Language Skills and the Demand for Segregation

Preliminary Draft Please Do Not Circulate

Ari Anisfeld *
University of Chicago
Job Market Paper

Pepi Pandiloski
University of Chicago

September 14, 2024
Click For Latest Draft

Abstract

This study investigates the impact of language skills on inter-group contact and the demand for segregation in segregated high schools in Macedonia, where students are divided by language of instruction. Using a lab-in-the-field experiment, we randomly assign students to cooperative tasks with either co-ethnic or cross-ethnic partners; random pairs provide variation in mutual language skills. Participants in cross-ethnic teams communicate less and rate the experience lower than those in co-ethnic teams. Nonetheless, inter-ethnic contact generally enhances social cohesion without a statistically significant effect on productivity. The cohesion effects are stronger among pairs with high language-skill alignment. Focusing on these effects, we use a weighting procedure to isolate match effects – the benefit of having shared language skills – from selection effects – the association between unobservables and language skills. The exercise suggests positive match effects account for two-thirds of the positive cohesion outcomes.

*Corresponding author: anisfeld@uchicago.edu. Thank yous coming soon ... 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 our implementation partners at the Center for Human Rights and Conflict Resolution, led by Luan Imeri and Safet Ballazhi.

1 Introduction

Segregation reduces the opportunities for social learning and productive exchange that are the hallmark of flourishing diverse societies (Alesina and Ferrara, 2005).[1] Yet, segregation can be desirable. For example, learning in one's mother tongue can be more efficient, particularly early in life, so segregation can support learning in schools serving distinct ethnolinguistic groups.[2] Indeed, the World Bank and other powerful international institutions have promoted mother tongue (L1) instruction as a good education policy. In the Macedonia high schools where our study is set: students of three different ethnicities go to the same schools but are segregated into separate classrooms based on language of instruction. This tradeoff between better learning and segregation poses a challenge in multi-ethnic countries from India to South Africa to the United States (e.g. Kumar et al., 2022; Heugh, 2013).[3]

Students separated in schools may develop skill sets that are specialized to segregated environments and embrace exclusionary identities, leading to a social dynamic in which segregation begets segregation. Linguistic segregation specifically, may erode the ability for communication between the groups, leading to even greater segregation (Lang, 1986).A common policy prescription is based on the contact hypothesis (Allport et al., 1954), which proposes that schools organize activities where students from different groups can cooperate toward a common goal. But contact between groups has had mixed results and has even backfired (Rao, 2019; Elwert et al., 2023; Scacco and Warren, 2018; Ghosh et al., n.d.). Local context ought to influence when and how inter-group contact leads to higher cohesion. Language skills developed in segregated environments may limit the effectiveness of contact-based interventions. In this paper, we ask how language skills shape inter-group contact. If two students have a better ability to communicate between themselves, do they have a better experience? Do they perform better on cooperative tasks? And how do measures of inter-group cohesion change as a result?

We investigate these questions in high schools in Macedonia where schools segregate ethnic Macedonian,

[1] Alesina and Ferrara (2005) also argue diversity can create frictions in the provision of public goods. On the other hand, when segregation occurs in unequal societies increased public good provision may exacerbate inequities (see e.g. Asher et al., n.d.).

[2] Other examples include lowering transaction costs, maintaining religious or cultural heritage, and forming social or political groups that serve the needs of particular affinities.

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

Albanian, and Turkish students by language of instruction. Field observations, interviews and focus groups suggest that segregated learning, media access and labor markets have increased segregation.[4] In 2001, an armed conflict between Albanian militia and Macedonian authorities was resolved with the Ohrid Agreement, which expanded the language rights of Albanians including providing public funding to Albanian-language colleges.While Albanians and Turks had access to home language instruction before 2001, the associate policy changes reduced access to and need for Macedonian language instruction (Fontana, 2017). Educators in our context point to two other drivers of segregation: media and labor-market orientation. Young people increasingly have access to home-language and international media. Further, many students aspire to work in richer European countries or the United States, where the returns to communication skills specialized to their hometown may not yield large returns. All these may erode mutual language skills, exacerbating already low cross-group interaction.

To study how identity and language skills drive outcomes during and after inter-group contact, we employ a randomized contact intervention and two waves of surveys. We randomize students into pairs to play "Overcooked! 2", a cooperative video game that requires teamwork and communication as players rush to prepare meals for hungry customers. We rely on two sources of variation. First, we randomize inter-group contact: control students get partners from the same language of instruction while treated students get partners from a different language of instruction. We stratify pairs by school, grade level and gender so that ethnicity is the most salient difference in cross-ethnic pairs. Second, by randomly pairing students, we generate random variation in shared linguistic skills. We use this underlying variation to understand the heterogeneity of effects due to shared language skills. Finally, within treated pairs, we subtly alter the framing of gameplay to get students to think about shared identities. Though segregated, schools potentially create cross-cutting identities (e.g. school membership, people who hang out in the courtyard between school shifts). Can shifting focus to a shared identity improve cohesion outcomes and gameplay?

We measure two intermediate and two ultimate outcomes. The intermediate outcomes capture the amount of communication between the players and the players' experience during the session. Both are measured using survey responses from the players and the session facilitators. The two main outcomes are cohesion and productivity. We measure cohesion with incentivized dictator games and an adaptive choice experiment (Drake et al., 2022). The first measure, common in the literature, is the difference in the amount of money shared

[4] See also Fontana (2017).

with co- and cross-ethnic peers. The second measures the willingness to pay (WTP) for a same-ethnicity partner. The adaptive choice experiment allows us to estimate individual-level parameters using an adaptive sequence of trade-offs akin to a binary search. Individual estimates let us understand the heterogeneity of effects by comparing demand curves induced by the WTP estimates. To measure productivity, facilitators provide reports of student scores and we also capture screen recordings with audio from pair gameplay. While we currently only employ reported scores, we are in the process of extracting other useful productivity measures from the audio-visual data.

We find that students with cross-ethnic partners have more challenges with communication and rate their experience lower than students with co-ethnic partners. Nevertheless, students with cross-ethnic partners show cohesion increases. On average, the cohesion index increases by 0.19 standard deviations. This change is driven by the students with initially average and high demand for co-ethnic partners. The improvements in cohesion do not come at the cost of lower productivity, as scores in the game are not statistically distinguishable between the treatment and control groups. Thus, students in cross-ethnic pairs overcome verbal communication challenges to perform as well as their peers in co-ethnic pairs in the game, while improving their attitudes toward peers from different ethnicities.

We then turn to the role of language skills. We divide the cross-ethnic pairs into two groups based on being above or below our measure of language skill match.[5] We find that pairs with a better ability to communicate unsurprisingly talk more to each other and have a better experience during the interaction. While language skills do not seem to shape productivity during the interaction, the students in pairs with higher communication ability appear to drive most of the improvements in cohesion outcomes. In the demand for segregation curves, the students in pairs with low language match are statistically indistinguishable from the control students. Meanwhile, the demand curve for students with a high-linguistic match bows sharply inward. Students in co-ethnic pairs share 22 additional Denars with anonymous recipients who share their mother tongue compared to anonymous recipients who do not. This number fell by 10 Denar for students with a high-language match cross-ethnic partner.

The heterogeneity by language match results are non-causal, with two major threats to our story. First, we

[5]We use self-reported language skills to determine a vector of skills in Albanian, Macedonian, Turkish and English on a scale from 0 to 10 and then determine language skill match as the maximum of the squared geometric mean (a number between 0 and 100).

may face "selection bias" as baseline language skills influence the probability an individual is in a high- or low-match pair and appear to be correlated with other factors, such as cultural awareness or sociability, that may make them more susceptible to treatment. To overcome the selection bias, we turn to inverse propensity weighting. Through simulations of the randomization procedure, we calculate the probability a given student will be assigned a high- or low-language skill match partner. Then, using weights, we get unbiased estimates of the local average treatment effects of being assigned a high language skill match partner compared to a low match or co-ethnic partner.[6] When re-estimating results accounting for probability weights, results maintain a similar direction but the estimates are attenuated.[7] While we still see a statistically significant impact in the cohesion index for high-match pairs, the difference between high and low match become statistically insignificant, as do cohesion sub-measures.[8] Second, as hinted above, the "high language match" indicator is a bundled treatment and the focus on language may be unwarranted. In future work, we will test competing theories using data at our disposal.

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

We complicate the contact hypothesis by considering the role of contextual factors such as skills and preferences. Under the contact hypothesis, integration produces positive outcomes if individuals from distinct groups work together toward a common goal where groups have equal status and the activities are condoned by authorities. (Allport et al., 1954; Pettigrew and Tropp, 2006; Paluck, Green, et al., 2019) We build on recent work with youth that has found mixed evidence — some showing strong positive results in attitudes and behaviors (Rao, 2019), others indicating changes in behavior but not attitudes (Scacco and Warren, 2018), and still others reporting generally positive results, which degrade as the proportion of out-group

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

[7] This could be due to a data factor that obfuscate the underlying mechanism; we must drop students with sure probability of high or low match for this analysis, limiting the sample size and also increasing the impact of measurement error since the language skill match is measured imperfectly. We are working on understanding this as we continue analysis.

[8] If these results hold as we account for the issues in the previous footnote and gather additional data, this would suggest that the beneficial effects of contact on cohesion are driven by traits associated with high language match probability more than the language match per se. The policy implication would be to focus on building these related skills (e.g. perspective taking a la Alan et al., 2021).

team members increases (Ghosh et al., n.d.). However, mechanisms for when and how contact works are not well understood. In a 2021 review, Paluck, Porat, et al. (2021) called for studies that unpack mechanisms, involve large sample sizes (the upper quartile is 78 participants in the extant research cited), and measure outcomes at least one day after contact. Earlier contact papers treat individuals as exchangeable based the study-defining group membership; such as racial category (Corno et al., 2022), religious background (Scacco and Warren, 2018; Mousa, 2020), caste (Lowe, 2021), economic class (Rao, 2019), or cross-cutting group identity(Bursztyn et al., 2024; Finseraas et al., 2019). Incorporating baseline skills, we show how individual heterogeneity and partner match effects can influence the individuals experiences and outcomes of integration activities. Since skills can be learned (e.g. Alan et al., 2021) and assignments to contact experiences can be targeted, understanding this heterogeneity is important.[9]

Finally, we provide insights for experimentalists interested in working with pairs and measuring productivity and cohesion. The use of pairs in experimental settings raises several complications and opportunities. We introduce novel techniques for implementing and analyzing such experiments. While there are a number of studies that focus on partners such as roommates (Boisjoly et al., 2006; Corno et al., 2022) or deskmates (Elwert et al., 2023), these studies tend to use simple randomization schemes and often focus on the results from a "focal" partner. We randomize in a manner that ensures specific proportions of partner types while minimizing the recruitment costs for a given sample size. Further, we borrow from the literature on spillovers to estimate match quality effects based on predetermined characteristics. Further 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 measures of productivity and effort in team setting.[10] Additionally, as described above adopted a measure from the labor economics literature on valuing workplace amenities to estimate the willingness to

[9]We also contribute to the literature on endogenous identities (Akerlof and Kranton (2000); Chen and Li (2009); Tajfel et al. (1971); see Shayo (2020) and Charness and Chen (2020) for reviews). In an additional manipulation, we frame game play around shared group identities–school and town-membership. While ethnicity and language are strong markers of identity, these group memberships can be as well. (Using the "inclusion of others in self" survey-based measure, students report closeness to their school and ethnicity. On average, ethnicity is roughly one point higher on the 7-point scale.) Previous research successfully primed aspects of individuals' multidimensional social identities to observe impacts on academic outcomes, prosocial behavior, honesty, and social preferences (Steele and Aronson, 1995; Shih et al., 1999; Nassir and Diesendruck, 2023; Cohn et al., 2017; Chen, Li, et al., 2014; Rubenson and Dawes, 2022). For example, Chen, Li, et al. (2014) found that priming school membership or ethnic identity among University of Michigan students of Chinese descent influenced cooperation, with common identity primes leading to increased cooperation with randomly assigned lab partners in behavioral games. In contrast, we find no evidence that priming shared identities influences behavior.

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

pay for segregation at an individual level (Drake et al., 2022).

2 Setting

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

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

Questions about language policy underscore long-standing political tensions about the role of *nationalities* and cohesion in the multi-ethnic state.[13] Indeed, the role of language in education has been contested since the expulsion of the Ottoman Empire in the early 1900s. Since at least 1950, governments have funded Albanian-language primary and secondary classrooms in the region, though at times government support was low.[14] After the dissolution of Yugoslavia, Albanians opened an "illegal" Albanian-language university

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

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

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

[14] For example, nationalistic policymakers reduced access to non-Macedonian language instruction through the Law of Secondary Education 1985 (Official Gazette of SRM No 16/85). The law required 30 students from a language group in a cohort to have a secondary school class in their preferred language. As a result, the number of high schools offering Albanian language classes fell

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

In practice, students from distinct ethnic backgrounds attend the same physical schools but are segregated by language of instruction from kindergarten through high school, limiting opportunities for cross-ethnic social interaction. As an illustrative example, consider two fictitious students: Amra, an Albanian girl, and Ana, a Macedonian girl, both born in the same year. Their parents enroll them in Goce Delcev Primary School where they attend through the eighth grade. Amra and Ana are rarely on campus simultaneously because the school day is divided into shifts organized by language.[17] Amra enrolls in the Technical High School, which also organizes shifts by ethnicity, providing minimal opportunities for cross-ethnic interactions. Ana attends the Gymnasium, a college preparatory high school, which organizes the day by grade level. Although she begins the day with all first-year students regardless of ethnicity, her classes are conducted in Macedonian, and her peers are predominantly ethnically Macedonian. Additionally, Macedonian classrooms are in different hallways from Albanian and Turkish classrooms. Thus, despite increased opportunities for cross-ethnic encounters, Ana's K-12 experience also remains largely segregated.[18] Outside of a school setting, students might have opportunities for cross-ethnic contact if they grow up in relatively integrated neighborhoods or participate in activities outside of school.[19]

from 39 in 1980/81 to 5 in 1990/91 (Ortokovski, see p. 362). The numbers rebounded post-independence.

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

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

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

[18]The Medical High School, the third high school in our study, also organizes shifts by cohort. These students take practical courses in the hospital, which may provide additional opportunities for mixing.

[19]Mixed families are rare–in our data, only 6 students report multiple ethnicities. However, home life could provide opportunities

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

In this segregated environment, baseline measures show low levels of social integration and cohesion (See Table 1). First, social networks are quite segregated. We elicit social networks using several methods including explicitly asking students to name up to six cross-ethnic peers. Only 45 percent of students name at least one peer who speaks a different mother tongue; compared to 96 percent who name at least one friend. In total, 88 percent of all distinct named links are co-ethnic.[23] Second, we find that students feel very close to students from their language-of-instruction (LOI) group and quite far from students from different LOIs. We measure this using the *inclusion of self-in-others* measure where a respondent selects one of 7 pairs of circles that range from barely touching to nearly completely overlapping (Aron et al., 1992). One is labeled "you" and the other has the name of a group (i.e. "Turks", "students from the Medical School"). We ask "what picture best describes your relationship with [group identifier]?". On average, students give their in-group LOI a 6.4 out of 7 while out-group LOIs average 2 out of 7 (barely overlapping circles). Finally, students share less with out-group members in dictator games. On average, there is a 12 percent gap (See section 5).

for mixing as well. We do not have data on the spatial extent of segregation in Gostivar–from discussions with locals, there are both mixed and highly segregated neighborhoods and villages.

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

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

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

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

3 Study Design

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

Recruitment and Timeline

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

Data comes from three sources: (1) a baseline survey, (2) video game sessions, and (3) an endline survey. Upon receiving parental consent, enumerators conducted the baseline survey that featured modules on language skills, social networks and social closeness, parents, video games, educational outlook, and real stakes dictator games.[24] Once both partners in a pair completed the baseline survey, we sent the session information to the facilitation team who organized the video game sessions described below.[25] Then at least two weeks later, enumerators conduct the endline survey asking students to sign up for our inter-ethnic course, and re-asking social network and behavioral measure questions from the baseline survey. Many of the same participants were invited to participate in our companion study, where students, in groups of 6, play a networked social learning game, passing information to peers via a website we built for the study (Anisfeld and Pandiloski, 2024). The second study involves limited cross-group interaction as students are instructed to remain quiet. Nevertheless, we strive to have students complete the first three steps before participating in Study 2.

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

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

Randomization

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

Main intervention: video game sessions

In the session, pairs of students learn to play "Overcooked! 2", a cooperative video game where players work as chefs to prepare as many orders as possible under time pressure. The main source of variation is the ethnic composition of the pair, or from the student's point-of-view whether the partner has the same or different ethnicity.

During the sessions, facilitators led students to a room with two computers prepped for survey taking and video game play.[27] Students introduce themselves and answer two ice breaker questions "What languages do you speak?" and "What is your favorite class?" Then, using Qualtrics on separate computers, students verified their name and birth month and proceeded to learn how to play the game from YouTube videos[28] and then played practice rounds in "solo mode", where an individual player controls both chefs.[29] The "solo mode" scores provide a measure of individual skill.[30] Then, students play together in "cooperative mode" on

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

[27] We hired teachers from the school to organize and run sessions. Each session had two teacher facilitators so that each student had at least one teacher who taught in their language of instruction, or was fluent in their language of instruction. We directed facilitators to interfere with game play as little as possible, and not at all during the "official play".

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

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

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

a single computer, each student controls one chef and they work together to provide sushi (level 1.3) or pasta (level 1.5) to hungry, impatient customers. Pairs play each level twice, a practice and an "official round", wherein students get payouts based on the team score. In addition to learning and playing "Overcooked! 2", students answered questions about their experience in the session, completed an adaptive choice experiment, provided additional behavioral measures. Based on their scores in "official rounds" and dictator game responses, we calculate payouts that the field team leaders deliver at a later date. The entire session lasts an hour on average, with pairs interactively playing the game for roughly 20 to 25 minutes.

"Overcooked! 2" has many advantages.[31] First, students work cooperatively towards a common goal on equal footing, matching the criteria set out by the contact hypothesis. Second, productivity in the game depends on good communication and cooperation between players. To give some flavor, level 1.5 has chefs prepare spaghetti with red sauce. To prepare one dish requires boiling pasta, chopping tomatoes, sauteing tomatoes, plating the pasta and the sauce before either one becomes "overcooked"[32], and finally delivering the plated dish to the window. As orders go out, plates get dirty and must be washed. On top of this, the kitchen is in hot air balloons that shift over time as wind gusts blow them and the chefs across the screen. While each chef could work independently, like Adam Smith's pin factory, there is profit in the division of labor. We hypothesized that language skills would have a meaningful impact on students' ability to divide labor and score points. Finally, video games are a culturally appropriate medium for youth contact. In our sample, two-thirds of students could recommend a video game by name, 55 percent said they played some days or more frequently and 72 percent said the expression "I love video games" describes them somewhat or more.[33] The figure is 87 percent for boys and 60 percent for girls.

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

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

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

4 Data and Measures

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

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

Above, we reported baseline cohesion measures and video game skills. The treatment group in the current sample has some hints of imbalance. They name .14 more cross-ethnic links than the control ($p < .05$) and have lower dictator game bias -4 denar ($p = .67$), yet they give less money to outgroup peers in dictator games (-15 denar, $p = .1$). In robustness checks, we include dictator game bias as a covariate for cohesion outcomes.[37]

Language skill match

Language skill match is a key component of the analysis. Partners may or may not share proficiency in English, Macedonian, Albanian, or Turkish. *A priori*, to what extent do partners have communication skills

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

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

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

[37] In additional checks, not included in the appendix, we find cross-ethnic links and outgroup giving do not impact the cohesion estimates. In future iterations, we will use a data-driven approach to incorporate many covariates to provide a robustness check accounting for imbalances.

to interact with each other in cooperative play, and how does this influence their experience, productivity, and cohesion?

To measure the language skill match, we employ a commonly used survey instrument the Language Experience and Proficiency (LEAP) Questionnaire (Marian et al., 2007). Students self-rate their English, Macedonian, Albanian, and Turkish skills using a scale of 0 to 10 in four domains: speaking, listening, reading, and writing. We average to provide one score per language. Then, for pair i and j, we determine a joint score in a given language by multiplying i and j's scores.[38] We define *language skill match* as the maximum joint score over the four languages. Formally, each student i has a vector of linguistic skills $f_i \in [0, 10]^4$. Then, for pair (i, j)

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

where $\odot$ denotes the element-wise (Hadamard) product and the $\max$ operator selects the larges element in the vector. The 'language skill match' reflects the two-way nature of communication and the inherent complementarity of skills required to communicate. The "max" captures first-order communication skills.[39] For co-ethnic pairs, $f_{\max}$ is typically $10 \times 10 = 100$. For cross-ethnic pairs, we see substantial variation in $f_{\max}$. It has a median of 75, an interquartile range of 62 to 88, a minimum of 10, and a maximum of 100.

Outcomes

We build 4 indices to measure communication, student experiences, productivity and cohesion. For each outcome type, we construct indices of 2 or 3 related variables. To make sure variables are comparable, we normalize by subtracting the control group's mean and dividing by the control group's standard deviation. Then, we take the simple average of all composite variables to generate the indices.

Outcomes: Cohesion Measures

The cohesion index comprises

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

[39] In future iterations, we may explore a multiple-dimensional measure, such as cosine similarity.

1. Willingness to pay for co-ethnic partners

2. Dictator game bias

Adaptive Choice Experiment: Willingness to pay for co-ethnic partners We measure individual-level willingness to pay for co-ethnic partners (segregation) and their partner's language skills using a choice experiment in which we cleverly select the next choice the student faces. Students make choices that trade-off qualities of a future partner for the end-of-year celebration, choosing between two options with specified ethnicity, languages spoken and a sure payment.[40] Intuitively, students who chose co-ethnic partners over a cross-ethnic partner with a larger bonus are willing to forgo money to maintain segregation. Students face 12 such questions; we use a Bayesian procedure to determine the next trade-off allowing us to estimate multiple individual-level parameters (Drake et al., 2022).[41] This method enhances our ability to understand heterogeneity relative to earlier studies that use a single choice decision or a "multiple price list" to target a price or demand curve for a single characteristic (e.g. Hedegaard and Tyran (2018), Rao (2019), and Ghosh et al. (n.d.)).

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

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

Where p is a sure price paid for choosing partner j with a utility normalized to 1. α_i is the WTP for a co-ethnic partner and β_i^L is the willingness to pay for student second language skills in L, English or Macedonian. In Appendix C, we detail the theoretical basis and estimation in detail.

The utility parameter α_i captures the willingness-to-pay for a co-ethnic partner holding language skills constant. The parameter cannot be interpreted as a direct measure of discrimination because students may have less pernicious utility gain for valuing the co-ethnic partner. For example, they may consider the possibility of matching with a friend and our network data shows that friendships tend to be co-ethnic. However,

[40] 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."

[41] We implement BACE using software developed by DPTT (2022).

changes in α (i.e. a treatment effect) can be interpreted as a change in the "price of prejudice".[42]

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

As a sanity test, we examine correlations between language skills and demand for students who speak English or Macedonian as a second language. Students with an additional point of fluency in a given language are willing to pay 3.5 to 5 denar for a partner who speaks that language.[43] Appendix Figure A2 shows demand for partners with specific language skills disaggregated by LOI. The median Macedonian LOI student is willing to pay over 100 denar for their partner to know Macedonian as a second language, whereas the median non-Macedonian LOI students has a WTP of 0.

Bias in dictator game sharing

Dictator games are widely used to measure prosocial preferences of altruism or fairness norms and bias. In a *dictator game*, each participant is asked to split 600 Denar with a peer. Each question reads: "Imagine you are splitting 600 denars between yourself and the randomly chosen **student from your high school with the same mother tongue as you**. They will not know who split the money. Decide how much of the 600 denars to send to the **student from your high school with the same mother tongue as you**. You will keep the rest of the money." Students make four splitting decisions varying the mother tongue ("the same" or "a different") and school membership ("from your high school" or "from another high school") of their otherwise anonymous partners. Some choices are realized, adding an incentive to for truthful reporting. These games have been used successfully with high school students and younger children (e.g. Ghosh et al., n.d.; List et al., 2017). We measure out-group bias by subtracting the average amount shared with a partner who shares their mother tongue from the amount shared with a partner who does not. In the baseline survey students share 29 Denar (13.5 percent) more with their co-ethnic partners.

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

[42] Under the assumption that the effects are not driven by updating statistical biases.

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

Real-world choice experiment–"Willingness to sign-up for inter-ethnic coursework": In the endline survey, students are asked if they want to participate in 5-session inter-ethnic English courses, hosted by the Center for Human Rights and Conflict Resolution (CHRCR). We hint that the course will be inter-ethnic by highlighting language support for students less comfortable with English ("While the course will use English for communication, there will be support for students in Macedonian, Albanian and Turkish language. And the skills you learn will be applicable in any language you speak!"). Any differences in sign-up rates across treatment arms provides additional evidence of effects on demand for segregation. While the course is free, we also elicit demand curves using a hypothetical take-it-or-leave price experiment, providing additional information.

Outcomes: Productivity

The productivity index comprises

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

Our primary measure of productivity is the total points scored in the "official rounds" of cooperative play. Players get points for successfully delivering dishes; they get tips for delivering dishes quickly and even more tips for delivering the dishes in the same order which they came in; conversely, players lose points for failing to deliver an order in time.[44] The total score encapsulates total productivity.[45] Because students play in solo mode first and we record their scores, we have access to a baseline video game ability for each student (though these may be influenced by treatment assignment.)

Outcomes: Communication

The communication index comprises

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

[45] In future work, we will use image detection algorithms to extract additional information from the 120 hours of screen recordings. While total score provides a holistic measure of productivity, the number of failed orders might provide a distinct measure of cooperative effort. We can also fairly easily extract the time series of points scored, which could provide a richer picture of playing styles. More challenging computationally, we could train a model to identify which actions go into making a dish and determine how much joint effort goes into the point totals.

1. Amount of communication, rated by facilitators

2. Ease of communication, rated by students

Facilitators rate the amount of communication and which languages they heard used. Facilitators rate "How much did the students communicate during the joint game?" on a seven-point scale. Inter-rater correlation is high (.78) and the measure is associated with student rating on easy of communication (.47), our second measure in the communication index.[46]

Outcomes: Experience

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

1. Enjoyed playing

2. Would play again with the same partner

3. Was partner fun?

The student measures are highly correlated within students, correlation coefficients range between 0.6 and 0.73. Facilitators also report, from their vantage, how much fun each player seemed to have and how frustrated they seemed to be. These measures correlate less tightly, but in the expected direction with student self-measures. Correlation between facilitator-assessed enjoyment and student self-measures range from .27 to .33, while that between facilitator-assessed frustration and student self-measures range from -.23 to -.30. The concordance between measures gives us confidence that student self-assessments are reliable.

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

5 Empirical Strategy

5.1 Main results

To estimate average treatment effects, we run the regression:

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

and

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

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

5.1.1 Testing identity treatment arms

We also run the first regression separately by identity treatment arms. Then τ is interpreted as the treatment effect of having a cross-ethnic partner with 1) no additional identity treatment, 2) with a focus on common identities or 3) with a focus on common identities and a common enemy.

5.1.2 Robustness with causal forests

We use causal forests, a "doubly robust" treatment effects estimator, to test the robustness of the main results against finite-sample bias due to chance imbalances and against biases due to variation in individual-level

treatment propensity scores (Athey, Tibshirani, et al., 2018). At a high level, we include predetermined covariates to address chance imbalances and use inverse propensity weights to address variation in treatment scores.[47] 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).[48] The estimator combines these in the causal forest to estimate conditional average treatment effects given x:

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

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

The assumptions required for the causal forest to consistently estimate treatment effects are selection on observables (conditional exogeneity) and overlap. We include all the observables that are included in the randomization procedures, so this is a use case where that assumption holds. For the main analysis, nearly all students have a probability of being assigned treatment or control due to our randomization procedure.[49] With these assumptions met, the algorithm returns unbiased estimates of the average treatment effects with

[47] 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.

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

[49] There are two students who are the only representative of their language of instruction within a strata and so we technically cannot assign them to the control condition.

asymptotic normal standard errors, which are clustered following Athey and Wager (2019).

5.2 Language skill match

We next turn to the role of language skill match, focusing on treated pairs with above or below median "language match". Recall, that our measure of language skill match is proportional to the maximum geometric mean of fluency in local languages or English. Researchers often conduct heterogeneity analyses using simple cuts of the data. Following this strategy, we start with a descriptive heterogeneity analysis, using the following regression to estimate treatment associations on high or low language match.

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

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

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

6 Results

6.1 Contact improves cohesion without lowering productivity

Figure 2 shows treatment effects for indices and the items that compose them in standard deviation units, along with the ATE estimates from the causal forests (shown in the lighter shade). In Table 5 and 6 we report the same results and also include covariates.

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

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

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

The final three rows of Figure 2 show that playing with a cross-ethnic partner improves cohesion measures. The cohesion index comprises dictator-game bias and willingness-to-pay for co-ethnic partners, multiplied by negative 1. Cohesion increases by .18 standard deviations, reflecting similar magnitude statistically significant declines in bias and willingness-to-pay. In levels, dictator game bias falls from 22.8 to 14.4 Denar, and students average. In total, we see an average decline in bias and willingness to pay for segregation despite students expressing less good opinions about their experience.

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

Appendix Figure A1 shows the effects disaggregated by the identity sub-treatment arms. Overall, while the estimates are noisy, each treatment arm elicits similar outcomes in terms of communication, experience, productivity and cohesion. We will wait for more data to make firm conclusions, but it appears that the common identity treatments are at best second order compared to ethnic identity effects and the language match effects discussed below.

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

Demand for segregation The adaptive choice experiment allows us to measure individual-level willingness-to-pay for a co-ethnic partner. In Figure 3a and 3b, we show the demand curve induced by the individual WTP for, respectively, a co-ethnic partner and for a partner who speaks Macedonian as a second language. The curve reports the empirical probability that an individual's willingness-to-pay is above some price over the support of prices, censored at -100 Denar.

The first thing to notice is that some people, roughly 10 percent, are not willing to pay for co-ethnic partners.[52] Nevertheless, many students are willing to pay for co-ethnic partners and demand is disbursed over the support from 0 to 400 Denar.[53] Across treatment and control roughly 20 percent of students have WTP above 300 Denar, roughly the expected value of participating in the video game session.[54]

While WTP declines by 20 percent on average, the curves show that the "cohesion gains" are not uniform. The demand curves show how the distribution of willingness to pay shifts in response to the intervention.[55] The cross-ethnic demand curve noticeably bends inward relative to the co-ethnic demand curve in the 25^{th} to 75^{th} percentile range. Further, we see a decline in demand at the highest prices. Interestingly, the demand for Macedonian as a second-language has a similar shape for students in Macedonian LOI, but we see no difference between treatment and control units from non-Macedonian LOI tracks. On average, we see no treatment effect on WTP for English (KS p = .32). In other words, among the Macedonian population we see an increased willingness to interact with other ethnicities when they do not speak Macedonian as a second language. This is consistent with the ethnicity results; suggesting marginally more openness to working across ethno-linguistic lines.

6.2 Language match

Figure 4 shows that language skill match is associated with large differences in the overall experience and cohesion outcomes, but not productivity results. First, while all inter-ethnic pairs communicate less than co-ethnic pairs, students in low-match pairs communicate significantly less and expresses more communication

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

[53] Roughly 0 to 8 USD

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

[55] A Kolmogorov-Smirnov test rejects the hypothesis that the demand curves were drawn from the same distribution (p = .016).

challenges than their high-match counterparts. Second, individuals in low-match pairs report significantly lower experience scores compared to those in high-match pairs. Third, despite the communication challenges, both groups show similar levels of productivity. The point estimate is actually higher for the low-match pairs (though statistically indistinguishable). Nevertheless, individuals in high-match pairs thought their partners were better at the game. Finally, we see significant differences in cohesion measures between individuals in high- and low-match pairs. Individuals in high-match pairs have cohesion index measure .27 standard deviation units (se = .05) above the control group and .18 standard deviation units (se = .07) above the individuals in low-match pairs. We see the same pattern in the index's sub-measures, particularly the willingness to pay for segregation. The individuals in high-match pair have a decline .3 standard deviation units (se = .09) or 23 percent lower WTP.

Figure 3c shows the demand for segregation. The biggest demand shift occurs for students in high-match pairs in the middle of the distribution. The students in the low-match pairs have demand curves very similar to the control until the highest prices. The Kolmogorov-Smirnov test rejects the null that high-match and control demand are drawn from the same distribution ($p < .001$) and fails to reject the null for low-match versus control (p = .69).

7 Towards Causal Heterogeneous Effects: Exposure Mapping

Since we randomly assign partners, we can improve on measuring heterogeneity via associations. Selection bias occurs because students with broader language skills are likelier to be in a high-language match pair. Yet, we can estimate how baseline skills influence propensity scores. Then, with almost exact propensity scores, we use weights to estimate causal effects. This is the idea behind "exposure mapping" (Aronow and Samii, 2017). In this section, we first provide an illusory example; then, formally define exposure mapping; and finally, bring the technique to our data and discuss limitations.

Illusory example: A toy example clarifies how selection bias might occur when splitting the sample based on a pre-determined characteristic and how probability weights derived from the randomized design address the bias. Consider a strata with two types of students in a trilingual environment. Bilingual students, labeled B, have fluency in their home language and another language, while monolingual students, labeled M, only

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

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

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

Exposure mapping: Here is how the toy example maps to the empirics from the previous section. We start by extending the potential outcome framework. 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

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

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

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

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

As shown above, we can account for selection bias using propensity scores $\pi_i(z,l)$. The key insight of Aronow and Samii (2017) is that the propensity score is *known* as long as we know the treatment assignment mechanism and the language skill vectors. Through simulations of the randomization procedure, we can estimate $\pi_i(z,l)$ with arbitrary precision. Run the randomization code R times and for each randomization collect the assignment $d(z,l)$ for each individual. Then, $\pi_i(z,l) = \frac{\sum^R \mathbb{1}_{z_i=z,l_i=l}}{R}$.

Finally, we need to discuss two assumptions that may influence the current set of results. *Assumption 1 Overlap:* $0 < \pi_i(z,l) < 1$. In other words, units with no possibility of being in a high or low language match must be dropped as the counterfactual is ill-defined. Roughly 14 percent of the current sample had 0 probability of being matched to a high match (46 students) or a low match (26 students). Appendix Figure A3 shows the distribution of language match assignment probabilities conditional on being assigned a cross-

ethnic partner. The distribution has mass throughout. *Assumption 2 Correct specification*: The exposure mapping is accurate. In future iterations, we will explore this assumption further.

With these, we can estimate the causal effect of being randomly assigned a high-language match partner versus a low-language match partner

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

Figure 7 shows estimation with exposure mapping.[57] The results are similar to those in the previous section but attenuated.[58] While we still see a statistically significant impact in our cohesion index for high-match pairs, the difference between high and low match become statistically insignificant, as do cohesion sub-measures.[59]

Causal forests provide similar results to exposure mapping with some slightly variation (See Figure **??**) and allow us to investigate treatment effects estimated for individuals. Theoretically, the results should be similar because both methods adjust for differences in treatment propensity.[60] The forests also produce [unbiased? consistent?] individual level estimates of treatment effects β_i^H and β_i^L. The range of treatment effects for a low match is between -0.01 and .26 standard deviations, while for high match the estimates range from .099 to .42. This is nice to see because it implies there are very few students for whom treatment "backfires" (at least when measured by the average of willingness to pay for co-ethnic partner and dictator game bias). When we look at individual estimates for

[57] While Aronow and Samii provide a formula for conservative standard errors, I bootstrap for now. The procedure uses the same exposure mapping weights for each iteration.

[58] This could be due to a data factor that obfuscates the underlying mechanism; we must drop students with a sure probability of high or low match for this analysis, limiting the sample size and also increasing the impact of measurement error since the language skill match is measured imperfectly. We are working on understanding this as we continue the analysis.

[59] If these results hold as we account for the issues in the previous footnote and gather additional data, this would suggest that the beneficial effects of contact on cohesion are driven by traits associated with high language match probability more than the language match per se. The policy implication would be to focus on building these related skills (e.g. perspective taking a la Alan et al., 2021).

[60] When using causal forests, we estimate propensity scores ($\pi(z,l)$) in first stage random forests that use sample splitting to avoid over fitting. This smooths the propensity scores away from zero.[61] This may be desirable, since idiosyncrasies in recruitment impacted the propensity scores (see Appendix B.1). Causal forests produce consistent estimates when potential outcomes are independent from treatment assignment conditional on observables $Y_i(1,1), Y_i(1,0), Y_i(0)||D_i|X_i$ and there is overlap. [Overlap failure...].Currently only using individual covariates, but ought to add randomization group or strata level leave-one-out covariates (e.g. mean Albanian skills of non-Albanians). Might lead to more propensity scores close to 0 or 1.

8 Policy Experiments

Next, we imagine a school implementing paired video game play as a tool to build cohesion across linguistic groups.[62] 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.[63] To sketch the idea, we create an edge between each cross-ethnic pair within a strata; then, we calculate the linguistic match, depending on the match type, we assign a weight to the edge that equals the sum of the treatment effect of the two students. Formally, for edge (i, j), if the linguistic match is above the threshold, we assign edge (i, j) weight $w_{ij} = \beta_i^H + \beta_j^H$. If the two students have a low match, we assign them the sum of their respective low-match treatment effects $w_{ij} = \beta_i^L + \beta_j^L$. Then, we find the optimal matching which maximizes the effects, thereby maximizing the cohesion gains.[64] We compare the optimal matching to the expected cohesion gains of a random matching and the infeasible scenario where every student has a high or low language match.

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

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

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

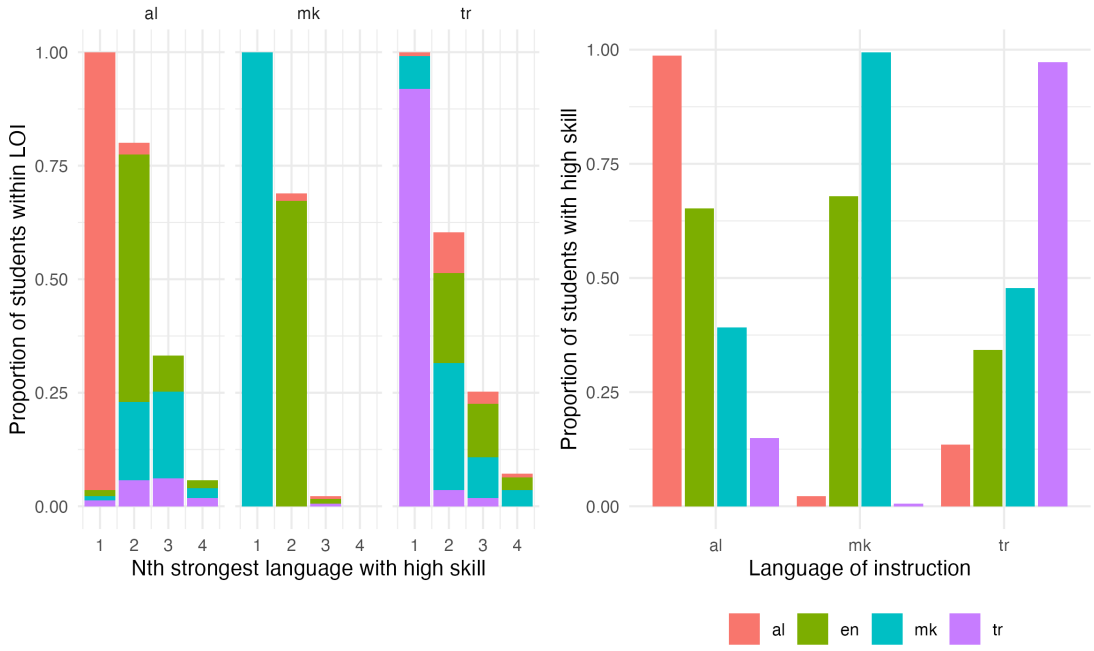
[64] In Appendix **??**, we describe the process in detail.

9 Conclusion

We investigate the role of language and identity in inter-ethnic contact using a lab-in-the-field experiment where students are randomly assigned the same or different ethnicity partners. In line with previous work, we find a moderate positive impact on cohesion outcomes after inter-ethnic contact. However, the main effects mask heterogeneity: contact has significantly stronger cohesion outcomes for students in pairs with high language-skill match. Focusing on these effects, exposure mapping suggests that language-skill match indeed has a causal effect. Still, some heterogeneity reflects "selection"–individuals with a higher probability of language-skill match also have better cohesion outcomes. While interethnic teams tend to have communication challenges, the high-language skill match groups often have similar overall experiences (enjoyment) and productivity as co-ethnic pairs. Nevertheless, the average willingness to pay for segregation is still a large positive number. This highlights that cohesion cannot be built from short one-off experiences, without repeated exposure, homophily forces are too strong. The counterfactual exercise also suggest that increasing the possibility of language match could further improve cohesion outcomes. Moving forward, we will investigate the role of *lingua francas* compared to a language that is one partner's home language to further unpack mechanisms.

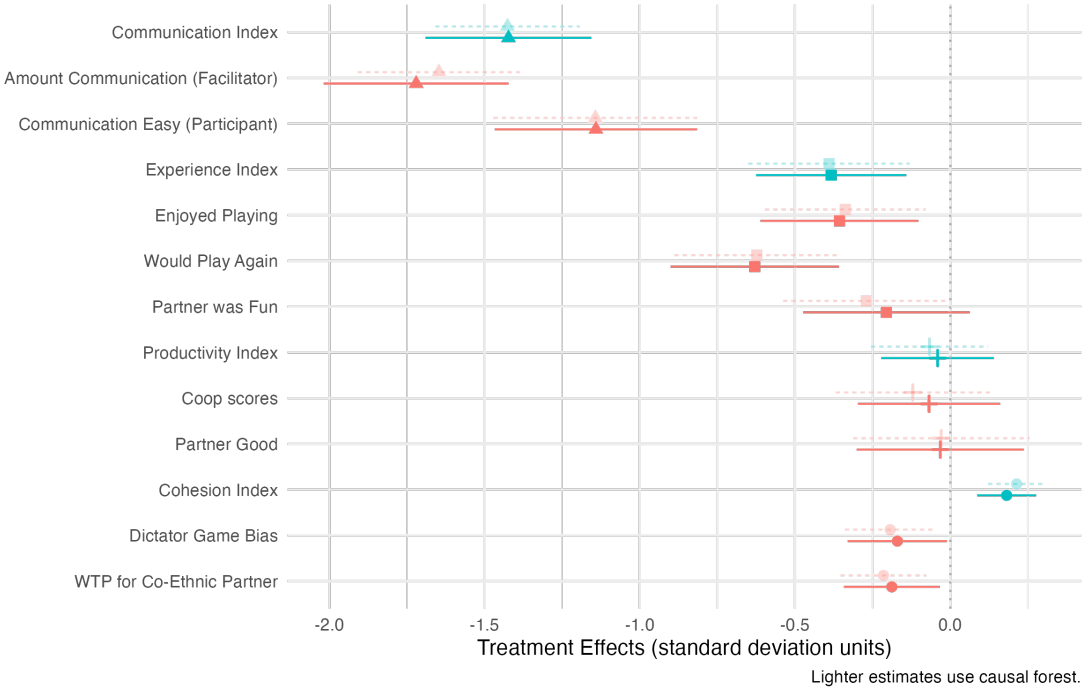
The results hint at a complementary link between language policy and cohesion, suggesting that policies promoting home language (L1) instruction should be paired with additional rigorous language training and, importantly, opportunities for interethnic contact. The causal arrow likely goes in both directions.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

Figure 1: Language skills by language of instruction



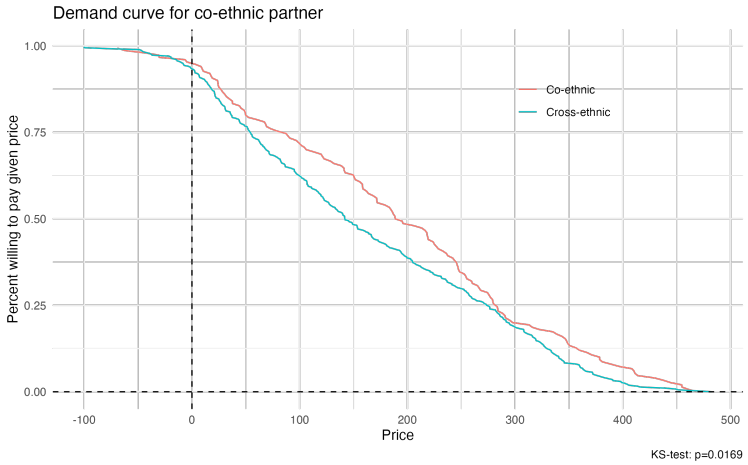
Notes: On the vertical axis both figures show the proportion of students within a language of instruction group who meet some criteria. The right plot shows self-reported language proficiency for the three local languages and English. Proficiency is defined as self-reporting an average of 7.5 fluency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

Figure 2: Treatment effects: Assigned cross-ethnic partner

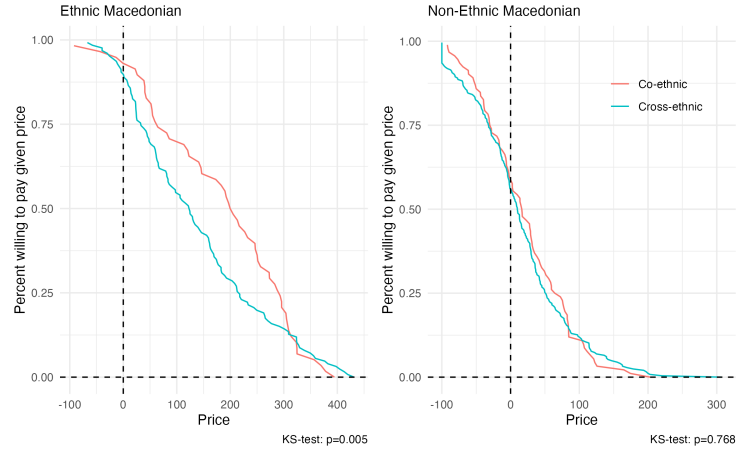


Notes: Coefficients on an indicator for any treatment in regression with no covariates. All sub-outcomes (red) are normalized around the control mean and control standard deviation. Indices (blue) are simple averages of normalized sub-outcomes. Standard errors clustered by strata.

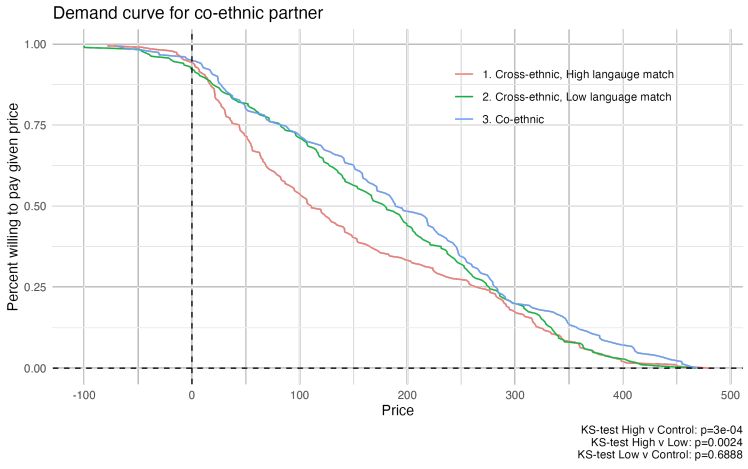
(a) Demand curve: Co-ethnic partner



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

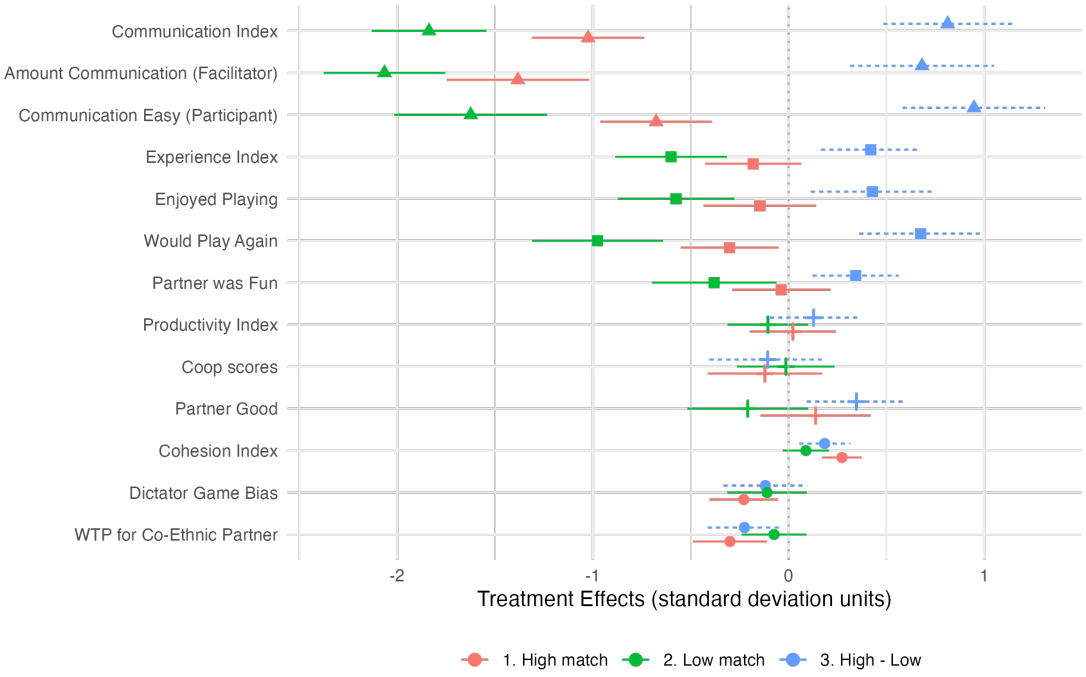


(c) Demand curve: Co-ethnic partner, by language match



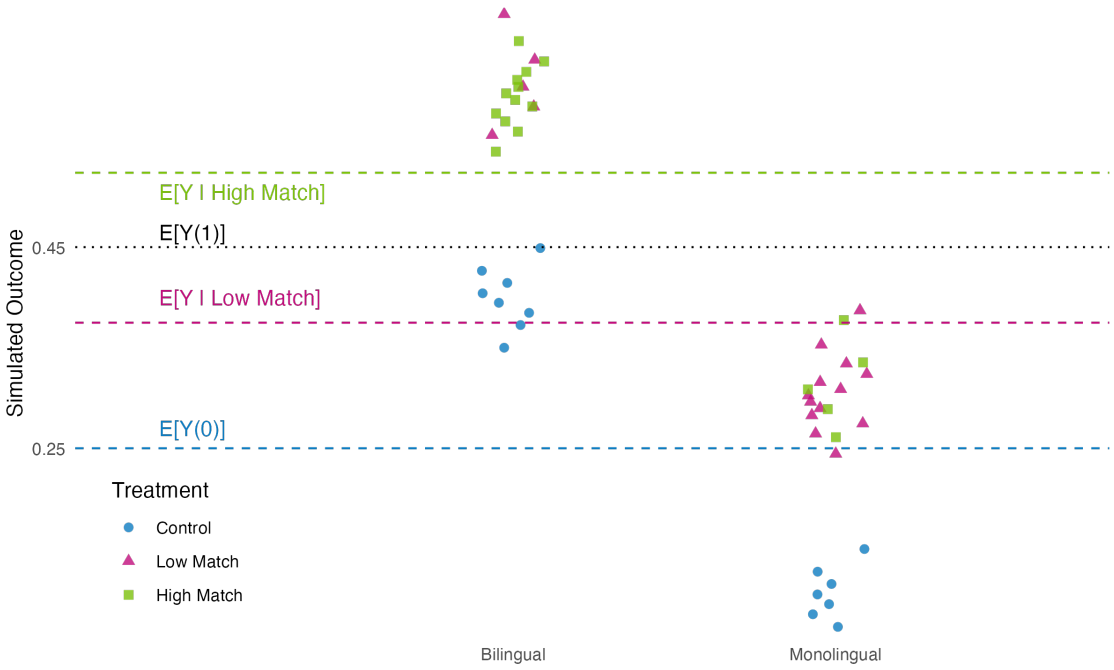
Notes: Demand curves are the cumulative distribution of individual-level willingness to pay estimates. High (low) language skill match defined as above (below) median language match score.

Figure 4: Heterogeneity by language match



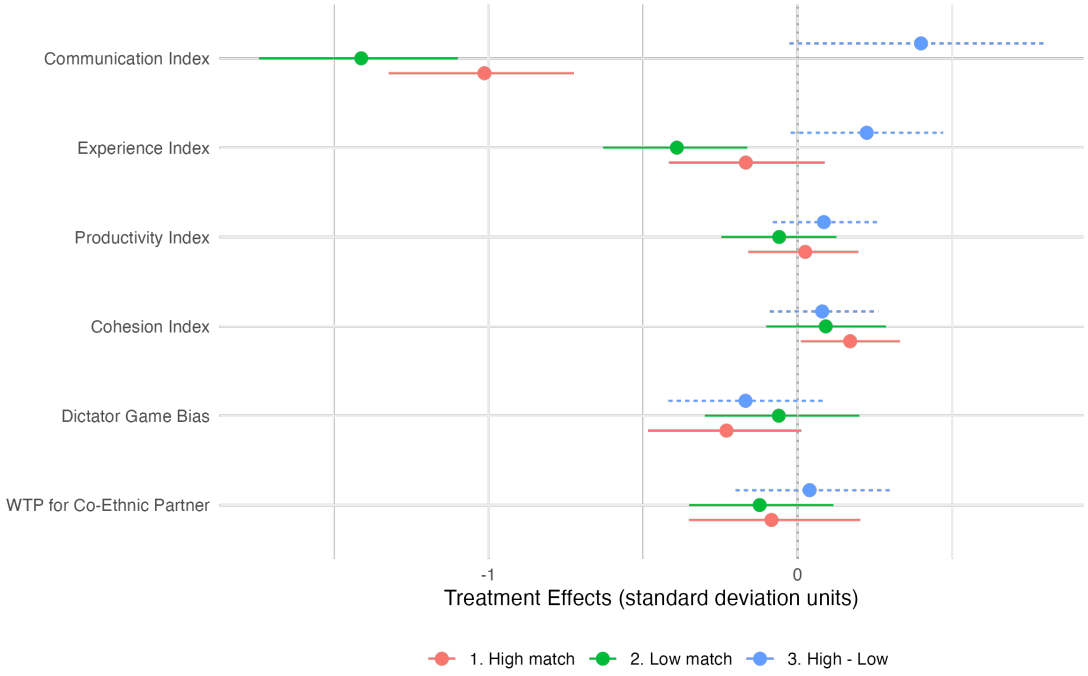
Notes: Coefficients on indicator for 1) high language skill match $\times$ treatment (β^H), 2) low language skill match $\times$ treatment (β^L) in regression with no covariates. High (low) language skill match defined as above (below) median language match score. All sub-outcomes are normalized around control mean and control standard deviation. Indices are simple averages of normalized sub-outcomes. Standard errors clustered by strata.

Figure 5: Simulation demonstrating selection bias when sample splitting



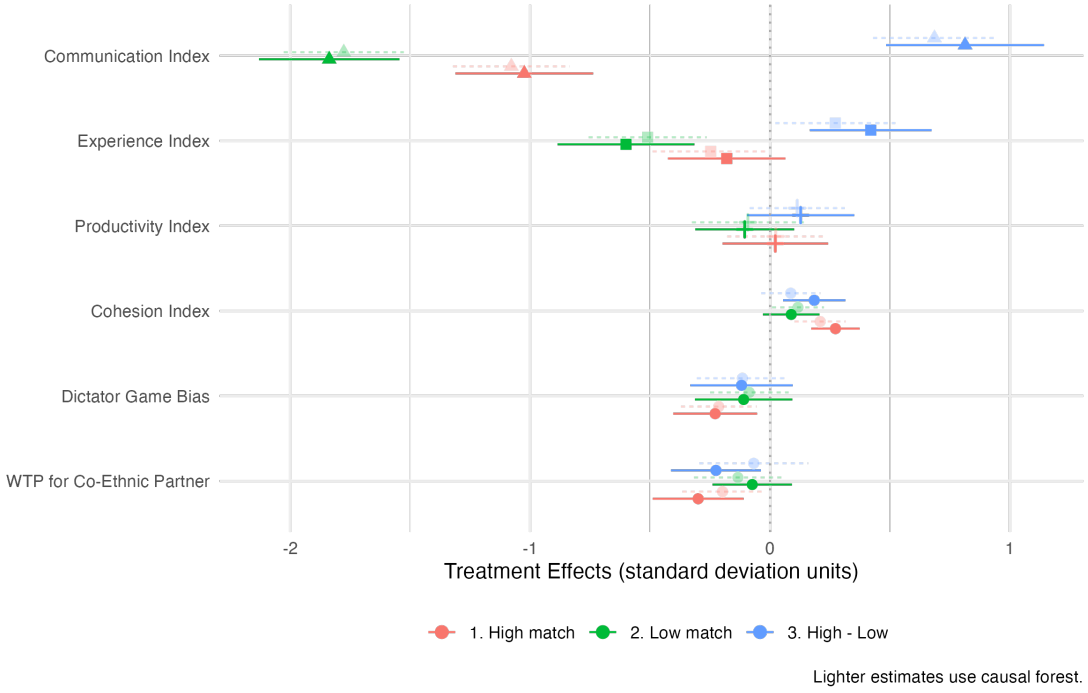
Notes: This stylized simulation shows a stratum with two types of students–bilingual or monolingual–in a trilingual environment. There is a constant treatment effect $\tau = .2$. Each point represents roughly 2000 students randomized into control, high match, or low match as shown with shapes and colors.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

Figure 6: Treatment effects disaggregated by language match (IPW)



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

Figure 7: Language Heterogeneity: Causal Forests and OLS



Notes: Full color estimates show same results from Figure 4. Lighter color estimates show causal forest CATE estimates that correct for differential propensity scores and chance imbalance. High (low) language skill match defined as above (below) median language match score. All sub-outcomes are normalized around control mean and control standard deviation. Indices are simple averages of normalized sub-outcomes.

Table 1: Demographics: Sample characteristics and balance

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 5: Communication and experience heterogeneity by language match

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

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

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

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

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

Table 6: Productivity and cohesion heterogeneity by language match

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

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

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

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

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

Table 7: Counterfactual matchings

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

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

A Appendix

Table A1: Full Sample Balance: Demographics

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

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

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

Table A2: Full Sample Balance: Other baseline characteristics

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

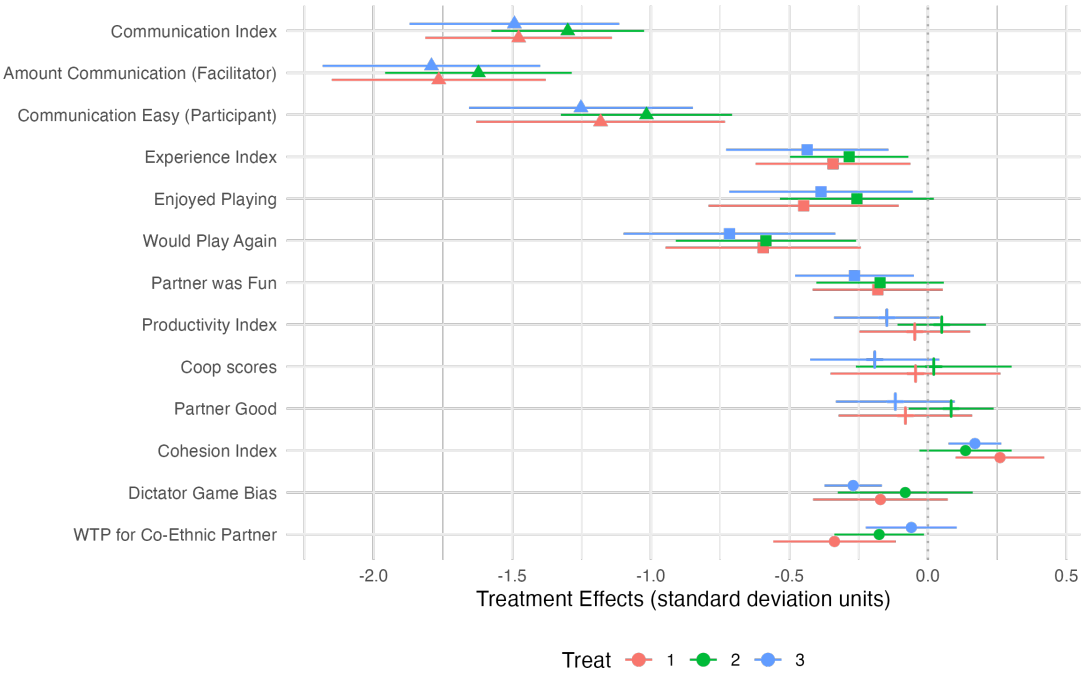


Figure A1: Treatment effects disaggregated by identity prime

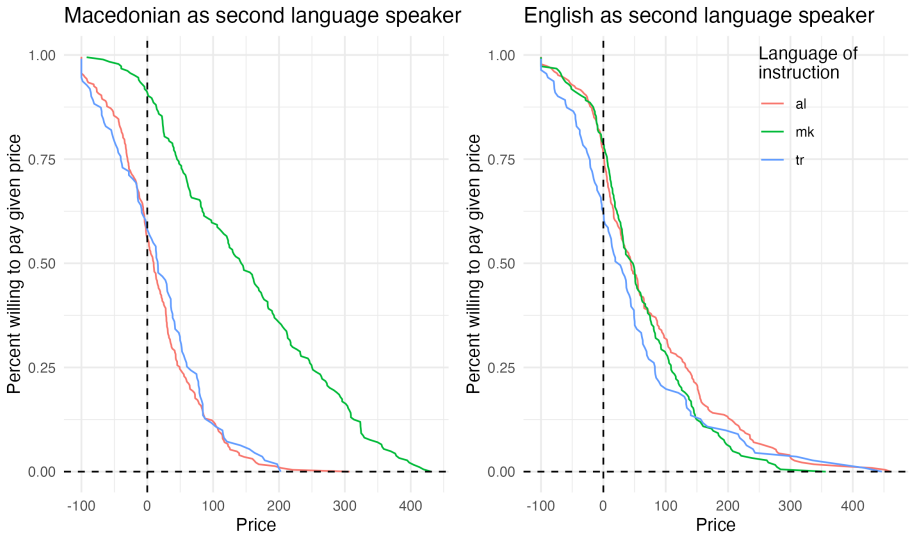


Figure A2: Caption

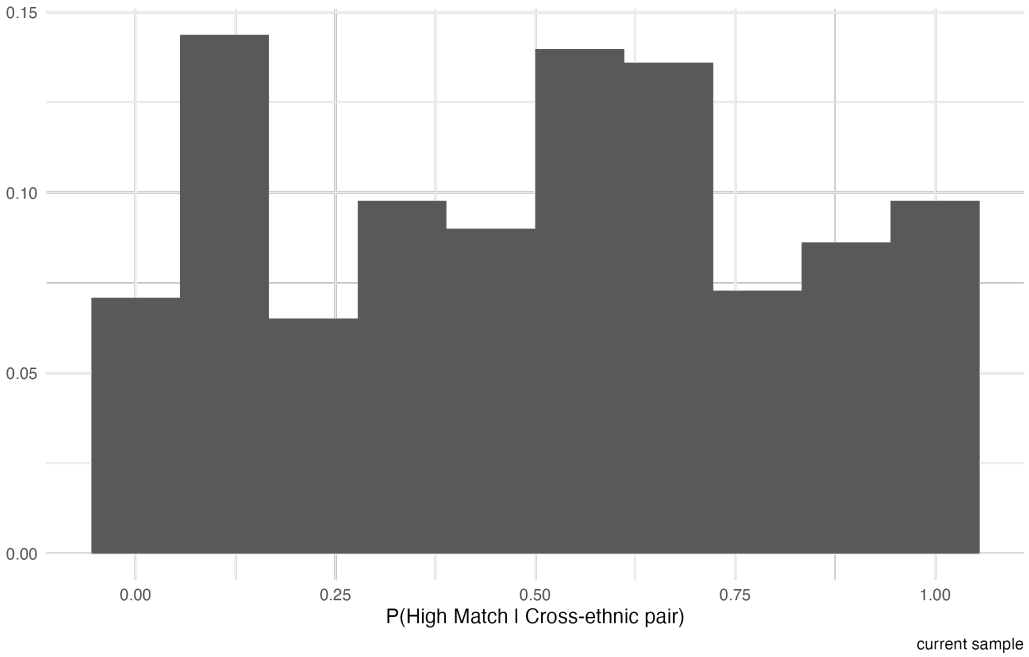
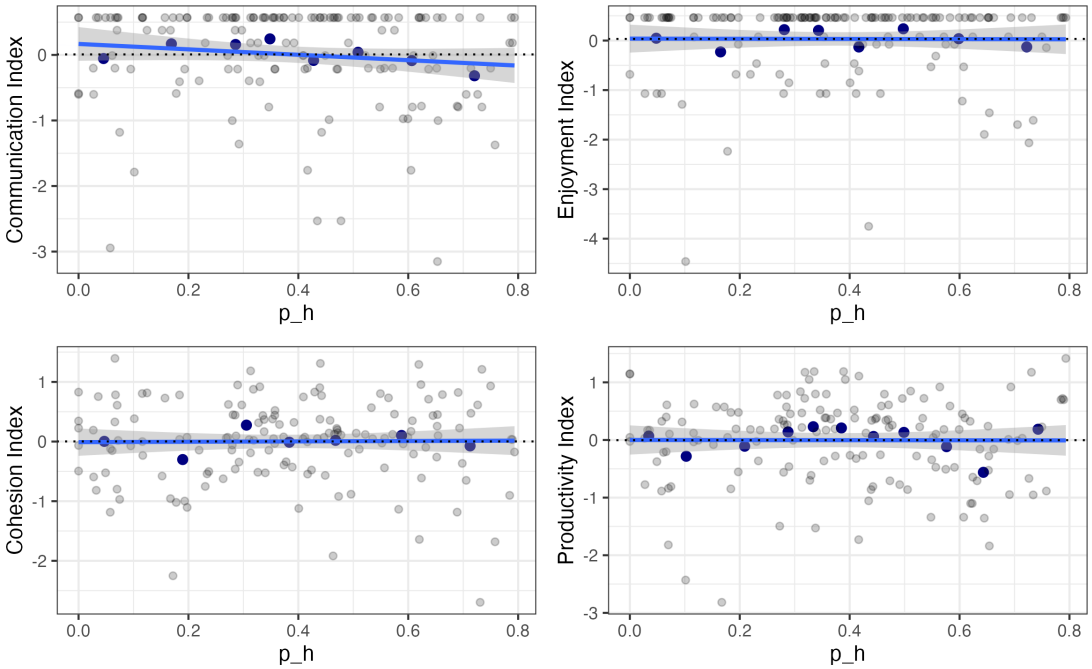


Figure A3: The probability distribution of a high language match conditional on being assigned a cross-ethnic partner. In the notation of exposure mappings this plot shows $\frac{\pi(1,1)}{\pi(1,1)+\pi(1,0)}$.

Figure A4: $Y(0)$ against $\pi(1,1)$



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

B Procedures

B.1 Randomization

Randomization into pairs that depend on baseline characteristics adds complexity. We face the following goals and constraints. First, we stratify by school, grade level, and gender; keeping gender and grade constant within pairs helps isolate identity and language's role. Moreover, we hope to split treatment-control groups at certain proportions to maintain power.[65] 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.[66] 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*

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

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

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

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

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

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

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

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

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

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

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

1. $\tilde{p} = p$

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

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

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

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

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

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

Assigning cross-ethnic pairs

After making control assignments, we assign the remaining students to treatment. To do so we arrange the list first by the size of the ethnic group in the strata and then by a random number. If we have an even number of students and the triangle inequality holds ($|E1| \leq |E2|+|E3|$), we split the list in half and match partners by their location in the list.[67] 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.

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

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

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

B.2 Data collection

Participation consisted of several activities. First, enumerators from our survey firm conducted baseline surveys, collecting data about language skills, social networks, behavioral measures, and questions about socio-economic status and well-being. For the most part, once both partners in a pair completed the baseline survey, we sent the session information to the facilitation team who organized the video game sessions described in this paper. In addition to learning and playing "Overcooked! 2", students answered questions about their experience in the session, completed the choice experiment, provided additional behavioral measures, and decided which pin to take as a reward either one reflecting the common identity (town or school insignia) or one representing the students' ethnic identities reflecting their mother tongue alphabet.[68] Finally, students participate in a midline survey where we ask students to sign up for the interethnic course and re-ask social network and behavioral measure questions from the baseline. These three steps constitute the major

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

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

C Appendix: Bayesian Adaptive Choice Experiment: Theory and Implementation

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

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

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

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

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

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

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

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

C.1 Implementation

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

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

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

Students chooses between A and B, choosing A if

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

and B otherwise. Equivalently, they choose A if

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

Likelihood: This corresponds to the logit likelihood:

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

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

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

Choice scenarios/Designs: The set of choices includes 1275 distinct designs, or vectors $(\Delta r, \Delta z)$. The reward Δr takes one of 25 values between -100 to 500 Denar (roughly -\$2 to \$10) with the options taking values divisible by 25 Denar (about 50 cents USD). There are dependencies between ethnic and linguistic factors in the design which lead to 51 distinct designs. Δz takes a value in $-1, 0, 1$ reflecting differences in indicator variables. An exemplary $\Delta z = (-1, 0, 1, 0, 1)$ with the z_1, z_2 and z_3 corresponding to ethnicities (Albanian, Macedonian and Turkish), z_4 corresponding to English ability and z_5 corresponding to Macedonian ability. Then, this example corresponds to option A being Turkish and a second-language Macedonian speaker. Option B is Albanian and we randomly decide if they both speak English or not.[69]

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

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

C.2 Preference estimation

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

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