Language, Identity, and Ethnic Cohesion in Schools

Ari Anisfeld *
University of Arkansas

Pepi Pandiloski
University of Porto

September 11, 2026

Abstract

Distinct group identities can create barriers for trade, investment in public goods, and social learning. This paper studies how integration impacts productivity and cohesion in settings with broad linguistic diversity, where language differences may deepen the divides created by group identities. We randomly assign high school students to same- or different-ethnicity partners to play a cooperative video game in Macedonia, where students are segregated by ethnicity. We measure productivity using video game scores, non-pecuniary experience with a survey, and assess social cohesion with two measures: the willingness to pay for a same-ethnicity partner (via an adaptive choice experiment) and ingroup bias (via a dictator game). We find that inter-ethnic contact improves social cohesion though with costs. Students report lower experience and we cannot rule out large negative productivity effects. To examine the role of language, we exploit random variation in the language skills overlap within pairs. We find that shared language skills are associated with greater inter-ethnic cohesion gains and lower non-pecuniary costs. These results reflect both a match effect (i.e., sharing a language is beneficial) and a correlation between language skills and positive potential outcomes. Additionally, cross-ethnic pairs match on one of the partner's home languages or a third neutral language. Cohesion gains are largest for participants who match on their own home language; if anything, matching on a neutral language is worse than matching on the partner's home language.

Keywords: Intergroup contact, Languages, Prejudice, Social interactions,
JEL: C90, C92, D83, D91, J15, J32, O15, Z13, Z18

*Corresponding author: anisfeld@uchicago.edu, arianisfeld.com. We would like to thank Steven Durlauf, Anjali Adukia, Raul Sanchez de la Sierra, and Lesley Turner for their thoughtful advice and gracious support, and to Oendrila Dube, James Robinson, and Leonardo Bursztyn for their invaluable guidance throughout the project. We also thank Violeta Petroska Beshka, Dan Black, Arun Chandrasekhar, Terence Chau, Josh Dean, Michael Dinerstein, Ben Feigenberg, Max Foxman, Rachel Glennester, Jens Ludwig, Alicia Menendez, Molly Offer-Westort, Goya Razavi, Max Tabord-Meehan, Mythili Vinnakota, Wendy Wong, 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 Nevenka Pandiloski, along with Pepi's extended family and friends, for a warm and welcoming introduction to Macedonia. This study was pre-registered in the AEA RCT Registry under the unique identifying number AEARCTR-0012559. The study was approved by the Institutional Review Board at the University of Chicago.

1 Introduction

Distinct group identities can create barriers to trade, investment in public goods, and social learning (Alesina and La Ferrara, 2005). These barriers are particularly pronounced when groups speak different languages, as linguistic differences can generate costly communication frictions and reinforce the salience of perceived identity divisions. Thus, early interventions that enhance social skills and shared linguistic human capital could potentially have large positive effects on prosperity through adult social and economic interactions (Boucher et al., 2024; Alan et al., 2021; Chyn, 2018). However, allocating resources towards integrated education can come at the cost of educational efficiency. Indeed, learning in one's mother tongue can be more efficient, especially early in life, which typically requires segregation.[1] To balance this trade-off, policymakers commonly propose cooperative contact interventions—organized activities intended to foster greater cohesion between groups.

This paper asks how integration impacts productivity and cohesion in the presence of linguistic diversity. We do so in high schools in Macedonia, a country with a recent history of ethnic conflict, where public schools segregate ethnic Macedonian, Albanian, and Turkish students by language of instruction. We randomize students into pairs to play a cooperative video game that requires teamwork and communication, with variation along two dimensions: whether partners share a common ethnicity and the overlap between partners' baseline language skills, which we call "communication potential". We measure cohesion with a choice experiment that elicits participants' willingness to pay for a same-ethnicity partner and measure ingroup bias in the amount of money shared during dictator games. Productivity measures come from video game scores.

We measure three key outcomes capturing distinct dimensions of integration success. First, we assess social preferences through an adaptive choice experiment that elicits willingness-to-pay for same-ethnicity partners in future interactions, as well as through dictator games measuring in-group bias. Second, we examine productivity costs by analyzing cooperative game scores and learning outcomes. Third, we measure non-pecuniary costs through self-reported enjoyment and experience quality. This comprehensive approach allows us to assess the trade-offs policymakers face.

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

Our main experimental results show that inter-ethnic contact improves social cohesion though at a cost. Cooperative contact with different-ethnicity partners significantly reduces preference for same-ethnicity partners in future interactions, with treated students willing to pay 24% less (45 denar) for co-ethnic partners at the median and 13 percent on average. Ingroup bias declines as well, leading to cohesion gains of 0.14σ, on average.[2] We find little evidence that productivity differs between same- and cross-ethnic pairs. However, the estimates are imprecise, and we cannot rule out economically meaningful negative effects as large as 0.3σ at the 95 percent confidence level. We do find clear evidence of non-pecuniary costs: students paired with different-ethnicity partners report significantly worse experiences. Understanding whether and how these costs can be reduced through policy design is essential for scaling integration interventions.

We then investigate whether shared language mediates both the benefits and costs of integration. In this multilingual setting, substantial variation in students' language abilities generates differences in communication potential across randomly assigned pairs.[3] However, analyzing these effects requires careful attention to selection: students with stronger multilingual skills are mechanically more likely to share a language with their randomly assigned partner, creating endogeneity in observed match quality. We address this through an "exposure mapping" approach that leverages the known randomization procedure to construct propensity scores, then employ augmented inverse propensity weighting with causal forests to recover unbiased estimates of how language complementarity affects integration outcomes. Results that rely on the communication potential are 'intent-to-treat' effects since it is an *a priori* measure that predicts but does not require communication between partners in a particular language.

Accounting for baseline skills, the correlation between communication potential and social cohesion can be decomposed into the benefit of being assigned a partner with high communication potential (match effects) and underlying individual heterogeneity correlated with language skills (selection). For example, language skills can reflect personality traits such as openness to experience that could also improve an individual's response to inter-ethnic contact. Alternatively, learning multiple languages can improve perspective-taking (Kinzler, 2020; Fan et al., 2015), which may be a fundamental building block for contact (Alan et al., 2021; Pettigrew and Tropp, 2006). The distinction is relevant for policy because they imply different solutions.[4] If

[2] σ is the control group standard deviations (see below).

[3] There is only linguistic variation in the different ethnicity pairs.

[4] Formally, we define potential outcomes indexed by high (above median) or low (below median) communication potential and estimate the average effect of going from a low to a high match. We use propensity re-weighting (detailed below) to account for selection bias, nevertheless, the match is a bundled treatment since pairs, in theory, could systematically match on communication

match effects drive the results, then a policymaker could increase cross-ethnic cohesion by funding activities similar to our intervention while carefully pairing students with overlapping language skills. More broadly, teaching mutual languages more effectively could complement future integration efforts. The decomposition suggests room for matching as a policy tool for building cohesion. Counterfactual analysis shows that optimal (feasible) matching could have improved the observed cohesion gains by roughly 30 percent–and the gains would be 45 percent higher if everyone could be assigned a high communication potential match.

Finally, coordinating on a language can complicate integration. Different ethnicity pairs either match on one of the partner's home languages or a neutral language (*lingua franca*) like English. Psychology research shows positive home language effects.[5] Using one player's home language implies the other partner is speaking a language where they are likely less fluent and may feel disadvantaged. The history of conflict and policy debates about the role of Macedonian language in schools (see Section 2) raise concerns about backlash from non-Macedonian students in cross-ethnic pairs that match on Macedonian. Thus, home language effects could lead to positive overall cohesion effects that are one-sided. Further, the contact hypothesis says contact is effective when participants have equal footing (Allport et al., 1954). So, it is ambiguous whether and how the language of the match might matter.

We find a positive home language effect that is not offset by backlash. When a cross-ethnic pair has the highest communication potential in one of the students' home languages, the home language speaker shows large cohesion gains. The dictator game bias drops to near 0. At the same time, there is no evidence of backlash for the second partner who continues 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.

Macedonia provides a close-to-ideal laboratory to study the role of language in inter-group contact due to the recent history of conflict, the increasing segregation in schools, and the variation in language skills across individuals and groups (Fontana, 2017; Bloodworth, 2020). In 2001, an armed conflict between Albanian militia and Macedoni fan authorities was resolved with the Ohrid Framework Agreement, which expanded the language rights of (large) minorities.While Albanians and Turks had access to home language instruction

potential and some other unobserved match characteristics that demonstrate complementarities.

[5]e.g. In a lab experiment, Jewish Israelis showed higher support for a peace plan when it was presented in Hebrew rather than English. The Hebrew effect held whether the proposal came from other Jewish Israelis or Palestinians (Grant et al., 2023).

before 2001, the Ohrid Framework policy changes reduced access to and need for the Macedonian language (Fontana, 2017). For example, a 2007 policy change made English the first *second language* taught in schools. Non-Macedonians study Macedonian language starting in fourth grade for a couple of hours per week. Many communities face tradeoffs between cohesion building and efficient learning in home languages.[6] 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.[7][8] Macedonia provides an example where language policy privileges home language instruction above nation-building throughout students' educational experiences.

We contribute to several strands of literature. First, we are the first paper to isolate the role of language in cooperative contact – and more broadly we show that the effectiveness of contact depends on participant skills, preferences, and the match between participants. We build on recent work with youth that has found mixed evidence of the effect of contact on cohesion: social preferences, attitudes, and behaviors toward outgroups. Some show strong positive results in attitudes and behaviors (Rao, 2019), others indicate changes in behavior but not attitudes (Scacco and Warren, 2018; Elwert et al., 2023), and still others report generally positive results, which degrade as the proportion of out-group team members increases (Ghosh et al., 2024). In work complementary to ours, Boucher et al. (2024) show that integration improves language acquisition of Syrian refugee pre-schoolers in Turkiye.[9] These and other studies of the effect of inter-group contact tend to 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). By incorporating a measure of baseline language skills, we show that inter-ethnic contact is less effective in improving cohesion if parties have communication barriers and also find that students who have higher language skills *ex ante* have better cohesion outcomes ("positive selection"). Since skills can be learned (e.g.

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

[7]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, 2004 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).

[8]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). 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 (Ginsburgh and Weber, 2020). Areas in Latin America where indigenous language speakers have fought for the rights to mother tongue instruction (Hynsjö and Damon, 2016) and so forth.

[9]In fact, language skill development is often a result of extended inter-group contact, a point we return to in the conclusion.

Alan et al., 2021) and assignments to contact experiences can be targeted, it is important to understand this heterogeneity.

Second, we contribute to the literature on the role of social identity on human behavior (Tajfel et al., 1971; Akerlof and Kranton, 2000). One strand of this literature seeks to understand why cleavages between groups become salient and why individuals identify with specific groups at specific times (Shayo, 2020; Posner, 2005). Previous work has highlighted the role of group status and group salience on identification with a group.[10] More recently, Atkin et al. (2021) show that costs can contribute to identification with a group. Whereas they refer to the market prices of group-specific foodstuffs and totems, we show that the language barriers impose non-pecuniary costs. Inter-ethnic pairs with higher communication potential demonstrate lower in-group favoritism and report lower non-pecuniary costs associated with contact. Thus, language differences increase individuals' identification with their ethnic group by increasing the costs of adopting a more inclusive identity. The fact that the cohesion gains are associated with a "home language effect"– suggests that this identification is not a function of being more "cosmopolitan" *a priori*.

Relatedly, we add to the literature on linguistic diversity and productivity. A common theoretical perspective is that linguistic diversity decreases productivity due to communication barriers (Lang, 1986), which is a finding supported by recent empirical work (Dale-Olsen and Finseraas, 2020; Guillouët et al., 2024). Others find that ethnic diversity in the workplace can lead to sabotage (Hjort, 2014). In our setting, communication barriers do not impact productivity and students have monetary incentives to provide effort for a joint payout. Yet, there may be non-pecuniary costs of inter-group contact and students are willing to pay to avoid these. The costs may reflect cognitive burden, in-group preference (homophily), or outgroup animus. We find that such costs decrease as communication potential increases, which highlights a positive externality to an individual's language learning, as in Lazear (1999). The result suggests that school systems that provide mother-tongue-based instruction create opportunities to build communication skills across groups through integrated classrooms for some portion of the day and extracurricular classes. More broadly, the work adds to our understanding of how language and communication influence social interactions (Durlauf and Ioannides, 2010).

Finally, we provide insights for experimentalists interested in working with paired experiments or measur-

[10] Some examples of endogenous identification based on status and salience include Nix and Qian (2015), Bazzi et al. (2019), Blouin and Mukand (2019), and Fouka (2020).

ing 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 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 a team setting. Finally, we show how a measure from the labor economics literature on valuing workplace amenities can be adapted to provide estimates of the willingness to pay for segregation at an individual level (Drake et al., 2022).

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

Macedonia offers a compelling backdrop for exploring the intersection of language and societal integration. This section delves into the nation's recent turbulent history, the pivotal role language policy played in the conflict, and the carefully negotiated resolution that sought to mend divides. From there, we show the implications of current language policy for the lived experiences of students in the three high schools of the study. For background, 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 less populous ethnicities.[11]

In 2001, Macedonia faced a crisis that led to the death of 200 people, and the displacement of nearly 7 percent of the population (Merdzanovic, 2015). Following Kosovo and other Yugoslav Wars, Albanian militants aimed to gain broader rights in the young republic[12] where Macedonian political leaders had centralized governing powers, and restricted access to bureaucratic jobs as well as expression of culture.[13] To avoid broader conflict, Albanian and Macedonian political leaders negotiated a power-sharing framework, 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.[14] These policy changes likely increased ethnic segregation in schools.

The Ohrid Framework Agreement addressed questions about language policy that underscore long-standing political tensions about the role of *nationalities* and cohesion in the multi-ethnic state. Since at least 1950, governments for what now makes up Macedonia have funded Albanian-language primary and secondary classrooms, though at times government support was low.[15] For example, in the mid-1990s when Albanians opened an "illegal" Albanian-language university in the city of Tetovo, the government bulldozed several buildings and arrested its leaders on "charges of sedition" (Fontana, 2017, pg. 101).After Ohrid, local governments gained more control over schools, secondary schools in Albanian and Turkish expanded, and the Albanian-language university gained full recognition. These policy changes increased the sense that the

[11]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]Formerly a constituent republic of Yugoslavia, Macedonians voted for independence in a 1991 referendum, which Albanian citizens largely boycotted.

[13]The flying of national flags provides flash point where these tensions become visible. On July 8 1997, the government passed a law restricting the use of "foreign" flags, aimed primarily at the Albanian's double-headed eagle and the Turkish flag. In defiance of the law, the Gostivar mayor Rufi Osmani flew both flags over the municipal building. On July 9, 1997, Macedonian police forces removed the flags and raided the municipality killing 3 and injuring more (*Police Violence in Macedonia* 1998). Today, in front of the Gostivar municipal government building fly the flags of the municipality, Macedonia, Albania, Turkey, the Roma, and the United States.

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

[15]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.

country had "a set of ... parallel education systems, depending on the language" according to 2010 policy document (cited in Fontana, 2017, pg. 257). The current approach to cohesion can be summarized as "if they don't interact, they won't fight."[16]

There is uncertainty about the role of Macedonian versus English as the language of common use in the country, which motivates our final set of research questions. English is considered a neutral language that could serve as a *lingua franca* if more students were to prioritize and master it. As of 2007, everyone begins English as a second language in first grade. Recent calls for integration focus on using full-time English-based classrooms in high schools, which are being piloted at several schools around the country (Bloodworth, 2020). In contrast, when the government attempted to make Macedonian the *first* second language in 2010, the policy faced severe backlash and was overturned by the courts (Fontana, 2017). Currently, Albanian- and Turkish-tracked students commence Macedonian language instruction in fourth grade.[17] A common complaint by non-ethnic Macedonians is that ethnic Macedonians are not required to and, typically, make no effort to learn Albanian (or Turkish). Theory predicts such communication free-riding by the majority (Lazear, 1999). We show, below, that there is a mixture of language learning strategies across groups–though Macedonians rarely learn local out-group languages.

Study Site: Our study site is an ethnically diverse town, in which the Albanians are the majority of the population, followed by Macedonians and Turks as large minorities.[18] Inverting the national population structure.

In the study site, public schools offer parallel tracks in Albanian, Macedonian, and Turkish.[19] In practice, students from distinct ethnic backgrounds attend the same physical schools but are segregated by the 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

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

[18] 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.

[19] As a note, Turkish identity reflects a remnant from the Ottoman colonial past, rather than recent immigration. (Ottomans held the territory of Macedonia for nearly 500 years until the First Balkan War of 1912.)

day is divided into shifts organized by language.[20] Next, Amra and Ana matriculate into the Gymnasium, a college preparatory high school, which organizes school shifts by grade level. Although they begin the day with all first-year students regardless of ethnicity, their classes are conducted in their home languages, and their peers are predominantly of the same ethnicity. Additionally, Macedonian classrooms are in different hallways from Albanian and Turkish classrooms. Thus, despite increased opportunities for cross-ethnic encounters, Amra and Ana's K-12 experience remains largely segregated. They would have similar experiences at the other two high schools in our setting. (The Medical High School also organizes shifts by cohort. The Technical School organizes shifts by ethnicity.) 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.[21]

Thus, high schools in Macedonia offer a unique setting to study language and integration. While students remain socially segregated, they attend the same schools, which lowers the direct costs of integration activities. Although students share some language skills, the level of overlap varies between pairs, revealing different communication barriers. Moreover, there is some political support for addressing these questions; policymakers wonder whether the Ohrid Agreement reforms have led to too much segregation (Bloodworth, 2020).

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.

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

[21] 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.

3.1 Recruitment and Timeline

All students at the three focal high schools were invited to participate in the study. From December 2023 to May 2024, enumerators visited classrooms, explained what participation would entail, and helped students complete the initial assent form. Participation required parental consent as well. Enumerators 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."*

Upon receiving parental consent, enumerators conducted the baseline survey. 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.Many of the same participants were invited to participate in our companion study, where students, in groups of 6, played a networked social learning game, where participants pass 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 strived to have students complete the first three steps before participating in Study 2.

3.2 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 ethnic identity. For efficient logistics and cost minimization, we randomized within strata on a rolling basis to provide steady work for the facilitation teams.[22] Randomization into pairs based on individual characteristics requires randomization that is more complicated than coin flips. Treatment is defined as having a partner of the same or different ethnicity, so we must ensure that individuals assigned to a specific treatment has an adequate match. Our procedure (detailed in Appendix B.1) adjusts *individual's* probability of being assigned a co-ethnic partner based on their group's relative size within strata to hold the *strata-wide* co-ethnic assignment proportion at 25 percent. The method

[22] 11 strata have one wave of randomization, 10 strata have two waves, and 2 strata have three waves

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.

3.3 Main intervention: video game sessions

In the video game 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.[23] Before starting the survey, students introduced themselves and answered two icebreaker 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, proceeded to learn how to play the game from YouTube videos,[24] and then practiced playing in "solo mode" where one player controls both chefs (described below).[25] The "solo mode" scores measure individual video game skill.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. After learning and playing "Overcooked! 2", students returned to individual computers to answer questions about their experience in the session and completed an adaptive choice experiment, which 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 interacting for roughly 20 to 25 minutes.

To provide some flavor of how the game works, level 1.5 has chefs prepare spaghetti with red sauce. To

[23] We hired teachers from the school to organize and run sessions. Each session had two teacher facilitators so that each student had at least one teacher who 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".

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

[25] 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. In addition to the solo mode, the treated pairs were assigned one of three identity primes inspired by Chen et al. (2014). Two arms primed common identity through town and school group membership, while a the third group got the control arm primes focused on cellphones. In Appendix G we present the results and argue that they are not meaningfully nor consistently different across arms. Hence, we pool the results across priming arms.

prepare one dish requires boiling pasta, chopping tomatoes, sauteing tomatoes, plating the pasta and the sauce before either one becomes "overcooked",[26] 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 shared language skills would have a meaningful impact on students' ability to divide labor and score points. Since many, if not most, real-life scenarios for integration require communication, this feature of the game helps mimic realistic conditions.

"Overcooked! 2" has many advantages for cooperative contact and studying social interactions.[27] First, students work cooperatively towards a common goal on equal footing, matching the criteria set out by the contact hypothesis (Allport et al., 1954). Second, productivity in the game mirrors a real-world kitchen (minus the knife skills). In some ways, it is more challenging than the real-world setting since players must self-organize their production process. Finally, video games are a culturally appropriate medium for youth contact. In our sample, three out of five students could recommend a video game by name, 52 percent said they played some days or more frequently and 71 percent said the expression "I love video games" describes them somewhat or more.[28] Boys tend to enjoy video games more; the figure is 86 percent for boys and 59 percent for girls. The boys score more points on average, but we do not find heterogeneous treatment effects by gender.

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

[27] 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.

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

4 Data and Measures

4.1 Data sources

Data primarily comes from two sources: (1) a baseline survey and (2) video game sessions. The baseline survey featured modules on language skills, social networks and social closeness, parents, video games, educational outlook, and a real stakes dictator games.[29] Additionally, we integrate administrative data from the school record keeping software, which provides additional covariates (GPA, absenteism). We also run an endline survey with questions parallel to the baseline. Due to field logistical constraints and the end of the school year, there is a lot of variance in the timing of the survey relative to the video game sessions.

We collect data from the video game sessions using a mixture of participant surveys, facilitator surveys, and screen recordings. The participants answer questions about their experience playing the game and their willingness to interact across groups. For each session, each facilitator held a worksheet to track the students scores in "Overcooked! 2", and evaluate the pairs communication and experience levels. Finally, we retained screen recordings from which we use computer vision technology to extract data about game play. Screen recordings provide audio and visual data of the cooperative game play. These provide both corroborating evidence for the facilitator reports and rich auxiliary data to explore mechanisms. Appendix C provides details on video processing. Overall, the rich and varied data sources allow us to understand the student experience and to cross-check reported data.

4.2 Sample characteristics, baseline integration, and balance

In Table 1, we present summaries of baseline characteristics and balance tests. 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.[30] In practice, 8 students whose LOI does not map to

[29] The baseline survey is available in Appendix **??**.

[30] There are two prevalent exceptions. First, There are 11 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.

their ethnicity are assigned to control unit "co-ethnic" pairs.[31] These hard-to-classify individuals attenuate treatment effect estimates as some control units may act like treated units.

The demographic and behavioral measures show balance assessed by running a regression on the treatment indicator with out controls (see Tables 1 and 2). The sample appears well-balanced as we observe approximately the same number of variables are significant at the 5% and 10% levels as what we would expect due to chance.

In this segregated environment, baseline measures show low 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 46 percent of students name at least one peer who speaks a different mother tongue; compared to 95 percent who name at least one friend. In total, 86 percent of all distinct named links are co-ethnic.[32] 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" 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 2.1 out of 7 (barely overlapping circles). Finally, students share less with out-group members in dictator games. On average, there is a 9.9 percent gap in baseline sharing (See section 5).

4.3 Communication potential

Communication potential 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

Using the Inclusion of Others in Self measure, Roma students feel close to Macedonians at a level slightly below what Macedonian students report (5.9 versus 6.7 out of 7). Second, roughly 8 percent of students who identify as Turkish in this study complete coursework in Albanian or Macedonian. One reason for this is that 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.

[31] There are 21 students whose LOI does not map to their ethnicity in treated pairs. Fortunately, all 21 of these students, in fact, have a cross-ethnic partner.

[32] This represents a lower bound as 127 students use all 6 slots for their friend clique.

mutual communication skills, which will help understand match effects.

Figure 1 shows that many but not all students report high skills in a second language.[33] The figure shows we can expect variation in whether two random students have complementary language skills. As expected, Macedonian students tend to have low second-language skills in local languages and specialize in English. Meanwhile, Albanians and Turks have more varied language skills. If language similarity predicts ease of learning an additional language, there is not a distinct advantage for any group in our setting as the four languages 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.

Operationalizing *communication potential* requires measures of individual language skills and a method to combine them to characterize the complementarity of skills. 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.[34] Then, for pair i and j, we determine a joint score in a given language by multiplying i and j's scores, corresponding to the geometric mean (squared). We define *"communication potential"* 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)

$$\underset{\text{communication potential}}{f_{ij}^{\max}} = \max(f_i \odot f_j)$$

where $\odot$ denotes the element-wise (Hadamard) product and the $\max$ operator selects the largest element in the vector. The "communication potential" 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 59 to 88. Students are likelier to use languages with high joint scores; less than 10 percent of cross-ethnic pairs have a second-place language match above 75 suggesting

[33] 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.

[34] Averaging helps reduce noise in the self-reported skills.

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; Roughly 4 out of 5 pairs that talk more than "a moderate amount" use the language with the highest communication potential (See Appendix E). By design, almost every control pair has a communication potential of 100. So analyses that rely on the communication potential measure focus on treatment pairs only.

4.4 Outcomes

We build 4 indices to measure cross-ethnic cohesion, productivity, student experiences (non-pecuniary benefits), and communication. Together these capture the main trade-offs in integration. Building cohesion or reducing prejudice is typically the purported goal of integration, but this can come at a cost to productivity or have non-pecuniary costs. Communication is a mechanism hypothesized to explain these costs (Dale-Olsen and Finseraas, 2020; Guillouët et al., 2024; Lang, 1986) and using it as an outcome also provides a proof of concept for the "communication potential" measure. For each outcome type, we construct indices based on 2 to 3 related variables. To allow for comparisons, we normalize each measure $\tilde{y}_i$ by subtracting the control group mean μ_0 and standard variation σ_0:

$$y_i = \frac{\tilde{y}_i - \mu_0}{\sigma_0}$$

Then, the index is the simple average of all composite variables.

4.5 Outcomes: Cross-Ethnic Cohesion Measures

The purpose of contact interventions is to build cross-group cohesion or reduce cross-group prejudices. Social ties across groups, prosocial preferences, and an absence of violence characterize a cohesive school environment. We measure these concepts using survey instruments:

1. Willingness to pay for same-ethnicity partners
2. Out-group bias measured with dictator games

Adaptive Choice Experiment: Willingness to pay for same-ethnicity partners

The amount a student is willing to pay for a co-ethnic partner 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). An individual's valuation may reflect ingroup preference (homophily), out-group bias (animus), or expected non-pecuniary benefits, such as lower cognitive burden or more fun. The expectation of non-pecuniary benefits is a form of statistical discrimination. Let us define prejudice as the sum of animus and the extent to which the students' expectations of non-pecuniary benefits are incorrect. Then, treatment effects net out the homophily channel and "correct" statistical discrimination, so treatment effects can be interpreted as a decline in prejudice.[35] In any case, the decrease in willingness to pay for a same-ethnicity partner implies an increase in willingness to cooperate across ethnic lines.

Students face 12 binary choice problems where they reveal their preferences regarding partner ethnicity, partner language skills, and a sure cash payment.[36] Bayesian updating determines the next trade-off in a process analogous to a binary search (Drake et al., 2024). In Appendix D, we detail the theoretical basis and estimation. The benefit of the machinery is that it allows us to estimate multiple parameters (see below) 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).

We focus on the parameters of the following linear utility function:

$$U_{ij} = p_j + \theta_i \mathbb{1}_{\{ethnicity_i = ethnicity_j\}} + \gamma_i^e \mathbb{1}_{ESL_j} + \gamma_i^m \mathbb{1}_{MSL_j} + (\gamma_i^{eo} \mathbb{1}_{ESL_j} + \gamma_i^{mo} \mathbb{1}_{MSL_j}) \times \mathbb{1}_{\{ethnicity_i = ethnicity_j\}} + \epsilon_{ij}$$

[35] This is under the assumption that homophily is unchanged. In general, changes in homophily and prejudice are not easily distinguished.

[36] 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.

Where p_j is a sure price paid for choosing partner j with a utility parameter normalized to 1. Terms $\mathbb{1}_X$ equal one if X is true. θ_i capture the WTP for a co-ethnic partner in denar units. γ_i^e and γ_i^o capture the willingness to pay for the partner's second language skills in English (ESL) or Macedonian (MSL). (Recall, English and Macedonian are taught in school and we are interested in the role of neutral *lingua franca* compared to the language of the dominant group.) γ_i^{eo} and γ_i^{mo} capture the additional valuation of a co-ethnic partner's second language skills.[37]

We estimate these parameters using population Monte Carlo in two steps. First, we estimate the vector of $\{\theta_i\}_{i \in N}$ 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 D).[38]

The utility parameter θ_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 differences in θ (i.e. a treatment effect) can be interpreted as a change in the "price of prejudice" Hedegaard and Tyran, 2018.

The individual willingness-to-pay parameters induce a demand curve that allows us to visualize heterogeneity in the population. 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). Note, as with any quantile treatment effects, the effects are only identified across distributions, not for individuals at a given quantile. To get at the latter, we need an additional strong assumption that individuals maintain their rank across treatment assignments.

To provide support for the validity of this measure, 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 (on a 10-point scale) in a given language are willing to pay 2.5 to 4.4 denar for a partner who speaks that language.[39]

[37] The term γ_i^{mo} is meaningless for Macedonian respondents as their co-ethnics speak Macedonian as a first language.

[38] 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.

[39] These come from linear regressions of willingness-to-pay on the fluency measure with strata fixed effects and with or without language of instruction fixed effects. (We drop Macedonian students from the Macedonian language regressions). In reality, the demand is non-linear in the fluency measure.

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 performed on networks show that people share more with people who they feel closer to (Goeree et al., 2010). For this reason, we focus on the difference 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.[40] 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 anonymous recipient's mother tongue ("the same" or "a different") and school membership ("from your high school" or "from another high school"). Some choices are realized, adding an incentive for truthful reporting.[41] 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 27.3 Denar (9.9 percent) more with their co-ethnic partners on average.

4.6 Outcomes: Productivity

The productivity index comprises of:

1. Cooperative game scores
2. Quiz scores on how to play the game

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

[40] We say 50 denar equals $1 when in reality the exchange rate is closer to 55 denar for $1. The national minimum wage in 2023 was 20,175MKD per month, which amounts to roughly 120 MKD per hour. Purchasing power is higher in Macedonia. An espresso cost between 40 and 60 MKD in Gostivar in mid-2024.

[41] These games have been used successfully with high school students and younger children (e.g. Ghosh et al., 2024; List et al., 2017).

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.[42] The total score encapsulates total productivity.

Students complete a quiz on how to play "Overcooked! 2". The quiz includes questions about the mechanics of the game controls and details about recipes and scoring. Learning is an important productivity outcome in a school setting. We explicitly teach some of the quiz content in students' home languages. This mimics learning in a bilingual classroom where bilingual course material is available.

4.7 Outcomes: Communication

The communication index comprises of:

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 (.80). The second measure uses students' ratings on ease of communication which correlates with facilitator scores (.39). In Appendix A3, we show that audio data from the screen recordings corroborates these measures. We use off-the-shelf audio processing algorithms to detect speech and find correlations of roughly .5 between the average facilitator score and the algorithm detected speech (a noisy measure of the underlying construct).

4.8 Outcomes: Student Experience (non-pecuniary costs/benefits)

Experience captures non-pecuniary costs or benefits of the "job". This captures a potential compensating differential that could affect wages and explain in part the demand for segregation.

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

The experience index comprises student self-reports on their experience playing with their partner on three dimensions[43]:

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 0.62 and 0.74. Facilitators also report, from their vantage point, 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 .26 to .31. Facilitator-assessed frustration and student self-measures are negatively correlated and range from -.21 to -.27. 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

$$y_i = \tau D_i + \phi_g + \beta X_i + \epsilon_i$$

y_i is an outcome for individual i (normalized as described above) or an average of these variables (index), D_i

[43]Occasionally students fail to answer all three questions. We retain the 24 students who skipped one question (averaging their responses to the remaining 2 questions to form the index) and drop 12 who answer one or fewer question.

is an indicator of being assigned a cross-ethnic partner, τ is the average 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. ϕ_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[44] 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 propensity scores.[45] 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).[46] The estimator combines these in the causal forest to estimate conditional average treatment effects given pre-determined covariate vector x:

[44] As a caveat, these are measured during the session and could be influenced by treatment assignment–e.g. if being in the same room changes effort. That said, the scores are measured before the main contact experience. Ultimately, the results are similar with or without the solo scores included. We present this in the results.

[45] 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.

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

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

The superscript $(-i)$ denotes the estimators are calculated using subsets of the data that exclude i (i.e. they are "out-of-bag" predictions) and make these so-called "honest" forests. 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. For the analysis of inter-ethnic contact, the treatment assignment indicator is exogenous and all but two students have a probability of being assigned treatment or control due to our randomization procedure.[47] 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).

These robustness checks also help account for differential attrition. We include the exact randomization propensity score (from the randomization procedure) alongside baseline covariates in the causal forest, which flexibly accounts for differential attrition under a missing-at-random (selection on observables) assumption. Since this assumption may be strong for attrition, Appendix H presents additional analyses: standard and covariate-stratified Lee (2009) bounds, which require only that treatment weakly reduces completion, not MAR, and results estimated on the low-attrition subsample excluding Tech school which accounts for the most attrition. These results are qualitatively consistent with the main findings.

5.2 Correlation between "Communication potential" and outcomes

We next turn to the role of language. The following regressions explore the heterogeneity between communication potential and outcomes across treated students (in cross-ethnic pairs). Recall there is almost zero variation in communication potential in the control group. We run:

[47] The two students were the only representatives of their language of instruction within a strata, so we technically cannot assign them to the control condition.

$$y_i = \alpha + \gamma * D_i + \beta * D_i * \frac{f_{ij}^{\max}}{IQR} + \epsilon_i$$

β, the parameter of interest, estimates how much outcome y changes on average when moving from the 25^{th} to 75^{th} percentile of the "communication potential" distribution among different-ethnicity pairs. The interquartile range (IQR) goes from 59 to 88 points or approximately 1.5 standard deviations.

To summarize effects, we also use the following specification, after splitting the treated sample at the median "communication potential".

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

We are interested in β^L and β^H, which represent the contrasts between the control group and the average y_i for individuals in pairs with Low and High language matches. These matches are defined as being below or above the median m, respectively.

6 Contact Results

6.1 Contact improves cohesion at the expense of the student's experience

We first present an overview of the "contact intervention" results. Figure 3 shows treatment effects (i.e. $\tau = E[Y_i(1) - Y_i(0)]$) 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 Tables 3 and 4, we report the communication results as well as the results shown in the figure with the inclusion of covariates.

Table 3 (panel 1) highlights cross-ethnic pairs had a harder time communicating than those in co-ethnic pairs. The communication index effects are negative 1.2 standard deviations. 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. 83 percent of co-ethnic pairs communicated more than "a moderate amount"

compared to 34 percent for treated groups.[48] The communication challenges motivate understanding the role of language in later analysis.

Students in cross-ethnic pairs also report having a aworse experience compared to students in a same-ethnicity pair (Figure 3, rows 4-7 and Table 3 (panel 2)). The experience index effects are negative 0.36 standard deviations. The experience index captures the non-pecuniary costs of working across ethnic lines. 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". 95 percent of control students said *agree* or *completely agree* compared to 86 percent of students in mixed pairs. Overall, the students in cross-ethnic pairs had generally agreeable experiences, but the students in co-ethnic pairs had an even better experience.

While students in cross-ethnic pairs had trouble communicating, we cannot reject a null effect on productivity measures (Figure 3 rows 8-10 and 4 (panel 1)). The productivity index effects are negative 0.06 standard deviations. Same-ethnicity pairs scored 327 points on average. Mixed-ethnicity pairs score 15 fewer points (s.e. 15) on average, accounting for strata fixed effects. The productivity loss is 4.5 percent, but indistinguishable from 0. The game score effects have a smaller effective sample size and wider standard errors and we cannot rule out effects as large as -44 point (13 percent, or .31 standard deviations). The effects are nearly identical when accounting for video game skills (as measured with solo scores). Treatment has tighter null on learning about the game. Students with co-ethnic partners score 69.8 percent on the short quiz about how the game works. Students with cross-ethnic partners score 69.5 percent.

The final three rows of Figure 3 show that playing with a cross-ethnic partner improves cohesion measures. The cohesion index comprises of ingroup bias measured through dictator games and willingness-to-pay for co-ethnic partners, multiplied by negative 1. Cohesion index effect is positive 0.14 standard deviations higher, reflecting similar magnitude statistically significant declines in sub-measures. In levels, ingroup bias falls from 19.8 to 14 Denar. This reflects a 29 percent drop in ingroup bias though is not statistically significant.[49] We discuss willingness-to-pay in the next section.

Treatment effect estimates are quite stable when estimated with causal forests (lighter points on 3). This

[48] 48 percent of co-ethnic pairs talked "Constantly" compared to 12 percent of cross-ethnic pairs.

[49] Another way to put the estimates in perspective is to compare them to other studies. In a recent study of a two-week inter-ethnic camp in India, participation reduced ingroup bias in dictator games by .28σ (Ghosh et al., 2024), roughly triple the effects found here.

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.

6.2 Contact shifts demand for same ethnicity partners inward

Figure 4a and 4b show demand curves induced by the individual valuations for a co-ethnic partner and for a partner who speaks Macedonian or English 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. Demand curves show the distribution of valuations which provides insights beyond mean differences.

The control group's demand for same-ethnicity partners is shown as a red solid line. Most students are willing to pay for co-ethnic partners with valuations high as 450 Denar, roughly $9 USD. Roughly two-thirds of students have WTP above 125 Denar which was equivalent to an hours work at minimum wage at the time of they study. Roughly 6 percent of students are willing to pay to have a cross-ethnic partner.[50] Taken together, the demand curves show highly homophilic preferences.

The demand curve for students assigned a cross-ethnic partner is shown in as the (blue dotted line). A Kolmogorov-Smirnov test rejects the hypothesis that the demand curves were drawn from the same distribution ($p = .031$). The distribution shifts inward starting roughly at the 75th percentile. So while WTP declines by 13 percent on average for students assigned to cross-ethnic partners, "cohesion gains" are not uniform. Willingness to pay is 24 percent lower at the median and the proportion willing to pay an hour of minimum wage drops 9 percentage points. Quantile treatment effects are significant throughout the middle of the distribution, but not in the tails.

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. In other words, among Macedonian students willingness to interact increases with other ethnicities when they do not speak Macedonian as a second language. This is consistent with the ethnicity results; suggesting more openness to working across

[50] 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.

ethno-linguistic lines. [51] At the same time, we see little difference in WTP for Macedonian as a second language between treatment and control students from non-Macedonian tracks. Finally, Figure 4c shows no treatment effect on WTP for English as a second language. The mean WTP for English is positive. Students value English skills on average, and the treatment does not appear to influence that valuation.

6.3 Language match predicts outcomes

Communication potential has significant positive associations with communication, experience, and cohesion. At the same time, we fail to reject the null that there is no relationship between communication potential and the productivity index. We characterize the relationship between communication potential and outcomes in two ways (see 5.2). Figure 7 shows the linear relationship and Tables **??** and **??** provide the non-parametric summary comparing students in pairs with below and above median communication paper.

First, The communication potential score is highly correlated with the communication index. This correlation supports using "communication potential" to measure a pair's ability to communicate, while only relying on each student's self-reported fluency vector. Every 30-point increase in communication potential$_{ij}$ is associated with a .65 (se =.11) standard deviation gain in the communication index. To put these numbers in perspective, when we compare the bottom to top quartile, the average amount of communication increases by 1.5 points on a 7-point Likert scale. That is the pairs talking move from below "a moderate amount" (3.2 out of 7) to above it (4.7 out of 7), where 4 is "a moderate amount". In comparison, the same ethnicity pairs talked much more: 5.8 out of 7. Second, students in pairs with higher communication potential report better experiences; while there is a substantial gap in experience reported for the lowest matches, the treated individuals approach the control group average as $f_{ij}^{\max}$ approaches 100. Third, cohesion is increasing in communication potential. Low matches have marginally significant gains in cohesion, while high matches have highly significant positive gains. Moving from the 25th to the 75th percentile of the communication potential distribution is associated with an extra 0.08 standard deviation in the cohesion index. While the high match pairs see statistically significant gains, the evidence is consistent with a story that says having low communication potential results in little to no reduction to ingroup bias. Finally, we cannot reject the

[51] Appendix Figure A4 shows demand for partners with specific language skills disaggregated by ethnicity (aggregating across treatment assignment). The median Macedonian student is willing to pay over 130 Denar for their partner to know Macedonian as a second language, whereas the median non-Macedonian student has a WTP of of 18 Denar.

null hypothesis that there is no relationship between communication potential and productivity.

Taken together, we find strong, but suggestive, evidence of communication potential influencing students' experiences and cohesion outcomes. These results reflect associations and in the next section we will make progress toward untangling selection from match effects.

7 Towards Causal Heterogeneous Effects: Exposure Mapping

Randomly assigning partners gives us hope that we can untangle two effects. One is the benefit of having a partner with a high "communication potential": we call these "match effects". The other is underlying individual heterogeneity. Students with better language skills are more likely to be in a high-language match pair. And at the same time, they may share traits that affect both their language skills and their cohesion outcomes.

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

7.1 Illustrative 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 can be used to estimate unbiased treatment effects. 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 $p(X) = \mathbb{E}[\text{High Match}|X]$. Meanwhile, co-ethnic pairs always have a high-language match.

To highlight the bias, Figure 5 shows results from a stylized simulation with constant treatment effects due to cross-ethnic contact. We set $\tau = \mathbb{E}[Y(1) - Y(0)] = 0.2$ for all pairs, but we assume heterogeneity in baseline cohesion measures. This is plausible. Bilingualism may endow B students with better perspective-taking (Kinzler, 2020; Fan et al., 2015), which could, in turn, increase their cohesion measures regardless of treatment (Alan et al., 2021).

Then, the true data-generating process is given by:

$$
\begin{aligned}
Y_i &= \beta_i^L D_i \times (1 - H_i) + \beta_i^H D_i \times H_i + Y_i(0) \\
\beta^L &= \mathbb{E}[Y_i(LowMatch) - Y_i(0)] \\
\beta^H &= \mathbb{E}[Y_i(HighMatch) - Y_i(0)]
\end{aligned}
$$

Where $Y_i(0)$ captures underlying heterogeneity, D_i is a treatment indicator, and H_i is an indicator of having a high language match. In the toy example, we set constant effects so $\beta_i^L = \beta_i^H = \tau$.

A few observations: First, while bilingual students are likelier to be in a high language match pair, 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 by language match, 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 conditional ATE (CATE) if we can estimate $\mathbb{E}[Y(1,1)] = \mathbb{E}[\mathbb{E}[Y(1,1)|\text{Match Type}]]$. Inverse propensity weights allow us to address the biased CATEs.[52] Moreover, we know the true propensity score because we know how baseline characteristics map to a "high" or "low" match given the randomization scheme.

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

7.2 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 potential outcome framework and then show how we can estimate potential outcomes of having a high or low language match partner. Then, the difference between these potential outcomes provides an ATE of having a high match versus a low match.

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:

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

with $\hat{\beta}^L$ defined analogously and $\hat{\gamma} \equiv \hat{\beta}^H - \hat{\beta}^L$. 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. While we do not observe potential outcomes $Y_i(1,1)$ and $Y_i(1,0)$ for the same student i, correlations between pre-determined covariates and match type suggest $\mathbb{E}[y(1,1)|(z_i,l_i)=(1,1)] \neq \mathbb{E}[y(1,1)]$. Going from the 25^{th} to 75^{th} quartile probability of being assigned to a high-match pair is associated with .56 additional named friends (13 percent increase) and .14 additional named cross-ethnic friends (28 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 $\pi_i(z,l) = Pr_i(z,l)$. The key insight of Aronow and Samii (2017) is that the propensity score is known because we control the treatment assignment mechanism and know the language skill vectors. Simulations of the randomization procedure provide estimates of $\pi_i(z,l)$. We run the randomization code R times and for each randomization collect the assignment $d(z,l)$ for each individual. Then, $\pi_i(z,l) = \frac{\sum_{z_i=z,l_i=l}^{R}}{R}$.

Then, under the following assumptions, we can estimate the average 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. To ensure common support, we implement a data-driven trimming procedure prior to estimation using empirical common support. In particular, we set the analysis region to include observations with $\pi_i(1,1) \in [\min(\pi_i(1,1)|l=1), \max(\pi_i(1,1)|l=0)]$. Applying these empirical cutoffs effectively trims the tails of the propensity score distribution, discarding 18% of the total sample (42 students from the lower bound and 44 from the upper bound). When we restrict the sample based on overlap, we estimate a LATE local to students who have some possibility of being in any given treatment state.

Assumption 2 Correct specification: The exposure mapping is accurate. The assumption says that $E[Y(1,1)]$ and $E[Y(1,0)]$ are estimands that capture the true causal relationship between communication potential and outcomes. Sävje (2024) argues this assumption is overly strong.[53]

[53] "Exposure mappings are primarily used to collect and describe sets of assignment vectors that share a similar interpretation

Assumption 3 Selection on Observables: $Y_i(1,1), Y_i(1,0), Y_i(0) \perp D_i | X_i$. This assumption is met by random assignment and because the exposure mapping uses X_i to define the potential outcomes.

To increase estimation efficiency, we use causal forests to estimate γ with an augmented inverse propensity weighted estimator (AIPW). As discussed in Section 5, causal forests combine flexible estimators of the mean and propensity score (estimated via nuisance functions). Causal forests provide similar results to exposure mapping. Theoretically, the results should be similar because both methods adjust for differences in treatment propensity. When using causal forests, we estimate propensity scores ($\hat{\pi}(z, l)$) in first-stage random forests that use the propensity scores attained from the randomization procedure and a rich set of baseline covariates. (The procedure uses sample splitting or "honesty" to avoid overfitting.) This smooths the propensity scores away from zero, which is desirable since idiosyncrasies in recruitment impact the propensity scores (see Appendix B.1). Causal forests produce consistent estimates when potential outcomes are independent of treatment assignment conditional on observables $Y_i(1,1), Y_i(1,0), Y_i(0) \perp D_i | X_i$ and there is overlap, which are subsets of the assumptions required for the exposure mapping argument. Overlap-adjustment changes the interpretation of the effects. Instead of estimating average treatment effects for the entire sample, we estimate the average treatment effect that are local to students in the overlap sample.

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. In this section, we consider two mechanisms. First, we focus on the communicative power of language and then consider the role of identity and language.

with respect to the application at hand. It is rare that exposures that serve this purpose are also correctly specified in the sense that they provide a complete description of the causal structure. The results herein show that conventional point estimators perform well even if the exposures are misspecified, provided that the specification errors are only weakly dependent." (Sävje, 2024).

8.1 Untangling match and selection effects

Recall that "communication potential" is associated with better experiences and cohesion outcomes (see Figure 7). To what extent do these patterns reflect match effects–benefits that arise when two individuals have a greater ability to communicate together–versus individual-level differences associated with language skills?

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 might complement future integration efforts. On the other hand, selection effects would require further investigation. For example, some underlying characteristic, such as openness to experience, could influence both language acquisition and cohesion outcomes. A third possibility is that acquiring additional language skills itself increases cohesion, perhaps by encouraging perspective-taking or greater exposure to other groups. In that case, language instruction could improve cohesion even apart from its effect on the quality of a particular match.

We find evidence for both match and selection effects. Figure 6 shows the estimated average treatment effects of high match or low match compared to the co-ethnic control.[54] Language match has a clear impact on how students react to and experience working with a different ethnicity peer. Students report better experiences in high-match pairs than in low-match pairs (95% CI [0.05, 0.55]) and communicate more (95% CI [0.24, 0.80]). The evidence for match effects is weaker for cohesion measures (95% CI [-0.01, 0.23]).

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 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, we estimate regressions of outcome indices against the probability of being assigned to a "high language match" ($\pi(1,1)$). Each plot in Figure 8 has a separate line for $Y(0)$, $Y(1,0)$, and $Y(1,1)$ allowing us to

[54] Appendix Figure A3 compares the estimation of the generalized random forest with Hájek (IPW) estimator with and without regression adjustment. The Hájek estimator uses propensity scores generated from re-randomization of the assignment process without any covariate adjustment. The results are similar to those obtained with the causal forest procedure, though noisier.

compare how potential outcomes vary over the distribution of types (at least in terms of the line of best fit). Recall that we normalize outcomes based on the control group distribution, so $Y(0)$ (solid red lines) is centered around 0 by construction.

Let us examine the plots in Figure 8 by first comparing levels and then slopes. In theory, the differences in levels integrate to the average treatment effects (shown in Figure 6).[55] For example, the difference between Y(1,0) (low match, green dashed lines) and Y(0) (control, red solid lines) in the communication index varies between -0.6 and -0.3 which averages to a bit over -.5, which is roughly the estimated ATE for high match pairs using the communication index.[56] The gaps between high match and low match lines show potential language match effects are consistently present throughout the distribution.

The slopes of the lines help clarify the selection effects. Steep upward sloping lines reflect positive selection, i.e. potential outcomes are increasing in the propensity score. Flat lines suggest constant effects with respect to the propensity score. Notice that the potential outcomes for co-ethnic pairs (i.e. $Y(0)$ or red lines) tend to be flatter. This pattern suggests that being a higher type is not associated with better outcomes i.e. outcomes for students in same ethnicity pairs bares little relationship to the probability of being in a high communication potential match. The cross-ethnic pairs tend to have positive slopes so that high propensity types have better treatment effects. This pattern suggests "selection on gains" for communication and experience. These patterns informed our interpretation of the results in Figure 6 described above.

The results suggest room for "benign self-selection" into inter-ethnic contact in the real world. Students with stronger propensity to a high language match, have higher non-pecuniary benefits and cohesion gains and so may be more likely to join an integration activity. With more high propensity types participating, the probability of a high match increases further and so during opt-in based integration activities, we likely see high matches among high propensity types. This evidence predicts that in smaller studies, where individuals select into participation, the estimated effects of contact will appear particularly high, and that is what meta analyses show (Paluck et al., 2019).[57]

To understand underlying heterogeneity, we turn to the personalized estimates of treatment effects β_i^H and

[55] The integration is with respect to the distribution of propensity scores, which is relatively flat.
[56] See the estimated "Low match" effect for the Experience Index in Figure 6 (blue circle).
[57] Alternatively, in smaller studies, the effort made by implementers might be particularly high.

β_i^L provided by causal forests. For cohesion measures, we care about sample characteristics such as the range and averages of the personalized treatment effects.[58] The range of treatment effects on the cohesion index for a low match is between -0.04 and .18 standard deviations. About 5.6 percent of the low-match point estimates are less than 0. For a high match, the estimates range from .09 to .34. These results suggest that for the majority of students contact has non-negative effects and so in this setting contact does not appear to backfire.

Finally, the correlation between propensity scores and baseline covariates can shed light on what underlying heterogeneity matters in this context. High propensity types look different than low propensity types. Tables A1 and A2 show the correlations between with the overlap sample which is defined based on the values for $\pi(1,1)$ where we see overlap in the data (between .04 and .95). For this sample, we report characteristics for pairs whose probability of a high language match approaches zero (Intercept), and for pairs whose probability of a high language match approaches 1 (Slope). In the overlap sample, the high types are more likely to be fluent in English, to have attended a middle school in Gostivar (a proxy for being from the town rather than villages or elsewhere), to plan to go to college and to be older. Moreover, the linear relationship suggest high propensity types are more likely to be female, less likely to be Turkish, and better at a local non-mother tongue language. At baseline, they also report higher sense of closeness to out groups in the Inclusion of Other in Self measure. Nevertheless, they are no more likely to have out-group friends and other behavioral measures, such as dictator game bias do not have a significant linear relationship with the propensity score (see Table A2).

8.2 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 (Allport et al., 1954) (to the best of our knowledge, this has not been formally tested). Further, language skill differences could upset this balance, creating a power dynamic that may cause backlash (Fouka, 2020). This is particularly concerning in this setting, given the fraught history of Macedonian as a second language (see Section 2).

[58] The personalized treatment effects are quite noisy individually because the data is small and the signal is not precise. Using sample splitting, we can learn about the quality of our estimates. The idea is that we hold out a portion of the sample when fitting the model and then can use predictions on the hold out sample to test the model quality. The forests do a good job estimating the average treatment effect, even though individual heterogeneity measures are not as good out of sample.

We address these issues with three sets of tests, which take the form of defining match-types based on the individual's relationship to the language that they and their partner have the highest communication potential in. There are three relationships within a pair. For each of relationship, we estimate $\beta^{Type} = E[Y(Type) - Y(0)]$ using the AIPW estimator from the causal forest.[59] First, a participant can be randomly assigned a partner with whom the highest communication potential is their *home language*. These effects are captured with $\beta^{Home} = E[Y(Home) - Y(0)]$. Second, the participant can be assigned a partner with whom their highest communication potential is with their partner's home language; we call these *second language*. These effects are captured with $\beta^{Second} = E[Y(Second) - Y(0)]$. Finally, there are pairs who match on a mutual second language *lingua franca*. The use of "communication potential" rather than actual language used gives these estimates an intent-to-treat interpretation. These effects are captured with $\beta^{LinguaFranca} = E[Y(LinguaFranca) - Y(0)]$.

Home language effect: Does using the home language lead to lower ingroup bias and willingness to pay for a co-ethnic partner relative to using other languages? The affirmative to this question is addressed with the hypothesis:

$$\textbf{H1: } \beta^{Home} > \beta^{Second} \text{ and } \beta^{Home} > \beta^{LinguaFranca}$$

There is evidence for a home language effect—-students who speak their own home language have higher cohesion outcomes ($\beta^{Home} - \beta^{Second}$: 95% CI [-0.049, 0.344]; $\beta^{Home} - \beta^{LinguaFranca}$: 95% CI [0.094, 0.391]).[60] They also report better communication and experiences and rate their partners as better at the game. Drilling down into sub-measures, the cohesion differences come from dictator game sharing. Home-language speakers' dictator bias is roughly 0.3 standard deviations below the co-ethnic-pair mean. In levels, the bias falls from 19.7 Denar for control students to 3.6 Denar for home-language speakers.[61] The results concord with recent psychology literature that show positive mother tongue effects when making cross-ethnic peace proposals (Grant et al., 2023).

[59] More formally, we estimate $\beta_i^{Type}(X_i)$ conditional on baseline covariates and propensity scores derived from the randomization procedure and then aggregate the personalized estimates to an average ITT effect.

[60] These are two-sided t-tests and use the same sample of individuals for whom we have non-zero probability of observing each of the potential outcome.

[61] These are raw means.

Is there backlash?: To test for backlash, we test the hypothesis that

$$\textbf{H2: } \beta^{SecondLanguage} < 0$$

.

While second-language speakers have significantly lower outcomes than home-language speakers, the evidence for "backlash" is slim. These students still have positive cohesion outcomes on average, though the two-sided confidence interval covers 0 (β^{Second}: 95% CI [-0.174, 0.277]).

English and Macedonian as lingua franca: If equal footing matters, then using a *lingua franca*—a shared second language—should yield desirable outcomes. This yields the hypotheses:

$$\textbf{H3: } \beta^{LinguaFranca} > 0$$

.

$$\textbf{H3 Strong: } \beta^{LinguaFranca} - \frac{\beta^{Home} + \beta^{Second}}{2} > 0$$

H3 says that it is beneficial to use a *lingua franca* for contact and H3 Strong says it is better to use a shared second language than some student's home language. We fail to reject the null in both cases.[62] Pairs who speak a *lingua franca* have similar communication potential, on average, to those who speak a second language with a home language speaker. These results 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 match 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 match with partners with their highest communication potential in either English or Macedonian. We define treatment effects where the potential outcomes are now indexed by the language of the language with the top communication potential.[63]

[62] For the cohesion index: H3: 95% CI [-.229, .143], H3 Strong: [-.307, .025]

[63] E.g. $\beta^{English} = E[Y(English) - Y(0)]$

and test the hypothesis

$$\textbf{H4: } \beta^{English} > \beta^{Macedonian}$$

This hypothesis is motivated by recent history that suggests that English would be preferred for contact. Figure 10 shows that we cannot reject the null hypothesis. Differences in communication, experience, and cohesion results are indistinguishable from 0.[64] Students using English with a partner have productivity gains. Directionally, teachers' predictions that English would outperform Macedonian holds for communication and experience indices. However, cohesion results are indistinguishable.

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 in-group/out-group bias measured in dictator games rather than through the 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.[65] 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. To sketch the idea, we create an edge between each cross-ethnic pair within a strata; then, we calculate the "communication potential", 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 $w_{ij} = \beta_i^H + \beta_j^H$.[66] 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 using Edmond's blossom algorithm (Edmonds, 1965).[67] Since we run the algorithm within strata, the language skills of the students within the strata constrain the

[64] We focus on the overlap sample with some probability of being assigned a top match in English or Macedonian.
[65] Video game events occur in the high schools, though they do not seem to be focused on integration, *per se*.
[66] These treatment effect estimates come from the causal forests
[67] In Appendix F, we describe the process in greater detail.

feasible matches (i.e. not everyone will get a high match in the optimal matching). We compare the optimal matching to the expected cohesion gains of random matching, and the infeasible scenario where every pair has a high communication potential or every pair has a low communication potential. One downside to this exercise is that the individual estimates are predictions that are potentially biased due to bias-variance trade offs. In a complementary exercise, we use augmented inverse propensity weighted (AIPW) scores which are mean unbiased by design, but suffer from high variance at the individual level. To make progress, we look for matchings that maximize (or minimize) the number of high language matches and use the share of matches to weigh a convex combination of the average score for low-matches and the average score for high matches.

Table **??** shows the range of estimates under various scenarios. Compared to the random benchmark, optimal matching gives a 26 to 33 percent boost to cohesion gains (and more than doubles the worst-case feasible match). These are small gains – on average, in levels for WTP it's roughly 5 denar or an extra 2 denar decline in ingroup bias in the dictator game). Nevertheless, this shows that educators with an eye toward integration may be able to 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.

10 Conclusion

We randomly assign students partners of the same or a different ethnicity to provide new evidence on the role of language and identity in inter-ethnic contact. In line with previous work, we find a moderate positive impact on cohesion outcomes after inter-ethnic contact. But these benefits come with non-pecuniary costs and while we cannot rule out no productivity costs, we also cannot rule out relatively large lose in productivity. We then bring in language: inter-group contact has better cohesion outcomes and lower costs for students assigned partners with whom they have high communication potential. These correlations are relevant in a world of endogenous friendship formation and "benign self-selection". Supposing cross-ethnic pairs meet with some probability and determine whether to match based on expected utility, then the relevant estimate is the unadjusted estimates for high-match pairs in the full sample.

Language plays a role as a communication tool and also conveyor of identities. We cannot rule out match effects based on communication potential or that heterogeneity reflects "selection"–individuals with a higher probability of a high communication potential match are more receptive to inter-group contact. Further, we find the strongest cohesion and experience outcomes are associated with having the highest communication potential in one's home language in a mixed-ethnicity pair.

Nevertheless, after inter-group contact, the average willingness to pay for segregation is still 160 denar, nearly the expected payout from participating in the game. Under strong assumptions, the counterfactual exercise suggests that improving language skills would lead to 45 percent larger cohesion gains, but in practice, the average willingness to pay would decrease by less than 10 denar. This analysis highlights that cohesion cannot be easily built from short one-off experiences. Without repeated exposure, preferences for homophily are too strong.

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 inter-ethnic contact. The causal arrow likely goes in both directions. Recent literature highlights the linguistic impact of integration. Rao (2019) find wealthier students in India substitute away from English when integrated with poorer students who tend to exclusively use Hindi at home. Alan et al. (2021) and Boucher et al. (2024) find Turkish skills increase for Syrian refugees after participating in the perspective-taking course or pre-school camp. Moreover, in Iraq, Christian soccer players start speaking Arabic more when Muslim soccer players are added to their team (Mousa, 2020). We provide a point-in-time snapshot of how schools might create cohesive communities, but language and integration may form a positive feedback loop whereby repeated cooperative inter-group contact becomes self-reinforcing.

Shared language alone does not create cohesive communities. Early proponents of nationalism presupposed that language was necessary and perhaps sufficient to form a cohesive state (Anderson, 1983).[68] During the negotiations to end World War I, Woodrow Wilson and his academic advisors advocated for the creation of the first Yugoslavia writing that: "the political and linguistic affiliations of the peoples whom it is proposed to unite in a single state are so well known that a detailed argument is unnecessary." (David Hunter Miller, 1924) Nevertheless, these same "peoples" took arms against one another during World War II and again at the

[68] See Chapter 5.

dissolution of the second Yugoslavia in the 1990s. Shared language is not a cure-all. Rather, shared language reduces the costs of integration–the benefits still must be high enough for integration to be worthwhile. 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.

References

Akerlof, George A. and Rachel E. Kranton (Aug. 1, 2000). "Economics and Identity*". In: *The Quarterly Journal of Economics* 115.3, pp. 715–753. DOI: 10.1162/003355300554881.

Alan, Sule et al. (Nov. 1, 2021). "Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking*". In: *The Quarterly Journal of Economics* 136.4, pp. 2147–2194. DOI: 10.1093/qje/qjab009.

Alesina, Alberto and Eliana La Ferrara (Sept. 2005). "Ethnic Diversity and Economic Performance". In: *Journal of Economic Literature* 43.3, pp. 762–800. DOI: 10.1257/002205105774431243.

Allport, Gordon Willard, Kenneth Clark, and Thomas Pettigrew (1954). "The nature of prejudice". In: Publisher: Addison-wesley Reading, MA.

Anderson, Benedict (1983). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso.

Angrist, Joshua D. and Victor Lavy (1997). "The Effect of a Change in Language of Instruction on the Returns to Schooling in Morocco". In: *Journal of Labor Economics* 15.1. Publisher: [University of Chicago Press, Society of Labor Economists, NORC at the University of Chicago], S48–S76.

Anisfeld, Ari and Pepi Pandiloski (2024). "Social Learning in Segregated Networks". In: *Working Paper*.

Aron, Arthur, Elaine N. Aron, and Danny Smollan (1992). "Inclusion of Other in the Self Scale and the structure of interpersonal closeness". In: *Journal of Personality and Social Psychology* 63.4. Place: US Publisher: American Psychological Association, pp. 596–612. DOI: 10.1037/0022-3514.63.4.596.

Aronow, Peter M. and Cyrus Samii (Dec. 1, 2017). "Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment". In: *The Annals of Applied Statistics* 11.4. DOI: 10.1214/16-AOAS1005. arXiv: 1305.6156[math,stat].

Athey, Susan, Julie Tibshirani, and Stefan Wager (Apr. 5, 2018). *Generalized Random Forests*. DOI: 10.48550/arXiv.1610.01271. arXiv: 1610.01271[econ,stat].

Athey, Susan and Stefan Wager (2019). "Estimating Treatment Effects with Causal Forests: An Application". In: *Observational Studies* 5.2, pp. 37–51. DOI: 10.1353/obs.2019.0001.

Atkin, David, Eve Colson-Sihra, and Moses Shayo (Apr. 2021). "How Do We Choose Our Identity? A Revealed Preference Approach Using Food Consumption". In: *Journal of Political Economy* 129.4. Publisher: The University of Chicago Press, pp. 1193–1251. DOI: 10.1086/712761.

Bazzi, Samuel et al. (Nov. 1, 2019). "Unity in Diversity? How Intergroup Contact Can Foster Nation Building". In: *American Economic Review* 109.11, pp. 3978–4025. DOI: 10.1257/aer.20180174.

Benatov, Joy, Rony Berger, and Carmit T. Tadmor (Jan. 1, 2021). "Gaming for peace: Virtual contact through cooperative video gaming increases children's intergroup tolerance in the context of the Israeli–Palestinian conflict". In: *Journal of Experimental Social Psychology* 92, p. 104065. DOI: 10.1016/j.jesp.2020.104065.

Bloodworth, Aryn (July 1, 2020). "Educational (de)segregation in North Macedonia: The intersection of policies, schools, and individuals". In: *European Educational Research Journal* 19.4. Publisher: SAGE Publications, pp. 310–328. DOI: 10.1177/1474904120907723.

Blouin, Arthur and Sharun W. Mukand (June 2019). "Erasing Ethnicity? Propaganda, Nation Building, and Identity in Rwanda". In: *Journal of Political Economy* 127.3. Publisher: The University of Chicago Press, pp. 1008–1062. DOI: 10.1086/701441.

Boisjoly, Johanne et al. (Dec. 2006). "Empathy or Antipathy? The Impact of Diversity". In: *American Economic Review* 96.5, pp. 1890–1905. DOI: 10.1257/aer.96.5.1890.

Boucher, Vincent et al. (2024). "Ethnic Mixing in Early Childhood: Evidence from a Randomized Field Experiment and a Structural Model". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3823627.

Bursztyn, Leonardo et al. (Feb. 2024). "The Immigrant Next Door". In: *American Economic Review* 114.2, pp. 348–384. DOI: 10.1257/aer.20220376.

Cappé, O et al. (Dec. 2004). "Population Monte Carlo". In: *Journal of Computational and Graphical Statistics* 13.4, pp. 907–929. DOI: 10.1198/106186004X12803.

Chen, Yan et al. (Mar. 1, 2014). "Which hat to wear? Impact of natural identities on coordination and cooperation". In: *Games and Economic Behavior* 84, pp. 58–86. DOI: 10.1016/j.geb.2013.12.002.

Churchill, Ward (Nov. 1, 2004). *Kill the Indian, Save the Man*. San Francisco, Calif: City Lights Publishers - City Lights Publishers. 158 pp.

Chyn, Eric (Oct. 2018). "Moved to Opportunity: The Long-Run Effects of Public Housing Demolition on Children". In: *American Economic Review* 108.10, pp. 3028–3056. DOI: 10.1257/aer.20161352.

Corno, Lucia, Eliana La Ferrara, and Justine Burns (Dec. 2022). "Interaction, Stereotypes, and Performance: Evidence from South Africa". In: *American Economic Review* 112.12, pp. 3848–3875. DOI: 10.1257/aer.20181805.

Dale-Olsen, Harald and Henning Finseraas (2020). "Linguistic diversity and workplace productivity - ScienceDirect". In: *Labour Economics* 64.

David Hunter Miller (1924). *My Diary at the Conference of Pairs. Volume IV. Documents 216-304.*

Drake, Marshall et al. (2022). *Bayesian Adaptive Choice Experiments*.

— (Feb. 2024). "Bayesian Adaptive Choice Experiments". In: *Working Paper*.

Durlauf, Steven N. and Yannis M. Ioannides (Sept. 4, 2010). "Social Interactions". In: *Annual Review of Economics* 2 (Volume 2, 2010). Publisher: Annual Reviews, pp. 451–478. DOI: 10.1146/annurev.economics.050708.143312.

Edmonds, Jack (1965). "Paths, Trees, and Flowers". In: *Canadian Journal of Mathematics* 17, pp. 449–467. DOI: 10.4153/CJM-1965-045-4.

Elwert, Felix, Tamás Keller, and Andreas Kotsadam (May 2023). "Rearranging the Desk Chairs: A Large Randomized Field Experiment on the Effects of Close Contact on Interethnic Relations". In: *American Journal of Sociology* 128.6. Publisher: The University of Chicago Press, pp. 1809–1840. DOI: 10.1086/724865.

Fan, Samantha P. et al. (2015). "The exposure advantage: Early exposure to a multilingual environment promotes effective communication". In: *Psychological Science* 26.7. Place: US Publisher: Sage Publications, pp. 1090–1097. DOI: 10.1177/0956797615574699.

Finseraas, Henning et al. (May 1, 2019). "Trust, ethnic diversity, and personal contact: A field experiment". In: *Journal of Public Economics* 173, pp. 72–84. DOI: 10.1016/j.jpubeco.2019.02.002.

Fontana, Giuditta (2017). *Education Policy and Power-Sharing in Post-Conflict Societies*. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-31426-6.

Fouka, Vasiliki (Jan. 1, 2020). "Backlash: The Unintended Effects of Language Prohibition in U.S. Schools after World War I". In: *The Review of Economic Studies* 87.1, pp. 204–239. DOI: 10.1093/restud/rdz024.

Ghosh, Arkadev et al. (2024). "Creating Cohesive Communities: A Youth Camp Experiment in India". Working Paper. Working Paper.

Ginsburgh, Victor and Shlomo Weber (June 1, 2020). "The Economics of Language". In: *Journal of Economic Literature* 58.2, pp. 348–404. DOI: 10.1257/jel.20191316.

Goeree, Jacob K. et al. (Feb. 2010). "The 1/d Law of Giving". In: *American Economic Journal: Microeconomics* 2.1, pp. 183–203. DOI: 10.1257/mic.2.1.183.

Grant, Leigh H., Ifat Maoz, and Boaz Keysar (May 1, 2023). "Lingua Franca as a Hidden Barrier to Conflict Resolution". In: *Journal of Conflict Resolution* 67.5. Publisher: SAGE Publications Inc, pp. 979–1006. DOI: 10.1177/00220027221123316.

Guillouët, Louise et al. (Aug. 22, 2024). "Language Barriers in Multinationals and Knowledge Transfers". In: *The Review of Economics and Statistics*, pp. 1–56. DOI: 10.1162/rest_a_01487.

Hedegaard, Morten Størling and Jean-Robert Tyran (Jan. 2018). "The Price of Prejudice". In: *American Economic Journal: Applied Economics* 10.1, pp. 40–63. DOI: 10.1257/app.20150241.

Heugh, Kathleen (Mar. 2013). "Multilingual Education Policy in South Africa Constrained by Theoretical and Historical Disconnections". In: *Annual Review of Applied Linguistics* 33, pp. 215–237. DOI: 10.1017/S0267190513000135.

Hjort, Jonas (Nov. 1, 2014). "Ethnic Divisions and Production in Firms *". In: *The Quarterly Journal of Economics* 129.4, pp. 1899–1946. DOI: 10.1093/qje/qju028.

Hynsjö, Disa and Amy Damon (Aug. 1, 2016). "Bilingual education in Peru: Evidence on how Quechua-medium education affects indigenous children's academic achievement". In: *Economics of Education Review* 53, pp. 116–132. DOI: 10.1016/j.econedurev.2016.05.006.

Jain, Tarun (June 2017). "Common Tongue: The Impact of Language on Educational Outcomes". In: *The Journal of Economic History* 77.2, pp. 473–510. DOI: 10.1017/S0022050717000481.

Kinzler, Katherine (2020). *How You Say It: Why We Judge Others by the Way They Talk—and the Costs of This Hidden Bias*. Boston: Houghton Mifflin Harcourt.

Kovalčík, Juraj, Magdaléna Švecová, and Michal Kabát (2023). "Viability of Using Digital Games for Improving Team Cohesion: A Systematic Review of the Literature". In: *Acta Ludologica* 6.1, pp. 46–65. DOI: 10.34135/actaludologica.2023-6-1.46-65.

Kumar, Hemanshu, Rohini Somanathan, and Mahima Vasishth (2022). "Language and learning in ethnically mixed communities". In: *Review of Development Economics* 26.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111 pp. 835–846. DOI: 10.1111/rode.12873.

Laitin, David D. and Rajesh Ramachandran (Aug. 2016). "Language Policy and Human Development". In: *American Political Science Review* 110.3, pp. 457–480. DOI: 10.1017/S0003055416000265.

— (May 1, 2022). "Linguistic diversity, official language choice and human capital". In: *Journal of Development Economics* 156, p. 102811. DOI: 10.1016/j.jdeveco.2021.102811.

Lang, Kevin (1986). "A Language Theory of Discrimination". In: *The Quarterly Journal of Economics* 101.2. Publisher: Oxford University Press, pp. 363–382.

Lazear, Edward P. (Dec. 1999). "Culture and Language". In: *Journal of Political Economy* 107 (S6). Publisher: The University of Chicago Press, S95–S126. DOI: 10.1086/250105.

List, Annika J., John A. List, and Anya Samek (Aug. 1, 2017). "Discrimination among pre-school children: Field experimental evidence". In: *Economics Letters* 157, pp. 159–162. DOI: 10.1016/j.econlet.2017.06.006.

Lowe, Matt (June 2021). "Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration". In: *American Economic Review* 111.6, pp. 1807–1844. DOI: 10.1257/aer.20191780.

Marian, Viorica, Henrike K. Blumenfeld, and Margarita Kaushanskaya (Aug. 2007). "The Language Experience and Proficiency Questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals". In: *Journal of speech, language, and hearing research: JSLHR* 50.4, pp. 940–967. DOI: 10.1044/1092-4388(2007/067).

Matsudaira, Jordan D (2004). "Evaluating the Impact of English Immersion versus Bilingual Education on Student Achievement". In.

Merdzanovic, Adis (Jan. 1, 2015). "A Short Postscript on other Cases: Macedonia and Kosovo". In: *Democracy by Decree : Prospects and Limits of Imposed Consociational Democracy in Bosnia and Herzegovina*. Stuttgart: ibidem.

Miguel, Edward (Apr. 2004). "Tribe or Nation? Nation Building and Public Goods in Kenya versus Tanzania". In: *World Politics* 56.3, pp. 327–362. DOI: 10.1017/S0043887100004330.

Mousa, Salma (Aug. 14, 2020). "Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq". In: *Science* 369.6505, pp. 866–870. DOI: 10.1126/science.abb3153.

Nix, Emily and Nancy Qian (Jan. 2015). *The Fluidity of Race: "Passing" in the United States, 1880-1940*. w20828. Cambridge, MA: National Bureau of Economic Research, w20828. DOI: 10.3386/w20828.

Ortakovski, Vladimir (1998). *Minorities in the Balkans*. Skopje: "2-ri August S" - Stip.

Paluck, Elizabeth Levy, Seth A. Green, and Donald P. Green (Nov. 2019). "The contact hypothesis re-evaluated". In: *Behavioural Public Policy* 3.2. Publisher: Cambridge University Press, pp. 129–158. DOI: 10.1017/bpp.2018.25.

Pettigrew, Thomas F. and Linda R. Tropp (2006). "A meta-analytic test of intergroup contact theory." In: *Journal of personality and social psychology* 90.5. Publisher: American Psychological Association, p. 751.

Piper, Benjamin, Stephanie S. Zuilkowski, and Salome Ong'ele (Nov. 2016). "Implementing Mother Tongue Instruction in the Real World: Results from a Medium-Scale Randomized Controlled Trial in Kenya". In: *Comparative Education Review* 60.4, pp. 776–807. DOI: 10.1086/688493.

Police Violence in Macedonia (1998). Vol 10. No 1 (D).

Posner, Daniel N. (June 6, 2005). *Institutions and Ethnic Politics in Africa*. 1st ed. Cambridge University Press. DOI: 10.1017/CBO9780511808661.

Rao, Gautam (Mar. 2019). "Familiarity Does Not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools". In: *American Economic Review* 109.3, pp. 774–809. DOI: 10.1257/aer.20180044.

Sävje, F (Mar. 1, 2024). "Causal inference with misspecified exposure mappings: separating definitions and assumptions". In: *Biometrika* 111.1, pp. 1–15. DOI: 10.1093/biomet/asad019.

Scacco, Alexandra and Shana S. Warren (Aug. 2018). "Can Social Contact Reduce Prejudice and Discrimination? Evidence from a Field Experiment in Nigeria". In: *American Political Science Review* 112.3. Publisher: Cambridge University Press, pp. 654–677. DOI: 10.1017/S0003055418000151.

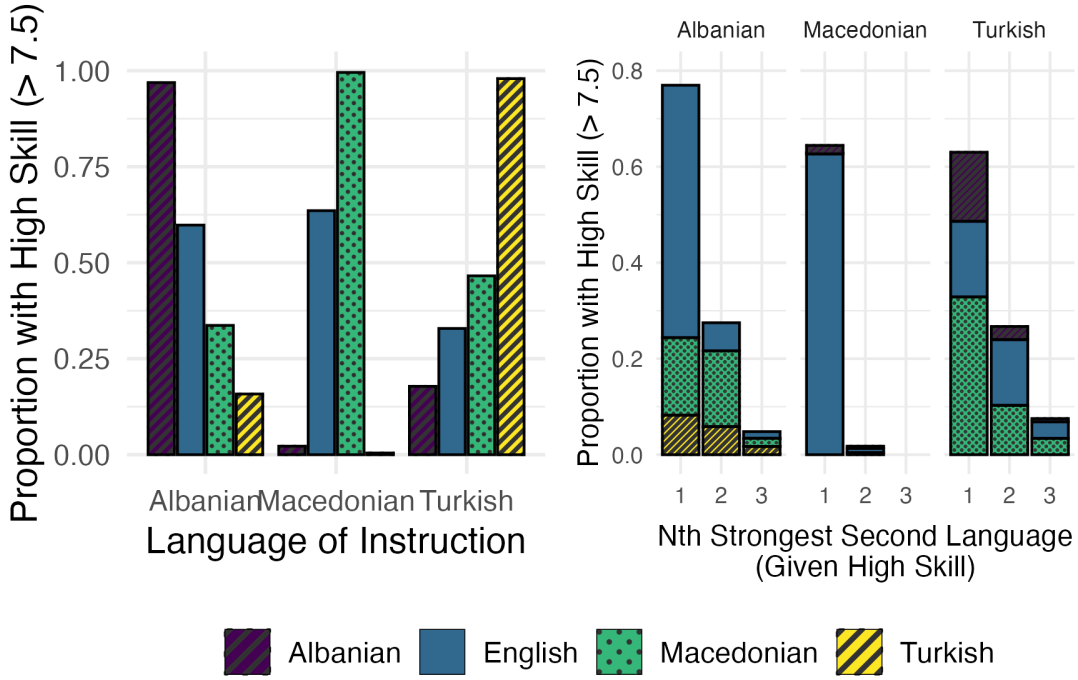
Shayo, Moses (Aug. 2, 2020). "Social Identity and Economic Policy". In: *Annual Review of Economics* 12.1, pp. 355–389. DOI: 10.1146/annurev-economics-082019-110313.

Tajfel, Henri et al. (1971). "Social categorization and intergroup behaviour". In: *European Journal of Social Psychology* 1.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2420010202, pp. 149–178. DOI: 10.1002/ejsp.2420010202.

Velez, John A et al. (Oct. 1, 2021). "A Recipe for Success: The Effect of Dyadic Communication and Cooperative Gameplay on Subsequent Non-gaming Task Performance". In: *Human Communication Research* 47.4, pp. 364–386. DOI: 10.1093/hcr/hqab008.

11 Figures

Figure 1: Second Language Skills by Language of Instruction

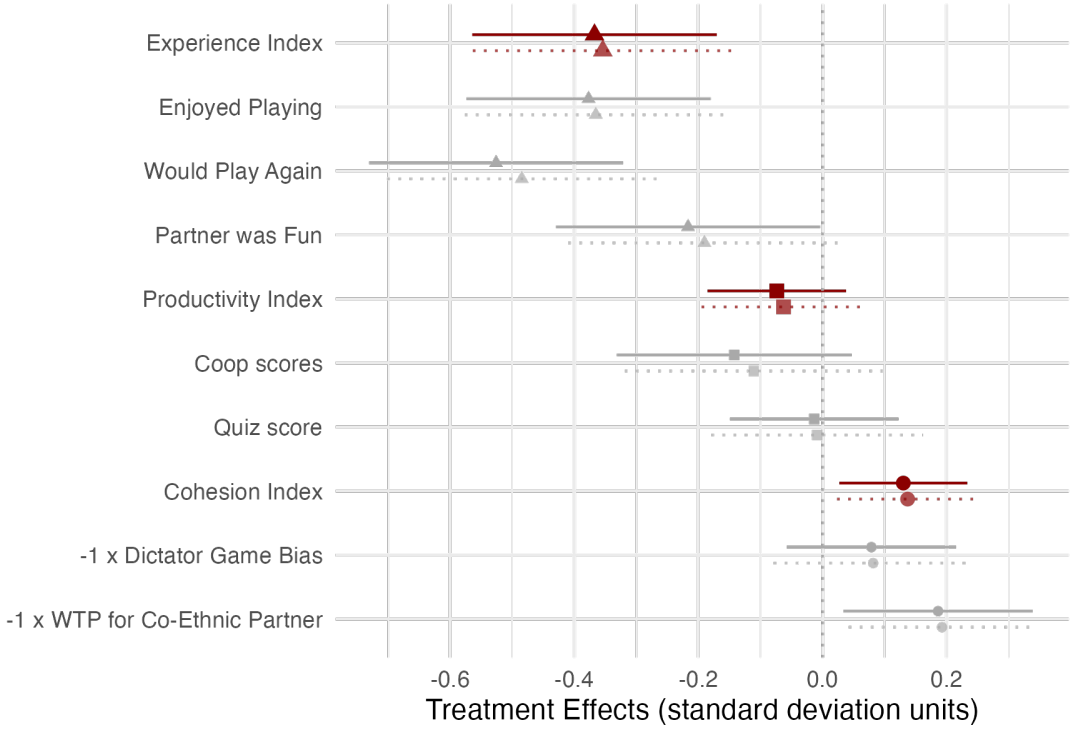


Notes: On the vertical axis both figures show the proportion of students within a language of instruction group who meet some criteria. The left plot shows self-reported language proficiency for the three local languages and English. Proficiency is defined as self-reporting an average of 7.5 fluency on a scale of 0 to 10 averaged over the four language domains: reading, writing, speaking, and listening. The left plot shows how those languages rank within student skill sets. The sample consists of students who completed the baseline survey and participated in the video game session. The sample is 44 percent Albanian, 34 percent Macedonian and 22 percent Turkish.

Figure 2: Screenshots from “Overcooked! 2” and participants playing a cooperative level



Figure 3: Treatment Effects: Assigned Cross-Ethnic Partner



Notes: Average treatment effects of having a different ethnicity partner compared to having a same ethnicity partner. Calculated with causal forests with covariates listed in paper (solid lines) and with regression with no covariates (dotted lines). All sub-outcomes (grey) are normalized around the control mean and control standard deviation. Indices (red) are simple averages of normalized sub-outcomes. Cohesion measures are multiplied by negative one so that positive numbers reflect gains and negative numbers loses. Standard errors clustered by strata. The sample consists of all students who participated in the video game sessions. For some sub-measures, students do not answer the questions and so the sample is restricted.

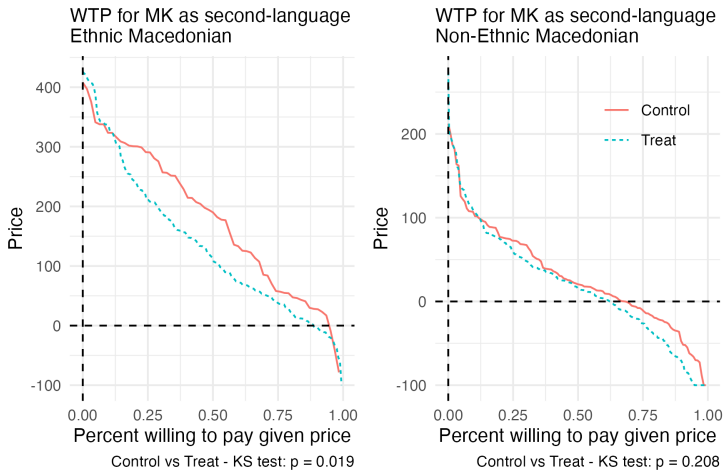
Outcomes: Amount communication reported by two facilitators on 7-point Likert scale. Communication Easy, Enjoyed Playing, Would Play Again, Partner Fun, and Partner Good reported by students on post-play survey on 5-point Likert scale. Cooperative video game scores come from the average scores of partners in the incentivized rounds of play of "Overcooked! 2". The mean solo scores normalized is the average of 3 solo rounds. The quiz score was derived from a short quiz on the video game rules and controls. Dictator game bias is the average difference in money shared in dictator games with anonymous representative of ingroup compared amount shared with anonymous representative of outgroup. WTP for Co-Ethnic Partner comes from the adaptive choice experiment.

Figure 4: Demand for partner characteristics

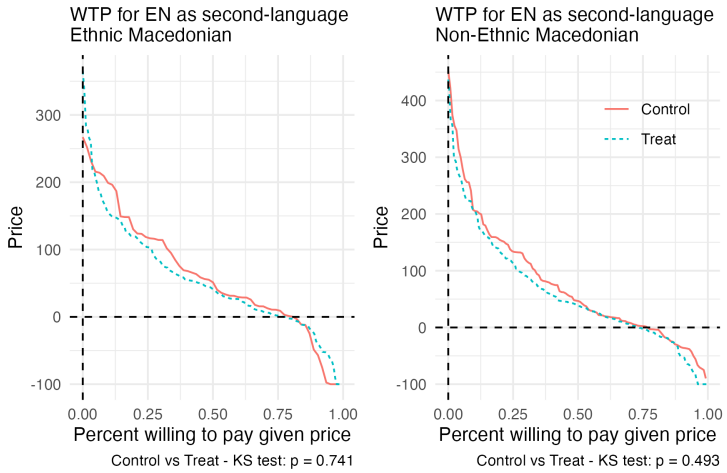
(a) Demand Curve: Co-ethnic Partner



(b) Demand Curve: Macedonian as a Second-language

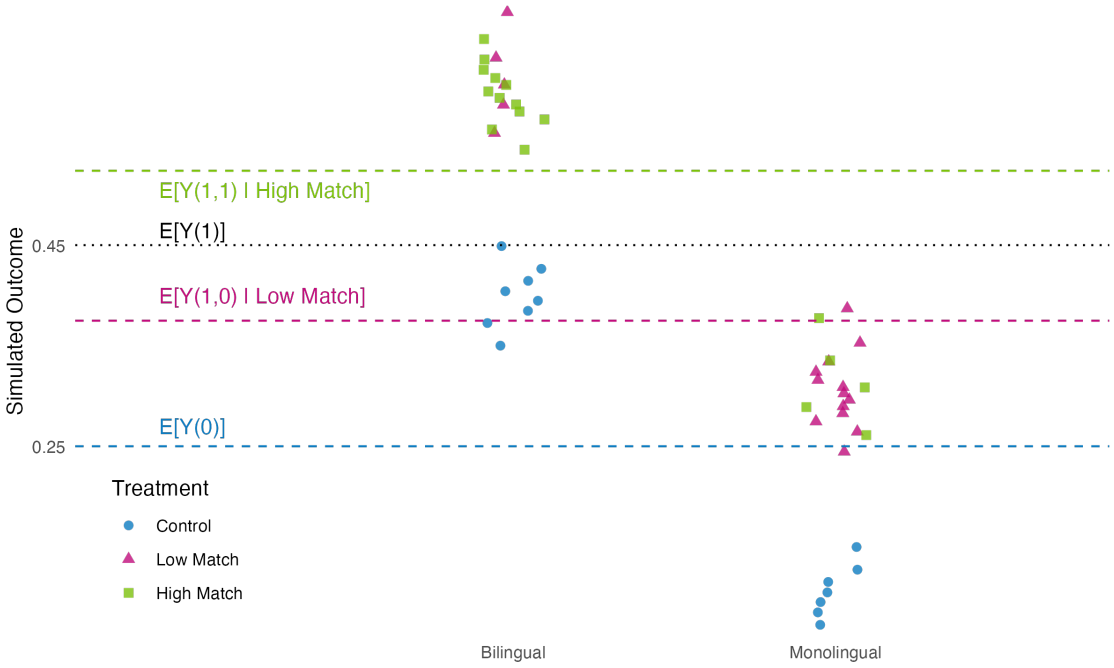


(c) Demand Curve: English as a Second-language



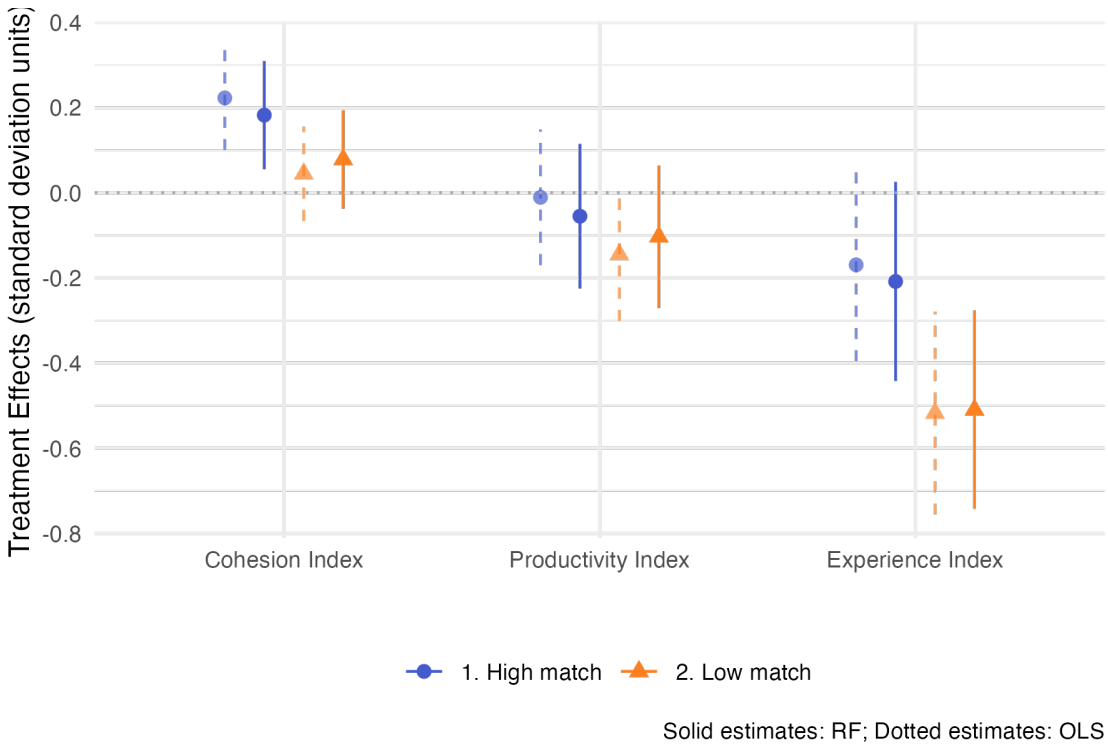
Notes: Demand curves are the cumulative distribution of individual-level willingness to pay estimates. Willingness to pay estimated from adaptive choice experiments in which respondents made binary choices between partners based on ethnicity, second language skills and a sure payout.

Figure 5: Simulation Demonstrating Selection Bias when Sample Splitting



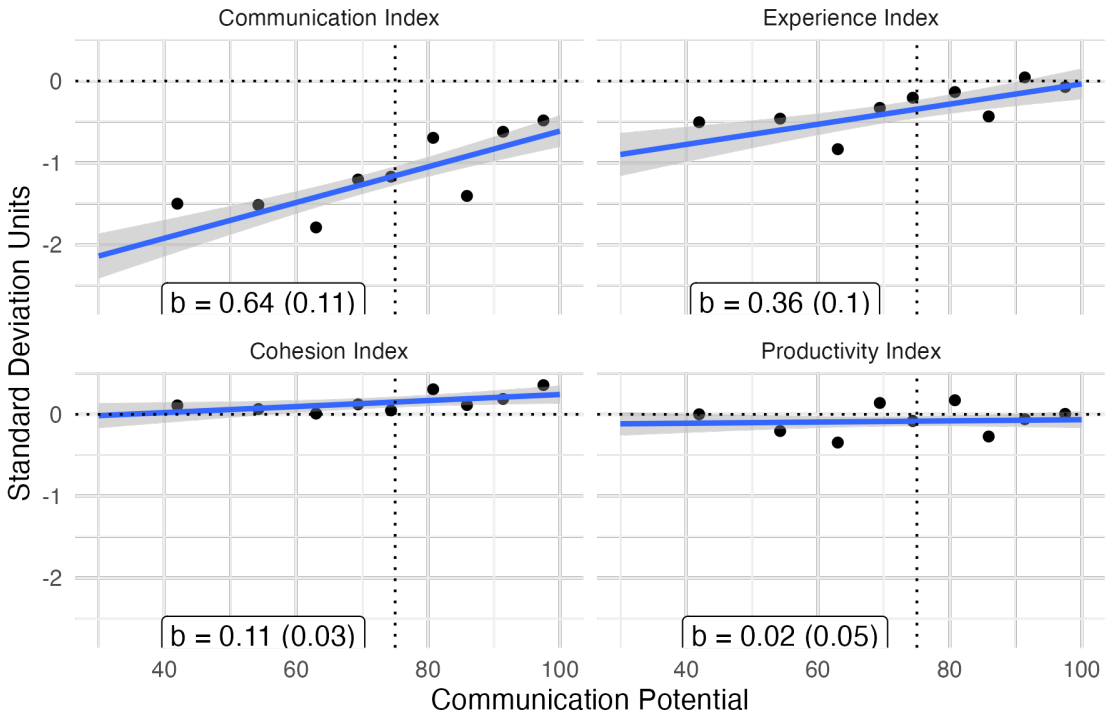
Notes: This stylized simulation shows a stratum with two types of students–bilingual or monolingual–in a trilingual environment. There is a constant treatment effect $\tau = .2$. Each point represents roughly 2000 students randomized into control, high match, or low match as shown with shapes and colors. 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: Heterogeneity by Communication Potential of Cross-Ethnic Pairs



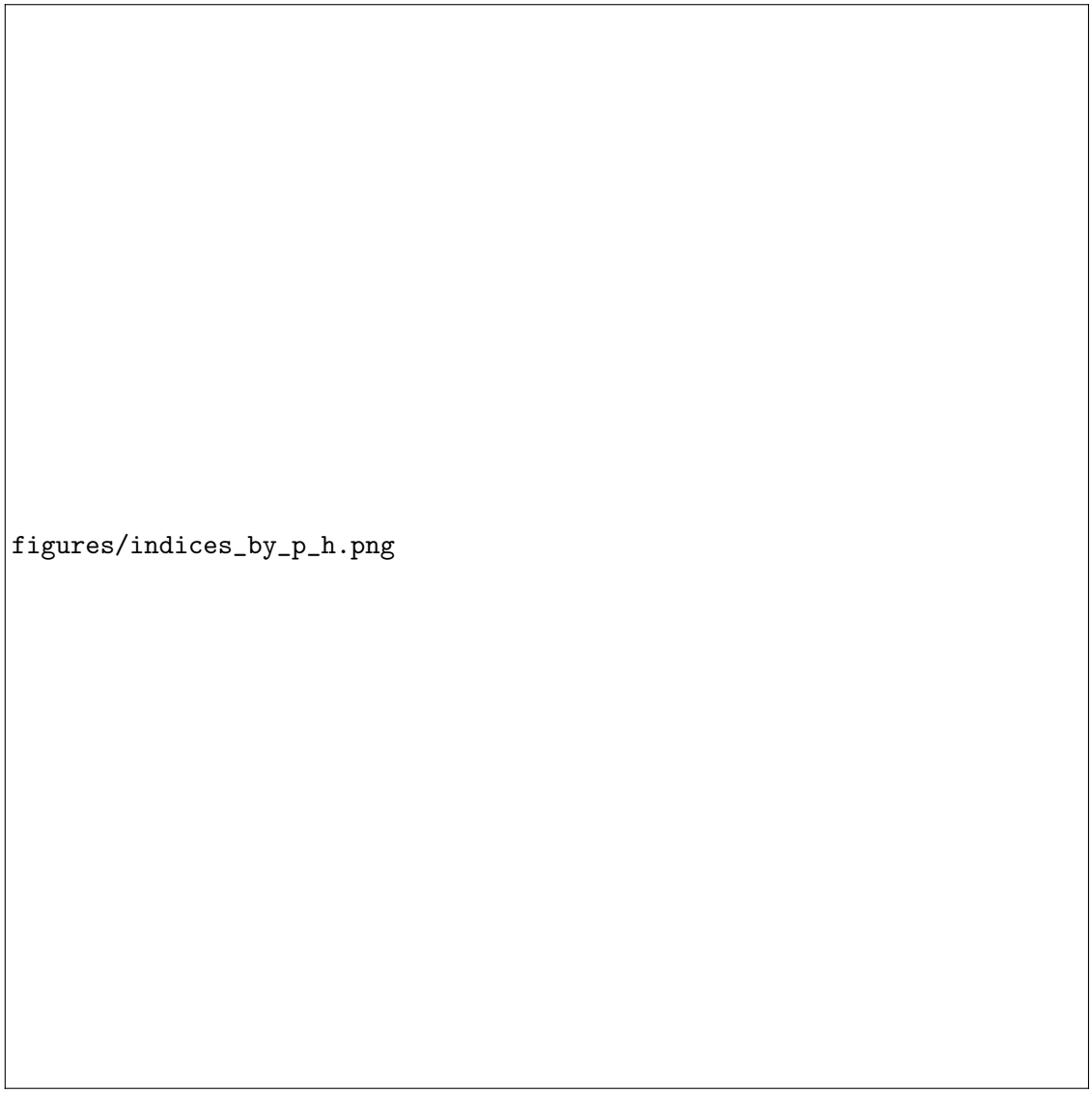
Notes: Estimates of average treatment effects of having a cross-ethnic partner with high or low language match compared to having a co-ethnic partner. Estimated with causal forests with covariates listed in the paper (bold color, solid lines) and with regression with indicators for match type and no covariates (light color, dashed lines). The sample is restricted to ensure overlap, such that students in the sub-sample have a non-zero probability of being in a high or low match. We define a high match as at or above the median "communication potential" score (75), and a low match as below the median. All sub-outcomes are normalized around the control mean and control standard deviation. Indices are simple averages of normalized sub-outcomes.

Figure 7: Associations Between Communication Potential and Outcomes



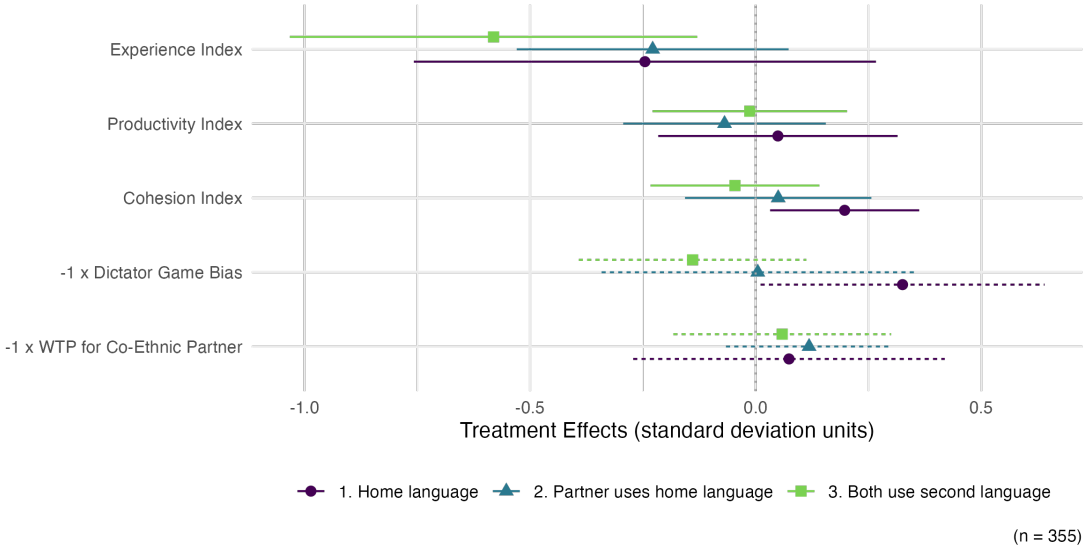
Notes: Lines show linear regression of indices against communication potential. Communication potential is calculated using student's self-reported fluency scores in Albanian, Macedonian, Turkish, and English. Within a pair, the communication potential is the largest geometric mean (squared) of fluency scores (i.e. the maximum of the individual's scores multiplied by each other component-wise).
The regression $\text{Outcome}_i = a + b\frac{\text{Communication Potential}_{ij}}{IQR} + \epsilon_i$. Where IQR is the interquartile range of communication potential. The estimated b shows the average change observed in an outcome when moving from the 25th to 75th percentile of communication potential (from roughly 60 to 90). While partners are random, $\text{Communication Potential}_{ij}$ is a function of endogenous skills; hence these results should be interpreted as associations. b are significantly different from zero for the Cohesion, Communication, Enjoyment Indices, but not the Productivity index. Cooperative video game scores and quiz scores are the components of the productivity index. Points show averages binned by dividing communication potential at quantiles. Standard errors clustered by strata.

Figure 8: Relationship Between Propensity to Have a High Match and Outcomes



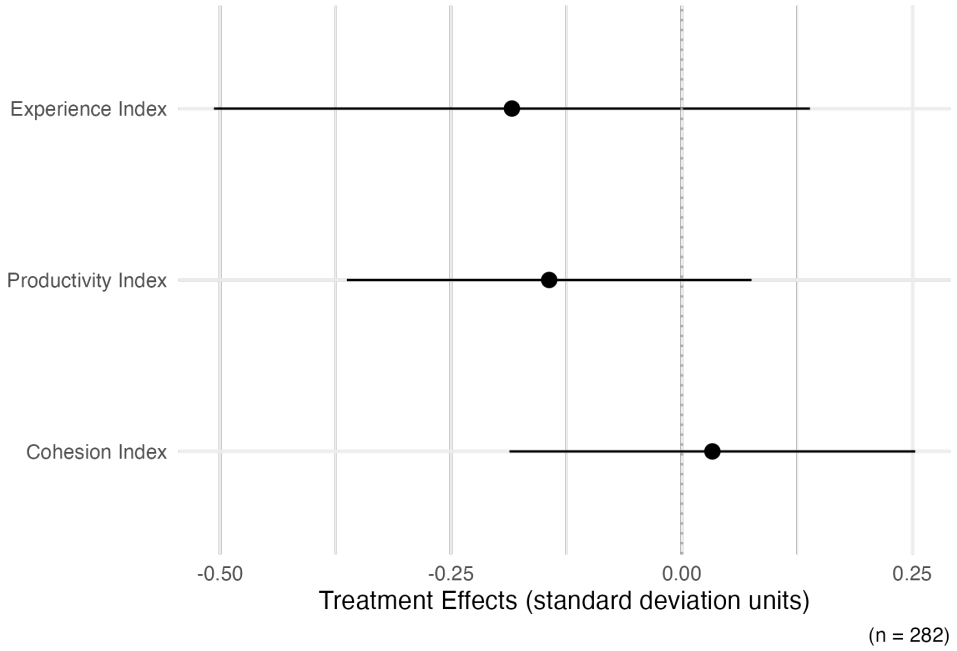
Notes: Lines show linear regression of indices against the probability of getting a high match partner $\pi(1,1)$. The regression Index$_i = \beta_{0,z,l} + \beta_{0,z,l}\pi_i(1,1) + \epsilon_i$ with ($\beta_{0,z,l}$, $\beta_{1,z,l}$) specific to treatment assignment (z,l). Each line shows the best linear projection of potential outcomes $Y(z,l)$ so that steep slopes imply selection and gaps between high match and low match capture match effects.

Figure 9: Home and Second Language Effects



Notes The plot shows treatment effects compared to the control group for students whose 1) top "communication potential" is the student's home language, 2) top "communication potential" is the student's partner's home language, or 3) top "communication potential" is both students' second language (*lingua franca*). Estimated with causal forests with covariates listed in the paper. Points with solid lines are indices and dotted lines are submeasures. The sample is not restricted.

Figure 10: Macedonian versus English as Second Language



Notes The plot shows Albanian and Turkish students' treatment effects when the "communication potential" is in Macedonian compared to English. Thus, negative effects imply matching in Macedonian language results in worse outcomes than English and positive effects imply matching in Macedonian language results in better outcomes. Estimated with causal forests with covariates listed in paper. The sample is restricted so all observations have a positive probability of matching on either English or Macedonian.

12 Tables

Table 1: Demographics: Sample characteristics and balance

Covariate	Sample Mean	(sd)	Control Mean	(se).	Treat Diff	(se)	p-value	N
Female	0.57	0.50	0.57	0.10	-0.01	0.06	0.88	662
Grade 1	0.36	0.48	0.30	0.09	0.08	0.05	0.09	662
Grade 2	0.31	0.46	0.35	0.11	-0.05	0.06	0.36	662
Grade 3	0.20	0.40	0.22	0.08	-0.03	0.04	0.51	662
Grade 4	0.13	0.34	0.13	0.06	0.00	0.02	0.89	662
Albanian LOI	0.44	0.50	0.47	0.04	-0.04	0.03	0.25	662
Macedonian LOI	0.34	0.47	0.33	0.04	0.01	0.03	0.66	662
Turkish LOI	0.22	0.41	0.20	0.04	0.03	0.03	0.39	662
LOI maps to ethnicity	0.96	0.20	0.96	0.01	0.00	0.02	0.92	662
English fluency	7.44	2.27	7.45	0.14	-0.01	0.16	0.97	662
Local language fluency (not LOI)	4.97	3.61	4.99	0.34	-0.03	0.28	0.92	662
Attended Gostivar middle school	0.67	0.47	0.66	0.04	0.02	0.04	0.65	662
Has Cellphone	0.97	0.18	0.95	0.02	0.02	0.01	0.12	662
Parent has college	0.45	0.50	0.39	0.04	0.08	0.04	0.05	662
Plans to go to college	0.88	0.32	0.88	0.02	0.00	0.02	0.96	662
Grade Point Average (previous year)	4.57	0.69	4.50	0.11	0.10	0.09	0.28	658
Absences	41.46	37.17	43.18	3.07	-2.40	3.05	0.44	658
Absences (inexcused)	5.83	9.80	7.06	1.31	-1.71	0.97	0.09	658
Life satisfaction (0-10)	8.46	2.02	8.31	0.23	0.21	0.20	0.29	662
Trust (0-10)	6.67	2.51	6.71	0.25	-0.05	0.23	0.84	662

Notes: Table shows full sample characteristics (Sample Mean), characteristics for pairs who are assigned a same ethnicity partner and have completed the video game session (Control Mean), and the mean differences between control students and students assigned different ethnicity partners and who have completed the video game (Treat Diff) along with associated p-values from a t-test of mean differences where standard errors are clustered by school-grade-gender-date of randomization.
Covariate details 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 with more than one local ethnicity). Self-reported English fluency and Local language fluency for non-LOI are on a scale from 0-10. Language fluency is imputed with a random forest regression for 1 student missing baseline data using information from the recruitment data.

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	5.07	3.07	5.09	0.30	-0.02	0.23	0.94	662
N out-group links named	0.65	0.97	0.62	0.07	0.04	0.07	0.51	662
Closeness to own LOI (baseline)	6.40	1.32	6.40	0.11	0.00	0.11	0.97	662
Closeness to out LOI (baseline)	2.10	1.16	2.06	0.06	0.05	0.07	0.44	662
Dictator game denar to outgroup (baseline)	220.58	127.83	228.86	12.79	-11.56	8.68	0.19	662
Dictator game bias (baseline)	27.34	78.07	29.65	8.44	-3.23	8.07	0.69	662
Enjoys video games	0.70	0.46	0.70	0.05	0.01	0.04	0.89	662
Makes video game rec	0.62	0.49	0.65	0.06	-0.05	0.03	0.20	662
Solo score	479.60	220.18	461.79	28.57	24.87	18.72	0.19	662
Solo score (demeaned)	0.26	212.76	-16.81	26.22	23.84	19.63	0.23	662

Notes: Table shows full sample characteristics (Sample Mean), characteristics for pairs who are assigned a same ethnicity partner and have completed the video game session (Control Mean), and the mean differences between control students and students assigned different ethnicity partners and who have completed the video game (Treat Diff) along with associated p-values from a t-test of mean differences where standard errors are clustered by school-grade-gender-date of randomization.

Covariate details Network data comes from 4 questions where students name 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 (28 percent) use all the slots for friend cliques. Closeness measures come from the Inclusion of Other-in-Self measures (scale from 1-7). Dictator game bias is the difference between the amount of money an individual shares with anonymous ingroup (same ethnicity) recipients compared to what is shared with anonymous outgroup recipients. Enjoys video games equals 1 if the student responded to the question of "How well does 'I love video games' describe you" with 5 or greater on a scale from 0 to 10. Makes video game rec equals 1 if the student names at least one video game when asked for a recommendation, and zero otherwise. Solo score is the average score in solo rounds of "Overcooked! 2". Solo score demeaned demeans the score by gender-specific solo score means.

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	0.0176		-1.77×10^{-15}		-5.86×10^{-15}	
	(0.0647)		(0.0954)		(0.0731)	
Cross-ethnic partner	-1.217***	-1.222***	-1.325***	-1.344***	-1.096***	-1.094***
	(0.1181)	(0.1233)	(0.1372)	(0.1416)	(0.1389)	(0.1461)
Fixed-effects						
school-grade-gender		Yes		Yes		Yes
Fit statistics						
Observations	640	640	662	662	640	640
R^2	0.18001	0.27738	0.19405	0.30424	0.08120	0.15403
Within R^2		0.19078		0.21192		0.08241

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	-0.0120		-2.68×10^{-16}		3.27×10^{-16}		-1.04×10^{-15}	
	(0.0651)		(0.0670)		(0.0749)		(0.0872)	
Cross-ethnic partner	-0.3584***	-0.3540***	-0.3693***	-0.3653***	-0.4965***	-0.4846***	-0.1964*	-0.1903*
	(0.1006)	(0.1070)	(0.1040)	(0.1078)	(0.1089)	(0.1108)	(0.1099)	(0.1122)
Fixed-effects								
school-grade-gender		Yes		Yes		Yes		Yes
Fit statistics								
Observations	650	650	649	649	641	641	642	642
R^2	0.01990	0.11705	0.01358	0.09319	0.02586	0.10556	0.00750	0.09670
Within R^2		0.02027		0.01358		0.02518		0.00729

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 the control mean and control standard deviation. Communication easy and experience sub-outcomes derived from participant survey items on a 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			Quiz score		
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variables									
Constant	1.38×10^{-16}			-1.54×10^{-16}			-1.87×10^{-15}		
	(0.0915)			(0.1361)			(0.0969)		
Cross-ethnic partner	-0.0850	-0.0628	-0.0701	-0.1447	-0.1108	-0.1194	-0.0114	-0.0086	-0.0145
	(0.0575)	(0.0674)	(0.0644)	(0.1072)	(0.1061)	(0.1003)	(0.0645)	(0.0871)	(0.0883)
Mean solo score (Normalized)			0.2303***			0.2714***			0.1847***
			(0.0336)			(0.0335)			(0.0593)
Fixed-effects									
school-grade-gender		Yes	Yes		Yes	Yes		Yes	Yes
Fit statistics									
Observations	662	662	662	662	662	662	662	662	662
R^2	0.00302	0.22719	0.29953	0.00460	0.32136	0.37422	2.83×10^{-5}	0.11335	0.13761
Within R^2		0.00199	0.09542		0.00371	0.08132		1.72×10^{-5}	0.02737

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

Dependent Variables:	Cohesion Index			-1 x Dictator Game Bias			-1 x WTP for Co-Ethnic Partner		
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variables									
Constant	-4.21×10^{-16}			-1.31×10^{-16}			2.76×10^{-16}		
	(0.0492)			(0.0681)			(0.0680)		
Cross-ethnic partner	0.1419**	0.1371**	0.1323**	0.1001	0.0816	0.0745	0.1836**	0.1926**	0.1902**
	(0.0526)	(0.0580)	(0.0588)	(0.0693)	(0.0818)	(0.0829)	(0.0768)	(0.0768)	(0.0766)
Baseline DG Bias			0.0970*			0.1462			0.0478
			(0.0535)			(0.0917)			(0.0492)
Fixed-effects									
school-grade-gender		Yes	Yes		Yes	Yes		Yes	Yes
Fit statistics									
Observations	662	662	662	662	662	662	662	662	662
R^2	0.00816	0.06024	0.07330	0.00214	0.06541	0.08106	0.00674	0.04240	0.04397
Within R^2		0.00754	0.02133		0.00143	0.01815		0.00720	0.00882

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 the control mean and control standard deviation. Cooperative video game scores come from the average scores of partners in the incentivized rounds of play of "Overcooked! 2". The mean solo scores normalized is the average of 3 solo rounds. The quiz score was derived from a short quiz on the video game rules and controls. Dictator game bias derives from the difference between the 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: Counterfactual matchings

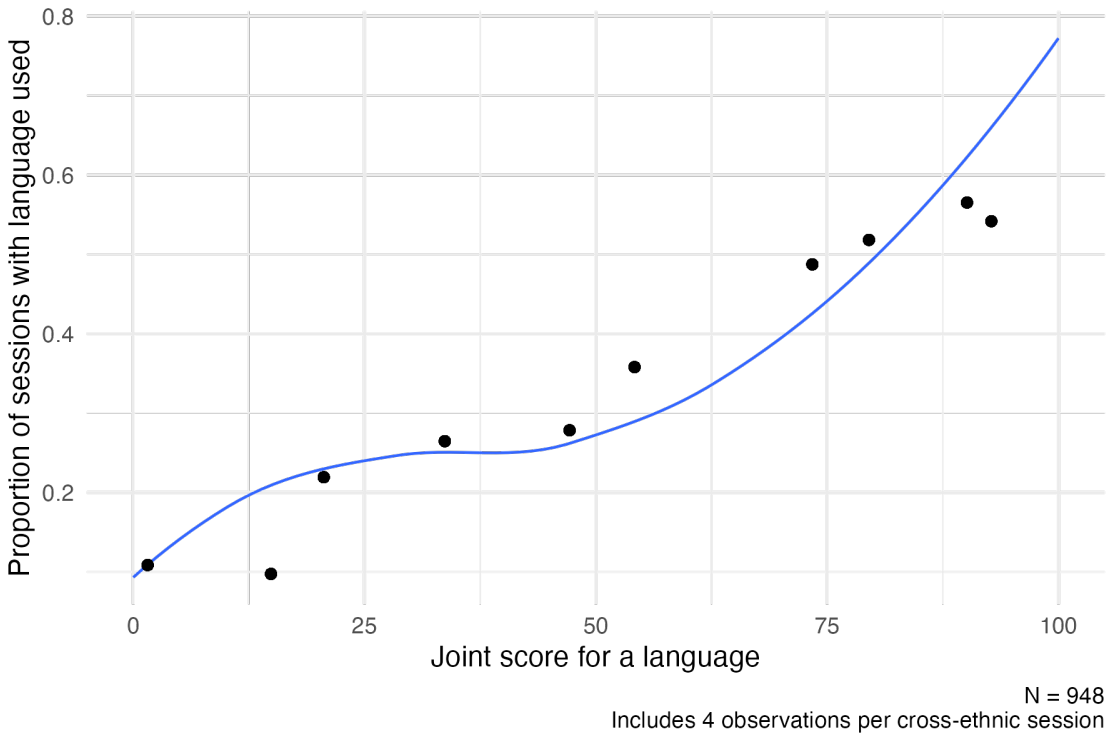
	n	All low match	Min matching	$E[\tau]$	Max matching	All high match
CATEs, in sample	382	0.074	0.092	0.138	0.184	0.199
AIPW $\times$ composition, in sample	382	0.069	0.091	0.124	0.157	0.174
		(0.072)	(0.060)	(0.050)	(0.053)	(0.059)
CATEs, full sample	470	0.070	0.093	0.132	0.173	0.194

Notes: This table presents the range of counterfactual average treatment effects based on feasible and infeasible matchings. In sample refers to students in the overlap sample who were treated. Full sample refers to all student who were treated. Both samples require full covariate coverage for the random forest. **CATES**: 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. Columns labeled "All low match" we take the average of all units' low-match statistics (β_i^L or the related AIPW score) and for "All high match", we average the high-match statistic (β_i^H or the related AIPW score). We approximate $E[\tau]$ from 1000 random feasible matchings; for each randomization r, we calculate a $\tilde{\tau}_r$ and average these to get $E[\tau]$.

AIPW scores: individual scores are unbiased but far too noisy to value any single matching directly, and selecting a matching to maximize a sum of scores would optimize over noise. Matchings reflect composition. "Max matching" selects the feasible within-strata matching that maximizes the *number* of high-match pairs (subject to matching the maximal number of students), and "Min matching" minimizes it; note that selecting the matching requires only language-skill data, not outcome estimates. For a matching with share s of matched students in high-match pairs, the table reports the composition-valued estimate $\tilde{\tau} = s, \bar{\Gamma}^H + (1-s), \bar{\Gamma}^L$, where $\bar{\Gamma}^H$ and $\bar{\Gamma}^L$ are the average AIPW scores, i.e., the pooled ATE estimates for high- and low-match pairing. "All low match" and "All high match" are $\bar{\Gamma}^L$ and $\bar{\Gamma}^H$; $E[\tau]$ applies the same valuation to the average composition of the random matchings. Standard errors (in parentheses) are computed by the delta method, treating the matching shares as fixed and clustering the score means at the strata level.

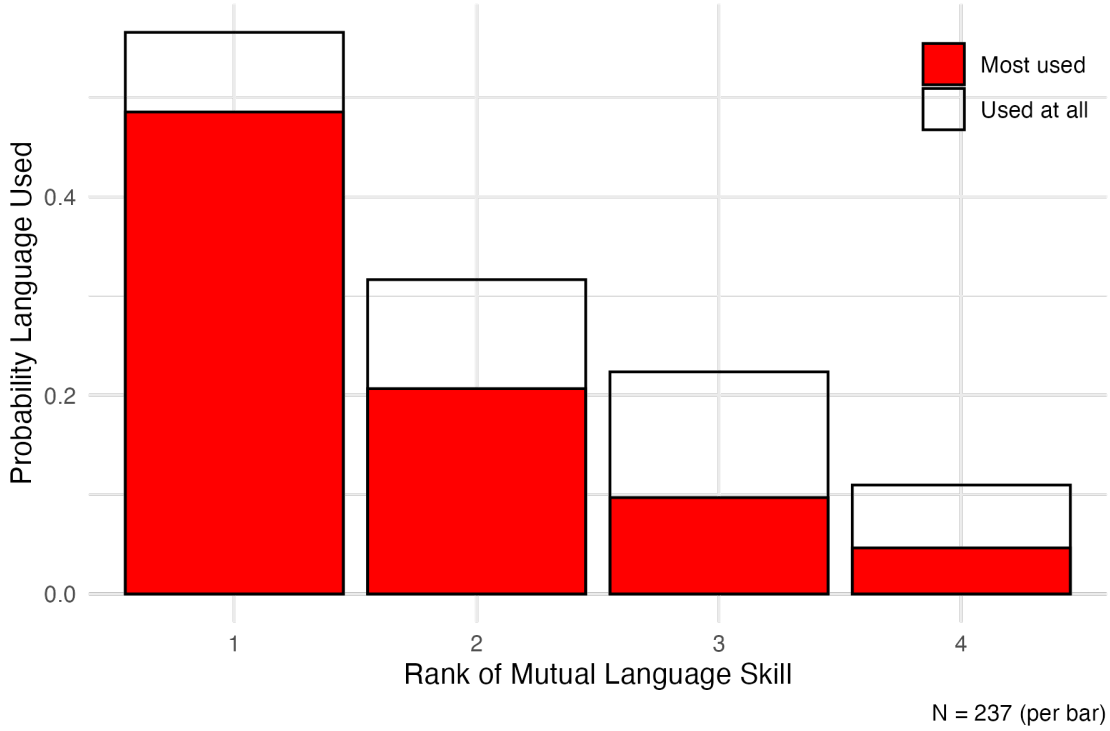
A Appendix

Figure A1: Language usage by joint score



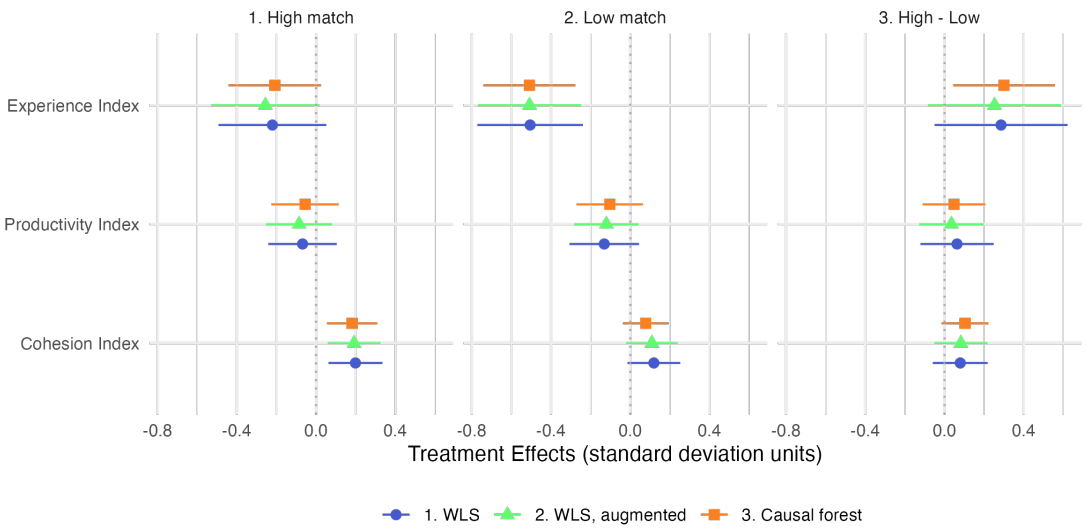
Notes: The plot shows the positive relationship between joint scores and language usage in the cross-ethnic sessions. Joint scores are the product of partners fluency measure for the language; each pair has a score for Albanian, English, Macedonian and Turkish.

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



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

Figure A3: Simple weighted means produce similar outcomes to random forests



Notes: Hájek (IPW) estimated with weighted least squares (WLS). WLS estimates include no covariate adjustment. Weights determined by simulating the treatment assignment mechanism. High (low) language skill match defined as above (below) median language match score. Indices are simple averages of normalized sub-outcomes. For WLS, augmented, we first model the outcome Y as a flexible function $\mu(X)$ of covariates using a regression forests; we then estimate WLS on $Y - \mu(X)$. Standard errors are clustered by school-grade-gender-date of randomization.

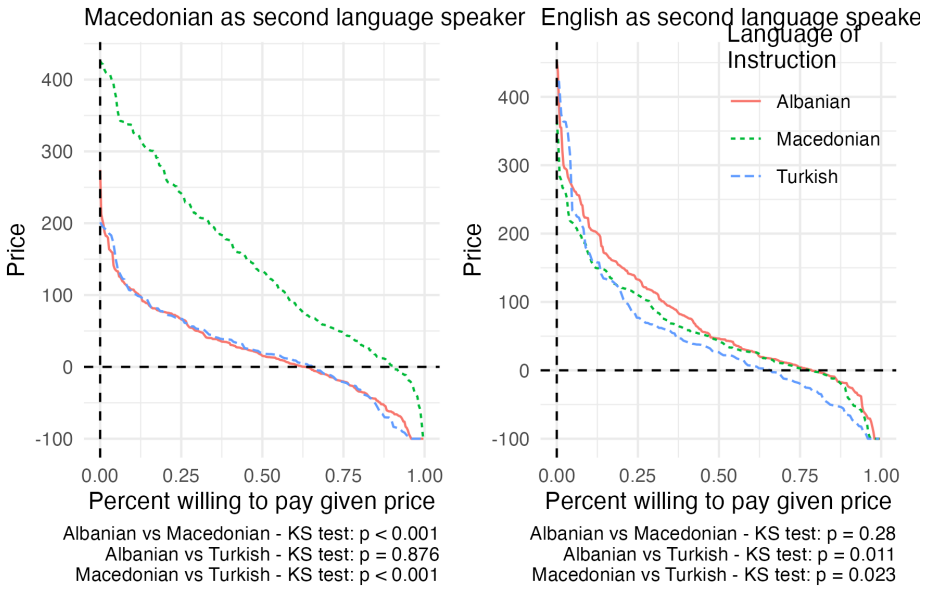


Figure A4: WTP for language skill as a second language by ethnicity (language of instruction)

Table A1: Baseline covariates and the propensity to have a high language match

Covariate	Sample Mean	(sd)	Intercept	(se).	Slope	(se)	p-value	N
Female	0.60	0.49	0.47	0.14	0.25	0.15	0.11	538
Grade 1	0.35	0.48	0.47	0.14	-0.23	0.15	0.15	538
Grade 2	0.30	0.46	0.31	0.11	-0.02	0.11	0.83	538
Grade 3	0.21	0.41	0.06	0.06	0.30	0.13	0.03	538
Grade 4	0.14	0.35	0.17	0.12	-0.05	0.13	0.70	538
Albanian LOI	0.40	0.49	0.42	0.07	-0.06	0.13	0.66	538
Macedonian LOI	0.41	0.49	0.29	0.08	0.24	0.13	0.07	538
Turkish LOI	0.19	0.39	0.28	0.06	-0.18	0.09	0.06	538
LOI maps to ethnicity	0.96	0.20	0.95	0.02	0.02	0.04	0.65	538
English fluency	7.69	2.10	5.60	0.17	4.21	0.36	0.00	538
Local language fluency (not LOI)	4.46	3.56	3.37	0.52	2.19	0.86	0.02	538
Attended Gostivar middle school	0.69	0.46	0.49	0.07	0.39	0.11	0.00	538
Has Cellphone	0.98	0.15	1.00	0.01	-0.05	0.03	0.10	538
Parent has college	0.46	0.50	0.40	0.04	0.12	0.09	0.18	538
Plans to go to college	0.89	0.31	0.78	0.04	0.22	0.07	0.00	538
Grade Point Average (previous year)	4.67	0.57	4.48	0.07	0.38	0.12	0.00	535
Absences	40.74	36.41	40.15	4.42	1.19	6.91	0.86	535
Absences (inexcused)	4.39	5.69	4.34	0.90	0.10	1.69	0.95	535
Life satisfaction (0-10)	8.46	2.03	8.45	0.31	0.02	0.44	0.96	538
Trust (0-10)	6.64	2.54	6.78	0.27	-0.28	0.47	0.56	538

Notes: Table shows *overlap sample* characteristics in column 1 and 2 (Sample Mean). In the next 5 columns, we report a continuous balance-like test. Under a linearity assumption, the *Intercept* shows the characteristics for individuals who have the probability of a high language match approach zero (though is still within the overlap definition). The probabilities are scaled conditional on treatment so that the *Slope* shows the difference as probability of a high language match approaches 1 (along with p-values from the test of the null hypothesis that the slope is 0). Standard errors are clustered by school-grade-gender-date of randomization. **Covariate details** 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 with more than one local ethnicity). Self-reported English fluency and Local language fluency for non-LOI are on a scale from 0-10. Language fluency is imputed with a random forest regression for 1 student missing baseline data using information from the recruitment data.

Table A2: Baseline cohesion and video game proclivities: Correlations with the propensity to have a high language match

Covariate	Sample Mean	(sd)	Intercept	(se).	Slope	(se)	p-value	N
N peers in named for network	5.19	3.18	4.83	0.51	0.73	0.81	0.37	538
N out-group links named	0.60	1.01	0.47	0.11	0.27	0.22	0.23	538
Closeness to own LOI (baseline)	6.42	1.27	6.18	0.16	0.47	0.28	0.11	538
Closeness to out LOI (baseline)	2.01	1.09	1.71	0.12	0.60	0.22	0.01	538
Dictator game denar to outgroup (baseline)	228.86	129.27	233.45	18.44	-9.25	26.51	0.73	538
Dictator game bias (baseline)	25.60	75.54	19.05	7.04	13.22	15.19	0.39	538
Enjoys video games	0.69	0.46	0.65	0.08	0.08	0.13	0.53	538
Makes video game rec	0.60	0.49	0.55	0.09	0.10	0.13	0.44	538
Solo score	475.79	220.35	470.38	44.57	10.91	58.36	0.85	538
Solo score (demeaned)	-0.29	211.27	-19.64	40.84	39.03	53.50	0.47	538

Notes: Table shows *overlap sample* characteristics in column 1 and 2 (Sample Mean). In the next 5 columns, we report a continuous balance-like test. Under a linearity assumption, the *Intercept* shows the characteristics for individuals who have the probability of a high language match approach zero (though is still within the overlap definition). The probabilities are scaled conditional on treatment so that the *Slope* shows the difference as probability of a high language match approaches 1 (along with p-values from the test of the null hypothesis that the slope is 0). Standard errors are clustered by school-grade-gender-date of randomization. **Covariate details** Network data comes from 4 questions where students name 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 (28 percent) use all the slots for friend cliques. Closeness measures come from the Inclusion of Other-in-Self measures (scale from 1-7). Dictator game bias is the difference between the amount of money an individual shares with anonymous ingroup (same ethnicity) recipients compared to what is shared with anonymous outgroup recipients. Enjoys video games equals 1 if the student responded to the question of "How well does 'I love video games' describe you" with 5 or greater on a scale from 0 to 10. Makes video game rec equals 1 if the student names at least one video game when asked for a recommendation, and zero otherwise. Solo score is the average score in solo rounds of "Overcooked! 2". Solo score demeaned demeans the score by gender-specific solo score means.

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.[69] 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.[70] 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*

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

[70] 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 $\frac{4}{20} = 20\%$, $\frac{6}{22} = 27\%$, $\frac{6}{18} = 33\%$ 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.[71] 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.

[71] 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 2000 simulations of the estimation procedure, conditional on any assignment there's a 25.1 percent probability of being assigned control on average across individuals. (There is a 4.9 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 2 strata with expected control probability within 0.5 percent of the target.

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.

Description	N	N treated	Prop. treated
Total sessions:	332	237	0.7138554
Sessions with any screen recording	320	231	0.7218750
Sessions with recordings of "Overcooked! 2"	292	213	0.7294521
Sessions with alternative windows shown	28	18	0.6428571
Sessions with any audio recording	313	224	0.7156550
Sessions with loud game audio	73	52	0.7123288

Table A3: Screen recording data available

C Appendix: Screen Recording Analysis

Screen recordings provide audio and visual data that provide both corroborating evidence for the facilitator reports and rich auxiliary data to explore mechanisms. Since our facilitators were not uniformly adept with technology, we aimed to make the process as simple as possible and only record one video per session. Prior to the cooperative rounds, the facilitator used the built-in Xbox Game Bar to start a screen recordings. After which, the facilitator would move the recording to a folder labeled with the session identifier. Finally, our field team used flash drives to collect the videos and then uploaded them to a secure server (using internet with excruciatingly slow upload speeds). In this appendix, we discuss the quality of screen recordings and how we extract data from them. The analysis was produced with the help of a ChatGPT Codex agent that had no knowledge about the broader project.

C.1 Video quality

For the vast majority of sessions, we have high quality audio and video data captured during pair play. High quality data includes the screen recording for the entire 4 levels played in cooperative mode and usable audio which includes voice along with not-too-loud sounds from the video game itself.

We do not have full coverage of the videos as outlined in Table A3. Some sessions have usable audio but not usable video and vice versa. Video data is either missing or we The recording software captures whichever software window is on screen when the recording starts; 27 recordings show Qualtrics screens or internet browsers. Some sessions (5) included false starts or late starts so that we have partial coverage of the four

levels. We are missing any screen recording for 12 sessions. Files are likely missing because the recording was not started, though it is possible the recording started, but the file was not moved to the proper location.[72]

All recordings have audio, but the audio is of varying quality. It is less obvious what fraction of the videos have high quality audio compared to high quality visuals. However, there are a couple of reasons audio data may be low quality. For a small set of recordings (7), there is no audio captured. A more common challenge is that video game audio is loud, the overall audio level is very low, or there is distortion. The difference in audio quality may reflect incidental adjustments in the recording software, differences in positioning of laptops in rooms, and/or participants adjusting game audio levels. Additionally, some sessions take place with background noise which appear to be hallway noise during passing periods. As described in detail below, we trained a model to classify occurrences of game music and sounds, and find 73 recordings with loud game audio.

Overall, there does not appear to be strong correlations between treatment status and recording quality.

C.2 Visual processing and quality analysis

Visual outcomes rely on computer vision to extract information about game play. At the end of each round of game play, a receipt screen shows the total score (our main productivity measure) as well as a break down of orders delivered (resulting in positive points) and orders failed (resulting in negative points). Additionally, we used computer vision to watch game play and count the number of fires, a highly visible coordination failure.

The data processing is similar to optical character recognition. First, we sample frames at 30 frames per minute. To extract scoring information, we run a Python script over the samples and for each frame test if it shows the receipt screen. If yes, we use template matching to extract the data and keep data from one complete receipt. In the case, that sampled frames do not include a complete receipt screen, we resample the video around the partially detected screen and re-run the script. In a few cases, students race through the receipt screen so that there is no complete screen. In this case, we use the "heads-up display" information which reports scores throughout the level to back out the final score, number of failed orders and so forth. This

[72]The facilitator protocol required the facilitator to attest to making and storing the recording in the proper place.

works very well. For levels where we have the receipt, we match almost 90 percent of the failed order counts and 99 percent of final scores.[73] To learn about fires, we take advantage of the fact that fires always start on stove tops and build a detector that looks for fire-colored pixels (yellow and orange with high saturation on an Hue-Saturation-Value scale) in the vicinity. We use stratified random samples of frames to adjust the detection rules and provide validation metrics. The fire detector matches human labels for 98% frames with flames flagged (1 of 60 false positives) with 100% of frames with no flames tagged (0 of 40 false positives). Since it is easy to miss false positives, we also looked at the 20 frames with no detected fire event, but with the high numbers of pixels in the target range. Of these 1 of 20 was a false positive. Taken together, we take this as evidence that the fire and score detectors perform at a high level.

The total game-level scores match the facilitator reports closely. In 1127 of 1168 game-levels observed, the scores match the facilitator reported scores exactly.[74]. Conditional on there being a discrepancy between the measurement the median difference is 7 points. In 10 game-levels score discrepancies are large (over 50 points). These 10 errors reflect facilitator errors. The high correspondence between facilitators' records and video extracted data gives us high confidence in both the facilitators' reporting and the quality of the visual data extraction pipelines.

C.3 Audio processing and quality analysis

Our main audio outcome is a prediction of the proportion of a session recording during which vocal communication occurring. We call this the "vocal-activity fraction" and use it to sanity check the facilitators' assessments of the amount of communication. Even for this relatively simple outcome, audio processing requires sophisticated machine learning models to produce usable data. First, we extract vocalizations from the recording using Hybrid Transformer Demucs **CITE Rouard**.[75] Then, in a second stage, we use Pyannote's speaker diarization community-1 model **CITE ***, a pretrained pipeline that for our purposes does voice activity detection (VAD) and segmentation. The voice detection looks for audio signals that appear

[73] We miss some events because multiple scoring events can occur simultaneously or at the very end of a level. So, these imputed measures are lower bounds.

[74] This number is marginally inflated, because we resolved roughly 10 discrepancies across facilitators using video data in early data processing.

[75] We use the off-the-shelf model, which is built for music source separation (e.g. extracting vocals, drums, and bass from music tracks). The vocal stem extraction (which includes voice-like noises such as whistles) performed better on an ad-hoc validation test compared to using the raw audio.

voice-like and so further reduce noise in measurement (e.g. ignoring audio artefacts created with htdemucs). In tests of community-1 on well known corpuses, community-1 achieves diarization error rates (DER) as low as 11.2% and as high as 44.6%.[76] As such, we treat the vocal-activity fraction as an exploratory automated proxy for the "vocal-activity fraction". We find that the vocal-activity fraction is correlated with facilitators' likert-scale reports of the amount of communication with Spearman correlations of .49 (bootstrapped 95-percent CI of .40 - .58). In Table A3, we report levels with loud game audio as a proxy of recording quality. To estimate these, we built a logistic classifier that categorizes recordings as having game audio that is "high", "moderate", "low", or "none". Using 8-second clips, we extract 69 audio features that describe the audio signature of the clip such as mel-frequency cepstral coefficients. This model performs moderately well when the model is targeting game related audio with F1 score of 0.87 on a human-annotated hold-out set.

[76] These reported error rates should be interpreted as an upper bar since the DER contain missed speech and false alarms, which matters for our use case, and speaker confusion (i.e. incorrectly labeling who said what), which does not matter for our use case. (see https://huggingface.co/pyannote/speaker-diarization-community-1). In internal testing using hand labeled samples, we find that the model does a good job at detecting vocal activity (presence/absence in 3 second clips), but appears to under count vocal-activity fraction, particularly when the vocalizations are quiet.

D 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 heavily on the accompanying paper (Drake et al. (2024); Hereafter DPTT).

The goal is to estimate preference parameters $\theta_i \in^k$ over k attributes for each individual $i \in N$. Specifically, we measure willingness to pay for partner ethnicity, second language skills, and an interaction between language and ethnicity. 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 (i.e. choice of option A or option B) and d are the choice scenarios that the agent faced.

BACE is a method to select choice scenarios that are maximally informative about the parameter space (DPTT, 2024). The researcher's problem is to choose the best discrete choice scenario to show next (d_{t+1}^*) given the current posterior $Pr(\theta_i|x_i^{(1:t)}, d_i^{(1:t)})$. d_{t+1}^* is best in that it maximizes the expected mutual information gained. BACE provides a multidimensional analogy to binary search: d_{t+1}^* is a choice scenario that, in expectation, partitions the current posterior into regions that predict different choice outcomes, so that the next observation is highly informative for updating beliefs about θ.

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

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 *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. Show choice scenario d^*_{t+1} to the student who responds with $x^{t+1} \in \{0, 1\}$

4. Update the prior (i.e. the posterior is the sampling distribution for the next iteration.)

We approximate the target distribution using population Monte Carlo (PMC). PMC is a sequential importance sampling method that draws candidate points from a proposal distribution and reweights them to obtain Monte Carlo estimates of posterior expectations Cappé et al., 2004. The algorithm is fast and parallelizable, making it well suited for computing and updating choice scenarios in real time.

D.1 Implementation

Students derive expected utility from a monetary reward r and a potential partner's ethnic and linguistic attributes, summarized by the vector z. When implementing BACE for data collection, we include detailed ethnicity indicators (e.g., Turkish, Albanian, Macedonian). For estimation and analysis, we collapse these indicators to a binary measure of whether the partner shares the student's ethnicity.

Utility is given by:

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

where z includes indicators for whether the potential partner is of the same ethnicity and whether they speak English, Macedonian, or neither as a second language. Additionally, z includes indicators for the interaction between shared ethnicity and second language skills, capturing the idea that second language skills may be less valuable.

We normalize the coefficient on the monetary reward to one, allowing the elements of θ to be interpreted in Denar units. Students are instructed to assume that both options correspond to partners from their cohort, of the same gender, and otherwise identical.

Students chooses between A and B, choosing A if

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

and B otherwise. Equivalently, option A is chosen if

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

We assume that ϵ follows a Type I extreme value distribution, capturing transitory, choice-specific utility shocks and decision noise.

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 scales utility differences and captures the degree of attention or indifference in choices. As μ gets large, even small utility differences translate into near-deterministic choices, while lower values of μ correspond to noisier decision-making and a higher frequency of choice inconsistencies.

Prior. When running the adaptive choice experiments, we specify a prior distribution over preference parameters for each element of z. Preferences are assumed to follow independent normal distributions centered at zero. We set the prior variance to 100 Denar for ethnicity match indicators and 200 Denar for language skill indicators. Centering the priors at zero is conservative, reflecting an initial belief of no systematic preference in either direction. The prior over the dispersion parameter μ is uniform on the interval $[0.01, 0.09]$.

Choice scenarios and designs. The design space consists of 1,275 distinct choice scenarios, each defined by a vector $(\Delta r, \Delta z)$. The reward difference Δr takes one of 25 values ranging from -100 to 500 Denar (approximately \$-2 to \$10 in 2023 dollars), in increments of 25 Denar (roughly \$0.50).

The attribute difference vector Δz takes values in $\{-1, 0, 1\}$, reflecting differences in indicator variables

across the two options. Due to logical dependencies between ethnicity and language attributes, the feasible design space for Δz consists of 51 distinct vectors.

As an illustration, consider $\Delta z = (-1, 0, 1, 0, 1)$, where z_1, z_2, and z_3 correspond to Albanian, Macedonian, and Turkish ethnicity indicators, respectively; z_4 indicates English as a second language; and z_5 indicates Macedonian as a second language. In this example, option A corresponds to a Turkish partner who speaks Macedonian as a second language, while option B corresponds to an Albanian partner. English language ability is randomly assigned to both options.[77]

At each iteration of the adaptive algorithm, we select the optimal design and translate it into a concrete choice scenario as described above. Each student faces 12 such choice scenarios. We chose this number to balance the need for sufficient information for estimation with students' attention constraints. Simulation results indicate that most of the reduction in mean squared error occurs within the first 10 choices, after which learning gains flatten substantially.

D.2 Preference estimation

After data collection, we estimate individual-specific preference parameters using population Monte Carlo (PMC). As a robustness check, we also estimate preferences using a mixed logit model. For estimation, we modify the attribute vector z so that it includes the two second-language indicators as before, along with a single "own ethnicity" indicator equal to one if the partner's ethnicity matches the student's language of instruction and zero otherwise.

For each student i, we observe a vector of binary choices $x_i^{(1:12)}$ and the corresponding matrix of choice designs $d_i^{(1:12)}$. Given these data, the likelihood function, and a prior, we apply the PMC algorithm to obtain an individual-specific posterior distribution over the preference parameter vector θ_i.

[77] The 51 feasible Δz vectors arise as follows. For ethnic pairings involving Macedonians, ethnicity can appear in either option A or B, yielding four distinct ethnic pairings. Each of these has two possible Macedonian-as-a-second-language configurations (as Macedonians cannot list Macedonian as a second language) and three possible English-as-a-second-language configurations, yielding $4 \cdot 2 \cdot 3 = 24$ scenarios. There are also three ethnic pairings that potentially exclude Macedonians (including the case where both options share the same ethnicity), each with three English and three Macedonian second-language configurations, yielding $3 \cdot 3 \cdot 3 = 27$ additional scenarios. Together, these generate 51 distinct Δz vectors. In cases where $\Delta z = (0, 0, 0, 0, 1)$ is selected, we randomly assign the same ethnicity to both options and suppress Macedonian as a second language when the ethnicity is Macedonian. This edge case affects approximately 2 percent of choice scenarios.

We implement estimation using a two-step procedure. In the first step, we estimate preferences using nuisance priors composed of independent distributions.[78] In the second step, we use the resulting set of estimates $\{\theta_i\}_{i=1}^N$ to construct empirical Bayes priors. Specifically, for each individual i, we define a normal prior with mean and variance computed from the leave-one-out sample of estimated parameters, where the prior mean is given by

$$\bar{\theta}_{-i} = \frac{1}{N-1} \sum_{k \neq i} \theta_k.$$

The empirical Bayes procedure is used to stabilize individual-level posterior estimates in a setting with a limited number of choices per student. To avoid mechanical reuse of the same data, the prior for each individual is constructed using a leave-one-out approach that excludes that individual's own choices. Importantly, the empirical Bayes step does not affect the likelihood, the adaptive design, or the identification of preferences, and serves only as a regularization device to reduce posterior variance.

D.3 Robustness to prior specification

Robustness to prior choice and estimation method. Table reports robustness checks that vary the nuisance priors used in the population Monte Carlo (PMC) leave-one-out estimation and compare PMC estimates to mixed logit estimates. Panel A shows that the population means of individual posterior estimates are stable across increasingly diffuse normal priors, with qualitative patterns unchanged as the prior variance increases from $N(0, 100)$ to $N(0, 300)$. In particular, preferences for same-ethnicity partners and for second-language skills are consistently positive and increase smoothly as the prior becomes less informative. Mixed logit estimates yield larger magnitudes but preserve the same ordering and sign patterns. Panel B reports correlations of individual-level estimates with those obtained under the baseline $N(0, 100)$ prior. For the main effects—same ethnicity and language skills—the correlations exceed 0.95 across all alternative specifications, indicating a high degree of stability at the individual level. In contrast, interaction terms exhibit lower correlations, reflecting limited identifying variation. Overall, these results indicate that the main preference parameters are robust to both prior specification and estimation approach.

[78] For the main effects, our preferred specification uses Normal draws centered at 0 with a standard deviation of 100. For interaction effects, we use normal centered at 0 with the standard deviation to 25.

Parameter	$N(0, 100)$	$N(0, 200)$	$N(0, 300)$	ML prior	ML estimates
Panel A: Population mean (SD) of individual posterior means; ML estimates show pop. mean and distributional SD					
Same ethnicity	167.1	217.3	247.4	254.1	302.9
	(131.9)	(190.3)	(223.5)	(231.6)	(257.1)
Speak English	61.8	70.7	76.4	76.8	86.2
	(103.5)	(122.4)	(132.3)	(115.0)	(117.1)
Speak Macedonian	61.5	79.2	88.2	93.0	113.6
	(110.7)	(154.5)	(174.2)	(165.9)	(186.7)
Speak English $\times$ Same eth.	-0.8	0.1	2.2	4.1	4.6
	(2.3)	(11.6)	(28.3)	(37.7)	(65.7)
Speak Macedonian $\times$ Same eth.	-0.4	-1.2	-1.1	-3.2	-3.2
	(1.3)	(7.5)	(20.1)	(0.0)	(0.1)
Panel B: Correlation with $N(0, 100)$ spec (individual-level)					
Same ethnicity	—	0.985	0.973	0.969	0.968
Speak English	—	0.985	0.968	0.971	0.950
Speak Macedonian	—	0.981	0.965	0.963	0.956
Speak English $\times$ Same eth.	—	0.754	0.728	0.716	0.730
Speak Macedonian $\times$ Same eth.	—	0.552	0.537	0.034	-0.054

Note: PMC specs use leave-one-out priors; Mixed Logit prior uses ML estimates, estimated with maximum simulated likelihood.

E Appendix: Language Skill Match

This section argues that $f_{\max}$ is an appropriate measure to capture the complementary language skills of a pair. Recall that we students self-rate their English, Macedonian, Albanian, and Turkish skills on a scale of 0 to 10. Then, for pair i and j, we determine a joint score in a given language k by multiplying i and j's scores: $f_k = f_{ik}\dot{f}_{jk}$. And we defined:

$$\underset{\text{language skill match}}{f_{\max}} = \max(f_a, f_e, f_m, f_t)$$

where the subscripts correspond to the first letter of the languages under consideration.

When f_k is large, students are more likely to use that language as reported by the session facilitators. Language use is distinct from the "amount of communication" outcome reported in the main result. The session facilitators record which languages they hear the students use during gameplay. They also report an intensity metric which is supposed to capture the proportion of time each language was used. Though in practice, roughly 60 percent of reports add up to 100 percent and often facilitator pairs are not in agreement on this measure. We average over the facilitators' measures and assume that the relative ranking is accurate. So, a language is "most used" if it has the highest usage proportion within a session. Figure A1 shows the proportion of sessions where language k is used given the joint score of a pair f_k. The plot shows a sharp increase in the likelihood a language is used as the joint score increases. Roughly one in three pairs use a language if their mutual skills are 50; this jumps to 76 percent when both students report top fluency in a language. On the other extreme, it may seem odd that 10 percent of pairs use a language when the joint score is 0; this may reflect a student who talks in a language despite not being understood.

While sessions are multilingual affairs, the ordering of the joint skills provides a strong signal as to whether students use the language. To explore this further, index the joint skills in order so that $f_{\max=f_1 \geq f_2 \geq f_3 \geq f_4}$. Figure A2 demonstrates two things: First, students are most likely to use the language they jointly know the best (black outlined bars). The probability a language is used declines along with its position in the ordering. Second, this pattern is more extreme when considering the "top match" among languages used (red bars). Pairs use their top language match the most roughly 55 percent of sessions.The second red bar shows that

roughly 19 percent of pairs use the language with a joint score f_2 more than their top language match. This tends to happen if f_1 and f_2 are very close; though this could reflect sub-optimal coordination. Roughly 18 percent of pairs use f_1 *and* f_2.

Taken together, the joint score f_k captures communicative capacity generally and $f_{\max}$ provides an easily understood one-dimensional measure that eases the analysis. When considering high or low language match, we lose very little by making this simplification. After all, when students rely on f_2, they typically do so when f_1 and f_2 are very close. What is more, we run regressions and use random forests to predict the amount of communication in a session using simple models that make use of f_1 through f_4. The model that includes f_1 alone outperforms nearly identical to models that use combinations of the joint scores (in terms of mean square error of out-of-bag predictions with random forests). On the downside, the use of $f_{\max}$ adds some measurement error when considering *lingua francas*.

F Appendix: Policy Counterfactuals

In this appendix, we provide details on the policy counterfactuals. In the counterfactual, we imagine that a decisionmaker can assign students to co-ethnic partners based on their observable characteristics. How does re-assignment impact expected changes in cohesion outcomes and non-pecuniary costs?

Let $Y_i(1, d(i,j))$ be student i's potential outcome after being assigned partner j where $d(i,j)$ is an indicator that equals 1 if the pair (i,j) has a high communication potential. Within gender by cohort strata S, we find the assignment vector $\{d(i,j)\}$ that maximizes the following problem:

$$\max_{\{d(i,j)\}: i,j \in S} \sum_{i \in S} Y_i(1, d(i,j))$$

In words, we maximize the aggregate outcome subject to the constraint that there is a feasible matching $\{d(i,j)\}$.

We approach estimating this in two ways. A direct approach that uses predictions from the generalized random forests and an indirect approach that uses averages of the score function.

In the direct approach, we use the model-implied estimates of $\beta_i^L = Y_i(1,0)$ and $\beta_i^H = Y_i(1,1)$ from the generalized random forest. For individuals, these are out-of-bag or out-of-sample predictions that reflect bias-variance tradeoffs. This becomes a maximum matching problem. The idea is to find the set of edges with the largest sum of weights in a graph such that no two edges share a vertex. To solve the problem, we use Edmond's Blossom algorithm.[79]

First, we create an edge connecting each set of cross-ethnic pairs. For the pair i, j, we calculate

$$d(i,j) = \mathbb{1}_{\text{Communication Potential} \geq m}$$

. Where m is the median communication potential. Then assign the weight to edge:

[79] Graph theory provides a suite of algorithms for solving maximum matching problems. Because we have three ethnicities, we cannot rely on simpler algorithms. For example, if this were a bipartite graph, we could use the Ford-Fulkerson max flow algorithm or the Hungarian algorithm.

$$w_{ij} = \begin{cases} \beta_i^H + \beta_j^H, & \text{if } d(i,j) = 1 \\ \beta_i^L + \beta_j^L, & \text{if } d(i,j) = 0 \end{cases}$$

Where β_i^H and β_i^L are the predicted personalized treatment effects returned from passing individual i's covariate profile to the generalized random forest. Edmond's Blossom algorithm then efficiently finds the set of pairs (matching) that maximizes (or minimizes) the sum of the weights.

The indirect approach uses averages of AIPW (augmented inverse propensity weighted) scores that are mean-unbiased for the ATE by construction. However, these scores are far too noisy to use individual estimates in an optimization routine. In this approach, we only consider the composition of a matching. "Max matching" selects the feasible within-strata matching that maximizes the *number* of high-match pairs (subject to matching the maximal number of students), and "Min matching" minimizes it. Selecting the matching requires only language-skill data, not outcome estimates. For a matching with share s of matched students in high-match pairs, we report the composition-valued estimate $\tilde{\tau} = s, \bar{\Gamma}^H + (1-s), \bar{\Gamma}^L$, where $\bar{\Gamma}^H$ and $\bar{\Gamma}^L$ are the average of AIPW scores, i.e., the pooled ATE estimates for high- and low-match pairing.

G Appendix: Variation in treatment

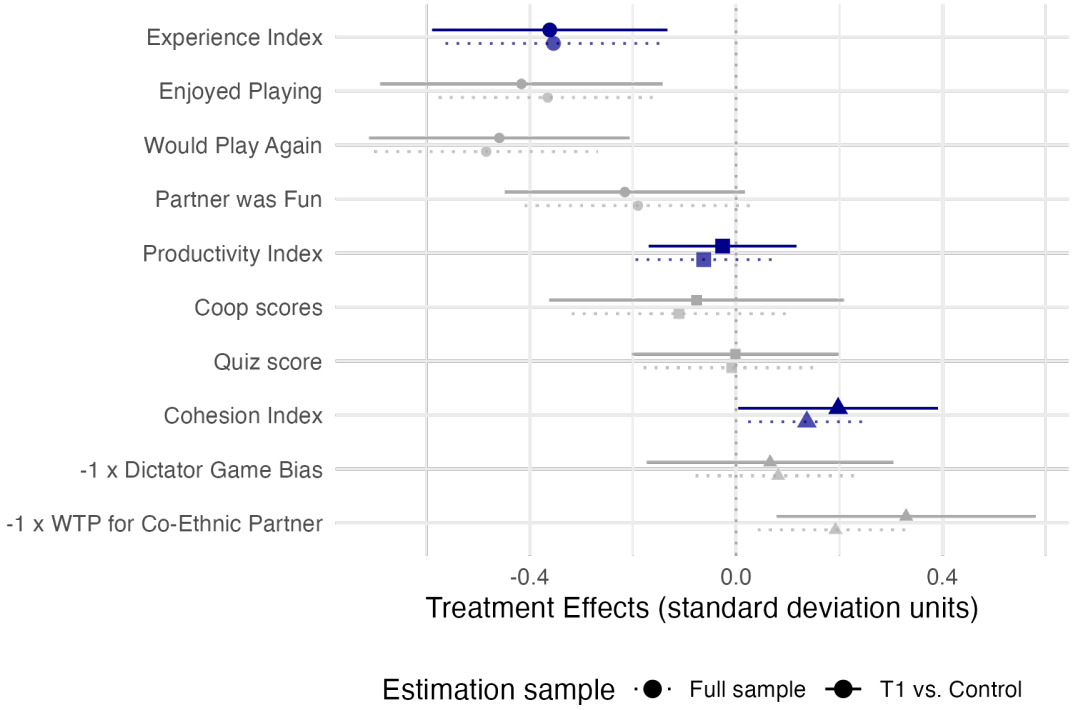
In the main results, we label all cross-ethnic pairs "treated". We randomize to these pairs a secondary priming treatment arm which leads to differences in the treatment experienced relative to the control group. In this appendix, we first show the main results maintain when we focus solely on students in cross-ethnic pairs who saw the same exact set of questions at the control group. Then, we discuss the priming treatments in detail and argue in favor of pooling across treatment arms.

Description of treatment arms The cross-ethnic pairs faced subtly distinct treatment arms that involved tweaks to the survey completed before playing the video game. Inspired by Chen et al. (2014), these were meant to prime identities. The first treatment (T1)–identical to the experience of the control group–was designed to be identity *neutral*. Individuals answered questions about their cellphone usage and saw pictures of cellphones. An exemplary questions is "What are your favorite things about your phone?". Directly before starting to partner play, they saw a GIF that said "Do your best!" (in their language of instruction).

We designed the second and third arms to prime shared identity relying on town- and school-spirit. Priming these shared identities might encourage individuals to see their partner as part of supra-ethnic groups. In both arms, the survey included the same pictures of Gostivar's most famed sites including Gostivar's clock tower, the Gostivar city crest, and the colorful large Gostivar sign along the Vardar river walk. The second arm (T2) further encouraged students to think about these identities by answering questions such as "What are your favorite things about Gostivar?" and "Why did you choice to attend your high school?" Further, directly before starting to partner play, they saw a GIF that said "Do it for high school name !" (in their language of instruction). Meanwhile, the third arm (T3) encouraged students to not only think of shared identity, but also of common rivals. In this case, students were asked to pick rival cities and high schools and answered questions such as "Why is Gostivar better than rival town selected ?". Directly before starting to partner play, they saw a GIF that said "Beat the rival high school name and make high school name proud!" We hypothesized that shared identity priming would lead to stronger cohesion results. Further, we hypothesized that adding an outgroup ("Us vs Them") would enhance the effects following a large literature in social science (e.g. Sharif, 1958).

Appendix Figure A5 compares the main (pooled) estimates with estimates where treatment is restricted to

Figure A5: Comparison of full sample results and restriction to T1

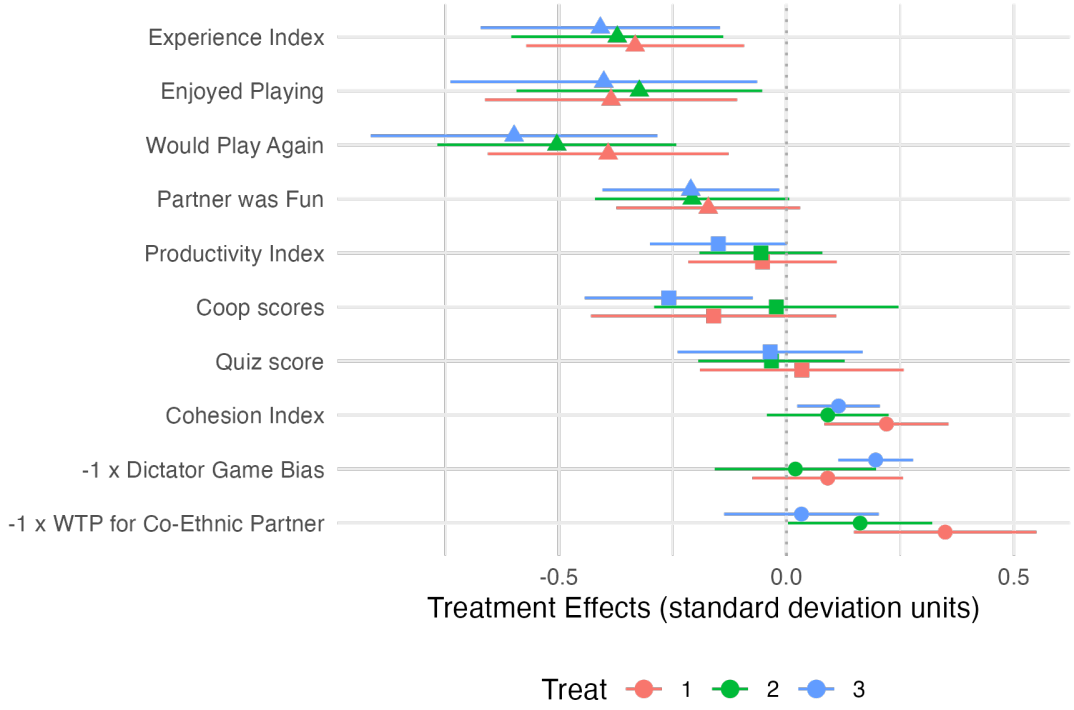


the T1 arm, which to reiterate is the same set of images and questions the co-ethnic pairs see. The plot shows OLS coefficients from regressions of treatment indicator and school-grade-gender fixed effects on outcomes. The T1 arm has slightly larger standard errors since the treatment sample size is roughly one-third (n=160) the total treated population (n = 474). Estimates are similar across the specifications for all outcomes and statistically indistinguishable with the possible exception of willing to pay for a co-ethnic partner. Students in T1 show willingness to pay effects that are .24 (se = .127, p ¡ .1) standard deviations larger than other students with cross-ethnic partners. The extent to which shared identity priming "backfires", this puts downward pressure on our results.

In Appendix Figure A6, we compare estimates of each of the treatments to control.

The results across priming arms are somewhat mixed. For example, the students primed to "Us vs Them" demonstrate the large positive effects on dictator game bias and the smallest effect on willingness to pay for a co-ethnic peer. Moreover, the ranking between T1, T2, and T3 is inconsistent across outcomes. Given this inconsistency in directionality, we choose not to focus on priming separately.

Figure A6: Effects by treatment T1, T2, T3



This treatment was inspired by Chen et al. (2014), who find that shared identity priming versus other-ing priming impact behavior of Chinese students paired with non-Chinese students at the University of Michigan. They compare China imagery ("other-ing" identity) to University of Michigan imagery (shared identity). Our experiment is weaker in that our "control" is a placebo rather than an "other-ing" prime. We chose this because we wanted at least one arm that was exactly the same as the co-ethnic control experience, and we felt strongly that the control face a neutral set of questions.

H Appendix: Attrition Analysis

This appendix provides detail on attrition in the experiment and analyses exploring the robustness of results to differential attrition.

H.1 Sources and timing of attrition

Table A4 documents the construction of the analysis sample and the extent of attrition at each stage of the study. Of the 822 students who consented to participate, 80.5 percent took part in the intervention. Twenty percent of non-completers do not participate because they (n=18) or their partner (n=16) do not complete baseline surveys. The remaining attritors and their partners both complete baseline surveys, but did not participate in the intervention.

We document evidence of differential attrition in Table A5. Completion rates are 90 percent among students assigned same-ethnicity partners (188 out of 208), compared with 77 percent among students assigned cross-ethnic partners (474 out of 614). This differential selection into participation is potentially concerning because the observed cross-ethnic sample may disproportionately include students who are more willing or better able to participate in cross-ethnic interactions, which could bias treatment effect estimates. Attrition also varies substantially across schools. In the Technical School, roughly half of students assigned cross-ethnic partners do not participate. This pattern partly reflects the organization of school shifts: at the Technical School, shifts are organized by ethnicity, whereas at the other schools they are organized by grade level. As a result, Albanian students at the Technical School attend class at a different time of day than Macedonian and Turkish students, making cross-ethnic participation more costly because at least one student would need to arrive early or stay late. Finally, Table A6 compares baseline characteristics of completers and non-completers by school. On average, completers and non-completers have similar observable characteristics in the Gymnasium and Medical School, but there are large and statistically significant differences at the Technical School. Technical School non-completers have lower baseline GPA, more absences, fewer cross-ethnic friendships, and lower Macedonian fluency than completers. While some baseline differences are imprecisely estimated, these patterns demand further analysis to understand the extent to which attrition influences the estimated treatment effects.

We assess robustness to attrition in several ways, including subsample analyses, Lee bounds, and regression adjustment. The evidence above suggests that attrition is most severe in the Technical School. We therefore first re-estimate the main treatment effects excluding this school. Figures A7 and A8 show that the resulting estimates are very similar to the main estimates, although the confidence intervals are wider because of the smaller sample size. This suggests that the main findings are not driven solely by differential attrition in the Technical School. However, excluding the Technical School does not fully resolve concerns about attrition. Because non-completers in the Technical School are negatively selected, complete-case estimates may overstate or understate the treatment effects for the full randomized sample.

We therefore implement Lee bounds to quantify how selective attrition could affect the complete-case estimates. The Lee bounds procedure imposes a monotonicity assumption on selection into completion. In this context, the assumption is that assignment to a cross-ethnic partner weakly lowers the probability of completion relative to assignment to a same-ethnicity partner for otherwise comparable students. Under this assumption, the excess completers in the arm with the higher completion rate are trimmed from the outcome distribution to equalize completion rates across arms.

We implement the bounds within school-by-grade-by-gender strata. For each stratum, we compute the difference in completion rates across arms and trim the corresponding share of observations from the arm with higher completion. The lower bound trims observations with the most favorable outcomes, while the upper bound trims observations with the least favorable outcomes. (In a few strata, the monotonicity assumption does not hold in sample, and we do not trim.) This yields conservative bounds on the treatment effect that allow for worst-case selection among the marginal completers.

Figure A9 reports Lee bounds for the main outcomes. The grey bars show unconditional Lee bounds, while the red bars implement the bounding exercise within school-by-grade-by-gender and then aggregate across strata using weights proportional to the number of treated units within the strata. The stratified bounds are generally narrower, consistent with the fact that attrition is highly concentrated in specific school and demographic cells. The results are broadly consistent with the complete-case estimates. In particular, the negative effect on the Experience Index remains robust to the Lee bounding exercise. For the Cohesion Index, WTP for a co-ethnic partner, and Dictator Game Bias, the bounded estimates remain directionally similar but the confidence intervals include zero, suggesting more caution in interpreting these effects. Overall,

the bounds indicate that differential attrition is unlikely to overturn the main qualitative patterns, though it weakens precision for some outcomes.

Step	N	Attrition
Full Sample	822	
Completed Baseline	804	-18
Both Partners Completed Baseline	788	-16
Participated in Intervention (Analysis Sample)	662	-126

Table A4: Attrition from Full Sample to Analysis Sample

School	Assigned (C)	Assigned (T)	Completed (C)	Completed (T)	Attrition % (C)	Attrition % (T)
Gymnasium	58	174	52	136	10.3	21.8
Medical	80	254	74	248	7.5	2.4
Tech	70	186	62	90	11.4	51.6
Overall	208	614	188	474	9.6	22.8

Table A5: Attrition by School

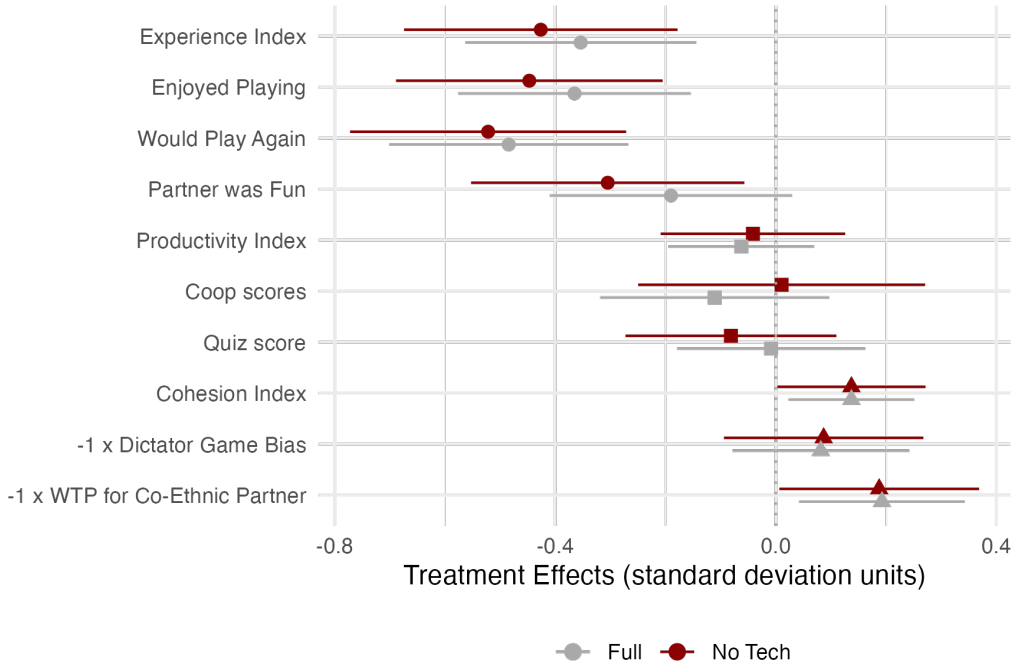


Figure A7: Main effects estimated without Technical School.

School	Variable	Completer mean	Attriter mean	Mean diff.	p-value	n (obs)	n (attr)
Tech	GPA	4.33	3.86	-0.468	0.000	151	101
Tech	Absences	40.46	52.07	11.612	0.010	151	101
Tech	Inexcused absences	9.74	18.25	8.512	0.003	151	101
Tech	Baseline DG bias (Denar)	41.12	57.07	15.947	0.334	152	92
Tech	Female	0.30	0.17	-0.130	0.329	152	104
Tech	Network degree	6.01	4.97	-1.046	0.043	152	92
Tech	Cross-ethnic links	0.68	0.42	-0.260	0.000	152	92
Tech	English fluency	7.54	7.08	-0.463	0.221	152	92
Tech	Macedonian fluency	7.99	7.15	-0.830	0.014	152	92
Gymnasium	GPA	4.51	4.58	0.068	0.669	188	44
Gymnasium	Absences	43.00	37.68	-5.318	0.361	188	44
Gymnasium	Inexcused absences	6.27	6.68	0.416	0.802	188	44
Gymnasium	Baseline DG bias (Denar)	15.29	-3.75	-19.043	0.024	188	40
Gymnasium	Female	0.62	0.77	0.156	0.287	188	44
Gymnasium	Network degree	4.24	3.70	-0.539	0.362	188	40
Gymnasium	Cross-ethnic links	0.71	0.90	0.187	0.362	188	40
Gymnasium	English fluency	7.10	7.01	-0.089	0.703	188	40
Gymnasium	Macedonian fluency	6.92	7.19	0.265	0.294	188	40
Medical	GPA	4.72	4.36	-0.359	0.207	319	11
Medical	Absences	41.02	49.64	8.614	0.230	319	11
Medical	Inexcused absences	3.73	4.82	1.091	0.509	319	11
Medical	Baseline DG bias (Denar)	27.87	20.00	-7.873	0.630	322	10
Medical	Female	0.66	0.50	-0.165	0.409	322	12
Medical	Network degree	5.11	4.60	-0.515	0.564	322	10
Medical	Cross-ethnic links	0.59	0.70	0.107	0.604	322	10
Medical	English fluency	7.60	7.53	-0.074	0.871	322	10
Medical	Macedonian fluency	7.86	7.35	-0.513	0.630	322	10

Table A6: Comparison of Completers and Attriters

Notes: Tables shows the average of baseline covariates for completers and attriters subset by school. P-values come from a bivarate regression of the baseline covariate on an indicator for completion.

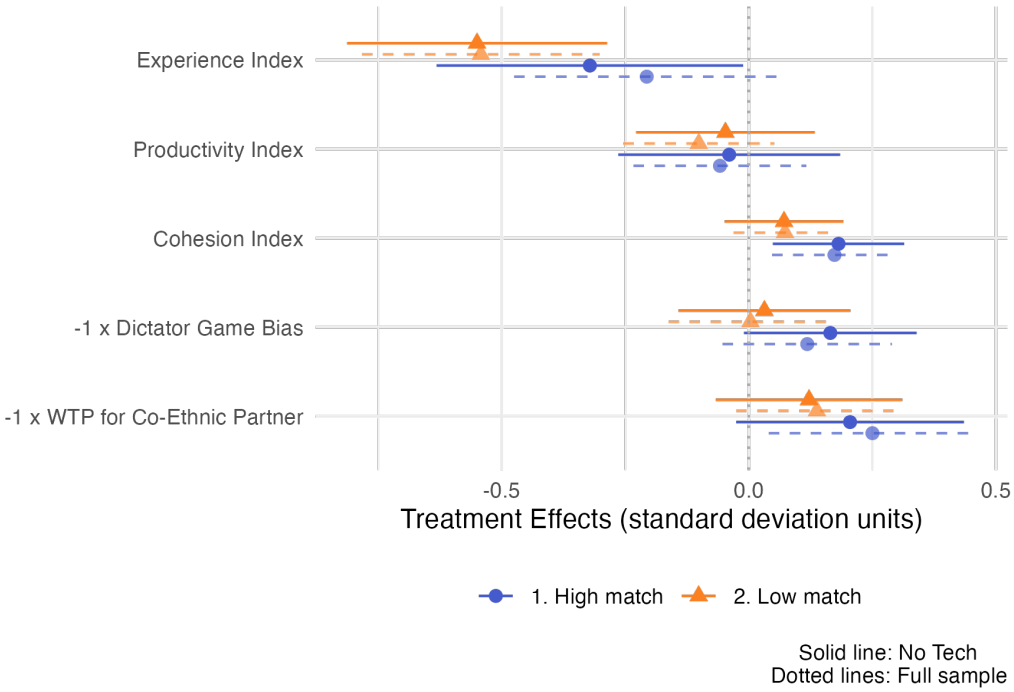


Figure A8: Language effects estimated without Technical School.

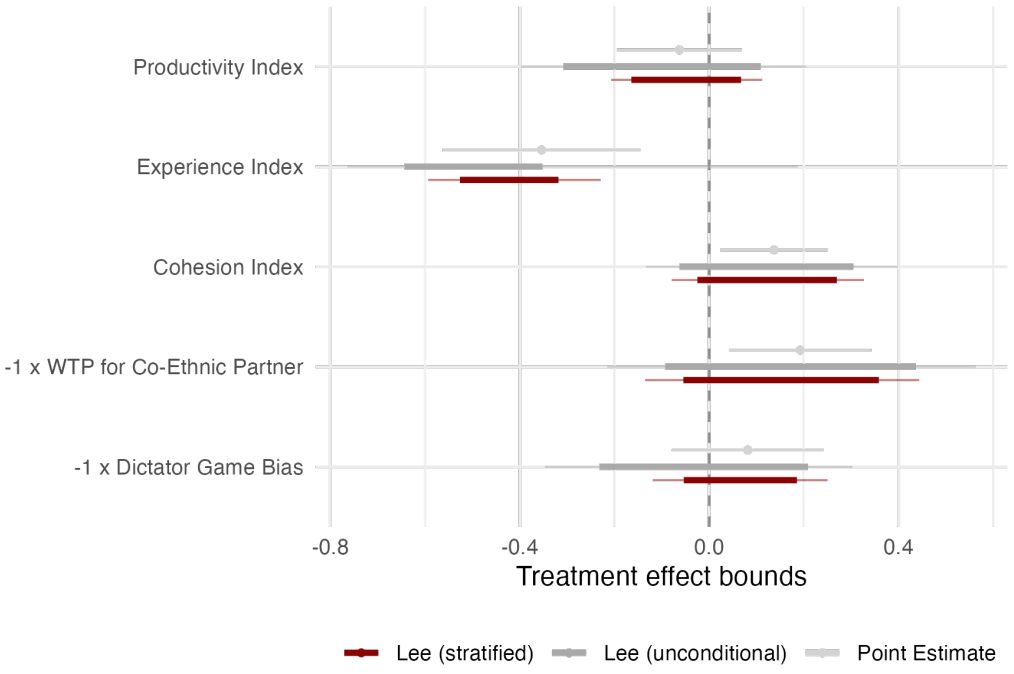


Figure A9: Lee bounds on main effects