Peer Diversity, College Performance and Educational Choices*

Arnaud Chevalier[ab] Ingo E. Isphording[a] Elena Lisauskaite[b]

This draft: March 2020
First draft: May 2017

Abstract

We study the effect of ethno-linguistic classroom composition on performance, educational choices and post-graduation migration of students. In the setting of a British University were assignment to seminars is plausibly exogenous, we define two measures of classroom composition: the share of non-native speakers and the linguistic diversity of the non-speaker group. Non-native speaker's performance benefits from greater linguistic diversity. Native speakers are unaffected by classroom composition. Survey information points to changed interaction patterns in the classroom as a likely mechanism. While ethno-linguistic seminar composition has no meaningful impact on subsequent course choice, the effect of initial diversity on grades persists until the final year of study. The results imply that avoiding segregation along language lines is key in providing education for an international student body.

JEL codes: I21, I24, J15
Keywords: higher education, diversity, peer effects, foreign students
Declaration of Interest: Declarations of interest: none

*We would like to thank Benjamin Elsner, Jonas Radbruch, Ulf Zölitz, the participants of the Workshop on Higher Education at the University of Essex, the Workshop on the Economics of Education and Expectations at RHUL, the 3rd IZA Workshop on the Economics of Education, the 10th International Workshop on Applied Economics of Education, the 18th IZA Transatlantic Meetings, the European Association of Labour Economists 2019 in Uppsala and two anonymous referees. Excellent research assistance was provided by Maximilian Mähr and Youpeng Zhang. A. Chevalier would also like to thank the Nuffield foundation (grant *EDU/42242*) for partially funding this work.

[a]IZA - Institute of Labor Economics
[b]Royal Holloway, University of London

Corresponding author: Ingo E. Isphording, IZA - Institute of Labor Economics, Schaumburg-Lippe-Str.5-9, 53113 Bonn, Germany, Phone: +49 3894 (204), isphording@iza.org

1 Introduction

The fast-paced internationalisation of the tertiary education system has drastically increased the number of international students from 2.0 million in 2000 to 5.3 million worldwide in 2017[1]. The UK is one of the largest recipient countries of foreign students. In 2019, it hosted about 486,000 international students, representing 25% of the student population in the country[2]. The increasingly diverse student population has sparked a scientific and political debate. Advocates of internationalisation argue that increasing diversity benefits both native and foreign students, while critics raise concerns about potential negative spillovers and native college flight. Empirical evidence on the effects of increasing numbers of foreign students is mostly restricted to primary and secondary education and points to ambiguous effects across different settings.[3] Tertiary education settings have received less attention to date.[4]

In this paper, we provide causal evidence on the effects of higher internationalisation and ethnic diversity on student grades and educational choices in a setting of exogeneous assignment to undergraduate seminars at a British university. We measure the ethnic composition of a seminar (small-scale learning group) by the share of non-native language background (labeled as the share of non-native speakers hereafter) but also by the ethnolinguistic diversity within the group of non-native speakers.

Ethno-linguistic diversity among non-native students is expected to influence student performance and choices as it changes incentives for English language use and assimilation, a mechanism that is straightforwardly derived from a model of language assimilation (Lazear, 1999). At a given share of non-native speakers, a non-native-speaking student from a smaller linguistic minority broadens her pool of potential learning partners if she communicates in English, albeit potentially at some cost. The smaller the pool of potential same-language learning partners, the higher the incentives to engage in English communication. This intuition directly maps into a theoretically-positive effect of diversity on classroom integration which in turn is expected to have positive spill-overs on the academic performance of non-native speakers. Further, language skills acquired through interaction with native peers can directly improve performance in English curricula, even in quantita-

[1] Numbers taken from UNESCO UIS database, uis.unesco.org

[2] Numbers taken from HESA database, www.hesa.ac.uk

[3] The results range between moderate negative effects on native students' performance (Ballatore et al., 2018; Brunello and Rocco, 2013; Jensen and Rasmussen, 2011; Gould et al., 2009) to zero effects (Geay et al., 2013; Ohinata and van Ours, 2013).

[4] Anelli et al. (2017) show that higher shares of foreign students in introductory maths courses reduce natives' likelihood of moving into STEM majors. Braakmann and McDonald (2018) point to ambiguous effects of diversity in UK universities depending on the level of aggregation.

tive courses. To the best of our knowledge, we are the first to study the effect of ethnic seminar diversity (beyond the simple share of immigrants) on student outcomes in a tertiary education setting.

We base our analysis on the administrative records of economics students at a university in the London Metropolitan area. This institutional environment provides a fitting "laboratory setting" for our analysis for two reasons. First, the student body that we analyse is characterised by a high degree of ethno-linguistic diversity. Over the 2006 to 2016 period, we observe on average a share of 56 percent of non-native students from 67 different non-English-speaking countries. The high degree of ethno-linguistic diversity ensures a sufficient common support for the share of non-native students and the diversity to identify the causal effect of both variables. This level of internationalisation is shared by other British institutions, and the institution that we analyse does not stand out in terms of selectivity or graduate earnings. Second, students of the economics programme are exogeneously allocated to seminars during the compulsory stage of their studies, which results in an exogenously varying ethno-linguistic composition that is consistent with an as-good-as-random assignment to seminars.

We describe three sets of results. First, grades of native speakers are unaffected by the share of non-native students and the linguistic diversity of a classroom. Non-native speakers are not penalised by higher shares of non-native students in the classroom, but benefit from a higher ethno-linguistic diversity. In a more diverse classroom, the academic performance of non-native speakers improves, especially for low-achieving students.

Second, we trace potential mechanisms through an in-class survey. Students change their pattern of classroom interaction across ethnicities in response to higher diversity. When diversity increases, non-native students become more likely to interact with native students while interaction patterns for native speakers remain unaltered. Diversity appears to counteract segregation between native and non-native speakers.

Third, in the longer run the initial share of non-native speakers increases the likelihood to choose courses that are again popular among non-native students. This segregation into courses, potentially, better suited for non-native students, might explain the persistent effect of share of non-native speakers on third year performance which is to some degree inconsistent with the contemporary results. With respect to early seminar diversity, and in line with contemporary results, though, third year grades are positively affected, suggesting some persistence of the contemporary effect of diversity on performance. Finally, a voluntary alumni survey suggests that being assigned to a larger share of non-native students in

the compulsory stage increases the likelihood of non-native speakers to be abroad at the point of interview of the post-graduation survey.

Taken together, our findings suggest that even at a high level of internationalisation, there are virtually no negative effects on the learning of native students through the exposure to non-native speakers in small-class teaching. Diversity improves the learning of non-native speakers while natives speakers' performance remains unaffected.

With this paper, we contribute to the literature on the effects of the internationalisation of education. In general, existing studies have focused on the share of foreign students in primary and secondary school classrooms. Ballatore et al. (2018) exploit rules of classroom formation to identify a sizeable negative effect of additional immigrant students in Italian primary schools, which are substantially larger for foreign rather than native students. Other studies describe small negative to zero to even slight positive effects of higher shares of foreign students using variation between cohorts or classes of the same school (Gould et al., 2009; Geay et al., 2013; Ohinata and van Ours, 2013; Figlio and Özek, 2017; Frattini and Meschi, 2017; Diette and Oyelere, 2017), regions (Jensen and Rasmussen, 2011; McHenry, 2015; Hunt, 2017) or between countries (Brunello and Rocco, 2013). With respect to diversity of primary/secondary school classrooms, Maestri (2017) finds a positive effect of more diverse classrooms on foreigners language acquisition, while Bredtmann et al. (2018) report negative correlations between diversity and the social integration of immigrants in German primary schools.

Only few studies have estimated the effects of internationalisation in tertiary education settings. Anelli et al. (2017) show that higher shares of foreign peers in introductory maths courses reduces the probability of native students moving into STEM majors, but find no direct effect on grades. Compared to our own study, their interest is on the effect of internationalisation on major choice in a US public university. In our setting, as common in European universities, students have already chosen their major at enrolment, and as such follow more comparable educational paths. Braakmann and McDonald (2018) describe a relationship between diversity at the course level and student performance for three UK-wide cohorts and deal with potentially endogenous course choices by exploiting within-programme variation across courses. They complement their main analysis with an IV strategy relying on network effects among foreign students. Their results point to ambiguous effects of diversity, depending on the level of analysis. Machin and Murphy (2017) find no evidence for a crowding out of domestic students in response to higher influx of foreign students in UK universities.

Our paper makes three contributions to these strands of literature. First, we examine the effect of diversity in higher education beyond the simple share of foreign students on a fine-grained seminar level. These seminars display relevant peer groups with meaningful social interactions. While Braakmann and McDonald (2018) demonstrate sensitivity of the level of observation with respect to student diversity, the classroom level has only been addressed by Maestri (2017) for primary school classrooms in the Netherlands and by Bredtmann et al. (2018) for German primary schools. Second, whereas the aforementioned studies have to deal with potential selection into classrooms and programs, we base our identification on a clean natural experiment relying on as-good-as-random allocation of students to small-scale seminars. This strategy alleviates many concerns about potential confounders through the selection of students into seminars. Third, we provide insights into plausible mechanisms by surveying students in the field about ethnic interactions and language use.

Our results have implications for education practitioners. Even in an environment where non-native speakers represent more than half of the students, we do not find negative effects of their share on the performance and educational choices of native students. This supports current policies of pursuing greater internationalisation in higher education and should caution against forces asking for stricter admission policies discriminating by origin. Moreover, to favour the integration of non-native speakers, diversifying early study environments by avoiding segregated classrooms appears to be a low-cost way to improve the educational performance of lower ability non-native-speaking students.

Beyond the educational setting, our results may speak to the effect of diversity in work-related settings, where existing literature has not yet reached a consensus about the effects of diversity on productivity (Hoogendoorn and Van Praag, 2012; Kahane et al., 2013; Trax et al., 2015; Dale-Olsen and Finseraas, 2019). Here, we provide complementary findings from "academic work groups" and show that higher levels of diversity indeed provide the potential to raise productivity and knowledge production.

2 Data and institutional setting

2.1 Institutional setting

We estimate the effect of the ethno-linguistic composition of seminars based on administrative data from an economics programme at a university in the London Metropolitan area. The university ranks among the top 30 universities worldwide with respect to the share

of foreign students.[5] The institution we analyse is typical of the British higher education sector in terms of its organisation. It is publicly funded, selective, and tuition fees for home students are set at the maximum specified by the regulator. Graduates from this institution earn the average graduate earnings five years after graduation (Britton et al., 2018). We focus on undergraduate students registered for any of the programme offered by the economics department, either solely or in conjunction with other departments.

Figure A1 describes the structure of undergraduate studies. An undergraduate degree lasts a minimum of three years. In any given year, students take four teaching units, spread over two teaching terms (Autumn and Spring), and lasting 10 weeks each. In our empirical analysis, we focus on compulsory teaching units that are taught over both Autumn and Spring terms, in either the first or second year. Teaching starts in the Autumn term. Seminar composition does not change between Autumn and Spring term. We focus on the mandatory courses of "Principles of Economics", "Quantitative Methods I", "Quantitative Methods II", "Microeconomics", and "Macroeconomics". In their third and final year, students can choose from about 20 elective courses. We regard third-year course choice as outcome variables in Section 4.4.

Courses comprise of weekly lectures taught by a faculty member and weekly small-group seminars where students discuss their coursework assignments. For each course, these groups meet for one hour per week over a period of 20 weeks, creating vast opportunities to interact. Seminars comprise of 26 students on average. Attendance at seminars is compulsory and monitored. Absenteeism may lead to exclusion from exams and in the case of non-EU students to visa revocation.[6]

Students are assigned to these seminars on an unsystematic basis. The administrator has little information on the personal characteristics of students apart from the program of study and only aims to balance seminar size. Students cannot influence seminar assignment. Switching to a different seminar group is prohibited. The main constraint of the administrator in allocating the students stems from time-tabling issues, especially for students on degrees that are joint with other departments. We discuss the unsystematic allocation property and provide balancing tests in Section 2.4 that show that assignment is consistent with an as-good-as-random process.

Seminars are taught by teaching assistants. Teaching material is the same for all seminars of a course and has been developed by the course leader. Teaching assistants are

[5] https://www.timeshighereducation.com/student/best-universities/international-student-table-2018-top-200-universities

[6] Following a change of administration in 2016, the no-switch policy was relaxed. We therefore disregard later cohorts for our main analysis.

usually PhD students in the economics programme. They are assigned to courses independently of the assignment of students. At the time of allocation, administrators and students have no information on which teaching assistant is going to be in charge of a given seminar. To alleviate further concerns, we later control for any effect of the seminar leader by including instructor fixed effects.

2.2 Sample description

We focus on 10 cohorts of students between 2006 and 2016. Each cohort comprises about 200 students, split almost equally between students from an English-speaking country (labeled native speakers) and students from non-English-speaking countries (labeled non-native speakers). We observe 2,184 individual students in 341 seminars, or 8,744 student $\times$ seminar observations in the five compulsory subjects.[7] The data contains information about a student's contemporary performance in compulsory courses and course choice in the non-compulsory stage of their studies, post-graduation migration from a post-graduate survey as well as background characteristics (gender, age and nationality). We further administered a survey on mechanisms in a separate contemporary cohort.

Performance. Our first main outcome is the final grade a student receives for each compulsory course in the first and second year of the studies. These grades are computed at the end of the year and include all of the coursework, mid-term exams and final exams. All coursework and mid-term exams are marked internally by the course leader and/or teaching assistants. Marking is undertaken anonymously whereby each marker gets allocated scripts independently of which seminar the student attended. Each script is identified by an exam number, which is different from the student identification number, and the linkage between exam number and student is done by administrators. The final exam, which carries the highest weight towards the final grade, is marked anonymously and independently by two internal graders and checked by an external moderator.

For our analysis of student performance, we construct three outcome variables. First, we standardise the final grade within each course per year. Courses with a final grade below 40 percent are defined as having been failed. Table 1 (Panel A) lists means and standard deviations of the performance variables. On average, 17% of students fail a course. At

[7]While Economics students take five compulsory courses, students on programmes joint with other departments only take between 2 and 4 of these courses. Therefore, the numbers of observation differ between individual students. Dropout between years is negligible and is not significantly related to diversity or share of non-native speakers. Regressions of a dummy of 1st/2nd to 3rd year attrition do not indicate a significant relationship between the seminar composition and attrition. The respective coefficients (and standard errors) are .003 (.025) for the share of non-native students and .037 (0.03) for the diversity index.

the other end of the grading scale, we observe whether a student finished the course *with honours*, i.e. with an average of 60% or above. About 60% of students in our sample have a final grade of 60 points or above. As a measure of mid-run performance, we further examine effects of first year seminar composition on average third-year grades.

Educational choices. We further analyse the effect of ethno-linguistic diversity on students' choices after the compulsory stage of their studies. In their third and final year of their studies, students choose four or more courses out of a set of about 20 different options. Course choice is not restricted by course or grade pre-requirements.

We describe a student's realised choice set by computing summary statistics of the chosen courses. First, we compute the share of numerical courses taken by a student. A course is considered to be numerical if its content is mainly quantitative and the assessments comprise of calculations rather than essay-type questions. Second, we compute the year-wise average share of non-native students in the chosen courses in that year as a measure of segregation. In other words, we compute a measure of average popularity of the chosen courses among non-native speakers. Third, we compute the average share of fails for the course as a measure of the difficulty of the realised choice set. To avoid that own choices mechanically determine this outcome, we base the computation on leave-me-out measures leaving out the current observation to ensure that own choice do not mechanically mechanically affect the outcome variables.

Table 1 (Panel A) describes these variables based on third-year information. The self-selection of students into courses induces considerable variation over the realised choice sets. The average share of non-native students varies by a standard deviation of 8.5 percentage points. The share of students failing in the chosen courses varies by a standard deviation of 4 percentage points. On average, 54% of students' choices in the final year are maths-intensive courses, this share varies by a standard deviation of 26 percentage points.

Post-graduation migration. To draw conclusions about a longer-term effect of ethno-linguistic seminar composition on students' post-graduation migration, we extract data from the Destinations of Leavers from Higher Education in the United Kingdom Survey (DLHE). The DLHE surveys recent graduates of each British institution six months after their graduation. Graduates are contacted by e-mail, post and telephone. Being administered within the timeframe of post-graduation job search, it is not very informative about graduate career success. However, we observe whether a student has left the country. We construct a binary indicator for being abroad at the time of interview. On average, 3% of

native graduates among survey respondents have left the country, compared with 16% of non-native graduates (Table 1, Panel A). These numbers might understate the true migration if migrants have lower response rates to the survey. We will later carefully examine the role of selective response when interpreting our results.

2.3 Ethno-linguistic composition of seminars

Language background by nationality. We do not have direct information on the language spoken by students. Instead, we classify the ethno-linguistic composition of seminars according to the language background of students by assigning language by the nationality of a student. We classify students from an English-speaking country where English is either the predominant or an official language as native English speakers. For non-English-speaking countries, we assign each student the predominant language of his/her nationality. While this is straightforward in most cases, we rely on a number of sources such as fact books and language encyclopaedias to determine the predominant language in the case of multilingual countries. Figure 1 and Table A1 in the appendix summarise the languages, related nationalities and number of speakers in our sample. Only 44 percent of our sample are classified as native speakers. The largest group of non-native students are Mandarin speakers (19 percent of the sample), followed by 5 percent Russian speakers, and 3 percent Italian speakers. Overall, our sample comprises students from 68 different language backgrounds.

Ethno-linguistic diversity. For each seminar in our sample, we compute the share of students originating from countries not having English as the predominant or official language which we label the *share of non-native speakers*. We then compute the diversity among the group of non-native speakers by seminar. To describe the diversity, we use the well-known Blau Diversity Index. The Blau index, which is closely related to the standard Herfindahl index of concentration, has a straightforward and intuitive interpretation: it measures the probability that two randomly-drawn non-native students within a seminar group have the same language background.[8] As such, it directly maps into the conceptual framework of Lazear (1999) where incentives to interact are related to the pool of potential same language partners. The Blau index is defined as

$$D = 1 - \sum_{k=1}^{K} p_k^2$$

[8]For very low numbers or shares of non-native students, both variables would have a discrete support. Given that only two seminars in our sample have five or fewer non-native speakers, we do not see discreteness of support as a threat to our empirical approach.

where p_k is the fraction of language group k speakers among the non-native speakers, excluding the current observation. Throughout the analysis, we use leave-me-out measures of share and diversity, where the current individual observation is kept out of the computation.

Theoretical maximum levels of the Blau Index depend on group size and the respective maximum number of potentially-distinct languages in the seminar. We account for this property by dividing the Blau Index by its respective maximum value

$$D_{max} = 1 - \frac{1}{n}$$

where n is the number of students among the group of non-native speakers in a seminar. This adjusted Blau Index has values ranging from 0, complete homogeneity, to 1, maximum possible heterogeneity of the non-native speakers group. In Section 4.3, we test the robustness of our results against alternative measures of diversity based on nationality only, and by assigning predominant native languages to countries with English as the second official language. We further test the robustness of our results against alternative ways of measuring diversity, such like the absolute number of languages in the classroom, the number or share of same-language peers, and an indicator for having at least one same-language peer in the classroom.

2.4 Testing for consistency with as good-as-random assignment to seminars

The identification of a causal effect of ethno-linguistic seminar composition relies on the idiosyncratic nature of the seminar assignment. The course administrator is instructed to assign students to seminars on a purely unsystematic basis. Deviations from this unsystematic assignment should only be due to scheduling conflicts due to already-assigned seminars and lectures from parallel study programmes. Indeed, only information about study programmes is known to administrators when making the allocation, but no further student characteristics. Unsystematic assignment is only sufficient to ensure exogeneity of the seminar composition when students cannot reallocate themselves between seminars. This is ensured as only attendance to the allocated seminar was recorded and thus visiting a different seminar would have been treated as non-attendance and would have led to exclusion.

Although we can rely on first-hand and in-depth institutional knowledge, in the following we provide supporting empirical evidence against students, teachers or administrators

circumventing the instructed unsystematic assignment out of reasons and through ways unknown to us. We therefore assess whether the observed patterns in the data are consistent with an as-good-as-random assignment applying two different tests.

First, we use the original data to simulate an artificial random group assignment and compare the actual distribution of peer compositions against this simulated random distribution. To account for the conditional nature of the assignment (conditional on course choice) as well as differences in average seminar size across courses, we permutate seminar assignments only within courses, thus maintaining the level of the original assignment as well as average seminar size by course. We then compute the share of non-native speakers and the ethno-linguistic diversity in each of these simulated seminars. In Figure 2, the white bars show the distributions of share and diversity under simulated random assignment within courses per cohort, based on 1,000 permutations. The grey bars show the actual observed distribution. Both distributions are of similar shape. Permutation-based p-values of a Wilcoxon rank sum test for equality of both distributions do not reject the null at conventional significance levels ($p = .892$ for share and $p = .549$ for diversity).[9] The observed distributions therefore are a plausible outcome of the assumed as-good-as-random assignment to seminars within courses.

Second, we formally test whether observable pre-determined average seminar characteristics predict left-out individual characteristics. Under random assignment, no such systematic relationship should exist. The results of a formal test of this hypothesis are summarised in Table 2. Each cell displays an estimate from a separate regression of the respective individual-level characteristic on leave-out-shares/averages of the same characteristic (the mean value of the variable within the peer group, not accounting for the individual observation itself). We account for fixed effects for courses per year as the level where the assignment takes place. Following Guryan et al. (2009) and Caeyers and Fafchamps (2016), we additionally control for the leave-out-share/average of the respective characteristic at the course/year level. This adjustment accounts for a mechanical negative correlation between own and peer characteristics that arises even under random assignment, as individuals cannot be their own peers. Significant correlations no longer appear as soon as we control for the level of assignment, for day/time fixed effects and study programme to account for potential deviations from unsystematic assignment due to scheduling conflicts,

[9]To determine the permutation-based p-values, we compare single simulated draws under the null as large as the observed population with the overall simulated population based on all 1,000 draws using a rank sum test. The empirical p-values are determined as the share of simulated draws which generate a test statistic as or more extreme than the one resulting from the comparison between actual observed distribution (grey bars) with the simulation-based population (white bars) displayed in Figure 2.

as expected from the institutional set-up.

Both implemented tests confirm that the observed seminar compositions in our data are consistent with the assumed as-good-as-random assignment of students to seminars. We therefore conclude that administrators, students, and teachers indeed do not infer in this process, and that we can maintain the necessary identification assumption of the as-good-as-random assignment of students to seminars throughout our analysis.[10]

3 Empirical strategy

3.1 Empirical model

We measure the ethno-linguistic classroom composition by the share of non-native speakers and the ethno-linguistic diversity among the group of non-native speakers. We identify the causal effect of these two core variables of interest by exploiting the exogeneous assignment of students to seminars. The resulting as-good-as-random treatment allocation allows us to assume differences in ethno-linguistic composition to be unrelated to students' observed or unobserved characteristics. Irrespective of their own language background, a student can experience different shares of and diversity among non-native speakers in their seminar. A student is not able to self-select into seminars by their composition, or to select another course, since we are focusing on compulsory first- and second-year courses.

Relying on the as-good-as-random assignment to seminars has advantages over alternative research designs used in the literature. Anelli et al. (2017) base their findings on arguably idiosyncratic variation over time within courses taught by the same teacher. Braakmann and McDonald (2018) use variation across courses in the same university, or within courses across universities. As the authors of these studies point out, these approaches do not entirely safeguard against selection issues that are alleviated in our setting. Nonetheless, relying on quasi-random variation has pitfalls of its own. Angrist (2014) cautions about peer effects estimated as spurious artefacts of measurement error. Feld and Zölitz (2017) instead show that under random assignment measurement error leads to an attenuation of the effect size. These issues play little role in our setting, as the variables of interest are based on information on nationality collected from administrative data which is arguably

[10] As an alternative balancing test, we additionally provide evidence in Table S1 in the supplementary web appendix that observable characteristics (age, gender, language of origin) do not jointly predict the share of non-native students or the diversity among those in all but one case. These results further support the hypothesis of as-good-as-random allocation of students to seminars. Out of the 48 coefficients reported, only one is significant at the 5 percent level (2 at the 10 percent level). In addition, in all but one case, we can reject that observable characteristics (age, gender, language of origin) jointly explain the treatment status at conventional significance levels.

measured with very little error.[11]

We estimate the effect of the share and diversity of non-native speakers via

$$y_{ics} = \beta_1 share_{non-native,cs} + \beta_2 D_{cs} + X_i'\gamma + Z_{cs}'\delta + \theta_{\mathbf{c}} + \epsilon_{ics}. \quad (1)$$

Here, y_{ics} denotes outcomes for student i, taking course c (the subscript c denotes a specific course in a specific year) and assigned to seminar s. The main variables of interest, $share_{non-native,cs}$ and D_{cs}, are the leave-me-out share of non-native speakers and ethno-linguistic diversity, respectively, assigned at the level of the seminar s. We additionally control for individual student characteristics (X_i): study program, age, gender, and whether they are non-native speakers. Z_{cs} is a vector of seminar-level characteristics including seminar leader fixed effects, size, time and day of the seminar. We additionally include course$\times$ year fixed effects to capture any unobservable characteristics that would be shared by all students attending a certain course in a specific year. This is also the level at which the seminar assignment takes place. ϵ_{ics} is the error term. We cluster standard errors at the seminar level relying on 332 clusters. We later corroborate our inference by clustering on larger levels of aggregation, using empirical p-values and accounting for multiple hypotheses testing in Section 4.3.

3.2 Identifying assumption and variation

We rely on a setting with as-good-as-random assignment of students to seminars. While this exogeneous assignment alleviates concerns about the self-selection of students into peer groups, our identification still relies on the assumption that peer ethnicity is, conditional on observable dimensions, unrelated to unobservable peer characteristics that affect the outcomes. We later provide empirical support for the validity of this assumption by examining the coefficient stability with respect to controlling for observable peer characteristics like gender, age and prior achievement.

Under this identification assumption, sufficient co-variation between share and diversity is needed to separately identify their effect with sufficient precision. Figure 3 provides a schematic description of the variation in our core variables. Seminars can differ in the share of non-native students (black symbols). Conditional on a specific share of non-native speakers, seminars can differ in their level of diversity among the group of non-native

[11] Another alternative research design in the estimation of the effect of foreign peers relies on variation by class size caps. (Ballatore et al., 2018) uses a Maimonides-type rule of class formation to estimate the effect of increasing the share of foreigners net of additional class size effects. In the absence of class size caps, such a strategy is not applicable in our context.

speakers. Comparing the hypothetical seminar B and C, it is easy to see that with a given share of non-native students, there can be variation in the diversity, here between 0 and 1. Moving from a group of non-native students that is fully homogeneous (seminar B, comprising speakers from one single group) to a seminar where the group of non-native speakers comprises many different language groups (seminar C) increases the diversity while keeping the share constant.

Figure 4 displays the empirical common support in both the share of non-native speakers and the Blau Index in our raw data (upper left panel). We observe considerable variation in the share of non-native speakers at the seminar level, ranging from about 20 to over 80 percent. Moreover, for each given share, there is also considerable variation in the diversity measure.

The upper right panel describes the respective variation in residuals after we account for fixed effects at the course×year level, individual and seminar characteristics. This is the relevant variation that is used in the quasi-experimental setup. Even after controlling for these factors, a considerable amount of variation remains both in the share and the Blau index. Table A2 in the appendix summarises the residual variation left in key variables after we control for the fixed effects, student and seminar characteristics according to equation 1. The standard deviation of the share of non-native speakers reduces between raw measures and residuals from 0.14 (raw) to 0.09 (residuals), and from 0.10 (raw) to 0.07 (residuals) for the Blau Index. Running a regression of the share of non-native students on the diversity and controlling for fixed effects according to equation (1) yields an insignificant coefficient of diversity of $\beta = 0.088[0.062]$. This low partial correlation ensures that effects of the share of non-native speakers and the diversity of the non-native speaker group can be identified separately with sufficient precision. For the remainder of the paper, we will refer to the standard deviations in residuals when describing effect sizes.

In the lower panels of Figure 4, the relationship between seminar size, share of non-native students and the diversity is depicted. Both measures of ethnic seminar composition are largely independent of seminar size and are not mechanically determined by it. For both small and large seminars, we observe sufficient variation and a comparable range of share and diversity. Nonetheless, to compare students from similarly sized seminars, we control for seminar size.

4 Results

We first present our main results on the effect of ethno-linguistic classroom composition on the contemporary performance of students in Section 4.1. We then use survey evidence to trace potential mechanisms behind the main results in Section 4.2. We test for the sensitivity of our results to different definitions of diversity and corroborate our inference with non-parametric permutation exercises in Section 4.3. We then turn to evidence on longer-term effects on third-year course choice and grades in Section 4.4 and post-graduation migration decisions in Section 4.5.

4.1 Performance

We start the discussion of the results with the effects of ethno-linguistic seminar composition on the contemporary performance of students. Both the share of non-native speakers and their diversity shape the learning environment of students. The share of non-native students might be related to lower average proficiency levels in the classroom. However, higher diversity is expected to increase incentives to engage in English conversation for non-native speakers and might lead to positive performance effects for this group.

The results of estimating equation (1) on contemporary grades as well as indicators for failing and receiving honours are summarised in Table 3. The table is organised in three panels describing the average results for all student-seminar observations (upper panel), native speakers only (middle panel), and students from a non-native language background (lower panel).

For the full sample, the share of non-native speakers is not significantly correlated to contemporary grades. We do observe a marginally significant positive effect of diversity on grades (column 1). A one standard deviation higher diversity increases grades by 1.7% ($0.07 * 0.242$) of a standard deviation. This effect appears mainly to affect marginal students. The probability to fail a course decreases significantly, while we do not see a similar effect for the probability of finishing a course with honours.

There are good reasons to expect that these average effects mask heterogeneity across native and non-native speakers as mainly non-native speakers are expected to react to changed incentives of English usage when diversity is high. The middle panel first summarizes separate effects on native speakers only. Indeed, the share of non-native speakers has no significant effect on the outcomes of native speakers.[12]

[12] As this lack of effects for native students is estimated rather imprecisely, it is useful to discuss the range of effect sizes we are able to detect given the estimator variance. The minimum detectable effect (MDE) of the share of non-native speakers lies at 4.6 percent of a standard deviation in grades for a 10 percent increase

The average effect appears primarily driven by non-native speakers, as summarised in the lower panel. For this subsample, a higher share of non-native speakers has a negative, albeit insignificant effect. Non-native speakers do, however, benefit from being assigned to a more diverse seminar. Increasing diversity by one standard deviation increases grades by 3.5% (0.07*0.496). This effect is driven by marginal students: A higher diversity by one standard deviation reduces the probability of failing by 1.1 percentage points. We find no effect of the class room linguistic composition on the probability of achieving an honour-level grade.

Taken together, we find that native students are largely insensitive to the linguistic composition of seminars. This is consistent with the moderate to zero effects of foreign students on native performance found in previous studies (Brunello and Rocco, 2013; Jensen and Rasmussen, 2011; Gould et al., 2009; Geay et al., 2013; Ohinata and van Ours, 2013). The institution studied here appears to be in line with other education settings regarding the effect of non-native students on natives' learning. Note however that the imprecision of estimates implies a certain range of negative effect sizes that would not be detectable in our setting. Further, we find insignificant negative effects of the share of non-native speakers on the performance of non-native speakers, which are counter-balanced by a higher linguistic diversity.[13]

4.2 Mechanisms

We now explore potential mechanisms for the contemporary effects of ethno-linguistic classroom composition on student performance. Non-native students' grades are affected by greater diversity, implying language-based mechanisms as the most plausible candidates. The findings of Section 4.1 on the effect of diversity support a model akin to Lazear (1999) where the return of investing in the majority language is greater to a member of a small linguistic minority than one of a large minority, conditional on the share of non-native speakers.[14] Analogously in our setting, minority students broaden their pool of potential learning partners if they communicate in English. Communicating in English bears its own costs, among others opportunity costs rising in the number of same-language peers. These opportunity costs fall in more diverse environments. We expect, conditional on the share of non-native students, that diversity is linked to a higher degree of communication

in the share. The MDE with respect to diversity indicates minimum detectable changes in performance of 3.1 percent of a standard deviation in grades in response to an increased diversity by 1 sd.

[13] Diversity and share of non-native speakers appear not to interact – the unconditional effect of share on outcomes is virtually identical to the effect conditional on diversity. Respective results are summarized in Tables S2 and S3 in the supplementary web appendix.

[14] Note that the Lazear model equates share and diversity as it focuses on the 2-languages case.

in English and a larger pool of potential learning partners from which a student can choose to interact with. Further, through learning-by-doing, English skills may increase and may lead to direct performance gains even in quantitative courses (Isphording et al., 2016).

To analyse in how far interaction patterns change in a way that is in line with the model outlined above, we examine how frequencies of interactions with native and non-native students change in response to linguistic seminar composition. We collected survey data on social interactions and language use of the most recent cohort of students attending the same courses as those investigated in the main analysis. We used an in-class written questionnaire asking students about the frequency of educational interactions with students by language background (native or non-native students), their English use and proficiency, and perceived quality of English in the classroom.

We obtained data from 538 student×seminar observations in the Spring term 2019, 222 of them with a non-native background. The overall response rate of the survey was 51 percent. Non-response is unrelated to a seminar's linguistic composition.[15] We merged survey responses to seminar composition variables of the share of non-native speakers and diversity. The survey questions were mainly asked on 1-5 scales regarding the frequency of interactions or quality of language proficiency, which we standardise for our empirical exercise.[16]

Table 4 summarises the results of estimating equation 1 for different items on language use and interaction by ethno-linguistic background as outcomes. The results support language use as an important mechanism behind the observed performance results. Non-native students are substantially less likely to interact with native speakers in seminars where the share of non-native students is larger (lower panel, columns 1) and are more likely to interact among each other (column 2). We cannot rule out that this effect is partially mechanical and stems from the proposed mechanism. Yet, we do not observe a similar pattern for the native speakers (top panel) which speaks against a pure mechanical effect. Coefficients are imprecisely estimated, though, and we cannot rule out that patterns are similar.[17]

[15] Regressing a binary indicator of non-response on the share of non-native students and diversity using equation 1 yields small and insignificant coefficients of $\beta_{share} = 0.15[0.22]$ and $\beta_{diversity} = 0.02[0.31]$.

[16] The questions are displayed in Table A10 in the appendix. Due to a change in the seminar assignment mechanism and administrative staff, we cannot rule out that students are assigned on the same unsystematic basis as our main sample. Nonetheless, a test for consistency with randomisation akin to the one in Section 2.4 does not reject the assumption of exogenous and as-good-as-random assignment (see Table S4 in the supplementary web appendix). Effects on performance differ to some degree from the analysis in the main sample and are, potentially due to smaller sample size, insignificant. While diversity has a similar positive effect on performance ($\beta = .67[.61]$)), the share of non-native speakers is negatively associated with performance ($\beta = -.91[.44]$).

[17] As the questionnaire does not specifically ask for interactions between own language and other language non-native students, we cannot examine in how far a higher diversity also increased interaction among non-native peers from different language background.

A larger pool of non-native students seems to lead to stronger segregation in classroom interactions. This effect is mitigated by a higher diversity among the non-native students: a more diverse classroom increases the interactions of non-native students with their native seminar peers. This direct evidence on the proposed mechanism links reasonably well with the observed slight negative effects of a higher share of non-native students on the performance of non-native speakers, as well as the mitigating effect of a higher diversity among the group of non-native speakers observed in Table 3.

Effects on direct measures on perceived comfort in English use, own proficiency and the quality of English in the classroom remain insignificant for non-native students. Native speakers marginally perceive a lower quality of English spoken in the classroom when assigned to a seminar with a higher share of non-native speakers.

Taken together, the survey evidence supports the idea that different incentives to engage in interactions with native students are one of the main mechanisms to explain the effects of diversity on performance.

4.3 Robustness checks

We now examine the sensitivity of our results to change in the set of controls, alternative definitions of language and diversity, the specific role of Mandarin speakers, non-linearity and alternative approaches to statistical inference.

Controlling for ability and correlated peer characteristics. We investigate in how far our results are sensitive to controlling for seminar characteristics that might be correlated to the ethno-linguistic seminar composition. The share of non-native students might, for example, be correlated with further dimensions of peer characteristics that have spillovers on choices and performance by themselves, e.g. gender and ability. We control for leave-me-out averages and shares of age and gender.

Unfortunately, our data does not contain information on pre-university ability, such as entry test scores or high-school GPA. Instead, we compute student ability as a leave-seminar-out average of those grades a student has received in further classes. We then test the stability of our estimated main effects towards further observable characteristics. Altonji et al. (2005) and Oster (2017) argue that movements of coefficients when controlling for observables are informative about selection on unobservables, too, as long as observable characteristics are a random subset of a larger set of characteristics. Table A4 in the appendix summarises coefficients for the share of non-native students and the diversity among them for four different specifications: only individual characteristics (1), additional

seminar characteristics (2), and additionally controlling for own (3) and peer achievement (4). The comparison of specifications with and without seminar characteristics (1 vs. 2, and 3 vs. 4) is informative about the role of seminar characteristics as potential confounders. Estimated coefficients of diversity are fairly stable when controlling for additional peer characteristics, and not statistically different between specifications. As in the base model, we find that native students are unaffected by seminar composition and that diversity but not share of non-native, has a positive effects on the grades of non-native students. This is not surprising: peer characteristics other than diversity appear not to play a role in explaining the outcome. The according R^2 values barely change between specifications. Hence, we are unable to construct formal parameter bounds following Altonji et al. (2005) and Oster (2017) but conclude that all peer observations that are observable in our data do not interact with the observed peer effects by linguistic background.[18]

While this robustness check does not hint at confounding unobserved peer characteristics, we cannot entirely rule out that observed effects are picking up variation in unobserved peer characteristics uncorrelated with observed peer controls. This issue of potentially confounding but unobserved peer characteristics is common to any peer effects study that relies on natural variation in peers. Confounding peer characteristics cannot be separated from a person, and therefore cannot be (quasi-)experimentally stimulated. However, one might argue that owing to this inseparability, estimated effects are the relevant policy parameter. Finally, we argue that while shares and averages in further unobserved dimensions might be correlated with the share of non-native students, this is much less likely to affect our main parameter of interest, namely the effect of diversity in the group of non-native speakers.

Alternative definitions of language and diversity. In the main specifications, we define diversity along the lines of language groups. As such, foreign-born students from English-speaking countries are defined as native speakers. This definition already anticipates language being a main mechanism in terms of how diversity affects student performance. Nonetheless, diversity could be defined along related but different dimensions.

In Table A5, we replicate the main findings of Table 3, column (1), using alternative definitions of language group. Specifically, we test two alternative definitions. Column (1) lists the baseline results. In column (2), we deviate from the main specification by assigning the predominant language to countries with English as an official language.[19]

[18] In Table A3, we have further explored the sensitivity of our results towards different sets of fixed effects. The coefficients of share and diversity turn out to be stable after controlling for necessary fixed effects for the level of group assignment and scheduling conflicts. Additionally controlling for individual and seminar characteristics as well as teaching assistant fixed effects has a negligible effect on the coefficients.

[19] Gambia, Kenya, Nigeria, South Africa, Trinidad & Tobago, Uganda

While the general pattern remains, the positive effect of diversity on the grades of non-native speakers is substantially reduced. One potential explanation is that we now include students who speak English very well (as it is an official language in their country of birth) as part of the non-native speakers' population. These students are potentially less affected by a larger diversity. In column (3), we define nativity and diversity solely on the basis of nationality. Accordingly, we compare UK students to non-UK students and disregard the language dimension; thus, estimates would mix the effect of language and of culture. As before, the pattern of estimated coefficients remains similar to our main specification. For the native-speaker group, alternative language definitions do not alter the base-line results. Taken together, patterns of results are fairly stable across different definitions of diversity and nativity, and are consistent with the effect being driven by language ability rather than culture.

We then explore alternative measures of diversity. In Table A6, column (2), we replace the Blau index by the number of languages spoken in the classroom. The number of languages is closely related to the diversity index (r = 0.414) and for all groups, the results indeed match the pattern found for the baseline results.

Rather than diversity, it might be the lack of opportunity to speak own language that provides incentives to learn English. We thus replace diversity by the number of own language speaker (Column 3), the share of own speakers (Column 4) and an indicator for having at least one person in the seminar speaking own language (Column 5). Note that these variables are negatively correlated with diversity; at a given share of non-native speakers, a greater share of own language speakers involves a reduced diversity; we thus expect the estimate to flip sign compares to the baseline specification. Indeed, having a greater share (number) of own speakers in the classroom reduces the grades of non-native students. As a further support for the hypothesis that the results are driven by changes in the demand for speaking language, we can see that a single own-language peer in the seminar is enough to reduce grades by 0.09 of a standard deviation. These specifications confirm that our baseline results are not sensitive to alternative measures of diversity

The role of Chinese students. Diversity and share of non native speakers are driven to a large degree by Chinese Mandarin speakers who represent 19 percent of the sample. We therefore test the robustness of our results with respect to controlling for the number or share of Mandarin speakers. We further examine heterogeneity in effects between Mandarin speakers and other non-native students.

Regressing the share of Mandarin speakers on the share of non-native speakers and

the diversity and controlling for fixed effects according to equation (1) yields coefficients of $\beta_{share} = 0.28[0.023]$ and $\beta_{diversity} = -0.63[0.029]$ confirming that Mandarin speakers largely influence the linguistic composition of seminars.

Table A7, column 1 displays the baseline results from Table 3. Despite the strong correlation between the share of Mandarin speakers and linguistic diversity, additionally controlling for the number or the share of Mandarin speakers in the seminar preserves the general pattern of the main results (columns 2 and 3). However the strong correlation between the share of Mandarin speakers and diversity reduces the precision of the diversity estimates that becomes insignificant. Splitting the sample between Mandarin speakers and other non-native students (columns 5 and 6) again yields patterns similar to our main results for both groups; the coefficients for diversity are less precisely estimated and insignificant, albeit likely due to sample size issues. We conclude from the stability of patterns that although Chinese play an important role in generating our results, the underlying mechanisms are to some degree independent of their presence.

Non-linearity The data at our disposal is characterised by a large support for the share of non-native speakers at the seminar level (ranging from 21% to 87%). This allows us to test for non-linearity in this effect, and assess whether there is a threshold at which the average level of English deteriorates to a level which makes learning less efficient. In Figure A2 we report estimates of the effect of share of non-native speakers defined in quartiles on grades. The estimates are small and not significantly different from zero for either population. Even at high level of internationalisation, the share of non-native speakers has no effect on contemporary learning. When similarly measuring diversity, the greater the level of diversity, i.e. the more difficult it is to find an own-language speaker, the more the grades of non-native students improve. This effect appears to be almost linear. This linearity suggests that the optimal policy would be to maximise diversity, especially since there is no negative effect of very high levels of diversity on the native speakers at such high levels of diversity. Instead, for native speakers, a certain u-shaped pattern emerges. Moderate levels of diversity have potential negative effects, while very low and very high levels do not affect native performance.

Robustness of inference. In the main specifications, we allow for clustering of error terms at the seminar level. In Table A8, we examine robustness to alternative inference corrections. Column (1) displays baseline results with errors clustered on the seminar level. Column (2) lists results when assuming i.i.d. error terms. Column (3) applies simple robust

standard errors. In the remaining specifications, we adjust the level of clustering to the course×year level (column 4) and the year level (column 5). Standard errors of the share parameter increase with higher levels and smaller numbers of clusters. Standard errors of the diversity parameter appear to be insensitive. Our conclusions remain unaffected by the choice of inference correction.

Further, we assess the robustness of inference by using non-parametric permutation tests. For this purpose, we randomly assign students within courses by cohort to placebo seminar IDs and re-run the analysis. We repeat this procedure 2,000 times. Distributions of the resulting simulated coefficients in relation to the originally-estimated parameter are summarised in Figure A3 in the appendix, focusing on the main results of Table 3, lower panel. The implied empirical p-values (p=0.19 and p=0.028 for share and diversity respectively) confirm the parametric significance levels.

Finally, in Table A9 in the appendix, we follow the procedure proposed by Romano and Wolf (2005) to examine in how far our result patterns are robust towards the issue of multiple hypothesis testing. This stepdown method adjusts p-values by taking into account the amount of outcomes that are tested within a subsample of native or non-native speakers. Comparing original and adjusted p-values across outcomes and subgroups shows that significance patterns remain largely robust. Only the long-term effect of diversity on grades (as defined below in Section 4.4) is no longer significant at the 5-percent level – its p-value increases from 0.046 to 0.08.[20]

4.4 Long-run Academic Effects

We now turn to the longer-run effects of ethno-linguistic seminar composition on final-year performance and course choice. Early seminar composition might have longer-run effects on future grades if language improvement, peer interactions or learning behaviour acquired in the compulsory stage continue to influence educational attainment.

Table 5 reports estimates of equation (1) for final-year grades. Note that, while long-term outcomes only differ on the student level, the analysis is conducted at the same student times seminar level of observation as in the analysis of contemporary performance effects in Table 3. Thus, coefficients are to be interpreted as the effect of being assigned to a single seminar in first or second year on long-run performance and choices. The table is organised into three panels separately for all students, native and non-native speakers.

We first investigate whether final year course choice is affected by early ethno-linguistic

[20]The tests were conducted using the Stata *rwolf* routine by Clarke et al. (2019).

seminar composition. There are at least three reasons why educational choices might be affected. First, the initial effect on grades affects students' perceived academic ability, which might sway their choices towards more or less demanding courses. In particular, non-native speakers might reinforce their perceived comparative advantage in more quantitative courses, while native speakers similarly might perceive their comparative advantage in English as being even greater. Second, students exposed to more individuals from other ethnicities change their patterns of interaction, which might change their attitudes towards these ethnicities (Boisjoly et al., 2006; Carrell et al., 2019), making them more willing to interact with the same ethnicities in future courses. Third, having experienced a linguistically-dissonant learning environment in their compulsory stage, students might opt for courses with a more quantitative curriculum where verbal communication plays a lesser role (Anelli et al., 2017).

We focus on three indicators describing a students' realised choice set among third-year non-compulsory courses: share of numerical courses, popularity among non-native speakers (measured as leave-me-out shares of non-native students in current year) and difficulty (measured as the share of fails in the chosen courses).

The results indicate only little influence of ethnic seminar composition on subsequent educational choices. Difficulty and share of quantitative content remain unaffected by seminar composition for native and non-native students alike. We do find, however, that non-native students exposed to more non-native speakers in the first year are slightly more likely to choose courses which are typically more popular among non-native students, leading to some degree of segregation of non-native speakers in third year courses. We do not observe any similar effect of the share of foreigners on natives' choices.

With respect to third year grades, having met a larger share of non-native speakers in the first two years is beneficial for both native and non-native speakers. The long-run effect of the share of non-native students on performance is not consistent with the contemporary lack of effects and points to different mechanisms in the long-run. Yet, it follows a similar logic as an enclave effect as it is investigated in the literature on labour market effects of ethnic enclaves (Edin et al., 2003; Damm, 2009). Ethnic enclaves, i.e. environments with large shares of non-native speakers, are in general detrimental for assimilation in the short run: they decrease incentives to acquire language skills. In the long run, though, this effect reverses. Potential reasons, both in the labour market as in our classroom setting, are better information dissemination in the group of non-native speakers and differential network formation that are valuable in the long run.

Finally, similar to the contemporaneous effect on grades, diversity has a long-run positive effect on grades of non-native students, too, which points to a certain persistence in the effect of diversity on performance. This long-term effect is in line with a human capital effect: non-native students who are exposed to a more diverse first year environment have higher incentives to invest in their English skills. This enables them to increase their performance contemporaneously but also in subsequent years, especially among low achievers.

4.5 Post-graduation migration

We now turn to the analysis of longer-term effects on post-graduation migration. While some foreign students invest in education abroad as a form of temporary migration before returning home, for others the returns to foreign education are greater if they remain in the country where they gain this education (Dustmann and Glitz, 2011). Since classroom linguistic composition affects both language acquisition and educational attainment, it might alter the relative returns of staying in the UK. Additionally, early exposure to more or less foreign students might affect the social network a student is able to build, which in turn affects migration decisions.

In Table 6, we examine the role of ethnic classroom composition during the compulsory stage of their study on post-education migration. Again, regressions are run on the level of a student times seminar as in the case of the analysis of contemporary performance effects and coefficients have to be interpreted as the effect of being assigned to a single seminar on post-graduation migration.

Information on student return migration is drawn from the Destinations of Leavers from Higher Education survey, an annual survey of recent graduates conducted by the Higher Education Statistical Agency. The survey response rate is around 30% but considerably lower for non-native speakers (20%) and is not related to early seminar composition for non-native speakers. Yet, native speakers are less likely to respond if they are exposed to more diverse seminars, although the effect is small. Being exposed to a higher diversity by one standard deviation reduces the response probability by 1.3% from a mean of 39.3 percentage points. Nonetheless, this effect of treatment on the response rate cautions interpreting the effect on native speakers as causal.

The results on migration decisions differ between native and non-native speakers. Non-native speakers who are exposed to more non-native speaking peers in their compulsory stage are more likely to have left the country at the point of the survey. A higher share

of non-native peers in a compulsory-stage seminar by 10 percentage points increases the probability of living abroad by 2.6%. This effect is in line with having fewer opportunities to build English-based networks when in contact with more non-native peers. We cautiously interpret these results as suggestive evidence of an effect of exposure to non-native speakers on their return or onward migration. Native speakers display a lower probability of migrating when having been exposed to a higher diversity. This effect is small, though; a 1 standard deviation increase in class-room diversity reduces migration by 0.6 percentage points.

5 Conclusion

Using data from a UK higher education institution that exogeneously allocates students to small classes, we do not find evidence of a negative effect of having a larger share of non-native students. In particular, native students are unaffected by the ethno-linguistic composition of seminars. Non-native students benefit from higher ethno-linguistic diversity in terms of their performance. The effect is linear in diversity and extends to grades in subsequent years. The effect is mostly driven by weaker students. Survey evidence implies that diversity augments the interaction among native and non-native students. Diversity does not alter final year course choice or the decision to migrate.

Our results are informative for the design of classroom assignment processes. Strategically avoiding an ethnically segregated early study environment by increasing classroom diversity may be a low-cost way to improve the educational performance of foreign students. Such re-assignment could for example be achieved through stratified assignment to seminars, where students are randomised within their own language groups.

The group-work-focused learning environment in the seminars may allow for generalisations to other settings of team production involving cognitive tasks. Evidence of the effect of diversity in production settings is scarce, restricted to either quasi-experiments based on sports data (Kahane et al., 2013), lab evidence (Hoogendoorn and Van Praag, 2012) or descriptive evidence from observational data (Trax et al., 2015), (Dale-Olsen and Finseraas, 2019). Here, we add causal field evidence from a setting sharing many features of collaborative environments, which are now standard in many workplace environments.

More generally, the effects of diversity on economic and social outcomes appear to differ by the level of aggregation and results are inconclusive so far. Alesina and Ferrara (2005) propose a model that allows for negative effects of diversity on public good provision and positive effects on productivity. The majority of the literature so far has focused on the

former negative effects, with evidence by Algan et al. (2016) demonstrating the negative effect of diversity on social cohesion in housing blocks in France. The latter positive effect on productivity has only recently gained attention by linking higher productivity (income) to birth place diversity (Ottaviano and Peri, 2006; Ager and Brückner, 2013). Against this broader literature on diversity, our results are informative about the positive effects of diversity on productivity on a much smaller level of peer groups with strong and meaningful social interactions.

References

Ager, P. and Brückner, M. (2013). Cultural diversity and economic growth: Evidence from the us during the age of mass migration. *European Economic Review* 64: 76–97.

Alesina, A. and Ferrara, E. L. (2005). Ethnic Diversity and Economic Performance. *Journal of Economic Literature* 43: 762–800.

Algan, Y., Hemet, C. and Laitin, D. D. (2016). The Social Effects of Ethnic Diversity at the Local Level: A Natural Experiment with Exogenous Residential Allocation. *Journal of Political Economy* 124: 696–733.

Altonji, J. G., Elder, T. E. and Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy* 113(1): 151–184.

Anelli, M., Shih, K. and Williams, K. (2017). Foreign Peer Effects and STEM Major Choice. *IZA Discussion Paper Series* 10743.

Angrist, J. D. (2014). The perils of peer effects. *Labour Economics* 30: 98–108.

Ballatore, R. M., Fort, M. and Ichino, A. (2018). Tower of Babel in the Classroom: Immigrants and Natives in Italian Schools. *Journal of Labor Economics* 36.

Boisjoly, J., Duncan, G., Kremer, M., Levy, D. and Eccles, J. (2006). Empathy or Antipathy? The Impact of Diversity. *American Economic Review* 96: 1890–1905.

Braakmann, N. and McDonald, S. (2018). The Impact of Student Diversity on Student Outcomes at University and Beyond - Evidence From English Administrative Data. *Unpublished manuscript* .

Bredtmann, J., Otten, S. and Vonnahme, C. (2018). Linguistic diversity in the classroom, student achievement, and social integration. *Ruhr Economic Paper No. 783* .

Britton, J., Belfield, C., Erve, L. van der, Vignoles, A., Walker, I. and Dearden, L. (2018). Estimating the Returns to College Degrees Using Linked Administrative Data. *Institute for Fiscal Studies Research Report* .

Brunello, G. and Rocco, L. (2013). The Effect of Immigration on the School Performance of Natives: Cross Country Evidence Using PISA Test Scores. *Economics of Education Review* 32: 234–246.

Caeyers, B. and Fafchamps, M. (2016). Exclusion Bias in the Estimation of Peer Effects. NBER Working Papers 22565, National Bureau of Economic Research, Inc.

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: 159–182.

Clarke, D., Romano, J. P. and Wolf, M. (2019). The Romano-Wolf Multiple Hypothesis Correction in Stata. IZA Discussion Papers 12845, Institute of Labor Economics (IZA).

Dale-Olsen, H. and Finseraas, H. (2019). Linguistic diversity and workplace productivity. *IZA Discussion Paper No. 12621* .

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

Diette, T. M. and Oyelere, R. U. (2017). Do limited english students jeopardize the education of other students? lessons from the north carolina public school system. *Education Economics* 25:5: 446–461.

Dustmann, C. and Glitz, A. (2011). Migration and Education. Norface Discussion Paper Series 2011011, Norface Research Programme on Migration, Department of Economics, University College London.

Edin, P.-A., Fredriksson, P. and Åslund, O. (2003). Ethnic Enclaves and the Economic Success of Immigrants - Evidence from a Natural Experiment. *The Quarterly Journal of Economics* 118: 329–357.

Feld, J. and Zölitz, U. (2017). Understanding Peer Effects: On the Nature, Estimation and Channels of Peer Effects. *Journal of Labor Economics* 35: 387–428.

Figlio, D. N. and Özek, U. (2017). Unwelcome Guests? The Effects of Refugees on the Educational Outcomes of Incumbent Students. NBER Working Papers 23661, National Bureau of Economic Research, Inc.

Frattini, T. and Meschi, E. (2017). The Effect of Immigrant Peers in Vocational Schools. *IZA Discussion Paper Series* 11027.

Geay, C., McNally, S. and Telhaj, S. (2013). Non-Native Speakers of English in the Classroom: What Are the Effects on Pupil Performance? *The Economic Journal* 123: 281–307.

Gould, E. D., Lavy, V. and Paserman, M. D. (2009). Does Immigration Affect the Long-Term Educational Outcomes of Natives? Quasi Experimental Evidence. *The Economic Journal* 119: 1243–1269.

Guryan, J., Kroft, K. and Notowidigdo, M. J. (2009). Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments. *American Economic Journal: Applied Economics* 1: 34–68.

Hoogendoorn, S. and Van Praag, M. (2012). Ethnic diversity and team performance: A field experiment. Tech. rep., inbergen Institute Discussion Paper No. 12-068/3.

Hunt, J. (2017). The Impact of Immigration on the Educational Attainment of Natives. *Journal of Human Resources* 52: 1060–1118.

Isphording, I. E., Piopiunik, M. and Rodríguez-Planas, N. (2016). Speaking in numbers: The effect of reading performance on math performance among immigrants. *Economics Letters* 139: 52–56, doi:10.1016/j.econlet.2015.11.

Jensen, P. and Rasmussen, A. W. (2011). The Effect of Immigrant Concentration in Schools on Native and Immigrant Children's Reading and Math Skills. *Economics of Education Review* 30: 1503–1515.

Kahane, L., Longley, N. and Simmons, R. (2013). The Effects of Coworker Heterogeneity on Firm-Level Output: Assessing the Impacts of Cultural and Language Diversity in the National Hockey League. *The Review of Economics and Statistics* 95: 302–314.

Lazear, E. P. (1999). Culture and Language. *Journal of Political Economy* 107: S95–S126.

Machin, S. and Murphy, R. (2017). Paying out and crowding out? The globalization of higher education. *Journal of Economic Geography* 17: 1075–1110.

Maestri, V. (2017). Can Ethnic Diversity Have a Positive Effect on School Achievement? *Education Economics* 25: 290–303.

McHenry, P. (2015). Immigration and the Human Capital of Natives. *Journal of Human Resources* 50: 34–71.

Ohinata, A. and Ours, J. C. van (2013). How Immigrant Children Affect the Academic Achievement of Native Dutch Children. *Economic Journal* 123: 308–331.

Oster, E. (2017). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics* .

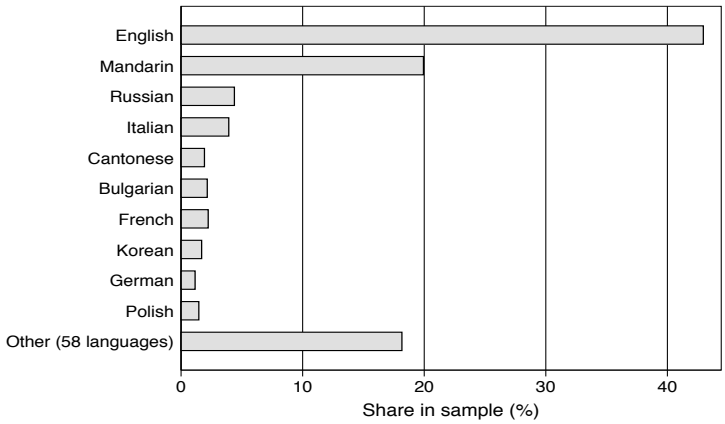
Ottaviano, G. I. and Peri, G. (2006). The economic value of cultural diversity: evidence from us cities. *Journal of Economic geography* 6: 9–44.

Romano, J. P. and Wolf, M. (2005). Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica* 73: 1237–1282.

Trax, M., Brunow, S. and Suedekum, J. (2015). Cultural Diversity and Plant-level Productivity. *Regional Science and Urban Economics* 53.

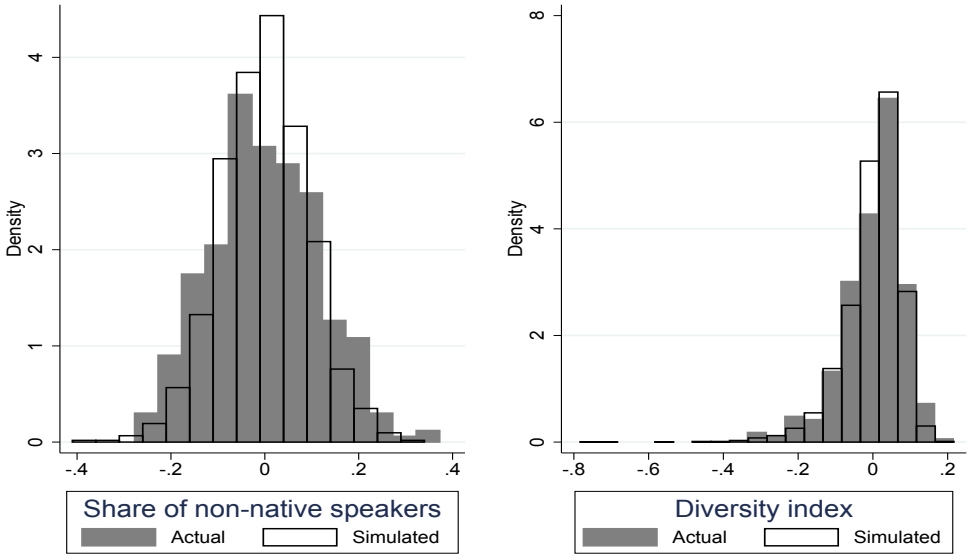
6 Tables & Figures

Figure 1: Sample composition by language background



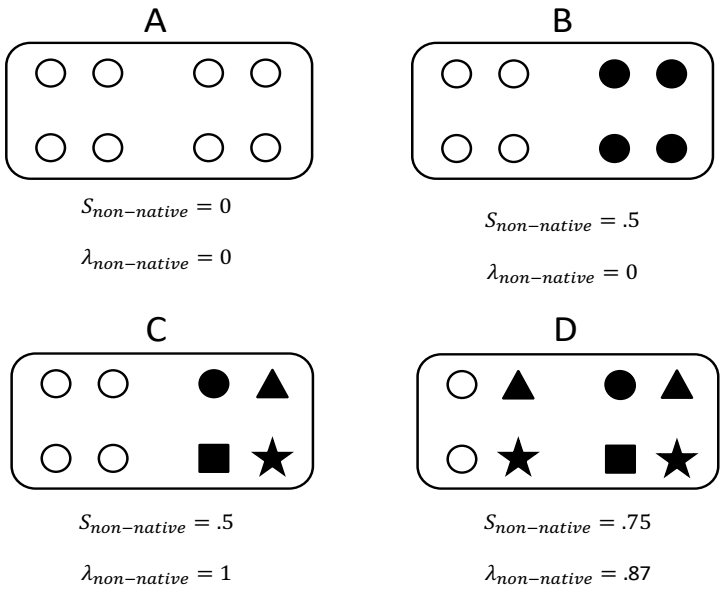
Notes: This figure displays the share of language groups in the individual sample (n=2,184). Languages are assigned by nationality: each student is assigned the predominant language of the country that the student reports as his/her nationality.

Figure 2: Simulated vs observed seminar composition



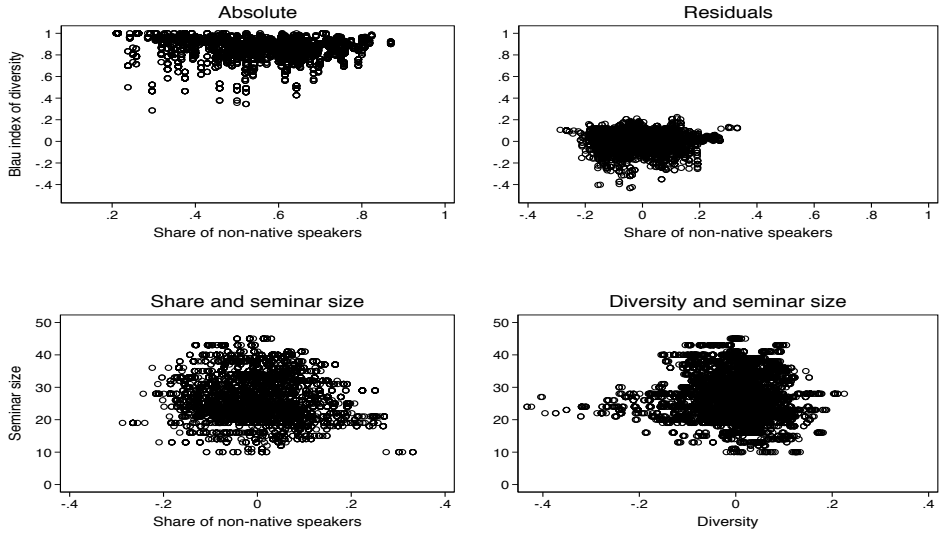
Notes: This figure compares observed distributions of the core variables of share of non-native speakers and diversity with simulated distributions based on pure random assignment based on 1,000 replications within courses, holding seminar sizes at observed levels. Variables are displayed as deviations from the $course \times term$ average. Permutation-based p-values of a Wilcoxon rank sum test cannot reject the null of equality between observed and simulated distribution.

Figure 3: Share vs diversity



Notes: This figure illustrates the difference between the share of non-native speakers and the Blau Diversity Index for example classrooms with eight students. Each symbol represents a student. White dots represent native students while black symbols are for non-native speakers; each shape represents a specific native language.

Figure 4: Variation in share of non-native speakers and diversity



Notes: This figure displays the variation in the share of non-native speakers and the ethnic diversity within the group of non-native speakers, in absolute levels (left panel) and in residuals after applying the within-transformation by course × year, study programme, day × hour, and seminar leader fixed effects (right panel) for the Blau Index. Standard deviations: share of non-native speakers 0.14 (absolute) and 0.09 (residuals), Blau Index 0.10 (absolute) and 0.07 (residuals).

Table 1: Sample descriptives

	Sample: Total				Sample: Native	Sample: Non-native
A. Dependent variables						
	Mean	*SD*	*Min*	*Max*	*Mean*	*Mean*
Performance						
Std. course grade	0.00	1.00	-4.75	3.14	0.08	-0.07
Course failed	0.17	0.37	0.00	1.00	0.15	0.19
With honors (above 60%)	0.40	0.49	0.00	1.00	0.43	0.38
No. of obs	*8744*				*4032*	*4712*
Educational choices						
Share of numerical courses in third year choices	0.54	0.26	0.00	1.00	0.49	0.58
Share of non-natives in third year courses	0.55	0.09	0.03	0.76	0.52	0.57
Difficulty of the courses chosen in third year	0.05	0.03	0.00	0.20	0.05	0.05
Std. average grade in third year	0.00	1.00	-5.45	2.50	0.14	-0.13
No. of obs	*7708*				*3645*	*4063*
Post-graduate outcomes						
Abroad	0.08	0.27	0.00	1.00	0.03	0.16
No. of obs	*2540*				*1583*	*957*
B. Individual characteristics						
	Mean	*SD*	*Min*	*Max*	*Mean*	*Mean*
Student's age	19.78	1.40	17.00	34.00	19.47	20.04
Female student	0.41	0.49	0.00	1.00	0.34	0.47
No. of obs	*8744*				*4032*	*4712*
C. Seminar characteristics						
	Mean	*SD*	*Min*	*Max*		
No. of students	*25.76*	6.83	10.00	45.00		
Share of non-native speakers	*0.55*	0.14	0.21	0.87		
Blau index of diversity	*0.87*	0.10	0.29	1.00		
No. of obs	*341*					

Notes: This table summarises descriptive statistics of individual and seminar characteristics and the dependent variables.

Table 2: Testing for consistency with random assignment

	(1)	(2)	(3)	(4)
Leave-me-out mean/share (seminar)				
Non-native speaker	0.376***	0.023**	0.011	0.001
	(0.052)	(0.009)	(0.010)	(0.012)
first year GPA[a]	0.449***	0.014	0.005	-0.014
	(0.071)	(0.013)	(0.015)	(0.017)
Student's age	0.139***	-0.006	-0.007	-0.009
	(0.053)	(0.009)	(0.010)	(0.011)
Gender: Female	0.194***	0.021**	0.012	0.007
	(0.056)	(0.009)	(0.011)	(0.012)
Language: Mandarin	0.293***	0.025***	0.015	0.014
	(0.055)	(0.009)	(0.011)	(0.013)
Language: Russian	0.232***	0.016	0.016	0.014
	(0.068)	(0.013)	(0.013)	(0.015)
Language: Italian	0.132*	-0.007	-0.015	-0.020**
	(0.071)	(0.009)	(0.009)	(0.010)
Leave-me-out share/mean (urn)	yes	yes	yes	yes
Course × year FE	no	yes	yes	yes
Study program FE	no	no	yes	yes
Day/Time FE	no	no	yes	yes
Seminar leader FE	no	no	no	yes
No. of observations[a]	8744	8744	8744	8744

Notes: This table summarises results of regressions of seminar-wise leave-me-out means/shares on observable student characteristics; each row represents a separate regression. Each regression includes the course/year-wise leave-me-out mean/share and fixed effects combinations as indicated in the table. [a]First year GPA is only available for 4,404 observations and does not generally enter our later specifications. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered at the individual level, are reported in parentheses.

Table 3: Diversity and educational performance

Sample	Grade	Fail	Honour
Total			
Share of non-natives	-0.043	0.050	0.015
	(0.117)	(0.036)	(0.053)
Blau Index	0.242*	-0.124***	-0.030
	(0.138)	(0.039)	(0.065)
Mean of dep. var.	0.000	0.167	0.401
R^2	0.04	0.08	0.12
No. of observations	8744	8744	8744
Native			
Share of non-natives	0.094	0.029	0.022
	(0.163)	(0.058)	(0.072)
Blau Index	-0.075	-0.043	-0.026
	(0.159)	(0.058)	(0.080)
Mean of dep. var.	0.079	0.146	0.427
R^2	0.08	0.10	0.15
No. of observations	4032	4032	4032
Non-native			
Share of non-natives	-0.165	0.074	-0.002
	(0.167)	(0.054)	(0.074)
Blau Index	0.496**	-0.186***	-0.033
	(0.197)	(0.060)	(0.085)
Mean of dep. var.	-0.068	0.185	0.379
R^2	0.07	0.10	0.15
No. of observations	4712	4712	4712
Course × year FE	yes	yes	yes
Study program FE	yes	yes	yes
Day/Time FE	yes	yes	yes
Seminar leader FE	yes	yes	yes
Individual controls	yes	yes	yes

Notes: This table summarises results of regressions of a set of outcome variables (standardised grade, indicator for failing a course, indicator for receiving an honour (60% or above) grade) on the seminar-wise leave-me-out share of non-native speakers and the diversity index. Results by language background (native/non-native speakers) are derived from split sample models. Controls contain seminar size, age, gender and whether they are a native speaker or not. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered at the seminar level, are reported in parentheses.

Table 4: Mechanisms

Sample	Interaction with native students	Interaction with non-native students	Feeling comfortable using English	Perceived quality of English	Own English proficiency
Native speakers					
Share of non-natives	-1.143	0.0952	-0.347	-1.868*	-0.101
	(0.895)	(0.864)	(0.351)	(0.990)	(0.414)
Blau Index	0.534*	-0.205	0.146	0.704	0.117
	(0.316)	(0.338)	(0.118)	(0.441)	(0.183)
Mean of dep. var.	0.28	-0.18	0.40	0.19	0.42
R^2	0.03	0.02	0.02	0.09	0.02
No. of observations	316	317	316	315	317
Non-native speakers					
Share of non-natives	-2.799**	1.947**	-0.958	-0.984	0.672
	(1.132)	(0.843)	(1.333)	(1.034)	(1.084)
Blau Index	1.454**	-1.281***	0.135	0.519	-0.500
	(0.547)	(0.386)	(0.643)	(0.462)	(0.539)
Mean of dep. var.	-0.40	0.26	-0.57	-0.27	-0.60
R^2	0.04	0.06	0.02	0.02	0.02
No. of observations	222	225	223	225	224

Notes: This table summarises results of regressions of a set of survey responses on potential mechanisms on the seminar-wise leave-me-out share of non-native speakers and the diversity index. Results by language background (native/non-native speakers) are derived from split sample models. Individual controls contain age, gender and whether they are a native speaker or not. The survey was administered in an adjacent cohort of the autumn semester 2018. The response rate of the survey was 51 percent. Outcomes are standardised from 1-5 scales (columns 1-2: Never to Very Often, column 4: Very uncomfortable to Very comfortable, columns 4 and 6: Very bad to Very good. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered at the seminar level, are reported in parentheses.

Table 5: Diversity and third-year choices

Sample	Share of Non-native-speaking	Share of Numerical courses	Difficulty	Grade
Total				
Share of non-natives	0.011*	-0.005	0.001	0.327***
	(0.006)	(0.025)	(0.003)	(0.095)
Blau Index	0.002	0.037	0.005*	0.258**
	(0.007)	(0.030)	(0.003)	(0.113)
Mean of dep. var.	0.549	0.537	0.051	0.000
R^2	0.68	0.33	0.62	0.09
No. of observations	7708	7708	7708	7708
Native				
Share of non-natives	0.001	-0.032	0.000	0.287*
	(0.008)	(0.038)	(0.003)	(0.163)
Blau Index	0.006	0.002	0.002	0.187
	(0.010)	(0.034)	(0.003)	(0.167)
R^2	0.63	0.33	0.67	0.11
No. of observations	3645	3645	3645	3645
Non-native				
Share of non-natives	0.022***	0.039	0.001	0.325**
	(0.009)	(0.033)	(0.004)	(0.149)
Blau Index	-0.005	0.051	0.006	0.326**
	(0.011)	(0.043)	(0.004)	(0.163)
R^2	0.68	0.33	0.62	0.09
No. of observations	4063	4063	4063	4063
Course × year FE	yes	yes	yes	yes
Study program FE	yes	yes	yes	yes
Day/Time FE	yes	yes	yes	yes
Seminar leader FE	yes	yes	yes	yes
Individual controls	yes	yes	yes	yes

Notes: This table summarises results of regressions of a set of outcome variables regarding course choices in third year on the seminar-wise leave-me-out share of non-native speakers and the diversity index. Results by language background (native/non-native speakers) are derived from split sample models. Controls contain seminar size, age, gender and they are a native speaker or not. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered at the seminar level, are reported in parentheses.

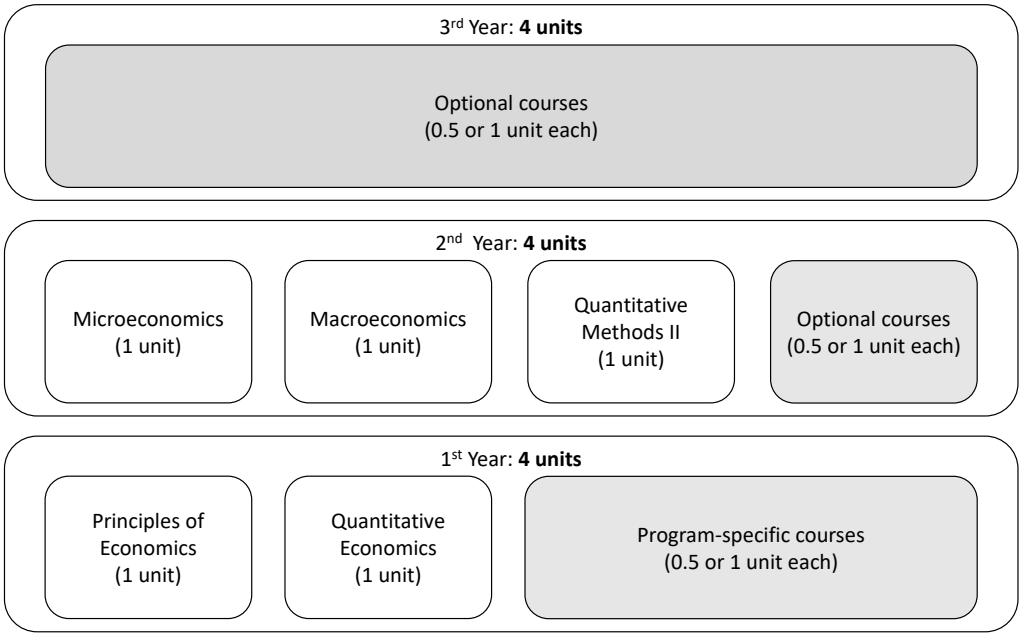
Table 6: Diversity and post-graduation migration

Sample	Response	Abroad
Total		
Share of non-natives	-0.052	0.055
	(0.033)	(0.042)
Blau Index	-0.064	-0.091*
	(0.054)	(0.048)
Mean of dep. var.	0.289	0.079
R^2	0.46	0.13
No. of observations	8744	2537
Native		
Share of non-natives	-0.022	0.015
	(0.054)	(0.032)
Blau Index	-0.181***	-0.090**
	(0.069)	(0.040)
Mean of dep. var.	0.393	0.027
R^2	0.58	0.12
No. of observations	4032	1581
Non-native		
Share of non-natives	-0.026	0.255**
	(0.047)	(0.100)
Blau Index	0.030	0.009
	(0.061)	(0.130)
Mean of dep. var.	0.203	
R^2	0.33	0.16
No. of observations	4712	956
Course × year FE	yes	yes
Study program FE	yes	yes
Day/Time FE	yes	yes
Seminar leader FE	yes	yes
Individual controls	yes	yes

Notes: This table summarises results of regressions on response rate to a post-graduation survey and to post-graduation migratory decision on the seminar-wise leave-me-out share of non-native speakers and the diversity index. Results by language background (native/non-native speakers) are derived from split sample models. Controls contain seminar size, age, gender and whether they are a native speaker or not. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered at the seminar level, are reported in parentheses.

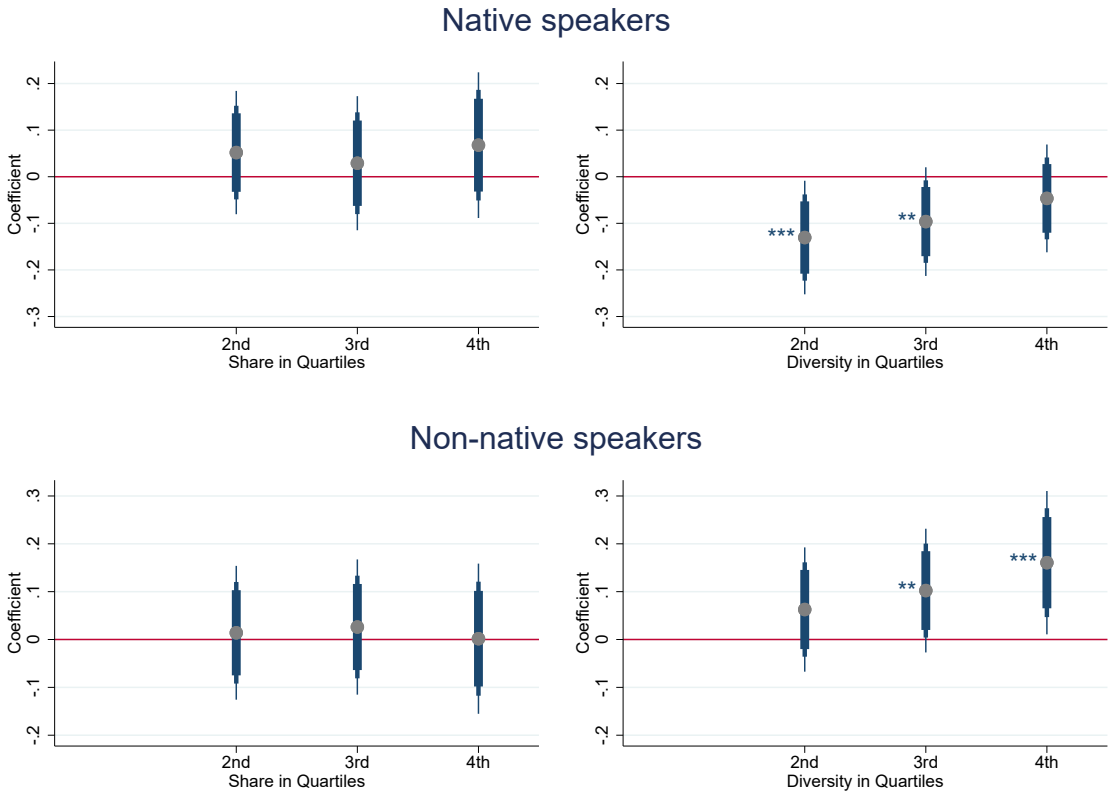
Appendix

Figure A1: Structure of teaching



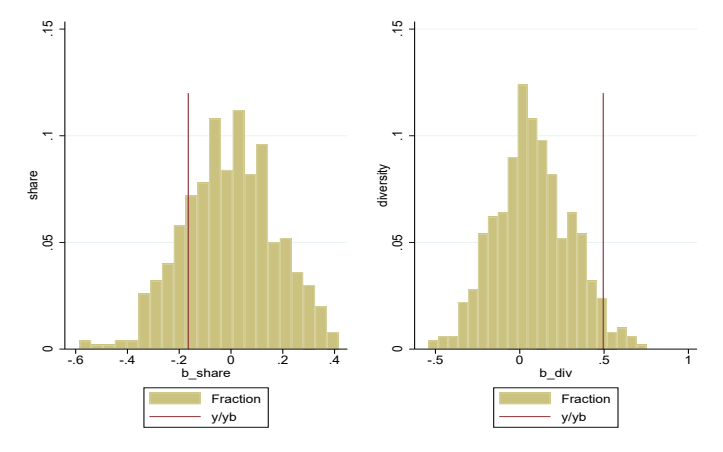
Notes: This figure describes the teaching structure of the institutional setting. Teaching happens in three consecutive years. Per year, students take four teaching units. In our specifications, we rely on exogeneous assignment into seminars within first- and second-year courses. Non-compulsory optional courses (grey) are not part of our sample. Third-year course choices are regarded as outcomes in Section 4.4.

Figure A2: Non-linear effects of Share and Diversity on Contemporaneous Grades



Notes: This figure shows the estimates of the effect of share of non-native speakers on grades, when share of non-native speakers and diversity is expressed in quartiles.

Figure A3: Distribution of placebo estimates



Notes: This figure shows the empirical distributions of placebo estimates for the share of non-native speakers (left) and ethno-linguistic diversity among those (right) on course grades. The cumulative distribution functions are based on 2000 estimates using a specification similar to the one displayed in column (2) of Table 3 lower panel and using random permutations of seminar ID to compute the treatments. The vertical line indicates the original estimate. Implied p-values are 0.19 (left) and 0.028 (right).

Table A1: Sample composition by language background

Language	Associated Nationalities	Number of Speakers	Share in sample (%)
ENGLISH	United States Ireland Australia New Zealand Kenya Uganda United Kingdom British Indian Ocean Territory British Overseas Citizen Nigeria Trinidad & Tobago Gambia Canada South Africa	971	44.46
MANDARIN	China Singapore Taiwan	420	19.23
RUSSIAN	Russia Kazakhstan	106	4.85
ITALIAN	Italy	68	3.11
CANTONESE	Hong Kong Macao	42	1.92
FRENCH	France	41	1.88
BULGARIAN	Bulgaria	39	1.79
KOREAN	North Korea South Korea	38	1.74
GERMAN	Germany Austria	33	1.51
POLISH	Poland	26	1.19
ARABIC	Bahrain Saudi Arabia Lebanon United Arab Emirates Libya Oman Morocco Kuwait Egypt Jordan Algeria	24	1.11
GREEK	Greece Cyprus	23	1.05
VIETNAMESE	Vietnam	21	0.96
SWEDISH	Sweden	20	0.92
PORTUGESE	Portugal Brazil Angola	18	0.82
SPANISH	Spain Mexico Columbia El Salvador	18	0.82
LITHUANIAN	Lithuania	17	0.78
WESTERN PUNJABI	Pakistan	17	0.78
AZERBAIJANI	Azerbaijan	17	0.78
HINDI	India	16	0.73
ALL OTHER (48)		209	9.57
Total Sample		2184	

Notes: This table gives the number of individual speakers of top 20 languages as well as the share of that particular language in our full sample.

Table A2: Raw and residual variation in key variables

	Mean	*SD*	*Min*	*Max*
Absolute				
Share of non-native speakers	0.54	0.14	0.21	0.87
Blau index of diversity	0.87	0.10	0.29	1.00
Residualised				
Share of non-native speakers	0.00	0.09	-0.29	0.33
Blau index of diversity	0.00	0.07	-0.43	0.23
No. of obs	*8744*			

Notes: This table shows variation in the share of non-native speakers and the diversity index, in absolute levels and in residualised after controlling for course×year, study programme, day×hour, and seminar leader fixed effects.

Table A3: Stability of coefficients

	(1)	(2)	(3)	(4)	(5)
Total					
Share of non-natives	-0.166*	-0.049	-0.067	-0.009	-0.043
	(0.095)	(0.096)	(0.101)	(0.118)	(0.117)
Blau Index	0.174	0.104	0.214	0.209	0.242*
	(0.132)	(0.131)	(0.139)	(0.136)	(0.138)
No. of observations	8744	8744	8744	8744	8744
Native					
Share of non-natives	0.129	0.111	0.091	0.078	0.094
	(0.134)	(0.137)	(0.148)	(0.167)	(0.163)
Blau Index	-0.059	-0.076	-0.058	-0.095	-0.075
	(0.170)	(0.170)	(0.165)	(0.165)	(0.159)
No. of observations	4032	4032	4032	4032	4032
Non-native					
Share of non-natives	-0.332**	-0.183	-0.196	-0.171	-0.165
	(0.132)	(0.133)	(0.139)	(0.170)	(0.167)
Blau Index	0.372**	0.251	0.428**	0.447**	0.496**
	(0.178)	(0.175)	(0.186)	(0.196)	(0.197)
No. of observations	4712	4712	4712	4712	4712
Course × term FE	yes	yes	yes	yes	yes
Study program FE	no	yes	yes	yes	yes
Day/Time FE	no	no	yes	yes	yes
Seminar leader FE	no	no	no	yes	yes
Individual controls	no	no	no	no	yes

Notes: This table summarises results of equation 1 regressing contemporary grades on the seminar-wise leave-me-out share of of non-native speakers and Blau Index. Columns differ by the sets of included fixed effects. Results by language background (native/non-native speakers) are derived from split sample models. Controls contain seminar size, age, gender and whether they are a native speaker or not. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered at the seminar level, are reported in parentheses.

Table A4: Robustness: Coefficient stability to seminar controls

	(1)	(2)	(3)	(4)
Sample	without ability controls		with ability controls	
	without Seminar Controls	with Seminar Controls	without Seminar Controls	with Seminar Controls
Total				
Share of non-natives	-0.043	-0.102	-0.048	-0.099
	(0.117)	(0.130)	(0.086)	(0.094)
Blau Index	0.242*	0.268*	0.205**	0.232***
	(0.138)	(0.139)	(0.092)	(0.088)
Mean of dep. var.	0.000	0.000	0.000	0.000
R^2	0.04	0.04	0.48	0.48
No. of observations	8744	8744	8680	8680
Native				
Share of non-natives	0.094	-0.033	0.028	-0.099
	(0.163)	(0.175)	(0.129)	(0.138)
Blau Index	-0.075	0.004	0.032	0.117
	(0.159)	(0.162)	(0.124)	(0.121)
Mean of dep. var.	0.079	0.079	0.079	0.079
R^2	0.08	0.08	0.46	0.46
No. of observations	4032	4032	4003	4003
Non-native				
Share of non-natives	-0.165	-0.121	-0.087	-0.063
	(0.167)	(0.178)	(0.119)	(0.138)
Blau Index	0.496**	0.455**	0.363***	0.339**
	(0.197)	(0.200)	(0.129)	(0.132)
Mean of dep. var.	-0.068	-0.068	-0.068	-0.068
R^2	0.07	0.07	0.51	0.51
No. of observations	4712	4712	4677	4677
Course × year FE	yes	yes	yes	yes
Study program FE	yes	yes	yes	yes
Day/Time FE	yes	yes	yes	yes
Seminar leader FE	yes	yes	yes	yes
Individual controls	yes	yes	yes	yes
Seminar controls	no	yes	no	yes
Individual ability	no	no	yes	yes
Peer ability	no	no	no	yes

Notes: This table summarises results of regressions of standardised grades on the seminar-wise leave-me-out share of non-native speakers and the diversity index. Results by language background (native/non-native speakers) are derived from split sample models. Individual controls contain age, gender and whether they are a native speaker or not. Seminar controls are share of females and average age. Own and peer ability is measured as leave-seminar-out GPA. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered at the seminar level, are reported in parentheses.

Table A5: Alternative language group definitions

	(1)	(2)	(3)
Sample	Baseline	Predominant languages	Nationality
Total			
Share of non-natives	-0.043	-0.051	-0.051
	(0.117)	(0.115)	(0.115)
Blau Index	0.242*	0.197	0.283*
	(0.138)	(0.146)	(0.162)
Mean of dep. var.	0.000	0.000	0.000
R^2	0.04	0.04	0.04
No. of observations	8744	8744	8744
Native			
Share of non-natives	0.094	0.057	0.023
	(0.163)	(0.161)	(0.163)
Blau Index	-0.075	-0.043	0.023
	(0.159)	(0.174)	(0.199)
Mean of dep. var.	0.079	0.084	0.100
R^2	0.08	0.08	0.08
No. of observations	4032	3970	3768
Non-native			
Share of non-natives	-0.165	-0.153	-0.110
	(0.167)	(0.167)	(0.160)
Blau Index	0.496**	0.406**	0.486**
	(0.197)	(0.197)	(0.205)
Mean of dep. var.	-0.068	-0.070	-0.076
R^2	0.07	0.07	0.07
No. of observations	4712	4774	4976
Course × year FE	yes	yes	yes
Study program FE	yes	yes	yes
Day/Time FE	yes	yes	yes
Seminar leader FE	yes	yes	yes
Individual controls	yes	yes	yes

Notes: This table summarises results of regressions of standardised grades on the seminar-wise leave-me-out share of non-native speakers using different assignments of language to nationality. Controls contain seminar size, age, gender and whether they are a native speaker or not. In column 2, students are given the predominant language of their country, and are not considered native speakers even if English is an official (but not-predominant) language. In column 3, only the UK nationals are considered to be native speakers. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered at the seminar level, are reported in parentheses.

Table A6: Alternative diversity definitions

	(1)	(2)	(3)	(4)	(5)
Sample	Baseline	No. of languages	No. of same lang.	Share of same lang.	At least one of same lang.
Total					
Share of non-natives	-0.043	-0.226*	–	–	–
	(0.117)	(0.134)	–	–	–
Diversity	0.242*	0.019***	–	–	–
	(0.138)	(0.006)	–	–	–
Mean of dep. var.	0.000	0.000	–	–	–
R^2	0.04	0.04	–	–	–
No. of observations	8744	8744	–	–	–
Native					
Share of non-natives	0.094	0.030	–	–	–
	(0.163)	(0.191)	–	–	–
Diversity	-0.075	0.007	–	–	–
	(0.159)	(0.008)	–	–	–
Mean of dep. var.	0.079	0.079	–	–	–
R^2	0.08	0.08	–	–	–
No. of observations	4032	4032	–	–	–
Non-native					
Share of non-natives	-0.165	-0.432**	-0.124	-0.113	-0.137
	(0.167)	(0.187)	(0.168)	(0.169)	(0.168)
Diversity	0.496**	0.027***	-0.013*	-0.466**	-0.089***
	(0.197)	(0.008)	(0.007)	(0.197)	(0.032)
Mean of dep. var.	-0.068	-0.068	-0.068	-0.068	-0.068
R^2	0.07	0.07	0.07	0.07	0.07
No. of observations	4712	4712	4712	4712	4712
Course × year FE	yes	yes	yes	yes	yes
Study program FE	yes	yes	yes	yes	yes
Day/Time FE	yes	yes	yes	yes	yes
Seminar leader FE	yes	yes	yes	yes	yes
Individual controls	yes	yes	yes	yes	yes

Notes: This table summarises results of regressions of standardised grades on the seminar-wise leave-me-out share of non-native speakers and different definitions of the diversity index and alternative measures for linguistic seminar composition. Controls contain seminar size, age, gender and whether they are a native speaker or not. In column 2, students are given the predominant language of their country, and are not considered native speakers even if English is an official (but not-predominant) language. In column 3, only the UK nationals are considered to be native speakers. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered at the seminar level, are reported in parentheses.

Table A7: The role of Mandarin speakers

	(1)	(2)	(3)	(4	(5)	(6)
	Share of Mandarin Speakers	Baseline	inc. Nbr Mandarin Speakers	inc. Share Mandarin Speakers	Mandarin Speakers only	Other Speakers only
Non-native						
Share of non-natives	0.291***	-0.165	-0.230	-0.179	-0.329	-0.156
	(0.021)	(0.167)	(0.185)	(0.209)	(0.228)	(0.340)
Blau Index	-0.634***	0.496**	0.715**	0.591*	0.411	0.815
	(0.030)	(0.197)	(0.319)	(0.339)	(0.443)	(0.604)
R^2	0.86	0.07	0.07	0.07	0.09	0.19
No. of observations	8744	4712	4712	4712	3202	1509

Notes: This table summarises results of different robustness checks examining the role of Chinese Mandarin speakers. Specification (1) explores the relationship between the share of Mandarin speakers and total share of non-native speakers as well as diversity. Specification (2) lists the baseline results similar to Table 3, column (2). Column (3) displays results controlling for the number of Chinese students in the seminar. Column (4) displays results controlling for the share of Chinese students. Columns (5) and (6) repeat this specification separately for Chinese and other non-native speakers. Controls contain seminar size, age, gender and whether they are a native speaker or not. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered at the seminar level, are reported in parentheses.

Table A8: Robustness of inference

	(1)	(2)	(3)	(4	(5)
	Baseline	i.i.d.	robust	course/year	year
Native					
Share of non-native	0.094	0.094	0.094	0.094	0.094
	(0.163)	(0.167)	(0.172)	(0.168)	(0.117)
Blau Index	-0.075	-0.075	-0.075	-0.075	-0.075
	(0.159)	(0.206)	(0.211)	(0.181)	(0.219)
R^2	0.08	0.08	0.08	0.08	0.08
No. of observations	4032	4032	4032	4032	4032
Non-native					
Share of non-natives	-0.165	-0.165	-0.165	-0.165	-0.165
	(0.167)	(0.163)	(0.163)	(0.193)	(0.274)
Blau Index	0.496**	0.496**	0.496**	0.496**	0.496**
	(0.197)	(0.205)	(0.205)	(0.215)	(0.231)
R^2	0.07	0.07	0.07	0.07	0.07
No. of observations	4712	4712	4712	4712	4712

Notes: This table summarises results of different robustness checks on inference. Specification (1) displays the baseline specification similar to Table 3, column (2) with standard errors clustered at the seminar level. Column (2) lists results assuming i.i.d. error terms. Column (3) lists results based on robust standard errors. Column (4) displays standard errors clustered on the course×year level. Column (5) applies standard errors clustered on the year level. Controls contain seminar size, age, gender and whether they are a native speaker or not. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A9: Romano-Wolf adjusted p-values

Total

	Share		Diversity	
	Baseline	Romano-Wolf	Baseline	Romano-Wolf
Grades, 1st year	0.4588	0.7285	0.0439	0.0758
Grades, 3rd year	0.0006	0.002	0.0226	0.0359
Share of foreigners	0.0652	0.0918	0.7416	0.7026
Share of quantitative courses	0.8556	0.8703	0.212	0.2615
Difficulty	0.7109	0.8703	0.0905	0.1138

Native speakers

	Share		Diversity	
	Baseline	Romano-Wolf	Baseline	Romano-Wolf
Grades, 1st year	0.2898	0.6088	0.511	0.9162
Grades, 3rd year	0.0787	0.1776	0.2628	0.6866
Share of foreigners	0.935	0.99	0.5778	0.9162
Share of quantitative courses	0.4025	0.6786	0.9535	0.9461
Difficulty	0.9065	0.99	0.4955	0.9162

Non-native speakers

	Share		Diversity	
	Baseline	Romano-Wolf	Baseline	Romano-Wolf
Grades, 1st year	0.9785	0.9661	0.006	0.01
Grades, 3rd year	0.0296	0.0259	0.046	0.0559
Share of foreigners	0.009	0.014	0.6494	0.5729
Share of quantitative courses	0.2305	0.3792	0.2406	0.2695
Difficulty	0.7927	0.9222	0.1395	0.1717

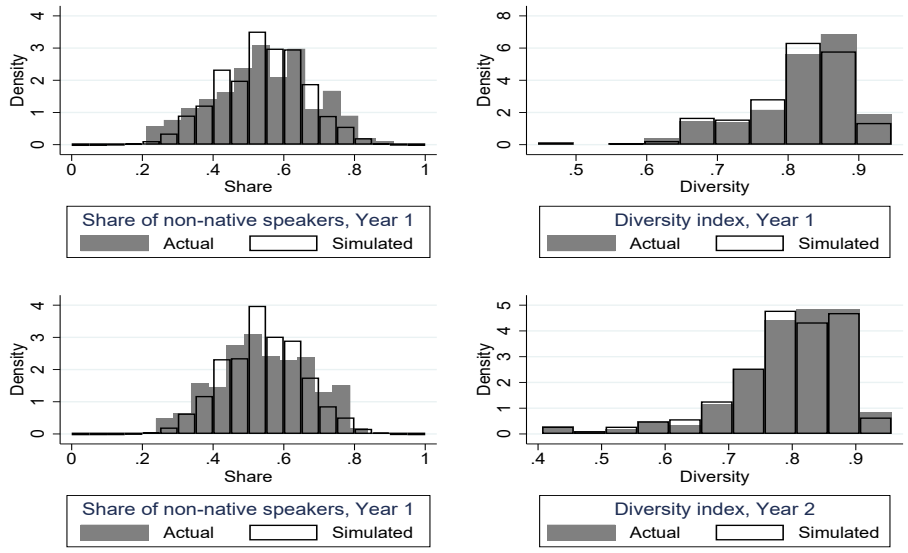
Notes: This table summarises p-values which are adjusted to account for multiple hypothesis testing within subsamples. P-values are adjusted using the Stata add-on *rwolf* by Clarke et al. (2019).

Table A10: Questions of the field survey

Question
How do you rate your English proficiency? **[Very bad (1) - Very good (5)]**
How comfortable do you feel speaking in English in this tutorial? **[Very uncomfortable (1) - Very comfortable (5)]**
For this course how often do you work with: Native English-speaking students **[Never (1) - Very often (5)]**
For this course how often do you work with: Non-native English-speaking students **[Never (1) - Very often (5)]**
How do you rate the level of English in the seminar discussions? **[Very bad (1) - Very good (5)]**

Supplementary online appendix

Figure S1: Simulated vs observed seminar composition by year



Notes: This figure compares observed distributions of the core variables of share of non-native speakers and diversity with simulated distributions based on pure random assignment based on 1,000 replications within courses, holding seminar sizes at observed levels. Variables are displayed as deviations from the $course \times term$ average.

Table S1: Seminar composition predicted by student characteristics

	Share of non-natives			Blau Index		
	Total	1st year	2nd year	Total	1st year	2nd year
Student's age	0.002	0.001	0.002*	0.001	0.000	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Gender: Female	0.002	0.004	0.001	-0.003	-0.003	-0.003
	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Language: Mandarin	0.007	0.004	-0.002	-0.000	0.007*	0.002
	(0.004)	(0.007)	(0.004)	(0.003)	(0.003)	(0.003)
Language: Russian	0.003	-0.000	-0.002	0.004	0.002	0.002
	(0.006)	(0.010)	(0.007)	(0.004)	(0.006)	(0.005)
Language: Italian	0.019**	0.013	0.009	0.003	0.005	0.002
	(0.007)	(0.011)	(0.008)	(0.006)	(0.005)	(0.009)
Language: Cantonese	0.012	0.013	-0.003	-0.013*	-0.005	-0.016*
	(0.008)	(0.013)	(0.009)	(0.006)	(0.008)	(0.008)
Language: French	0.006	-0.004	-0.001	-0.002	-0.002	0.001
	(0.008)	(0.012)	(0.009)	(0.005)	(0.007)	(0.007)
Language: Other	0.005	0.009	-0.003	-0.001	-0.001	-0.002
	(0.004)	(0.007)	(0.004)	(0.003)	(0.003)	(0.003)
N	8744	3672	5072	8744	3672	5072

Notes: This table summarises results of regressions of share of non-native speakers and Blau Index on observable student characteristics by year of study; each row represents a separate regression. Each regression includes the course/year-wise leave-me-out mean/share and fixed effects as indicated in the table. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table S2: Share of non-native speakers and educational performance

Sample	Grade	Fail	Honour
Total			
Share of non-natives	-0.044	0.051	0.015
	(0.118)	(0.037)	(0.053)
Mean of dep. var.	0.000	0.167	0.401
R^2	0.04	0.08	0.12
No. of observations	8744	8744	8744
Native			
Share of non-natives	0.095	0.030	0.022
	(0.164)	(0.058)	(0.072)
Mean of dep. var.	0.079	0.146	0.427
R^2	0.08	0.10	0.15
No. of observations	4032	4032	4032
Non-native			
Share of non-natives	-0.163	0.073	-0.002
	(0.169)	(0.055)	(0.074)
Mean of dep. var.	-0.068	0.185	0.379
R^2	0.07	0.10	0.15
No. of observations	4712	4712	4712
Course × year FE	yes	yes	yes
Study program FE	yes	yes	yes
Day/Time FE	yes	yes	yes
Seminar leader FE	yes	yes	yes
Individual controls	yes	yes	yes

Notes: This table summarises results of regressions of a set of outcome variables (standardised grade, indicator for failing a course, indicator for receiving an honour (60% or above) grade) on the seminar-wise leave-me-out share of non-native speakers only. Results by language background (native/non-native speakers) are derived from split sample models. Controls contain seminar size, age, gender and whether they are a native speaker or not. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered at the seminar level, are reported in parentheses.

Table S3: Share of non-native speakers and third-year choices

Sample	Share of Non-native speakers	Share of Numerical courses	Difficulty	Grade
Total				
Share of non-natives	0.011*	-0.005	0.001	0.321***
	(0.006)	(0.025)	(0.003)	(0.096)
Mean of dep. var.	0.549	0.537	0.051	0.000
R^2	0.68	0.33	0.62	0.08
No. of observations	7708	7708	7708	7708
Native				
Share of non-natives	0.000	-0.032	0.000	0.280*
	(0.008)	(0.038)	(0.003)	(0.164)
R^2	0.63	0.33	0.67	0.11
No. of observations	3645	3645	3645	3645
Non-native				
Share of non-natives	0.022***	0.039	0.001	0.322**
	(0.009)	(0.033)	(0.004)	(0.150)
R^2	0.68	0.33	0.62	0.09
No. of observations	4063	4063	4063	4063
Course × year FE	yes	yes	yes	yes
Study program FE	yes	yes	yes	yes
Day/Time FE	yes	yes	yes	yes
Seminar leader FE	yes	yes	yes	yes
Individual controls	yes	yes	yes	yes

Notes: This table summarises results of regressions of a set of outcome variables regarding course choices in third year on the seminar-wise leave-me-out share of non-native speakers only. Results by language background (native/non-native speakers) are derived from split sample models. Controls contain seminar size, age, gender and they are a native speaker or not. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors, clustered at the seminar level, are reported in parentheses.

Table S4: Testing for random assignment in survey sample

	Share of non-natives	Blau Index
Student's age	-0.00261**	-0.000759
	(0.00119)	(0.00107)
Gender: Female	0.0000784	-0.00350
	(0.00660)	(0.00409)
Language: Mandarin	-0.00912	0.0259***
	(0.00906)	(0.00926)
Language: Russian	-0.0339***	-0.00740
	(0.00690)	(0.0116)
Language: Italian	-0.0172	-0.0154**
	(0.0128)	(0.00673)
Language: French	-0.0196	0.00187
	(0.0120)	(0.0105)
Language: Other	-0.0112	-0.00754
	(0.00895)	(0.00705)
R^2	0.55	0.84
No. of observations	1025	1025

Notes: This table summarises results of regressions of share of non-native speakers and Blau Index from survey data on observable student characteristics; each row represents a separate regression. Each regression includes the course/year-wise leave-me-out mean/share and fixed effects as indicated in the table. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.