Small-Group Interactions Reduce Homophily by Muting Directed Search*

Andrea Galeotti
London Business School
Reem Qamar
London Business School
Adam Szeidl
Central European University and CEPR

July 19, 2026

Abstract

We study how the quasi-random assignment of MBA students to different peer groups shapes their homophily by gender and ethnicity. We show that students form substantially more cross-trait friendships within small groups than in the broader MBA cohort. We then decompose homophily into the contributions of peer-group composition, preferences, and directed search. Directed search, in which students approach peers with particular characteristics, can substantially amplify even weak preferences for same-trait friends. Survey evidence indicates only weak same-trait preferences but strong same-trait bias in approach decisions. Consistent with this pattern, structural estimates imply that directed search accounts for much of overall homophily. This result helps explain why homophily is lower in small groups, where directed search is muted, and suggests that homophily can be reduced through peer-group design.

*Emails: agaleotti@london.edu, rqamar@london.edu, szeidla@ceu.edu. We thank the London Business School MBA program team, Career Centre, and Student Association for sharing data. We thank Randal Peterson for sharing personality traits data and Ee Hwee Lau for anonymizing the data. We thank Matthew Jackson, Ben Golub, Jean Pierre Benoit, Sanjeev Goyal, Eduard Talamas, Paolo Surico, Elias Papaioannou, and seminar and conference participants for useful comments. Galeotti gratefully acknowledges funding for the project titled "Social Structure, Homophily and Outcomes" from the European Research Council Advanced Grant, financed by the UK Research and Innovation Council, no. EP/Z534079/1. Szeidl is grateful for funding from the Austrian Science Fund (FWF) under grant F100600.

1 Introduction

Homophily—the tendency of people to form ties with others like themselves, particularly along dimensions such as gender and ethnicity—is a fundamental feature of social networks.[1] Because social networks shape access to information, collaboration, and economic opportunities, homophily has important economic consequences. By limiting interactions across groups, it restricts the flow of information and opportunities, contributing to differences in beliefs and inequalities in employment, income, and social mobility.[2] Homophily can arise from two distinct sources: people may prefer similar friends, and people may be more likely to meet others similar to them (Currarini et al., 2009). Distinguishing between these sources is important because institutional design can affect meeting opportunities. However, little is known about what determines whom people meet, and therefore about how institutions can promote integration in social networks.

A potentially important determinant of meeting opportunities is individuals' own search behavior. In large social groups, people influence whom they meet through their choices about whom to approach. This behavior, which we refer to as *directed search*, can amplify even modest preferences for same-trait friends. When people base approach decisions on easily observable traits, such as gender or ethnicity, small differences in tastes may translate into disproportionately many meetings with same-trait peers. As a result, modest preferences for same-trait friendships may coexist with substantial homophily. Because this homophily reflects search behavior rather than preferences alone, it may be malleable to peer-group design. Specifically, in small groups, repeated interaction naturally generates meetings among all group members, limiting the scope for directed search. Small-group interactions may therefore reduce homophily, making peer-group design a potential tool for promoting integration.

In this paper we ask whether small-group interactions reduce homophily by muting directed search. We exploit the quasi-random assignment of MBA students at London Business School to several overlapping peer groups. Students are assigned to streams of about 80 students, study

[1]Homophily was introduced in the classic contribution of Lazarsfeld and Merton (1954) and has been shown to be important in a wide range of settings (McPherson et al., 2001; Jackson, 2009; Pin and Rogers, 2016).

[2]See Golub and Jackson (2012) for beliefs, Montgomery (1991), Calvó-Armengol and Jackson (2004) and Ioannides and Loury (2004) for employment and income, Chetty et al. (2022) for social mobility, and Katz (2026) for missing markets for opportunities.

groups of 4–6 students within each stream, and classroom seating arrangements that generate adjacency between students. Because the same students are exposed to large, small, and spatially proximate peer environments, the setting allows us to compare how these different environments shape friendship formation. The assignment algorithms balance demographic and background characteristics, so exposure to particular peers is only quasi-random. Following Borusyak and Hull (2023), we recover the random component of peer-group exposure from the assignment algorithms. This allows us to estimate how joint participation in different peer environments causally affects friendship formation between same-type and different-type students.

Using network surveys covering roughly 1,400 students in three MBA cohorts, we estimate these causal effects. We find that the friendship networks generated by exposure to smaller, more structured peer groups exhibit substantially lower homophily by gender and ethnicity. Homophily is high among friendships generated by stream exposure and among friendships formed across streams. It is much lower among friendships generated by study-group exposure and classroom adjacency. For example, friendship formation is about 130% more likely among same-gender than among different-gender pairs in the stream, but only about 50% more likely in the study group. For ethnicity, the corresponding premia are about 200% in the stream and 40% in the study group. Thus, interactions in smaller groups generate substantially lower homophily.

We next ask whether the lower homophily generated within study groups also reduces homophily in students' broader friendship networks. Students could, in principle, offset more diverse study-group friendships by forming additional same-trait friendships elsewhere. We cannot estimate the overall effect of study-group assignment because all students are assigned to a study group. We can, however, test for such compensating behavior using quasi-random variation in the share of same-trait peers within study groups. We find that students assigned to more mixed study groups exhibit significantly lower overall homophily. Thus, the cross-trait friendships generated within small groups are not fully offset by additional same-trait friendships elsewhere in the network.

We emphasize that our estimates do not identify the effect of group size in isolation. The randomized variation comes from which peers students encounter within groups of a given size, not from variation in group size itself. Moreover, the peer environments in our setting differ along several dimensions. Study groups and classroom adjacency are not only smaller than streams but also generate more frequent, structured, and harder-to-avoid interactions. We therefore interpret our estimates as identifying the effects of exposure to different peer environments, of which group size is one important dimension but not the only one.

To understand why exposure to these smaller groups reduces homophily, we decompose homophily into three channels and use several approaches to understand their contributions. The first channel is *peer-group composition*: students may encounter same-trait peers more frequently because they are assigned to, or choose, environments with many similar peers. The second is *preferences*: conditional on meeting, students may be more likely to form friendships with same-trait peers. The third is *directed search*: within a peer environment, students may selectively approach same-trait peers. Distinguishing between these channels is important because they imply different approaches to promoting integration.

We next present reduced-form evidence on the three channels. First, we ask whether students sort into peer environments with similar peers. We measure this channel using registrations for activities outside the assigned stream: Student Association club events, Career Centre events, and tailored courses. These activities are important venues for MBA social interaction. If peer-group composition were an important source of homophily, same-gender or same-ethnicity pairs should be more likely to choose the same activities. We find little evidence of such sorting. We then turn to preferences and directed search. To measure them, we use a survey module in which respondents evaluate hypothetical peer profiles that vary in gender, region, and extracurricular activities.[3] Respondents report both their willingness to approach the peer and their desire to form a friendship. We find that students express only a slight premium in friendship preferences for same-trait profiles, but a much larger premium in their decision to approach a same-trait profile. These findings suggest that homophily reflects substantial directed search but only modest same-trait preferences.[4]

To quantify the contributions of the three channels, we structurally estimate a model of network

[3]We use region rather than ethnicity for ethical reasons, but in our data region and ethnicity are highly correlated and exhibit very similar homophily patterns.

[4]Ingram and Morris (2007) provide direct evidence on selective interaction using electronic name tags to track face-to-face interactions at a business mixer.

formation that incorporates peer-group composition, preferences, and directed search. Given the institutional structure of study groups, we assume that all members of a study group meet one another. This assumption allows differences in homophily between streams and study groups to identify the contribution of directed search. Separately, the survey-based measures of friendship preferences and approach decisions provide an independent source of identifying variation. Whether we identify the model using variation in homophily between streams and study groups or using survey moments on friendship preferences and approach decisions, the qualitative results are similar: directed search consistently emerges as a major driver of homophily. In our preferred specification, directed search accounts for 45% of homophily by gender and 80% of homophily by ethnicity within streams. These findings indicate that the lower homophily observed in small groups arises primarily because small-group interactions mute directed search.

A remaining question is whether the lower homophily observed in small groups reflects a temporary reduction in the cost of maintaining different-trait friendships, rather than reduced directed search. If this were the main force, different-trait friendships formed in study groups should fade after study groups dissolve, causing the difference between stream and study-group homophily to shrink among lasting friendships. Instead, both the reduction in homophily and the estimated contribution of directed search remain similar when we restrict attention to lasting friendships. Moreover, this alternative model predicts substantially less directed search and substantially stronger same-trait preferences than observed in our survey. These patterns favor the directed-search interpretation over the view that small groups reduce homophily simply by creating temporary low-cost connections.

The paper contributes to the literature on the sources of homophily. Existing work shows that homophily can arise from both preferences and meeting opportunities (Currarini et al., 2009, 2010). It also shows that the structure of interaction opportunities shapes patterns of homophily and integration (Kossinets and Watts, 2006; Boucher et al., 2020), and that homophily persists as social networks evolve (Jackson et al., 2022). We contribute to this literature by decomposing meeting opportunities into peer-group composition and directed search and showing that directed search is a major source of homophily in our setting. We further show that small-group interactions reduce homophily, plausibly because they mute directed search. Together, these findings suggest that institutions can foster more integrated social networks by organizing repeated interactions in small, mixed peer groups.

We also contribute to research studying the effects of peer-group assignment on social networks. Prior work shows that random peer-group assignment affects friendship formation, network position, and cross-group relationships (Marmaros and Sacerdote, 2006; Boisjoly et al., 2006; Carrell et al., 2013; Hasan and Bagde, 2015; Carrell et al., 2019; Lowe, 2021; Chetty et al., 2022; Corno et al., 2022). Our contribution is to show that small-group interactions causally reduce homophily and that they do so by muting directed search, highlighting a potential approach to creating more integrated social networks.

The rest of the paper proceeds as follows. Section 2 describes the institutional setting, the data, and the empirical design. Section 3 presents the causal effects of exposure on friendship formation and homophily. Section 4 studies mechanisms and provides the quantitative decomposition.

2 Context, Data, and Design

2.1 Institutional Context

The MBA program at London Business School (LBS) is a two-year comprehensive degree program targeted toward mid-career professionals. LBS assigns students to three main types of peer groups. First, all incoming students are partitioned into five to six *streams*, comprising 75–86 students each. Students take all twelve mandatory "core" courses within their stream, including all courses in their first six months and a majority of courses in the next six months.[5] The remaining courses in the second six months of the first year, and all courses in the second year, are "electives" chosen by the students.[6] Second, LBS partitions the students in each stream into *study groups* of 4–6 students. Study groups jointly work on group assignments and collaborative projects in four out of twelve core courses. Students remain in the same study group throughout their core courses. Third, LBS

[5] In the first half of the first year, students attend nine core courses for a total of 51 sessions. In the second half, they attend three core courses for a total of 25 sessions. Each session lasts 2.45 hours.

[6] Electives are generally across MBA cohorts and across degree programs.

assigns students to *seating charts* for the lecture theaters where the core courses take place. The seating charts are changed three times during the first year: the first seating chart is implemented as soon as the students start their MBA program in August for 21 sessions, the second chart begins in mid-September for 40 sessions, and finally, the third chart is implemented in January of the following year for 25 sessions. We define adjacency in the seating charts as the third type of peer group.

These three nested assignments generate three distinct margins of exposure that differ sharply in group size and interaction intensity. Furthermore, the assignment of students to streams, study groups, and seating charts is quasi-random to ensure that a given set of student characteristics is evenly balanced across streams and study groups. These characteristics include gender, nationality, prior industry, GMAT score, age, work experience, native English status, and other background characteristics. The seating charts are designed to minimize instances of people of the same gender, nationality, or study group being seated next to each other. Appendix A.2 details the exact process of each algorithm.

Throughout their MBA journey, students also participate in events organized by the Student Association and the LBS Career Centre.

2.2 Data Description

For this paper, we use data from the MBA 2025, MBA 2026, and MBA 2027 cohorts,[7] with an average class size of 467 students. We obtained data from multiple sources; the data were merged based on students' LBS-assigned email addresses by an independent research assistant working for LBS. The data shared with us contain only anonymous student identifiers. Our main datasets are as follows: administrative data provide predetermined characteristics, surveys provide network outcomes and additional predetermined characteristics, and club/career data and elective enrollment data provide observed non-core-classroom meeting opportunities.

Administrative data. Fixed and predetermined characteristics of each student are collected

[7]The cohort label corresponds to the graduation year. The MBA 2025 cohort began the program in August 2023 and graduated in July 2025; analogously, the MBA 2026 and MBA 2027 cohorts began in August 2024 and August 2025 and graduate in July 2026 and July 2027, respectively.

directly from the LBS School Database (SchoolDB). These include basic demographic details, such as gender, nationality, native English status, and age. Further, SchoolDB also stores detailed information on students' professional and academic backgrounds.

Surveys. For each cohort, the survey schedule includes four surveys during the MBA journey. The first survey of a cohort takes place in August, right after the MBA orientation, which marks the commencement of the program for the new cohort. The second survey is in December of the first year, before the winter break. The third survey takes place at the end of the first year, in June, before classes break for the summer. The fourth survey is administered in June at the end of the second year. We have collected all four surveys for the MBA 2025 cohort, the first three surveys for the MBA 2026 cohort, and the first two surveys for the MBA 2027 cohort.[8]

The surveys were conducted online: all students in the cohort received an email invitation and a few reminders to complete the surveys. To maximize participation, we offered a £25 incentive for an estimated 15-minute completion time. The average response rate across all surveys was nearly 87 percent, with an average median completion time of roughly 17 minutes.

The primary role of the surveys was to collect information on students' networks within their MBA cohort. We asked students to nominate others in their cohort with whom they have connections of different types:

- Career network (*With whom do you most frequently discuss career matters?*),
- Social network (*With whom do you socialise the most (eating out, concerts, theatre, etc.)?*),
- Personal help network (*Who do you turn to talk about personal problems or when you feel under pressure?*),
- Close friendship network (*Who are your closest friends?*).

Respondents could choose up to 6 other students from their cohort for each connection type. To choose a student, they had to start typing the name; after entering a few characters, the interface offered matching names in a drop-down menu. Only students in the respondent's cohort appeared

[8]Online Appendix C reports the survey questions.

as options and could be selected. Our sample has an average of just under 4 nominations per respondent per question and an average reciprocation rate of nearly 40 percent.[9]

The survey also collected information on students' income before the MBA and their socioeconomic background (SEB). To measure socioeconomic background, students were asked to imagine a ladder representing where people stand within their MBA cohort, with the top of the ladder representing the best off and the bottom representing the worst off in terms of money, education, and jobs. They were then asked to indicate where they think they stand on the ladder relative to their MBA cohort. While an individual's socioeconomic background is a complex trait, the response to this question provides insight into how students view themselves relative to their peers.

Clubs. The Student Association at LBS has numerous student-run clubs that provide students with the opportunity and space to connect over mutual interests. These clubs organize events throughout the year, which students can attend upon registering and paying the registration fee (if any). Clubs are classified as either "social", "sports", "regional", or "career". We obtained from the Student Association the list of clubs and the list of registered students for each club event. Overall, we have data on 111 student-run clubs and 1,820 distinct events from May 2024 to May 2026; the data include approximately 40,619 registrations from students in MBA 2025, MBA 2026, and MBA 2027, and nearly 99 percent of students registered for at least one event.

In addition to student-run clubs, the Career Centre at LBS hosts events to help students achieve their career goals. We obtained the registration and attendance data from all these events from the Career Centre. The Career Centre hosted 291 distinct events in the calendar year 2025, with nearly 84 percent of the students attending at least one event.[10]

Core sample. The core sample consists of all students in SchoolDB, with a total of 1,400 individuals. A large part of our empirical analysis is done at the pair level. For this, we create a dyadic dataset within each cohort, giving us a total of $\sum_c n_c(n_c - 1)/2 = 328,671$ (unordered) pairs for $c = \{2025, 2026, 2027\}$. For each survey s, we consider the subset of pairs (i, j) such that both i and j have responded to survey s. All surveys are then pooled together to give us

[9] These nomination and reciprocation rates are in line with comparable name-generator network surveys, including the Caltech study in Jackson et al. (2022), the Add Health friendship data studied by Currarini et al. (2009, 2010), and the friendship networks analyzed by Ball and Newman (2012).

[10] Data from the Career Centre are collected at the end of each calendar year.

$\sum_{c,s} n_{cs}(n_{cs} - 1)/2 = 680,898$ (unordered) pairs for all cohorts and surveys.

2.3 Summary Statistics

Table 1: Peer Group Characteristics

	Number	Size		
		Average	Min	Max
Cohort	3	466.7	411	514
Stream	17	82.4	75	86
Study group	249	5.6	4	6

Notes. Entries report the number of cohorts, streams, and study groups in the core sample, together with the mean, minimum, and maximum number of students in each type of peer group.

Table 1 provides summary statistics on key peer groups in our data. Our core sample includes three cohorts, 17 streams, and 249 study groups. All streams and all study groups are fairly similar in size, with about 82 students per stream and about 5.6 students per study group.

Table 2 reports summary statistics for the students in our sample. Alongside overall means, we report the minimum and maximum across the 17 streams. Students enter the MBA program at about age 29 and are diverse in background: about 56% are male, about 40% are classified as having native English status, and the cohort is ethnically diverse.[11] Students have, on average, just under 5.5 years of pre-MBA work experience, and over half previously worked in consulting or finance. The small range across streams shows that streams are similar along these dimensions.

Table 3 reports summary statistics on the network. Throughout the paper, we focus on the "OR" network, in which a pair of students (i, j) in survey s is connected if either i nominates j or j nominates i in at least one of the network questions in s. When no confusion arises, we refer to these connections simply as friendships. The table shows that the average student is connected to about 10 other students. Roughly half of these ties are outside the student's stream, while the

[11] There are two sources of information about a student's ethnicity: the survey and SchoolDB. We ask students to report their racial/ethnic affiliation in the first survey of each cohort, and SchoolDB collects this information at the time of admission to the program. While both collect the same information, our surveys use a broader classification structure than SchoolDB. For students who responded to the first survey (nearly 96 percent), we confirm that the ethnicity reported in the survey matched closely with that in SchoolDB. For the remaining students, we align the ethnicity stored in SchoolDB with our survey classification structure.

Table 2: Student Characteristics

	Average	Min	Max
A. Demographics			
Age (years)	28.7	28.5	29
Male (%)	56.2	54.3	61.3
Native English status (%)	40.3	34.9	45.7
B. Ethnicity (%)			
Black	3	1.2	6.2
East/SE Asian	17.7	13.6	21.3
Hispanic	12.9	8.6	17.6
Middle Eastern	7.7	4	13.1
South Asian	19.2	14.1	26.2
White	30.8	20	42
Others	8.6	1.2	25.9
C. Prior work experience			
Years worked	5.5	5.3	5.7
Consulting (%)	33.1	27.2	37.2
Finance (%)	22.7	19	27.9

Notes. All reported demographics are averages, minimums, and maximums across the 17 streams in our sample. Age and work experience are as reported at the beginning of the MBA program. Native English status is based on whether a student reports English as their first language.

Table 3: Network Characteristics

	Overall	Outside Stream	Stream not Group	Study Group
Average number of links	10.1	5.1	4	1
Share same gender (%)	66.7	69.1	68.8	52.5
Share same ethnicity (%)	46.7	56.3	43.9	16.9

Notes. The first row reports the average number of links per student in different peer groups. The second and third rows report the percentages of those links that are with a student of the same gender and of the same ethnicity, respectively.

other half are within the stream. Among the approximately five within-stream connections, about four are outside the student's study group and about one is within the study group. These patterns already suggest an important role for peer exposure: streams and study groups account for only about 18% and 1% of the cohort, respectively, yet they account for a much larger share of observed friendships.

The table also reports the composition of friendships by gender and ethnicity. Among the roughly 10 friendships of the average student, about 66% are with someone of the same gender and about 46% are with someone of the same ethnicity. These shares are substantially above the corresponding cohort shares—about 50% for gender and about 20% for ethnicity—indicating substantial inbreeding homophily. Most notably, the share of same-trait friendships is highest outside the stream, lower within the stream, and lowest in the study group. For example, the share of same-gender friendships declines from 69% outside the stream to 52% within the study group. Thus, even at the descriptive level, smaller peer groups appear to generate a disproportionate number of cross-trait ties.

These descriptive patterns should not, however, be given a causal interpretation. Streams and study groups differ not only in size and interaction intensity, but also in how students are assigned to them. In principle, the assignment procedure could place students together who are especially good matches along dimensions other than gender or ethnicity, thereby generating both higher friendship rates and lower same-trait shares in smaller groups. To distinguish this explanation from a causal effect of peer-group exposure itself, we develop an empirical strategy that isolates the random component of the assignment process.

2.4 Empirical Strategy

Our objective is to estimate the effect of peer-group exposure—stream, study group, and seating adjacency—on friendship formation. The challenge is that these exposures are not generated by a purely random assignment. As discussed above, the program office deliberately balances demographic and background characteristics when constructing streams, study groups, and seating charts. Conditional on these balancing constraints, however, assignment is quasi-random. We exploit this quasi-randomness by building on the recentered IV approach of Borusyak and Hull (2023), adapted to the sequential structure of the LBS assignment process.

Let the sample consist of all unordered pairs of students (i, j) in cohort c observed in survey wave s. Let Y_{ijcs} denote an indicator equal to one if students i and j are connected in wave s. We consider three exposure measures. First, S_{ijc} equals one if the pair is assigned to the same stream. Second, G_{ijc} equals one if the pair is assigned to the same study group. Third, A_{ijcs} equals one if the pair has ever been seated adjacently by survey wave s.[12]

Basic idea. Consider first the effect of being assigned to the same stream on friendship formation:

$$Y_{ijcs} = \beta_S \cdot S_{ijc} + \varepsilon_{ijcs}. \tag{1}$$

The challenge with estimating equation (1) in OLS is that, because S_{ijc} is not fully randomly assigned, the identifying condition

$$\mathbb{E}[S_{ijc}\varepsilon_{ijcs}] = 0$$

need not hold. To address this challenge, following Borusyak and Hull (2023), we isolate the systematic component of stream assignment as the conditional expectation of S_{ijc} over the realizations of the assignment process. We denote this object by

$$\mathbb{E}[S_{ijc} \mid w^S],$$

where the information set w^S contains all the predetermined information used by the assignment rule. We measure this object by drawing counterfactual stream assignments from the original assignment algorithm, holding fixed the predetermined characteristics used by the algorithm, and averaging the resulting realizations of S_{ijc}. This conditional expectation captures the on-average bias induced by the assignment process. The residual

$$\widetilde{S}_{ijc} := S_{ijc} - \mathbb{E}[S_{ijc} \mid w^S]$$

[12]Since seating charts change over the first year, A_{ijcs} is cumulative: seating chart 2 enters from survey 2 onward, and seating chart 3 enters from survey 3 onward.

therefore isolates the random component of stream exposure.

Under the key identifying condition that, conditional on w^S, the random component of the stream assignment is orthogonal to the error term, we can use $\widetilde{S}_{ijc}$ as an instrument for S_{ijc} in a recentered IV (RCIV) approach. The corresponding first stage is

$$S_{ijc} = \gamma \widetilde{S}_{ijc} + \vartheta_{ijcs}.$$

Nested assignments. Our setting is richer than this basic example because we have three assignments—stream, study group, and adjacency—and they are nested in the sense that each layer is constructed conditional on the previous one. In particular, study groups are assigned within realized streams, while seating charts are assigned after both streams and study groups are realized. In this setting, we wish to estimate

$$Y_{ijcs} = \beta_S \cdot S_{ijc} + \beta_G \cdot G_{ijc} + \beta_A \cdot A_{ijcs} + \varepsilon_{ijcs}. \tag{2}$$

In adopting the Borusyak–Hull approach to this setting, the key issue is how to construct the correction terms for the study-group and adjacency layers. The right approach is to construct these corrections conditional on the realized allocations of the previous layers of assignment. Formally, let w^X denote the information set relevant for assignment $X = S, G, A$. The nested assignment structure implies that the realized stream assignment is part of w^G and w^A, while the realized study-group assignment is part of w^A. We then construct, for each layer $X = S, G, A$, the mean exposure

$$\mathbb{E}[X_{ijc} \mid w^X]$$

and the corresponding recentered exposure

$$\widetilde{X}_{ijc} := X_{ijc} - \mathbb{E}[X_{ijc} \mid w^X].$$

Hence, study-group exposure is adjusted using its mean conditional on the realized stream assignment and the predetermined characteristics relevant for the study-group algorithm; similarly, adjacency exposure is adjusted using its mean conditional on the realized stream and study-group assignments and the predetermined characteristics relevant for the seating algorithm. The intuition is that each downstream layer of assignment is constructed conditional on the previous layer(s), so those earlier realized allocations are part of the information that must be partialed out in order to isolate the random innovation in the downstream exposure.

Formally, let g^X denote the randomization shock used by the algorithm that determines exposure $X \in \{S, G, A\}$. For this baseline specification, the identifying condition is

$$\mathbb{E}[\varepsilon_{ijcs} \mid g^X, w^X] = \mathbb{E}[\varepsilon_{ijcs} \mid w^X], \qquad X \in \{S, G, A\}. \tag{IC}$$

This condition states that, once we condition on the information relevant for the assignment of exposure X, the remaining assignment variation is mean independent of the structural error. Appendix A.1.1 shows that, together with the assignment-mechanism property used to construct the recentered exposure, condition (IC) implies the orthogonality conditions required for RCIV identification using the instruments $\widetilde{X}_{ijc}$.

Appendix A.1.2 also clarifies why conditional counterfactual corrections are needed for downstream exposures. If one were instead to correct study-group or adjacency exposure using counterfactuals that ignore the realized upstream assignments, the resulting recentered variables would generally retain systematic variation correlated with those upstream exposures, and this residual component can violate the required orthogonality conditions.

Heterogeneous effects. Our main interest is not only the causal effect of exposure, but also how these effects vary with traits such as gender and ethnicity. Let T_{ij} be an indicator equal to one if pair (i, j) shares a given trait, such as gender or ethnicity. We are then interested in the model

$$Y_{ijcs} = \beta_S \cdot S_{ijc} + \beta_G \cdot G_{ijc} + \beta_A \cdot A_{ijcs} + \rho \cdot T_{ij} + \omega_S \cdot T_{ij} S_{ijc} + \omega_G \cdot T_{ij} G_{ijc} + \omega_A \cdot T_{ij} A_{ijcs} + \varepsilon_{ijcs}. \tag{3}$$

For this specification, the instruments are

$$(\widetilde{S}_{ijc}, \widetilde{G}_{ijc}, \widetilde{A}_{ijcs}) \quad \text{and} \quad (T_{ij}\widetilde{S}_{ijc}, T_{ij}\widetilde{G}_{ijc}, T_{ij}\widetilde{A}_{ijcs}).$$

If T_{ij} is measurable with respect to w^X, then conditioning on T_{ij} adds no information beyond conditioning on w^X, and condition (IC) is sufficient for identification. Otherwise, Appendix A.1.1 shows that for identification we need to use the strengthened condition

$$\mathbb{E}[\varepsilon_{ijcs} \mid g^X, w^X, T_{ij}] = \mathbb{E}[\varepsilon_{ijcs} \mid w^X, T_{ij}], \qquad X \in \{S, G, A\}. \tag{IC-T}$$

The possibility of heterogeneous effects also points to a potential source of misspecification: the error term may contain heterogeneous effects with respect to characteristics that are either unobserved or not included in the regression. As is well known from Heckman and Vytlacil (1998), instrumental variables need not recover a common treatment effect in the presence of such omitted slope heterogeneity. In our setting, Appendix A.1.3 characterizes the corresponding RCIV estimand.[13] In the simple case with a single exposure treatment, it takes the form of a variance-weighted average treatment effect; in the case with multiple exposure treatments, each coefficient is in general a linear combination of heterogeneous exposure treatment effects because omitted-interaction bias propagates across exposures through the covariance structure of the recentered exposure variables.

In practice, this concern appears limited in our data. When we estimate specification (3) using same-gender and same-ethnicity as trait variables, either separately or jointly in the same regression, we draw very similar quantitative conclusions across specifications.[14] This suggests that omitted-interaction bias from the other observed trait is small in our setting.

Implementation and balance tests. We compute the correction terms for each level of exposure by drawing 200 counterfactual realizations for the assignment corresponding to that level of exposure and taking their mean; we discuss the details of how we conduct counterfactuals in Appendix A.2. In all regressions, we cluster standard errors at the pair level.[15]

We validate our empirical strategy with balance tests. For each pair, we consider five exposure measures: being in the same stream; being in the same study group; and being adjacent in each of the three seating charts assigned to the cohort, denoted by Adjacent in A1, Adjacent in A2,

[13] This derivation is closely related to Heckman and Vytlacil (1998) and to Appendix C.1 of Borusyak and Hull (2023). We discuss this formally in Appendix A.1.3.

[14] See Table A3 and Figure A1 in Appendix A.5.

[15] We experimented with study-group-level clustering, where a cluster is the set of pairs belonging to the same study group, and with dyadic-robust clustering; neither materially affected the results.

and Adjacent in A3. The outcome variables are measures of similarity in the characteristics used to stratify the relevant allocation. For stream and study-group assignment, these are gender, nationality, industry, native English status, age, GMAT score, and work experience. For seating-chart adjacency, they are gender, nationality, and study-group membership. In Appendix Table A1, we report OLS estimates in the top panel and RCIV estimates in the bottom panel, both with cohort controls.

The OLS results confirm that exposure is not randomly assigned: roughly half of the coefficients are statistically significant at conventional levels. By contrast, after applying the RCIV correction, only one coefficient remains significant at the 10 percent level. These results support the interpretation that the recentered exposures isolate the quasi-random component of assignment and can therefore be used to identify the causal effect of peer-group exposure on friendship formation and friendship composition.

3 Effect of Exposure on Friendship Formation and Composition

3.1 Exposure and Friendship Formation

We estimate the effect of exposure on friendship by estimating equation (2) using RCIV in the core sample, pooling all cohorts and survey waves. We measure exposure at three levels: whether a pair of students belongs to the same stream, to the same study group, or is seated adjacently. Because seating charts change over the course of the first year, the adjacency indicator equals one if the two students had ever been seated adjacently by the relevant survey wave.[16]

Column 1 of Table 4 reports the results. The "control mean," that is, the average friendship rate among pairs in different streams, is about 1.6 percent. The regression coefficients capture the incremental effects of being in the same stream, being in the same study group, and being seated adjacently. All three effects are precisely estimated and economically large.[17] Given the quasi-random assignment and our use of the RCIV specification, these estimates imply that exposure

[16] This means that seating chart 2 is added from the second survey onward, and seating chart 3 from the third survey onward.

[17] We do not include individual fixed effects in this specification, but Appendix Table A2 shows that the results are robust to doing so.

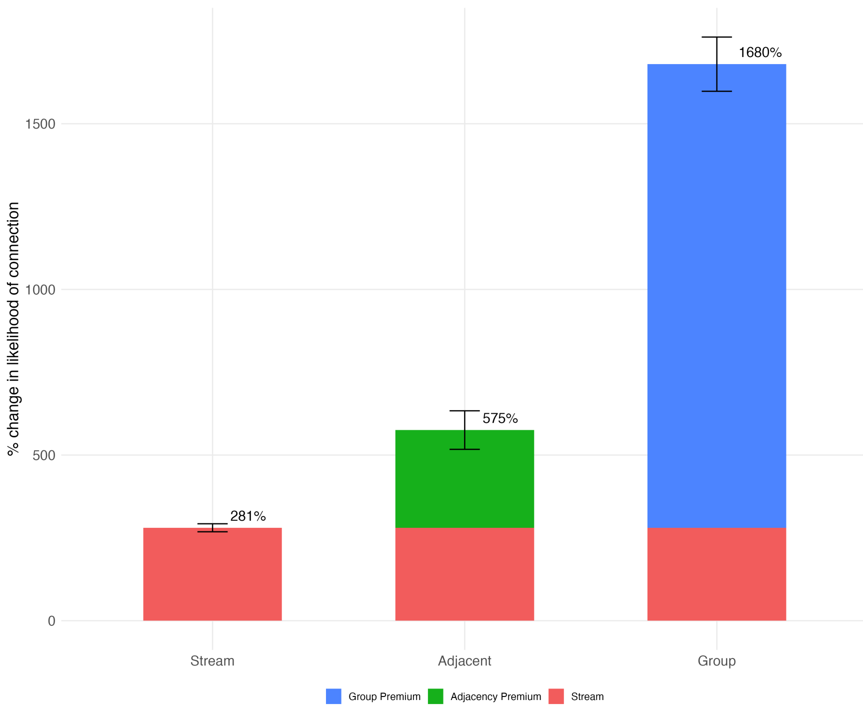
Table 4: Probability of Friendship

	Link			
	(1)	(2)	(3)	(4)
Stream	0.044***	0.026***	0.034***	0.012***
	(0.0010)	(0.001)	(0.0009)	(0.001)
Study group	0.220***	0.192***	0.218***	0.184***
	(0.007)	(0.008)	(0.007)	(0.012)
Adjacent	0.046***	0.047***	0.046***	0.046***
	(0.005)	(0.005)	(0.005)	(0.007)
Stream × Trait		0.036***	0.051***	0.178***
		(0.002)	(0.003)	(0.009)
Study group × Trait		0.070***	0.023	0.313***
		(0.014)	(0.021)	(0.089)
Adj × Trait		0.0001	0.003	0.005
		(0.013)	(0.014)	(0.049)
Trait		0.011***	0.035***	0.115***
		(0.0004)	(0.0008)	(0.002)
Trait	None	Gender	Ethnicity	Similarity
Control Mean	0.016	0.01	0.009	
Observations	680,898	680,898	680,898	680,898
Cluster	Pair	Pair	Pair	Pair

Notes. RCIV estimates of exposure on friendship and friendship composition. Each specification includes survey × cohort fixed effects. Standard errors clustered at the pair level are reported in parentheses. Baseline probability is defined for each specification as follows: (1) probability of friendship outside the stream, (2) probability of friendship for a different-gender pair outside the stream, (3) probability of friendship for a different-ethnicity pair outside the stream. Because similarity is continuous, we do not define a baseline probability for specification (4). $^{*}p < 0.1$;$^{**}p < 0.05$;$^{***}p < 0.01$.

causally increases the probability of a connection, consistent with earlier causal evidence from Marmaros and Sacerdote (2006) and Hasan and Bagde (2015).

Figure 1: Exposure Premium Relative to the Friendship Rate for Pairs in Different Streams



Notes. Percentage change in the likelihood of forming a friendship based on exposure in the peer group, with 95% confidence intervals. The baseline probability is the mean friendship rate between pairs in different streams. Exposure boosts are calculated using specification (1) from Table 4.

Figure 1 translates these coefficients into exposure premia, defined as the proportional increase in friendship probability relative to the baseline rate for pairs of students who are in different streams. Sharing a stream nearly triples the probability of friendship. Adjacency raises it by more than a factor of 5, and being assigned to the same study group raises it by a factor of almost 17. These are large causal effects.[18]

[18] Adjacent and study-group pairs are also same-stream pairs, so the premia for these smaller peer environments are cumulative: they combine the stream effect with the additional effect of adjacency or study-group assignment.

3.2 Exposure and Friendship Composition: Gender and Ethnicity

We next ask whether exposure changes which friendships form. To do so, we estimate our interaction regression (3), allowing the effect of exposure to differ depending on whether the two students share a given trait. Interpreting these regression coefficients is complicated by the fact that both the different-trait friendship rate and the same-trait friendship rate change with different levels of exposure. To help with the interpretation, we introduce the notion of the *homophily premium*, which is the proportional increase in friendship probability for same-trait versus different-trait pairs for a given level of exposure. Formally, for a trait T and exposure category X, let $p_X^T(\text{same})$ and $p_X^T(\text{diff})$ denote the regression-predicted probability of friendship for a same-trait pair and a different-trait pair, respectively. The homophily premium is

$$H\text{Premium}_X^T := \frac{p_X^T(\text{same}) - p_X^T(\text{diff})}{p_X^T(\text{diff})}.$$

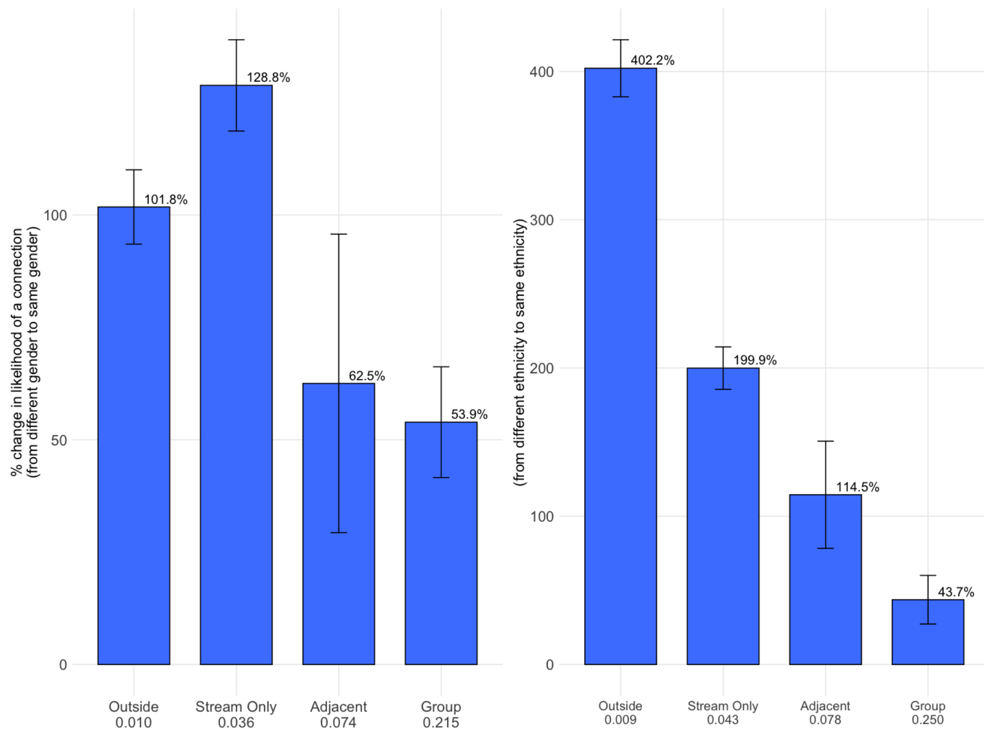
A value of zero corresponds to no homophily.

Consider now column 2 of Table 4, which focuses on gender. The uninteracted trait coefficient is positive and highly significant, implying that among pairs in different streams, same-gender friendships form at a substantially higher rate than different-gender friendships. In fact, across streams, the friendship rate for same-gender pairs is more than twice that for different-gender pairs, implying a cross-stream homophily premium of about 102%. This is plotted as the first bar in Panel A of Figure 2.

How does gender homophily change for different levels of exposure? We answer this question by computing the homophily premium for each level of exposure using the probabilities predicted by column 2 of Table 4. The second bar in Panel A of Figure 2 shows that within streams—but outside study groups and adjacency—the homophily premium is, if anything, slightly higher. By contrast, for adjacent pairs and for pairs in the same study group, the premium is only about 50%, significantly lower than across streams and within streams only. That is, the ties formed in these smaller peer groups are much less gender-homophilous than those formed in the larger groups.

The pattern is even sharper for ethnicity, which is defined using the categories in Panel B of

Figure 2: Homophily Premium



Notes. Panels A and B report the homophily premium by peer group for gender and ethnicity, respectively, with 95% confidence intervals. For each trait and exposure category, the premium is defined as the percentage increase in the probability of friendship for same-trait pairs relative to different-trait pairs. The baseline probabilities (reported under each bar) are the predicted probabilities of friendship between individuals of different genders or ethnicities within each peer group. The underlying probabilities are computed from specifications (2) and (3) of Table 4.

Table 2. In column 3 of Table 4, the Trait variable equals one when the pair has the same ethnicity. The uninteracted Trait coefficient implies very high homophily across streams: Panel B of Figure 2 shows a premium above 400%. Within streams—but outside study groups and adjacency—this declines to about 200%; among adjacent students, to about 100%; and among students in the same study group, to about 40%.[19]

Taken together, these results show that the composition of the network changes sharply with the nature of exposure. Homophily is high outside and within streams, but falls substantially for adjacent and study-group pairs. Small groups do not simply create more friendships; they create

[19]We also estimate equation (3) including both same-gender and same-ethnicity simultaneously. As Table A3 and Figure A1 show, the interaction coefficients and homophily premia are very similar to those reported here.

disproportionately more cross-gender and cross-ethnicity friendships.

3.3 Exposure and Friendship Composition: Similarity Index

Gender and ethnicity are natural focal traits, but they need not be the only dimensions generating homophily in friendship. To capture a broader notion of similarity, we construct a similarity index in four steps. First, we use all baseline characteristics available in the data: gender, ethnicity, geographic region, industry, age, years of work experience, GMAT, self-reported income, and socioeconomic background. Second, for each characteristic we define pairwise similarity. For discrete variables—gender, ethnicity, region, and industry—a pair is similar if both students have the same value. For continuous variables—GMAT, income, age, and work experience—a pair is similar if both students lie on the same side of the cohort median. Third, we regress friendship in the core sample on these pair-level similarities and on all pairwise interactions among them, and use a Lasso procedure to select the relevant predictors (see Table A4 in Appendix A.6).[20] Finally, we regress friendship on the selected subset and use the fitted values as our similarity index. Appendix A.6 reports the details of the construction.

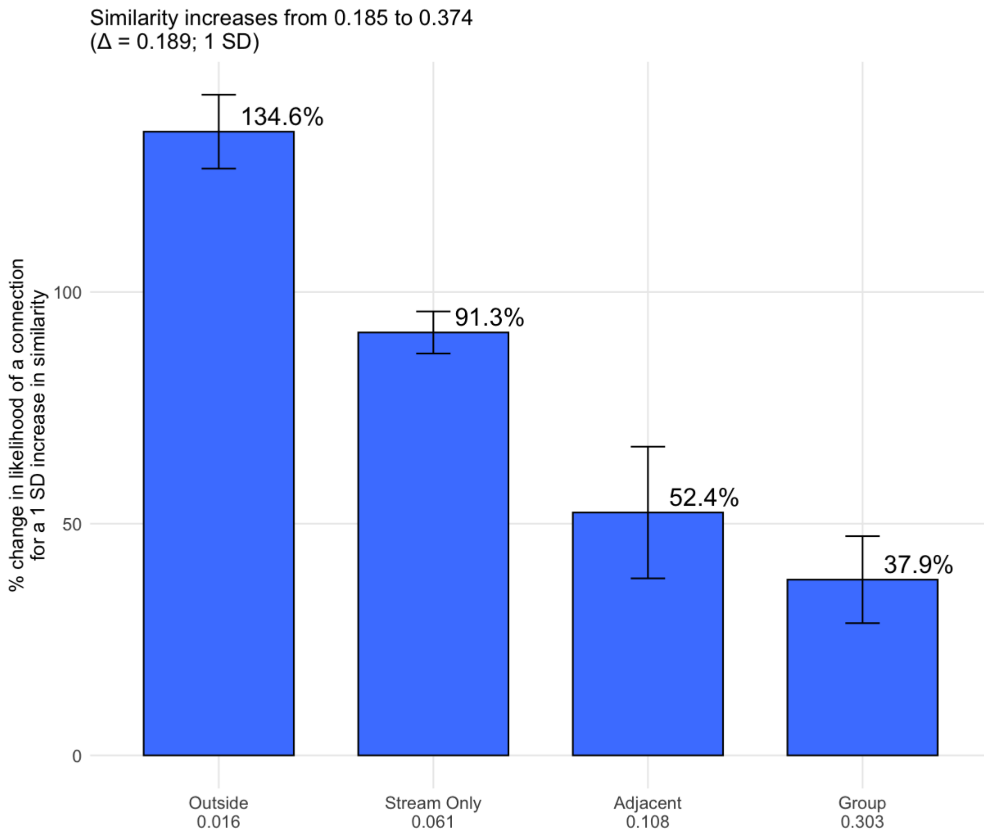
Column 4 of Table 4 and Figure 3 repeat the homophily-premium exercise using this broader similarity index. The qualitative pattern is unchanged. Similarity-based homophily is high outside small groups, remains high within streams, falls sharply for adjacent pairs and, especially, for pairs in the same study group. Thus the reduction in homophily within small groups is not confined to gender and ethnicity alone; it also appears under a broader measure of baseline similarity.

3.4 Inbreeding Homophily

The homophily premium is a convenient way to summarize the regression estimates, but it is not the standard measure used in the homophily literature. We therefore report the traditional index of inbreeding homophily, introduced by Coleman (1958) and used in recent economics work by Currarini et al. (2009). The index compares the realized share of same-trait links with the share

[20]The Lasso selects gender, ethnicity, and region, together with the interactions of gender with ethnicity and region, the interactions of region with industry, age, work experience, and income, and the interactions of ethnicity with GMAT score and income, as the main predictors of assortative friendship.

Figure 3: Similarity Premium



Notes. The figure reports the homophily premium, with 95% confidence intervals, by peer group for similarity. For each trait and exposure category, the premium is defined as the percentage increase in the probability of friendship associated with a one-standard-deviation increase in similarity. The baseline probabilities (reported under each bar) are the predicted friendship probabilities at the mean similarity level in each peer group. The marginal increase in probabilities is computed from specification (4) of Table 4.

that would arise under random matching, and normalizes this difference by the maximum scope for additional same-trait links.

Definitions. To fix ideas, consider a binary trait such as gender, so student i has type $t_i \in \{F, M\}$. Consider the connections among a set L of pairs of students. For example, if we consider homophily in the network of students in the same stream but outside the same study group, then L is the set of pairs of students who belong to the same stream but to different study groups. The number of links in L between students of type t and t' is $K_{tt'} = \sum_{ij \in L} Y_{ij} \mathbf{1}\{t_i = t, t_j = t'\}$.

Same-gender homophily is then the share of same-gender links

$$H_{\text{sg}} = \frac{K_{FF} + K_{MM}}{K_{FF} + K_{MM} + K_{FM} + K_{MF}}.$$

Under random matching, the composition of same-gender friendships would coincide with the share of same-gender pairs in the relevant peer group among all potential pairs; let this benchmark be denoted by R_{sg}.[21] The corresponding inbreeding-homophily index is

$$IH_{\text{sg}} = \frac{H_{\text{sg}} - R_{\text{sg}}}{1 - R_{\text{sg}}}. \tag{4}$$

In the binary case, this aggregate index can be written as a weighted average of female and male inbreeding homophily, with weights proportional to the total number of links that females and males have.

These measures are descriptive, because they do not account for the fact that assignment to streams and study groups is not fully random. As a result, differences in homophily across peer groups could in principle reflect systematic differences in group composition rather than differences in the nature of exposure itself. To address this concern, we also compute inbreeding homophily implied by the causal estimates in Table 4. Recall that $p_X^T(\text{same})$ and $p_X^T(\text{diff})$ are the predicted probabilities that a same-trait pair and a different-trait pair are connected under a given exposure treatment X, respectively.[22] Multiplying these predicted probabilities by the number of potential pairs of each type yields predicted link counts. We then compute predicted inbreeding homophily by replacing the raw link counts in the formulas above with these predicted counts.

Results. Table 5 reports the raw and regression-implied measures of inbreeding homophily.[23] The raw and predicted levels are not identical, but the broad patterns are very similar.[24] For

[21] Formally, $R_{\text{sg}} = \sum_{t \in \{M,F\}} \sum_{ij \in L} \mathbf{1}\{t_i = t_j = t\} / \sum_{tt' \in \{M,F\}^2} \sum_{ij \in L} \mathbf{1}\{t_i = t, t_j = t'\}$.

[22] For example, the predicted probability that a same-gender pair who share a stream, but are neither in the same study group nor adjacent, forms a link is obtained by summing the constant, stream, trait, and stream $\times$ trait coefficients in column 2 of Table 4.

[23] We do not include adjacency in this table, because its definition is changing over time as new seating charts are introduced, complicating the analysis.

[24] Confidence intervals (not reported) show that the raw and predicted measures are sometimes significantly different, but the economic magnitude of the differences is small.

Table 5: Inbreeding Homophily from Raw Data and Regression Estimates

	Inbreeding Homophily							
	Overall		Outside Stream		Stream not group		Study group	
	Raw	Pred	Raw	Pred	Raw	Pred	Raw	Pred
Gender	0.31	0.32	0.34	0.34	0.35	0.37	0.16	0.20
Ethnicity	0.32	0.33	0.44	0.44	0.26	0.28	0.04	0.06

Notes. "Raw" reports inbreeding homophily computed directly from the observed network using expression (4), with the analogous multi-category version for ethnicity. "Pred" reports predicted inbreeding homophily implied by the regression estimates of columns (2) and (3) in Table 4 for gender and ethnicity homophily, respectively. Study-group ethnicity homophily is restricted to study groups with at least one same-ethnicity respondent pair.

gender, inbreeding homophily is high outside the study group, remains high within the stream but outside the study group, and falls sharply within study groups; Appendix Table A6 shows that the same qualitative pattern holds separately for males and females. For ethnicity, inbreeding homophily is highest outside the stream, lower within the stream, and lowest in the study group. These patterns closely mirror those obtained from the homophily-premium analysis.

An advantage of the inbreeding-homophily index is that its magnitude is directly interpretable. For gender, the study-group value is low but still clearly positive, indicating that gender homophily persists even in small-group environments. By contrast, ethnicity inbreeding homophily in the study group is close to zero, suggesting that small groups largely eliminate ethnicity-based homophily.

One concern is that the near-absence of study-group ethnicity homophily could arise mechanically because some study groups contain no same-ethnicity peers. Note that this concern does not apply for the homophily premium results of Figure 2 because they report friendship probabilities conditional on same-ethnicity pairs. To limit its influence on the inbreeding homophily measure, when computing both raw and predicted ethnicity homophily we restricted the study-group analysis to groups that contain at least one pair of same-ethnicity students who both responded to the survey. This leaves 211 of the 249 study groups in the sample. The near-absence of study-group ethnicity homophily is therefore not driven simply by a lack of potential same-ethnicity ties.

3.5 Overall Homophily

Our design identifies the effect of small-group exposure on homophily within those groups. To ask whether this local effect carries over to students' broader networks, we exploit quasi-random variation in the composition of study groups. In particular, it is possible that students compensate for the lower share of same-trait friendships within the study group by forming more same-trait friendships elsewhere. A strong compensation effect of this sort would imply that our estimates do not reflect the impact of small-group exposure on overall homophily.

The cleanest design would be to compare students who do versus do not participate in study groups. Because study groups are mandatory, this design is not feasible in our context. However, variation in the *composition* of study groups allows us to explore whether, holding study-group size fixed, being assigned more same-trait peers in the study group affects a student's overall homophily.

Let OT_i^G denote the number of other students in i's study group who share i's trait, and let Size_i^G denote the size of i's study group. We then estimate

$$IH_{ics}^X = \text{const} + \gamma_1 \cdot OT_i^G + \gamma_2 \cdot \text{Size}_i^G + \varepsilon_{ics}. \tag{5}$$

Here the dependent variable measures student i's inbreeding homophily in peer group X, calculated using equation (4) but using the number of other students of the same and different traits to create an individual-level homophily measure. The coefficient of interest γ_1 measures the extent to which student i's homophily, measured in different layers of peer groups, responds to the number of same-trait peers in i's study group. Because study group sizes differ, we control for the number of students in the group. Since the right-hand-side variables are only quasi-random, we use our study-group-level recentering approach and estimate (5) using RCIV.[25]

Table 6 reports the results. Columns 1–4 focus on gender. Column 1 shows that increasing the number of same-gender peers in the study group by one raises a student's inbreeding homophily within the study group by about 8 percentage points, although the estimate is not statistically significant. We expected this effect to be positive: given our finding that homophily in study

[25]Specifically, we use the conditional study-group counterfactuals to compute $\mathbb{E}[OT_i^G \mid w^G]$ and $\mathbb{E}[\text{Size}_i^G \mid w^G]$, and form the residuals $\widetilde{OT}_i^G = OT_i^G - \mathbb{E}[OT_i^G \mid w^G]$ and $\widetilde{\text{Size}}_i^G = \text{Size}_i^G - \mathbb{E}[\text{Size}_i^G \mid w^G]$, which are our instruments.

Table 6: Effect of Study Group Composition on Homophily

	Inbreeding Homophily							
	Gender				Ethnicity			
	Study Group	Stream not group	Outside Stream	Overall	Study Group	Stream not group	Outside Stream	Overall
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
No. same-trait peers in study group	0.079	0.002	0.007	0.044*	0.068	0.019	-0.008	0.030*
	(0.051)	(0.035)	(0.033)	(0.024)	(0.088)	(0.023)	(0.023)	(0.017)
Study-group size	0.020	0.002	0.004	-0.014	-0.013	-0.019	0.005	-0.018
	(0.048)	(0.032)	(0.029)	(0.022)	(0.052)	(0.021)	(0.022)	(0.018)
Observations	1,978	3,306	3,361	3,486	905	3,288	3,361	3,486
Cluster	Student	Student	Student	Student	Student	Student	Student	Student

Notes. Each specification uses RCIV and regresses individual inbreeding homophily, calculated using (4), on the number of peers in the student's study group who share the student's trait and the number of people in the student's study group. Each specification includes survey × cohort fixed effects. Standard errors, clustered at the individual level, are reported in parentheses. $^{*}p < 0.1$;$^{**}p < 0.05$;$^{***}p < 0.01$.

groups is low, changes in the composition of peers should map into changes in the composition of friends. However, the estimate is imprecise.

Columns 2 and 3 show little evidence of compensation outside the study group. The estimated effects on homophily within the stream but outside the study group, and outside the stream altogether, are an order of magnitude smaller than the study-group effect. Column 4 shows that one additional same-gender peer in the study group increases overall homophily by 4.4 percentage points. Thus, changes in study-group composition affect overall homophily, suggesting that any compensatory response outside the study group is not large enough to undo the direct effect of the study group.

Columns 5–8 show similar patterns for ethnicity. In particular, we again find a positive and statistically significant effect on overall homophily. Finally, the size of the study group has no detectable effect on either gender or ethnicity inbreeding homophily, within the study group or outside it. Study groups range from 4–6 students, and our estimates suggest that variation in group size within this range does not meaningfully affect the process of friendship formation.

In summary, although the within-study-group estimates are imprecise, the evidence is consistent with indirect compensatory effects being relatively weak, suggesting that the "local" homophily-reducing effects of small-group exposure do translate into changes in overall homophily.

3.6 Discussion of Identification, Interpretation, and Robustness

We next discuss three robustness checks and interpretive issues: lasting friendships, less detectable traits, and alternative network definitions.

Lasting friendships. A possible interpretation of our results is that the small-group effects reflect temporary friendships. Since peers in the study group are easily accessible, students may befriend different-trait peers whom they otherwise would not. This logic could explain lower homophily in the study groups, while suggesting that those friendships are less important and will likely fade after study groups are dissolved. This interpretation also suggests that our results should weaken or disappear for friendships that last beyond the dissolution of the study groups at the end of the first year. To explore this, we exploit the panel structure of the data and zoom in on lasting ties. We define a lasting connection for a pair $\{i, j\}$, $Y_{ijc} = 1$, as a friendship present in survey 2 and remaining through the last observed survey.[26] We then re-estimate the same empirical strategy using this outcome and compute predicted inbreeding homophily across peer groups. The results are in Table 7.[27]

Table 7: Inbreeding Homophily of Lasting Friendships

	Inbreeding Homophily							
	Overall		Outside Stream		Stream not group		Study group	
	Raw	Pred	Raw	Pred	Raw	Pred	Raw	Pred
Gender	0.43	0.44	0.4	0.39	0.54	0.55	0.2	0.27
Ethnicity	0.39	0.39	0.56	0.55	0.31	0.32	0.02	0.08

Notes. "Raw" reports inbreeding homophily of lasting connections computed directly from the observed network using expression (4), with the analogous multi-category version for ethnicity. "Pred" reports inbreeding homophily implied by columns (1) and (4) of Appendix Table A7 for gender and ethnicity homophily, respectively.

[26] Formally, for cohort c with last observed survey S_c, $Y_{ijc} = \prod_{s=2}^{S_c} Y_{ijcs}$. We do not start with survey 1 to make sure enough meaningful interactions have taken place, but the results are robust for starting with survey 1.

[27] Table A7 in Appendix A.8 reports the regression estimates for lasting connections.

Although the levels of homophily are generally higher, the qualitative patterns are very similar to our benchmark results in Section 3.4. For gender, the inbreeding homophily of lasting friendships is substantially lower in study groups than outside them. For ethnicity, homophily declines monotonically as we move to smaller peer groups and becomes almost zero for study groups. In summary, our main findings do not appear to be driven by temporary friendships and also hold for lasting ones.

Detectable and less detectable traits. Gender and ethnicity are highly visible traits, so visibility may partly explain why they generate strong homophily. To explore this, we re-estimate specification (3) for a broader set of characteristics: region, pre-MBA industry, age, work experience, income, socioeconomic background, and the Big Five personality traits. Appendix A.9 describes the construction of these variables and reports the full results in Table A9. Region is strongly correlated with ethnicity, and the estimates for region exhibit a pattern similar to those for ethnicity. For the remaining—less directly observable—traits, the baseline trait coefficients show some evidence of homophily, but the magnitudes are much smaller than those of gender, ethnicity, and region. This pattern is in line with related evidence such as Jackson et al. (2022), and consistent with the interpretation we develop below that much of homophily is driven by search that operates along salient traits.

Career, social, personal-help, and close-friendship networks. As described in Section 2.2, we collected data on four types of connections: career ties, social ties, personal-help ties, and close friendships. To examine whether our results differ across these networks, we re-estimate our main specifications for gender and ethnicity separately for each network type. The results are reported in Appendix Table A10, and Appendix Table A12 reports the corresponding homophily premia. Homophily is positive in all four networks. For ethnicity, the pattern closely mirrors the main results: homophily is much lower among adjacent and study-group pairs than outside these smaller peer groups. For gender, the same qualitative pattern is present, although the decline in study groups is less pronounced in close friendships.

Reciprocated connections. Our main analysis uses the OR network, which includes all links claimed by at least one of the two students. To examine whether this choice matters, we repeat the analysis using the AND network. That is, a link for a pair (i, j) in survey s is included only if i nominates j and j nominates i in at least one question in survey s. Appendix Table A11 reports the estimates, and Appendix Table A12 reports the corresponding homophily premia. Reciprocated links are rarer, so the estimated exposure effects are smaller in levels. The homophily-premium pattern is broadly consistent with the main results for ethnicity. For gender, the study-group premium is lower than the stream-only premium, but the decline is less pronounced than in the OR network.

4 Mechanisms and Decomposition

Why does small-group exposure reduce homophily? To explore this, consider the mechanisms that drive homophily. Currarini et al. (2009) distinguish between two drivers: choice and chance, that is, preferences for same-trait peers and a meeting process biased toward same-trait peers.[28] We find it useful to further decompose biases in the meeting process into those arising from *peer-group composition* and those arising from *directed search*. The former refers to biases in how students sort, or are sorted, into peer groups. It includes institutional assignments such as the LBS study groups, as well as students' choices of which courses, clubs, or events to attend. These groupings are determined partly by institutional factors exogenous to the student, partly by students' preferences over activities (e.g., the "wine club"), and partly by preferences over the type of people they meet in different groups. In contrast, biases driven by directed search refer to students' choices of whom to seek out in these settings in order to establish a friendship. These choices are plausibly shaped by which individuals, or which already interacting sets of peers, students find familiar or compatible.

A key implication of directed search is that—for the traits of gender and ethnicity—it can turn a small preference bias into a large meeting bias for same-trait peers. This follows because of two properties. The decision of whom to approach is discrete, and traits other than gender and ethnicity are poorly observable before that decision is made. The latter implies that gender and ethnicity play an important role in determining whom to approach, while the former implies that even a small preference bias for same-gender or same-ethnicity friendships can have a large effect

[28] See also Bramoullé et al. (2012) for a similar mechanism in a population with multiple types.

on friendship composition. For example, a male student with a slight preference for male friends may, because of the slightly higher expected value associated with such matches, systematically choose to approach male peers in his stream when seeking a new friendship, thereby generating a large meeting bias and, consequently, a disproportionate representation of male friends.

In summary, we consider three mechanisms driving homophily: peer-group composition, preferences, and directed search. We first discuss reduced-form evidence that sheds light on the importance of these mechanisms, then develop a formal network accounting framework that disciplines the role of each in the decomposition, and then structurally estimate this model to quantify their contributions to overall homophily.

4.1 Reduced-Form Evidence on Mechanisms

Peer-group composition. Besides the LBS-assigned peer groups that were the focus of our analysis so far, students participate in many other peer groups. We have data on three major types of such groups: events organized by the Student Association, events hosted by the Career Centre, and tailored courses. These groups are important forums for interaction. Recall that the Student Association runs 1,820 distinct events from May 2024 to May 2026, with nearly 99% of students registering for at least one club event; the Career Centre data cover 291 events in the calendar year 2025, with nearly 84% of students attending at least one event. Students sort themselves into these events. They similarly sort themselves into tailored courses, which are taken in the second half of the first year. While tailored courses are conducted within a cohort, they do not have predetermined streams and do not respect the stream or study-group structure.[29]

To assess bias in peer-group composition in these events, we exploit registration data for club events, attendance data for career events, and enrollment data for tailored courses. For each type of peer group, we define a measure of co-registration: a pair (i, j) is defined as co-registered if both students register for the same club event, attend the same Career Centre event, or enroll in the same tailored course, respectively. We then compute inbreeding homophily for co-registration as in

[29] In the second year of the MBA program students choose elective courses; these are open to students from different degree programs and cohorts. Since we focus on connections across MBA students in the same cohort, we have not explored electives.

Section 3.4.

Table 8: Co-Registration Inbreeding Homophily in Clubs, Events, and Courses

	Student Association Club Events					Career Centre Events	Tailored Courses
	All	Regional	Social	Career	Sports	Attendance	Tailored
Gender	0.004	0.001	0.0	0.091	0.287	0.008	0.008
Ethnicity	0.002	0.067	-0.002	0.002	0.128	-0.007	0.006

Notes. Inbreeding homophily is calculated as in Section 3.4 from co-registration. Data for Student Association events cover May 2024–May 2026; data for Career Centre events cover the calendar year 2025; data for tailored courses cover the MBA 2025 and MBA 2026 cohorts.

Table 8 reports the results. Co-registration in Student Association events, Career Centre events, and tailored courses exhibits very low homophily, suggesting little sorting into these peer groups along these dimensions.[30] Thus, biases in the composition of these peer groups are unlikely to be an important driver of overall homophily. In principle, there could be more sorting in other peer groups that we do not observe. However, the fact that co-registration homophily is consistently low across the three different types of peer groups we do observe suggests that sorting is also limited in unobserved groups. We conclude that peer-group composition bias is unlikely to be the main driver of homophily in our setting.

Directed search and preference for similarity. To explore directed search and same-trait preferences, we added a section to Survey 2 for the MBA 2027 cohort to elicit both.[31] In this section, every student is presented with two hypothetical scenarios. Each scenario involves two hypothetical student profiles, each with three randomly assigned characteristics: gender (male or female), region (East Asian, European, Middle Eastern, North American, South American, or South Asian), and activity (music or traveling, sports or foodie). We chose region rather than ethnicity because they are correlated and we thought students would be more comfortable expressing their preferences over region.[32] In each scenario, all three characteristics of the two profiles differ.

In each scenario, we asked respondents a series of questions to infer their search behavior and

[30]The only exception is the events organized by sports clubs, a large share of which are gender-segregated.
[31]The complete survey is in Appendix C.
[32]Indeed, re-estimating Table 4 using region, rather than ethnicity, gives qualitatively similar results (Table A9).

their preferences. To infer their search behavior, we told respondents that the two profiles were two hypothetical students in their stream, that they had a few minutes before class started to have a chat with one of them, and asked them to choose the profile they would approach. To infer preferences, we randomly chose one of the two profiles offered in the directed search question and asked the respondent to picture a scenario in which, in recent months, they had had a chance to interact occasionally with that person. We then asked the respondent to report the likelihood that they would want to become friends with that person.

We use these measures to assess the strength of directed search and preference bias by estimating

$$Y_{ips} = \delta \cdot \text{Same trait}_{ip} + \text{Controls} + \varepsilon_{ips}. \tag{6}$$

Each observation consists of a student i and a profile p in a scenario s offered to that student. On the right-hand side, we include an indicator for the student having the same trait (gender or region) as the profile. The outcome variable measures choice or preference. For choice, the outcome is simply an indicator for the student choosing that profile. For preference, the outcome is the probability of friendship—conditional on having met the profile—reported by the student. The controls include respondent, scenario, and profile characteristics.

Because the choice and preference measures have different units, to make the results comparable, it is helpful to normalize the regression coefficients by the mean outcome of profiles that do not have the same trait. That is, we compute the choice premium as the proportional increase in the probability of choosing a same-trait rather than a different-trait profile. We compute the preference premium as the proportional increase, conditional on having met, in the probability that the student would want to become friends with a same-trait rather than a different-trait profile.

The regression results in Table 9 show that there is a significant effect of same gender and same region on both our choice (column 1) and preference (column 2) measures. Comparing the associated premia in Figure 4 reveals that the search premia are substantially and significantly larger than the preference premia. The likelihood of choosing to meet a same-gender student is about 34% higher than that of choosing to meet a different-gender student. In contrast, the reported likelihood of wanting to become friends with a same-gender student is only 4% higher than that for a different-gender student. Similarly, for region, the choice premium for same region is 33%, while the preference premium is only 11%. These findings are consistent with the view that same-trait preferences are relatively mild, but are amplified by directed search.

Table 9: Directed Search and Preference for Gender and Region

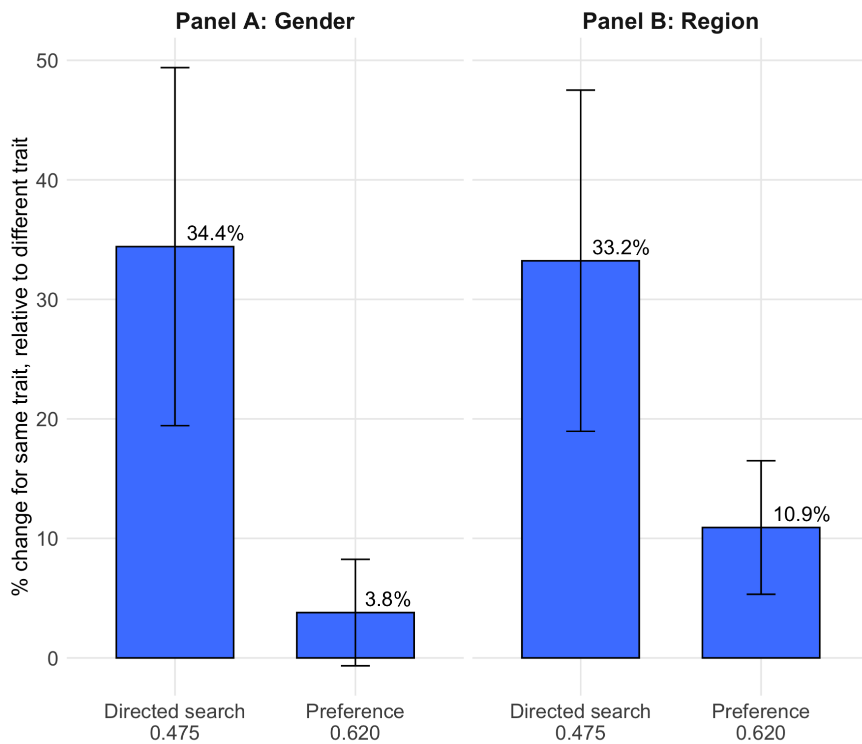
	Choose profile x	Preference for profile x
	(1)	(2)
Profile x has respondent's gender	0.147***	0.025*
	(0.032)	(0.014)
Profile x has respondent's region	0.158***	0.068***
	(0.034)	(0.018)
Sample	All	All
Observations	1,880	940

Notes. Each observation is a respondent-profile pair. Column (1) uses the directed-search outcome, equal to one if the respondent chooses profile x. Column (2) uses the respondent's reported probability of wanting to become friends with profile x, scaled from 0 to 1; profile x is randomly selected from the two profiles in the directed-search question. For the same-gender row, the indicator equals one if profile x has the respondent's gender; for the same-region row, it equals one if profile x has the respondent's region. All respondents see one same-gender and one different-gender profile; profile region is randomized across six regions, so some respondents see no same-region profile. Both specifications control for scenario, respondent gender and region, and profile activity. Standard errors, clustered at the individual level, are reported in parentheses. $^{*}p < 0.1;^{**} p < 0.05;^{***} p < 0.01$.

One concern is that stated friendship preferences may be more sensitive to social-desirability bias than approach decisions. However, both questions use the same profile information and closely parallel wording, so any stigma attached to expressing same-trait bias should affect the two measures similarly.

Another concern with these results is that our survey, which is based on unincentivized behavior in hypothetical scenarios, may not accurately capture real-world search behavior and preferences. Although we cannot fully address this concern, we provide two checks on the internal and external validity of our measures. First, Appendix Table B1 shows that respondents with a stronger tendency to choose same-trait profiles also report a stronger preference for such profiles. Second, Appendix Table B2 shows that directed search is positively associated with a student's gender

Figure 4: Search and Preference Premia Using Coefficients from Table 9



Notes. The figure reports search and preference premia computed from the coefficients in Table 9. For each trait, the search premium is the percentage increase in the probability of choosing a same-trait profile relative to a different-trait profile. The preference premium is the percentage increase in the reported probability of wanting to become friends with a same-trait profile relative to a different-trait profile, conditional on having met the profile. The baseline probabilities shown below the category labels are the corresponding predicted probabilities for different-trait profiles. Bars report point estimates; vertical lines show 95% confidence intervals.

homophily outside the study group, but not within the study group.[33] This pattern is consistent with our interpretation that directed search is muted inside study groups and more relevant in larger peer environments. The same table also shows that preference is generally positively related to homophily, though the evidence is less uniform across different layers of exposure.

Summary of reduced-form evidence. In sum, the evidence so far suggests limited bias in peer-group composition, and only modest same-gender and same-region preferences. By contrast, we do observe a significant tendency to direct search toward same-gender and same-region peers. The

[33] We do not report an analogous region-homophily version of Appendix Table B2 because this specification would leave very few observations for the relevant comparisons.

prominence of directed search is consistent with the findings of Ingram and Morris (2007), who document directed search by gender and race in a business-mixer event. However, it is unclear how much each mechanism contributes to observed homophily. To explore this, we develop and estimate a model that incorporates all three mechanisms.

4.2 A Network-Formation Accounting Framework

We develop an accounting framework to study friendship formation. Our framework does not derive behavior from first principles, but it allows us to decompose network formation into the three mechanisms of homophily highlighted above. The logic is as follows. Students attend their stream and some Student Association club events.[34] In each of these forums, they engage in directed search, approaching peers based on their traits. When two students meet, a friendship link forms if they both view the match favorably. Students also attend their study group; the small size and collaborative nature of this forum mute directed search, so links form whenever both students view the match favorably.

This framework captures our three mechanisms as follows. Peer-group composition is captured by the composition of peers in the stream and in the clubs; preferences for the same trait are captured by friendship choices conditional on meeting; and directed search is captured by students' choices of whom to approach. This structure naturally lends itself to a decomposition of homophily into the three mechanisms. Importantly, our framework does not identify the causal drivers of each mechanism: for example, choices in directed search could also reflect preferences. But it distinguishes the direct effects of preferences from any indirect effects they may have through the other mechanisms.

To clarify the distinction between model parameters and data moments, we denote any quantity Q observed in the data by Q^{obs}. We focus on friendships that form among two sets of possible links, L^{obs}_{Stream} and L^{obs}_{Group}. These correspond, respectively, to pairs of students who are in the same stream but not in the same study group, and to pairs of students who are in the same study group. We focus on three traits: gender, ethnicity, and region. For gender, students can have one of two trait

[34] For simplicity, we ignore the Career Centre events and the tailored courses.

values, F or M. For ethnicity and region, students can have one of the corresponding categorical trait values in the data. The trait value of student i is denoted by t_i.

We now describe the probability that a pair $i, j \in L^{obs}_{Stream}$ meets. Within the stream, meetings arise through directed search as follows. Let m_{ij} denote the probability that student i initiates a meeting with student j. Since either student can initiate the meeting, the probability that the pair meets in the stream is

$$m_{ij} + (1 - m_{ij})m_{ji}.$$

Within Student Association events, meetings arise through the same directed-search process, but only conditional on the two students co-registering. Let c^{obs}_{ij} be an indicator equal to one if students i and j register for at least one common event, and zero otherwise.[35] Let M_{ij} be an indicator equal to one if i and j meet. The probability that the pair meets either within the stream or through clubs is

$$\Pr[M_{ij} = 1 \mid ij \in L^{obs}_{Stream}] = [m_{ij} + (1 - m_{ij})m_{ji}] \left[1 + (1 - m_{ij})(1 - m_{ji})c^{obs}_{ij}\right].$$

Conditional on meeting, each student receives a signal about the quality of the match. Let Γ_{ij} denote the probability that student i receives a positive signal upon meeting student j. A friendship forms only if both students receive positive signals. Hence, under the simplifying assumption that these signals are i.i.d., the unconditional probability that the pair forms a friendship is

$$\Pr[Y_{ij} = 1 \mid ij \in L^{obs}_{Stream}] = \Pr[M_{ij} = 1 \mid ij \in L^{obs}_{Stream}] \cdot \Gamma_{ij}\Gamma_{ji}. \tag{7}$$

We now consider students $i, j \in L^{obs}_{Group}$ who belong to the same study group. We assume that, for these pairs, there are no meeting frictions; that is, $\Pr[M_{ij} = 1 \mid ij \in L^{obs}_{Group}] = 1$. This implies that

$$\Pr[Y_{ij} = 1 \mid ij \in L^{obs}_{Group}] = \Gamma_{ij}\Gamma_{ji}. \tag{8}$$

To make the model parsimonious, we assume that both directed-search intensities and prefer-

[35] For simplicity, we ignore repeated co-registrations.

ences for connections depend only on whether the two students share the same trait.

Assumption 1. Preferences for connections and directed-search intensities for a pair i, j depend only on whether the two students share the same trait; that is,

$$m_{ij} = \begin{cases} m_{ST} & \text{if } t_i = t_j \text{ and } ij \in L^{obs}_{Stream}, \\ m_{DT} & \text{if } t_i \neq t_j \text{ and } ij \in L^{obs}_{Stream}, \end{cases}$$

$$\Gamma_{ij} = \begin{cases} \Gamma_{ST} & \text{if } t_i = t_j \text{ and } ij \in L^{obs}_{Stream} \cup L^{obs}_{Group}, \\ \Gamma_{DT} & \text{if } t_i \neq t_j \text{ and } ij \in L^{obs}_{Stream} \cup L^{obs}_{Group}, \end{cases}$$

where ST and DT signify same trait and different trait, respectively.

Under Assumption 1, we can rewrite expression (7) as

$$\Pr[Y_{ij} = 1 \mid ij \in L^{obs}_{Stream}] = \begin{cases} \Gamma^2_{ST} \left[m_{ST} + (1 - m_{ST}) m_{ST}\right] \left[1 + (1 - m_{ST})^2 c^{obs}_{ij}\right] & \text{if } t_i = t_j, \\ \Gamma^2_{DT} \left[m_{DT} + (1 - m_{DT}) m_{DT}\right] \left[1 + (1 - m_{DT})^2 c^{obs}_{ij}\right] & \text{if } t_i \neq t_j, \end{cases} \tag{9}$$

and expression (8) as

$$\Pr[Y_{ij} = 1 \mid ij \in L^{obs}_{Group}] = \begin{cases} \Gamma^2_{ST} & \text{if } t_i = t_j, \\ \Gamma^2_{DT} & \text{if } t_i \neq t_j. \end{cases} \tag{10}$$

4.3 Moments and Estimation

For a given trait, such as gender, the model has four deep parameters summarized in the vector

$$\theta = (\Gamma_{ST}, \Gamma_{DT}, m_{ST}, m_{DT})'.$$

We estimate θ by two-step Generalized Method of Moments (GMM), with standard errors computed from the robust asymptotic covariance matrix; see Hansen (1982). In our main approach, we use four degree moments, giving a just-identified system.[36] Our moments measure the intensity

[36] For the just-identified specifications, the weighting matrix affects neither the point estimates nor their first-order asymptotic covariance; we nevertheless use the same two-step GMM procedure throughout for consistency with the

with which same-trait and different-trait links are formed in the stream and in the study group, respectively. To introduce these moments formally, we let $n^{obs}_{Stream} = |L^{obs}_{Stream}|$ and $n^{obs}_{Group} = |L^{obs}_{Group}|$ denote the cardinalities of the stream and study-group pair sets, respectively, and let Y^{obs}_{ij} indicate a link between i and j in the data.

Same-trait and different-trait links in L^{obs}_{Stream}. Our first two moments are the share of possible stream links that are actually formed between same-trait and different-trait students, respectively:

$$\frac{1}{n^{obs}_{Stream}} \sum_{ij \in L^{obs}_{Stream}} Y^{obs}_{ij} \mathbf{1}\{t_i = t_j\} = \frac{1}{n^{obs}_{Stream}} \sum_{ij \in L^{obs}_{Stream}} \Pr[Y_{ij} = 1 \mid ij \in L^{obs}_{Stream}] \mathbf{1}\{t_i = t_j\}, \quad (11)$$

$$\frac{1}{n^{obs}_{Stream}} \sum_{ij \in L^{obs}_{Stream}} Y^{obs}_{ij} \mathbf{1}\{t_i \neq t_j\} = \frac{1}{n^{obs}_{Stream}} \sum_{ij \in L^{obs}_{Stream}} \Pr[Y_{ij} = 1 \mid ij \in L^{obs}_{Stream}] \mathbf{1}\{t_i \neq t_j\}. \quad (12)$$

The left-hand side computes the number of same-trait (different-trait) same-stream links that are formed in the data, normalized by the number of possible links in L^{obs}_{Stream}. The right-hand side computes the predicted value of this share in the model, where $\Pr[Y_{ij} = 1 \mid ij \in L^{obs}_{Stream}]$ is given by expression (9).

Same-trait and different-trait links in L^{obs}_{Group}. These are the analogous moments for the study group:

$$\frac{1}{n^{obs}_{Group}} \sum_{ij \in L^{obs}_{Group}} Y^{obs}_{ij} \mathbf{1}\{t_i = t_j\} = \frac{1}{n^{obs}_{Group}} \sum_{ij \in L^{obs}_{Group}} \Gamma^2_{ST}, \quad (13)$$

$$\frac{1}{n^{obs}_{Group}} \sum_{ij \in L^{obs}_{Group}} Y^{obs}_{ij} \mathbf{1}\{t_i \neq t_j\} = \frac{1}{n^{obs}_{Group}} \sum_{ij \in L^{obs}_{Group}} \Gamma^2_{DT}. \quad (14)$$

The left-hand side computes the number of same-trait (different-trait) same-group links that are formed in the data, normalized by the number of possible links in L^{obs}_{Group}. The right-hand side computes the predicted value of this share in the model, where $\Pr[Y_{ij} = 1 \mid ij \in L^{obs}_{Group}]$ is given by expression (10).

When conducting the estimation, an issue we need to resolve is that we measure links for students in a given cohort in multiple surveys. Our approach is to take the average across surveys.

overidentified specifications.

That is, we define $Y_{ij}^{obs} = \frac{1}{s_{ij}^{obs}} \sum_s Y_{ij,s}^{obs}$, where $Y_{ij,s}^{obs}$ indicates a link between i and j in survey s and s_{ij}^{obs} is the number of surveys in which (i, j) are present. Note that $Y_{ij}^{obs} \in [0, 1]$.[37]

4.4 Decomposing Stream Homophily

Table B3 in Appendix B.2 reports the estimates $\hat{\theta}$ for gender, ethnicity, and region. We include both ethnicity and region because in the survey evidence on mechanisms in Section 4.1 we used region as a proxy for ethnicity. All model parameters are probabilities; reassuringly, the estimated parameters all lie in $[0, 1]$. Since the GMM is just identified, the moments are matched exactly.

We use the estimated model to decompose homophily in the stream into the contributions of peer-group composition, preferences, and directed search. We proceed in three steps: we first compute homophily with peer-group composition only, then add preferences, and finally add directed search. Our decomposition starts with a counterfactual that mutes biases in preferences and search, by setting $\Gamma_{ij} = (\hat{\Gamma}_{ST} + \hat{\Gamma}_{DT})/2$ and $m_{ij} = (\hat{m}_{ST} + \hat{m}_{DT})/2$ in the model. Since preferences and search no longer depend on the trait profile, in this counterfactual only peer-group composition affects inbreeding homophily. We then add back preferences by setting Γ_{ST} and Γ_{DT} to their estimated values, but maintaining no bias in search. Finally, we also add back the bias in directed search by setting m_{ST} and m_{DT} to their estimated values. We compute the model-implied inbreeding homophily in each counterfactual, and interpret the incremental changes as the contributions of preferences and directed search. Because the model is nonlinear, the order can in principle affect the channel contributions, but empirically this order dependence is small and has only minor effects on our quantitative results.

Table 10 reports the results. For gender, we find that peer-group composition makes a negligible contribution to inbreeding homophily, while preferences and directed search each account for about half. For ethnicity, peer-group composition again contributes negligibly, preferences account for 20%, and directed search accounts for 80%. The results for region are very similar to those for ethnicity, which is reassuring since we used region as a proxy for ethnicity in Section 4.1. These results indicate that directed search is a major driver of homophily in the stream.

[37] We replicated the structural analysis using only data from the second survey from all cohorts, and obtained very similar results.

Table 10: Decomposing Stream Inbreeding Homophily

	Gender		Ethnicity		Region	
Channel	Contribution	%	Contribution	%	Contribution	%
Club peer composition	0.002	0.6	0.000	0.1	0.002	0.7
	$[0.002, 0.002]$	$[0.5, 0.6]$	$[0.000, 0.000]$	$[0.1, 0.1]$	$[0.002, 0.002]$	$[0.6, 0.7]$
Preferences	0.190	53.9	0.053	20.5	0.045	13.8
	$[0.152, 0.228]$	$[42.9, 64.9]$	$[0.028, 0.078]$	$[11.0, 30.0]$	$[-0.011, 0.102]$	$[-3.4, 31.0]$
Directed search	0.161	45.5	0.207	79.4	0.282	85.5
	$[0.118, 0.203]$	$[34.6, 56.5]$	$[0.179, 0.236]$	$[69.9, 88.9]$	$[0.222, 0.342]$	$[68.3, 102.7]$
Total	**0.353**	**100.0**	**0.261**	**100.0**	**0.329**	**100.0**

Notes. The table decomposes model-implied stream inbreeding homophily into the contributions of club peer composition, preferences, and directed search. Contributions are measured as the incremental effect of each channel on inbreeding homophily. Percentage contributions are computed by dividing each channel's contribution by total model-implied homophily. 95% confidence intervals for the channel rows are reported in brackets below the corresponding estimates. These intervals are computed by the delta method, treating each reported decomposition quantity as a function of the four estimated GMM parameters and using the full two-step GMM covariance matrix.

The large role of directed search reflects the fact, documented in Table 5, that inbreeding homophily is substantially higher in the stream than in the study group. Our moments capture this difference in homophily across two environments, so the model must account for it through one of the three channels. Preferences are assumed to be the same across environments, while peer-group composition differs little between the stream and study group in the data. As a result, the model attributes most of the additional homophily observed in the stream to directed search, which is assumed to operate only in that environment. This logic also explains why preferences account for a larger share of gender homophily than of ethnicity or region homophily. Since directed search is assumed not to operate in study groups, the homophily that remains in study groups is attributed to preferences. Because study groups largely eliminate ethnicity and region homophily but not gender homophily, the model assigns a larger role to preferences for gender.

Identifying from data on preferences and search. The estimates in Table 10 are identified using variation in friendship composition between the study group and the stream. As discussed in Section 4.1, we also collected direct measures of preferences and directed search in the survey. These provide three additional moments, formally defined in Appendix B.3, measuring the same-trait preference premium and directed search premium in the stream. Because this survey was conducted only for the MBA 2027 cohort, these moments are available only for that sample, and because we did not collect these measures for ethnicity, we can only use them for gender and region.

The additional moments allow us to re-estimate the model using a different source of variation: the difference between the preference and directed search premia. Combined with the stream composition moments (11) and (12), the survey moments yield an overidentified system that exploits this additional source of variation. We can also combine the seven moments in other ways, or include all of them. Appendix Tables B4 and B5 report the results for different combinations of moments. There is some tension across the different sources of variation, which causes the J-test to reject some specifications. However, every specification matches stream homophily well and supports the same qualitative conclusion. The share of inbreeding homophily attributed to directed search ranges from 25% to 75% for gender and from 53% to 96% for region. Thus, the conclusion that directed search is a major driver of homophily in the stream is robust across two independent sources of identifying variation.

Lasting friendships. A potential concern with the results so far is that the full network may include temporary or superficial connections, whereas preferences may play a larger role in the formation of lasting friendships. To address this concern, we replicate our main structural specification and homophily decomposition using only lasting friendships. Appendix Tables B6 and B7 show that the estimates are qualitatively similar: directed search remains an important driver of gender, ethnicity, and region homophily, indicating that our conclusions are not driven by short-lived relationships.

4.5 Alternative Explanation: Study Group Reduces Cost of Connections

An alternative explanation for the gap between stream homophily and study-group homophily, already introduced in Section 3.6, is that the two peer environments differ in the cost of a connection. This alternative predicts that the study-group reduction in homophily should fade among lasting ties; empirically, it does not. The model allows us to formulate this explanation more precisely.

We can think of the likelihood that i would like to become friends with j as $\Gamma_{ij} = \Pr[v_{ij} - c_{ij} \geq$

0], where v_{ij} is the gross per-period value to i and c_{ij} is the per-period maintenance cost to i. Assumption 1 restricts $v_{ij} - c_{ij}$ to depend only on t_i and t_j, and not on the exposure treatment of ij. However, the cost c_{ij} may be lower for pairs in the same study group during the period in which study groups are active, namely the first nine months of the program. Moreover, this cost reduction may be larger for different-trait pairs if repeated exposure is especially helpful for reducing coordination frictions when those frictions are initially large. In this explanation, the difference between stream and study group homophily is driven by preferences, not search.

This explanation implies that the same-trait preference premium should be larger in the stream than in the study group. However, the maintenance-cost advantage created by active study groups should disappear once the study groups dissolve. Therefore, the gap in the same-trait preference premium between the stream and the study group should shrink or disappear when we focus on lasting friendships. As we noted in Section 3.6, this does not happen (Table 7), providing evidence against the alternative explanation. The structural model allows us to make this argument quantitatively.

Specifically, we estimate a version of the model in which preferences depend both on trait similarity and on whether two students are assigned to the same study group. We shut down directed search: all pairs are assumed to meet with probability one. Formally, this specification replaces Assumption 1 with Assumption 2 (Appendix B.4). This version of the model also has four moments, corresponding to the same-trait and different-trait preferences in the stream and in the study group. We derive the analogs of moments (11)–(14) under Assumption 2, stated in (B3). We estimate this specification in both the full network and the network of lasting connections. The results are reported in Appendix Table B8.

The results do not support the temporary maintenance-cost explanation. Although this specification matches the homophily patterns in the full network, it fails to match the key prediction that the preference gap should shrink or disappear among lasting connections. In the full network, the same-trait preference premium is larger in the stream than in the study group: 44% versus 20.9% for gender, and 92.5% versus 12.2% for region. However, this gap does not seem to narrow once we focus on lasting connections, where the same-trait preference premium in the stream versus the group is 82.9% versus 25.6% for gender, and 116.8% versus 31.7% for region.

This quantification also contrasts with the survey data on search and preferences. Because this model attributes all homophily to preferences, it implies lower search premia and higher preference premia than those observed in the survey. Specifically, this model implies zero same-trait search premia in the stream, whereas Figure 4 documents survey-based search premia of 34% by gender and 33% by region. It also implies same-trait preference premia in the stream of 44% for gender and 92.5% for region, compared with survey-based preference premia of only 3.8% and 10.9%. Taken together, these results favor the directed-search explanation over the temporary maintenance-cost explanation.

5 Conclusion

This paper studies how peer-group assignment shapes homophily in friendship networks. Using the quasi-random assignment of MBA students to streams, study groups, and adjacent seats, we show that exposure in smaller groups substantially reduces homophily by gender and ethnicity. To understand the underlying mechanism, we combine reduced-form evidence, survey measures of friendship preferences and approach decisions, and a structural model of network formation. We find little evidence that peer-group composition is an important source of homophily, and only modest same-trait preferences. By contrast, both the survey evidence and structural estimates indicate that directed search is an important source of homophily in the stream. These findings suggest that small-group interactions reduce homophily primarily by muting directed search, thereby creating opportunities for cross-trait friendships that would otherwise not arise.

Our result that preferences play a modest role in driving homophily suggests that policies that reduce homophily may help realize potential benefits from more diverse social networks without substantially conflicting with individuals' friendship preferences. Our results further suggest that peer-group design may be an effective way to achieve that goal. Peer-group design does not require changing preferences or the composition of the population. Instead, it works by encouraging interactions that otherwise might not occur. An important next step in this research agenda is to conduct randomized interventions that vary group size, generating direct evidence on the full equilibrium effects of peer-group design on network formation and homophily.

References

Ball, B. and Newman, M. E. J. (2012). Friendship networks and social status. *arXiv preprint arXiv:1205.6822*.

Boisjoly, J., Duncan, G. J., Kremer, M., Levy, D. M., and Eccles, J. (2006). Empathy or antipathy? the impact of diversity. *American Economic Review*, 96(5):1890–1905.

Borusyak, K. and Hull, P. (2023). Nonrandom exposure to exogenous shocks. *Econometrica*, 91(6):2155–2185.

Boucher, V., Tumen, S., Vlassopoulos, M., Wahba, J., and Zenou, Y. (2020). Ethnic mixing in early childhood: Evidence from a randomized field experiment and a structural model. Technical report, CEPR Discussion Papers.

Bramoullé, Y., Currarini, S., Jackson, M. O., Pin, P., and Rogers, B. W. (2012). Homophily and long-run integration in social networks. *Journal of Economic Theory*, 147:1754–1786.

Calvó-Armengol, A. and Jackson, M. O. (2004). The effects of social networks on employment and inequality. *American Economic Review*, 94(3):426–454.

Carrell, S. E., Hoekstra, M., and West, J. E. (2019). The impact of college diversity on behavior toward minorities. *American Economic Journal: Economic Policy*, 11(4):159–182.

Carrell, S. E., Sacerdote, B. I., and West, J. E. (2013). From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica*, 81(3):855–882.

Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R. B., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., Johnston, D., Koenen, M., Laguna-Muggenburg, E., Mudekereza, F., Rutter, T., Thor, N., Townsend, W., Zhang, R., Bailey, M., Barbera, P., Bhole, M., and Wernerfelt, N. (2022). Social capital i: measurement and associations with economic mobility. *Nature*, 608(7921):108–121.

Coleman, J. S. (1958). Relational analysis: The study of social organizations with survey methods. *Human Organization*, 17(4):28–36.

Corno, L., La Ferrara, E., and Burns, J. (2022). Interaction, stereotypes, and performance: Evidence from south africa. *American Economic Review*, 112(12):3848–3875.

Currarini, S., Jackson, M. O., and Pin, P. (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045.

Currarini, S., Jackson, M. O., and Pin, P. (2010). Identifying the roles of race-based choice and chance in high school friendship formation. *Proceedings of the National Academy of Sciences*, 107(11):4857–4861.

Golub, B. and Jackson, M. O. (2012). How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, 127(3):1287–1338.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Hasan, S. and Bagde, S. (2015). Peers and network growth: Evidence from a natural experiment. *Management Science*, 61(10):2536–2547.

Heckman, J. J. and Vytlacil, E. J. (1998). Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources*, 33(4):974–987.

Ingram, P. and Morris, M. W. (2007). Do people mix at mixers? structure, homophily, and the "life of the party". *Administrative Science Quarterly*, 52(4):558–585.

Ioannides, Y. M. and Loury, L. D. (2004). Job information networks, neighborhood effects, and inequality. *Journal of Economic Literature*, 42(4):1056–1093.

Jackson, M. O. (2009). Networks and economic behavior. *Annual Review of Economics*, 1:489–511.

Jackson, M. O., Nei, S. M., Snowberg, E., and Yariv, L. (2022). The dynamics of networks and homophily. *NBER Working Paper*.

Katz, L. F. (2026). Neighborhood effects and missing markets for opportunity. Technical report, National Bureau of Economic Research.

Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311(5757):88–90.

Lazarsfeld, P. F. and Merton, R. K. (1954). Friendship as a social process: A substantive and methodological analysis. In Berger, M., Abel, T., and Page, C. H., editors, *Freedom and Control in Modern Society*, pages 18–66. Van Nostrand, New York.

Lowe, M. (2021). Types of contact: A field experiment on collaborative and adversarial caste integration. *American Economic Review*, 111(6):1807–1844.

Marmaros, D. and Sacerdote, B. (2006). How do friendships form? *The Quarterly Journal of Economics*, 121(1):79–119.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444.

Montgomery, J. D. (1991). Social networks and labor-market outcomes: Toward an economic analysis. *American Economic Review*, 81(5):1407–1418.

Pin, P. and Rogers, B. W. (2016). Stochastic network formation and homophily. In Bramoullé, Y., Galeotti, A., and Rogers, B. W., editors, *The Oxford Handbook of the Economics of Networks*, pages 138–166. Oxford University Press.

Appendix

Appendix A provides additional material on the institutional context, data, and empirical design. Appendix B contains the material supporting Section 4. Appendix C reproduces the survey instruments used to measure directed search, preferences, connections, and socioeconomic background.

A Appendix to Section 2: Context, Data, and Design

Appendix A.1 gives the empirical-strategy details; Appendix A.2 describes the assignment algorithms; Appendix A.3 reports the balance tests; Appendices A.4–A.7 report the individual-fixed-effects, joint-trait, similarity-index, and male/female inbreeding-homophily specifications; and Appendices A.8–A.10 report the lasting-friendship, less-detectable-trait, and alternative-network robustness exercises.

A.1 Discussion of Empirical Strategy

This appendix formalizes three points that are discussed in Section 2.4. First, we state assumptions under which recentered instruments based on conditional counterfactuals identify the causal effect of each exposure. This is an application of Borusyak and Hull (2023) to a setting in which the assignment process is nested: streams are assigned first, study groups are assigned within streams, and seating charts are assigned conditional on both (Section A.1.1). Second, Section A.1.2 explains why conditional recentering is required for downstream exposures: if the regression omits an interaction with an upstream exposure, conditional counterfactuals still satisfy the orthogonality conditions needed to identify downstream coefficients, whereas unconditional counterfactuals generally do not. Third, Section A.1.3 characterizes the RCIV estimand when the regression omits an interaction between exposure and a predetermined pair characteristic.

Throughout, we index an unordered pair of students by $\ell = (i, j)$ and suppress cohort and survey-wave indices whenever doing so simplifies notation.

A.1.1 Identification with a Nested Assignment Algorithm

Let

$$D_\ell := (S_\ell, G_\ell, A_\ell)'$$

collect the three exposure treatments, let T_ℓ denote a predetermined pair characteristic (or a vector of pair characteristics) that is included in the regression, such as whether the pair shares the same gender, and let C_ℓ collect other exogenous controls. The specified interacted model is

$$Y_\ell = \alpha + D_\ell'\beta + \rho T_\ell + (T_\ell D_\ell)'\omega + \lambda' C_\ell + \varepsilon_\ell, \tag{A1}$$

where $\beta = (\beta_S, \beta_G, \beta_A)'$ and $\omega = (\omega_S, \omega_G, \omega_A)'$. This is analogous to our main specification (3).

For each assignment stage $X \in \{S, G, A\}$, write exposure as

$$X_\ell = f_\ell^X(g^X; w^X), \qquad X \in \{S, G, A\}, \tag{A2}$$

where g^X is the randomization shock used by the X-assignment algorithm and w^X is the full set of predetermined inputs held fixed when that algorithm is run. In our setting, w^S contains the predetermined characteristics used by the stream algorithm; w^G contains the predetermined characteristics used by the study-group algorithm together with the realized stream allocation; and w^A contains the predetermined characteristics used by the seating-chart algorithm together with the realized stream and study-group allocations.

For each stage $X \in \{S, G, A\}$, let $F^X(\cdot \mid w^X)$ denote the conditional distribution of the assignment shock induced by the stage-X algorithm. Since w^X contains the full set of inputs used by the algorithm, the distribution of the assignment shock does not depend on T_ℓ once w^X is fixed:

$$F^X(g^X \mid w^X, T_\ell) = F^X(g^X \mid w^X).$$

Therefore, since

$$X_\ell = f_\ell^X(g^X; w^X),$$

we have

$$\mathbb{E}[X_\ell \mid w^X, T_\ell] = \int f_\ell^X(\gamma; w^X)\, dF^X(\gamma \mid w^X, T_\ell) = \int f_\ell^X(\gamma; w^X)\, dF^X(\gamma \mid w^X) = \mathbb{E}[X_\ell \mid w^X].$$

The following is the main identifying assumption:

Assumption A1 (Weak shock exogeneity). For each stage $X \in \{S, G, A\}$

$$\mathbb{E}[\varepsilon_\ell \mid g^X, w^X, T_\ell] = \mathbb{E}[\varepsilon_\ell \mid w^X, T_\ell].$$

Assumption A1 is the analog of weak shock exogeneity in Borusyak and Hull (2023). It states that, conditional on the information used by the assignment algorithm and on the pair characteristic, the assignment shock is mean independent of the structural error.[38]

Combining Assumption A1 with the assignment-mechanism property above, we obtain the orthogonality conditions:

$$E[Z_\ell \widetilde{X}_\ell \varepsilon_\ell] = 0 \quad \text{for all} \quad Z_\ell \in \{1, T_\ell\}$$

.

Define the stage-specific expected exposure

$$\mathbb{E}[X_\ell \mid w^X] = \int f_\ell^X(\gamma; w^X)\, dF^X(\gamma \mid w^X), \qquad X \in \{S, G, A\},$$

and the corresponding recentered exposure

$$\widetilde{X}_\ell := X_\ell - \mathbb{E}[X_\ell \mid w^X].$$

These are precisely the objects measured in the data by repeatedly rerunning the stage-specific assignment algorithm while holding fixed the appropriate inputs.

[38] In the basic specification of Borusyak and Hull (2023), there is no interaction between treatment X and pair characteristic T_ℓ, as in our specification (2). In that case, identification requires the weaker condition $E[\varepsilon_\ell \mid g^X, w^X] = E[\varepsilon_\ell \mid w^X]$. Assumption A1 strengthens it to account for the interacted term $T_\ell X_\ell$. We note that if T_ℓ is measurable with respect to w^X, the strengthened condition is implied by the baseline condition.

Since $\widetilde{X}_\ell$ is a function of (g^X, w^X), we have

$$E[Z_\ell \widetilde{X}_\ell \varepsilon_\ell] = E\left[Z_\ell \widetilde{X}_\ell E[\varepsilon_\ell \mid g^X, w^X, T_\ell]\right].$$

Assumption A1 implies

$$E[Z_\ell \widetilde{X}_\ell \varepsilon_\ell] = E\left[Z_\ell \widetilde{X}_\ell E[\varepsilon_\ell \mid w^X, T_\ell]\right].$$

Because Z_ℓ and $E[\varepsilon_\ell \mid w^X, T_\ell]$ are measurable with respect to (w^X, T_ℓ), iterated expectations give

$$E[Z_\ell \widetilde{X}_\ell \varepsilon_\ell] = E\left[Z_\ell E[\varepsilon_\ell \mid w^X, T_\ell] E[\widetilde{X}_\ell \mid w^X, T_\ell]\right] = 0.$$

The last equality follows because

$$E[\widetilde{X}_\ell \mid w^X, T_\ell] = E[X_\ell - E[X_\ell \mid w^X] \mid w^X, T_\ell] = E[X_\ell \mid w^X, T_\ell] - E[X_\ell \mid w^X] = 0.$$

Hence, under instrument relevance, the coefficients in (A1) are identified by 2SLS using the instrument vector

$$\left(\widetilde{S}_\ell,\, \widetilde{G}_\ell,\, \widetilde{A}_\ell,\, T_\ell \widetilde{S}_\ell,\, T_\ell \widetilde{G}_\ell,\, T_\ell \widetilde{A}_\ell\right).$$

A.1.2 Why Unconditional Counterfactuals Fail in the Misspecified Model

We now explain why conditional recentering is required for downstream exposures. To keep the argument transparent, consider the study-group effect. Suppose the true model is

$$Y_\ell = \alpha + \beta_S S_\ell + \beta_G G_\ell + \rho T_\ell + \omega T_\ell S_\ell + \varepsilon_\ell, \tag{A3}$$

and suppose the estimated regression omits T_ℓ and $T_\ell S_\ell$. The omitted component is

$$\eta_\ell := \rho T_\ell + \omega T_\ell S_\ell + \varepsilon_\ell.$$

The conditional study-group residual is $\widetilde{G}_\ell := G_\ell - \mathbb{E}[G_\ell \mid w^G]$. Since S_ℓ is measurable with respect to w^G and since the study-group algorithm only conditions on w^G, we have

$$E[G_\ell \mid w^G, T_\ell, S_\ell] = E[G_\ell \mid w^G].$$

Therefore,

$$E[\widetilde{G}_\ell \mid w^G, T_\ell, S_\ell] = E[G_\ell - E[G_\ell \mid w^G] \mid w^G, T_\ell, S_\ell] = 0.$$

It follows that

$$E[\widetilde{G}_\ell T_\ell S_\ell] = E\left[T_\ell S_\ell E[\widetilde{G}_\ell \mid w^G, T_\ell, S_\ell]\right] = 0.$$

Similarly,

$$E[\widetilde{G}_\ell T_\ell] = E\left[T_\ell E[\widetilde{G}_\ell \mid w^G, T_\ell]\right] = 0.$$

Assumption A1 gives

$$E[\widetilde{G}_\ell \varepsilon_\ell] = 0.$$

Hence

$$E[\widetilde{G}_\ell \eta_\ell] = \rho E[\widetilde{G}_\ell T_\ell] + \omega E[\widetilde{G}_\ell T_\ell S_\ell] + E[\widetilde{G}_\ell \varepsilon_\ell] = 0.$$

Now consider instead an *unconditional* study-group counterfactual correction, obtained by omitting the realized stream allocation from the information set used to redraw study groups. Let $\bar{w}^G$ denote this reduced information set and define

$$\widetilde{G}_\ell^u := G_\ell - \mathbb{E}[G_\ell \mid \bar{w}^G].$$

Abstracting from any correlation between $\widetilde{G}_\ell^u$ and ε_ℓ, the omitted interaction term alone generates

$$\begin{aligned} E[\widetilde{G}_\ell^u \eta_\ell] &= \rho E[\widetilde{G}_\ell^u T_\ell] + \omega E[\widetilde{G}_\ell^u T_\ell S_\ell] \\ &= \rho E[\mathrm{Cov}(G_\ell, T_\ell \mid \bar{w}^G)] + \omega E[\mathrm{Cov}(G_\ell, T_\ell S_\ell \mid \bar{w}^G)] \neq 0. \end{aligned}$$

The first term is zero if T_ℓ is measurable with respect to $\bar{w}^G$, but the second term is generically nonzero because the nested structure of the assignment process implies that S_ℓ is correlated with G_ℓ. Hence, the unconditional correction violates the RCIV orthogonality condition.

This is the formal reason study-group counterfactuals must be drawn conditional on the realized stream assignment. Unconditional recentering removes the component of study-group exposure predictable from the predetermined information in $\bar{w}^G$, but it leaves in the instrument the component predictable from the realized upstream assignment S_ℓ. That residual component is exactly what generates the correlation with the omitted term $T_\ell S_\ell$.

The same logic applies one step further down the assignment sequence, from study groups to seating-chart exposure.

A.1.3 Misspecification: What RCIV Identifies

We now characterize the RCIV estimand when the regression omits an interaction between exposure and a predetermined pair characteristic. Throughout this subsection, we maintain Assumption A1 and consider the well-specified conditional counterfactual corrections. This allows us to isolate misspecification of the outcome equation from the separate issue of misspecifying the counterfactual correction, which is studied in Section A.1.2 above.

To make the interpretation of RCIV under misspecification transparent, begin with a single exposure X_ℓ assigned according to (A2). Suppose the true model is

$$Y_\ell = \alpha + \beta X_\ell + \rho T_\ell + \omega T_\ell X_\ell + \varepsilon_\ell, \tag{A4}$$

but the estimated regression omits T_ℓ and $T_\ell X_\ell$:

$$Y_\ell = \alpha + b X_\ell + u_\ell. \tag{A5}$$

Under Assumption A1

$$b^{RCIV} = \frac{\mathbb{E}[\widetilde{X}_\ell Y_\ell]}{\mathbb{E}[\widetilde{X}_\ell X_\ell]} = \beta + \omega \frac{\mathbb{E}[\widetilde{X}_\ell T_\ell X_\ell]}{\mathbb{E}[\widetilde{X}_\ell X_\ell]} = \beta + \omega \frac{\mathbb{E}[T_\ell \operatorname{Var}(X_\ell \mid w^X)]}{\mathbb{E}[\operatorname{Var}(X_\ell \mid w^X)]}.$$

Hence

$$b^{RCIV} = \beta + \zeta\omega, \qquad \zeta := \frac{\mathbb{E}[T_\ell \operatorname{Var}(X_\ell \mid w^X)]}{\mathbb{E}[\operatorname{Var}(X_\ell \mid w^X)]}. \tag{A6}$$

Equation (A6) shows that RCIV does not recover the baseline effect β in the misspecified scalar model. Instead, it identifies a variance-weighted average treatment effect. If T_ℓ is binary, then $\zeta \in [0, 1]$, so b^{RCIV} lies between the treatment effect for $T_\ell = 0$ pairs and the treatment effect for $T_\ell = 1$ pairs. If the conditional variance $\operatorname{Var}(X_\ell \mid w^X)$ is constant, then $\zeta = \mathbb{E}[T_\ell]$, and the scalar RCIV coefficient reduces to the average exposure effect across values of T_ℓ.

Relation to correlated random coefficient IV. This is closely related to the logic of Heckman and Vytlacil (1998) and to Appendix C.1 of Borusyak and Hull (2023), which extends classic results on IV identification in the presence of heterogeneous treatment effects in the recentered IV approach. In Heckman and Vytlacil (1998)'s correlated-random-coefficient framework, IV does not in general recover a common structural coefficient when treatment effects are heterogeneous; instead, it recovers a weighted average of heterogeneous effects, with the weights determined by how the instrument generates identifying variation. Our setting is a special case of that logic. The difference is that here the instrument is the recentered assignment innovation, so the weights take the form of conditional variances of exposure.

Formally, specification (A4) can be rewritten as

$$Y_\ell = \alpha_\ell + B_\ell X_\ell + \varepsilon_\ell, \qquad \alpha_\ell := \alpha + \rho T_\ell, \quad B_\ell := \beta + \omega T_\ell.$$

Thus the coefficient on exposure is a random coefficient as in Heckman and Vytlacil (1998). For any valid instrument q_ℓ, the IV estimand is

$$b(q) = \frac{\mathbb{E}[q_\ell B_\ell X_\ell]}{\mathbb{E}[q_\ell X_\ell]},$$

that is, a weighted average of the heterogeneous slopes B_ℓ. In our case, $q_\ell = \widetilde{X}_\ell := X_\ell - \mathbb{E}[X_\ell \mid w^X]$, so the weights are proportional to $\mathrm{Var}(X_\ell \mid w^X)$. This yields

$$b^{RCIV} = \frac{\mathbb{E}[B_\ell \,\mathrm{Var}(X_\ell \mid w^X)]}{\mathbb{E}[\mathrm{Var}(X_\ell \mid w^X)]} = \beta + \omega \frac{\mathbb{E}[T_\ell \,\mathrm{Var}(X_\ell \mid w^X)]}{\mathbb{E}[\mathrm{Var}(X_\ell \mid w^X)]}.$$

Hence, our setting is a structured correlated-random-coefficient model in which the RCIV weights are induced by the conditional variance of exposure.

Extending the logic to the full three-exposure model. Let

$$D_\ell = (S_\ell, G_\ell, A_\ell)', \qquad \widetilde{D}_\ell = (\widetilde{S}_\ell, \widetilde{G}_\ell, \widetilde{A}_\ell)'.$$

Suppose the model is

$$Y_\ell = \alpha + D_\ell'\beta + \rho T_\ell + (T_\ell D_\ell)'\omega + \varepsilon_\ell,$$

but the estimated regression omits T_ℓ and $T_\ell D_\ell$. Then the RCIV coefficient vector satisfies

$$b^{RCIV} = \beta + Q^{-1} M \omega, \tag{A7}$$

where

$$Q := \mathbb{E}[\widetilde{D}_\ell D_\ell'], \qquad M := \mathbb{E}[\widetilde{D}_\ell (T_\ell D_\ell)'].$$

Thus, in the presence of multiple correlated exposures, omitted-interaction bias need not remain within a single coefficient. Instead, each RCIV coefficient generally becomes a linear combination of the heterogeneous exposure effects.[39]

Finally, suppose the regression includes interactions of the exposure treatments with one predetermined characteristic T_ℓ, but omits interactions with another characteristic Z_ℓ. Then the same omitted-variable logic applies to the included endogenous block $(D_\ell, T_\ell D_\ell)$. The coefficients on the included exposure terms and the included interaction terms need not recover their baseline counterparts; rather, they are shifted by a linear combination of the omitted heterogeneous effects

[39] Analogously to the discussion of the correlated random coefficient (IV) above, in this multivariate case, a simplification obtains under the assumption that $\mathbb{E}[\widetilde{D}_\ell (T_\ell D_\ell)'] = \bar{t}\,\mathbb{E}[\widetilde{D}_\ell D_\ell']$ for some scalar $\bar{t}$, in which case (A7) becomes $b^{RCIV} = \beta + \bar{t}\,\omega$.

associated with $Z_\ell D_\ell$. This is the sense in which an omitted interaction can affect not only the exposure coefficients, but also the coefficients on the interactions that remain in the regression.

A.2 Counterfactuals: Stream, Study-Group, and Seating-Chart Algorithms

LBS uses quasi-random procedures to allocate students to streams, study groups, and seating charts. These allocations must satisfy all requirements set by the MBA program office to ensure that student characteristics are evenly balanced across streams and study groups. To implement these requirements, the allocation is modeled as a binary programming problem. For our sample, there were four kinds of constraints for the stream assignment:

1. Each student is assigned only once.

2. Maximum stream size for each cohort.

3. For each stream, s, the average value of a continuous characteristic, c, must be such that $\bar{c}_s$ is within $\bar{c} \pm b_c$, where $\bar{c}_s$ is the stream-level average, $\bar{c}$ is the cohort-level average, and b_c is the deviation permitted by the program office for characteristic c. Examples of continuous characteristics are age, total GMAT score, and work experience.

4. For each stream, s, the number of students with discrete characteristic d must be such that either $d_s = \bar{d}$ or d_s is within $\bar{d} \pm 1$, where d_s is the number of students with characteristic d in stream s and $\bar{d}$ is the total number of students with characteristic d divided by the number of streams. Examples of discrete characteristics are prior industry, firm, region, nationality, gender, and native English status.

After the stream allocation, each student is assigned to a study group. The allocation of study groups follows the same structure as that of streams, with a broader range of constraints for continuous and discrete characteristics reflecting the composition of the stream.

Seating charts are generated after every student is allocated to a stream and a study group. In the program's first year, there are three iterations of seating charts. The first seating chart applies to all core courses in Term 1, Part 1, with 21 sessions. For Term 1, Part 2, which has 40 sessions, students are assigned a second seating chart. The third seating chart is used for core courses in Term 2, with 25 sessions. For the MBA 2025 and MBA 2026 cohorts, two different algorithms were used for the seating charts, one for chart 1 and the other for charts 2 and 3. This was standardized for the MBA 2027 cohort, and the second algorithm was used for all three seating charts.

The first algorithm, used to generate chart 1, takes the list of students in any stream and scrambles the order up to $5,000$ times. In each iteration, it calculates the number of clashes. Here, a clash is defined as an instance where two students assigned adjacent seats have the same gender, have the same nationality, or belong to the same study group. The algorithm uses a weighting scheme where the weights of clashes in gender, nationality, and study groups are 1, 2, and 3, respectively. The algorithm stops when either the weighted sum of the clashes is zero or $N = 5,000$ iterations are complete. If the algorithm reaches the Nth draw, then the allocation with the lowest clash score is implemented.

The second algorithm follows the same demographic criteria set by the program office but follows a different procedure. Instead of a weighted clash score, the algorithm counts the number of red flags. A red flag is defined as any instance where two students assigned adjacent seats have the same gender, have the same nationality, or belong to the same study group. Note that in this case, all red flags hold the same weight, and there can be at most one red flag for a pair of students, even if they clash on more than one dimension. The program first takes the list of students in a stream, assigns them a random seat, and calculates the total number of red flags generated in the completely random assignment. Then, it selects two students at random and checks whether swapping those two students is "feasible." Here, a swap is considered feasible if either student does not generate any red flags in their new seats. If the swap is deemed feasible, then the swap is implemented if the total number of new red flags is no more than the number of red flags before the swap. Once the swap is implemented or rejected, the program continues to repeat this process until either the total number of red flags is zero or $1,500$ swaps have been implemented.

A.3 Recentered Exposure: Balance Tests

Table A1: Balance Tests

Spec		Gender	Nat.	Ind.	GMAT	Age	Native English status	Work ex.	Study group
OLS	Stream	-0.006**	-0.009***	-0.007***	-0.002	-0.004	-0.006**	-0.001	
		(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	
	Study group	-0.092***	-0.063***	-0.118***	-0.036***	-0.044***	-0.088***	-0.047***	
		(0.009)	(0.004)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	
	Adj. in A1	-0.245***	-0.052***						-0.0003
		(0.014)	(0.007)						(0.003)
	Adj. in A2	-0.286***	-0.049***						0.001
		(0.010)	(0.005)						(0.002)
	Adj. in A3	-0.289***	-0.049***						0.002
		(0.008)	(0.004)						(0.002)
RCIV	Stream	0.0001	6.94×10^{-5}	-0.0002	-0.0004	-0.001	4.75×10^{-5}	0.0007	
		(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	
	Study group	-0.010	-0.0007	-0.008	-0.011	-0.003	-0.003	0.009	
		(0.009)	(0.004)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	
	Adj. in A1	-0.006	-0.004						-0.0007
		(0.014)	(0.007)						(0.003)
	Adj. in A2	-0.017*	-0.001						-0.0010
		(0.010)	(0.005)						(0.002)
	Adj. in A3	-0.011	-0.001						-0.001
		(0.008)	(0.004)						(0.002)

Notes. All regressions were estimated using data from a sample of 1,400 students, which produced 328,671 observations based on unordered dyadic pairs. The OLS specification uses equation (2), and the RCIV specification uses the corresponding recentered exposure variables as instruments. The outcome variables indicate whether both members of each pair shared characteristics in terms of gender, nationality, and pre-MBA industry affiliation; whether both individuals had (quantitative) GMAT scores above or below the cohort median; whether both were above or below the average cohort age; whether both shared native English status; whether both possessed work experience above or below the cohort median; and whether both were assigned to the same study group. Independent variables are different levels of exposure, where "Adjacent in At" denotes pair (i, j) being adjacent at least once in or before seating chart t for $t = \{1, 2, 3\}$. *p<0.1; **p<0.05; ***p<0.01.

Table A1 tests balance only for the variables used to stratify each allocation. Stream and study-group assignments use gender, nationality, industry, GMAT, age, native English status, and work experience; seating-chart adjacency uses only gender, nationality, and study-group membership.

A.4 Probability of Friendship: Individual Fixed Effects

Table A2 reports the same regressions as Table 4 but adds individual fixed effects. Table A2 shows that the results in Table 4 are robust to individual fixed effects.

Table A2: Probability of Friendship with Individual Fixed Effects

	Link			
	(1)	(2)	(3)	(4)
Stream	0.044***	0.026***	0.034***	0.034***
	(0.004)	(0.001)	(0.003)	(0.0009)
Study group	0.221***	0.192***	0.218***	0.222***
	(0.014)	(0.008)	(0.013)	(0.007)
Adjacent	0.047***	0.047***	0.046***	0.047***
	(0.014)	(0.006)	(0.013)	(0.005)
Stream × Trait		0.036***	0.051***	0.145***
		(0.002)	(0.008)	(0.007)
Study group × Trait		0.070***	0.023	0.313***
		(0.014)	(0.038)	(0.061)
Adj × Trait		-0.0003	0.004	-0.033
		(0.014)	(0.030)	(0.039)
Trait		0.011***	0.039***	0.112***
		(0.0006)	(0.003)	(0.002)
Trait	None	Gender	Ethnicity	Similarity
Observations	680,898	680,898	680,898	680,898
Controls	Individual	Individual	Individual	Individual
Cluster	Pair	Pair	Pair	Pair

Notes. RCIV estimates of exposure on friendship and friendship composition. Each specification includes individual fixed effects. Standard errors clustered at the pair level are reported in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

A.5 Probability of Friendship: Extending Table 4 using gender and ethnicity jointly

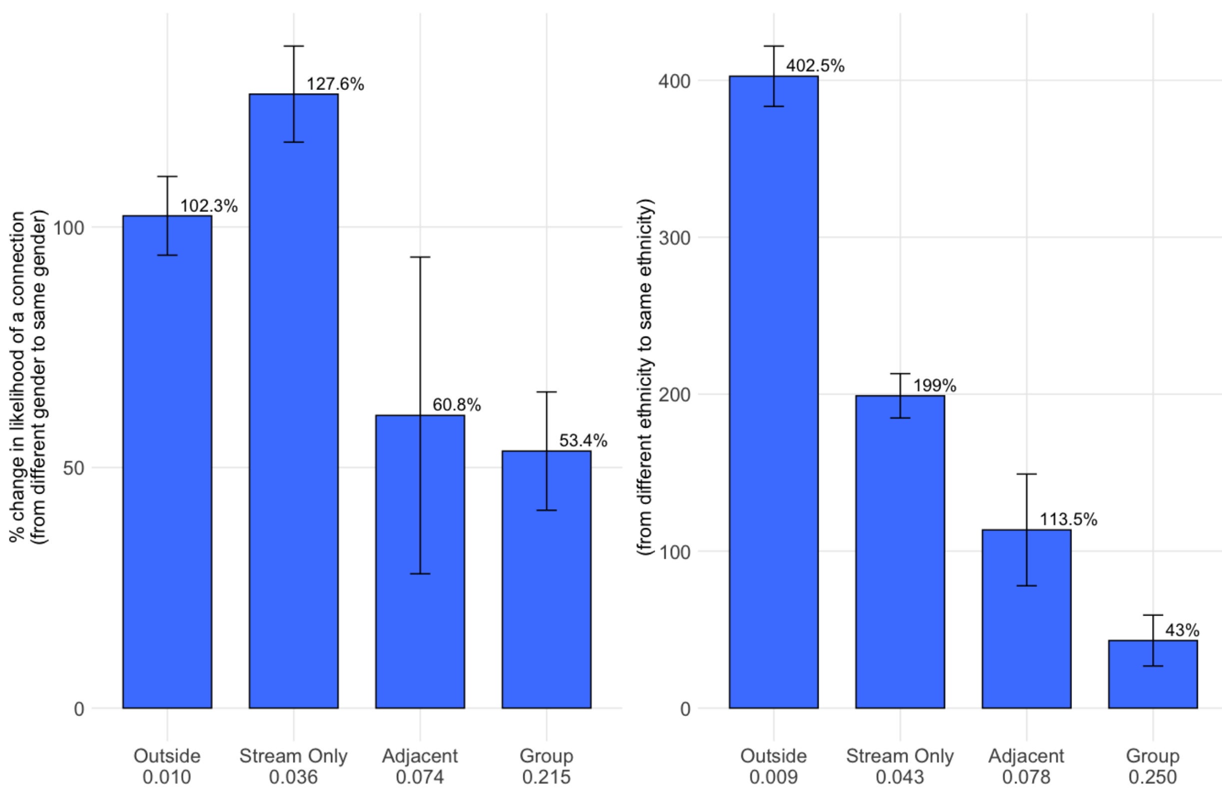
Table A3: Probability of Friendship Using Gender and Ethnicity Jointly

	Link		
	(1)	(2)	(3)
Stream	0.026***	0.034***	0.016***
	(0.001)	(0.0009)	(0.001)
Study group	0.192***	0.218***	0.189***
	(0.008)	(0.007)	(0.008)
Adjacent	0.047***	0.046***	0.047***
	(0.005)	(0.005)	(0.005)
Stream × Gender	0.036***		0.035***
	(0.002)		(0.002)
Study group × Gender	0.070***		0.069***
	(0.014)		(0.014)
Adj × Gender	0.0001		-0.0007
	(0.013)		(0.013)
Gender	0.011***		0.011***
	(0.0004)		(0.0004)
Stream × Ethnicity		0.051***	0.051***
		(0.003)	(0.003)
Study group × Ethnicity		0.023	0.022
		(0.021)	(0.021)
Adj × Ethnicity		0.003	0.002
		(0.014)	(0.014)
Ethnicity		0.035***	0.035***
		(0.0008)	(0.0008)
Trait	Gender	Ethnicity	Gender and Ethnicity
Baseline	0.01	0.009	0.005
Observations	680,898	680,898	680,898
Cluster	Pair	Pair	Pair

Notes. RCIV estimates of exposure on friendship and friendship composition. Each specification includes survey × cohort fixed effects. Standard errors clustered at the pair level are reported in parentheses. Columns (1)–(2) are equivalent to columns (2)–(3) of Table 4. Baseline probability is defined for each specification as follows: (1) probability of friendship for a different-gender pair outside the stream, (2) probability of friendship for a different-ethnicity pair outside the stream, (3) probability of friendship for a pair that differs in both gender and ethnicity outside the stream. $^{*}p < 0.1$;$^{**}p < 0.05$;$^{***}p < 0.01$.

Columns (1) and (2) of Table A3 are the same as columns (2) and (3) of Table 4 and are reported for convenience. Column (3) of Table A3 considers a specification in which the traits Gender and Ethnicity and their respective interactions with the exposure variables enter together. We observe that the point estimates for Gender and Ethnicity and their interactions with the exposure variables in column (3) are (essentially) the same as the respective estimates in column (1) and column (2). This is reassuring in view of the discussion of heterogeneous effects in Section 2.4 and Appendix A.1.3.

Figure A1: Homophily Premium Based on Column (3) of Table A3



Notes. Panels A (B) show the percentage change in the likelihood of forming a friendship between a pair with the same gender (ethnicity) and a pair with different genders (ethnicities), across all peer groups. The baseline probabilities (reported under each bar) are the predicted probabilities of friendship between individuals of different genders or ethnicities within each peer group. The probability increases for same-gender and same-ethnicity friendships are taken from specification (3) of Table A3.

A.6 Similarity Index

To construct the similarity index, we first define pair-level similarity on all available individual baseline demographics: gender, ethnicity, geographic region, industry, age, years of work experience, GMAT score, self-reported income, and socioeconomic background. We regress friendship (i.e., Y_{ij}) on all pair-level similarities and pairwise interactions, and then use a Lasso procedure to select the relevant predictors.

The Lasso output shortlists the following variables, listed here in decreasing order of importance (Lasso logit coefficients in parentheses): region (0.756), ethnicity (0.422), gender $\times$ region (0.396), gender $\times$ ethnicity (0.19), gender (0.184), region $\times$ age (0.153), region $\times$ work exp (0.074), ethnicity $\times$ GMAT (0.04), region $\times$ industry (0.038), ethnicity $\times$ income (0.017), and region $\times$ income (0.004). Similarity is then estimated using the predicted value from the specification in Table A4, normalized so that *Similarity* $\in [0, 1]$.

Table A4: Construction of Similarity

	Link			Link	
Variable	Coef.	S.E.	Variable	Coef.	S.E.
Constant	0.008^{***}	(0.0005)	Ethnicity $\times$ GMAT	0.008^{***}	(0.002)
Gender	0.008^{***}	(0.0004)	Ethnicity $\times$ Income	0.006^{***}	(0.002)
Ethnicity	0.009^{***}	(0.002)	Region $\times$ Gender	0.034^{***}	(0.002)
Region	0.013^{***}	(0.002)	Region $\times$ Industry	0.010^{***}	(0.002)
Ethnicity $\times$ Gender	0.014^{***}	(0.002)	Region $\times$ Age	0.014^{***}	(0.002)
Region $\times$ Work Exp	0.009^{***}	(0.002)	Region $\times$ Income	0.003	(0.002)
Observations	680,898		Cluster	Pair	

Notes. This table reports the estimates used to construct the similarity index. The similarity index is calculated by first obtaining the predicted value from the specification above and then normalizing it so that *Similarity* $\in [0, 1]$. Standard errors, clustered at the pair level, are reported in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

A.7 Inbreeding Female and Male Homophily

We now compute predicted male and female inbreeding homophily. Female homophily is

$$H_F = \frac{K_{FF}}{K_{FF} + K_{FM}}.$$

Under random matching, the composition of female friendships would coincide with the share of females in the relevant peer group. Let this benchmark be denoted by R_F.[40] Female inbreeding homophily is then

$$IH_F = \frac{H_F - R_F}{1 - R_F}. \tag{A8}$$

This index is equal to zero under random matching and is positive whenever female-female friendships are overrepresented relative to the population share of female peers. Male inbreeding homophily is defined analogously.

Table A5 reports the regressions used to compute predicted male and female inbreeding homophily. Columns (1) and (2) are used to compute, respectively, the predicted male and female values reported in Table A6 below.

Table A5: Regressions for Estimating Inbreeding Homophily

	Link	
	(1)	(2)
Stream	0.030^{***}	0.030^{***}
	(0.0005)	(0.0001)
Study group	0.188^{***}	0.188^{***}
	(0.008)	(0.008)
Stream × Trait	0.030^{***}	0.039^{***}
	(0.002)	(0.003)
Study group × Trait	0.075^{***}	0.067^{***}
	(0.016)	(0.021)
Trait	0.010^{***}	0.011^{***}
	(0.0005)	(0.0006)
Trait	Gender	Gender
Observations	551,713	464,780
Sample	At least one male	At least one female
Cluster	Pair	Pair

Notes. The outcome is an indicator equal to one if a pair of students are friends. Stream equals one if the pair belongs to the same stream, and Study group equals one if the pair belongs to the same study group. Trait equals one if the two individuals in the pair share the trait indicated in the row "Trait." Column (1) restricts the sample to pairs with at least one male, and column (2) restricts the sample to pairs with at least one female. Standard errors, clustered at the pair level, are reported in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

[40] Formally, $R_t = \sum_{ij \in L} \mathbf{1}\{t_i = t_j = t\} / \sum_{ij \in L} (\mathbf{1}\{t_i = t, t_j = F\} + \mathbf{1}\{t_i = t, t_j = M\})$.

Table A6: Male and Female Inbreeding Homophily from Raw Data and Regression Estimates

	Inbreeding Homophily							
	Overall		Outside Stream		Stream not group		Study group	
	Raw	Pred	Raw	Pred	Raw	Pred	Raw	Pred
Male	0.31	0.33	0.35	0.35	0.35	0.38	0.17	0.22
Female	0.29	0.31	0.32	0.33	0.34	0.36	0.15	0.16

Notes. "Raw" reports male and female inbreeding homophily computed directly from the observed network using the definitions in Section A.7. "Pred" reports predicted inbreeding homophily implied by the regression estimates of columns (1) and (2) in Table A5 for male and female homophily, respectively.

A.8 Inbreeding Homophily for Lasting Friendships

Table A7 reports the regressions used to estimate predicted inbreeding homophily for lasting friendships. In particular, columns (1) and (4) are used to compute the gender and ethnicity values reported in Table 7, respectively. Columns (2) and (3) are used to compute, respectively, the male and female values reported in Table A8 below. A lasting friendship is defined as a friendship between a pair $\{i, j\}$ present in survey 2 and remaining through the last observed survey, i.e., for cohort c with last observed survey S_c, $Y_{ijc} = \prod_{s=2}^{S_c} Y_{ijcs}$.

Table A7: Regressions for Predicted Inbreeding Homophily of Lasting Friendships

	Lasting Connection			
	(1)	(2)	(3)	(4)
Stream	0.014***	0.014***	0.014***	0.024***
	(0.002)	(0.001)	(0.001)	(0.001)
Study group	0.088***	0.088***	0.088***	0.107***
	(0.005)	(0.005)	(0.005)	(0.004)
Stream × Trait	0.036***	0.034***	0.039***	0.040***
	(0.002)	(0.002)	(0.003)	(0.003)
Study group × Trait	0.044***	0.051***	0.030***	0.007
	(0.008)	(0.009)	(0.011)	(0.012)
Trait	0.006***	0.006***	0.007***	0.021***
	(0.0009)	(0.0010)	(0.001)	(0.001)
Trait	Gender	Gender	Gender	Ethnicity
Sample	All	At least one male	At least one female	All
Observations	86,670	70,703	58,587	86,670

Notes. Columns (1)–(3) use gender as the relevant trait: column (1) includes all pairs, column (2) restricts the sample to pairs with at least one male, and column (3) restricts the sample to pairs with at least one female. Column (4) uses ethnicity as the relevant trait and includes all pairs. IID standard errors are reported in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A8: Male and Female Inbreeding Homophily of Lasting Friendships

	Inbreeding Homophily							
	Overall		Outside Stream		Stream not group		Study group	
	Raw	Pred	Raw	Pred	Raw	Pred	Raw	Pred
Male	0.44	0.45	0.4	0.39	0.56	0.56	0.23	0.31
Female	0.42	0.42	0.39	0.39	0.53	0.53	0.19	0.21

Notes. "Raw" reports male and female inbreeding homophily of lasting connections computed directly from the observed network using the definitions in Section A.7. "Pred" reports predicted inbreeding homophily implied by columns (2) and (3) of Table A7 for male and female homophily, respectively.

A.9 Detectable and Less Detectable Traits

Table A9: Probability of Friendship – Traits Different from Gender and Ethnicity

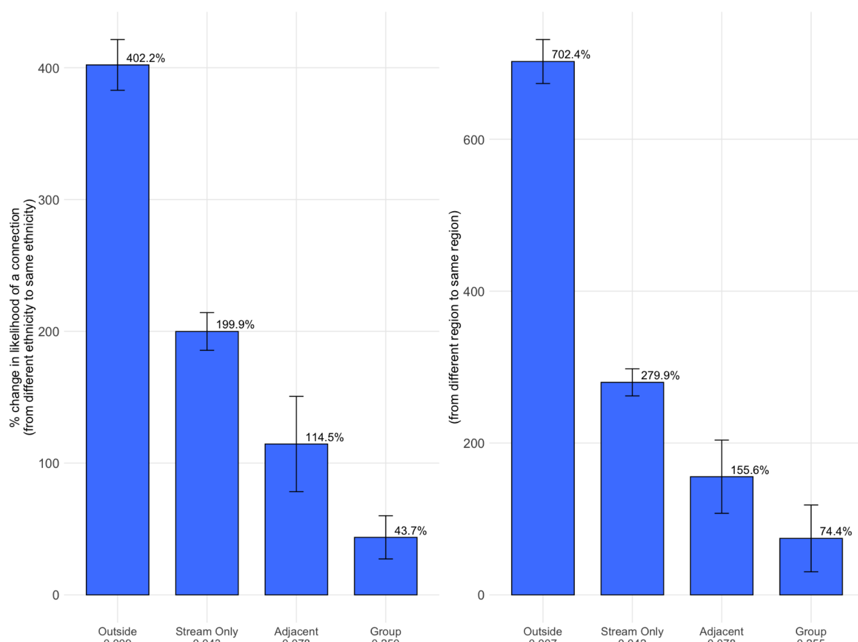
	Link					
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.006***	0.013***	0.012***	0.013***	0.013***	0.012***
	(0.0004)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Stream	0.034***	0.044***	0.041***	0.041***	0.043***	0.043***
	(0.0009)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Study group	0.219***	0.221***	0.227***	0.227***	0.216***	0.221***
	(0.007)	(0.008)	(0.009)	(0.009)	(0.009)	(0.010)
Adjacent	0.046***	0.043***	0.050***	0.052***	0.046***	0.049***
	(0.005)	(0.006)	(0.007)	(0.007)	(0.006)	(0.008)
Stream × Trait	0.067***	0.001	0.006***	0.006***	0.003	0.002
	(0.004)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Study group × Trait	0.073	-0.004	-0.013	-0.013	0.010	-0.001
	(0.057)	(0.016)	(0.013)	(0.013)	(0.013)	(0.013)
Adjacent × Trait	0.004	0.010	-0.006	-0.011	0.001	-0.004
	(0.019)	(0.010)	(0.009)	(0.009)	(0.009)	(0.010)
Trait	0.051***	0.004***	0.005***	0.003***	0.003***	0.004***
	(0.001)	(0.0005)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Trait	Region	Industry	Age	Work Exp	GMAT	Income
Baseline	0.007	0.014	0.013	0.014	0.014	0.013
Observations	680,898	680,898	680,898	680,898	680,898	680,898
Cluster	Pair	Pair	Pair	Pair	Pair	Pair
	(7)	(8)	(9)	(10)	(11)	(12)
Constant	0.014***	0.014***	0.014***	0.014***	0.013***	0.013***
	(0.0005)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Stream	0.045***	0.041***	0.044***	0.042***	0.042***	0.044***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Study group	0.232***	0.226***	0.227***	0.224***	0.219***	0.203***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.009)
Adjacent	0.046***	0.053***	0.037***	0.039***	0.048***	0.044***
	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)
Stream × Trait	-0.002	0.008***	0.001	0.005**	0.006***	0.0009
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Study group × Trait	-0.022*	-0.015	-0.019	-0.010	-0.002	0.032**
	(0.013)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Adjacent × Trait	0.002	-0.017*	0.015	0.012	-0.007	0.001
	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Trait	0.001***	0.001**	0.0007	0.001***	0.002***	0.002***
	(0.0004)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Trait	SEB	O	C	E	A	N
Baseline	0.015	0.015	0.015	0.015	0.015	0.015
Observations	680,898	627,991	627,991	627,991	627,991	627,991
Cluster	Pair	Pair	Pair	Pair	Pair	Pair

Notes. RCIV estimates of exposure on friendship and friendship composition. Each specification includes survey × cohort fixed effects. Standard errors, clustered at the pair level, are reported in parentheses. Baseline probabilities are the average friendship rates for pairs in different streams who do not share the corresponding trait. Specifications (8)–(12) assess the Big Five traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

In the main text, we considered gender and ethnicity. Table A9 reports estimates of the probability of friendship using other traits. Column (1) considers region. We also consider the following individual baseline demographics as trait variables: industry, age, years of work experience, GMAT score, self-reported income, and socioeconomic background (SEB).

We also consider the Big Five personality traits (or the OCEAN model). This model organizes personality traits into five dimensions: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. These traits are collected through the IPIP-120 questionnaire, a set of 120 questions that assesses these five traits. This questionnaire is administered to students as part of a core course during the first week of classes. Whenever the trait is continuous, such as self-reported income or years of work experience, we consider a pair to share the trait if both students are above or below the median of that variable.

Figure A2: Homophily Premium for Ethnicity and Region



Notes. Panels A and B report the homophily premium by peer group for ethnicity and region, respectively, with 95% confidence intervals. For each trait and exposure category, the premium is defined as the percentage increase in the probability of friendship for same-trait pairs relative to different-trait pairs. The baseline probabilities (reported under each bar) are the predicted probabilities of friendship between individuals of different ethnicities or regions within each peer group. The underlying probabilities for ethnicity are computed from specification (3) of Table 4, and the probabilities for region are computed from specification (1) of Table A9.

A.10 Career, Social, Personal-Help, Close-Friendship, and Reciprocated Connections

Table A10 reports the RCIV estimates of exposure on friendship and friendship composition with respect to gender and ethnicity in different networks: the career network (columns 1 and 5), social network (columns 2 and 6), personal-help network (columns 3 and 7), and close-friendship network (columns 4 and 8).

Table A10: Probability of Friendship: Career, Social, Personal-Help, and Close-Friendship Networks

	Link							
	Career	Social	Personal help	Close friendship	Career	Social	Personal help	Close friendship
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Stream	0.016***	0.015***	0.010***	0.012***	0.021***	0.021***	0.013***	0.018***
	(0.0008)	(0.0008)	(0.0007)	(0.0008)	(0.0007)	(0.0007)	(0.0006)	(0.0007)
Study group	0.118***	0.098***	0.075***	0.076***	0.143***	0.117***	0.092***	0.099***
	(0.006)	(0.006)	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)	(0.005)
Adjacent	0.033***	0.017***	0.019***	0.019***	0.034***	0.016***	0.020***	0.020***
	(0.004)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)
Stream × Trait	0.023***	0.028***	0.018***	0.025***	0.036***	0.035***	0.026***	0.032***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.003)	(0.003)	(0.002)	(0.003)
Study group × Trait	0.067***	0.050***	0.046***	0.064***	0.018	0.008	0.019	0.026
	(0.011)	(0.011)	(0.010)	(0.011)	(0.018)	(0.017)	(0.015)	(0.017)
Adj × Trait	0.008	0.001	0.006	0.006	0.004	0.004	0.005	0.003
	(0.011)	(0.010)	(0.009)	(0.010)	(0.012)	(0.012)	(0.011)	(0.012)
Trait	0.007***	0.008***	0.005***	0.007***	0.022***	0.025***	0.018***	0.022***
	(0.0003)	(0.0004)	(0.0003)	(0.0003)	(0.0007)	(0.0007)	(0.0006)	(0.0007)
Trait	Gender	Gender	Gender	Gender	Ethnicity	Ethnicity	Ethnicity	Ethnicity
Baseline	0.006	0.007	0.005	0.006	0.005	0.006	0.004	0.005
Observations	680,898	680,898	680,898	680,898	680,898	680,898	680,898	680,898

Notes. Each specification includes survey × cohort fixed effects. Standard errors clustered at the pair level are reported in parentheses. Baseline probabilities are defined for each specification as follows: columns (1)–(4) report the probability of friendship in the respective network for a different-gender pair outside the stream, and columns (5)–(8) report the probability of friendship in the respective network for a different-ethnicity pair outside the stream. $^{*}p < 0.1$;$^{**}p < 0.05$;$^{***}p < 0.01$.

Table A11 reports RCIV estimates of exposure on friendship and friendship composition with respect to gender and ethnicity for reciprocated friendship. That is, in these regressions the outcome variable for pair $\{i, j\}$ in survey s is a dummy variable that equals one if i nominates j and j nominates i in any of the sociometric questions.

Table A11: Probability of Friendship: Reciprocated Friendship

	Reciprocated Link	
	(1)	(2)
Stream	0.008***	0.012***
	(0.0006)	(0.0005)
Study group	0.046***	0.060***
	(0.004)	(0.004)
Adjacent	0.011***	0.012***
	(0.003)	(0.003)
Stream × Trait	0.018***	0.025***
	(0.001)	(0.002)
Study group × Trait	0.043***	0.027*
	(0.009)	(0.015)
Adj × Trait	0.008	0.0005
	(0.009)	(0.010)
Trait	0.005***	0.016***
	(0.0003)	(0.0006)
Trait	Gender	Ethnicity
Baseline	0.004	0.003
Observations	680,898	680,898

Notes. Each specification includes survey × cohort fixed effects. Standard errors clustered at the pair level are reported in parentheses. Baseline probability is defined for each specification as follows: (1) probability of reciprocated friendship for a different-gender pair outside the stream, (2) probability of reciprocated friendship for a different-ethnicity pair outside the stream. $^{*}p < 0.1;^{**} p < 0.05;^{***} p < 0.01$.

Table A12: Homophily Premium in Career, Social, Personal-Help, Close-Friendship, and Reciprocated Connections

	Overall	Career	Social	Personal-help	Close friendship	Reciprocated
Panel A: Gender						
Outside stream	101.8	113.6	114.7	118.0	124.1	130.6
Stream only	128.8	137.4	168.7	168.3	183.7	203.0
Adjacent	62.5	77.5	110.6	102.5	111.8	163.0
Study group	53.9	73.5	77.0	84.4	109.2	126.5
Panel B: Ethnicity						
Outside stream	402.2	450.0	450.3	502.3	469.4	511.1
Stream only	199.9	220.0	219.6	253.5	230.6	269.4
Adjacent	114.5	116.8	177.9	152.6	152.9	179.3
Study group	43.7	46.7	49.8	61.3	68.9	95.5

Notes. Entries report the homophily premium, in percent, by peer group and network type. For career, social, personal-help, and close-friendship connections, the homophily premium is derived using the estimates in Table A10. For reciprocated connections, the homophily premium is derived using the estimates in Table A11.

B Appendix to Section 4: Mechanisms and Decomposition

Appendix B.1 reports the survey-based validation exercises. Appendix B.2 reports the baseline GMM estimates and decompositions. Appendices B.3 and B.3.1 report the MBA 2027 survey-moment specifications and lasting-connection GMM exercises. Appendix B.4 reports the preference-heterogeneity specification.

B.1 Reduced-Form Evidence on Mechanisms from Survey Questions

Table B1: Correlation Between Same-Trait Directed Search and Same-Trait Preference for Friendship

	Choose Same Trait	
	(1)	(2)
Male	-0.103**	0.118**
	(0.047)	(0.057)
Preference for same trait − Preference for different trait	0.401***	0.433***
	(0.116)	(0.143)
Trait	Gender	Region
Observations	211	117

Notes. Column (1) uses gender and column (2) uses region. The sample includes respondents to MBA 2027 Survey 2 who saw one same-trait and one different-trait profile. Both specifications control for respondent gender, respondent region, and the sum of stated preferences for the two profiles. IID standard errors are reported in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table B2: Correlation Between Inbreeding Homophily and Same-Gender Directed Search and Same-Gender Preference for Friendship

	Inbreeding Homophily					
	Study group		Stream not group		Outside Stream	
	(1)	(2)	(3)	(4)	(5)	(6)
Choose same gender	0.037		0.115***		0.107***	
	(0.047)		(0.037)		(0.032)	
Preference		-0.047		-0.087		-0.160
		(0.158)		(0.120)		(0.102)
Preference × Same gender		0.383*		-0.169		0.297**
		(0.226)		(0.156)		(0.143)
Observations	1,146	1,146	1,710	1,710	1,702	1,702
Cluster	Individual	Individual	Individual	Individual	Individual	Individual

Notes. The sample consists of respondents to MBA 2027 Survey 2. Choose same gender is a binary variable that is 1 if the respondent chooses a same-gender profile in the question intended to capture directed search. Preference is the respondent's reported preference in the question intended to capture preferences for same-trait interaction. Same gender is a binary variable that is 1 if the profile presented to the respondent in the question about preference has the same gender as the respondent. Standard errors, clustered at the individual level, are reported in parentheses. $^{*}p < 0.1$;$^{**}p < 0.05$;$^{***}p < 0.01$.

Columns (1), (3), and (5) of Table B2 show that there is a positive association between individuals who choose the same-gender profile in the directed search question and their inbreeding homophily outside their study group, but this association is not present within their study group. This pattern is consistent with our hypothesis that study groups mute directed search: choosing the same-gender profile is not associated with homophily within the study group, whereas it is positively associated with homophily outside the study group.

Columns (2), (4), and (6) attempt to assess the external validity of the preference measure. We regress an individual's inbreeding homophily in a given subnetwork on the respondent's stated likelihood of wanting to befriend a candidate profile and its interaction with an indicator for whether the profile has the respondent's same gender. Formally, each observation is a respondent–profile pair, so each respondent contributes two observations, one for each randomized profile. The coefficient on *Preference* captures the association between stated preference and same-gender homophily when the evaluated profile is of the opposite gender, while the coefficient on *Preference* × *Same* *gender* captures how that association changes when the evaluated profile is of the same gender. Because profile characteristics are randomized, identification comes both from within-respondent variation—comparing the same respondent's evaluations of two different profiles—and from across-respondent variation, since some respondents see one same-gender and one different-gender profile, while others see two same-gender profiles. The estimates in columns (2), (4), and (6) therefore ask whether respondents who express a stronger relative preference for same-gender profiles also exhibit higher same-gender homophily in the observed network.

In the study-group and outside-stream subnetworks, the interaction term is positive, implying that stronger stated preference for a same-gender profile is associated with greater same-gender homophily when the evaluated profile is of the same gender; by contrast, the corresponding estimate in the stream-but-not-group subnetwork is imprecise and does not show the same pattern.

B.2 GMM: Estimates and Decomposition

Table B3: GMM Estimates in the OR Network

Estimates	Gender	Ethnicity	Region
m_{ST}	0.07^{***}	0.12^{***}	0.13^{***}
m_{DT}	0.05^{***}	0.05^{***}	0.04^{***}
Γ_{ST}	0.57^{***}	0.56^{***}	0.60^{***}
Γ_{DT}	0.47^{***}	0.50^{***}	0.53^{***}

Notes. The table reports GMM estimates in the core sample for the OR network based on the four moments (11)–(14). The parameters are estimated separately for gender, ethnicity, and region using four moments: the share of possible same-trait and different-trait links formed within L^{obs}_{Stream}, and the share of possible same-trait and different-trait links formed within L^{obs}_{Group}. Inference uses the robust asymptotic GMM covariance matrix described in Section 4.3. $^{*}p < 0.1$;$^{**}p < 0.05$;$^{***}p < 0.01$.

B.3 GMM: Estimates and Decomposition in Alternative Specifications

For the MBA 2027 cohort, we also observe survey-based measures of preferences and directed-search intensity for same-gender and same-region peers from the same stream. These additional measures allow us to introduce two further sets of moments.

Preferences survey. Our second set of moments concerns preferences for same-trait and different-trait friendships, which are captured in the model by the parameters Γ_{ST} and Γ_{DT}. For the data analog, we define $\Gamma^{\text{obs}}_{iST}$ to be student i's self-reported probability, elicited in the survey, of becoming friends with a profile that has the same trait as i. When the trait is gender, each individual faces such a choice. However, when the trait is region, some individuals never face a profile that shares their region. We therefore denote by $\mathcal{N}^{obs,Pref}_{ST}$ the set of individuals for whom this variable is observed, and let $n^{obs,Pref}_{ST} = |\mathcal{N}^{obs,Pref}_{ST}|$. The quantities $\Gamma^{\text{obs}}_{iDT}$, $\mathcal{N}^{obs,Pref}_{DT}$, and $n^{obs,Pref}_{DT}$ are defined analogously. Our preference moments therefore take the form

$$\begin{aligned} \frac{1}{n^{obs,Pref}_{ST}} \sum_{i \in \mathcal{N}^{obs,Pref}_{ST}} \Gamma^{obs}_{i,ST} &= \Gamma_{ST}, \\ \frac{1}{n^{obs,Pref}_{DT}} \sum_{i \in \mathcal{N}^{obs,Pref}_{DT}} \Gamma^{obs}_{i,DT} &= \Gamma_{DT}. \end{aligned} \tag{B1}$$

Directed search survey. Our third set of moments concerns directed search. In the model, among students with trait value t, the relative probability of choosing a profile with trait value t is $m_{ST}/(m_{ST} + m_{DT})$. This normalization is necessary to map the model to the data, where students were forced to choose one of two profiles, so that the empirical probabilities add up to one. Let $\mathcal{N}^{obs,Search}$ be the set of individuals i who, in the survey, faced one profile sharing their trait and another profile with a different trait, and let $n^{obs,Search} = |\mathcal{N}^{obs,Search}|$. For each $i \in \mathcal{N}^{obs,Search}$, denote by $m^{obs}_{i,ST}$ the probability that student i chooses the same-trait profile in the directed-search survey question. Our search moment is thus

$$\frac{1}{n^{obs,Search}} \sum_{i \in \mathcal{N}^{obs,Search}} m^{obs}_{i,ST} = \frac{m_{ST}}{m_{ST} + m_{DT}}. \tag{B2}$$

The directed-search moment for different-trait profiles is linearly dependent on the same-trait moment and is therefore omitted.

Table B4 provides estimates of the parameter vector θ for different sets of moment conditions.

Table B4: GMM Estimates by Moment Specification

	Gender (1)	Gender (2)	Gender (3)	Gender (4)	Gender (5)	Region (1)	Region (2)	Region (3)	Region (4)	Region (5)
Parameter										
m_{ST}	0.07***	0.05***	0.05***	0.07***	0.05***	0.24***	0.09***	0.09***	0.14***	0.10***
m_{DT}	0.06***	0.04***	0.04***	0.06***	0.04***	0.05***	0.03***	0.03***	0.07***	0.03***
Γ_{ST}	0.57***	0.65***	0.64***	0.56***	0.64***	0.50***	0.71***	0.70***	0.61***	0.68***
Γ_{DT}	0.48***	0.61***	0.58***	0.50***	0.58***	0.48***	0.62***	0.61***	0.43***	0.61***
Moment conditions used										
Stream links	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Study-group links	✓	×	✓	✓	✓	✓	×	✓	✓	✓
Preference survey	×	✓	✓	×	✓	×	✓	✓	×	✓
Directed-search survey	×	✓	✓	✓	×	×	✓	✓	✓	×
J-test	0	1.226	39.374	1.86	39.478	0	13.592	31.749	4.511	17.862
p-value	NA	0.268	< 0.001	0.173	< 0.001	NA	< 0.001	< 0.001	0.034	< 0.001

Notes. The table reports GMM estimates for the OR network for Survey 2 of the MBA 2027 cohort. The parameters are estimated separately for gender and region. The moments used in each of the five specifications are listed in the lower panel. Inference uses the robust asymptotic GMM covariance matrix described in Section 4.3. $^{*}p < 0.1;^{**} p < 0.05;^{***} p < 0.01$.

Table B5: Gender and Region Decomposition by Moment Specification

	(1)		(2)		(3)		(4)		(5)	
Channel	Contribution	%	Contribution	%	Contribution	%	Contribution	%	Contribution	%
Gender										
Club peer composition	0.003	1.3	0.003	1.4	0.003	1.4	0.003	1.3	0.003	1.4
	[0.003, 0.003]	[1.1, 1.6]	[0.003, 0.003]	[1.1, 1.7]	[0.003, 0.003]	[1.1, 1.7]	[0.003, 0.003]	[1.0, 1.6]	[0.003, 0.003]	[1.1, 1.7]
Preferences	0.165	74.0	0.054	25.4	0.087	39.7	0.117	51.1	0.086	39.1
	[0.079, 0.250]	[34.7, 113.2]	[0.019, 0.089]	[10.3, 40.6]	[0.055, 0.119]	[26.1, 53.0]	[0.066, 0.168]	[32.0, 70.3]	[0.048, 0.124]	[20.9, 57.3]
Directed search	0.055	24.7	0.156	73.2	0.129	58.9	0.109	47.6	0.131	59.5
	[−0.036, 0.146]	[−14.6, 64.1]	[0.112, 0.200]	[58.0, 88.3]	[0.088, 0.171]	[45.6, 72.6]	[0.061, 0.156]	[28.4, 66.6]	[0.075, 0.187]	[41.2, 77.8]
Total	**0.223**	**100.0**	**0.213**	**100.0**	**0.219**	**100.0**	**0.229**	**100.0**	**0.220**	**100.0**
Region										
Club peer composition	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0	0.000	0.0
	[−0.000, −0.000]	[−0.1, −0.1]	[−0.000, −0.000]	[−0.1, −0.1]	[−0.000, −0.000]	[−0.1, −0.1]	[−0.000, −0.000]	[−0.1, −0.1]	[−0.000, −0.000]	[−0.1, −0.1]
Preferences	0.012	3.6	0.056	18.3	0.060	19.9	0.155	47.3	0.041	12.4
	[−0.112, 0.137]	[−33.6, 41.2]	[0.039, 0.074]	[13.1, 24.0]	[0.043, 0.078]	[14.4, 25.4]	[0.097, 0.213]	[31.7, 62.9]	[0.022, 0.061]	[6.5, 18.5]
Directed search	0.321	96.4	0.250	81.7	0.242	80.1	0.173	52.7	0.290	87.6
	[0.194, 0.447]	[58.9, 133.7]	[0.218, 0.281]	[76.1, 87.0]	[0.211, 0.273]	[74.7, 85.7]	[0.124, 0.223]	[37.2, 68.5]	[0.250, 0.329]	[81.6, 93.6]
Total	**0.333**	**100.0**	**0.306**	**100.0**	**0.302**	**100.0**	**0.328**	**100.0**	**0.331**	**100.0**
Moment conditions used										
Stream links	✓		✓		✓		✓		✓	
Study-group links	✓		×		✓		✓		✓	
Preference survey	×		✓		✓		×		✓	
Directed-search survey	×		✓		✓		✓		×	

Notes. The table decomposes model-implied gender and region inbreeding homophily in the OR network for Survey 2 of the MBA 2027 cohort. Contributions are measured incrementally as in Table 10. Percentage contributions are computed by dividing each channel's contribution by total model-implied homophily. 95% confidence intervals for the channel rows are reported in brackets below the corresponding estimates. These intervals are computed by the delta method, treating each reported decomposition quantity as a function of the four estimated GMM parameters and using the full two-step GMM covariance matrix. The final panel reports the moment conditions used in each specification.

B.3.1 GMM for Lasting Connections: Estimates and Decomposition

Table B6: GMM Estimates for Lasting Connections

Estimates	Gender	Ethnicity	Region
m_{ST}	0.10^{***}	0.17^{***}	0.14^{**}
m_{DT}	0.04^{***}	0.05^{***}	0.05^{***}
Γ_{ST}	0.41^{***}	0.39^{***}	0.48^{***}
Γ_{DT}	0.32^{***}	0.37^{***}	0.36^{***}

Notes. The table reports GMM estimates in the core sample for lasting connections based on the four moments (11)–(14). The parameters are estimated separately for gender, ethnicity, and region using four moments: the share of possible same-trait and different-trait links formed within L^{obs}_{Stream}, and the share of possible same-trait and different-trait links formed within L^{obs}_{Group}. The sample includes the MBA 2025 and MBA 2026 cohorts. Lasting connections are defined as friendships present in survey 2 and remaining through the last observed survey. Inference uses the robust asymptotic GMM covariance matrix described in Section 4.3. $^{*}p < 0.1;^{**}p < 0.05;^{***}p < 0.01$.

Table B7: Decomposition of Inbreeding Homophily: Lasting Connections

	Gender		Ethnicity		Region	
Channel	Contribution	%	Contribution	%	Contribution	%
Club peer composition	0.002	0.4	0.001	0.3	0.003	0.7
	$[0.002, 0.002]$	$[0.3, 0.4]$	$[0.001, 0.001]$	$[0.2, 0.3]$	$[0.002, 0.003]$	$[0.6, 0.7]$
Preferences	0.226	41.5	0.020	6.5	0.118	29.4
	$[0.091, 0.361]$	$[16.5, 66.7]$	$[-0.075, 0.115]$	$[-24.2, 37.1]$	$[-0.074, 0.311]$	$[-18.7, 77.7]$
Directed search	0.316	58.1	0.289	93.2	0.280	69.8
	$[0.172, 0.459]$	$[33.0, 83.1]$	$[0.186, 0.392]$	$[62.7, 123.9]$	$[0.082, 0.478]$	$[21.6, 118.1]$
Total	**0.544**	**100.0**	**0.310**	**100.0**	**0.401**	**100.0**

Notes. The table decomposes model-implied inbreeding homophily for lasting connections into the contributions of club peer composition, preferences, and directed search. Contributions are measured as the incremental effect of each channel on inbreeding homophily. Percentage contributions are computed by dividing each channel's contribution by total model-implied homophily. 95% confidence intervals for the channel rows are reported in brackets below the corresponding estimates. These intervals are computed by the delta method, treating each reported decomposition quantity as a function of the four estimated GMM parameters and using the full two-step GMM covariance matrix.

B.4 GMM: Preference Heterogeneity Without Matching Frictions

In this section, we replace Assumption 1 with the following assumption.

Assumption 2. We assume that

$$\Gamma_{ij} = \begin{cases} \Gamma_{ST}^{Stream} & \text{if } ij \in L_{Stream}^{obs} \text{ and } t_i = t_j, \\ \Gamma_{DT}^{Stream} & \text{if } ij \in L_{Stream}^{obs} \text{ and } t_i \neq t_j, \\ \Gamma_{ST}^{Group} & \text{if } ij \in L_{Group}^{obs} \text{ and } t_i = t_j, \\ \Gamma_{DT}^{Group} & \text{if } ij \in L_{Group}^{obs} \text{ and } t_i \neq t_j, \end{cases}$$

and that $m_{ij} = 1$ for all ij.

In this specification, we estimate $\theta = \{\Gamma_{ST}^{Stream}, \Gamma_{DT}^{Stream}, \Gamma_{ST}^{Group}, \Gamma_{DT}^{Group}\}$ using GMM in the core sample by matching the degree moments within the stream and the study group. Let $n_{Stream}^{obs} = |L_{Stream}^{obs}|$ and $n_{Group}^{obs} = |L_{Group}^{obs}|$ denote the cardinalities of the stream and study-group pair sets, respectively. These moments are:

$$\begin{aligned} \frac{1}{n_{Stream}^{obs}} \sum_{ij \in L_{Stream}^{obs}} Y_{ij}^{obs} \mathbf{1}\{t_i = t_j\} &= \frac{1}{n_{Stream}^{obs}} \sum_{ij \in L_{Stream}^{obs}} \left(\Gamma_{ST}^{Stream}\right)^2 \\ \frac{1}{n_{Stream}^{obs}} \sum_{ij \in L_{Stream}^{obs}} Y_{ij}^{obs} \mathbf{1}\{t_i \neq t_j\} &= \frac{1}{n_{Stream}^{obs}} \sum_{ij \in L_{Stream}^{obs}} \left(\Gamma_{DT}^{Stream}\right)^2, \\ \frac{1}{n_{Group}^{obs}} \sum_{ij \in L_{Group}^{obs}} Y_{ij}^{obs} \mathbf{1}\{t_i = t_j\} &= \frac{1}{n_{Group}^{obs}} \sum_{ij \in L_{Group}^{obs}} \left(\Gamma_{ST}^{Group}\right)^2 \\ \frac{1}{n_{Group}^{obs}} \sum_{ij \in L_{Group}^{obs}} Y_{ij}^{obs} \mathbf{1}\{t_i \neq t_j\} &= \frac{1}{n_{Group}^{obs}} \sum_{ij \in L_{Group}^{obs}} \left(\Gamma_{DT}^{Group}\right)^2. \end{aligned} \tag{B3}$$

Table B8: GMM Estimates of Preference Heterogeneity Between Study-Group and Stream Friendships

	OR network		Lasting connections	
	Gender	Region	Gender	Region
Γ_{ST}^{Stream}	0.28***	0.39***	0.25***	0.34***
Γ_{DT}^{Stream}	0.20***	0.20***	0.14***	0.16***
Γ_{ST}^{Group}	0.57***	0.60***	0.41***	0.48***
Γ_{DT}^{Group}	0.47***	0.53***	0.32***	0.36***
Preference Premium within Stream	44.0%	92.5%	82.9%	116.8%
Preference Premium within Group	20.9%	12.2%	25.6%	31.7%

Notes. The table reports GMM estimates for the core sample of preference heterogeneity between study-group and stream friendships using the four moments in (B3). The parameters are estimated separately for gender and region. The sample for the OR network includes the MBA 2025, MBA 2026, and MBA 2027 cohorts. The sample for lasting connections includes the MBA 2025 and MBA 2026 cohorts. Inference uses the robust asymptotic GMM covariance matrix described in Section 4.3. $^{*}p < 0.1;^{**}p < 0.05;^{***}p < 0.01$.

C Survey Appendix

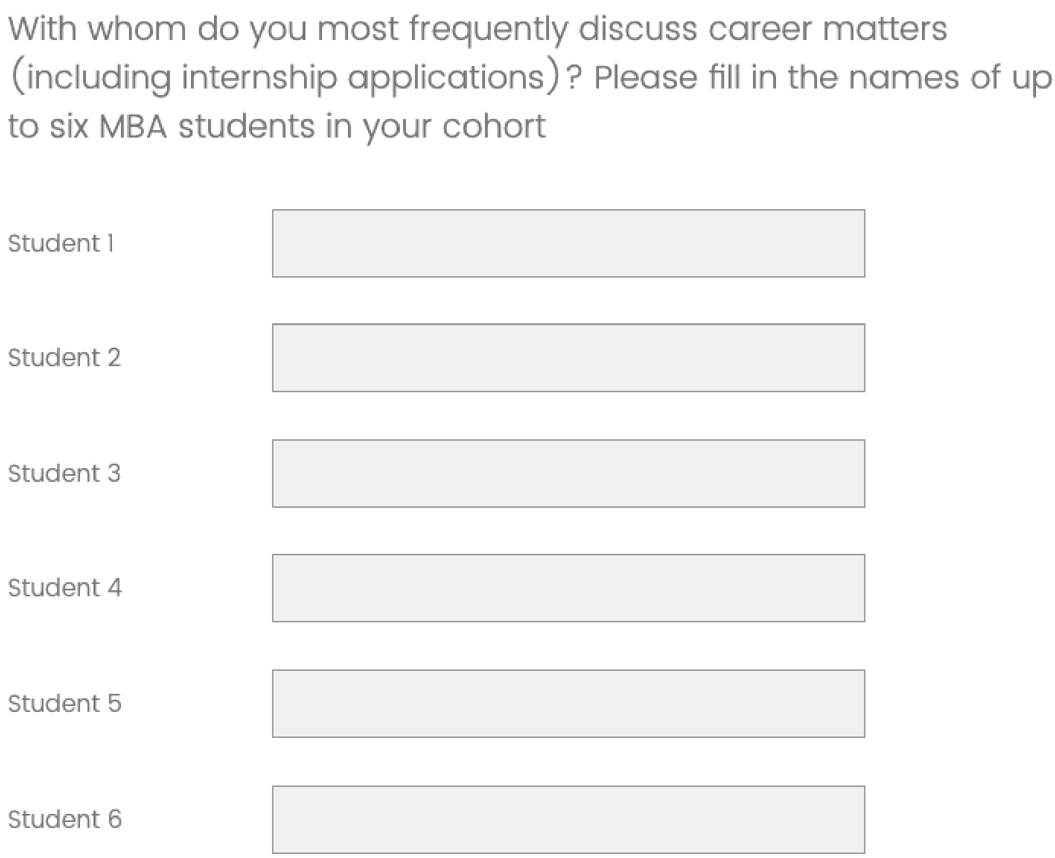
This appendix reports the survey materials in three parts: connection questions, socioeconomic background, and directed search and preferences.

Connection Questions

Figure C1: Connection Nominations: Instructions for Naming MBA Cohort Peers

We would like you to answer some questions about your social relationships with MBA students in your cohort. For each question, please type the names of up to six MBA students from your cohort. Start by typing either their first name (or known as) or their last name, then select the correct autocompleted name. You will not be able to list any names of people outside of your LBS MBA cohort.

Figure C2: Career-Network Nomination Question



With whom do you most frequently discuss career matters (including internship applications)? Please fill in the names of up to six MBA students in your cohort

Student 1

Student 2

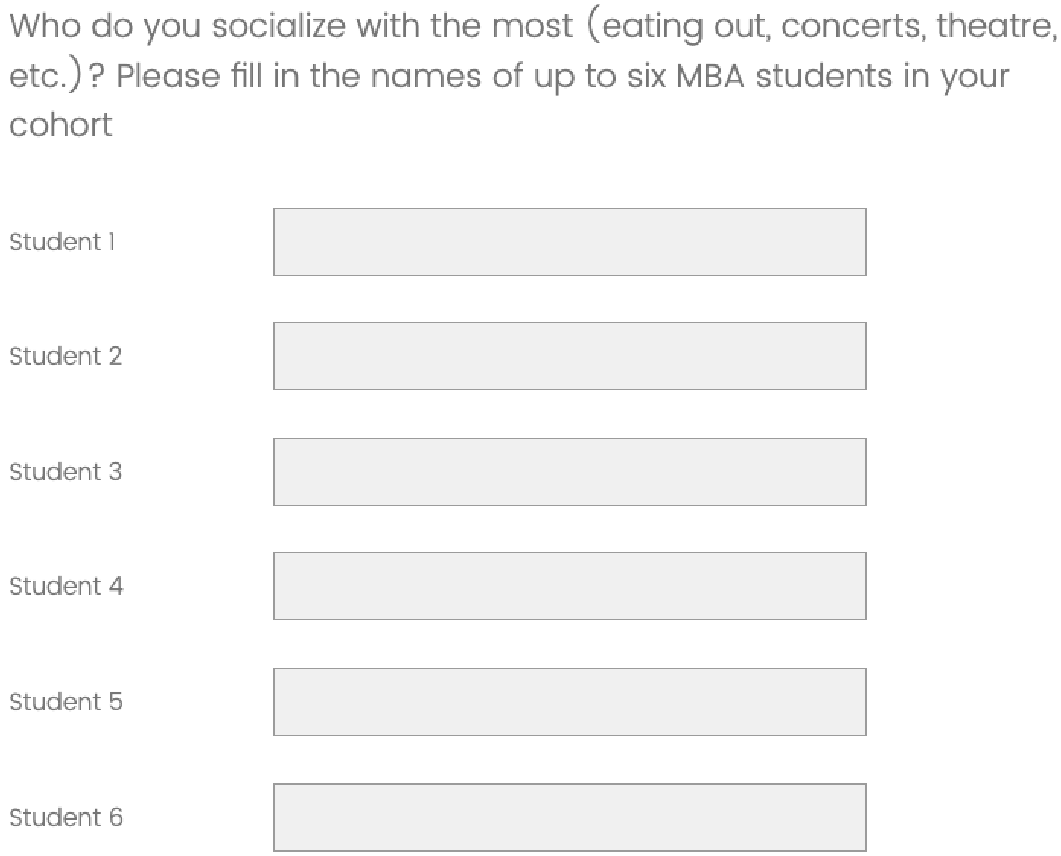
Student 3

Student 4

Student 5

Student 6

Figure C3: Social-Network Nomination Question


Who do you socialize with the most (eating out, concerts, theatre, etc.)? Please fill in the names of up to six MBA students in your cohort

Student 1

Student 2

Student 3

Student 4

Student 5

Student 6

Figure C4: Personal-Help-Network Nomination Question

Who do you turn to talk about personal problems or when you feel under pressure? Please fill in the names of up to six MBA students in your cohort

Student 1

Student 2

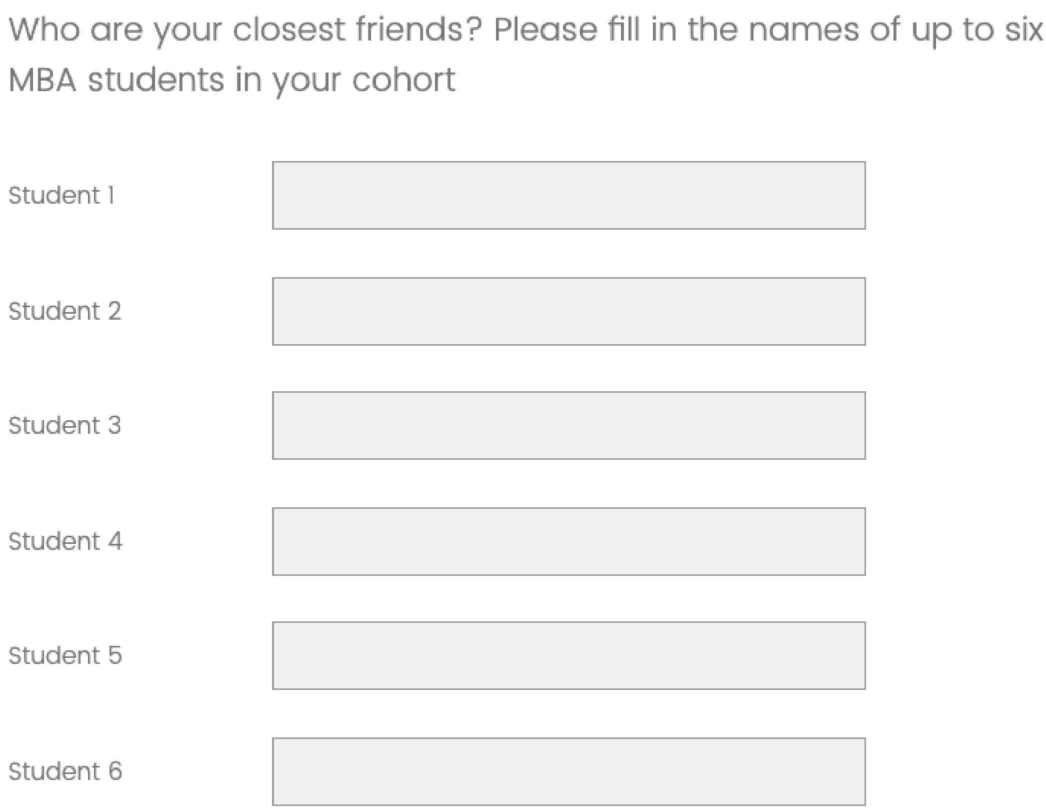
Student 3

Student 4

Student 5

Student 6

Figure C5: Close-Friendship Nomination Question



Who are your closest friends? Please fill in the names of up to six MBA students in your cohort

Student 1

Student 2

Student 3

Student 4

Student 5

Student 6

Socioeconomic Background

Figure C6: Socioeconomic-Background Ladder Question Relative to the MBA Cohort

Please imagine another ladder with steps numbered from zero at the bottom to 10 at the top. **This ladder represents where people stand within your LBS MBA cohort**. At the TOP of the ladder are the people who are the best off - those who have the most money, the most education, and the most respected jobs. At the BOTTOM are the people who are the worst off - who have the least money, least education, and the least respected jobs or no job. The higher up you are on this ladder, the closer you are to the people at the very top; the lower you are, the closer you are to the people at the very bottom. Please indicate the step (0 is the lowest, 10 is the highest) where you think YOU stand at this time in your life, relative to other people in your LBS MBA cohort.

0 1 2 3 4 5 6 7 8 9 10

Step

Directed Search and Preferences

Figure C7: Directed-Search Scenario and Profile-Choice Question

In this section, you will read about two different scenarios describing a possible interaction with some students in your stream. These scenarios take place just before the first lecture of a core course during your first term of your MBA. Please note that the students in the scenarios are hypothetical.

There are no right or wrong answers. Please choose as you would in real life.

Scenario 1

You arrive ten minutes before the lecture starts and see two MBA students from your stream standing nearby. They are both by themselves and not busy. While you recognise both of them, you have only exchanged greetings in the past. You would like to have a conversation, but you only have time to approach one of them.

Student A: A woman from North America, and you have heard that she likes music.
Student B: A man from East Asia, and you have heard that he likes travelling.

Whom do you approach?

○ Student A

○ Student B

Figure C8: Preference Question: Likelihood of Wanting to Become Friends After Repeated Interaction