform better on cooperative tasks? And how do measures of inter-group cohesion change as a result?

[1] The World Bank and related international institutions have promoted mother tongue (L1) instruction as good education policy. (See also Piper et al. (2016) and Laitin and Ramachandran (2016)).

[2] The tradeoffs also consider controlling costs and developing language skills to access regional or international markets.

[3] Lawmakers most recently passed bans in California (1998), Arizona (2000), and Massachusetts (2002), though only Arizona's law still stands as of 2024. See Matsudaira, n.d. for a review. A tragic example comes from U.S. Indian policy where as recently as 1980 Native American students were sent to boarding schools where school officials beat them for speaking their languages (Churchill, 2004).

[4] The list goes on to include Pakistan, Indonesia, Kazakhstan, Ukraine, ethnically diverse regions such as Quebec in Canada, Catalonia in Spain, Bolzano in northern Italy or Luxembourg that support multi-lingual education systems. Areas in Latin America where indigenous language speakers have fought for the rights to mother tongue instruction (Hynsjö and Damon, 2016) and so forth.

[5] An additive bilingual or multilingual approach provides a middle ground. In such a setting, home language instruction is the primary language of instruction in early grades, but not the only language. Later students take coursework in additional language(s). Of course, schools must be integrated for this to work. Heugh (2013) discusses the challenges of implementing this framework in post-Apartheid South Africa.

Moreover, language is more than a "communication technology" as it has a deep connection to identity. We further ask how identity considerations affect the process, if at all.

We investigate these questions in high schools in Macedonia where schools segregate ethnic Macedonian, Albanian, and Turkish students by language of instruction. Field observations, interviews, and focus groups suggest that segregated learning, media access, and labor markets have increased segregation.[6] In 2001, an armed conflict between Albanian militia and Macedonian authorities was resolved with the Ohrid Framework Agreement, which expanded the language rights of Albanians including providing public funding to Albanian-language colleges.While Albanians and Turks had access to home language instruction before 2001, the associated policy changes reduced access to and need for the Macedonian language (Fontana, 2017). At the same time, the cost-benefit calculation for language learning has shifted due to increased access to non-Macedonian media and aspirations to work in richer European countries or the United States. These factors can erode mutual language skills, exacerbating already low cross-group interaction (Lang, 1986).

To study how identity and language skills drive outcomes during and after inter-group contact, we employ a randomized contact intervention and two waves of surveys. We randomize students into pairs to play "Overcooked! 2", a cooperative video game that requires teamwork and communication as players rush to prepare meals for hungry customers. The design provides two sources of variation. First, control students get partners from the *same language* of instruction, while treated students are paired with students from a *different language* of instruction. Stratifying pairs by school, grade level, and gender increases the salience of ethnic difference. Second, random pairs generate random variation in the "communication potential" across pairs. Some pairs fluently speak a shared language while others may struggle to communicate in any language. We develop a measure called the "max language match" which predicts what language the pair has the best chance of mutual understanding. Students self-reported language skills in Albanian, Macedonian, Turkish, and English on a scale from 0 to 10. To calculate *"max language match"*, we take the (squared) geometric mean of the pairs' skills for each language and then pick the largest number. The measure does a good job of predicting the amount of communication and which language the pair will use, if any. We use this underlying variation in a reduced-form, or 'intent-to-treat', framework to understand the heterogeneity of effects due to shared language skills.

[6] See also Fontana (2017).

We measure two intermediate and two ultimate outcomes. The intermediate outcomes capture the amount of *communication* between the players and the players' *experience* during the session. These are measured through survey responses from the players and the session facilitators. The two main outcomes are *cohesion* and *productivity*. We measure cohesion with incentivized dictator games and a Bayesian adaptive choice experiment (Drake et al., 2022). The first measure, common in the literature, is the difference in the amount of money shared with co- and cross-ethnic peers. The second measures the willingness to pay (WTP) for a same-ethnicity partner. Dictator games are a natural cohesion measure since they capture a notion of social closeness. The adaptive choice experiment allows us to estimate individual-level parameters using an adaptive sequence of trade-offs akin to a binary search. Individual estimates let us understand the heterogeneity of effects by comparing demand curves induced by the WTP estimates. To measure productivity, facilitators provide reports of student scores and we also capture screen recordings from pair gameplay.[7]

We find that students with cross-ethnic partners have more challenges with communication and rate their experience lower than students with co-ethnic partners. Nevertheless, students with cross-ethnic partners show cohesion gains. On average, the cohesion index increases by 0.19 standard deviations. "Demand for segregation" shifts inward starting at moderately low willingness-to-pay levels and continuing throughout the demand curve. The improvements in cohesion do not come at the cost of lower productivity, as scores in the game are not statistically distinguishable between the treatment and control groups. Thus, students in cross-ethnic pairs overcome verbal communication challenges to perform as well as their peers in co-ethnic pairs in the game, while improving their attitudes toward peers from different ethnicities. The overall cohesion effects echo the results from a typical contact intervention (Paluck, Green, et al., 2019).

To understand language effects, we borrow an idea from the econometrics of spillovers extending the potential outcome framework to define several language-specific treatments such as high versus low "max language match"(Aronow and Samii, 2017). As in the spillover setting, predetermined characteristics influence the likelihood an individual is assigned to a treatment. The key insight is that the propensity score is *known* for specific match types that correspond to pre-determined observables. Simulations of the randomization procedure provide the likelihood of matching with other students who have known skill vectors. Pairing these propensity scores with augmented inverse propensity weighting (AIPW) provides unbiased estimates of the

[7] We currently only employ reported scores but are in the process of extracting other productivity measures from the audio-visual data.

local average treatment effects.[8]

Before accounting for selection bias, "max language match" strongly correlates with the cohesion outcomes as well as the communication and experience measures. Accounting for differential propensity scores shows that the associations combine match effects–it is useful to have shared language skills–and some underlying unobservable type that makes students with high propensity scores more likely to communicate, enjoy playing, and show higher cohesion measures. Though noisy, this suggests there is room for matching as a policy tool for building cohesion–indeed, optimal matching would have improved the observed cohesion gains by roughly 10 percent. And we would see 30 percent gain, if everyone could have a high "max language match."[9]

Finally, we ask how the language the students match on matters. Contact theory suggests having equal status will improve cohesion outcomes. Does it help if both students are speaking a second language? Does it hurt if one student is speaking their home language? Further, due to the recent history of conflict and fights about learning Macedonian in school (see Section 2), does using a neutral language like English help build cohesion? We find that the language of the match matters, but not in the manner suggested by theory and contextual history. The results show strong *home language effects*–having the home language be the "max language match" leads to strong cohesion outcomes. The dictator game bias drops to nearly 0. We take this as evidence of language as a marker of shared cultural identity. At the same time, there is no evidence of "backlash" for their partners who continue to have positive cohesion effects. Finally, while noisy, there does not appear to be a "neutral language" effect–if anything using Macedonian increases cohesion outcomes relative to English for the non-Macedonian students.

Our study's main contribution is twofold. First, we introduce nuance into our understanding of intergroup contact. In a context with groups that have different native languages, we find that contact is ineffective in improving cohesion if parties have barriers to interaction. Second, we use these insights to inform education policymakers in our context to best design integrative policy. We show that targeting students who can better communicate during activities designed to foster cohesion can lead to better outcomes.

We complicate the contact hypothesis by considering the role of contextual factors such as skills and preferences. Under the contact hypothesis, integration produces positive outcomes if individuals from distinct

[8]Local to "compliers" who have a non-zero probability of a low or high match.

[9]This exercise makes use of personalized treatment effects estimated with generalized random forests.

groups work together toward a common goal where groups have equal status and the activities are condoned by authorities. (Allport et al., 1954; Pettigrew and Tropp, 2006; Paluck, Green, et al., 2019) We build on recent work with youth that has found mixed evidence — some showing strong positive results in attitudes and behaviors (Rao, 2019), others indicating changes in behavior but not attitudes (Scacco and Warren, 2018), and still others reporting generally positive results, which degrade as the proportion of out-group team members increases (Ghosh et al., n.d.). However, mechanisms for when and how contact works are not well understood. In a 2021 review, Paluck, Porat, et al. (2021) called for studies that unpack mechanisms, involve large sample sizes (the upper quartile is 78 participants in the extant research cited), and measure outcomes at least one day after contact. Earlier contact papers treat individuals as exchangeable based the study-defining group membership; such as racial category (Corno et al., 2022), religious background (Scacco and Warren, 2018; Mousa, 2020), caste (Lowe, 2021), economic class (Rao, 2019), or cross-cutting group identity(Bursztyn et al., 2024; Finseraas et al., 2019). Incorporating baseline skills, we show how individual heterogeneity and partner match effects can influence the individuals' experiences and outcomes of integration activities. Since skills can be learned (e.g. Alan et al., 2021) and assignments to contact experiences can be targeted, understanding this heterogeneity is important.

Finally, we provide insights for experimentalists interested in working with pairs and measuring productivity and cohesion. The use of pairs in experimental settings raises several complications and opportunities. We introduce novel techniques for implementing and analyzing such experiments. While several studies focus on partners such as roommates (Boisjoly et al., 2006; Corno et al., 2022) or deskmates (Elwert et al., 2023), these studies tend to use simple randomization schemes and often focus on the results from a "focal" partner. We randomize in a manner that ensures specific proportions of partner types while minimizing the recruitment costs for a given sample size. Further, we borrow from the literature on spillovers to estimate match quality effects based on predetermined characteristics. Moreover, we innovate on the measurement of social preferences and cohesion by integrating novel data sources and methods from outside of this literature. First, by using video games, a culturally appropriate medium for youth contact, we obtain a novel measure of productivity and effort in team setting.[10] Additionally, as described above, we adopted a measure from the labor economics literature on valuing workplace amenities to estimate the willingness to pay for segregation at an individual level (Drake et al., 2022).

[10]While we do not make use of it in this draft, the video game screen captures provide a rich source of data we are excited to learn from.

Outline The paper proceeds by describing the institutional context (section 2). Next, we discuss the experiment (section 3) and data and data generation (section 4). The analysis takes part in two steps: First, we introduce the empirical strategy for studying the effects of an inter-ethnic partner (section 5) and present results (section 6). Then, we introduce the empirical strategy for studying the effects of language match (section 7) and show the role of match effects and the importance of which language students match on (section 8). Finally, we conclude (section 10).

2 Setting

Post-conflict Macedonia provides a useful setting to study the role of language and integration. As detailed below, in the wake of the Yugoslav Wars, the country faced a crisis that led to the death of 200 people, and the displacement of nearly 7 percent of the population Merdzanovic, 2015. Avoiding broader conflict, Albanian and Macedonian political leaders negotiated a power-sharing framework that increased language rights and also likely increased segregation in the country. In this section, we provide general background on Macedonia, elaborate on the role of language policy and segregation in education, and share baseline data demonstrating the extent of segregation in the study sites.

Macedonia is a middle-income Balkan country of 2 million citizens with diverse ethnicities, languages, and religions. Nationally, ethnic Macedonians are the majority with Albanians the largest minority, followed by Turks, Roma, and other smaller ethnicities.[11] Our study site is an ethnically diverse town, but here the Albanians are the majority of the population, followed by Macedonians and Turks as large minorities.[12]

Questions about language policy underscore long-standing political tensions about the role of *nationalities* and cohesion in the multi-ethnic state.[13] Indeed, the role of language in education has been contested since

[11] To put numbers on it, in the 2021 census at least 58 percent identified as Macedonian, 24 percent Albanian, and 4 percent Turkish with contingents of Roma, Serbs, Vlachs, among others. 7 percent of the population are "persons whose data are taken from administrative sources" and do not have an ethnic affiliation in the census. Censuses in Macedonia are politically fraught. https://makstat.stat.gov.mk/PXWeb/pxweb/mk/MakStat/MakStat__Popisi__Popis2021_NaselenieVkupno__Naselenie__EtnoKulturniKarakteristiki/T1008P21.px/

[12] Gostivar census figures stand at 55 percent Albanian, 21 percent Macedonian, 12 percent Turkish with 6 percent not participating. Students from other municipalities attend high school in Gostivar lowering the Albanian share.

[13] The flying of national flags provides another flash point where these tensions become visible. In 1990s post-Yugoslav Macedonia, the government passed a law restricting the use of "foreign" flags, aimed primarily at Albanian's double-headed eagle. In defiance of the law, the Gostivar mayor Rufi Osmani flew the Albanian flag over the municipal building. On July 9, 1997, Macedonian police forces raided the municipality killing 3 and injuring more to remove the flag. Today, in front of the Gostivar municipal

the expulsion of the Ottoman Empire in the early 1900s. Since at least 1950, governments have funded Albanian-language primary and secondary classrooms in the region, though at times government support was low.[14] After the dissolution of Yugoslavia, Albanians opened an "illegal" Albanian-language university in Tetovo as political tensions ratcheted. In 2001, violence broke out between the Albanian militia and Macedonian authorities, but a large-scale conflict was averted with the Ohrid Framework Agreement of 2001. The Framework guided future constitutional amendments and policy changes that increased the rights of Albanians and other minorities. Notably, it made Albanian an official national language, provided for Albanian-language tertiary schools, and guaranteed bureaucratic jobs assigned by ethnic quotas controlled by the "Balancer" mechanism.[15] The policy changes increased the sense that the country had "a set of ... parallel education systems, depending on the language" according to 2010 policy document (cited in Fontana, 2017, pg. 257). Recently, there have been political calls for increased integration, such as through English language tracks in high school (Bloodworth, 2020). Nevertheless, the current approach to cohesion can be summarized as "If they don't interact, they won't fight."[16]

In practice, students from distinct ethnic backgrounds attend the same physical schools but are segregated by language of instruction from kindergarten through high school, limiting opportunities for cross-ethnic social interaction. As an illustrative example, consider two fictitious students: Amra, an Albanian girl, and Ana, a Macedonian girl, both born in the same year. Their parents enroll them in Goce Delcev Primary School for kindergarten through 8^{th} grade. Amra and Ana are rarely on campus simultaneously because the school day is divided into shifts organized by language.[17] Amra enrolls in the Technical High School, which also organizes shifts by ethnicity, providing minimal opportunities for cross-ethnic interactions. Ana attends the Gymnasium, a college preparatory high school, which organizes the day by grade level. Although she begins the day with all first-year students regardless of ethnicity, her classes are conducted in Macedonian, and her peers are predominantly ethnically Macedonian. Additionally, Macedonian classrooms are in differ-

government building fly the flags of the municipality, Macedonia, Albania, Turkey, the Roma, and the United States.

[14]For example, nationalistic policymakers reduced access to non-Macedonian language instruction through the Law of Secondary Education 1985 (Official Gazette of SRM No 16/85). The law required 30 students from a language group in a cohort to have a secondary school class in their preferred language. As a result, the number of high schools offering Albanian language classes fell from 39 in 1980/81 to 5 in 1990/91 (Ortakovski (1998), see p. 362). The numbers rebounded post-independence.

[15]For a recent NY Times discussion see https://www.nytimes.com/2024/05/15/world/europe/north-macedonia-ethnic-balancer.html

[16]Private conversation with a Macedonian education policymaker in October 2023. See also Fontana (2017, p. 257) citing the 2010 policy document *Strategy for Integrated Education*: "ethnic separation is seen as a measure of conflict prevention."

[17]The schools alternate which ethnicity takes the morning shift each week; during Ramadan, Albanian students take the afternoon shift for the entire month.

ent hallways from Albanian and Turkish classrooms. Thus, despite increased opportunities for cross-ethnic encounters, Ana's K-12 experience also remains largely segregated.[18] Outside of a school setting, students might have opportunities for cross-ethnic contact if they grow up in relatively integrated neighborhoods or participate in activities outside of school.[19]

While completely segregated, the school system provides opportunities to study several additional languages in school. As of 2007, everyone begins English as a second language in first grade. In fourth grade, Albanian- and Turkish-tracked students commence Macedonian language instruction, as mandated by the Constitution.[20] In 2010, the government attempted to make Macedonian the *first* second language, but the policy faced severe backlash and was overturned by the courts (Fontana, 2017). A common complaint is that Macedonians are not required to and typically make no effort to learn Albanian (or Turkish). Theory predicts such communication free riding by the majority (Lang, 1986). Figure 1 shows that many but not all students report high skills in a second language.[21] The figure shows we can expect variation in whether two random students have complementary language skills. As expected, Macedonian students tend to have no second language skills in local languages and specialize solely in English. Meanwhile, Albanians and Turks have more varied language skills.[22]

In this segregated environment, baseline measures show low levels of social integration and cohesion (See Table 1). First, social networks are quite segregated. We elicit social networks using several methods including explicitly asking students to name up to six cross-ethnic peers. Only 45 percent of students name at least one peer who speaks a different mother tongue; compared to 96 percent who name at least one friend. In total, 88 percent of all distinct named links are co-ethnic.[23] Second, we find that students feel very close to students from their language-of-instruction (LOI) group and quite far from students from different LOIs. We measure this using the *inclusion of self-in-others* measure where a respondent selects one of 7 pairs of circles that range from barely touching to nearly completely overlapping (Aron et al., 1992). One is labeled "you"

[18] The Medical High School, the third high school in our study, also organizes shifts by cohort.

[19] Mixed families are rare–in our data, only 6 students report multiple ethnicities. We do not have data on the spatial extent of segregation in Gostivar–from discussions with locals, there are both mixed and highly segregated neighborhoods and villages.

[20] "In schools where education is carried out in another language, the Macedonian language is also studied" Article 48 of the Constitution of Macedonia (1991)

[21] That is they report an average of 7.5 on a scale from 0 to 10 for skills in that language. See below for details on the measure.

[22] Linguists say language similarity predicts ease of learning; here, the languages all come from distinct branches of the language tree; Albanian, Macedonian and English are Indo-European languages that are Albanian, Slavic, and Germanic respectively; Turkish is a Turkic language and has the least similar grammar.

[23] This represents a lower bound as 148 students use all 6 slots for their friend clique.

and the other has the name of a group (i.e. "Turks", "students from the Medical School"). We ask "what picture best describes your relationship with [group identifier]?". On average, students give their in-group LOI a 6.4 out of 7 while out-group LOIs average 2 out of 7 (barely overlapping circles). Finally, students share less with out-group members in dictator games. On average, there is a 12 percent gap in baseline sharing (See section 5).

3 Study Design

In this section, we describe the recruitment and data collection timeline, how we randomized students into co- and cross-ethnic pairs and the procedures for the lab-in-the-field experiment.

Recruitment and Timeline

All students at the three focal high schools could participate in the study. From December 2023 to May 2024, enumerators hired by the survey firm visited classrooms, explained what participation would entail, and helped students complete the initial assent form. (Participation required parental consent as well.) Enumerators typically shared the mother tongue of the students they recruited and were always fluent speakers of the students' language of instruction. To avoid experimenter demand effects and reduce self-selection into the study, we framed the study around friendship and social learning. The consent language stated: *"We are conducting a study to understand how students make friends and learn about educational opportunities. Participation includes answering questions and participating in games. As part of the study, you may have the opportunity to participate in free tutoring/school enrichment in English."*

Data comes from three sources: (1) a baseline survey, (2) video game sessions, and (3) an endline survey. Upon receiving parental consent, enumerators conducted the baseline survey that featured modules on language skills, social networks and social closeness, parents, video games, educational outlook, and a real stakes dictator games.[24] Once both partners in a pair completed the baseline survey, we sent the session information to the facilitation team who organized the video game sessions described below.[25] Then at least two weeks later, enumerators conduct the endline survey asking students to sign up for our inter-ethnic course,

[24] The baseline and endline surveys are available upon request.

[25] A handful of students participated in the video game session before the baseline survey.

and re-asking social network and behavioral measure questions from the baseline survey. Many of the same participants were invited to participate in our companion study, where students, in groups of 6, play a networked social learning game, passing information to peers via a website we built for the study (Anisfeld and Pandiloski, 2024). The second study involves limited cross-group interaction as students are instructed to remain quiet. Nevertheless, we strive to have students complete the first three steps before participating in Study 2.

Data collection is ongoing. As of September 1, we have completed nearly all baseline surveys, 262 video game sessions, and roughly 340 endline surveys. Since we do not have endline data for the current population, we do not include an analysis of it in this draft.

Randomization

We randomize students into pairs stratified by school, grade level, and gender; keeping gender and age constant within pairs to help isolate the role of language and identity. For efficient logistics and cost minimization, we randomized within strata on a rolling basis to provide steady work for the facilitation teams.[26] Randomization into pairs based on individual characteristics requires randomization that is more complicated than coin flips. In particular, we must ensure that individuals assigned to a specific type of pair (treatment status) have an adequate match. Our procedure (detailed in Appendix B.1) adjusts *individual* control assignment probability within strata downward to hold the *strata-wide* control assignment proportion at 25 percent. The method maximizes the sample size given the recruitment effort. The procedure is fully automated, which allows us to determine *de facto* assignment probabilities and the probability of language skill match for an individual within a strata. We did not have access to self-reported ethnicity when conducting randomization and relied on language of instruction as a proxy. In practice, 96 percent of students' ethnicity matches the language of instruction.

Main intervention: video game sessions

In the session, pairs of students learn to play "Overcooked! 2", a cooperative video game where players work as chefs to prepare as many orders as possible under time pressure. During the sessions, facilitators led students to a room with two computers prepped for survey taking and video game play.[27] Students introduce

[26] 11 strata have one wave of randomization, 10 strata have two waves, and 2 strata have three waves

[27] We hired teachers from the school to organize and run sessions. Each session had two teacher facilitators so that each student

themselves and answer two ice breaker questions: "What languages do you speak?" and "What is your favorite class?" Then, using Qualtrics on separate computers, students verified their name and birth month and proceeded to learn how to play the game from YouTube videos[28] and then played practice rounds in "solo mode", where an individual player controls both chefs.[29] The "solo mode" scores provide a measure of individual skill.[30] Then, students play together in "cooperative mode" on a single computer, each student controls one chef and they work together to provide sushi (level 1.3) or pasta (level 1.5) to hungry, impatient customers. Pairs play each level twice, a practice and an "official round", wherein students get payouts based on the team score. In addition to learning and playing "Overcooked! 2", students answered questions about their experience in the session, completed an adaptive choice experiment, provides additional behavioral measures. Based on their scores in "official rounds" and dictator game responses, we calculate payouts that the field team leaders deliver at a later date. The entire session lasts an hour on average, with pairs interactively playing the game for roughly 20 to 25 minutes.

"Overcooked! 2" has many advantages.[31] First, students work cooperatively towards a common goal on equal footing, matching the criteria set out by the contact hypothesis. Second, productivity in the game depends on good communication and cooperation between players. To give some flavor, level 1.5 has chefs prepare spaghetti with red sauce. To prepare one dish requires boiling pasta, chopping tomatoes, sauteing tomatoes, plating the pasta and the sauce before either one becomes "overcooked"[32], and finally delivering the plated dish to the window. As orders go out, plates get dirty and must be washed. On top of this, the kitchen is in hot air balloons that shift over time as wind gusts blow them and the chefs across the screen. While each chef could work independently, like Adam Smith's pin factory, there is profit in the division of labor. We hypothesized that language skills would have a meaningful impact on students' ability to divide labor and score points. Finally, video games are a culturally appropriate medium for youth contact. In our

had at least one teacher who taught in their language of instruction, or was fluent in their language of instruction. We directed facilitators to interfere with gameplay as little as possible, and not at all during the "official play".

[28] We produced these in each of the local languages.

[29] Students play level 1.1, watch a second video, and then play level 1.2 twice. The second time is an "official round", which counts toward their payout.

[30] These measures may be affected by treatment assignment, so our main specification does not use them.

[31] For researchers considering using a cooperative game for a research project, we considered a few off-the-shelf alternatives. "Keep Talking and Nobody Explodes" gives players asymmetric information as partners race to diffuse a bomb. The game instructions and manual are too English-centric for our use case. Some puzzle games require cooperation, such as "Pico Park", but productivity is harder to measure in these games. We also considered "Moving out". Earlier research in communication studies used "Overcooked! 2" in a lab setting (Velez et al., 2021).Benatov et al. (2021) devise a bespoke curriculum for "contact through gaming" using Minecraft and similar work could be conducted in Roblox. Kovalčík et al. (2023) discuss additional games used to study teamwork and cohesion.

[32] In which case, the kitchen catches fire!

sample, two-thirds of students could recommend a video game by name, 55 percent said they played some days or more frequently and 72 percent said the expression "I love video games" describes them somewhat or more.[33] The figure is 87 percent for boys and 60 percent for girls.

4 Data and Measures

Sample characteristics and balance In Table 1, we present summaries of baseline characteristics and balance tests for the sample of students who have participated in the experiment through June 2024.[34]

The ethnic composition of the sample reflects that of the town. The plurality of students are in Albanian language of instruction (LOI) tracks, followed by Macedonian LOI and then Turkish LOI. Ninety-six percent of participants report an ethnicity that matches the language of instruction.[35] In practice, 7 students whose LOI does not map to their ethnicity are assigned to control unit "co-ethnic" pairs.[36] The hard-to-classify individuals attenuate treatment effects as some control units may act like treated units. The main demographic variables are balanced across treatment and control except Albanian LOI. In the current sample, the treatment group skews Turkish.

Above, we reported baseline cohesion measures and video game skills. The treatment group in the analysis sample has hints of imbalance. They name .14 more cross-ethnic links than the control ($p < .05$) and have lower dictator game bias -4 denar ($p = .67$), yet they give less money to outgroup peers in dictator games (-15 denar, $p = .1$). We use generalized random forests with these and other baseline characteristics to account for chance imbalance.

[33] At least 5 on a scale from 0 to 10.

[34] Table A1 and A2 contain data for the full sample who have been randomized a partner. The full sample appears more "balanced" as the p-values appear closer to a uniform distribution.

[35] There are two prevalent exceptions. First, There are 15 Roma students in our study. Roma are the fourth largest ethnic group in the Municipality (<4 percent, Per census data. https://makstat.stat.gov.mk/PXWeb/pxweb/mk/MakStat/MakStat__Popisi__Popis2021__NaselenieVkupno__Naselenie__EtnoKulturniKarakteristiki/T1008P21.px/). Students from Roma backgrounds tend to complete schooling in Macedonian and report the strongest sense of belonging to Macedonian students. Using the Inclusion of Others in Self measure, Roma students feel close to Macedonians at a level slightly below what Macedonian students report (5.6 versus 6.4 out of 7). Second, roughly 10 percent of students who identify as Turkish in this study complete coursework in Albanian or Macedonian. Several tracks within the technical high school (e.g. "architecture") are not offered in Turkish. Additionally, some ethnically Turkish students speak a different local language as a home language.

[36] There are 16 students in this group altogether in the current sample but the 'treated' units fortunately have a truly "cross-ethnic" partner. The classification issues are slightly larger in the full sample affecting 36 of 822 students. In future robustness tests, we will reclassify pairs and BACE analysis to reflect ethnicity.

Max language match

"Max language match" is a key component of the analysis. Partners may or may not share proficiency in English, Macedonian, Albanian, or Turkish. *A priori*, we would like to know to what extent partners have mutual communication skills, which will help understand match effects.

To measure the language skills, we employ a commonly used survey instrument the Language Experience and Proficiency (LEAP) Questionnaire (Marian et al., 2007). Students self-rate their English, Macedonian, Albanian, and Turkish skills using a scale of 0 to 10 in four domains: speaking, listening, reading, and writing. We average to provide one score per language. Then, for pair i and j, we determine a joint score in a given language by multiplying i and j's scores.[37] We define *"max language match"* as the maximum joint score over the four languages. Formally, each student i has a vector of linguistic skills $f_i \in [0, 10]^4$. Then, for pair (i, j)

$$\underset{\text{Max language match}}{f_{ij}^{\max}} = \max(f_i \odot f_j)$$

where $\odot$ denotes the element-wise (Hadamard) product and the $\max$ operator selects the larges element in the vector. The "max language match" reflects the two-way nature of communication and the inherent complementarity of communication skills. The "max" captures first-order communication skills. For co-ethnic pairs, $f_{ij}^{\max}$ is typically $10 \times 10 = 100$. For cross-ethnic pairs, we see substantial variation in $f_{ij}^{\max}$. It has a median of 75, an interquartile range of 62 to 88. Students are likelier to use languages with high joint scores; less than 10 percent of pairs have a second-place language match above 75 suggesting that the language associated with the top score is the most likely to be used for communication. Facilitators report on pairs language usage which provides supporting evidence (See Appendix **??**). For control pairs, we set the "max language match" to 100 since this is generally true and the variance of the measure is close to 0.

Outcomes

We build 4 indices to measure communication, student experiences, productivity, and cohesion. For each outcome type, we construct indices of 2 or 3 related variables. To make sure variables are comparable, we

[37] This corresponds to the geometric mean (squared).

normalize by subtracting the control group mean and dividing by the control group standard deviation. Then, we take the simple average of all composite variables to generate the indices.

Outcomes: Cohesion Measures

The cohesion index comprises

1. Willingness to pay for co-ethnic partners
2. Dictator game bias

Adaptive Choice Experiment: Willingness to pay for co-ethnic partners

The amount a student is willing to pay for co-ethnic partners provides a snapshot of how much the student values segregation. Intuitively, students who chose co-ethnic partners over a cross-ethnic partner with a larger bonus are willing to forgo money to maintain segregation Hedegaard and Tyran, 2018. This measure captures a mixture of prejudice and homophily – students might be willing to pay more because they have many friends in their class and think it is fun to play with their friends. Treatment effects net out the homophily channel so the effects can be interpreted as a decline in prejudice, making it a natural cohesion measure.

Students face 12 binary choice problems where they reveal their preferences regarding partner ethnicity, partner language skills, and a sure cash payment.[38] Bayesian updating determines the next trade-off in a process analogous to a binary search. (In Appendix C, we detail the theoretical basis and estimation in detail.) The benefit of the machinery is that it allows us to estimate multiple parameters at the individual level where previous work has relied on take-it-or-leave-it offers which provide a single data point to understand the demand for segregation/integration (Hedegaard and Tyran, 2018; Rao, 2019; Lowe, 2021).

In the main analysis, we focus on parameters of the following linear utility function:

[38]The choice problem is framed as an opportunity to provide input for an end-of-year event. Students read: "At the end of the study, Gostivar Municipality hopes to host a competition where you can play 'Overcooked! 2' again. We will let you choose the characteristics of your partner and give a cash bonus for playing with them." We stress that there is no payment for gameplay scores and the partners should be considered equal in every other way.

$$U_{ij} = p_j + \alpha_i 1\{\text{same ethnicity(i, j)}\} + \beta_i^E 1\{\text{speaks English(j)}\} + \beta_i^M 1\{\text{speaks Macedonian as a second language(j)}\} + \epsilon_{ij}$$

Where p is a sure price paid for choosing partner j with a utility normalized to 1. α_i is the WTP for a co-ethnic partner and β_i^L is the willingness to pay for student second language skills in L, English or Macedonian. We estimate these parameters using population Monte Carlo in two steps. First, we estimate the vector of $\{\alpha_i\}_{i \in N}$ using a nuisance prior and then we determine an empirical prior using the mean and standard deviation of the leave-one-out set of parameter estimates (See Appendix C).[39]

The utility parameter α_i captures the amount i is willing to pay for a co-ethnic partner holding language skills constant. As stated above, the parameter cannot be interpreted as a direct measure of discrimination. However, average changes in α (i.e. a treatment effect) can be interpreted as a change in the "price of prejudice".

The individual willingness-to-pay parameters induce a demand curve that allows us to visualize heterogeneity in the population and heterogeneity of effects. We build the demand curve by counting what fraction of the population is willing to pay for a characteristic at a given price level. (The curve is one minus the CDF of the distribution of WTP measures).

As a sanity test, we examine correlations between language skills and demand for students who speak English or Macedonian as a second language. Students with an additional point of fluency in a given language are willing to pay 3.5 to 5 denar for a partner who speaks that language.[40] Appendix Figure A5 shows demand for partners with specific language skills disaggregated by LOI. The median Macedonian LOI student is willing to pay over 100 Denar for their partner to know Macedonian as a second language, whereas the median non-Macedonian LOI student has a WTP of 0.

Bias in dictator game sharing Dictator games are widely used to measure prosocial preferences of altruism or fairness norms and bias. Regardless of the underlying preference parameter, dictator games on networks show that people share more with people who they feel close to. For this reason, we focus on the difference

[39] In practice, the nuisance prior has a limited effect on the estimates; the correlations between a nuisance prior with standard deviation 100 versus one with standard deviation 200 is roughly .98.

[40] These come from simple linear regressions of willingness-to-pay price on the fluency measure with or without language of instruction fixed effects. In reality, the demand is non-linear in the fluency measure.

between the amount shared with co-ethnic versus cross-ethnic pairs.

In a *dictator game*, each participant is asked to split 600 Denar (roughly $12 USD) with a peer.[41] Each question reads: "Imagine you are splitting 600 denars between yourself and the randomly chosen **student from your high school with the same mother tongue as you**. They will not know who split the money. Decide how much of the 600 denars to send to the **student from your high school with the same mother tongue as you**. You will keep the rest of the money." Students make four splitting decisions varying the mother tongue ("the same" or "a different") and school membership ("from your high school" or "from another high school") of the otherwise anonymous partners. Some choices are realized, adding an incentive for truthful reporting.[42] We measure out-group bias by subtracting the average amount shared with a co-ethnic partner from the amount shared with a cross-ethnic partner. In the baseline survey, students share 29 Denar (13.5 percent) more with their co-ethnic partners on average.

We conduct both games during a baseline survey, directly after the contact intervention, and again during the endline survey, which allows us to measure persistence or fade-out.

Outcomes: Productivity

The productivity index comprises

1. Cooperative game scores
2. Survey question about the quality of partner

Our primary measure of productivity is the total points scored in the "official rounds" of cooperative play. Players get points for successfully delivering dishes; they get tips for delivering dishes quickly and even more tips for delivering the dishes in the same order in which they came in; conversely, players lose points for failing to deliver an order in time.[43] The total score encapsulates total productivity.[44] Because students

[41] We say 50 denar equals $1 when in reality the exchange rate is closer to 55 denar for $1. Purchasing power is higher in Macedonia. An espresso cost between 80 to 100 denar in Skopje in mid-2024.

[42] These games have been used successfully with high school students and younger children (e.g. Ghosh et al., n.d.; List et al., 2017).

[43] Each order has a timer which is visualized as a bar that turns from green to red as the bar shrinks.

[44] In future work, we will use image detection algorithms to extract additional information from the 120 hours of screen recordings. While the total score provides a holistic measure of productivity, the number of failed orders might provide a distinct measure of cooperative effort. We can also fairly easily extract the time series of points scored, which could provide a richer picture of playing

play in solo mode first and we record their scores, we have access to a baseline video game ability for each student (though these may be influenced by treatment assignment.)

Outcomes: Communication

The communication index comprises

1. Amount of communication, rated by facilitators
2. Ease of communication, rated by students

Facilitators rate the amount of communication and which languages they heard used. Facilitators rate "How much did the students communicate during the joint game?" on a seven-point scale. Inter-rater correlation is high (.78) and the measure is associated with student rating on ease of communication (.47), our second measure in the communication index.[45]

Outcomes: Experience

The experience index comprises student self-reports on their experience playing with their partner on three dimensions:

1. Enjoyed playing
2. Would play again with the same partner
3. Was partner fun?

The student measures are highly correlated within students, correlation coefficients range between 0.6 and 0.73. Facilitators also report, from their vantage, how much fun each player seemed to have and how frustrated they seemed to be. These measures correlate less tightly but in the expected direction with student self-measures. Correlation between facilitator-assessed enjoyment and student self-measures ranges from .27 to

styles. More challenging computationally, we could train a model to identify which actions go into making a dish and determine how much joint effort goes into the point totals.

[45] In future iterations, we hope to analyze audio from the screen recordings. This may allow us to unpack the quality and content of communication. However, we are concerned that sound quality is not uniformly good.

.33, while that between facilitator-assessed frustration and student self-measures range from -.23 to -.30. The concordance between measures gives us confidence that student self-assessments are reliable.

5 Empirical Strategy

5.1 Main results

To estimate average treatment effects, we run the regression:

$$y_i = \alpha + \tau D_i + \epsilon_i$$

and

$$y_i = \alpha + \tau D_i + \gamma_g + \beta X_i + \epsilon_i$$

y_i is an outcome normalized as described above or an index of normalized variables, D_i is an indicator of being assigned a cross-ethnic partner, τ is the treatment effect and object of interest reflecting the causal effect of participating with a cross-ethnic partner rather than a co-ethnic partner. α is the control group mean, which typically equals 0 or is close to zero due to normalization. γ_g is a strata fixed effect (school $\times$ grade $\times$ gender) and X_i is a baseline measure of the outcome if it exists. Specifically, we include solo scores for productivity measures and baseline dictator game bias for cohesion measures. We cluster the standard errors at the strata by randomization wave level.

5.1.1 Robustness with causal forests

We use causal forests, a "doubly robust" treatment effects estimator, to test the robustness of the main results against finite-sample bias due to chance imbalances and against biases due to variation in individual-level treatment propensity scores (Athey, Tibshirani, et al., 2018). At a high level, we include predetermined covariates to address chance imbalances and use inverse propensity weights to address variation in treatment

scores.[46] We use the `grf` package in R to implement the causal forest, following Algorithm 1 from Athey and Wager (2019) who recommend a data-driven approach for covariate selection which is said to improve the stability of the estimates.

The causal forest provides estimates of $\tau(X_i) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$ using random forests to estimate an adaptive kernel weights $\alpha_i(x)$ as well as auxiliary regression functions for the mean outcome value $\hat{\mu}(X_i) = \mathbb{E}[Y_i|X_i = x]$ and the propensity score $\hat{p}(X_i) = \mathbb{E}[D_i|X_i = x]$ (Athey, Tibshirani, et al., 2018).[47] The estimator combines these in the causal forest to estimate conditional average treatment effects given x:

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x)(Y_i - \hat{\mu}^{(-i)}(X_i))(D_i - \hat{p}^{(-i)}(X_i))}{\sum_{i=1}^{n} \alpha_i(x)(D_i - \hat{p}^{(-i)}(X_i))^2}$$

The superscript $(-i)$ denotes the estimators are calculated using subsets of the data that exclude i (i.e. they are "out-of-bag" predictions). Finally, the individual-level predictions are aggregated to provide average treatment effects using an augmented inverse-propensity weighting (see equations 8 in Athey and Wager (2019)).

The assumptions required for the causal forest to consistently estimate treatment effects are selection on observables (conditional exogeneity) and overlap. We include all the observables that are included in the randomization procedures, so this is a use case where that assumption holds. For the main analysis, nearly all students have a probability of being assigned treatment or control due to our randomization procedure.[48] With these assumptions met, the algorithm returns unbiased estimates of the average treatment effects with asymptotic normal standard errors, which are clustered following Athey and Wager (2019).

[46] Covariates include: *demographics*: indicators for grade level, school, gender, the language of instructions, whether the ethnicity matches the language of instruction; *re-randomization based propensity score* probability of being assigned control, probability of a high language match; *language ability*: fluency measure for Albanian, English, Macedonian and Turkish, as well as the max fluency measure for non-mother tongue local language; *socio-economic indicators*: indicators for cellphone ownership, whether a parent attended college, whether the student went to a Gostivar middle school, and if the student intends to get a graduate degree; *baseline measures*: trust score, life satisfaction score, number of friends named, number of cross-ethnic friends named, baseline dictator game bias, baseline dictator game sharing with the outgroup; *IOS measures*: measure for own ethnicity, mean of other ethnicities, own school and mean of other schools; *video game*: Likert score for "enjoy video games", indicator for whether they make a recommendation.

[47] The random forest regressions also use the adaptive kernel weights, so that $\hat{\mu}(x) = \sum_j^n \alpha_i(x) Y_i$

[48] Two students are the only representatives of their language of instruction within a strata, so we technically cannot assign them to the control condition.

5.2 Correlation between "Max language match" and outcomes

We next turn to the role of language. To understand how the language match correlates with outcomes, we run the linear regression:

$$y_i = \alpha + \gamma * D_i + \beta * D_i * \tilde{f}_{ij}^{\max} + \epsilon_i$$

Where $\tilde{f}_{ij}^{\max} = \frac{f_{ij}^{\max} - \bar{f}^{\max}}{\sigma_{f_{ij}^{\max}}}$ is the normalized "max language match". This gives γ the interpretation as the treatment effect at the mean level of "max language match" and β measures how much y changes on average with a 1 standard deviation[49] change in "max language match".

To summarize effects, we also use the following specification, after splitting the treated sample at the median "max language match".

$$y_i = \alpha + \beta^L D_i \times \mathbf{1}_{\text{match}_{p(i)} < m} + \beta^H D_i \times \mathbf{1}_{\text{match}_{p(i)} \geq m} + \epsilon_i$$

We are interested in β^L (β^H) which provides contrasts between the control group and the average y_i for individuals in pairs with Low (High) language matches, defined as being below or above the median m. We also report $\gamma = \beta^H - \beta^L$ and its confidence interval to test whether the differences in outcomes for participants in high and low language match pairs are likely due to chance.[50]

As discussed below, these heterogeneous effects estimates may suffer from selection bias. After presenting these results, we consider how to take advantage of random peer assignment to estimate a causal effect of a high linguistic match.

[49] 19 points in the joint score

[50] We retrieve γ from the regression: $y_i = \alpha + \beta^L T_i + \gamma T_i \mathbf{1}$

6 Contact Results

6.1 Contact improves cohesion without lowering productivity

Focusing first on the "contact intervention", students show cohesion gains after playing with a cross-ethnic partner compared to a co-ethnic partner. Co-ethnic pairs communicate substantially more and report a better overall experience. Nevertheless, treated and control pairs are similarly productive. The following paragraphs flesh out these results and provide context by highlighting some results in levels. Figure 2 shows treatment effects for indices and the items that compose them in standard deviation units, along with the ATE estimates from the causal forests (shown in the lighter shade). In Table 5 and 6 we report the same results and also include covariates.

Students in mixed-ethnicity pairs had a harder time communicating than those in co-ethnic pairs (Figure 2 rows 1-3) and had a worse experience (Figure 2 rows 4-7). To put treatment effects in perspective, facilitators report the amount of communication on a 7-Likert scale with "a moderate amount" of communication as the midpoint. 88 percent of co-ethnic pairs communicated more than "a moderate amount" compared to 36 percent for treated groups.[51] To put the experience results in perspective, we asked students to what degree they agreed with the statement "I would like to play again with my partner". 97 percent of control students said *agree* or *completely agree*compared to 85 percent of students in mixed pairs.

While students had trouble communicating, we cannot reject a null effect on productivity or perceived-ability measures (Figure 2 rows 8-10). Mixed-ethnicity pairs score 9 fewer points (s.e. 16) on average than co-ethnic pairs who score 328 points on average. The productivity loss is less than 3 percent and indistinguishable from 0. When accounting for video game skills (as measured with solo scores), the effects get slightly larger but remain indistinguishable from 0. Moreover, students perceive their partner's skills similarly across treatment arms, even after accounting for the partner's solo scores.

The final three rows of Figure 2 show that playing with a cross-ethnic partner improves cohesion measures. The cohesion index comprises dictator-game bias and willingness-to-pay for co-ethnic partners, multiplied by negative 1. Cohesion increases by .18 standard deviations, reflecting similar magnitude statistically sig-

[51] 55 percent of co-ethnic pairs talked "Constantly" compared to 11 percent of cross-ethnic pairs.

nificant declines in sub-measures. In levels, dictator game bias falls from 22.8 to 14.4 Denar.

Treatment effect estimates are quite stable when estimated with causal forests (lighter points on 2). This is reassuring, but not surprising given the data come from a randomized experiment where the propensity scores are very similar across units and the samples are relatively balanced on observables.

Demand for segregation bows inward

Demand curves show the distribution of valuations providing insights beyond mean differences. Figure 3a and 3b show demand curves induced by the individual valuations for a co-ethnic partner and for a partner who speaks Macedonian as a second language, respectively. The curves report the empirical probability that an individual's willingness to pay is above some price over the support of prices, censored at -100 Denar.

Starting with the control group's demand (red line), roughly 10 percent of students are willing to pay to have a cross-ethnic partner.[52] Most students are willing to pay for co-ethnic partners and demand is disbursed over the support from 0 to 500 Denar.[53] Roughly 20 percent of students have WTP above 300 Denar, roughly the expected value of participating in the video game session.[54]

While WTP declines by 20 percent on average, the curves show that the "cohesion gains" are not uniform. The demand curves show how the distribution of willingness to pay shifts in response to the intervention.[55] The cross-ethnic demand curve noticeably bends inward relative to the co-ethnic demand curve in the 25^{th} to 75^{th} percentile range. Further, we see a decline in demand at the highest prices. Assuming the ordering of valuation does not shift dramatically, the results are consistent with a model where students update their preferences – those who initially are close to indifferent due to low prejudice are not impacted by the experience.

The language results provide corroborating evidence. The demand for Macedonian as a second-language has a similar shape to the co-ethnicity demand curve for Macedonian students, but we see no difference between treatment and control units from non-Macedonian LOI tracks. On average, we see no treatment effect on WTP

[52] We use language of instruction as 'ethnicity' in this analysis. As noted earlier, for a small portion of students, the language of instruction is not their ethnicity. In the analysis not shown, we drop these students and get visually similar demand curves.

[53] Roughly 0 to 10 USD

[54] 25 percent of the expected value of participating in the study in total.

[55] A Kolmogorov-Smirnov test rejects the hypothesis that the demand curves were drawn from the same distribution (p = .017).

for English (KS p = .32). In other words, among the Macedonian population we see an increased willingness to interact with other ethnicities when they do not speak Macedonian as a second language. This is consistent with the ethnicity results; suggesting marginally more openness to working across ethnolinguistic lines.

6.2 Language match predicts outcomes

Figure 4 shows that "max language match" is positively associated with communication, experience, and cohesion. First, there is a tight correlation between the max language match score and communication–this is evidence in favor of using "max language match" as a measure of communication potential. Every standard deviation increase in $f_{ij}^{\max}$ is associated with a .53 (se =.089) standard deviation gain in the communication index. Second, students with higher match scores report better experiences; while there is a substantial gap at low match scores, the treated individuals approach the control group average as $f_{ij}^{\max}$ approaches 100 (slope = .28 (se =.074)). Next, the cohesion index shows a statistically significant relationship showing that there is a (slope = .078 (se = .029)). Finally, while productivity shows a similar pattern; it is not statistically distinguishable from 0 (slope = .089 (se = .06)).

Taken together, we find strong, but suggestive, evidence of language match influencing students' experiences and cohesion outcomes. Intriguingly, the cohesion and experience/communication indices are not tightly correlated. Meanwhile, the productivity index is tightly correlated with experience and communication (as is the cooperative score alone). To put numbers on this, the regression of productivity on communication has a slope coefficient of .2 (se =.05) while the regression of the cohesion on communication has a slope coefficient of .04 (se = .022), and the correlations are more extreme when regressing on the experience index. These correlations go against the notion of a simple functional role for language impacting cohesion by making communication easier.

Figure 3c shows the demand for segregation. Here we split students at the median "max language match" and call students above $f_{ij}^{\max}$ of 75 a "high match" and below a "low match". The biggest demand shift occurs for students in high-match pairs in the middle of the distribution. The students in the low-match pairs have demand curves very similar to the control until the highest prices. The Kolmogorov-Smirnov test rejects the null that high-match and control demand are drawn from the same distribution ($p < .001$) and fails to reject

the null for low-match versus control (p = .69). However, as we have already hinted at, the inward bowing of the "high match" versus "low match" demand curves can reflect heterogeneous treatment effects or selection bias.

In the next section, we will discuss how to move toward causal estimands of mechanisms. Recall that while partners are randomly assigned, the fluency match score $f_{ij}^{\max}$ is a function of endogenous predetermined skills; hence the plots show correlations. On the other hand, if there were no correlation between the potential outcome $Y(0)$ for a given index and the fluency match score, the x-axis of figure 4 would show the control group potential outcomes, since the indices are normalized based on the control group mean and standard deviation. For the communication index, this may seem like a reasonable assumption –in which case, we might be tempted to read conditional average treatment effects off the regression line. However, as we will discuss in the next section, that may still be misleading since the measure is conditional on the observation ($\mathbb{E}[Y(1, f_{ij}^{\max})|F = f_{ij}^{\max}]$) and not the unconditional potential outcome.

7 Towards Causal Heterogeneous Effects: Exposure Mapping

Random assignment of partners gives us hope at disentangling match effects–the benefit of having a partner with a high "max language match"–and underlying heterogeneity–the fact that students with broader language skills are likelier to be in a high-language match pair and may have correlated characteristics that co-determine language skills and cohesion outcomes.

We do this by estimating propensity scores from baseline skills. Then, with almost exact propensity scores, we use weights to estimate causal effects. This is the idea behind "exposure mapping" a technique used to measure spillovers in network settings (Aronow and Samii, 2017). In this section, we first provide an illusory example; then, formally define exposure mapping; and finally, bring the technique to our data and discuss limitations.

Illusory example: A toy example clarifies how selection bias *could* occur when splitting the sample based on a pre-determined characteristic and how probability weights derived from the randomized design address the bias. Consider a strata with two types of students in a trilingual environment. Bilingual students, labeled

B, have fluency in their home language and another language, while monolingual students, labeled M, only know their home language. Though we are equally likely to get $(M, M), (B, M), (M, B), (B, B)$ pairs, the probability a pair is a high language match is not equal. Bilingual students are likelier to get a high language match pair than monolingual students. In cross-ethnic pairs, two monolingual students (M, M) will *never* have a high language match, while two bilingual students (B, B) will *always* be a high language match. Depending on the particulars, (B, M) or (M, B) pairs may or may not be language matches. In our toy example, the $B - M$ pairs have a 50 percent chance of being a high match so that the overall propensity scores are $p(B) = \frac{3}{4}$ and $p(M) = \frac{1}{4}$, where $p(X) = \mathbb{E}[\text{High Match}|X]$.

To highlight the bias, Figure 5 shows results from a stylized simulation with constant treatment effects. We set $\tau = 0.2$ for all B and M, but assume heterogeneity in baseline cohesion measures. This is plausible. Bilingualism may endow B students with better perspective-taking (**kinzler**; **mamood**), which could, in turn, increase their cohesion measures regardless of treatment (Alan et al., 2021).

A few observations: First, the simulations show bilingual students are likelier to be in a high language match pair. Nevertheless, the pooled results return $\tau = \mathbb{E}[Y(1) - Y(0)] = 0.2$ since this is a randomized experience. Next, when we split the sample, we get biased estimates–it appears there are heterogeneous treatment effects because the conditional treatment effects use the unconditional expected outcome of the control group $\beta^{\text{Match Type}} = \mathbb{E}[Y(1,1)|\text{Match Type}] - \mathbb{E}[Y(0)]$. This insight suggests we can recover the CATE if we can estimate $\mathbb{E}[Y(1,1)] = \mathbb{E}[\mathbb{E}[Y(1,1)|\text{Match Type}]]$. Following well-known results, inverse propensity score weights allow us to address the bias.[56] Moreover, we know the true propensity score because we know how baseline characteristics map to a "high" or "low" match given the randomization scheme.

Exposure mapping: Here is how the toy example maps to the empirics from the previous section. First, recall in Section 5 we proposed a regression summarizing the effect of a high or low match:

$$y_i = \alpha + \beta^L D_i \times \mathbf{1}_{\text{match}_{p(i)}<m} + \beta^H D_i \times \mathbf{1}_{\text{match}_{p(i)}\geq m} + \epsilon_i$$

β^L and β^H may suffer from selection bias as in the toy example. To unpack the estimands, we extend the

[56] In this simple example, controlling for bilingual ability also works, but that will not translate so easily to the data.

potential outcome framework and then show how we can estimate causal estimands that capture the treatment effect of having a high language match partner. Later, we will discuss the interpretation of the estimand.

We define $d(z, l)$ where index z is 1 if the student is in a co-ethnic pair and the index l is 1 if the pair has a high-language match. This produces three potential treatment assignments

- $d(0) = d(0, 1)$: co-ethnic pair
- $d(1, 0)$: cross-ethnic pair, low language match
- $d(1, 1)$: cross-ethnic pair, high language match

When we run the regression of an outcome on a vector of ones and indicators for match type, the coefficients give the mean contrasts:

$$\begin{aligned}\hat{\beta}^H &= \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{(z_i,l_i)=(1,1)}} \left(\sum_{i=1}^{n} y_i \mathbb{1}_{(z_i,l_i)=(1,1)} \right) - \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{z_i=0}} \left(\sum_{i=1}^{n} y_i \mathbb{1}_{z_i=0} \right) \\ &\xrightarrow{p} \mathbb{E}[y(1,1)|(z_i, l_i) = (1, 1)] - \mathbb{E}[y(0)]\end{aligned}$$

with $\hat{\beta}^L$ defined analogously and $\hat{\gamma} = \hat{\beta}^H - \hat{\beta}^L$. We have bias in $\hat{\beta}^H$ unless $\mathbb{E}[y(1,1)|(z_i, l_i) = (1, 1)] = \mathbb{E}[y(1, 1)]$. Selection bias depends on the correlation between language skills and the distribution of potential outcomes. The experimental design allows us to identify the correlations between $Y(0)$ and outcomes. Appendix figure **??** shows these correlations are indistinguishable from 0. Nevertheless, we do not observe potential outcomes $Y(1, 1)$ and $Y(1, 0)$, correlations between pre-determined covariates and match type suggests $\mathbb{E}[y(1,1)|(z_i, l_i) = (1, 1)] \neq \mathbb{E}[y(1, 1)]$. Going from the 25^{th} to 75^{th} quartile probability of being assigned to a high-match pair is associated with .52 additional named friends (11 percent increase) and .13 additional named cross-ethnic friends (27 percent increase).If this sociability makes the partner more likely to have a higher $Y_i(1, 1)$ then our estimators are biased as we over-sample higher $Y_i(1, 1)$ individuals when estimating the average outcome $\bar{Y}(1, 1)$.

We can account for selection bias using propensity scores $\pi_i(z, l)$. The key insight of Aronow and Samii

(2017) is that the propensity score is *known* given we control the treatment assignment mechanism and know the language skill vectors. Simulations of the randomization procedure provide estimates of $\pi_i(z,l)$ with arbitrary precision. Run the randomization code R times and for each randomization collect the assignment $d(z,l)$ for each individual. Then, $\pi_i(z,l) = \frac{\sum^R \mathbb{1}_{z_i=z,l_i=l}}{R}$.

Then, under the following assumptions, we can estimate the causal effect of being randomly assigned a high-language match partner versus a low-language match partner

$$\gamma = \frac{1}{N}\left(\sum \mathbf{1}_{z_i=1,l_i=1}\frac{y_i(d(1,1))}{\pi_i(1,1)} - \sum \mathbf{1}_{z_i=1,l_i=0}\frac{y_i(d(1,0))}{\pi_i(1,0)}\right)$$

Assumption 1 Overlap: $0 < \pi_i(z,l) < 1$. In words, units with no possibility of being in a high or low language match must be dropped as the counterfactual is ill-defined. Roughly 14 percent of the current sample had 0 probability of being matched to a high match (46 students) or a low match (26 students). *Assumption 2 Correct specification*: The exposure mapping is accurate. In future iterations, we will explore this assumption further.

To increase power, we estimate γ with an augmented inverse propensity weighted estimator (AIPW), using causal forests to estimate these. Causal forests provide similar results to exposure mapping with some slight variation (discussed next section). Theoretically, the results should be similar because both methods adjust for differences in treatment propensity.[57]

8 The Role of Language

Now we return to the key questions about whether and how language skills influence cohesion building. Section 6 showed language skill match is associated with better experiences and stronger cohesion outcomes.

Language match and selection

[57] When using causal forests, we estimate propensity scores ($\pi(z,l)$) in first-stage random forests that use sample splitting to avoid overfitting. This smooths the propensity scores away from zero.(The minimum $\pi(1,0) \simeq .08$ and $\pi(1,1) \simeq .05$ across estimators.) This may be desirable since idiosyncrasies in recruitment impacted the propensity scores (see Appendix B.1). Causal forests produce consistent estimates when potential outcomes are independent from treatment assignment conditional on observables $Y_i(1,1), Y_i(1,0), Y_i(0)||D_i|X_i$ and there is overlap.

Recall that "max language match" is associated with better experiences and cohesion outcomes (see Figure **??**). To what extent do these patterns reflect match versus selection effects? The distinction is relevant for policy because they imply different solutions. If match effects drive the results, then a principal could increase cohesion through activities like this one by carefully pairing students with similar language skills. More broadly, an additional focus on teaching mutual languages would complement future integration efforts. On the other hand, selection effects would require further investigation. It could be some factor, such as openness to experience, that influences both individual language skills and cohesion measures, or it could be that learning languages itself enhances cohesion (e.g., through perspective-taking). In this case, the policy implications are less clear. Even more broadly, this distinction might be a useful tool for theorizing when and why integration efforts work.

We find evidence for both match and selection effects. Figure **??** shows the estimated average treatment effects of high match or low match compared to the co-ethnic control.[58] Language match has a clear impact on the process. Students report better experiences in high-match pairs than in low-match pairs (95% CI [-0.01, 0.43]) and communicate more (95% CI [0.237, 0.842]). The evidence for match effects is weaker for cohesion measures (95% CI [-0.0629, 0.289]).

Knowing the propensity score allows us to gain valuable insights about the underlying effects. The probability of being assigned a "high language match" is known for every individual and it indexes types; *i.e.* students with a 40 percent probability of being assigned a high match are likely to be similar in expectation. If we had much more data, we would be able to estimate marginal treatment effects non-parametrically. Given limited power, Figure 7 shows the regression lines of outcome indices against the probability of being assigned to a "high language match" ($\pi(1,1)$). Each plot has a separate line for $Y(0)$, $Y(1,0)$ and $Y(1,1)$ allowing us to compare how potential outcomes vary over the distribution of types (at least in terms of the line of best fit). Notice that the potential outcomes for co-ethnic pairs (i.e. $Y(0)$ or red lines) tend to be flatter, which suggests that being a higher type is not associated with better outcomes *a priori*. Students with a 30 percent chance of getting a high language match partner have similar cohesion outcomes as those with a 60 percent chance, *when assigned a co-ethnic partner*. While the results are noisy, the cohesion plot hints at both selection on gains (the lines appear upward-sloping and faster than the co-ethnic lines) and a language match effect (there

[58] Appendix Figure A3 shows estimation with exposure mapping. While Aronow and Samii provide a formula for conservative standard errors, I bootstrap for now. The procedure uses the same exposure mapping weights for each iteration. The results are similar to those in the random forests, though noisier.

is a consistent gap between the high match and low match lines). These results suggest room for "benign self-selection" into contact in the real world. For example, if we organized an activity without randomization and students who can better communicate mutually selected in, the results might be particularly good.

The causal forests produce personalized estimates of treatment effects β_i^H and β_i^L conditional on covariates. The range of treatment effects for a low match is between -0.01 and .26 standard deviations, while for a high match, the estimates range from .099 to .42. This is nice to see because it implies there are very few students for whom treatment "backfires". These estimates should be treated with caution due to the limited power.

Does the language used matter?

Next, we turn to mechanisms, focusing on hypotheses driven by theory and context. Contact theory predicts that when individuals are on equal footing, cohesion outcomes improve more. Language skill differences could upset this balance, creating a power dynamic that may cause resentment. This is particularly concerning in this setting, given the fraught history of Macedonian as a second language (see Section 2).

We address these issues with three sets of tests.

Home language versus second language: In pairs where one student speaks their home language, the second language speaker is at a communicative disadvantage. There is indeed a home language effect—students who speak their own home language have higher cohesion outcomes. They also report better communication and experiences and rate their partners as better at the game. While second-language speakers have significantly lower outcomes, there is no evidence of "backlash." These students still have positive cohesion outcomes on average. Drilling down into sub-measures, the cohesion differences come from dictator game sharing. Home-language speakers' dictator bias is 0.5 standard deviations below the co-ethnic-pair mean. In real terms, the bias falls from 23 Denar for control students to 1.5 Denar for home-language speakers. If dictator games measure perceived social closeness, this result suggests that speaking in someone's home language can significantly reduce gaps and extend to the entire out-group (at least in the short term).

English and Macedonian as lingua franca: If equal footing matters, then using a *lingua franca*—a shared second language—should yield desirable outcomes. However, this is not the case. Pairs who speak a *lingua franca* have similar fluency match scores to those who speak a second language with a home language

speaker, but lower communication levels, lower reported experiences, and lower cohesion outcomes. While experience and cohesion results do not reach conventional significance levels, the directional pattern suggests there is no specific gain from students having equal status in terms of the language used.

Macedonian or English as the language used: The previous results pooled English and Macedonian as *top languages*. Discussions with teachers and school administrators suggested that English was preferred by students. Moreover, the history of language policy in schools suggests there might be a distaste for using Macedonian. To test this, we focus on Albanian and Turkish students who are randomly paired with a student whose top match language is either English or Macedonian. History and context suggest that English would lead to better outcomes, but it does not. Differences in communication, experience, and cohesion results are indistinguishable from 0.[59] Directionally, teachers' predictions hold for communication and experience indices. However, if anything, students have better cohesion results when using Macedonian. These noisy differences come from the dictator games, where the average bias is -0.203 (se = 0.228) standard deviation units lower for students paired with a Macedonian-speaking top match than with an English-speaking one.

In summary, the language used matters, but not in the manner suggested by theory and contextual history. We find evidence that using local languages—particularly one's home language—leads to stronger cohesion outcomes. Since these results are driven by dictator game bias rather than demand for segregation, we theorize this reflects an identity channel. Hearing your home language or even a local language appears to be associated with lower bias in dictator games.

9 Policy Experiments

Finally, we imagine a school implementing paired video gameplay as a tool to build cohesion across linguistic groups.[60] We estimate the range of potential average treatment effects under the possible matching of partners. We draw on matching algorithms from network theory to find the optimal pairings.[61] To sketch the idea, we create an edge between each cross-ethnic pair within a strata; then, we calculate the "max language

[59] We focus on the overlap sample with some probability of being assigned a top match in English or Macedonian.

[60] Video game events occur in the high schools, though they do not seem to be focused on integration, *per se*.

[61] Specifically we use Edmond's Blossom algorithm for weighted graphs which is implemented in the Python project *networkx* (**networkx**; Edmonds, 1965).

match", depending on the match type, we assign a weight to the edge that equals the sum of the treatment effect of the two students. Formally, for edge (i, j), if the linguistic match is above the threshold, we assign edge (i, j) weight $w_{ij} = \beta_i^H + \beta_j^H$. If the two students have a low match, we assign them the sum of their respective low-match treatment effects $w_{ij} = \beta_i^L + \beta_j^L$. Then, we find the optimal matching which maximizes the effects, thereby maximizing the cohesion gains.[62] We compare the optimal matching to the expected cohesion gains of a random matching and the infeasible scenario where every student has a high or low "max language match".

Table **??** shows the range of treatment effects under various scenarios. Compared to the randomized benchmark, optimal matching gives a 10 percent boost to cohesion gains (and nearly 50 percent better than the worst-case feasible match). This shows that educators with an eye toward integration could improve the impact of local cohesion-building initiatives by attending to student skill sets. While we use unobserved treatment effect heterogeneity here, the number of high language matches is the main difference between the best and worst matching. Within a cohort, the optimal pairing could be approximated by hand. On the other hand, the optimal feasible matching only provides one-third of the cohesion gains of the scenario where every student has a high match partner. This points to an additional benefit to longer-term investments in skills, an idea we return to in the conclusion.

10 Conclusion

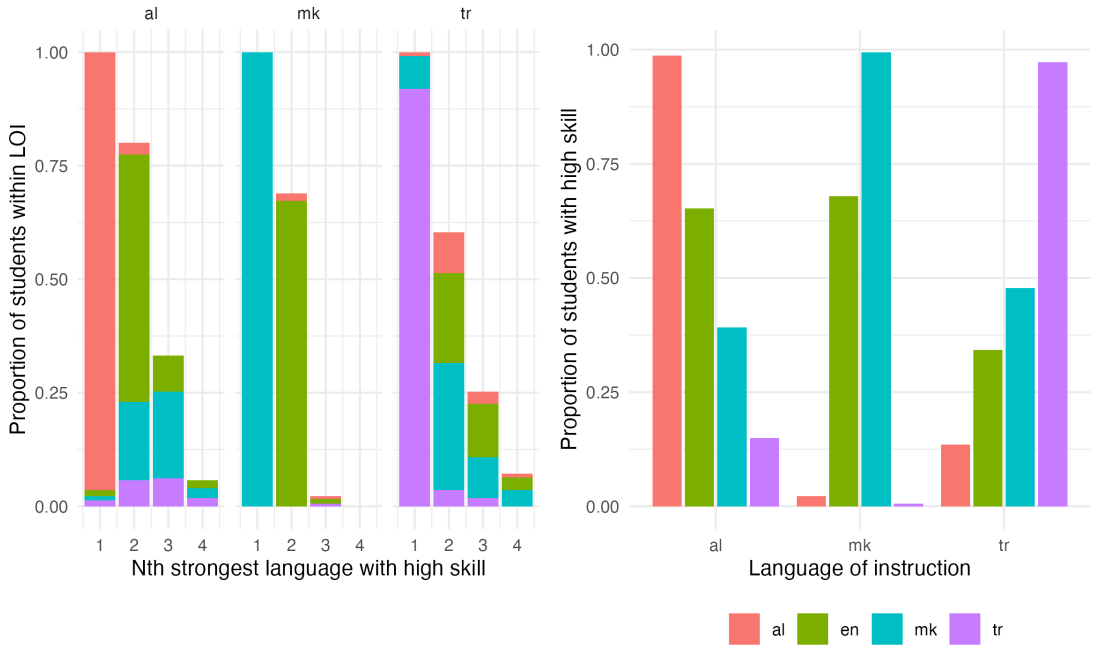
We investigate the role of language and identity in inter-ethnic contact using a lab-in-the-field experiment where students are randomly assigned the same or different ethnicity partners. In line with previous work, we find a moderate positive impact on cohesion outcomes after inter-ethnic contact. However, the main effects mask heterogeneity: contact has significantly stronger cohesion outcomes for students in pairs with high language-skill match. Focusing on these effects, exposure mapping suggests that language-skill match indeed has a causal effect. Still, some heterogeneity reflects "selection"–individuals with a higher probability of language-skill match also have better cohesion outcomes. While interethnic teams tend to have communication challenges, the high-language skill match groups often have similar overall experiences (enjoyment)

[62] In Appendix **??**, we describe the process in detail.

and productivity as co-ethnic pairs. Nevertheless, the average willingness to pay for segregation is still a large positive number. This highlights that cohesion cannot be built from short one-off experiences, without repeated exposure, homophily forces are too strong. The counterfactual exercise also suggests that increasing the possibility of language match could further improve cohesion outcomes.

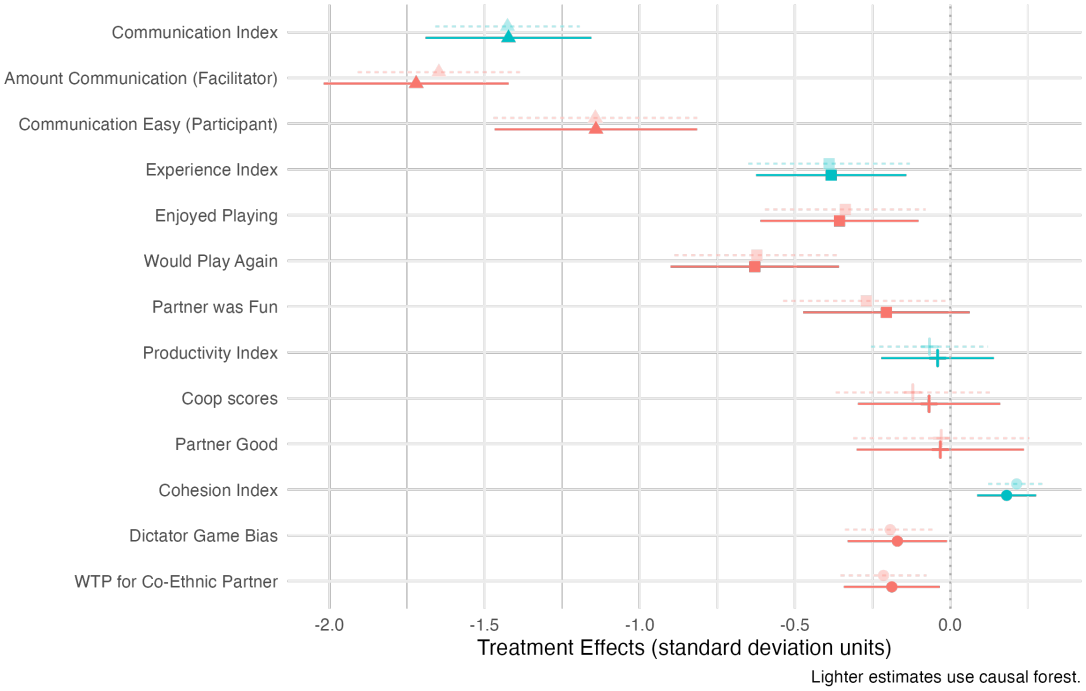
The results hint at a complementary link between language policy and cohesion, suggesting that policies promoting home language (L1) instruction should be paired with additional rigorous language training and, importantly, opportunities for interethnic contact. The causal arrow likely goes in both directions. Recent literature highlights the linguistic impact of integration. **rao˙famiarlity˙2019** find wealthier students substitute away from English when integrated with poorer students who tend to exclusively use Hindi at home. Alan et al. (2021) find Turkish skills increase for Syrian Refugees after participating in the perspective-taking course. Moreover, in Iraq, Christian soccer players start speaking Arabic more when Muslim soccer players are added to their team (Mousa, 2020). This insight may be useful in various settings where members of various language groups coexist, which is an increasingly common feature in communities around the world.

Figure 1: Language skills by language of instruction



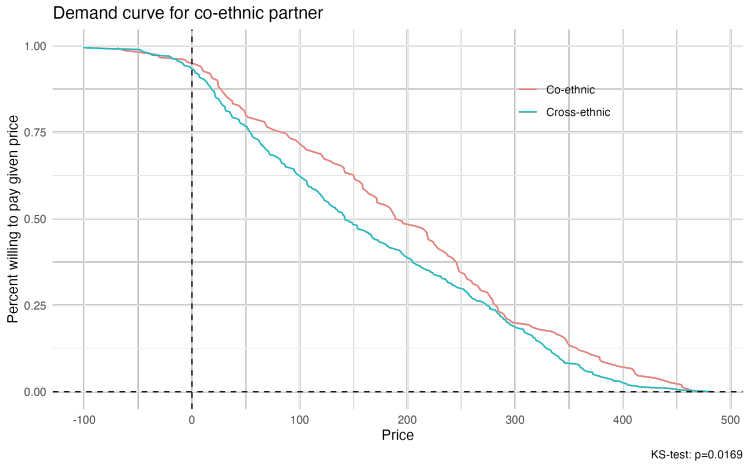
Notes: On the vertical axis both figures show the proportion of students within a language of instruction group who meet some criteria. The right plot shows self-reported language proficiency for the three local languages and English. Proficiency is defined as self-reporting an average of 7.5 fluency on a scale of 0 to 10 over the four language domains: reading, writing, speaking, and listening. The left plot shows how those languages rank within student skill sets. This plot shows that almost all students report the highest score for their language of instruction and that the majority of students know a second language.

Figure 2: Treatment effects: Assigned cross-ethnic partner

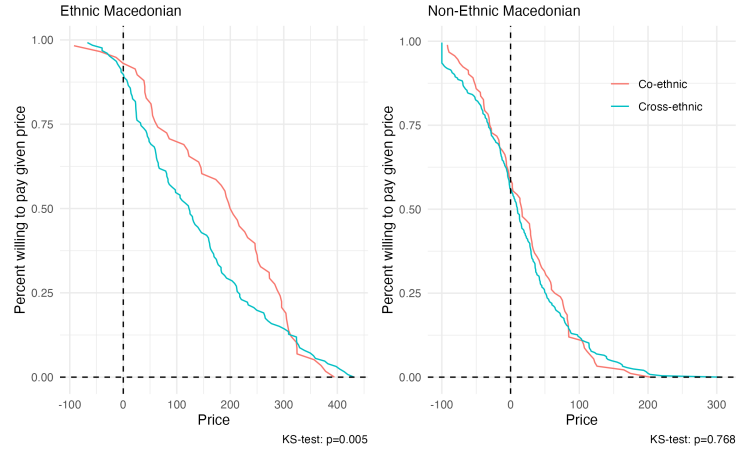


Notes: Coefficients on an indicator for any treatment in regression with no covariates. All sub-outcomes (red) are normalized around the control mean and control standard deviation. Indices (blue) are simple averages of normalized sub-outcomes. Standard errors clustered by strata.

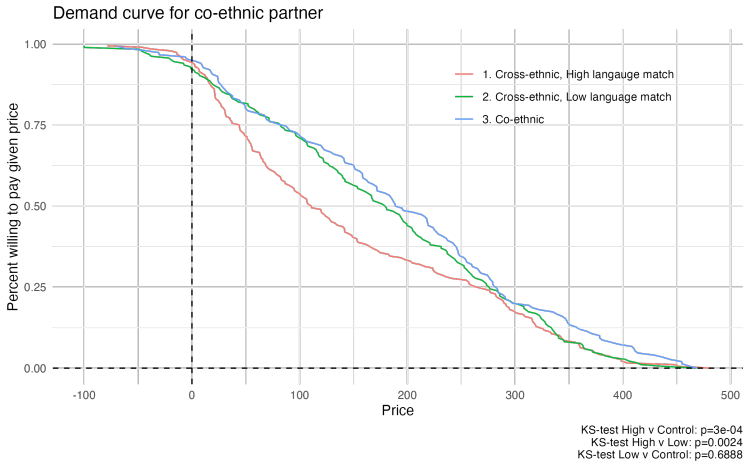
(a) Demand curve: Co-ethnic partner



(b) Demand curve: Macedonian as a second-language

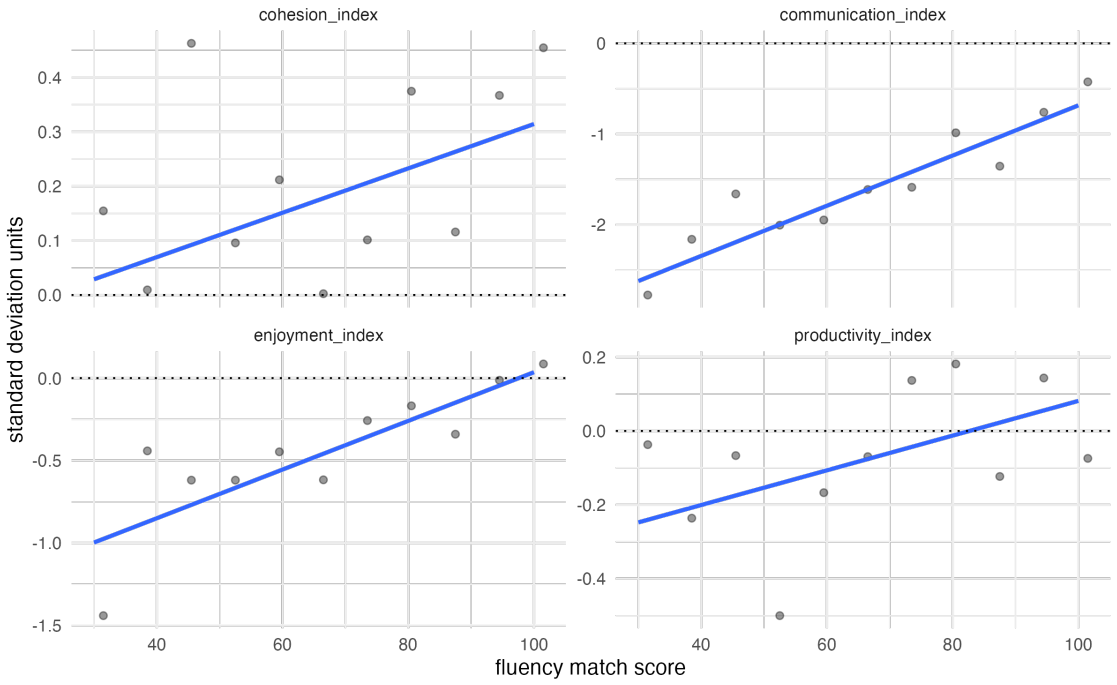


(c) Demand curve: Co-ethnic partner, by language match



Notes: Demand curves are the cumulative distribution of individual-level willingness to pay estimates. High (low) language skill match defined as above (below) median language match score.

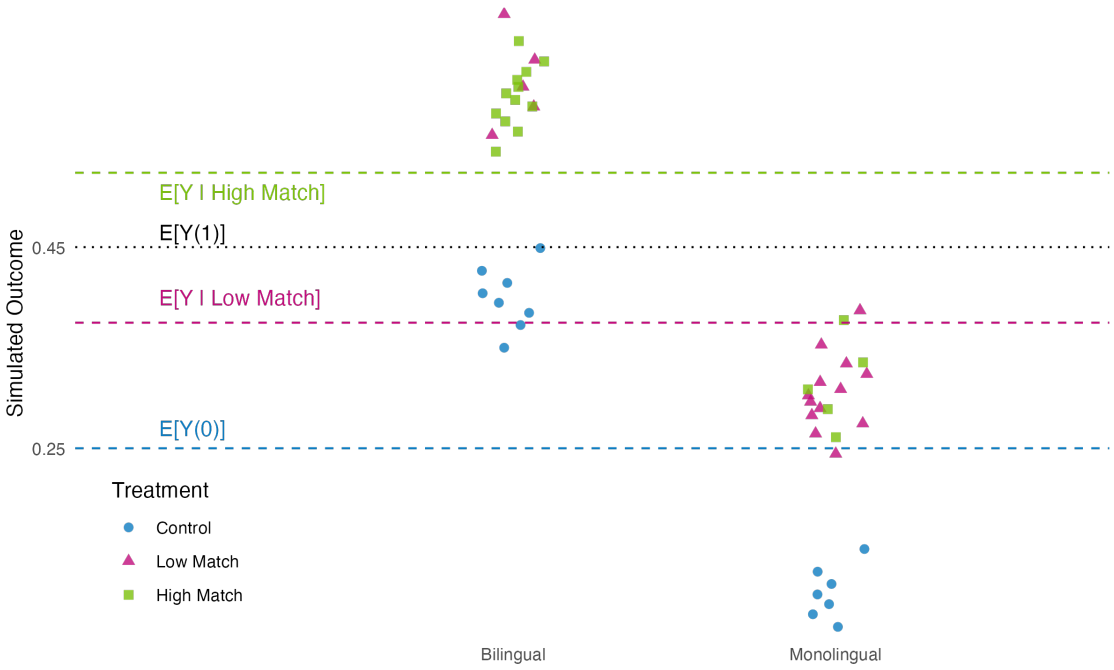
Figure 4: Associations between fluency match score and outcomes



Notes:

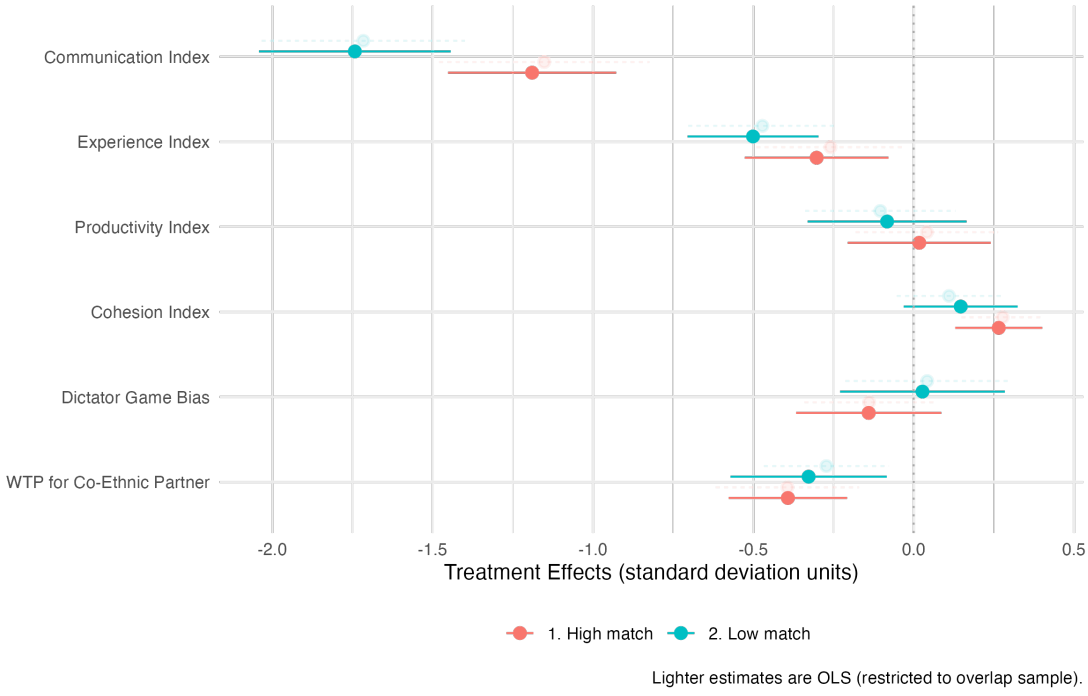
Lines show linear regression of indices against fluency match scores (defined above). The regression $\text{Index}_i = \beta_0 + \beta_1 \tilde{f}_{ij} + \epsilon_i$ are run against the normalized fluency matches. The estimated (β_0, β_1) are significantly different from zero for the Cohesion, Communication and Enjoyment indices, but not productivity. While partners are random, f_{ij} is a function of endogenous skills; Hence these results are interpreted as associations. Points show binned averages within roughly 6 point bins. Standard errors clustered by strata.

Figure 5: Simulation demonstrating selection bias when sample splitting



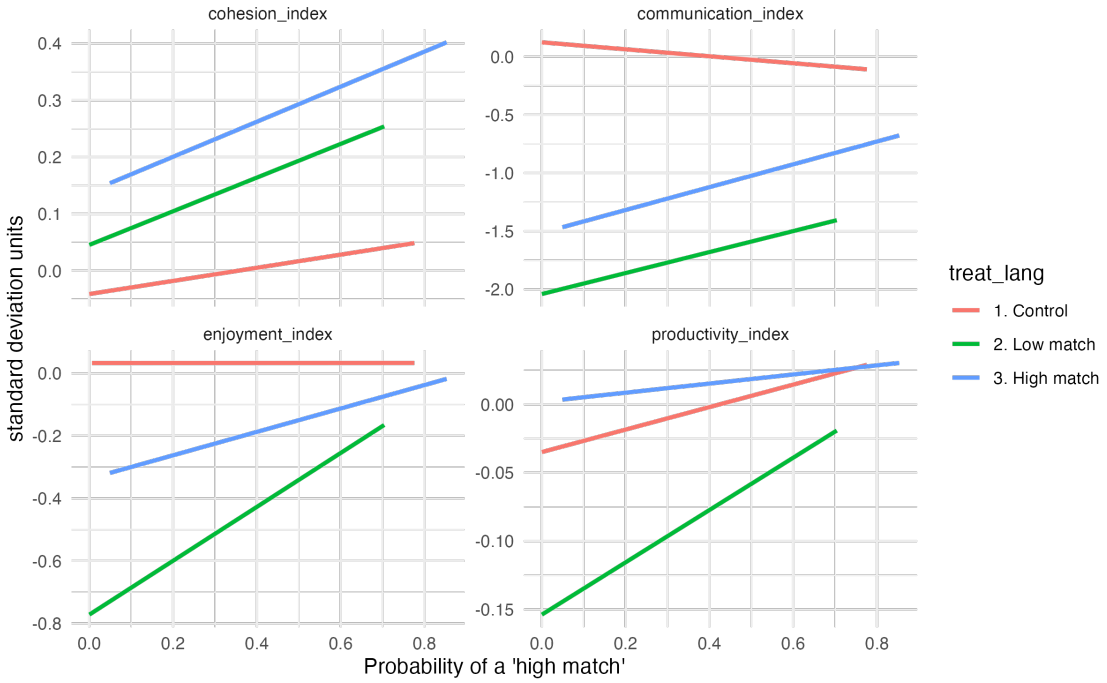
Notes: This stylized simulation shows a stratum with two types of students–bilingual or monolingual–in a trilingual environment. There is a constant treatment effect $\tau = .2$. Each point represents roughly 2000 students randomized into control, high match, or low match as shown with shapes and colors. The x-axis divides students by type and the y-axis shows the realized outcome. A regression run on the whole sample returns $\hat{\tau} = .2$, but the regression $y_i = \alpha + \beta^L \mathbf{1}_{lowmatch} + \beta^H \mathbf{1}_{highmatch} + \epsilon$ returns $\beta^H = \mathbb{E}[Y(1)|highmatch] - E[Y(0)] > \tau > \mathbb{E}[Y(1)|lowmatch] - E[Y(0)] = \beta^L$. Using inverse propensity weighting where $p = \mathbb{E}[highmatch|type]$ addresses the bias.

Figure 6: Language Heterogeneity: Causal Forests and OLS



Notes: Full color show causal forest ATE estimates that correct for differential propensity scores and chance imbalance. Lighter color show unweighted mean differences estimated on sample with overlap. High (low) language skill match defined as above (below) median language match score. All sub-outcomes are normalized around control mean and control standard deviation. Indices are simple averages of normalized sub-outcomes.

Figure 7: Relationship between propensity score and outcomes



Notes:
Lines show linear regression of indices against the probability of getting a high match partner $\pi(1,1)$. The regression $\text{Index}_i = \beta_{0D} + \beta_{0D}\tilde{\pi}_i(1,1) + \epsilon_i$ are run against the centered probabilities $\tilde{\pi}_i(1,1) = \pi_i(1,1) - \mathbb{E}[\pi_i(1,1)]$ and β_{0D}, β_{1D} are specific to treatment assignment.

Figure 8: How relationships with language impact outcome

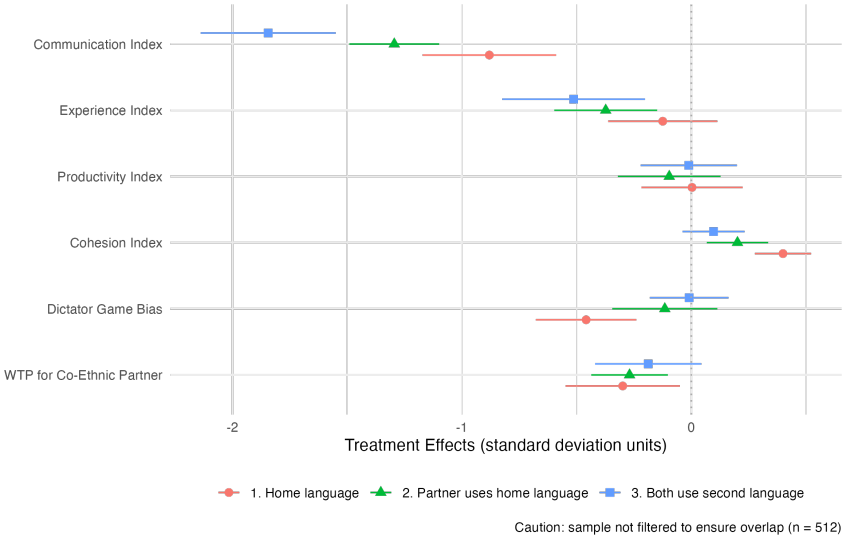
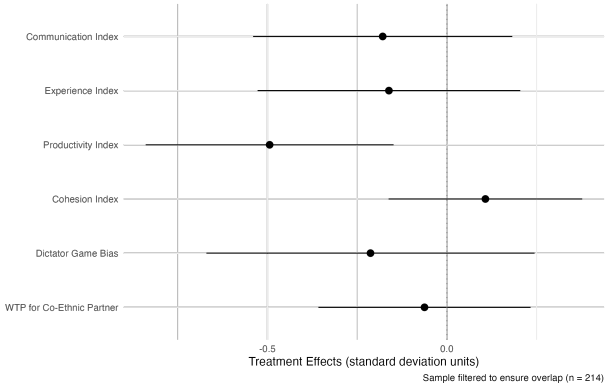


Figure 9: Macedonian versus English as second language



Notes The top plot shows treatment effects compared to the control group for students whose 1) top match is their home language (red circles), 2) top match is their partners home language (green triangles) or 3) top match is both students' second language (blue squares). The bottom plot shows Albanian and Turkish students outcomes when matching on Macedonian - English. The sample is filtered so all observations have positive probability of matching on either English or Macedonian.

Table 1: Demographics: Sample characteristics and balance

Covariate	Sample Mean	(sd)	Control Mean	(se).	Treat Diff	(se)	p-value	N
Female	0.56	0.50	0.57	0.11	-0.01	0.06	0.81	522
Grade 1	0.37	0.48	0.31	0.10	0.08	0.05	0.11	522
Grade 2	0.28	0.45	0.31	0.10	-0.05	0.06	0.38	522
Grade 3	0.18	0.39	0.21	0.08	-0.04	0.05	0.43	522
Grade 4	0.16	0.37	0.16	0.08	0.01	0.03	0.81	522
Albanian LOI	0.43	0.50	0.47	0.04	-0.04	0.04	0.32	522
Macedonian LOI	0.35	0.48	0.39	0.05	-0.05	0.04	0.28	522
Turkish LOI	0.21	0.41	0.15	0.04	0.09	0.03	0.01	522
LOI maps to ethnicity	0.96	0.18	0.95	0.02	0.02	0.02	0.33	513
English fluency	7.48	2.25	7.56	0.16	-0.11	0.18	0.56	522
Local language fluency (not LOI)	4.87	3.59	4.70	0.41	0.23	0.37	0.54	522
Attended Gostivar middle school	0.70	0.46	0.70	0.04	0.00	0.04	0.97	513
Has Cellphone	0.97	0.17	0.96	0.01	0.02	0.01	0.30	513
Parent has college	0.43	0.50	0.38	0.04	0.08	0.05	0.10	513
Plans on college	0.89	0.32	0.87	0.03	0.02	0.03	0.47	513
Life satisfaction (0-10)	8.46	1.98	8.23	0.26	0.32	0.20	0.12	513
Trust (0-10)	6.69	2.51	6.88	0.26	-0.26	0.25	0.29	513

Notes: LOI is the language of instruction. LOI maps to ethnicity equals 1 if the student self-identifies with the ethnicity associated with their language of instruction (the measure excludes students who identify with more than one local ethnicity). English fluency and Local language fluency for non-LOI are on a scale from 0-10 described in the text. Language skills are imputed for 9 students missing baseline data. Clustered by strata.

Table 2: Baseline cohesion and video game proclivities: Sample characteristics and balance

Covariate	Sample Mean	(sd)	Control Mean	(se).	Treat Diff	(se)	p-value	N
N peers in named for network	5.11	3.08	5.20	0.36	-0.13	0.26	0.62	513
N out-group links named	0.64	0.96	0.54	0.06	0.14	0.06	0.03	513
Closeness to own LOI (baseline)	6.44	1.26	6.43	0.13	0.02	0.14	0.90	513
Closeness to out LOI (baseline)	2.04	1.13	2.00	0.07	0.06	0.07	0.40	513
Dictator game denar to outgroup (baseline)	224.61	131.48	235.47	13.87	-15.27	8.89	0.10	513
Dictator game bias (baseline)	27.92	80.97	30.91	10.26	-4.20	9.79	0.67	513
Enjoys video games	0.50	0.50	0.54	0.06	-0.06	0.04	0.17	513
Makes video game rec	0.62	0.48	0.68	0.06	-0.07	0.04	0.09	513
Solo score	477.40	220.41	452.43	33.01	35.04	17.54	0.05	522
Solo score (demeaned)	0.34	211.79	-23.40	30.41	33.31	19.46	0.10	522

Notes: Network data comes from 4 questions where we ask students to elect other students as friends or sources of information and academic support. One question specifically asks for cross-ethnic links. Each question has 6 slots. 148 students use all the slots for friend clique. Closeness measures come from the Inclusion of Other-in-Self measure (scale from 1-7). Dictator game measures are as described in the text. Solo score is the average score in solo rounds of "Overcooked! 2". Solo score demeaned demeans the score by gender-specific solo score means. Clustered by strata.

Table 3: Communication and experience treatment effects: Assigned cross-ethnic partner

Dependent Variables:	Communication Index		Amount Communication (Facilitator)		Communication Easy (Participant)	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Constant	0.0075		5.96×10^{-15}		-4.33×10^{-15}	
	(0.0732)		(0.0984)		(0.0906)	
Cross-ethnic partner	-1.423***	-1.394***	-1.720***	-1.677***	-1.141***	-1.120***
	(0.1363)	(0.1591)	(0.1520)	(0.1750)	(0.1663)	(0.1853)
Fixed-effects						
school-grade-gender		Yes		Yes		Yes
Fit statistics						
Observations	505	505	524	524	505	505
R^2	0.21055	0.30706	0.24012	0.34495	0.09297	0.16826
Within R^2		0.20584		0.23730		0.08733

Clustered (strata) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Dependent Variables:	Experience Index		Enjoyed Playing		Would Play Again		Partner was Fun	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Constant	0.0328		-2×10^{-15}		-2.18×10^{-15}		-1.37×10^{-15}	
	(0.0876)		(0.0865)		(0.0954)		(0.1084)	
Cross-ethnic partner	-0.3833***	-0.3575***	-0.3567***	-0.3294**	-0.6296***	-0.5972***	-0.2055	-0.2126
	(0.1235)	(0.1299)	(0.1299)	(0.1377)	(0.1385)	(0.1495)	(0.1369)	(0.1491)
Fixed-effects								
school-grade-gender		Yes		Yes		Yes		Yes
Fit statistics								
Observations	492	492	512	512	504	504	508	508
R^2	0.02339	0.12662	0.01319	0.09745	0.03408	0.10960	0.00832	0.11192
Within R^2		0.02011		0.01098		0.02966		0.00889

Clustered (strata) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Indices are simple averages of normalized sub-outcomes. All sub-outcomes are normalized around control mean and control standard deviation. Communication easy and experience sub-outcomes derived from participant survey items on 5-point Likert scale. Amount communication is derived from the average of two facilitator responses on a 7-point Likert scale. Standard errors clustered by strata.

Table 4: Productivity and cohesion treatment effects: Assigned cross-ethnic partner

Dependent Variables:	Productivity Index			Coop scores			Partner Good		
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variables									
Constant	-0.0034			-1.66×10^{-16}			2.62×10^{-16}		
	(0.1010)			(0.1431)			(0.1203)		
Cross-ethnic partner	-0.0405	-0.0069	-0.0195	-0.0679	-0.0301	-0.0664	-0.0317	-0.0017	0.0074
	(0.0926)	(0.0995)	(0.0967)	(0.1170)	(0.1221)	(0.1134)	(0.1375)	(0.1536)	(0.1535)
Mean solo score (Normalized)			0.1203***			0.3248***			-0.0860
			(0.0361)			(0.0412)			(0.0563)
Fixed-effects									
school-grade-gender		Yes	Yes		Yes	Yes		Yes	Yes
Fit statistics									
Observations	508	508	508	524	524	524	508	508	508
R^2	0.00054	0.21449	0.22886	0.00104	0.33635	0.40713	0.00017	0.08467	0.08840
Within R^2		1.77×10^{-5}	0.01831		0.00027	0.10690		4.42×10^{-7}	0.00408

Clustered (strata) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Dependent Variables:	Cohesion Index			Dictator Game Bias			WTP for Co-Ethnic Partner		
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variables									
Constant	0.0022			-3.76×10^{-16}			-3.18×10^{-16}		
	(0.0490)			(0.0764)			(0.0754)		
Cross-ethnic partner	0.1819***	0.1436***	0.1337***	-0.1702**	-0.1444	-0.1303	-0.1880**	-0.1447*	-0.1370*
	(0.0485)	(0.0441)	(0.0462)	(0.0816)	(0.0873)	(0.0874)	(0.0788)	(0.0760)	(0.0783)
Baseline DG Bias			-0.1429**			0.2016*			0.0843*
			(0.0619)			(0.1184)			(0.0495)
Fixed-effects									
school-grade-gender		Yes	Yes		Yes	Yes		Yes	Yes
Fit statistics									
Observations	513	513	513	513	513	513	522	522	513
R^2	0.01163	0.08034	0.10843	0.00459	0.06706	0.09222	0.00748	0.06105	0.07261
Within R^2		0.00701	0.03733		0.00316	0.03004		0.00421	0.01036

Clustered (strata) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Indices are simple averages of normalized sub-outcomes. All sub-outcomes are normalized around control mean and control standard deviation. Coop scores derives from the average scores of partners in the incentivized rounds of play of "Overcooked! 2". Mean solo scores normalized is the average of 3 solo rounds. Partner good derives from participant reports in 5-point Likert scale. Dictator game bias derives from the difference between amount shared with same- and distinct-mother-tongue students. Baseline dictator game bias is the same measure taken at baseline. Willingness to pay for a co-ethnic partner derives from response to 12-questions in the adaptive choice experiment. Standard errors clustered by strata.

Table 5: Communication and experience heterogeneity by language match

Dependent Variables:	Communication Index		Amount Communication (Facilitator)		Communication Easy (Participant)	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Constant	0.0075		5.03×10^{-15}		-1.09×10^{-14}	
	(0.0733)		(0.0985)		(0.0907)	
Low match	-1.838***	-1.766***	-2.065***	-2.016***	-1.624***	-1.519***
	(0.1495)	(0.1612)	(0.1585)	(0.1744)	(0.1993)	(0.2098)
High match	-1.024***	-1.065***	-1.383***	-1.367***	-0.6770***	-0.7668***
	(0.1466)	(0.1669)	(0.1855)	(0.1979)	(0.1458)	(0.1718)
Fixed-effects						
school-grade-gender		Yes		Yes		Yes
Fit statistics						
Observations	505	505	524	524	505	505
R^2	0.26998	0.34383	0.27263	0.36954	0.14832	0.19732
Within R^2		0.24798		0.26593		0.11921

Clustered (strata) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Dependent Variables:	Experience Index		Enjoyed Playing		Would Play Again		Partner was Fun	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Constant	0.0328		-3.81×10^{-15}		-4.76×10^{-15}		-2.06×10^{-15}	
	(0.0877)		(0.0866)		(0.0955)		(0.1085)	
Low match	-0.6004***	-0.4975***	-0.5751***	-0.4668***	-0.9763***	-0.8663***	-0.3802**	-0.3396**
	(0.1456)	(0.1351)	(0.1520)	(0.1480)	(0.1708)	(0.1675)	(0.1624)	(0.1618)
High match	-0.1806	-0.2367*	-0.1465	-0.2072	-0.3017**	-0.3587**	-0.0374	-0.0990
	(0.1250)	(0.1357)	(0.1469)	(0.1581)	(0.1274)	(0.1425)	(0.1283)	(0.1437)
Fixed-effects								
school-grade-gender		Yes		Yes		Yes		Yes
Fit statistics								
Observations	492	492	512	512	504	504	508	508
R^2	0.04801	0.13458	0.02977	0.10253	0.06828	0.12580	0.02832	0.12016
Within R^2		0.02903		0.01655		0.04732		0.01808

Clustered (strata) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Indices are simple averages of normalized sub-outcomes. All sub-outcomes are normalized around control mean and control standard deviation. Communication easy and experience sub-outcomes derived from participant survey items on 5-point Likert scale. Amount communication is derived from the average of two facilitator responses on a 7-point Likert scale. Standard errors clustered by strata.

Table 6: Productivity and cohesion heterogeneity by language match

Dependent Variables:	Productivity Index			Coop scores			Partner Good		
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variables									
Constant	-0.0034			-1.89×10^{-16}			2.97×10^{-16}		
	(0.1011)			(0.1432)			(0.1204)		
Low match	-0.1056	-0.0691	-0.0754	-0.0138	-0.0306	-0.0534	-0.2084	-0.1215	-0.1166
	(0.1053)	(0.1141)	(0.1116)	(0.1272)	(0.1405)	(0.1298)	(0.1575)	(0.1676)	(0.1679)
High match	0.0221	0.0486	0.0309	-0.1209	-0.0296	-0.0784	0.1382	0.1054	0.1192
	(0.1124)	(0.1165)	(0.1118)	(0.1491)	(0.1496)	(0.1367)	(0.1436)	(0.1609)	(0.1614)
Mean solo score (Normalized)			0.1178***			0.3253***			-0.0914*
			(0.0352)			(0.0418)			(0.0531)
Fixed-effects									
school-grade-gender		Yes	Yes		Yes	Yes		Yes	Yes
Fit statistics									
Observations	508	508	508	524	524	524	508	508	508
R^2	0.00513	0.21776	0.23152	0.00325	0.33635	0.40724	0.01741	0.09084	0.09505
Within R^2		0.00418	0.02170		0.00027	0.10706		0.00674	0.01134

Clustered (strata) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Dependent Variables:	Cohesion Index			Dictator Game Bias			WTP for Co-Ethnic Partner		
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variables									
Constant	0.0022			-1.07×10^{-15}			1.22×10^{-16}		
	(0.0490)			(0.0765)			(0.0755)		
Low match	0.0884	0.0687	0.0521	-0.1100	-0.0761	-0.0527	-0.0741	-0.0714	-0.0514
	(0.0602)	(0.0538)	(0.0531)	(0.1036)	(0.1157)	(0.1065)	(0.0846)	(0.0736)	(0.0774)
High match	0.2728***	0.2121***	0.2079***	-0.2287**	-0.2068*	-0.2009*	-0.2995***	-0.2112**	-0.2150**
	(0.0515)	(0.0600)	(0.0625)	(0.0893)	(0.1051)	(0.1111)	(0.0966)	(0.0970)	(0.0975)
Baseline DG Bias			-0.1454**			0.2039*			0.0868*
			(0.0617)			(0.1183)			(0.0484)
Fixed-effects									
school-grade-gender		Yes	Yes		Yes	Yes		Yes	Yes
Fit statistics									
Observations	513	513	513	513	513	513	522	522	513
R^2	0.02200	0.08557	0.11458	0.00652	0.06902	0.09473	0.01682	0.06405	0.07667
Within R^2		0.01265	0.04398		0.00524	0.03272		0.00739	0.01470

Clustered (strata) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Indices are simple averages of normalized sub-outcomes. All sub-outcomes are normalized around control mean and control standard deviation. Coop scores derives from the average scores of partners in the incentivized rounds of play of "Overcooked! 2". Mean solo scores normalized is the average of 3 solo rounds. Partner good derives from participant reports in 5-point Likert scale. Dictator game bias derives from the difference between amount shared with same- and distinct-mother-tongue students. Baseline dictator game bias is the same measure taken at baseline. Willingness to pay for a co-ethnic partner derives from response to 12-questions in the adaptive choice experiment. Standard errors clustered by strata.

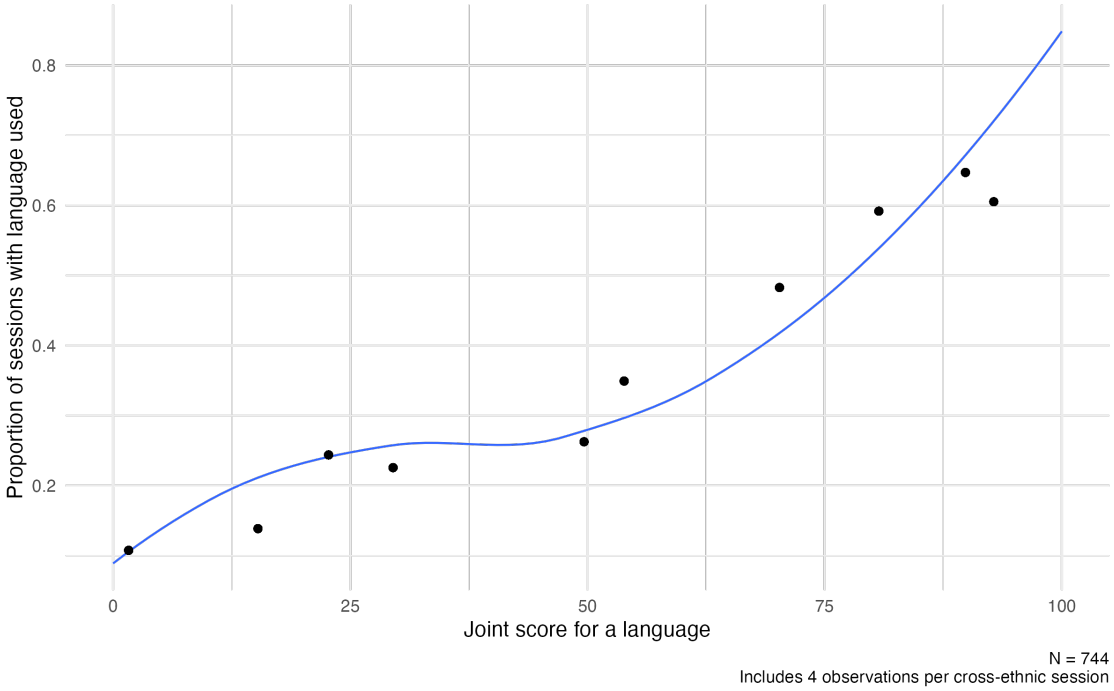
Table 7: Counterfactual matchings

All low match	Min matching	$E[\tau]$	Max matching	All high match
0.121	0.133	TBD: Assumes $E[\tau] = \hat{\tau} = .173$ for now	0.190	0.223

Notes: This table presents the range of counterfactual average treatment effects based on feasible and infeasible matchings. For each pair of cross-ethnic students within a strata, we assign a weight that is the sum of their estimated treatment effects (from the causal forest). If the pair have high mutual language skills we assign $w_{ij} = \beta_i^H + \beta_j^H$ and if they are a low-match we assign $w_{ij} = \beta_i^L + \beta_j^L$. The 'Max matching' is the matching that selects edges that maximize the weights and the table presents $\tilde{\tau} = \frac{1}{N} \sum_{(i,j) \in \text{Max matching}} w_{ij}$. Min matching is defined similarly. We approximate $E[\tau]$ from re-randomizations of the treatment assignment. For each randomization r, we calculate a $\tilde{\tau}_r$ and average these to get $E[\tau]$. Finally, we take the average of β_i^H for "All high match" and the average of β_i^L for "All low matches".

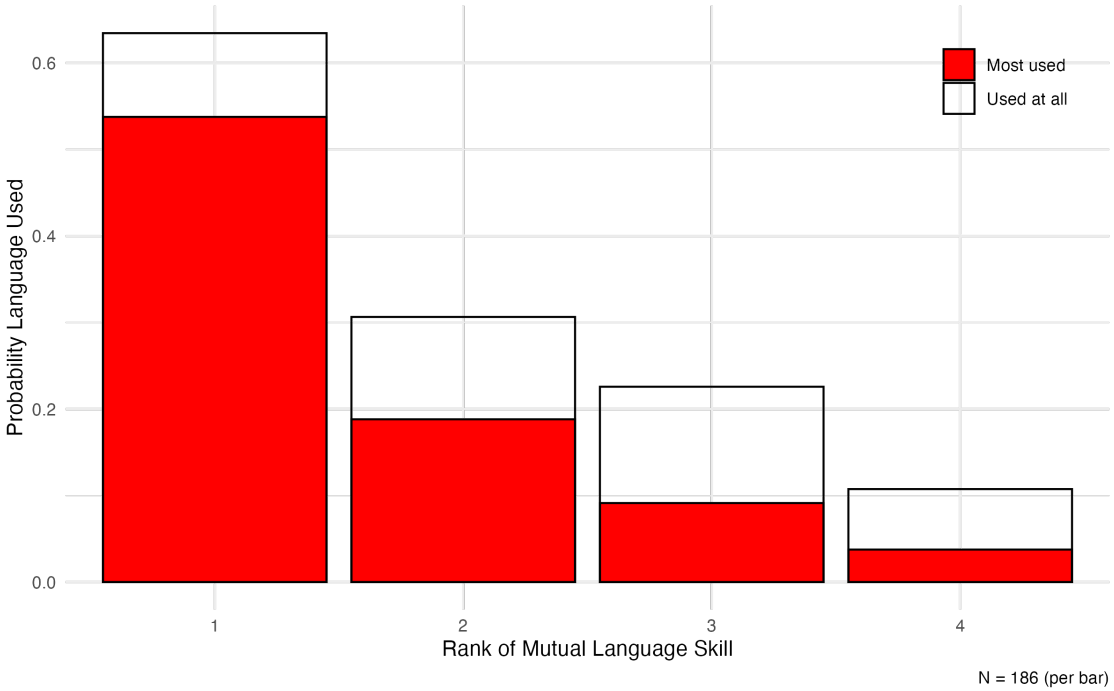
A Appendix

Figure A1: Language usage by joint score



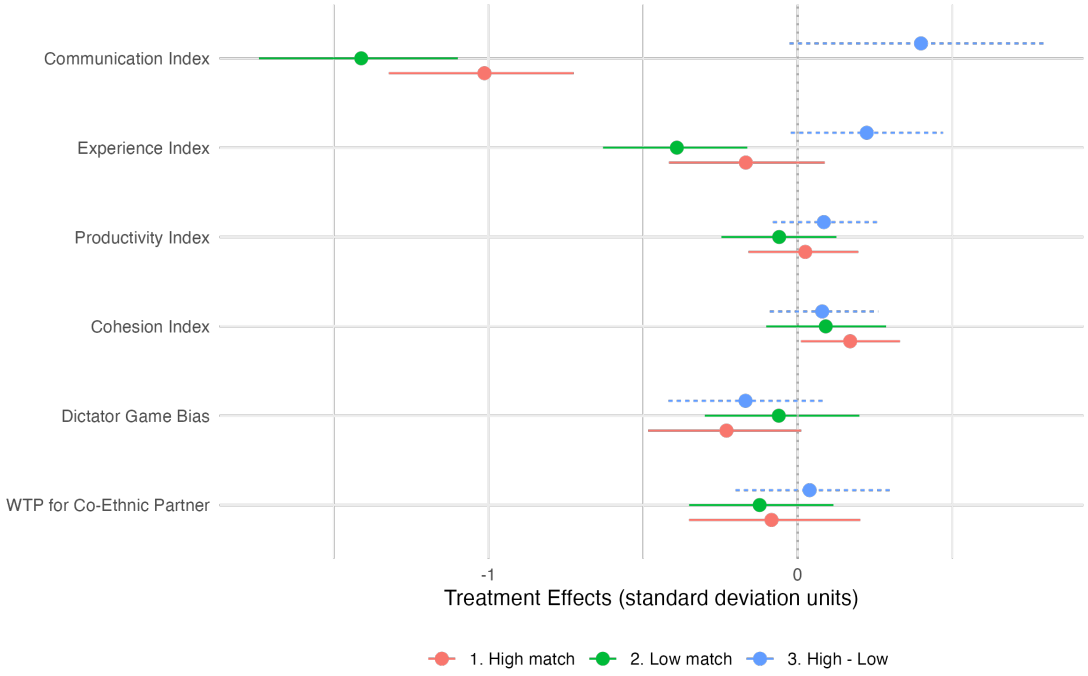
Notes: The plot shows the positive relationship between joint scores and language usage in the cross-ethnic sessions. Joint scores are the product of partners fluency measure for the language; each pair has a score for Albanian, English, Macedonian and Turkish.

Figure A2: Language usage based on the rank of mutual language skills



Notes: With in a pair, we rank mutual language skills based on joint scores (x-axis). The y-axis shows the proportion of sessions in which a language is used (black outlines) or most used (red).

Figure A3: Treatment effects disaggregated by language match (IPW)



Notes: Inverse-probability weighted mean contrasts. Weights determined by simulating the treatment assignment mechanism. High (low) language skill match defined as above (below) median language match score. All sub-outcomes are normalized around control mean and control standard deviation. Indices are simple averages of normalized sub-outcomes. Bootstrapped standard errors.

Table A1: Full Sample Balance: Demographics

Covariate	Sample Mean	(sd)	Control Mean	(se).	Treat Diff	(se)	p-value	N
Female	0.53	0.50	0.54	0.10	-0.01	0.03	0.67	822
Grade 1	0.32	0.47	0.30	0.09	0.03	0.03	0.29	822
Grade 2	0.33	0.47	0.36	0.10	-0.04	0.03	0.23	822
Grade 3	0.22	0.42	0.22	0.08	0.00	0.02	0.99	822
Grade 4	0.13	0.34	0.13	0.07	0.01	0.02	0.81	822
Albanian LOI	0.44	0.50	0.48	0.03	-0.05	0.02	0.05	822
Macedonian LOI	0.33	0.47	0.32	0.03	0.02	0.02	0.38	822
Turkish LOI	0.22	0.42	0.20	0.03	0.03	0.02	0.19	822
LOI maps to ethnicity	0.94	0.23	0.95	0.01	-0.01	0.02	0.52	754
English fluency	7.36	2.31	7.43	0.13	-0.09	0.15	0.53	763
Local language fluency (not LOI)	4.85	3.61	5.00	0.31	-0.19	0.24	0.43	763
Attended Gostivar middle school	0.67	0.47	0.66	0.04	0.01	0.04	0.74	754
Has Cellphone	0.95	0.21	0.94	0.02	0.01	0.01	0.25	754
Parent has college	0.42	0.49	0.38	0.04	0.06	0.04	0.15	754
Plans on college	0.87	0.34	0.88	0.02	-0.01	0.02	0.58	754
Life satisfaction (0-10)	8.43	2.06	8.31	0.21	0.16	0.16	0.34	754
Trust (0-10)	6.65	2.52	6.82	0.23	-0.23	0.20	0.27	754

Notes: LOI is the language of instruction. LOI maps to ethnicity equals 1 if the student self-identifies with the ethnicity associated with their language of instruction (the measure excludes students who identify with more than one local ethnicity). English fluency and Local language fluency for non-LOI are on a scale from 0-10 described in the text. Language skills are imputed for 9 students missing baseline data. Clustered by strata.

Covariate	Sample Mean	(sd)	Control Mean	(se).	Treat Diff	(se)	p-value	N
N peers in named for network	4.95	2.95	5.01	0.31	-0.08	0.20	0.69	754
N out-group links named	0.61	0.90	0.62	0.06	-0.01	0.06	0.81	754
Closeness to own LOI (baseline)	6.39	1.36	6.39	0.11	-0.01	0.11	0.95	754
Closeness to out LOI (baseline)	2.07	1.18	2.06	0.07	0.01	0.07	0.89	754
Dictator game denar to outgroup (baseline)	218.14	130.63	227.32	12.86	-12.36	7.98	0.13	754
Dictator game bias (baseline)	29.38	82.36	31.96	8.44	-3.48	7.57	0.65	754
Enjoys video games	0.51	0.50	0.53	0.05	-0.03	0.03	0.39	754
Makes video game rec	0.65	0.48	0.67	0.06	-0.03	0.03	0.40	754

Table A2: Full Sample Balance: Other baseline characteristics

Notes: Network data comes from 4 questions where we ask students to elect other students as friends or sources of information and academic support. One question specifically asks for cross-ethnic links. Each question has 6 slots. 148 students use all the slots for friend clique. Closeness measures come from the Inclusion of Other-in-Self measure (scale from 1-7). Dictator game measures are as described in the text. Solo score is the average score in solo rounds of "Overcooked! 2". Solo score demeaned demeans the score by gender-specific solo score means. Clustered by strata.

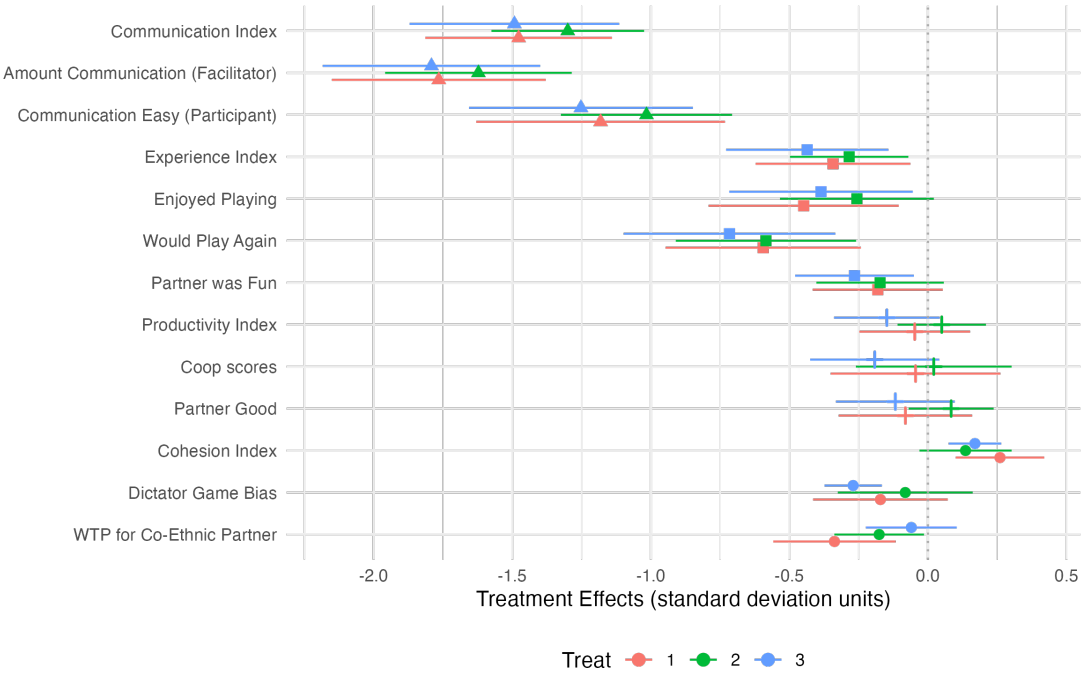


Figure A4: Treatment effects disaggregated by identity prime

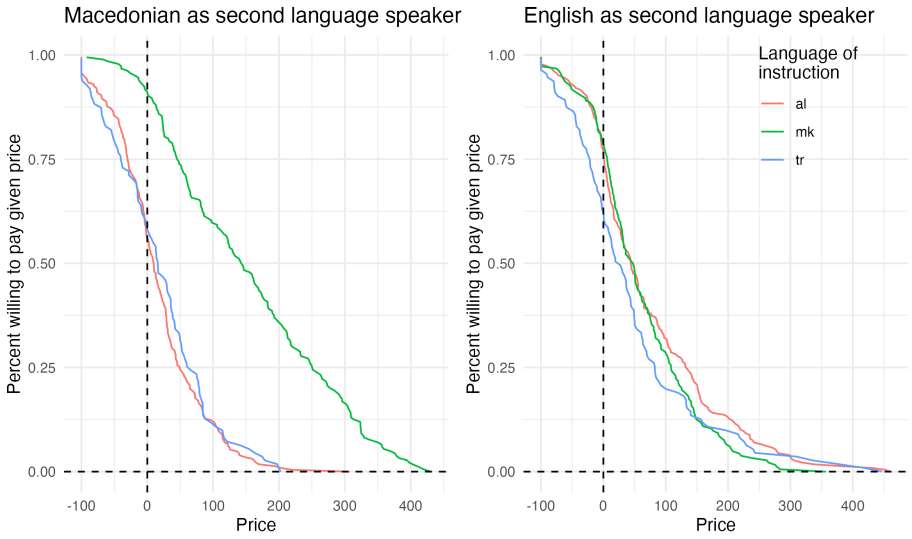


Figure A5: Caption

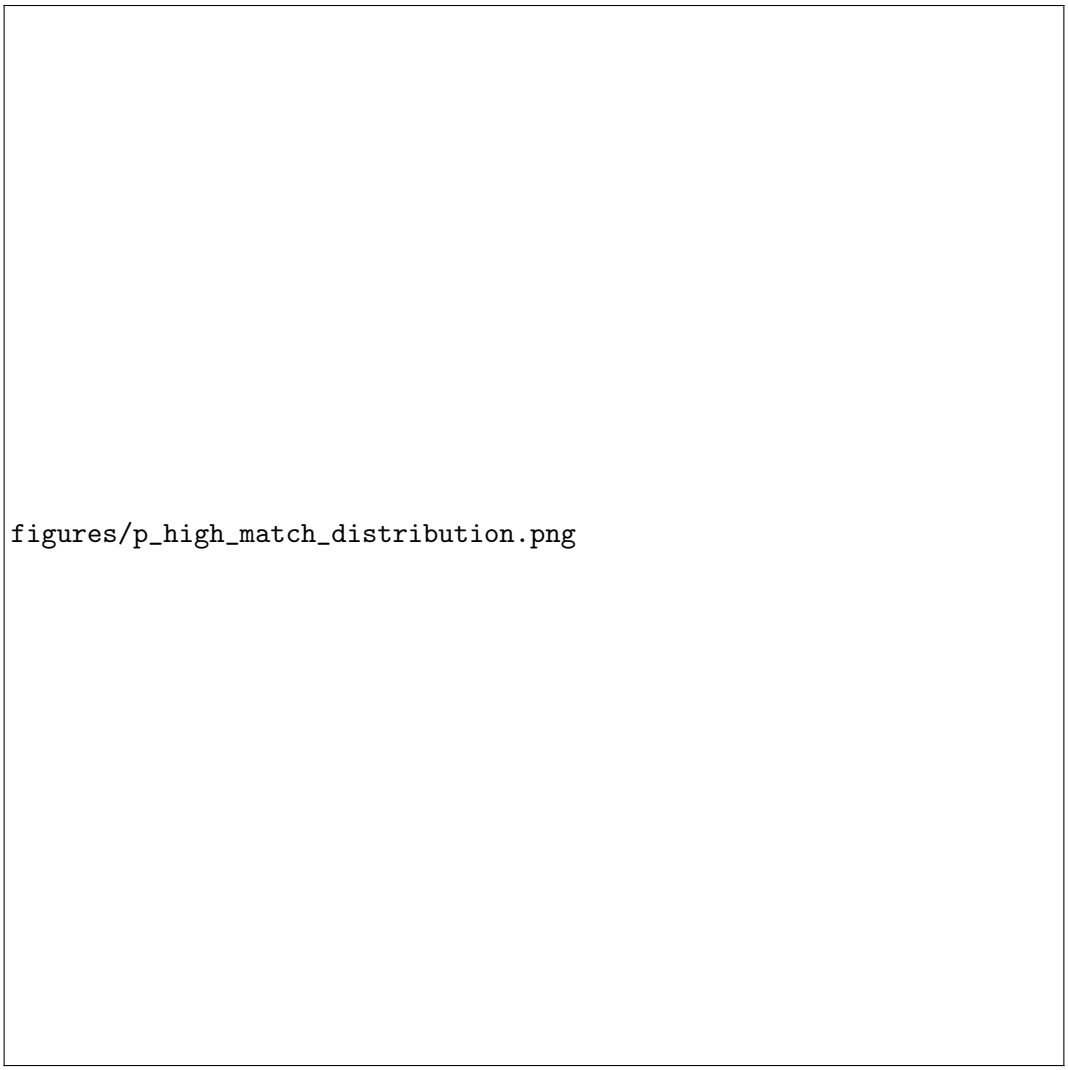
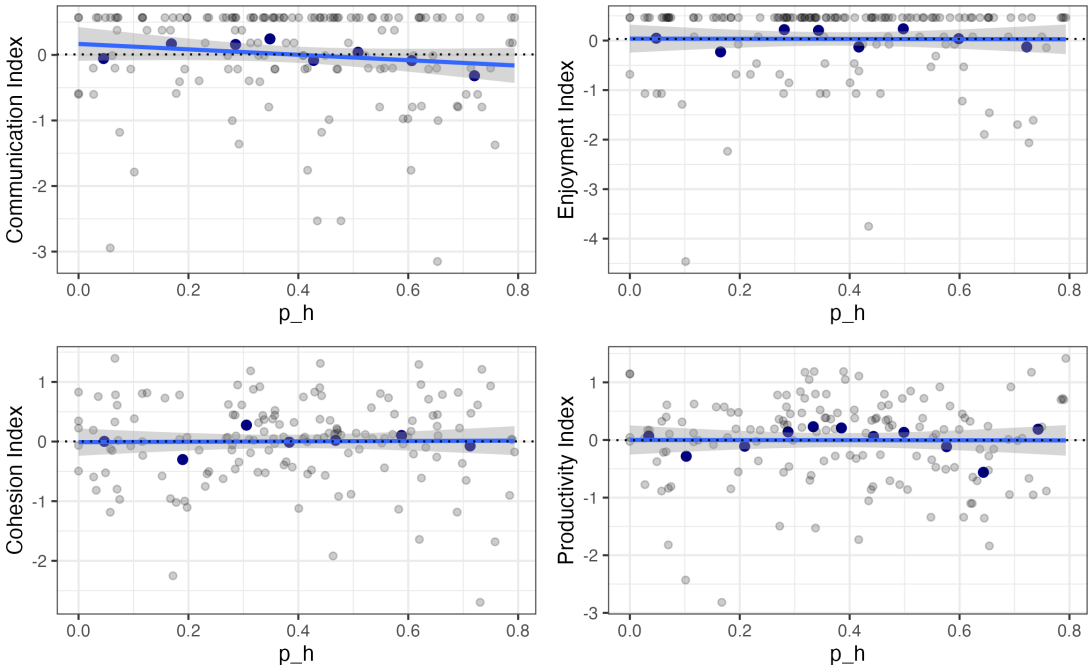


Figure A6: The probability distribution of a high language match conditional on being assigned a cross-ethnic partner. In the notation of exposure mappings this plot shows $\frac{\pi(1,1)}{\pi(1,1)+\pi(1,0)}$.

Figure A7: $Y(0)$ against $\pi(1,1)$



Notes: Binscatter shows local means and faint dots show raw data. The dotted line shows $\mathbb{E}[Y(0)]$ which is roughly 0 by design. The thick line shows the linear model that fits $y(0)$ on $\pi(1,1)$.

B Procedures

B.1 Randomization

Randomization into pairs that depend on baseline characteristics adds complexity. We face the following goals and constraints. First, we stratify by school, grade level, and gender; keeping gender and grade constant within pairs helps isolate identity and language's role. Moreover, we hope to split treatment-control groups at certain proportions to maintain power.[63] Finally, due to tight timelines, we randomized within strata on a rolling basis to provide steady work for the facilitation teams.

The first problem arises when some ethnic group is a supermajority meaning that the number of recruits from one ethnicity is larger than the other two combined. This occurs in 9 of 23 grade x school x gender strata.[64] To see why, consider an exaggerated case with 96 participants from E1 and 8 from E2. If one were to assign individuals to co- or cross-ethnic pairs at random with $\frac{3}{4}$ probability, there would be 72 E1's assigned cross-ethnic partners, but only 6 E2's to match them with. We can thus make 6 treatment pairs. Meanwhile, we create 13 control pairs (12 from E1 and 1 from E2). So, naively randomizing in this way, the realized probability of assignment to control is $\frac{13}{19} = 68\%$. This problem arises because we do not match students "assigned" to treatment in the largest ethnic group within the strata.

To overcome this challenge, we designed an algorithm that ensures participants have an equal *a priori* probability of assignment to *control* condition regardless of the composition of the strata. To address the issue mentioned above, we *adjust the control assignment probability* to avoid "over-assignment" to control. An additional nuisance is that strata may not be evenly divisible into pairs. To address this we use *probabilistic rounding to round to the nearest even number* while maintaining the target assignment probabilities.

Problem 1: Adjusting the assignment probabilities

As seen in the example above, we need a solution to randomization when we have a super-majority so that the control group size is roughly proportional to the target assignment probability p. We could either (1) *a*

[63] Moreover, maintaining equal assignment probabilities in expectation within strata helps reduce biases in estimation CITE AZEEM.

[64] This may reflect the underlying cohort populations, recruitment effort of the field team, truancy, or the preferences of students. Reflecting the underlying population, these tend to be Albanian supermajorities. Macedonians or Turks represent a plurality of students in 6 of the strata.

priori randomly drop students from the largest ethnic group or (2) adjust the initial assignment probability to $\tilde{p} \leq p$, where we select $\tilde{p}$ so that in expectation the control group size is proportional to p.

We consider both approaches and their implications using the running example. Under approach (1), we drop 96 - 8 students from ethnicity E1. Now we have 8 E1 and 8 E2. We then assign 4 to control and 12 to treatment. Under approach (2) we assign control units with probability $\tilde{p} \simeq .05$. Then 4.8 E1 units and .4 E2 units are assigned to control. We use probabilistic rounding (described below) to deal with non-even numbers, which provides a range of scenarios, with realized $p_{control}$ as one of $4\frac{}{20=20\%, \frac{6}{22}=27\%, \frac{6}{18}=33\% and \frac{8}{20}=40\%.}$

We use the second approach. While the solution is more complicated it allows us to maximize the sample for a given level of recruitment effort. In each scenario, we use more than 16 recruits. Both solutions provide expected assignment probability of roughly $\frac{1}{4}$.

To determine $\tilde{p}$, the following equation must hold

$$\tilde{p}N = p * (N - N_{dropped})$$

Where $N = |E1| + |E2| + |E3|$ is the total sample within strata with the $E1$ the largest group, p is the desired assignment probability and $N_{dropped}$ is the number of individuals left unassigned.

1. If $|E1| \leq |E2| + |E3|$ and N is even, we assign all units ($N_{dropped} = 0$).
2. If $|E1| \leq |E2| + |E3|$ and N is odd, we assign all but one unit ($N_{dropped} = 1$).
3. When $|E1| > |E2| + |E3|$, let $L = [|E1| - (|E2| + |E3|)]$ and $N_{dropped} = (1 - \tilde{p})L$

Notice that when $|E1| \geq |E2| + |E3|$, all cross-ethnic pairs have one member of E1 (i.e. there are no E2-E3 pairs).

Then we define $\tilde{p}$ for each of the scenarios as follows:

1. $\tilde{p} = p$

2. $\tilde{p}N = p(N-1) \implies \tilde{p} = p\frac{N-1}{N}$

3. $\tilde{p}N = p(N-(1-\tilde{p})L) \implies \tilde{p} = p * \frac{N-L}{N-p*L}$

Probabilistic rounding We use probabilistic rounding to address strata whose control units do not divide evenly into pairs. Say the strata of size N and target probability is p. If pN is not divisible by 2, we round up with a probability that is proportional to the remainder of pN divided by 2. There is an even number n such that $pN \in [n, n+2]$. Then, q is the probability of rounding up that solves the equation $(n+2)q+n(1-q) = p*N$. We round up with probability $q = \frac{pN-n}{2}$. Say $N = 13$. Then, we round up with probability $q = .25*(3.25-2)/2 = .625$ and down with probability $(1-q) = .375$.

For purposes of coding, we use $q = p*r/2$, where $r = mod(N, \frac{2}{p})$ is the number of remaining people in a block after the first $\frac{2}{p}$ students have been assigned to pairs. This avoids the need to determine n.

Returning to our running example, we round up E1 with probability $q_1 \simeq .8/2$ and we round up E2 with probability $q_2 \simeq .4/2$, so that the expected control probability is

$$\frac{4}{20}*(1-q_1)*(1-q_2) + \frac{6}{22}*q_1*(1-q_2) + \frac{6}{18}*(1-q_1)*q_2 + \frac{8}{20}*q_1*q_2 \simeq .25$$

Assigning cross-ethnic pairs

After making control assignments, we assign the remaining students to treatment. To do so we arrange the list first by the size of the ethnic group in the strata and then by a random number. If we have an even number of students and the triangle inequality holds ($|E1| \leq |E2|+|E3|$), we split the list in half and match partners by their location in the list.[65] When $|E1| \geq |E2|+|E3|$, we drop the first $|E1|-(|E2|+|E3|)$ students from E1 and match the remaining students from $E1$ with the students from the other ethnicities, matching by list order. Similarly, we drop one student from $E1$ if the total treatment population is odd but the triangle inequality holds. Ordering by ethnic group size first ensures the matches are cross-ethnic. We then randomly assign the pair a specific treatment assignment. The assignments are then input into Qualtrics.

[65] i.e. The first person in the top half of the list matches with the first person in the bottom half of the list, the second with the second, and so on.

Assessment of randomization procedure On average, the procedure achieved its goal: The realized control assignment probability is 25.3 percent. Moreover, in 5000 simulations of the estimation procedure, conditional on any assignment there's a 24.8 percent probability of being assigned control on average across individuals. (There is a 5.5 percent probability that an individual is not assigned treatment or control due to oversubscription of some ethnic group within some strata or lack of parity.) Moreover, within strata, we simulate control probability conditional on any assignment and find all but 4 strata with expected control probability within 0.5 percent of the target (min = 22.5%, max = 25.2%).

Individuals have control probabilities that are not equal to 25 percent by design and also due to logistical constraints in the field. Recall that $\tilde{p}$ is the individual level treatment probability and so within imbalanced strata, individuals will face treatment probabilities less than p. Further, for 12 of 23 strata, we ran randomization in waves–in follow-up waves, we included E1 students who were passed over to maximize our sample size while staying within budget. Such students have a slightly higher expected probability of control assignment. When using inverse propensity weighting we account for these discrepancies. Finally, 6 individuals were the sole representatives of their ethnicity within a strata during a randomization wave. Dropping these individuals does not affect the results.

B.2 Data collection

Participation consisted of several activities. First, enumerators from our survey firm conducted baseline surveys, collecting data about language skills, social networks, behavioral measures, and questions about socio-economic status and well-being. For the most part, once both partners in a pair completed the baseline survey, we sent the session information to the facilitation team who organized the video game sessions described in this paper. In addition to learning and playing "Overcooked! 2", students answered questions about their experience in the session, completed the choice experiment, provided additional behavioral measures, and decided which pin to take as a reward either one reflecting the common identity (town or school insignia) or one representing the students' ethnic identities reflecting their mother tongue alphabet.[66] Finally, students participate in a midline survey where we ask students to sign up for the interethnic course and re-ask social network and behavioral measure questions from the baseline. These three steps constitute the major

[66] Initially we thought to provide pins with the flags of Albania, Macedonia or Turkey, which are commonly used to represent ethnicity in Macedonia. Our field partners advised us against this, as the flags are divisive.

data sources for this paper. Many of the same participants were invited to participate in our companion study, wherein students in groups of 6 participated in a social learning game, where students pass information via a website we built for the study. The second study involves very limited cross-group interaction as students are instructed to remain quiet. Nevertheless, we strive to have students complete the first three steps before participating in Study 2.

C Appendix: Bayesian Adaptive Choice Experiment: Theory and Implementation

Here we present the theory behind the Bayesian adaptive choice experiment (BACE) and our implementation thereof. We implement BACE using software developed by Drake et al. (2022) and our discussion of theory relies on the accompanying paper (Drake et al. (2024); Hereafter DPTT).

The goal is to estimate preference parameters $\theta_i \in \mathbb{R}^k$ for individual $i \in N$ over k attribute classes. In our case, we measure willingness to pay for partner ethnicity and second language skills. We estimate θ in a Bayesian framework retrieving posterior $Pr(\theta_i|x_i^{(1:T)}, d_i^{(1:T)})$ where x is histories of choices (option A or option B) and d are the choice scenarios that the agent faced. (*e.g.* Option A provides 50 more Denar than Option B but unlike Option B, Option A does not speak English and otherwise they are the same.)

DPTT provides a method to efficiently search the parameter space. The researcher's problem is to choose d_{t+1}^*, the optimal choice scenario to show next, given a prior $Pr(\theta_i|x_i^{(1:t)}, d_i^{(1:t)})$. d_{t+1}^* is best in that it maximizes the mutual information gained (or, alternatively, minimizes entropy). The solution provides a multidimensional analogy to binary search in that d_{t+1}^* splits the current prior into two "half-spaces" in which the student is equally likely to choose either option in the menu.

$$d_{t+1}^* = \arg\max_{d_{t+1}} \int_\theta \sum_{x^{t+1}\in\{0,1\}} \left[\log \frac{Pr(\theta|x^{(1:t+1)}, d_{t+1})}{Pr(\theta|x^{(1:t)})}\right] Pr(x^{t+1}|\theta, d_{t+1}) Pr(\theta|x^{(1:t)}) d\theta$$

To help focus on the current design choice d_{t+1}, we suppress reference to the design history $d^{(1:t)}$ on which all the $Pr(.|.)$ objects depend. We also suppress reference to the decision maker i though each individual has their own parameters and choice scenarios.

The learning process follows the following *Algorithm* 1) sample θ from the current prior distribution, 2) numerically integrate over θs over a range of possible designs to determine d_{t+1}^*, 3) provide choice scenario d_{t+1}^* to the student to learn x^{t+1} and 4) update the prior. (ie. the posterior is the prior for the next iteration.)

To sample from the distribution, we use population Monte Carlo (PMC). Section 9.2 of DPTT (2024) provide

a thorough explanation of the method; we highlight a few key points. At a high-level, PMC provides a sample "approximately" from the target distribution and "approximately unbiased" estimates of an integral under the distributionCappé et al., 2004. It does so by drawing thousands of points from the proposal distribution and then "pruning" them using importance sampling weights. This is fast enough to compute on a server and send the next choice to the decision maker.

C.1 Implementation

Students derive expected utility from a reward r and a partner's ethnic and linguistic attributes captured in z. When running BACE for data collection, we include specific ethnicity markers (i.e. Turkish, Albanian, Macedonian); While we collapse these to own ethnicity or not for analytic purposes.:

$$u(z) = 1 * r + \theta' z + \epsilon$$

z includes an indicator for whether or not the potential partner is from the same ethnicity and whether they speak English, Macedonian, or neither as a second language. We normalize the coefficient on the reward to 1, which puts θ into Denar units. Students are told to assume both options are students in their cohort, who have the same gender and are otherwise identical. ϵ is a Type-1 extreme value shock term, which we interpret as the decision makers' uncertainty and attention

Students chooses between A and B, choosing A if

$$1 * r_A + \theta' z_A + \epsilon_A > 1 * r_B + \theta' z_B + \epsilon_B$$

and B otherwise. Equivalently, they choose A if

$$(r_A - r_B) + \theta'(z_A - z_B) > -(\epsilon_A - \epsilon_B)$$

Likelihood: This corresponds to the logit likelihood:

$$Pr(x^{t+1} = 1|\theta) = \frac{1}{1 + \exp(-\mu(\Delta r + \theta' \Delta z))}$$

we replace differences with Δ, and introduce μ, a dispersion parameter that we estimate that captures attention or indifference. As μ gets large, it magnifies small differences in utility so that the probability the student picks the option with higher utility goes to 1. We estimate μ is smaller if students make more "mistakes".

Prior: The prior for each z is a normal distribution centered at 0. We set the variance to 100 Denar for ethnic matches and 200 Denar for language skills. This prior is somewhat conservative as we center the prior estimates around 0. The prior over the dispersion parameter μ is uniform between 0.01 and 0.09.

Choice scenarios/Designs: The set of choices includes 1275 distinct designs, or vectors $(\Delta r, \Delta z)$. The reward Δr takes one of 25 values between -100 to 500 Denar (roughly -\$2 to \$10) with the options taking values divisible by 25 Denar (about 50 cents USD). There are dependencies between ethnic and linguistic factors in the design which lead to 51 distinct designs. Δz takes a value in $-1, 0, 1$ reflecting differences in indicator variables. An exemplary $\Delta z = (-1, 0, 1, 0, 1)$ with the z_1, z_2 and z_3 corresponding to ethnicities (Albanian, Macedonian and Turkish), z_4 corresponding to English ability and z_5 corresponding to Macedonian ability. Then, this example corresponds to option A being Turkish and a second-language Macedonian speaker. Option B is Albanian and we randomly decide if they both speak English or not.[67]

We learn individuals preferences using the algorithm discussed above. At each iteration, we take the optimal design and translate it into a choice as described above. They are told that to imagine the students are identical in other respects and equally good at playing "Overcooked! 2". Students face 12 choice scenarios using the decision problem discussed above to select the next menu of options. We chose 12 to balance the need to collecting sufficient data for estimation with the attention span of the students. Simulations show a large gain in terms of mean square error over the first 10 choices after which the learning curve flattens out.

[67] Why are there 51 options? For each ethnic pairing, an ethnicity can be shown in option A or option B (1 or -1). Thus, there are four sets of ethnic pairings with Macedonians. Each of these pairings have 2 MK as second language options (the Macedonian cannot be a second language speaker) and 3 ESL options $4 \cdot 2 \cdot 3 = 24$ such choice scenarios and 3 sets that (potentially) exclude Macedonians (including the option where ethnicity is the same for either option) $3 \cdot 3 \cdot 3 = 27$ choice scenarios providing 51 total possible vectors that are z. If we pull vector of $z = (0, 0, 0, 0, 1)$ we randomly choose an ethnicity for both option A and B and if it's Macedonian we ignore the Macedonian as a second language. In practice this quirk affects roughly 2 percent of choice scenarios.

C.2 Preference estimation

For estimation, we use a similar population Monte Carlo procedure to estimate individual-level parameters. We adjust the choice problem so that z now includes the two second language indicators as before and an "own ethnicty" indicator, that equals 1 if the ethnicity shown matches the students language of instruction and 0 otherwise. For each student i, we have a vector $x_i^{(1:T)}$ of binary choice indicators and a T-by-3 matrix of designs $d_i^{(1:T)}$. We feed these data, the likelihood function and a prior through a procedure which returns the posterior distribution for the parameter vector θ_i.

We use a two-step procedure. We first run the analysis with a prior that is composed of three independent distributions $N(0, 100)$. Then, we take the resulting set of $\{\theta_i\}_{i \in N}$ to create priors that are normally distributed with mean and variance calculated from the leave-one-out set of thetas (*e.g.* For individual i, the prior is centered at $\bar{\theta}_{-i} = \frac{1}{N-1} \sum_{-i} \theta_k$).