Social Learning in Segregated Networks

Ari Anisfeld
University of Chicago

Pepi Pandiloski
University of Chicago
Job Market Paper *

January 28, 2025

CLICK HERE FOR THE MOST RECENT VERSION
PRELIMINARY AND INCOMPLETE - PLEASE DO NOT CIRCULATE!

Abstract

How does segregation influence social learning? On one hand, if one group has information that is beneficial to another, segregation may harm learning by preventing information flow. On the other hand, if individuals learn better from in-group members, segregation may increase learning. Using a model of network formation and social learning, we formalize these tensions into two hypotheses. To test them, we design a social learning game for groups of 6 players and implement it in ethnically segregated schools in Macedonia. In the computer game, 3 students of 2 distinct ethnicities need to guess an unknown binary state of the world. The game manipulates the parameters of the model: private initial signals, the network on which players observe each other, and bias, by randomizing the number of high-bias players in the 6-person group. Experimental results provide evidence for the theoretical hypotheses. First, we find evidence that bias decreases the cost of segregation. In low-bias groups, students learn more in the integrated networks than in the segregated networks, avoiding echo chambers, but the same is not true in the high-bias groups. Second, integration leads to more learning than segregation in games in which one ethnicity gets mostly incorrect signals and the other ethnicity that gets mostly correct signals. To go beyond the lab, we fit our model to the experimental data and find that students are 45% more biased against different-ethnicity players than same-ethnicity players. We combine these bias estimates with data on real-life social networks, and characterize conditions under which a counterfactual policy of one integrated classroom per cohort would increase social learning in the high schools.

*Corresponding author: predragpandiloski@gmail.com. Pepi would like to extend an immense gratitude to Leonardo Bursztyn, Oeindrila Dube, Steven Durlauf, and James Robinson for their patient mentorship and generous help in developing the idea for this paper. We also thank Anjali Adukia, Ozan Candogan, Arun Chandrasekhar, Josh Dean, Claire Fan, Rachel Glennester, Donovan Harvey, Eduardo Montero, Jose Miguel Pascual, Violeta Petroska-Beshka, Raul Sanchez de la Sierra, Daniel Sonnenstuhl, as well as seminar participants at the Harris School of Public Policy PhD Workshop, the University of Chicago Development Economics Workshop and to the participants at the many brown bag lunches who provided feedback and moral support over the years. This research is funded by the JPAL European Social Inclusion Initiative, The Pearson Institute, the Development Economics Center and the Becker Friedman Institute at the University of Chicago. We'd like to thank Luan Imeri, Safet Ballazhi, and Zana Jusufi-Osmani at the Center for Human Rights and Conflict Resolution for excellent project implementation in the field, Jasmina Mihailovska and Tea Todorova at M-Prospect for coordinating data collection, and to 20 enumerators and 21 session facilitators whose diligent work made this study possible. We would like to thank Michael Plunkett and Federico Dominguez Molina for software development assistance in building the lab game platform, and Alex Chen for incredible research assistance in the field.

1 Introduction

Segregation within a population – whether driven by explicitly segregationist policy or individuals' preferences – is ubiquitous in societies around the world. How does segregation influence social learning? On one hand, if one group has information that is beneficial to another, segregation may harm social learning by restricting access to useful information. Indeed, conventional wisdom from the theoretical literature suggests that segregation can slow information spread or create echo chambers[1]. On the other hand, segregation typically delineates groups with different identities – be it class, race, ethnicity, or even gender – and identity may play a role in how individuals learn from each other[2]. If individuals learn better from similar people, segregation may have the opposite effect and improve social learning.

This question can have important policy implications. In Macedonia, an armed ethnic conflict in 2001 de-escalated with a peace agreement that extended rights to minorities, including native language education at all levels. Now, students of different ethnicities go to the same schools, but are segregated into different classrooms native language. Native language instruction might be desirable: it gives cultural recognition to minority groups and may increase human capital through better access to education. However, such parallel institutions have resulted in chronic segregation in the high schools we study: about 55% do not know any student of a different ethnicity, and about 96% of all reported friendships are within the same ethnicity. In part to address these issues, policymakers have talked about opening an integrated high school track in English[3]. Social networks can be very powerful in shaping student outcomes, so when policymakers make decisions to segregate or integrate, how might this affect social learning?

How segregation affects social learning is a notoriously difficult question to study empirically. First, observed relationships do not just capture avenues of social influence, but also pre-existing similarities between individuals[4]. Then, it could be commonalities among the individuals, rather than social influence, that cause

[1] A few seminal studies that seem to agree on this point are Acemoglu et al., 2011, Golub and Jackson, 2012, Chandrasekhar et al., 2020

[2] Distinction as a social force has long been recorded and studied across the social sciences. Anthropologist Gregory Bateson documents schismogenesis – the proclivity of groups to act in ways that are opposite of different groups – as early as 1936 (Bateson, 1958). In sociology, Bourdieu's *Distinction* has become a classic about this phenomenon, in which he documents how tastes diverge across groups divided by class Bourdieu, 1984). In recent work in economics, Conlon et al., n.d. find that wives adopt information from husbands, but not vice versa.

[3] In conversation with the Mayor of the town our study is based in, he said that the Municipality is exploring the policy with the Ministry of Education.

[4] For example, neighborhoods or classrooms may present opportunities to meet and form social ties. Individuals also tend to

similarities in observed behavior, an issue known as the correlated effects problem[5]. Second, groups may have asymmetric preferences, such that the "true state of the world" is different for different groups[6], and researchers may not know to what degree that is the case.

To address these two core difficulties, we combine a theoretical model with a lab-in-the-field experiment to study this question. This offers three core advantages. First, the model conceptualizes segregation and social learning, and generates testable hypotheses for the experiment. In the network formation stage of the model, *segregation* grows as same-ethnicity links become more common than different-ethnicity links, holding fixed the average number of links[7]. In the social learning stage, agents get noisy random signals, and then learn from their network links by taking a weighted average of all observed network actions. The groups have *correlated priors* if one group has a more correct initial signal than the other in expectation[8]. This model generates two hypotheses: 1) bias *decreases* the costs of segregation on learning, and 2) the correlation of priors *increases* the costs of segregation on learning.

The second advantage of our approach is that the model informs the design of the experiment. We designed a computer game that manipulates parameters of the model to test the theoretical hypotheses while also dealing with the two core empirical issues outlined above. The game *"Red or Blue"* is played by a group of 6 students – 3 students from 2 distinct ethnicities. In this game there are two bags, a red bag (with 4 red balls and 2 blue balls) and a blue bag (with 4 blue balls and 2 red balls). The computer picks one bag at random, and the goal of the game is to guess whether the chosen bag is red or blue. Once the bag is chosen, each player randomly gets one of the 6 balls in the bag as a "private signal". Thus, 4 students get a correct signal and 2 students get an incorrect signal about the true color. In the first round, the players see their ball and make a guess. In subsequent rounds, the players observe each other on a social network. Each player can see their own previous-round guess and the previous-round guesses of their network neighbors, after which they guess again.[9] If students pick the correct color in the last round of the game they get a cash reward, otherwise they

form relationships with those who are similar to them.

[5]See Bramoullé et al., 2020 on the correlated effects problem in the identification of peer effects, and common solutions to it.

[6]For example, it would be a mistake to interpret the observation that Albanians don't eat pork as a failure of social learning from Macedonians; rather, Albanians and Macedonians have different religious prescriptions that imply different utility from eating pork.

[7]This holds constant the overall density of the network, which could weakly increase learning under the model, and instead focuses on how the identity of the individuals across the links matters. Intuitively, segregation arises from "rewiring links" rather than adding or removing links in the network.

[8]This creates an environment in which one ethnicity can stand to learn from the other. Agents do not know ex ante the precision of the initial signal.

[9]The computer shows the players' guesses as they were actually realized in the previous round, and players do not communicate their guess to other players, precluding lying.

get nothing. The game repeats for 3 to 5 rounds, stopping randomly to avoid end-of-experiment effects; each sextet plays 12 games.

We find that learning increases by 9% on average across games. This means that the number of players who correctly guess the color of the bag increases from 4 at the onset of the game (those who get the correct signal) to about 4.54 players who correctly guess at the end of the game. However, there is a large heterogeneity across games. The modal outcome is perfect learning – in about 45.3% of the games all six players correctly guess the color. Yet, in almost 28% of all games we observe "mislearning", such that 3 or fewer players correctly guess the color – an outcome worse if everyone just followed their own signal without observing others. These disparate outcomes motivate our treatment arms in the experiment, which manipulate parameters of the model in order to empirically test the theoretical hypotheses.

First, we induce random segregation by varying the observation network during each game. Figure 2 represents the *Segregated* and *Integrated* treatments. Segregation is varied by fixing the amount of information available to players, but changing the ethnic identity of the individuals that are observed. Second, we vary the correlation of signals. Since four out of six players get a correct signal, how these are distributed across ethnicities can help us vary the correlation of priors. The *Uncorrelated priors* arm consists of games in which both ethnicities have the same prior information: 2 players get a correct signal and 1 player gets an incorrect signal. The *Correlated priors* arm consists of games in which one ethnic group needs to learn from the other: one ethnicity gets 2 incorrect and 1 correct signal, while the second receives 3 correct signals. Third, we manipulate bias at group-level by randomizing the types of players that compose the group. Specifically, in our baseline survey, only about half of the participants report a cross-ethnic friendship, which we use as a proxy for low bias. By randomizing group members we ensure that some groups have less biased players than other groups by chance. A median split on this group-level variable helps us manipulate bias: *High bias* groups are those in which 3 or fewer students have a friend from a different ethnicity, while the *Low bias* groups are those in which 4 or more students have a friend from a different ethnicity. This design addresses the empirical problems associated with our question. By randomizing the network structure and the initial signals, we deal with the correlated effects problem since we eliminate the possibility that any common characteristics drive initial actions and preclude endogenous selection of peers. Additionally, students across ethnicities show similar preferences for red and blue, alleviating concerns about asymmetric preferences.

To test the two hypotheses, we compare learning across different treatment conditions. First, we find support for the hypothesis that bias decreases the cost of segregation. We find that in the low-bias groups, players learn the true color more often overall. In low-bias groups, integrated networks lead to more learning than the segregated networks, but the same is not true in high-bias groups. Most importantly, we can reject the null hypothesis that bias increases the cost of segregation at the 10% significance level, leading us to conclude that our experimental data is consistent with the first hypothesis. This result suggests that the benefits of integration critically depend on the *types* of individuals in the network. A limitation of this manipulation of bias is that it may be confounded. For example, it may be that students with cross-ethnic friends are more trusting generally, and that this trust interacts with the treatment instead. In exploratory analyses, we show that our result is unlikely to be driven by general trustworthiness or cognitive skills. Second, we also find evidence for the second hypothesis that correlated priors increase the cost of segregation. In games with correlated priors, integration leads to more learning than segregation, while the same is not the case for games with uncorrelated priors. Again, we are able to reject the null hypothesis that correlated priors decrease the cost of segregation at the 10% significance level, leading us to conclude that the experimental data is also consistent with the second hypothesis. These results imply nuanced effects of segregation: it can be harmful if ethnic groups need to learn from one another, but beneficial if groups are sufficiently biased against one another.

The third advantage of connecting a model and a lab experiment, is that we can go beyond the lab game and ask how learning changes if policymakers integrate classrooms. To do this we combine the bias estimates from the experiment with data on real-life social networks in the high schools. Inspired by conversations with local policymakers, we consider a counterfactual policy in which schools integrate one classroom per cohort. Using a model of network formation, in which linking decisions depend on ethnicity and classroom, we generate counterfactual social networks under this policy, and compare learning between the status-quo and the counterfactual networks. In the simulations on the counterfactual networks we use two bias distributions. The first one is the estimated parameter from the lab game, assuming integration does not interact with bias. However, students may decrease their bias once they are integrated, as a large literature on inter-group contact suggests (Pettigrew, 1998, Mousa, 2020, Lowe, 2021). Indeed, in a separate study in the same high schools, we find that cooperative integration reduces bias by 0.19 standard deviations (Anisfeld and Pandiloski, 2024). Therefore, we also consider learning in the counterfactual networks under the assumption that integration

decreases bias by 0.19 standard deviations, as suggested by Anisfeld and Pandiloski, 2024.

Our bias estimation results suggest that individuals discount others' actions regardless of ethnicity. We estimate same-ethnicity bias of about 0.6 on average, meaning that players weigh actions of same-ethnicity players about 0.4 as much as they value their own individual action[10]. However, participants have a bias of 0.87 toward members of different ethnicities, meaning that they value their action 0.13 as much as they value their own action – 45% increase from the same-ethnicity bias. In the counterfactual simulations, we find that the benefits of integrating one classroom per cohort depend on the distribution of initial actions. If ethnic groups have the uncorrelated priors, forming an integrated classroom decreases social learning, since students are more biased against different-ethnicity individuals. This gap is only marginally repaired under the assumption that integration lowers bias in the integrated classrooms. However, if the groups interact with correlated priors, integration can increase learning over the status-quo networks. This is because the integrated classrooms creates opportunities for cross-ethnic ties, that allows useful information to travel across the ethnic clusters in the network. Importantly however, this benefit of integration happens only if sufficiently small fraction of individuals in the population take the correct initial action.

Our study contributes to at least two broad strands of literature. A large empirical literature studies social learning across domains[11]. While the empirical literature has considered how identity might shape social learning (e.g. Munshi, 2004, Miller and Mobarak, n.d., Bayer et al., 2008, Conlon et al., n.d.), the question of how network structure interacts with identity has eluded empirical treatment. It may be important to study social learning in a network of individuals, since networks allow to capture spillovers and group-level dynamics, while paired learning designs may fail to account for the emergent properties of segregation and networked learning. To the best of our knowledge, we are the first to empirically study how identity and segregation interact in social learning on networks.

There is also a large theoretical literature in economics that studies social learning in networks (see Golub and Sadler, n.d. for a review), and some papers have even studied how segregation might affect learning

[10]Interestingly, this estimate is similar to other estimates in the literature: both Mobius et al., 2015 and Conlon et al., 2022 find similar weights of 0.4, even though they have very different empirical strategies and populations of interest.

[11]Examples include the spread of agricultural technology (Conley and Udry, 2010, Foster and Rosenzweig, 1995), take up of insurance (Cai et al., 2015) and microfinance (Banerjee et al., 2013), financial decisions (Bursztyn et al., 2014), spread of health products (Kremer and Miguel, 2007, Dupas, 2014, Oster and Thornton, 2012), retirement savings (Duflo and Saez, 2002, Duflo and Saez, 2003), job referrals (Granovetter, 1983, L. A. Beaman, 2012, L. Beaman and Magruder, 2012, Goel and Lang, 2019, Bayer et al., 2008, Bayer et al., 2008, Damm, 2009), voting behavior (Lazarsfeld et al., 1944, Zelizer, 2019), and crime (Glaeser et al., 1996)

(e.g. Golub and Jackson, 2012, Chandrasekhar et al., 2020, Acemoglu et al., 2014). This literature has mainly focused on network structure, but bias in updating has received less attention. In real life however, segregation delineates different identities, and individuals may be biased against other groups. In Section 2, we extend the model of Golub and Jackson, 2012 to allow for agents to discount players with different identity. In this extended model, we show how both segregation (extensive margin network links) as well as bias (intensive margin network links) shape social learning. Our study also represents a step to bridge the empirical and theoretical studies of social learning, a gap that has been flagged by previous work (Mobius and Rosenblat, 2014). In our study, there is a correspondence between the model and the lab game, which allows us not only to test model-implied hypotheses, but also to estimate relevant parameters in the model and conduct counterfactual simulations on real-life networks.

The second literature that we contribute to is on identity and political economy. Both conformism[12] and in-group bias[13] have been documented in many domains. Little is knows about conformism and in-group bias in the process of social learning however. We add to this literature by quantifying the extent to which individuals conform to in-group and out-group members, and by documenting in-group bias in the process of social learning. Finally, a large literature has focused on inter-group contact (see Pettigrew, 1998 for a review), and recent experimental work integrates groups in cooperative tasks to improve cohesion (e.g. Corno et al., 2022, Lowe, 2021, Mousa, 2020). Little is known about how such interventions might shape social learning. Our second contribution to the political economy literature is to characterize environments in which contact between groups may lead to higher social learning than the status quo. We show that integration may be more optimal for social learning if the individuals have low learning bias or if one group has better information then the other.

The paper proceeds as follows. Section 2 introduces the model and generates theoretical hypotheses. Section 3 details the local context in Macedonia, while Section 4 the study design and data. Section 5 tests the hypotheses, Section 6 structurally estimates the bias in the experiment, and Section 7 presents the counterfactual policy simulations. Section 8 concludes and outlines next steps.

[12]Conformism has been studied in assessment and attitudes (e.g. Mackie et al., 1990, Morgan et al., 2012), mimicry (Gump and Kulik, 1997, Schalk et al., 2011), preferences (e.g. Benjamin et al., 2010), and education (e.g. Bursztyn and Jensen, 2015, Bursztyn et al., 2019), among other domains

[13]In-group bias has been studied in economic allocations (e.g. Tajfel et al., 1971, Leider et al., 2009, Chen et al., 2010), trust and cooperation (e.g. Falk and Zehnder, 2013, Charness et al., 2007), education (e.g. Fisman et al., 2018, Kramon and Posner, 2016 productivity (e.g. Bandiera et al., n.d., Marx et al., n.d.), and public policy (Bar and Zussman, 2020, Burgess et al., 2015).

2 Model and hypotheses

Network formation

The network formation unfolds first in the model. Consider a population of agents $N = \{1, 2, ..., n\}$, such that each individual i has one of two types: $k_i \in \{1, 2\}$. Each type is represented equally in the population[14]. The agents are connected in a social network described by an $n \times n$ symmetric adjacency matrix $\mathbf{G} \in \{0, 1, 1-b\}^{n \times n}$. If agents i and j are not connected in the network, then $G_{ij} = 0$; if they are connected and belong to the same group ($k_i = k_j$) then $G_{ij} = 1$; if they are connected and belong to different groups $k_i \neq k_j$ then $G_{ij} = 1 - b$, where $b \in [0, 1)$.[15] The parameter b here captures bias, and allows the tie across different groups to be "weaker" – the larger the bias the weaker the tie. The strength of the link will come into play in the social learning stage later.

Each possible pair ij meets once[16], and at this meeting, they form a link with probability:

$$\mathbb{P}(ij \text{ form a link }) = \begin{cases} p_s & \text{if } k_i = k_j \\ p_d & \text{if } k_i \neq k_j \end{cases} \tag{1}$$

We treat $\mathbf{G}(n, b, p_s, p_d)$ as a random object and $\mathbf{G}$ as the realization of the random process. Since we want to explore consequences of segregation, we focus on networks in which links within each group are at least as likely as links across groups: $p_s \geq p_d > 0$. Conceptually we think of **segregation** as being captured by ratio p_s/p_d, holding fixed the average number of links in expectation. If $p_s/p_d = 1$ then there is no segregation, and links across groups are as common as links within groups; the larger p_s/p_d is, the more segregated the networks are. Note that this definition corresponds to a "rewiring" of the network.[17]

The network formation model is often referred to as the stochastic block model and belongs to one of the

[14] We take n to be even, and focus on the 2 group case to most simply convey the intuitions. Extensions to m groups of arbitrary sizes n_m are straightforward.

[15] If b=1 then the networks are behaving as completely separate islands and is an uninteresting edge case that we exclude here.

[16] For simplicity in the analyses, we treat self-links ii as any other link. The results don't change if we impose $G_{ii} = 0$ or $G_{ii} = 1$ in either the simulations or the results on asymptotic learning.

[17] We do not consider a model of segregation, in which segregation goes down if we add more cross-group links. In this case, the model predicts more learning because there is more access to information, which is a trivial and uninteresting case. This definition also conflates measurements of segregation with overall density in the network.

most widespread family of models in the literature.[18] The advantage of this modelling approach is that it is simple and tractable, while capturing differential linking across groups that is essential to our analysis[19].

Social learning

Once the network is formed, at $t = 0$, each agent i receives a private binary signal $\omega_i \in \{-1, 1\}$ such that:

$$\mathbb{P}(\omega_i \text{ is correct}) \quad = \quad \begin{cases} s_1 & \text{if } k_i = 1 \\ s_2 & \text{if } k_i = 2 \end{cases} \tag{2}$$

with $s_1, s_2 \in [0, 1]$. While the precision of the signal can be different by group, most individuals in the network receive the correct signal, i.e. $\frac{s_1+s_2}{2} > 0.5$. The different precision of the signals allows us to conceptually distinguish when priors are correlated. Intuitively, if group 1 has a stronger signal than group 2, i.e. $s_1 > s_2$, then individuals from group 2 can stand to learn from individuals from group 1, since they have better information on average. We call such priors **correlated**: holding fixed average precision ($\frac{s_1+s_2}{2}$), the larger $|s_1 - s_2|$, the stronger the correlation.

At the end of period $t = 0$, each agent's action is simply their private signal, resulting in an initial vector of actions $\mathbf{a}(0) = \omega$. In each period thereafter, agents play a myopic coordination game with their network neighbors. In this game, agents get utilities from coordinating with the actions with other people around them, because that is how they learn about the true state of the world, but they also have preferences for *distinction*, that is, disutility from taking the same action with an individual from a different group. Formally, agents' utilities are:

$$u_i = -\left(\frac{1}{d_i} \sum_{j:k_i=k_j} |a_i - a_j| + \frac{1-b}{d_i} \sum_{j:k_i \neq k_j} |a_i - a_j| \right) \tag{3}$$

[18] For example, the seminal model by Erdős and Rényi, 1959 is a special case of our model, in which $p_s = p_d$ and all agents are of the same type; the latent space model pioneered by Hoff et al., 2002, is a generalization, in which each agents has their own type, captured by a coordinate in a multi-dimensional space, and linking probabilities to all other agents decrease with distance between pairs of agents. These models are known as conditionally independent dyad (CID) models, since each pair links independently of any other pairs, and have found myriad applications in economics and beyond. Chandrasekhar, 2016 is a great review of these models and their use in economics and beyond.

[19] A limitation of conditionally-independent dyad models is that they tend to fall short in generating the rate of clustering observed in real social networks. Real-world networks exhibit a lot of "triangles": if agent i is linked to agent j, and agent j is linked to agent k, then a link between i and k is more likely. Since links are formed independently, they cannot match the clustering that characterizes real-world networks. While clustering may have important implications on learning patterns, delving deeper into these caveats is beyond the scope of this paper.

where $d_i = \sum_j G_{ij}$. One way to view the bias parameter b, is as as a preference for *distinction*, a disutility from taking the same action as someone from a different group. Distinction as a social force has long been recorded and studied across the social sciences. Anthropologist Gregory Bateson documents schismogenesis – the proclivity of groups to act in ways that are opposite of different groups – as early as 1936 (Bateson, 1958). In sociology, Bourdieu's *Distinction* has become a classic about this phenomenon, in which he documents how tastes diverge across groups divided by class (Bourdieu, 1984). Conceptually, we think of **bias** as captured by the parameter b, and measuring the discounting of the actions of individuals from different groups.[20]

Before characterizing the best-reply dynamics to this game, define an $n \times n$ updating matrix $\mathbf{T}(\mathbf{G})$ with characteristic entry

$$T_{ij} = \frac{G_{ij}}{\sum_i G_{ij}} \tag{4}$$

i.e. $\mathbf{T}$ is a row-normalized version of the adjacency matrix $\mathbf{G}$ that represents the network. Then, the social learning rule for $t \geq 1$ that arises as a best-reply to the coordination game with utilities in (3) is:

$$\mathbf{a}(t) = sgn\left[\mathbf{T}(\mathbf{G})\mathbf{a}(t-1)\right] \tag{5}$$

where *sgn* is the signum function that is applied element-wise; this returns 1 if the element is positive, -1 if the element is negative, and the previous-period action if the element is zero. That is, for a typical agent i this updating rule implies a weighted average of previous-period actions $a_i(t) = sgn\left[\sum_j T_{ij}(\mathbf{G})a_j(t-1)\right]$. Importantly, the weighted average is proportional to the weights of the links in the network $\mathbf{G}$; this means that the larger the bias parameter b, the more agents discount the actions of individuals from different groups.

This social learning model is often called the DeGroot model and has a long history – see French Jr., 1956, Harary, 1959, DeGroot, 1974 for the earliest formulations of the rule, and Golub and Sadler, 2016 and Jackson, 2008 for great reviews of this model and its relation to economics. We favor this model for two main reasons. First, there is growing empirical evidence that Bayesian learning is not a good predictor of real life behavior, and that the DeGroot model may do a better job (Massey and Wu, 2005, Weizsäcker, 2010,

[20] In our model, we move b and segregation p_s, p_d independently, so the results we present here may not extend to cases when bias and segregation are dependent, for example, if integration lowers bias.

Chandrasekhar et al., 2020).[21] Second, this model has a natural interpretation of the parameters, and we can use it to estimate the bias b. The model is also computationally tractable in large real-life networks, and facilitates simulations and counterfactual excesses that are of interest to our study.

Hypotheses

To formalize testable hypotheses, we are interested in how the effects of segregation p_s/p_d on social learning might be shaped by 1) bias b and 2) the correlation of priors $|s_1 - s_2|$. The simulation algorithm is detailed in the Algorithm 1 box below. The inputs for the algorithm are population size n, probability of same-type and different-type links p_s, p_d, precision of priors for each group s_1, s_2, bias b, number of updating rounds L, and number of iterations for each parameter set I. The simulations process mirrors the model: first, we generate random networks, then, we seed random priors, and finally play out learning using the DeGroot updating process. The outcome of interest is the fraction of agents that learn the true state of the world.

Algorithm 1 Theoretical Simulations of Social Learning

Inputs: $\{n, p_s, p_d, b, s_1, s_2, L, I\}$
for $i \in \{1, 2, ..., I\}$ **do**
    Generate a random network $\mathbf{G}(n, b, p_s, p_d)$ as described in (1)
    Generate random vector of initial beliefs $\omega(s_1, s_2)$ as described in (2)
    Generate an updating matrix $\mathbf{T}(\mathbf{G}(n, b, p_s, p_d))$ as described in (3)
    Assign $\mathbf{a}(0) = \omega$, and calculate learning vector at period L recursively as $\mathbf{a}(L) = sgn[\mathbf{T}\mathbf{a}(L-1)]$
    Calculate network-level learning for iteration i as $a_i = \frac{\mathbf{1}'\mathbf{a}(L)}{n}$
    Store (i, a_i)
end for
Calculate average learning over iterations $a = \frac{\sum_{i \leq I} a_i}{I}$
Store the mapping $(n, p_s, p_d, b, s_1, s_2, L) \to a$

We set $I = 1000$, $L = 20$. In general, these parameters should be as large as possible for the law of large numbers to kick in (in the case of I) or learning to stabilize (in the case of L). We find that these choices yield stable estimates.[22]

[21] Massey and Wu, 2005 show that players sometimes under- or over-value the signals of other in isolated ball-and-urn lab game, something that is not consistent with rational Bayesian learning. Weizsäcker, 2010 conducts a meta-analysis of 13 studies on social learning, concluding that agents systematically contradict Bayesian predictions. Chandrasekhar et al., 2020 conduct an experiment designed to adjudicate between Bayesian and DeGroot learning types, and find stronger evidence for DeGroot-type learning. In a future iteration of the paper, we will attempt to adapt the insights from Chandrasekhar et al., 2020 to directly test in our context whether our participants behave more like DeGroot or Bayesian learners.

[22] Though note that in the case of discrete learning, there may not be convergence in beliefs. Consider if there is an isolated link ij who only copy each other, and assume that $a_i(0) = -1, a_j(0) = 1$. Then under this model i and j will keep alternating their guesses. In practice, we see that the fraction of agents who correctly guess the state of the world converges "globally", though sometimes there are "local" oscillations depending on the specific network structure.

Since $b \in [0,1]$, we explore a grid of possible bias values $b = \{0, 0.25, 0.5, 0.75, 1\}$. We vary segregation through p_s, p_d such that we fix degree (expected number of links per node), while varying the ratio $\frac{p_s}{p_d}$. Specifically, for a given degree d, we consider the grid: $(p_s, p_d) \in \left\{ \left(\frac{u}{(u+2)n} d, \frac{1}{(u+2)n} d \right) : u \in \mathbb{N}, 1 \leq u \leq 10 \right\}$. For example, when $u = 1$ then $p_s/p_d = 1$ and there is no segregation since probability of linking is the same, as u increases, the networks get more segregated.

We set average precision of the initial signal as $\frac{2}{3}$. In general, the larger this is the more learning overall, and vice versa. For large enough or small enough precision, we create a corner case in which there is perfect or no learning. In these cases, segregation, bias, and correlated priors do not matter and the patterns "collapse". We chose an overall precision of $2/3$ in relation to expected degree (to avoid the corner cases), and in a way that best illustrates the relationship between segregation, bias, and priors. Given the average precision, we consider a grid of signalling schedules: $(s_1, s_2) \in \left\{ (\frac{2}{3}, \frac{2}{3}), (\frac{2}{3} - 0.1, \frac{2}{3} + 0.1), (\frac{2}{3} - 0.2, \frac{2}{3} + 0.2), (\frac{1}{3}, 1) \right\}$. In our main simulation results we set $n = 100, d = 5$. In Appendix A.2 we explore how the main simulation results change as we vary n and d. The appendix simulations show that the main patterns remain unchanged.[23]

In Figure 1 we present the main results of the simulations. On the x-axis we plot *segregation*, captured by p_s/p_d, which ranges from 1 to 10. On the y-axis the fraction of individuals who learn the true state of the world (output a from Algorithm 1), averaged over iterations. In Panel A we explore how bias interacts with learning, such that each of the 5 lines plotted represents a different level of bias (darker colors mean larger bias). On the left, we plot the relationship between segregation and learning when each ethnicity has correlated priors; on the right, we plot the same relationship, but for uncorrelated priors. In Panel B the x- and y-axis remain unchanged, however here we explore how correlated interact with segregation. Each line represents a different schedule of signals, with darker lines denoting higher correlation. On the left side, we consider a case with high bias, and on the right a case with low bias.

First, we can notice that bias shifts learning down. In Panel B, we can notice that the relationship between segregation and learning depends on the correlation of priors. For low correlation (similar information across groups) we can notice that the relationship between segregation and learning is positive, even under small bias. As the correlation grows, and one group can stand to learn from the other, the relationship flips and becomes negative – segregation harms learning because it prevents useful information from flowing.

[23] Though for some parameters, the patterns exhibit "corner cases", like when agents have many links and always learn the true state. In these cases, the interactions between bias, segregation, and information disappear.

To fix ideas, consider the "benefits of integration" (or equivalently, the "cost of segregation") measured as $\Delta a = a(I) - a(S)$, for some $S > I$ on the x-axis; this captures how much more learning there is under integration vs. segregation. What happens to Δa as bias increases? Visually inspecting Figure 1 Panel A, we see first that larger bias shifts learning curves down. Furthermore, we can see that for any two biases $b_h > b_s$, the following holds: $\Delta a(b_h) \leq \Delta a(b_s)$: bias either decreases the benefits of integration (under correlated priors), or increases the cost of integration (under uncorrelated priors). This motivates our first hypothesis:

Hypothesis 1. *Higher bias decreases overall learning and decreases the benefits of integration on social learning.*

Social learning is maximized when agents incorporate more information from others around them. If agents are biased against those others, learning decreases. When an agent has a finite number of links that convey information, learning is maximized if the bias on those links is minimized.

Next, how do the benefits of integration Δa change as priors become more correlated? Visually inspecting Figure 1 Panel B, we can see that increasing the correlation of priors shifts learning down, but increases the benefits of integration, eventually making integration more optimal. This motivates the second hypothesis.

Hypothesis 2. *Higher correlation of priors decreases overall learning and increases the benefits of integration on learning.*

The intuition here is simple too. The more correlated the priors, the more one group is dependent on the other to learn. As this correlation increases, the benefits that accrue to integration because it facilitates information flow outweigh the cost of integration due to bias, eventually making the relationship between integration and learning positive.

In Appendix A.1, we present an extension of the model in Golub and Jackson, 2012, a seminal paper that studies how segregation harms learning. While our modelling strategy mirrors their, there are three key differences. They consider asymptotic learning, such that the society grows infinite, and use continuous updating with uncountable set of actions. These two features allow for tractable theoretical treatment of this question. Importantly, in their their model there is no bias ($b = 0$). The main result of Golub and Jackson shows that the larger the second eigenvalue of the adjacency matrix of the social network, the slower the

learning. In their model, the second eigenvalue captures segregation in the network, so the interpretation is that segregation slows learning. In the extended model where $b \geq 0$, we show that the second eigenvalue is a function of both segregation and bias, introducing some nuance in the interpretation of this seminal finding. Furthermore, we also show how the asymptotic learning case does not generate our hypotheses, as it cannot induce the trade-offs that arise in finite networks. For example, while our simulations show that segregation can *improve* learning under high enough bias, such conclusion is not possible in the asymptotic case, where agents' network links grow to infinity.

3 Background and context

Our setting is Macedonia, a small but diverse country in the heart of the Balkan peninsula. Macedonians, Albanians, Turks, and other ethnic groups have lived in this region together for hundreds of years. Nationally, Macedonians are the majority, with about 60% of the population, followed by Albanians who compose about 25%, and Turks which come at 4%; however, in the municipality in which we conducted our research, Albanians are the majority, followed by Macedonians and Turks. We conduct our game in 3 multi-ethnic high schools in one municipality.

Between the end of World War II and 1991, Macedonia was a composite republic of communist Yugoslavia. During this time, Albanians had limited access to political participation and little cultural rights. These policies of exclusion continued well into the 1990s, during the early years of independent Macedonia. Cycles of Albanian protests and Macedonian-majority government suppression of these protests escalated into an armed conflict between Albanian militia and Macedonian forces in 2001. This conflict in Macedonia was resolved with the Ohrid Agreement, which granted minorities more political and cultural rights, including education in their native languages at all levels.

In the inter-ethnic high schools that serve as the setting of our study, the effects of this agreement are apparent. Albanians, Macedonians, and Turks, go to the same schools, but are all segregated into different classrooms, which instruct in the three local languages. Students have fixed classrooms, such that different subject teachers rotate to a fixed group of students in a fixed room. Often, different ethnic groups are organized in different shifts, and even when they are integrated within the same shift, often have different wings in

the schools assigned to them. Therefore, different ethnicities have few opportunities to interact. Our baseline network survey shows the extent of segregation. About 55% of students do not know any peers of a different ethnicity, and only 3.9% of all network links are across individuals of different ethnicities (Table 2).

We have some evidence that students prefer real-life advice from individuals from the same ethnicity. In a survey we asked students to what extent they are likely to heed some advice to visit a new cafe, to buy a cheap and good laptop, or to a apply for a talent scholarship. In each of these three questions, we presented the advice-giver as a random name that signals a specific ethnic group. In all three questions, students express higher likelihood to learn from others when those others share their ethnicity.[24] Naturally, these answers are confounded by other factors, such as different beliefs, asymmetric preferences, or experimenter demand effects.

Our paper is motivated by this segregated environment to ask how this social structure, priors, and inter-ethnic bias can shape learning. The Mayor has talked about a prospective policy of creating integrated English-language classrooms in the high schools, and that they are negotiating this policy with the Ministry. If a classroom in each cohort is integrated are integrated, how might this policy change social learning through shifts in the social network? Our study is partly motivated by this question, and in Section 7 we conduct counterfactual exercises in which we examine how social learning might change under this policy.

4 Study design and data

Study design

We conducted the lab game in 3 inter-ethnic high schools in Macedonia, and participation in the study was open to all students during recruitment. To recruit students, local enumerators fluent in the students' native language went to their classrooms and presented the opportunity to participate in the study. The enumerators then helped students fill an assent form, after which their parent received an automatic email with a link to fill in parental consent online. To avoid experimenter demand effects and reduce self-selection into the study, we

[24] All three questions were asked on a 7-point Likert scale, from strongly disagree (1) to strongly agree (7). We observe preferences for same ethnicity advice-give in cafe visits (5.31 vs. 5.09), laptop buying (5.16 vs. 4.97), and scholarship application (5.31 vs. 5.11), though none of the differences are statistically significant.

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."*

The lab game is a variation of a ball-in-urn problem we developed for the study. In the lab game, a randomly composed group of 6 students – 3 students from 2 distinct ethnicities – play a learning game called *"Red or Blue"*. For example, in some sessions, groups were composed of 3 Macedonians and 3 Albanians, in others 3 Albanians and 3 Turks, etc. A summary statistics of participants is presented in Table 1. We built a custom platform for playing this game[25]. The platform allowed us to set-up sessions with a specific composition of players, distribute URL links to 6 participants, and simultaneously allowed them to play the game. The platform also facilitates the randomization as described below. We conducted 95 sessions with unique 6-student groups.

During each session, a pair of two facilitators, one from each ethnicity represented in the group, facilitated the process. The facilitators were teachers from the school who were hired and trained according to a protocol for conducting the sessions. Prior to each session, we loaded the list of randomly chosen students into our game platform. Session facilitators were only given a session ID, which included unique URL links for each student in that session. The facilitators then distributed these URL links to each participant accordingly. The sessions happened with the students in the same room playing simultaneously, and students could use either a computer or a smartphone, both of which are compatible with our platform.

In case of a participant no-show, the facilitators were provided with two randomly chosen reserve players, one from each ethnicity. If a reserve player joined the session, the field team coordinators created a session on the platform with the new player composition, and gave the facilitators the new session ID and URL links. In case the reserve players were also unavailable, the session was rescheduled. To ensure compliance on sessions composition, we asked participants before the game about their birth month and year, and compared that report with their birthday report at registration. We find that 96.03% of participants have a perfect match on both their birth month and year reported at registration and during the session, suggesting high compliance on the random session composition.

[25]The python-based platform is available on GitHub here: https://github.com/prijatelilab/PrijateliTree

At the beginning of each session, the participants watched an animated video in their native language that explains the rules of the game; the English language version of the video is available on this Youtube link[26]. During the video, participants are told that there are two bags, a red bag (with 4 red balls and 2 blue balls) and a blue bag (with 4 blue balls and 2 red balls). The computer picks one bag at random, and the goal of the game is to guess the correct color of the chosen bag. To help them guess, the computer randomly distributes one ball from the bag to each of the 6 students. Thus, 4 students get a correct signal and 2 students get an incorrect signal about the true state of the world. In the first round the students can see their ball and guess the color of the bag. In subsequent rounds, each player sees their own guess from the previous round, along with the guesses that two or three other players made in the previous round. Students observe the actual guess of their peers on the computer, shutting down the possibility of cheating. After observing this, each participant makes a guess in the current round. In each round, students who respond quickly will wait until everyone submits their guess to proceed in the next round, such that all guesses are aggregated and shown appropriately to the participants. To maximize comprehension, we also included a board-game-style visualized rule card in the session, and allowed students to play two practice games to get familiar with the process. We discard data from the practice games in all our analyses.

During the sessions, participants are only allowed to watch their screens and communication about the game is strictly forbidden. If students break this rule, the facilitator is responsible to disqualify them, though this did not happen in practice and teachers report little talking[27]. The game repeats for 3 to 5 rounds, stopping randomly to avoid end-of-experiment effects. The students are incentivized to guess the correct color: if they pick the correct color in the last round of the game they get a cash reward, otherwise they get nothing. Every 250 points are worth 50 MKD (0.9 USD). Points sum over the games, and students win rewards for all collected points. On average, each game win paid 27 MKD (or about 0.45 USD). To put this in context, with 3 wins out of the 12 games, a student has a little bit more money than is needed for a coffee in a typical coffee shop. Students earn on average 211 MKD. Each group plays 12 games during the session.

The lab game incorporates a randomization design that varies the three parameters in the model, and helps us test the hypotheses. First, we randomize the network structure on which the 6 students can observe each other. Figure 2 shows the *Segregated* and *Integrated* treatments. Students are represented by nodes and ob-

[26] https://www.youtube.com/watch?v=h4S1J68MifE

[27] 51% say there was no talking at all, and 93% of facilitators rank the amount of student talking to be 3 or less on a 1-6 scale. When asked whether the talking was explicitly about the game and broke the rules, and received only 2 reported instances.

servations by links (each link denotes bidirectional observation). Distinct ethnicities are denoted by different colors/shapes. Across the two networks, we hold fixed the number of links for each network position (degree distribution is fixed), while varying the identity of the observed player. Therefore, the amount of information available to players is in expectation the same, but the identity of the observed individuals changes. We administer this manipulation at the game level, and each arm is selected with 50% likelihood. To ensure comparability we also randomize each players network position across games.

It is important to note that under no bias, where each observed information is given equal weight, the model predicts that the integrated treatment should lead to more learning than the segregated treatment. Under the model, the average learning in the integrated network over all possible initial signals is 91%, while the average learning in the segregated network is 80%. This is because the segregated network is more prone to create echo chambers: if one ethnicity gets 2 incorrect and 1 correct signals in the network, the model predicts that each cluster averages to a different action and perpetually gets stuck there. This happens with smaller likelihood in the integrated network since information can flow more easily among players. Therefore, the experiment induces an overall environment in which integration is optimal if students have no bias.

Second, in each game we also randomize initial signals to the players. Since 4 out of 6 players get a correct signal, how these signals are distributed across ethnicities varies the correlation of priors. The *uncorrelated priors* arm are the games in which both ethnicities have the same information, i.e. each group recieves 2 correct and 1 incorrect signals. In the *correlated priors* arm, one ethnicity will receive 3 correct signals, while the other ethnicity will have 2 incorrect signals and 1 correct; in this random treatment the second ethnicity can stand to learn from the first. This treatment varied at the game level. Since we distribute signals randomly, there are 6 possible schedules of signals that belong in the correlated priors arm, and 9 possible schedules that are in the uncorrelated priors arm. Therefore, the first treatment is administered with 40% likelihood and the second with 60% likelihood.

Lastly, we induce random manipulation of bias by randomly composing the groups and inducing a variation in the "types" of players. Specifically, we administered a network survey in which students report friendships to other students in their school. Only about half of participants report a cross-ethnic friendship. We use this variation in pre-existing individual integration as a proxy for bias. Since groups are randomly formed, some groups have more players with cross-ethnic friends than other groups by chance. This measure of group-level

integration varies between 0 (if no players in the group have a cross-ethnic friend), to 6 (if all players in the group have a cross-ethnic friend). We split the groups by the median, resulting in about 60% groups that have 3 or fewer players with a cross-ethnic friend, and 40% of groups that have 4 or more players with a cross-ethnic friend. The former constitute the high bias groups while the latter the low bias groups[28].

This manipulation of bias comes with a limitation however. While the composition of players is random, whether one has cross-ethnic friends can be correlated with other characteristics. For example, if having more cross-ethnic friendships is also associated with higher general trust, it may be trust rather than lower bias that matters. In the results below, we unpack what other observables are associated with having a cross-ethnic friend, and rule out some competing explanations.

The lab game has two core advantages. First, it addresses the empirical problems associated with studying our question. By randomizing network structure and initial signals, we deal with the correlated effects problem, since we eliminate common characteristics driving initial actions as well as endogenous selection of peers that can confound estimates. Additionally, students in our sample show similar preferences for the colors, alleviating concerns about asymmetric preferences. The average percentages of pressing Red-Blue are 54%-46% for Albanians, 54.4%-45.6% for Macedonians, and 54.75%-45.25% for Turks. The second advantage of the game is that it mirrors our model. This allows for a clear mapping between model-implied hypotheses and empirical tests. Additionally, we can use the model to estimate the bias parameter from the observed data. This allows us to combine these parameter estimates with data on real-world networks of the high school students, and simulate learning in the real world.

Data

We use several administrative and original data sources in our analyses. First, we administered a baseline survey before the games.[29] Beyond a standard demographic module, the baseline survey included:

1. An inter-group preferences module including 1) 4 incentivized dictator games in which students play

[28] In a second manipulation of bias, we randomly revealed student names (signalling ethnicity) or anonymized shown guesses (e.g. "Peer 1"). Once we collect all the data, we will analyze these arms and include the results in future drafts.

[29] The baseline survey happened before the games for 99.6% of the participants with only one exception to the protocol. The baseline survey commenced in January 2024 and we have surveyed about 820 students, or about 93% of our sample.

against anonymous peers who speak their own mother tongue (in-group) and a different mother tongue (out-group), and 2) an Inclusion of Others in Self psychometric question (Aron et al., 1992) for each local ethnicity.

2. A language module, in which students list and rank, in order of fluency, Macedonian, Albanian, Turkish, and English; for English, and languages they say they speak, we also ask detailed questions about fluency and exposure.

3. A parent module, with questions on occupation, age, and highest level of education.

4. A module on attitudes including own ambitions for education, general trust, and life satisfaction.

5. A social networks module containing several types of questions. In the ones pertinent to this study, we ask students to list up to 6 peers in four types of ties: close friends (clique), peers they go to for school help, peers they go to to get information, and cross-ethnic friends. We aggregate all four questions to construct school-cohort level networks, and present network statistics in Table 2.

Our main data source is the online game platform which we built and designed with custom features to fit our needs. This platform exports complete data about all features and interactions during each game: an ID to link participants from sessions to other data, data about the game including all the random treatments (true color, signals, treatment conditions, network positions, etc) and each participant's action in each round. The sessions began in April 2024 and were finalized in December 2024.

We complement the game data with three short Qualtrics surveys that are embedded in the gaming platform. When the students open the platform, they start with pre-game survey. This survey includes the animated explainer about the game, simple tips about playing the game, asks participants to confirm their birthdays, and gives them a few questions to check comprehension. Then the students continue to play the game on the platform. After they complete 2 practice and 12 official rounds, the platform connects them to a second, post-game survey. In this survey students answer a few questions about their game strategies and beliefs during the game play, and gives them random questions about real-world advice-taking. We repeat a subset of the comprehension questions to check for game comprehension after the game. After the second survey, the students are free to go. The third survey comes from the sessions facilitators. After each session, the facilitators submit a survey in which they indicate whether students broke communication rules during the

session, the degree to which each student seemed to understand the game, and the degree to which each student seemed to enjoy the play.

5 Testing the hypotheses

How much learning is there?

We find that on average, games end with 4.54 of the 6 players guessing correctly. Since the game starts with 4 correct and 2 incorrect signals, this means that the fraction of agents who correctly guess goes from 66.66% to 75.66%, an increase of 9%. Moreover, there is considerable heterogeneity across games. Figure 3 plots the histogram of the number of agents who correctly guess in the last round. The modal outcome of the games is perfect learning: all six players guess correctly in 45.3% of the games! Yet, almost 28% of all games end in mislearning: fewer than 4 players guess the color correctly. This outcome is worse than if players guessed in isolation and followed only their private signal. This heterogeneity in outcomes motivates our analyses: when are agents more likely to end in a world of perfect learning, rather than in the world of mislearning?

Testing Hypothesis 1: Bias, Integration, and Learning

Recall that Hypothesis 1 stated that bias lowers learning and decreases the benefits of integration. Specifically, we want to test that:

$$L(HighBias) \leq L(LowBias) \tag{6}$$

$$L(HighBias, Int.) - L(HighBias, Seg.) \leq L(LowBias, Int.) - L(LowBias, Seg.) \tag{7}$$

where $L()$ represents average learning under the condition within the brackets. To test these two implications, we focus on network-level learning in the last round of each game, which varies from 0 to 1. As detailed in Section 4, we compare groups in which 3 or fewer students have at least one friend from a different ethnicity (*High bias*) with groups in which 4 or more students have at least one friend from a different ethnicity (*Low bias*). This represents a below-above median split with respect to our bias proxy, which we manipulate at the

session-level by randomly composing the group of 6 players.

Figure 4 plots the raw means associated with the inequalities (6) and (7). To test the implication in (6) we plot the means by bias in Panel A. Panel B gets at claim (7) by splitting the means by bias and segregation. Visually inspecting Figure 4 it seems that the means are consistent with the model predictions. To formally test the hypothesis, we run the following regressions:

$$Learning_{ig} = \beta_0 + \beta_H High_i + X_i + \sigma_i + \epsilon_{ig} \tag{8}$$

$$Learning_{ig} = \beta_0 + \beta_{HS} HighXSeg_{ig} + \beta_{LI} LowXInt_{ig} + \beta_{LS} LowXSeg_{ig} + X_i + \sigma_i + \epsilon_{ig} \tag{9}$$

where the subscript i denotes session and subscript g game; $High$ (Low) are dummies that take the value of 1 when the group in the session have below-median (above-median) cross-ethnic friends; and Seg (Int) are dummies that take the value of 1 when the game is played on a segregated (integrated) network. The regressions also add fixed effects X_i for school-by-grade randomization strata, and ethnic pairing (Macedonian-Albanian, Albanian-Turkish, and Turkish-Macedonian). We cluster standard errors at the session-level, which is the largest level of randomization in our design. Essentially, the regression in (8) maps onto Figure 4 Panel A, while the regression in (9) on Panel B, but they allow us to make more precise estimates as we can control for randomization strata and correct standard errors.

Table 3 shows the results of these regressions. We take a standard hypothesis testing approach to test the model implications. Specifically we test:

$$\begin{aligned} H_0 &: \ \theta > 0 \\ H_A &: \ \theta \leq 0 \end{aligned} \tag{10}$$

where $\theta = \beta_H$ in the regression in (8), while $\theta = \beta_{LS} - \beta_{LI} - \beta_{HS}$ in the regression in (9). Below each regression in Table 3, we present the estimated θ and the $p-$value associated with the one sided hypothesis test that the observed data is generated under the null. As predicted by the model, all the estimated θ have a negative sign. Moreover, we can reject the null that the observed data is generated under the null consistently at the 10% level. We conclude that the observed behavior in the experiment is consistent with the theory: higher bias reduces learning, and decreases the benefits of integration. This result also highlights that the

benefits of integration can critically depend on the characteristics of the individuals being integrated.

It is important to note that this result does not have a causal interpretation: we cannot conclude that making a cross-ethnic friend makes one learn better in integrated networks. Rather, our interpretation is that there are heterogeneous effects of integration by type. We think of each individual as having a *type* defined by their bias, as proxied by cross-ethnic friendships. We contend that whether integration or segregation are better depends on the bias-type of the student. Still, it may be that other characteristics interact with the integration treatment. For example, if students with cross-ethnic friends also have higher trust, it may be that we are capturing a trust interaction. In Table 5 we unpack what is in the type we capture. We find that having a cross-ethnic friend is related to lower in-group bias in a Dictator game, speaking local languages better, and perceiving other local ethnicities as closer on the Inclusions of Other in self metric. These are all characteristics that we expect from someone with prior interaction with other ethnicities, consistent with the type we are trying to delineate.

The table also suggests that we can reject at least two competing explanations for our results. First, individuals with cross-ethnic friends are not more trusting generally; on the contrary, they show lower levels of generalized trust, though this relationship is not significant. Therefore, the optimality of integration over segregation for the low bias types cannot be explained by trust. Second, we see that having a cross-ethnic friend is associated with having less educated parents, lower educational ambitions, and lower GPA. To the extent that these capture cognitive abilities, we can reject the alternative explanation that cognitive skills lead to better learning in integrated networks.

Testing Hypothesis 2: Correlated Priors, Integration, and Learning

We take an analogous approach to testing Hypothesis 2. Hypothesis 2 suggested that correlated priors lower learning, and increase the benefit of integration. Specifically, we want to test that:

$$L(CorrPri) \leq L(UncorrPri) \tag{11}$$

$$L(UncorrPri, Int.) - L(UncorrPri, Seg.) \leq L(CorrPri, Int.) - L(CorrPri, Seg.) \tag{12}$$

Figure 5 plots the means associated with the two parts of Hypothesis 2 stated above – splitting learning by whether priors are correlated or uncorrelated in Panel A, and then additionally by integrated vs. segregated in Panel B. We put these raw means into the analogous regression forms:

$$Learning_{ig} = \beta_0 + \beta_C Cor_{ig} + X_i + \sigma_i + \epsilon_{ig} \tag{13}$$

$$Learning_{ig} = \beta_0 + \beta_{CS} CorXSeg_{ig} + \beta_{CI} CorXInt_{ig} + \beta_{UI} UncorXInt_{ig} + X_i + \sigma_i + \epsilon_{ig} \tag{14}$$

where Cor ($Uncor$) are dummy variables that equal 1 when the game has correlated (uncorrelated) priors, and Int (Seg) are dummies that are 1 when the game has we include the same fixed effects in X_i and again cluster at the session level. Table 4 shows the output of these regressions. We take the same approach of testing the one-sided null hypothesis as in (10). In this case, we want to set $\theta = \beta_C$ in the regression in (13), while $\theta = \beta_{UI} + \beta_{CS} - \beta_{CI}$ in the regression in (14).

Below each model in Table 4 we present the estimated θ and the p-value associated with the one sided hypothesis test that the observed data is produced under the null. Again, as predicted by the theory, all estimated θ are negative. While we do not find statistically significant evidence for the first part of the claim that correlated priors reduce learning, we do find statistically significant evidence at the 10% level for the claim that correlated priors increase the benefits of integration.

Overall, our formal tests show that the data behaves consistently with the hypotheses, and that this is generally statistically significant at the 10% level. These empirical findings confirm the theoretical intuitions and imply a nuanced effect of integration. First, integration is better when groups have correlated priors, such that one group has to learn from the other. Second, the benefits of integration fall as player bias increases, so the characters.

6 Estimating bias

One advantage of the correpondence between the model and the lab game, is that we can use the experiment to estimate the bias preference. In this section, we fit the DeGroot learning model to the experimental data to estimate players' bias, as captured by the model parameter b. Let $\boldsymbol{\omega} \in \{-1, 1\}^6$ be a 6×1 vector of initial

signals, such that each entry ω_i corresponds to the random signal that player i gets. In the bias estimation, we plug in the randomly distributed signals $\boldsymbol{\omega}$ as the initial actions of the players. Without loss of generality, we code the true state of the world as 1, while the false state as -1.

We also randomize the networks of observation as in Figure 2. Let $\boldsymbol{G} \in \mathbb{R}^6 \times \mathbb{R}^6$ be the matrix representing the social network administered randomly in the lab. We parametrize the bias by allowing each link to be represented by $G_{ij} \in \{0, 1-b_s, 1-b_d, 1\}$. Here, the link is valued 0 if ij cannot observe each other in the game, we normalize the value of self links G_{ii} to 1, and $1-b_s, 1-b_d$ represent the links to same and different-ethnicity players respectively. The bias toward players of same- and different-ethnicities is captured by b_s, b_d respectively; the higher this bias, the less players value the information of others from that group, relative to their own information. We impose the network and signals in the lab game, but students apply the bias. Our goal in this chapter is to estimate that bias as captured by (b_s, b_d). Note that here we depart slightly from the model in Section 2, and we allow individuals to have bias against everyone, including members of their own ethnicity. We normalize own-bias to 0, and estimate the bias against others relative to own bias.[30] This is appropriate since there is some empirical evidence that individuals discount everyone's information relative to their own (e.g. Conlon et al., 2022), so here we focus on how that discounting may differ toward individuals of the same- vs. different-ethnicities.

Point identification may not be possible under this model. To illustrate, if the bias one assigns to others is greater than $\frac{2}{3}$, then because agents observe at most 3 other players, they will always follow their own signal and not learn from others. Therefore, any bias between $\frac{2}{3}$ and 1 are empirically indistinguishable. This is a problem in general since any bias between $(\frac{d-1}{d}, 1]$, with d being the highest number of network links, are observationally equivalent. Therefore we must rely on set identification to retrieve bias estimates. In a future iterations of this paper we wish to formally explore set identification of the bias in the model. In this version, we rely on a data-driven process to retrieve a distribution $F(b_s, b_d)$ of most likely biases that the participants apply during the game. We do this by using an approach that captures the bias parameters that minimize the distance between data moments and model moments.

More formally, let $\boldsymbol{m}^{data} \in \{-1, 1\}^{gri}$ be a $gri \times 1$ vector of data moments that capture behavior in the

[30] A key assumption in this parametrization is that the ratio of other to own bias is constant across players, regardless of how many players they observe or the identity of these players. More formally, holding k_i, k_j constant, we assume that G_{ij}/G_{ii} is the same for all i and j.

lab game. Here, g denotes the number of games in the data of interest, r denotes the number of rounds in the game, and i denotes each of the six players playing the game. Each element in this vector, for example m_{gri}^{data} represents the observed action that player i took in game g and round r. On the other hand, let $\boldsymbol{m}^{model}(\boldsymbol{b}) \in \mathbb{R}^{gri}$ be the analogous vector of model moments that are implied by the model parameters $\boldsymbol{b} = (b_s, b_d)$. More specifically, given parameters (b_s, b_d), a typical element of the moment vector is:

$$\boldsymbol{m}_{g,r,i}^{model}((b_s, b_d)) = \theta \left[Y_{g,r-1,i} + (1 - b_s) \sum_{j \in N_{is}} Y_{g,r-1,j} + (1 - b_d) \sum_{k \in N_{id}} Y_{g,r-1,k} \right] \tag{15}$$

$$\text{with } \theta = \frac{1}{1 + |N_{is}|(1 - b_s) + |N_{id}|(1 - b_d)} \tag{16}$$

where N_{is} and N_{id} are the sets of same-ethnicity and different ethnicity players that i observes respectively, and $|N_{is}|, |N_{id}|$ denote the carnality of each set respectively; θ essentially normalizes the weighted average such that our assumption of row-normalized matrix with identical slopes holds. This is the weighted average updating rule analogous to the DeGroot model described in Section 2.

In our estimation procedure, we wish to retrieve the bias estimates that make the model-implied actions most consistent with the observed data. Therefore, for a given set of proposal biases, we define a criterion function $c(b_s, b_d)$ that captures the distance between the model and the data as follows:

$$c(b_s, b_d) = \boldsymbol{e}'(b_s, b_d)\boldsymbol{e}(b_s, b_d) \tag{17}$$

where

$$\boldsymbol{e}(b_s, b_d) = \boldsymbol{m}^{model}(b_s, b_d) - \boldsymbol{m}^{data} \tag{18}$$

Our estimation algorithm is formally detailed in the Algorithm 2 box below. To retrieve a distribution $F(b_s, b_d)$ of biases that minimize the criterion, we first calculate the criterion function on a discrete grid of possible biases. This results in a surface of criterion values for each pair of biases from the grid. Then, we select the set of biases that are equal to or smaller than some percentile value of all the criterion values, therefore selecting the biases that best predict the data. Once these pairs of biases and their associated criterion values are selected, we convert them into a distribution, such that the probability of each pair of bias pair is proportional to the inverse criterion. Thus, the bias parameters that achieve smaller criterion are more likely under this distribution.

Algorithm 2 Estimation of Bias Parameters Distribution

1: **for** $(b_s, b_d) \in X^2$ **do**
2: Calculate set of criterion values $c(b_s, b_d)$ and store the biases and criterion values
3: Calculate the q^{th} percentile of all criterion values and define it as c_{q,X^2}
4: Keep the subset of biases $Y \subset X^2$ such that $\forall (b'_s, b'_d) \in Y : c(b'_s, b'_d) \leq c_{p,X^2}$
5: **end for**
6: **for** $(b_s, b_d) \in Y$ **do**
7: Define the probability of each pair of bias as $p(b_s, b_d) = \frac{c^{-1}(b_s, b_d)}{\sum_{(b_s, b_d) \in Y} c^{-1}(b_s, b_d)}$
8: Store (b_s, b_d) and $p((b_s, b_d))$ into a probability distribution $F(b_s, b_d)$.
9: **end for**
10: **return** $F(b_s, b_d)$

In our estimation we set X to be a discrete set of 50 equidistant points on the line $[0, 1]$, i.e. search for optimal bias parameter on a 50×50 grid on the $[0, 1]^2$ set. Additionally, we set $q = 2$, therefore, we select the 2% of biases that minimize the criterion the most[31].

Figure 6 visualizes the results of this estimation algorithm. We estimate the bias using only the named treatment arms, when ethnicity is salient to the participants. In Panel A, we plot the criterion function for all pairs in the grid of biases, i.e. the outcome of step 2 in our algorithm. It is noticeable that there are clear delineated regions with uniform performance – hinting at the idea that there are sets of biases which are observationally equivalent. It is interesting to note that the shaded regions intersect at $1/k$ points where $k \in \{0, 1, 2, 3\}$, which is the amount of people (from a given ethnicity) a player observes. This illustrates how features of the experiment design – how many individuals one observes – delineates the regions that are observationally equivalent, and constrains point identification.

Panel B in Figure 6 plots the estimated distribution $F(b_s, b_d)$. Specifically, here we plot the 2% of bias parameters that minimize the criterion function. Darker shades denote higher probability, which as detailed in the algorithm, is proportional to the criterion. The criterion plot suggests a conjecture that a likely candidate for the set of biases consistent with the observed data may lie on the line defined by the points $b_s = \frac{1}{2}, b_d = 1$ and $b_s = \frac{2}{3}, b_d = \frac{2}{3}$. Using a more formal set-identification estimation approach is one of the key ways this paper can be improved.

Table 6 presents the means and variances of the estimated distribution $F(b_s, b_d)$. We estimated this distri-

[31] Ideally, we'd like to get as tight of a set of possible biases as possible by setting the percentage to be as small as possible. The 2% is somewhat arbitrary, though we tried other cutoffs and they yield qualitatively similar results.

bution pooling all data. We see that the bias toward individuals of the same ethnicity is about 0.6. This means that players put 0.4 as much weight on the actions of players from their own ethnicity, as much as they put on their own actions. This estimate is similar in magnitude to other estimates from the literature. For example, Mobius et al., 2015 and Conlon et al., 2022 both find similar weights, even though they have very different empirical approaches and subject populations. We see a substantial increase in bias against the other ethnicity, estimating a bias of 0.87, a 45% increase. This means that on average players value the actions of individuals from other ethnicities 0.13 times as much as they value their own action.

7 How does integrating classrooms change social learning?

The Mayor of the Municipality in which we conducted our study has talked about proposals to open an integrated English-language classrooms, and that they are discussing this policy with the Ministry of Education. In this section, we use model-based counterfactuals to formally explore the implications of this policy. Needless to say that our estimates in this section rely on modelling assumptions and abstractions from complexities in real life. To the extent that such complexities may be important, our estimates should be taken less as precise predictions, and more as suggestive implications. In other settings, policymakers can conduct similar exercises under different assumptions that are appropriate in the given context to inform how social learning can change as a consequence of shifts in the social networks.

To conduct the counterfactuals we need three ingredients: 1) bias, 2) initial signals, and 3) counterfactual social networks. For the bias we consider the estimated distribution $F(b_s, b_d)$ from the previous section to inform how agents learn from each other. However, students may decrease their bias once they are integrated (Pettigrew, 1998, Mousa, 2020, Lowe, 2021). We find the same result in a separate study, in which we find that playing cooperative video games can decrease bias, measured through a Dictator Game and a Willingness to Pay for same-ethnicity partner. We find that mixed pairs of players decrease their bias by 0.19 standard deviations compared to same-ethnicity pairs (Anisfeld and Pandiloski, 2024). In the counterfactual networks we therefore consider two bias regimes: one in which bias remains unchanged under integration, i.e. using $F(b_s, b_d)$ from the previous section, and another in which bias shifts by 0.19 standard deviations, i.e. $F'(b_s - 0.19\sigma_{b_s}, b_d - 0.19\sigma_{b_d}) = F(b_s, b_d)$.

Second, we need prior signals, including an average precision (what fraction of agents get correct signals) as well as the distribution of these signals (how are the correct signals distributed across ethnicities). For the average precision, i.e. the fraction of agents who have a true prior, we use use a grid of 10 equidistant points in the range $\mathbb{P}(\omega = 1) \in [0.5, 1]$. For a given average precision we also consider two possible schedules of distribution. Under uncorrelated priors priors, each ethnic group gets the same precision, such that everyone has the same quality of information regardless of ethnicity. Under correlated priors we distribute the signals unevenly, comparing Albanians on one end with Macedonians and Turks on the other, since the number of Albanians is usually similar to the number of Macedonians and Turks combined. While signals may be unevenly distributed in many ways, focusing on two similarly sized groups allows for simple manipulation and keeps things consistent with the rest of the paper. We set group-level precision as follows. First, we identify whether the number of Albanians or Macedonians and Turks is smaller. The smaller group gets a correct signal with certainty (i.e. everyone has a correct prior), while the larger group gets the rest of the correct signals (if they exist) such that the ethnicity-level average precision adds up to the corresponding precision at the network level. To illustrate, suppose there are 30 Albanians, 20 Macedonians, and 15 Turks in one cohort network, and suppose that 0.7 of all agents in the network get the correct signal. Under uncorrelated priors, each ethnicity has a precision of 0.7; under correlated priors, Albanians get a precision of 1, while the Macedonians and Turks get a precision of 0.4428.

Third, we need a way to construct counterfactual networks. We expand the stochastic block model from Section 2 by allowing that linking probabilities depend not only on ethnicity, but also on classroom assignments. Specifically, we imagine that the network is formed such that each pair meets once and a link forms with one of the following probabilities: $p \in \{p_s^1, p_s^0, p_d^1, p_d^0\}$. The subscript here denotes same- or different-ethnicity, while the superscript is coded as 1 if the pair of students are from the same classroom, and 0 if they are from different classrooms. So, if a pair that meets is of the same ethnicity and is in the same classroom, they link with probability p_s^1, if they are of different ethnicity and different classrooms, p_d^0.

As we detail in Section 4, during our baseline survey, we elicited data on social ties from our participants, yielding 12 real-life networks, one for each school-grade level.[32] Using this data on possible and realized links, we can directly estimate p_s^1, p_s^0, p_d^0. To construct counterfactual networks we need an assumption about

[32]In this section, we define a link between individuals i and j if either i or j listed the other person in at least one of the social survey questions (clique, cross ethnic friends, school help, information). In Table 2, we present summary statistics for the real-life networks.

what p_d^1 may be after classrooms are integrated, as we do not observed in the data. To retrieve this quantity, we assume that the ratios of own-to-different classroom linking probabilities are the same across ethnicity:

$$\frac{p_s^1}{p_s^0} = \frac{p_d^1}{p_d^0} \implies p_d^1 = p_d^0 \frac{p_s^1}{p_s^0} \tag{19}$$

Finally, we need to specify how the classrooms are assigned in the world of the counterfactual policy. We do this by adding an additional classroom in each cohort, and then assign students in our sample to this classroom with independent probability $1/c$ where c is the total number of classrooms in that cohort. Therefore, the classroom in expectation reflects the ratios of ethnicities in the cohort.

Equipped with an estimation of bias, a range of initial signals, as well as counterfactual networks, we can play out the evolution of learning with the model. Our main outcome in this section is the fraction of agents who guess the correct state of the world at round $t = 10$ of the social learning process. Let learning at the school-cohort level sc as $l_{sc} = \frac{\sum_{i \in N_{sc}} \mathbb{1}[\text{i guessed correctly at t=10}]}{N_{sc}}$. For each network, we run 1000 iterations within which we pick random bias point from the estimated distribution, and generate a random networks (if using the model). Then, we average over iterations, yielding $\bar{l}_{sc}$, which we average again across networks $L = \frac{1}{12} \sum_{sc} \bar{l}_{sc}$.

In Figure 7 we present the results of the counterfactual exercises. On the x-axis we plot the initial beliefs and on the y-axis we plot the average change in learning L relative to the initial true actions. Few patterns emerge. First, we can see an inverse-U shape of change in learning as a function of initial fraction of true actions. Second, when each ethnicity has the same information, integrating a classroom decreaes social learning. This is not surprising, since the students in the mixed classrooms are linking with different-ethnicity student (who they are more biased against), rather than the same-ethnicity students (who they would've learned more from). We see that if integration decreases bias by 0.19 standard deviations, learning improves under the counterfactual networks, but a large gap still remains. Third, looking at the changes in learning under correlated priors we notice an interesting trend. When the initial correct actions are high, the curves have the similar pattern as under uncorrelated priors. However, for low enough initial true actions, we see that the integrated counterfactual networks outperform the segregated status quo networks. Wen true information is scarce on the aggregate, and students stand to learn from others of different ethnicities, integration can enable flow of information and improve learning. This exercise showcases how integration, bias, and the

informational environment matter in shaping social learning. In particular, we illustrate how our modelling approach can be combined with our experimental results to examine how a policy counterfactual of integrated classrooms may shape social learning forces in real life.

8 Conclusion and next steps

How does segregation affect social learning? Do correlated priors make segregation harmful? Does bias make it beneficial? We tackle these questions with a combination of theory and empirics. A model of network formation and social learning implies two formal hypotheses. First, bias lowers learning and makes the benefits of segregation higher. Second, correlated priors can erode learning, but make the benefits of segregation higher. To formally test these two hypotheses, we design a variant of a ball-in-urn game for 6 people, and implement it in interethnic high schools in Macedonia.

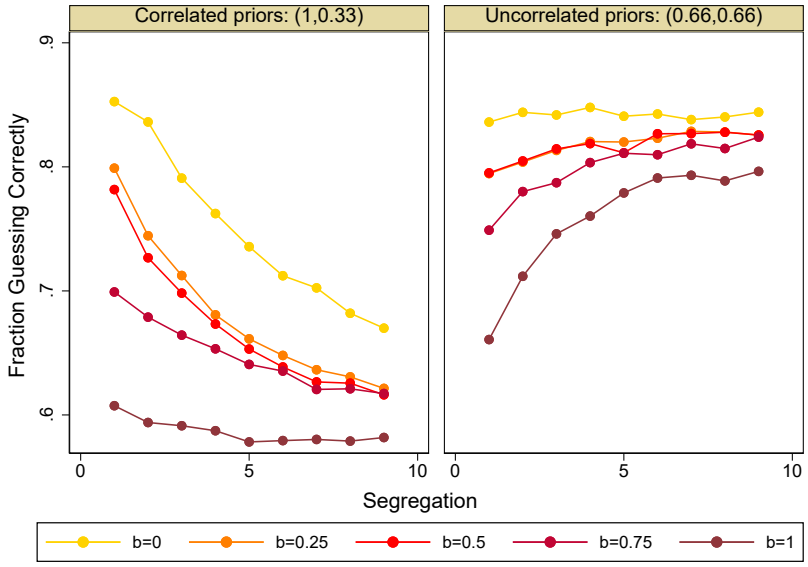
We find compelling evidence for both hypotheses. First, when groups are composed of more low-bias players, participants learn better when they are integrated than when they are segregated, but this is not true in the high-bias groups. We reject the null hypothesis that bias increases the benefits of integration at the 10% level. We also find that correlated priors cause higher learning in the integrated networks than in the segregated networks, but again, this is not true under uncorrelated priors. We also reject the null that correlated priors decrease the benefits of integration at the 10% level. Both the theory and empirics agree that there is a nuanced effect of integration. It can be beneficial if one group has beneficial information for another, but harmful if the groups are sufficiently biased against one another.

How should policymakers think about these trade-offs? To illustrate the implications of our results, we conduct counterfactual simulations to analyze social learning under a counterfactual policy of integrated classrooms. We first structurally estimate the bias individuals have against others in the learning game. We find that individuals' bias against the actions of out-group members is 45% higher than the bias against in-group members. We then combine these estimates with data on real-life networks and a model of network formation to construct counterfactual social networks under the new policy. We find that the counterfactual integrated networks lead to more learning than the status-quo networks only if initial signals are highly correlated and sufficiently low number of individuals take the correct actions at the onset of the game.

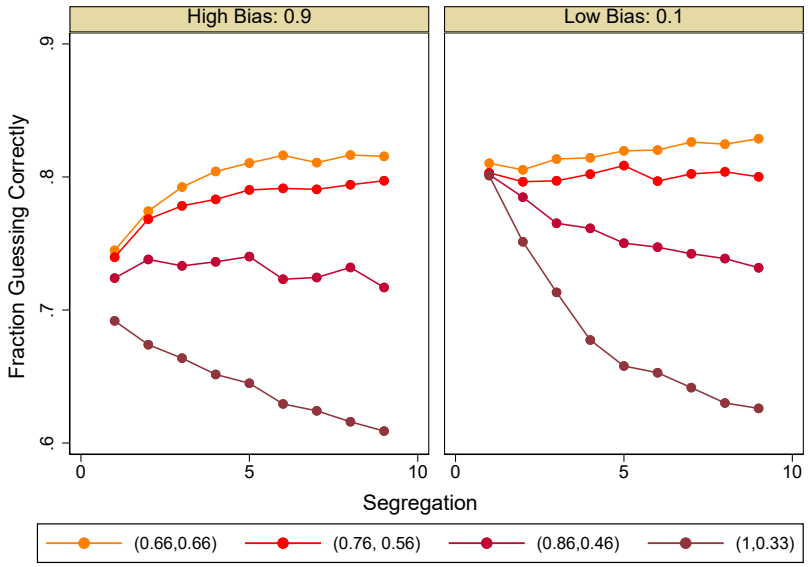
We contribute to two large strands of literature. Social learning has been studied widely in economics, but the literature has either considered identity without network structure, or network structure without identity. Our study aims to fill this gap by simultaneously considering both in the process of social learning. We find that whether segregation is good or bad depends on both the correlation of priors as well as the bias. While the former is a conventional wisdom in the theoretical literature of social learning on networks, the latter is a contribution of our study. We also add to a large literature on identity that considers both out-group bias and in-group conformity, and show evidence of both in the process of social learning.

We see two general directions for improving the current version of the paper. First, we need to formalize the estimation of the structural model. While the general principles of our estimation procedure are well-established, the process includes some arbitrary decisions. A more appropriate approach that we are implementing is using a discrete choice framework such as mixed logit, that can estimate the bias even at the individual level. Estimating bias at the individual level can be useful in describing how bias in learning is related to other observed characteristics. Second, we can expand on our counterfactual simulations by exploring how the results change under different modelling assumptions. Once we estimate individual-level bias parameters, we can ask alternative questions about how targeting low-bias students for the integrated classrooms compares to the other policy counterfactuals that integrate but do not target.

Tables and Figures



(a) Panel A: Segregated learning with bias



(b) Panel B: Segregated learning with correlated priors

Figure 1: Theoretical simulations and hypotheses. This figure plots the results from the theoretical simulations. On the x-axis we plot the extent of segregation, p_s/p_d. On the y-axis we plot the fraction of agents who guess the correct state of the world, which is calculated using Algorithm 1. Panel A plots segregated learning over bias. Each different line represents the relationship between segregation and learning for different values of bias (darker shades denote higher bias). On the left, we plot these relationships for correlated priors ($s_1 = 1, s_2 = 0.33$), and on the right for uncorrelated priors ($s_1 = s_2 = 0.66$). Panel B plots segregated learning over correlation of priors. Each different line represents the relationship between segregation and learning for different values of correlation (darker shades denote higher correlation). On the left, we plot these relationships for high bias ($b = 0.9$), and on the right for low bias ($b = 0.1$)

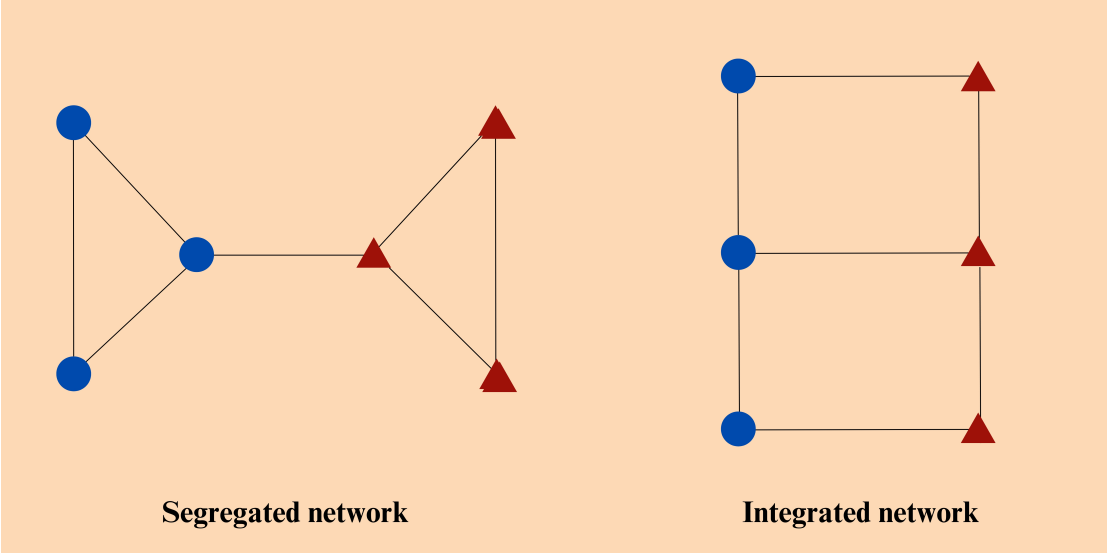


Figure 2: Observation networks in the game. This figure plots the network treatment in the lab game, composed of 6 players from 2 distinct ethnicities. Nodes denote players, and ethnicity is marked by color and shape. Lines represent mutual observation, so any two nodes that are linked can see each other's action's during the game.

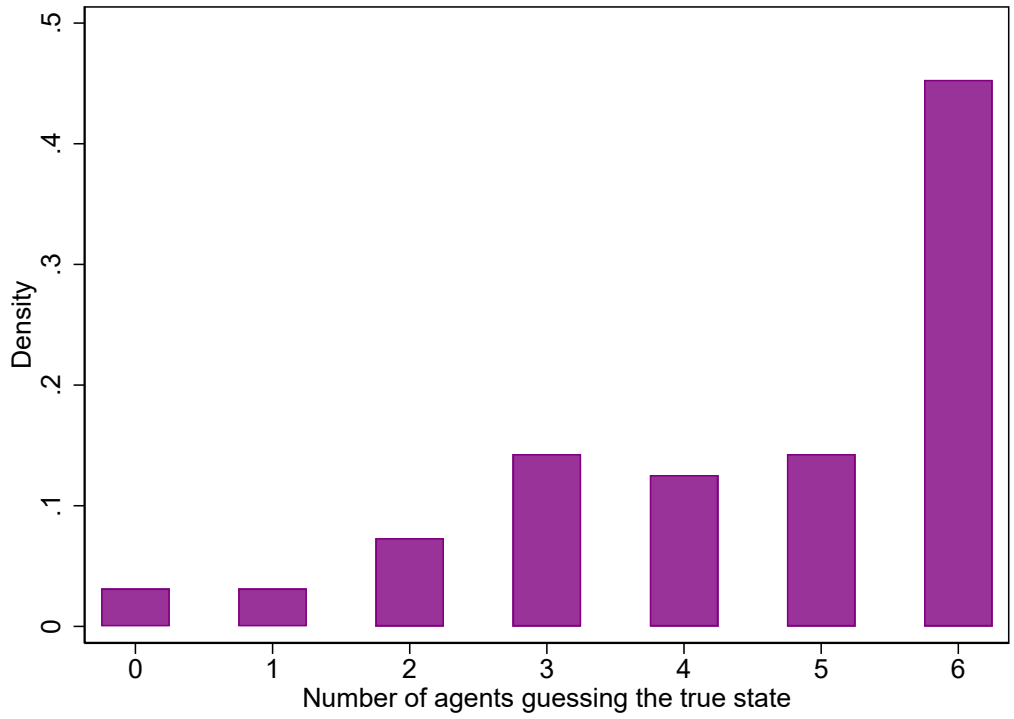
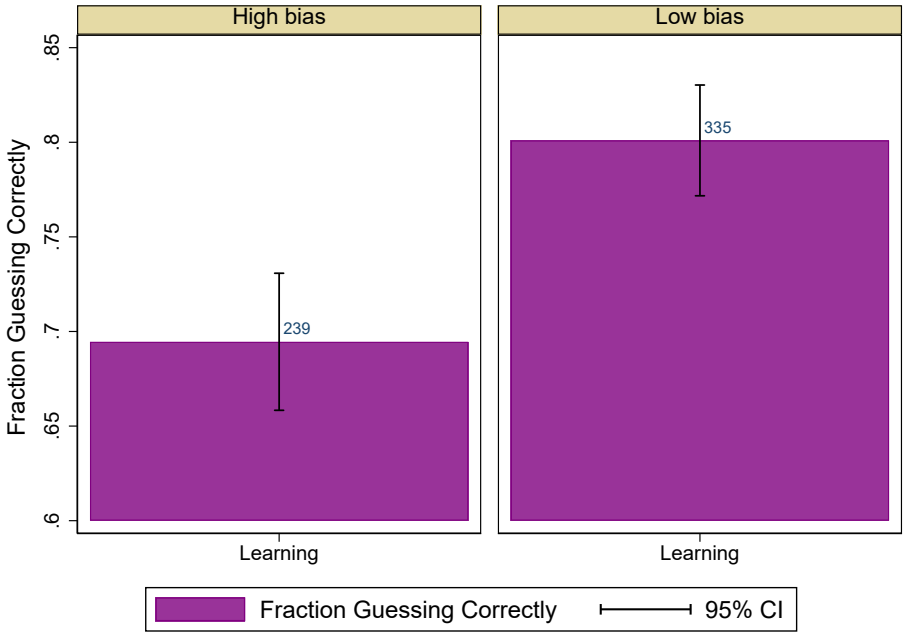
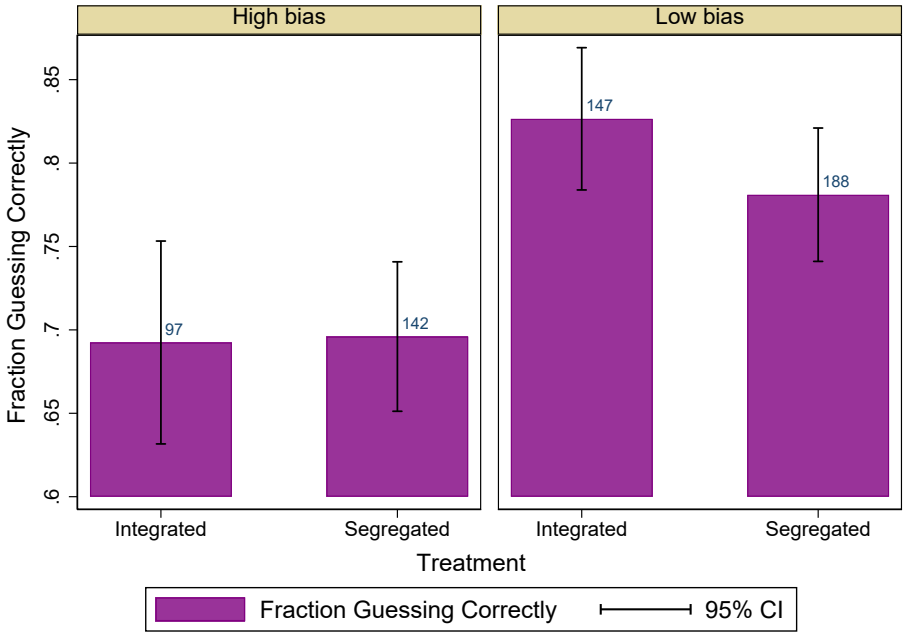


Figure 3: Histogram of learning outcomes. This histogram plots the learning outcome at the network-game level. The *x-axis* presents the number of players that correctly guess the true color in the last round, and the *y-axis* the fraction of network-game outcomes that take that outcome.

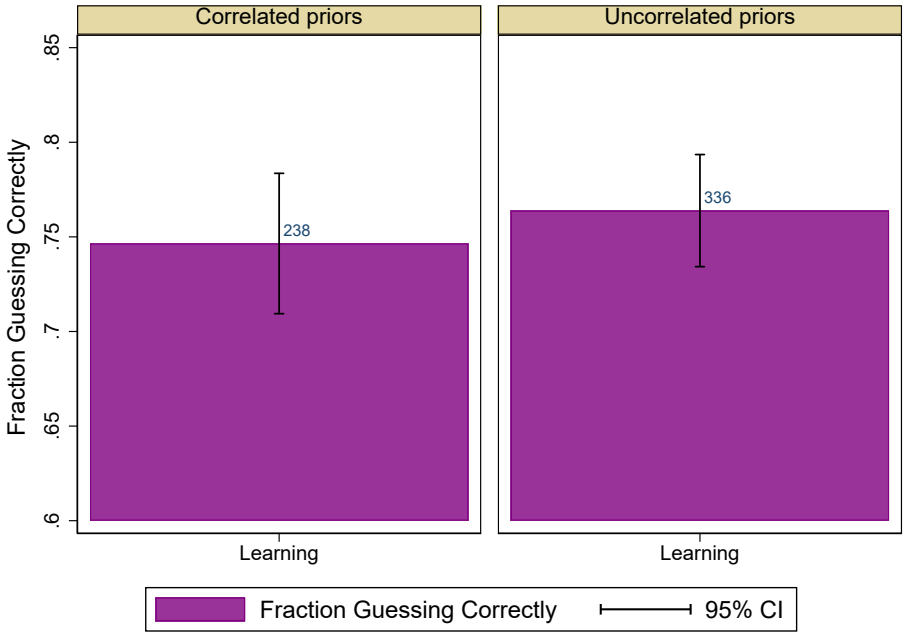


(a) Panel A: Overall learning by bias

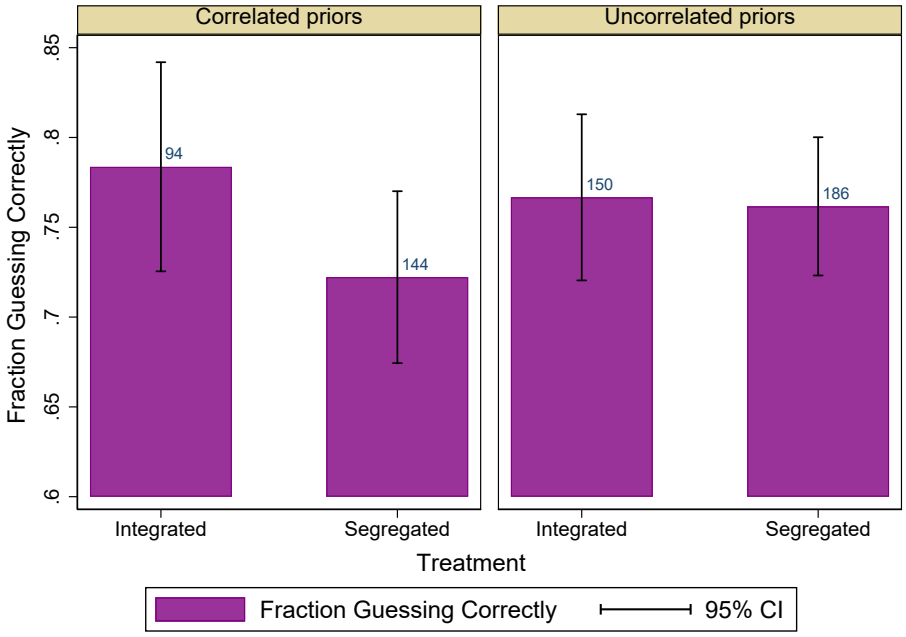


(b) Panel B: Learning by bias, *Segregated* vs. *Integrated* conditions.

Figure 4: Testing Hypothesis 1. This figure plots the mean of learning over network-game observations. Learning is defined as the fraction of the 6 players who correctly guess the color in the last round of the game (N=574). In Panel A, we plot learning by bias. The left panel plots learning for 6-person groups in which half or fewer students have a cross-ethnic friend, while the right panel plots learning for 6-person groups in which more than half of the players have a cross ethnic friendship. For each player, *cross-ethnic friendship* is equal to 1 if the student names at least one friend from a different ethnicity in the baseline social network survey. Panel B is analogous to Panel A, but it additionally splits learning separately for the *Segregated* and *Integrated* conditions. Each figure plots the means, 95% confidence interval, and indicates sample size on top of the bar. We exclude practice rounds.

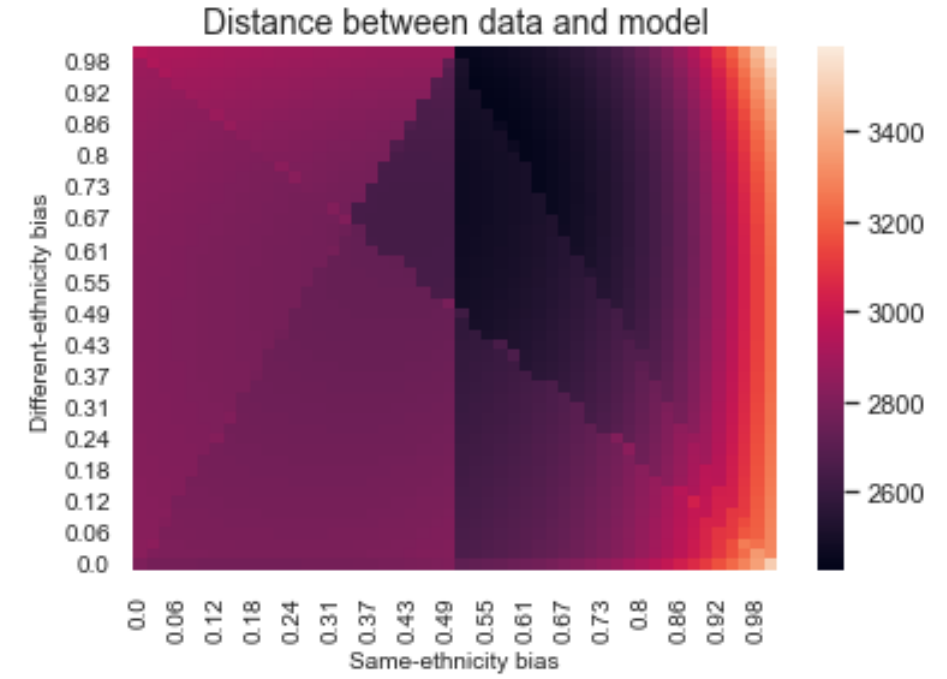


(a) Panel A: Overall learning by correlated priors

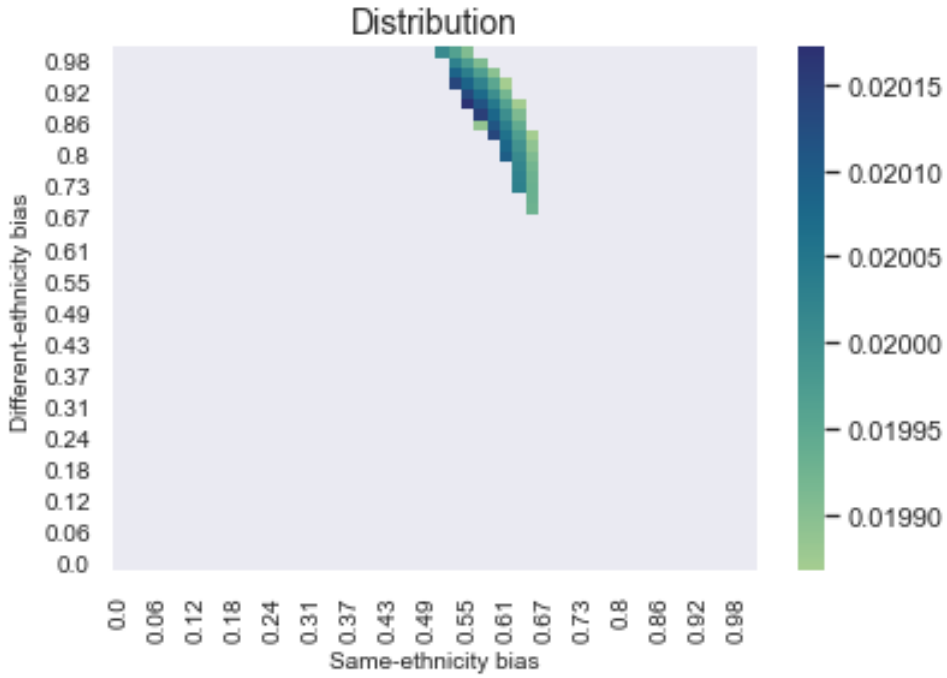


(b) Panel B: Learning by correlated priors, *Segregated* vs. *Integrated* conditions.

Figure 5: Testing Hypothesis 2. This figure plots the mean of learning over network-game observations. Learning is defined as the fraction of the 6 players who correctly guess the color in the last round of the game (N=574). In Panel A, we plot the average true guess by correlated priors. The right panel plots learning for games in which both ethnicities received 2 correct and 1 incorrect signals (uncorrelated priors), while the left panel plots learning for games in which one ethnicity gets 3 correct signals, while the other 2 incorrect and 1 correct signal (correlated priors). In each game signals are distributed to players randomly. Panel B is analogous to Panel A, but additionally splits learning separately for the *Segregated* and *Integrated* conditions. Each figure plots the means, 95% confidence interval, and indicates sample size N on top of the bar. We exclude practice rounds.



(a) Panel A: Moment distances



(b) Panel B: Distribution of the weights that achieve 2% lowest moment distances

Figure 6: Estimating bias. This figure visualizes the bias estimation procedure. The estimation is done over a 50×50 grid of same-ethnicity bias b_s (x-axis) and different-ethnicity bias b_d (y-axis). In Panel A, for each pair of bias, we calculate the criterion, i.e. the distance between the model-predicted moments and experimental data, as in Equation (17). Darker shades indicate smaller distances and therefore better fit. In Panel B, we form a distribution using the 2% of bias pairs that best predict the data, where the probability of each (b_s, b_d) is proportional to criterion (better fit means higher probability). This estimation procedure is detailed in Algorithm 2.

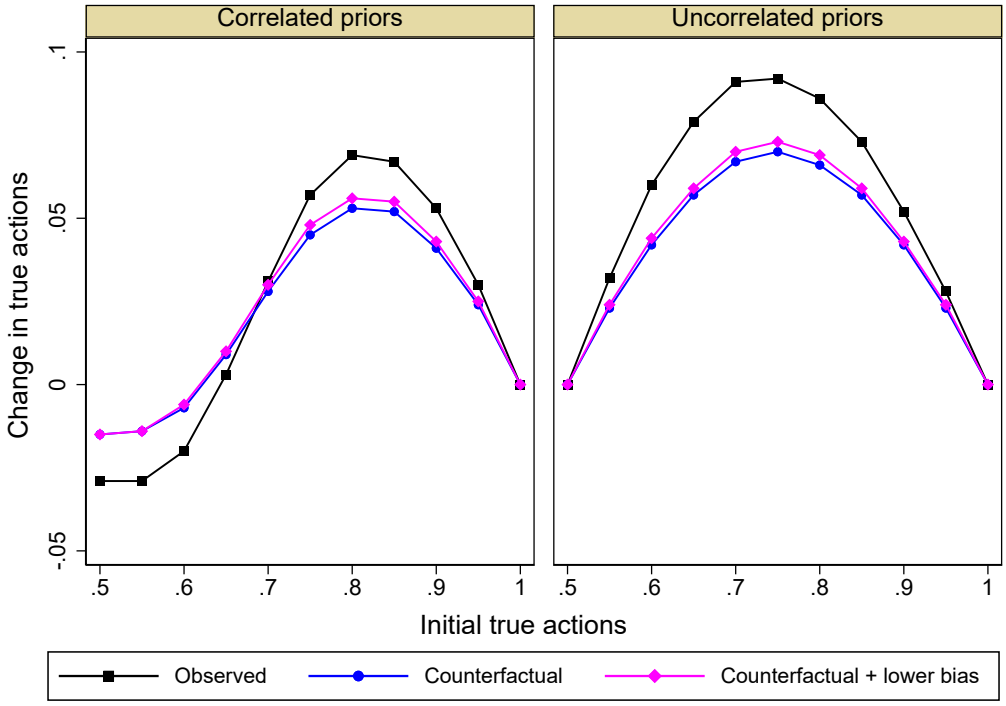


Figure 7: Counterfactuals. This figure plots the results from the counterfactual exercises. On the *x-axis* we plot the precision, i.e. the initial fraction of agents who take the correct action. On the *y-axis* we plot change in learning relative to the initial . We calculate the change in learning using the estimated bias distribution from the experiment. On the right panel we plot the simulations with uncorrelated priors, such that each ethnicity has the same initial true action, and the left panel presents learning under correlated priors, in which there are differences in the precision of the priors. In each line we plot a different type of network and bias used to simulate social learning. The *Observed* line plots the change in learning using the measured status-quo networks. The *Counterfactual* line plots the change in learning using model-generated networks under the counterfactual scenario, in which one integrated classroom is introduced in each cohort. The *Counterfactual+lower bias*, takes the same counterfactual network, but assumes that integration also lowers bias in learning in the same magnitude as in Anisfeld and Pandiloski, 2024. The network formation model and the assumptions made about counterfactual parameters are detailed in Section 7.

Characteristic	N	Mean	SD	Min	Max
Female	527	0.49	0.50	0	1
Age at registration	499	16.31	1.12	13	23
Macedonian	527	0.40	0.49	0	1
Albanian	527	0.41	0.49	0	1
Turkish	527	0.19	0.39	0	1
At least one cross-ethnic friend	527	0.45	0.50	0	1
Dictator Game bias	511	32.63	87.00	-300	500
Fluency in other local languages	511	3.19	2.88	0	10
Proximity to other local ethnicities	511	2.08	1.16	1	7
Generalized Trust	511	6.60	2.46	1	12
At least one parent has college degree	521	0.44	0.50	0	1
Both parents have college degrees	527	0.18	0.39	0	1
Both parents have college degrees	515	0.50	0.50	0	1
Grade Point Average (2023-2024)	524	4.54	0.72	2.1	5

Table 1: Summary statistics of individual characteristics. This table shows the characteristics of participants. There are 527 unique participants in this study, though some questions have missing responses since they were not mandatory and/or data is missing for some students. Female is a binary indicator equal to 1 if the enumerator-reported gender is female. Age at registration is derived from the self-reported birth month and year at registration. Macedonian, Albanian, and Turkish are binary variables that equal 1 if the participant is enrolled in a Macedonian, Albanian, or Turkish language instruction classrooms. At least one cross ethnic friend is coded as 1 if the student reports at least one friend from another instruction language in the baseline survey. Dictator game bias is defined as the average amount sent to a co-ethnic minus the average amount sent to a cross-ethnic peer on a Dictator Game. Fluency in other local languages is the average reported fluency (1-10 rank) in local languages that is not the participant's mother tongue. Proximity to other local ethnicities is the average reported proximity on an Inclusion of Others in Self psychometric test. Trust is a self report (0-10 rank) on the general trustworthiness of people in society. Parent education and own educational attainment is measured by a survey report, in which students report the highest educational level each of their parents attained, and the highest educational level they want to attain. GPA comes from administrative school data, and students are graded from 1 (fail) to 5 (excellent).

School	Grade	N	Average Degree	Clustering	Components	Cross-Ethnic Links (%)
School 1	1	94	5.13	0.45	6	7.88
	2	84	4.05	0.39	11	9.41
	3	54	2.56	0.41	12	1.45
	4	22	2.55	0.45	4	0.00
School 2	1	70	2.63	0.31	19	4.35
	2	77	3.92	0.40	9	3.97
	3	40	3.30	0.46	6	1.52
	4	52	2.69	0.49	14	0.00
School 3	1	92	4.07	0.35	3	6.95
	2	101	5.41	0.44	2	5.13
	3	85	3.81	0.41	7	4.94
	4	49	2.49	0.48	11	1.64

Table 2: Summary statistics of social networks. This table shows network statistics for the 12 real-life networks. Here we use all students with baseline surveys (N=820), not just those who have played the game. Average degree is the average number of reported links. Clustering represents the probability that two people are linked given that they have a mutual friend. Components is the number of disconnected clusters in the network. Cross-ethnic links represents the percentage of links in the network that are across individuals of different ethnicities.

	Network-level Learning			
	Hypothesis 1.1		Hypothesis 1.2	
	(1)	(2)	(3)	(4)
High Bias	-0.094*	-0.094*		
	(0.04)	(0.04)		
High Bias X Segregated			0.008	0.014
			(0.04)	(0.04)
Low Bias X Integrated			0.129**	0.134**
			(0.05)	(0.05)
Low Bias X Segregated			0.076	0.077
			(0.05)	(0.05)
Constant	0.643***	0.643***	0.544***	0.542***
	(0.05)	(0.05)	(0.06)	(0.06)
Strata FE	Y	Y	Y	Y
Ethnic pairing FE	N	Y	N	Y
R^2	0.116	0.140	0.121	0.146
N	574	574	574	574
θ	-0.094	-0.094	-0.060	-0.071
$\Pr(\theta > 0)$	0.005***	0.008***	0.079*	0.048**

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Testing Hypothesis 1. This table presents the regressions that test Hypothesis 1. In every column, the outcome is the network-level learning in the last round of each game, measured as the fraction of the 6 students that correctly guessed the color. Columns (1) and (2) test whether correlated priors decrease learning, and map onto the raw means presented in Figure 4 Panel A. Columns (3) and (4) test whether bias decreases the benefits of integration, and map onto the raw means presented in Figure 4 Panel B. Columns (1) and (3) include school-by-cohort strata fixed effects, while columns (2) and (4) add a fixed effect for the ethnic pairing in the session (Macedonian-Albanian, Albanian-Turkish, Turkish-Macedonian). At the bottom of each column, we show the estimated θ, which theory predicts $\theta \leq 0$. Below it, we present the p-value associated with the one-sided test of the null hypothesis that $\theta > 0$.

	Network-level Learning			
	Hypothesis 2.1		Hypothesis 2.2	
	(1)	(2)	(3)	(4)
Correlated Signals	-0.019	-0.024		
	(0.02)	(0.02)		
Uncorrelated Priors X Integrated			0.004	0.006
			(0.02)	(0.02)
Correlated Priors X Segregated			-0.043	-0.047
			(0.03)	(0.03)
Correlated Priors X Integrated			0.021	0.017
			(0.04)	(0.04)
Constant	0.617***	0.619***	0.617***	0.618***
	(0.02)	(0.02)	(0.03)	(0.03)
Strata FE	Y	Y	Y	Y
Ethnic pairing FE	N	Y	N	Y
R^2	0.094	0.124	0.099	0.129
N	574	574	574	574
θ	-0.019	-0.024	-0.060	-0.058
$\Pr(\theta > 0)$	0.216	0.163	0.053*	0.058*

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Testing Hypothesis 2. This table presents the regressions that test Hypothesis 2. In every column, the outcome is the network-level learning in the last round of each game, measured as the fraction of the 6 students that correctly guessed the color. Columns (1) and (2) test whether bias decreases learning, and map onto the raw means presented in Figure 5 Panel A. Columns (3) and (4) test whether bias decreases the benefits of integration, and map onto the raw means presented in Figure 5 Panel B. Columns (1) and (3) include school-by-cohort strata fixed effects, while columns (2) and (4) add a fixed effect for the ethnic pairing in the session (Macedonian-Albanian, Albanian-Turkish, Turkish-Macedonian). At the bottom of each column, we present the estimated θ which theory predicts $\theta \leq 0$. Below it, we present the p-value associated with the one-sided test of the null hypothesis that $\theta > 0$.

	No cross-ethnic friends			Any cross ethnic friend			
	N	Mean	SD	N	Mean	SD	Difference
Dictator Game bias	291	34.71	93.60	220	29.89	77.54	-4.822
Fluency in other local languages	291	2.21	2.48	220	4.48	2.87	2.266***
Proximity to other local ethnicities	291	1.87	1.15	220	2.34	1.12	0.466***
Trust	291	6.71	2.41	220	6.44	2.52	-0.274
At least one parent has college degree	287	0.47	0.50	234	0.41	0.49	-0.060
Both parents have college degrees	291	0.20	0.40	236	0.17	0.37	-0.034
Wants to attain graduate degree	286	0.52	0.50	229	0.47	0.50	-0.046
GPA	288	4.62	0.66	236	4.44	0.78	-0.183***

Table 5: What's in a "type"? This table shows the characteristics of participants with no cross-ethnic friends and at least one cross-ethnic friend, and the difference on these characteristics between these two groups (coefficient from a regression of given characteristic on any cross friend indicator). Dictator game bias is defined as the average amount sent to a co-ethnic minus the average amount sent to a cross-ethnic peer on a Dictator Game. Fluency in other local languages is the average reported fluency (1-10 rank) in local languages that is not the participant's mother tongue. Proximity to other local ethnicities is the average reported proximity on an Inclusion of Others in Self psychometric test. Trust is a self report (0-10 rank) on the general trustworthiness of people in society. Parent education and own educational attainment is measured by a survey report, in which students report the highest educational level each of their parents attained, and the highest educational level they want to attain. GPA data comes from administrative data sets in the schools.

Parameter	b_s	b_d
Mean	0.60	0.87
Variance	0.01	0.15

Table 6: Estimated biases. This table plots the means and variances of the same-ethnicity (b_s) and different-ethnicity (b_d) bias using the estimation procedure described in **Algorithm 2**, and visualized in Figure 6

References

Acemoglu, D., Bimpikis, K., & Ozdaglar, A. (2014). Dynamics of information exchange in endogenous social networks: Endogenous social networks. *Theoretical Economics*, *9*(1), 41–97. https://doi.org/10.3982/TE1204

Acemoglu, D., Dahleh, M. A., Lobel, I., & Ozdaglar, A. (2011). Bayesian Learning in Social Networks. *The Review of Economic Studies*, *78*(4), 1201–1236. https://doi.org/10.1093/restud/rdr004

Anisfeld, A., & Pandiloski, P. (2024). Language, Identity and Ethnic Cohesion in Schools. *Working Paper*.

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*(4), 596–612. https://doi.org/10.1037/0022-3514.63.4.596

Bandiera, O., Barankay, I., & Rasul, I. (n.d.). SOCIAL PREFERENCES AND THE RESPONSE TO INCENTIVES: EVIDENCE FROM PERSONNEL DATA. *QUARTERLY JOURNAL OF ECONOMICS*.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2013). The Diffusion of Microfinance. *Science*, *341*(6144), 1236498. https://doi.org/10.1126/science.1236498

Bar, R., & Zussman, A. (2020). Identity and Bias: Insights from Driving Tests. *The Economic Journal*, *130*(625), 1–23. https://doi.org/10.1093/ej/uez048

Bateson, G. (1958). *Naven: A Survey of the Problems suggested by a Composite Picture of the Culture of a New Guinea Tribe drawn from Three Points of View*. Stanford University Press.

Bayer, P., Ross, S. L., & Topa, G. (2008). Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes. *Journal of Political Economy*, *116*(6), 1150–1196. https://doi.org/10.1086/595975

Beaman, L., & Magruder, J. (2012). Who Gets the Job Referral? Evidence from a Social Networks Experiment. *The American Economic Review*, *102*(7), 3574–3593.

Beaman, L. A. (2012). Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the U.S. *The Review of Economic Studies*, *79*(1), 128–161. https://doi.org/10.1093/restud/rdr017

Benjamin, D. J., Choi, J. J., & Strickland, A. J. (2010). Social Identity and Preferences. *American Economic Review*, *100*(4), 1913–1928. https://doi.org/10.1257/aer.100.4.1913

Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste* (R. Nice, Trans.). Harvard University Press.
OCLC: 10323218.

Bramoullé, Y., Djebbari, H., & Fortin, B. (2020). Peer Effects in Networks: A Survey. *Annual Review of Economics*, *12*(Volume 12, 2020), 603–629. https://doi.org/10.1146/annurev-economics-020320-033926

Burgess, R., Jedwab, R., Miguel, E., Morjaria, A., & Padró i Miquel, G. (2015). The Value of Democracy: Evidence from Road Building in Kenya. *American Economic Review*, *105*(6), 1817–1851. https://doi.org/10.1257/aer.20131031

Bursztyn, L., Ederer, F., Ferman, B., & Yuchtman, N. (2014). Understanding Mechanisms Underlying Peer Effects: Evidence From a Field Experiment on Financial Decisions. *Econometrica*, *82*(4), 1273–1301. https://doi.org/10.3982/ECTA11991

Bursztyn, L., Egorov, G., & Jensen, R. (2019). Cool to be Smart or Smart to be Cool? Understanding Peer Pressure in Education. *The Review of Economic Studies*, *86*(4), 1487–1526. https://doi.org/10.1093/restud/rdy026

Bursztyn, L., & Jensen, R. (2015). How Does Peer Pressure Affect Educational Investments? *. *The Quarterly Journal of Economics*, *130*(3), 1329–1367. https://doi.org/10.1093/qje/qjv021

Cai, J., De Janvry, A., & Sadoulet, E. (2015). Social Networks and the Decision to Insure. *American Economic Journal: Applied Economics*, *7*(2), 81–108.

Chandrasekhar, A. G. (2016, April). Econometrics of Network Formation. In *The Oxford Handbook of the Economics of Networks* (pp. 302–357). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199948277.013.21

Chandrasekhar, A. G., Larreguy, H., & Xandri, J. P. (2020). Testing Models of Social Learning on Networks: Evidence From Two Experiments. *Econometrica*, *88*(1), 1–32. https://doi.org/10.3982/ECTA14407

Charness, G., Rigotti, L., & Rustichini, A. (2007). Individual Behavior and Group Membership. *American Economic Review*, *97*(4), 1340–1352. https://doi.org/10.1257/aer.97.4.1340

Chen, Y., Harper, F. M., Konstan, J., & Li, S. X. (2010). Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens. *American Economic Review*, *100*(4), 1358–1398. https://doi.org/10.1257/aer.100.4.1358

Conley, T. G., & Udry, C. R. (2010). Learning about a New Technology: Pineapple in Ghana. *American Economic Review*, *100*(1), 35–69. https://doi.org/10.1257/aer.100.1.35

Conlon, J. J., Mani, M., Rao, G., Ridley, M., & Schilbach, F. (n.d.). Learning in the Household.

Conlon, J. J., Mani, M., Rao, G., Ridley, M. W., & Schilbach, F. (2022, August). Not Learning from Others. https://doi.org/10.3386/w30378

Corno, L., La Ferrara, E., & Burns, J. (2022). Interaction, Stereotypes, and Performance: Evidence from South Africa. *American Economic Review*, *112*(12), 3848–3875. https://doi.org/10.1257/aer.20181805

Damm, A. P. (2009). Ethnic Enclaves and Immigrant Labor Market Outcomes: Quasi-Experimental Evidence. *Journal of Labor Economics*, *27*(2), 281–314. https://doi.org/10.1086/599336

DeGroot, M. H. (1974). Reaching a Consensus. *Journal of the American Statistical Association*, *69*(345), 118–121. https://doi.org/10.2307/2285509

Duflo, E., & Saez, E. (2002). Participation and investment decisions in a retirement plan: The influence of colleagues' choices. *Journal of Public Economics*, *85*(1), 121–148. https://doi.org/10.1016/S0047-2727(01)00098-6

Duflo, E., & Saez, E. (2003). The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment. *The Quarterly Journal of Economics*, *118*(3), 815–842.

Dupas, P. (2014). Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment. *Econometrica*, *82*(1), 197–228.

Erdős, P., & Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, *6*(290-297), 18.

Falk, A., & Zehnder, C. (2013). A city-wide experiment on trust discrimination. *Journal of Public Economics*, *100*, 15–27. https://doi.org/10.1016/j.jpubeco.2013.01.005

Fisman, R., Shi, J., Wang, Y., & Xu, R. (2018). Social ties and favoritism in Chinese science. *Journal of Political Economy*, *126*(3), 1134–1171. https://doi.org/10.1086/697086

Foster, A. D., & Rosenzweig, M. R. (1995). Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture. *Journal of Political Economy*, *103*(6), 1176–1209.

French Jr., J. R. P. (1956). A formal theory of social power. *Psychological Review*, *63*(3), 181–194. https://doi.org/10.1037/h0046123

Glaeser, E. L., Sacerdote, B., & Scheinkman, J. A. (1996). Crime and Social Interactions*. *The Quarterly Journal of Economics*, *111*(2), 507–548. https://doi.org/10.2307/2946686

Goel, D., & Lang, K. (2019). Social Ties and the Job Search of Recent Immigrants. *ILR Review*, *72*(2), 355–381. https://doi.org/10.1177/0019793917729350

Golub, B., & Sadler, E. (n.d.). LEARNING IN SOCIAL NETWORKS.

Golub, B., & Jackson, M. O. (2010). Naïve Learning in Social Networks and the Wisdom of Crowds. *American Economic Journal: Microeconomics*, *2*(1), 112–149. https://doi.org/10.1257/mic.2.1.112

Golub, B., & Jackson, M. O. (2012). How Homophily Affects the Speed of Learning and Best-Response Dynamics. *The Quarterly Journal of Economics*, *127*(3), 1287–1338. https://doi.org/10.1093/qje/qjs021

Golub, B., & Sadler, E. (2016, June). Learning in Social Networks. In *The Oxford Handbook of the Economics of Networks* (Yann Bramoullé, Andrea Galeotti, and Brian W. Rogers).

Granovetter, M. (1983). The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, *1*, 201–233. https://doi.org/10.2307/202051

Gump, B. B., & Kulik, J. A. (1997). Stress, affiliation, and emotional contagion. *Journal of Personality and Social Psychology*, *72*(2), 305–319. https://doi.org/10.1037/0022-3514.72.2.305

Harary, F. (1959). A criterion for unanimity in French's theory of social power. In *Studies in social power* (pp. 168–182). Univer. Michigan.

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098. https://doi.org/10.1198/016214502388618906

Jackson, M. O. (2008). *Social and Economic Networks*. Princeton University Press.

Kramon, E., & Posner, D. N. (2016). Ethnic Favoritism in Education in Kenya. *Quarterly Journal of Political Science*, *11*(1), 1–58. https://doi.org/10.1561/100.00015005

Kremer, M., & Miguel, E. (2007). The Illusion of Sustainability*. *The Quarterly Journal of Economics*, *122*(3), 1007–1065. https://doi.org/10.1162/qjec.122.3.1007

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). *The people's choice*. Duell, Sloan & Pearce.

Leider, S., Möbius, M. M., Rosenblat, T., & Do, Q.-A. (2009). Directed Altruism and Enforced Reciprocity in Social Networks *. *Quarterly Journal of Economics*, *124*(4), 1815–1851. https://doi.org/10.1162/qjec.2009.124.4.1815

Lowe, M. (2021). Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration. *American Economic Review*, *111*(6), 1807–1844. https://doi.org/10.1257/aer.20191780

Mackie, D. M., Worth, L. T., & Asuncion, A. G. (1990). Processing of persuasive in-group messages. *Journal of Personality and Social Psychology*, *58*(5), 812–822. https://doi.org/10.1037//0022-3514.58.5.812

Marx, B., Pons, V., & Suri, T. (n.d.). Diversity and Team Performance in a Kenyan Organization.

Massey, C., & Wu, G. (2005). Detecting Regime Shifts: The Causes of Under- and Overreaction. *Management Science*, *51*(6), 932–947. https://doi.org/10.1287/mnsc.1050.0386

Meyer, C. D. (2023, January). *Matrix Analysis and Applied Linear Algebra, Second Edition*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611977448

Miller, G., & Mobarak, A. M. (n.d.). GENDER DIFFERENCES IN PREFERENCES, INTRA-HOUSEHOLD EXTERNALITIES, AND THE LOW DEMAND FOR IMPROVED COOKSTOVES.

Mobius, M., Phan, T., & Szeidl, A. (2015, March). Treasure Hunt: Social Learning in the Field. https://doi.org/10.3386/w21014

Mobius, M., & Rosenblat, T. (2014). Social Learning in Economics. *Annual Review of Economics*, *6*(Volume 6, 2014), 827–847. https://doi.org/10.1146/annurev-economics-120213-012609

Morgan, T. J. H., Rendell, L. E., Ehn, M., Hoppitt, W., & Laland, K. N. (2012). The evolutionary basis of human social learning. *Proceedings. Biological Sciences*, *279*(1729), 653–662. https://doi.org/10.1098/rspb.2011.1172

Mousa, S. (2020). Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq. *Science*, *369*(6505), 866–870. https://doi.org/10.1126/science.abb3153

Munshi, K. (2004). Social learning in a heterogeneous population: Technology diffusion in the Indian Green Revolution. *Journal of Development Economics*, *73*(1), 185–213. https://doi.org/10.1016/j.jdeveco.2003.03.003

Oster, E., & Thornton, R. (2012). Determinants of Technology Adoption: Peer Effects in Menstrual Cup Take-Up. *Journal of the European Economic Association*, *10*(6), 1263–1293. https://doi.org/10.1111/j.1542-4774.2012.01090.x

Pettigrew, T. F. (1998). INTERGROUP CONTACT THEORY. *Annual Review of Psychology*, *49*(Volume 49, 1998), 65–85. https://doi.org/10.1146/annurev.psych.49.1.65

Schalk, J., Fischer, A., Doosje, B., Wigboldus, D., Hawk, S., Rotteveel, M., & Hess, U. (2011). Convergent and Divergent Responses to Emotional Displays of Ingroup and Outgroup. *Emotion (Washington, D.C.)*, *11*, 286–98. https://doi.org/10.1037/a0022582

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202

Weizsäcker, G. (2010). Do We Follow Others When We Should? A Simple Test of Rational Expectations. *American Economic Review*, *100*(5), 2340–2360. https://doi.org/10.1257/aer.100.5.2340

Zelizer, A. (2019). Is Position-Taking Contagious? Evidence of Cue-Taking from Two Field Experiments in a State Legislature. *American Political Science Review*, *113*(2), 340–352. https://doi.org/10.1017/S0003055419000078

A Appendices

A.1 The case of asympotic learning with continuous updating

Our modelling approach here mirrors Golub and Jackson, 2012, a seminal theoretical study of how segregation slows learning in social networks (henceforth GJ). They too employ a stochastic block model as a network formation process as well as a DeGroot model as a social learning process. We departs in two key ways. First, we extend their model by introducing a bias parameter b, such that individuals are allowed to be biased against individuals from different groups; in GJ, $b = 0$ and so individuals do not have any bias against other types. Second, GJ consider a continuous action space in an asymptotic learning framework, that is, as society grows to infinity $(n \to \infty)$ and with an updating rule analogous to Equation (4) but without the signum function (given the continuous action space). Building on their work, we generate two propositions that mirror the two main hypotheses in the previous subsection. Namely, both bias and correlated priors slow asymptotic learning. In the end of this subsection, we discuss why certain features of the modelling approach in GJ preclude us from examining the precise frictions we are interested in.

Let's consider this extended model with bias. Continuous updating and asymptotic learning come with some convenient results from the theory of Markov chains. Importantly, with continuous updating, convergence to the same belief is guaranteed if the social network is strongly connected and aperiodic[33]. The assumptions that GJ impose on the networks[34] ensure there is always convergence to the same action in the network.[35] Since disagreement is precluded, the outcome of interest is the *speed* of learning, captured by the following measure of consensus time *CT*. For a given $\epsilon > 0$ and an interaction matrix $\mathbf{G}$, define:

$$CT(\epsilon, \mathbf{G}) = \sup_{\mathbf{a}} \min\{t : CD(t; \mathbf{G}, \mathbf{a}) < \epsilon\} \tag{20}$$

where $CD(t; \mathbf{G}, \mathbf{a}) = ||\mathbf{T}(\mathbf{G})^t\mathbf{a} - \mathbf{T}(\mathbf{G})^\infty\mathbf{a}||^2_{\mathbf{s}(\mathbf{G})}$ is the consensus distance, which is a weighted sum of

[33] Strongly connected means that the there is a path between any i and j in the network, and aperiodic means that the lowest common divisor of all cycles in the network is 1, which is true if for at least one i, $A_{ii} = 1$). See Golub and Sadler, n.d. for a more detailed discussion, and Meyer, 2023 for technical details.

[34] The focus on networks with four features: sufficient density, no vanishing groups, interior homophily, and comparable densities between groups

[35] Additional assumptions are needed to ensure that the convergence is to the true state of the world: 1) the majority of agents get the correct signal, 2) no subnetworks have outsized influence, 3) no subnetworks ignores most of the network. (Golub and Jackson, 2010)

distances between the belief at time t and the asymptotic belief, with weights $\mathbf{s}(\mathbf{G}) = (\frac{d_1}{D(\mathbf{G})}, ..., \frac{d_n}{D(\mathbf{G})})$, where $d_i(\mathbf{G}) = \sum_{j \in N} G_{ij}$ the weighted degree of agent i and $D(\mathbf{G}) = \sum_i d_i(\mathbf{G})$ as the sum of weighted degrees in the network.

Define an 2×2 matrix of *effective connectivity* $\mathbf{F}(b, p_s, p_d)$ with characteristic entries

$$F_{zq}(b, p_s, p_d) \quad = \quad \begin{cases} \frac{p_s}{p_s + p_d(1-b)}, & \text{if } z = q \\ \frac{p_d(1-b)}{p_s + p_d(1-b)}, & \text{if } z \neq q \end{cases} \tag{21}$$

The matrix $\mathbf{F}$ has the same dimension as the number of groups, and each entry F_{zq} measures the fraction of weighted degrees that agents of group z have with agents of group q. The main contribution of GJ is to show that the second eigenvalue[36] of the matrix $\mathbf{F}$ bounds the convergence time in the DeGroot model, i.e., for any $\epsilon > 0$, there is a large enough n such that:

$$\frac{(1-\epsilon)}{2} \left(\frac{log(n)}{log\left(\frac{1}{|\lambda_2(\mathbf{F})|}\right)} \right) \leq CT(\epsilon; \mathbf{G}) \leq 2(1+\epsilon) \left(\frac{log(n)}{log\left(\frac{1}{|\lambda_2(\mathbf{F})|}\right)} \right) \tag{22}$$

Essentially, the speed of learning is sandwiched by functions of the second eigenvalue of the matrix $\mathbf{F}$. The larger the second eigenvalue, the slower the learning. The proofs for the main results in GJ are technical, and we refer the interested reader to their paper for more details. The main results translate in our case, since the arguments rely on large-n dynamics as the realized links in the network approach the expected links, which is true in the model extension with bias presented here. The following proposition relies on the connection between second eigenvalue of $\mathbf{F}$ and speed of learning.

Proposition 1. Bias slows learning:

$$\frac{\partial \lambda_2(\mathbf{F}(b, p_s, p_d))}{\partial b} > 0 \tag{23}$$

Proof. First we show that the second eigenvalue in this case is identical to:

$$\lambda_2(\mathbf{F}(b, p_s, p_d)) = \frac{p_s - p_d(1-b)}{p_s + p_d(1-b)} \tag{24}$$

[36] If we calculate the eigenvalues associated with the matrix $\mathbf{F}$ and ordered them by absolute value from largest to smallest, the second eigenvalue is the second on that list. The largest eigenvalue of a stochastic matrix is unique and is unity, so the second eigenvalue is always smaller than 1in absolute value.

Note that we can decompose

$$\mathbf{F} = \frac{p_d(1-b)}{p_s + p_d(1-b)}\mathbf{O} + \frac{p_s - p_d(1-b)}{p_s + p_d(1-b)}\mathbf{I} \tag{25}$$

where $\mathbf{O}$ is a 2×2 matrix of ones and $\mathbf{I}$ is the 2×2 identity matrix. Since $\mathbf{O}$ and $\mathbf{I}$ commute (are simultaneously diagonalizable), the eigenvalues of F are the sum of the respective eigenvalues of $\mathbf{O}$ and $\mathbf{I}$. Since $\mathbf{O}$ has rank 1, it has only one non-zero eigenvalue, so the second eigenvalue of this matrix is zero. The second matrix is diagonal, so its eigenvalues are equal to its diagonal entries, so all the eigenvalues are equal to $\frac{p_s - p_d(1-b)}{p_s + p_d(1-b)}$. Adding the two vectors of eigenvalues, we get that the second largest eigenvalue of $\mathbf{F}$ is $\frac{p_s - p_d w}{p_s + (m-1)p_d w}$.

Taking derivative of this quantity we obtain:

$$\frac{1}{\partial b}\left(\frac{p_s - p_d(1-b)}{p_s + p_d(1-b)}\right) = \frac{p_s p_d}{(p_s + p_d(1-b))^2} > 0 \tag{26}$$

□

This characterization of the eigenvalue some intuitive interpretations. First, the speed of convergence is a function of both segregation and bias. Just like in GJ, the second eigenvalue λ_2 increases in segregation, through either increasing p_s or decreasing p_d, slow convergence; in our extended model, bias also slows convergence.

Importantly, in this extended model, the second eigenvalue is not only a measurement of segregation, as is the case in GJ with $b = 0$. Rather here the second eigenvalue captures someas a measurement of "effective connectivity", that is a function of both segregation and bias. To illustrate the consequences of this re-interpretation, consider a scenario in which a social planner increases cross-group linkages. In the initial segregated network the parameters describing society are p_s, p_d^s, b^s, while after integration the society is characterized by p_s, p_d^i, b^i, such that $p_d^i > p_d^s$. In the model of GJ, $b^i = b^s = 0$, and the integration would always lead to a faster convergence. However, if this integration also sufficiently *increases* bias, such that $b^i > 1 - (1-b^s)\frac{p_d^s}{p_d^i}$, then integration will *slow* convergence.

We can also analyze how the initial distribution of signals impacts the time to consensus. GJ develop a "representative agent theorem", which shows that asymptotically we can treat each group in the network

as one individual. Intuitively, as $n \to \infty$, the amount of information each agent receives grows large, and convergence within type happens in only 1 period. From the first period on, the speed of learning in the asymptotic case depends on how different groups are connected and what their priors are. Recall that s_1, s_2 represent the average precision for each group, and let $\overline{a}_k(t)$ represent the average action of group k at time t. Then, we can employ the representative agent theorem to show that the disagreement between the two groups at time t is:

Proposition 2. (Golub and Jackson, 2012) The disagreement between the groups at time t is proportional to the second eigenvalue and the difference in priors. Specifically:

$$\overline{a}_1(t) - \overline{a}_2(t) = \lambda_2^t(s_1 - s_2) \tag{27}$$

Proof. This proof uses the "representative agent theorem", that implies that we can asymptotically treat each type of agent as a single individual. Essentially, this means focusing on the expected links to agents of each type. We will first show that:

$$\overline{a}_i(t) = \frac{1}{2}(1 + \lambda_2^t)s_i + \frac{1}{2}(1 - \lambda_2^t)s_{-i} \tag{28}$$

First, note that $\overline{a}_i(0) = s_i$, and that $\frac{1}{2}(1 + \lambda_2) = \frac{p_s}{p_s + p_d(1-b)}$ and $\frac{1}{2}(1 - \lambda_2) = \frac{p_d(1-b)}{p_s + p_d(1-b)}$, which is straightforward to check. The proof proceeds by induction. For $t = 1$:

$$\begin{aligned}
\overline{a}_i(1) &= \frac{p_s}{p_s + p_d(1-b)}\overline{a}_i(0) + \frac{p_d(1-b)}{p_s + p_d(1-b)}\overline{a}_{-i}(0) && (29)\\
&= \frac{1}{2}(1 + \lambda_2)s_i + \frac{1}{2}(1 - \lambda_2)s_{-i} && (30)
\end{aligned}$$

Now, assume this is true for $\overline{a}_i(t)$. Then for $t + 1$ we have:

$$\begin{aligned}
\overline{a}_i(t+1) &= \frac{p_s}{p_s + p_d(1-b)}\overline{a}_i(t) + \frac{p_d(1-b)}{p_s + p_d(1-b)}\overline{a}_{-i}(t) && (31)\\
&= \frac{1}{2}(1 + \lambda_2)\left(\frac{1}{2}(1 + \lambda_2^{t-1})s_i + \frac{1}{2}(1 - \lambda_2^{t-1})s_{-i}\right) && (32)\\
&+ \frac{1}{2}(1 - \lambda_2)\left(\frac{1}{2}(1 + \lambda_2^{t-1})s_{-i} + \frac{1}{2}(1 - \lambda_2^{t-1})s_i\right) && (33)\\
&= \frac{1}{2}(1 + \lambda_2^t)s_i + \frac{1}{2}(1 - \lambda_2^t)s_{-i} && (34)
\end{aligned}$$

Which proves the claim. Calculating the difference yields:

$$\begin{aligned} \overline{a}_1(t) - \overline{a}_2(t) &= \left(\frac{1}{2}(1+\lambda_2^t)s_1 + \frac{1}{2}(1-\lambda_2^t)s_2\right) - \left(\frac{1}{2}(1+\lambda_2^t)s_2 + \frac{1}{2}(1-\lambda_2^t)s_1\right) \quad (35) \\ &= \lambda_2^t(s_1 - s_2) \quad (36) \end{aligned}$$

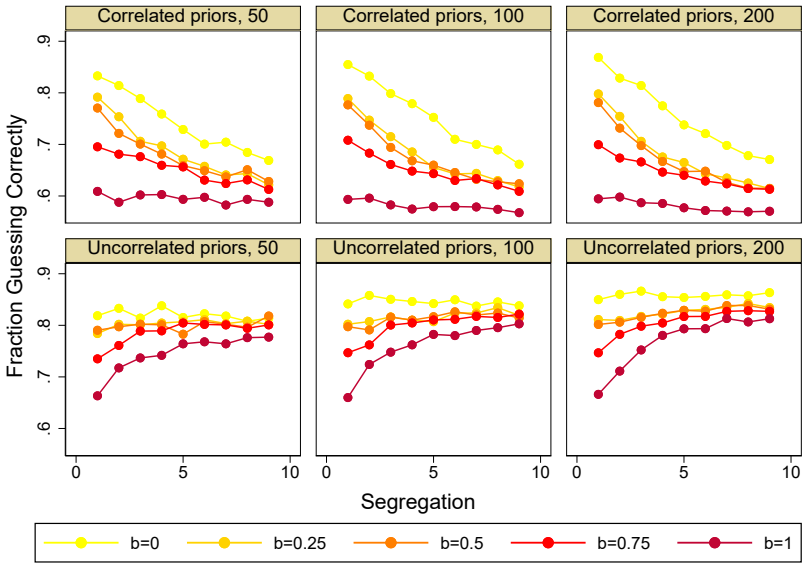
□

There are two main implication of this proposition. First, it shows that an increase in correlation of priors ($|s_1 - s_2|$) slows learning. While the disagreement still tends to zero as t increases[37], it can persist for longer as $s_1 - s_2$ grows. Second, it shows that if the groups have exactly the same priors $s_1 = s_2$, segregation does not matter asymptotically, since convergence happens in one period. Intuitively, if each group has the same prior, as number of agents grows to infinity, agents can learn just as well from their own group as they would from the other group.

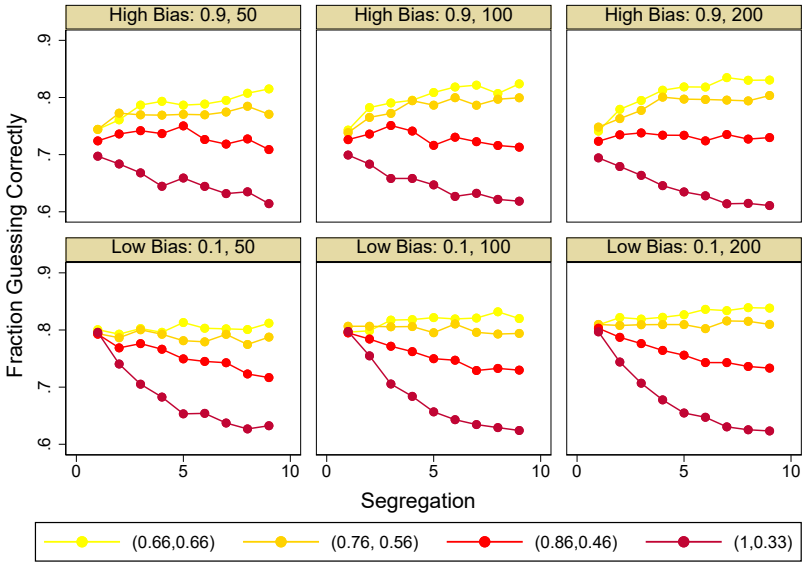
Proposition 1 and 2 show how bias and correlated priors each slow learning in the case of asymptotic learning with continuous updating. However, this approach cannot generate disagreements (failure of convergence), or the *interactions* between segregation, bias, and correlated priors that we generate in the previous subsection. That is, under this framework, if priors are not uncorrelated, segregation always harms the speed of learning. In the previous subsection we showed that segregation can *increase* learning, even for small bias. The asymptotic model cannot generate these frictions since the number of links each individual has grows infinite. With this abundance of information, there are no trade-offs in how links are distributed across types. These trade offs are key in generating the patters from our simulations.

[37] Since $|\lambda_2| < 1$ in any stochastic matrix

A.2 Additional simulations

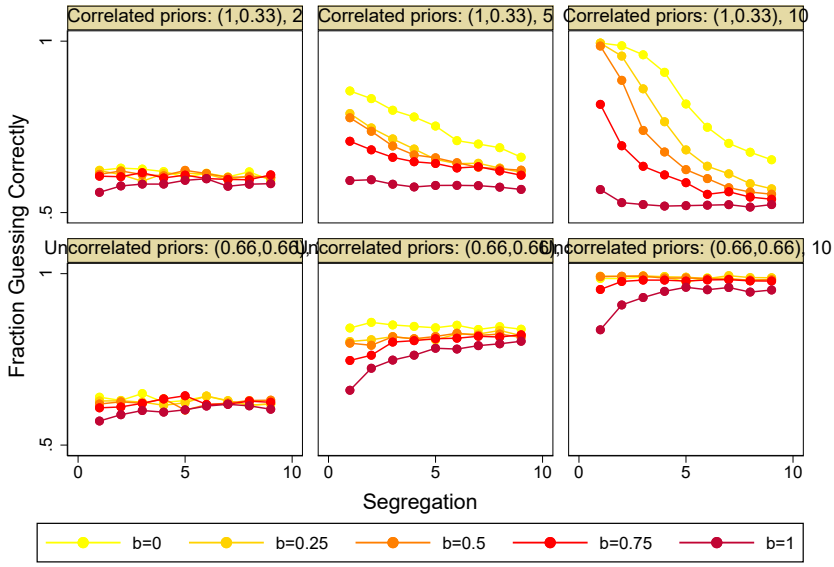


(a) Panel A: Segregated learning with bias

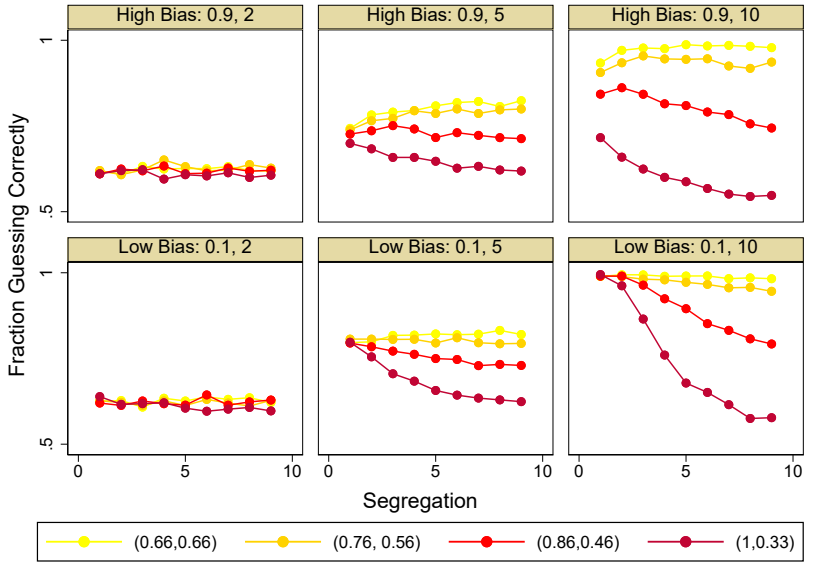


(b) Panel B: Segregated learning with correlated priors

This figure plots the results from the theoretical simulations, over different values of population size n. On the x-axis we plot the extent of segregation, p_s/p_d. On the y-axis we plot the fraction of agents who guess the correct state of the world, which is estimated using Algorithm 1. Both panels present the theoretical results for $n = \{50, 100, 200\}$ (recall that in the paper simulations $n = 100$). Panel A plots segregated learning over bias. Each different line represents the relationship between segregation and learning for different values of bias (darker shades denote higher bias). Above, we plot these relationships for correlated priors ($s_1 = 1, s_2 = 0.33$), and below for uncorrelated priors ($s_1 = s_2 = 0.66$). Panel B plots segregated learning over correlation of the priors. Each different line represents the relationship between segregation and learning for different values of correlation (darker shades denote higher correlation). Above, we plot these relationships for high bias ($b = 0.9$), and below for low bias ($b = 0.1$).



(a) Panel A: Segregated learning with bias



(b) Panel B: Segregated learning with correlated priors

This figure plots the results from the theoretical simulations, over different values of degree d (number of expected links per individual). On the x-axis we plot the extent of segregation, p_s/p_d. On the y-axis we plot the fraction of agents who guess the correct state of the world, which is estimated using Algorithm 1. Both panels present the theoretical results for $d = \{2, 5, 10\}$ (recall that in the paper simulations $d = 5$). Panel A plots segregated learning over bias. Each different line represents the relationship between segregation and learning for different values of bias (darker shades denote higher bias). Above, we plot these relationships for correlated priors ($s_1 = 1, s_2 = 0.33$), and below for uncorrelated priors ($s_1 = s_2 = 0.66$). Panel B plots segregated learning over correlation of the priors. Each different line represents the relationship between segregation and learning for different values of correlation (darker shades denote higher correlation). Above, we plot these relationships for high bias ($b = 0.9$), and below for low bias ($b = 0.1$).