Education Exports and Human Capital Flows: Evidence from a Tuition Lottery

Pre-Analysis Plan

May 29, 2024

Daniel Firoozi
Claremont McKenna College

1. Introduction

Governments vie for skilled workers through strategic immigration and tax laws. Education exports have drawn considerable attention in this competition, with rich countries weighing policies to streamline the immigration process for nonresident college students. Yet, policy levers that change the direct cost of education exports, like nonresident tuition, attract relatively less scrutiny despite complex dynamics and theoretic trade-offs. In the short-run, nonresident students may cross-subsidize resident students through supplemental tuition fees. In the long-run, high skill nonresidents may influence the economic outcomes of residents through their impact on tax revenue, labor supply, labor demand, and innovation. Colleges price discriminate on residency to balance these competing objectives but, critically, lack information and an incentive to act on the externalities of their choices.

The result is considerable variation in nonresident tuition policies among research universities worldwide. At one end, institutions like Princeton and the University of Zurich post uniform tuition policies regardless of residency. At the other end, institutions like UC Berkeley and the University of Toronto discriminate sharply on residency, charging supplemental tuition regardless of whether a student is from another country or national subdivision. Intermediate approaches like those at Oxford or the Paris School of Economics offer lower rates for domestic and select international students while imposing higher fees on others.

It is striking that there are no experimental estimates of the long-run social benefits and costs to nonresident supplemental tuition as universities navigate the trade-offs. Namely, higher nonresident tuition could plausibly aid in cost recovery for instruction but risks deterring human capital inflows from people who have or will develop in-demand entrepreneurial, research, or executive skills. The misaligned incentives between universities and governments are a further challenge, particularly when the former do not internalize the fiscal externalities their policies deliver to the latter.

I plan to advance the literature on human capital and migration by offering the first experimental evidence on the long-run social benefits and costs of nonresident tuition. I will

use a pre-analysis plan and data from a randomized tuition lottery implemented in 2012 at a member of the Association of American Universities (AAU), an elite group of North America's top 71 research institutions. Under the lottery, 1,333 international and domestic nonresident students from 45 countries who were admitted to the elite research university were randomly assigned waivers that would reduce nonresident supplemental tuition by 20,000 dollars, 30,000 dollars, or 40,000 dollars over four academic years. More than 80 percent of these students were neither American citizens nor legal permanent residents, including close to half of all "out-of-state" students. Because nonresident tuition was assigned by a computer-randomized lottery, identification of causal effects follows from a simple comparison of the outcomes across treatment arms akin to a randomized control trial (RCT).

2. Research Objectives

The primary objective of this research paper is to examine the social costs and benefits of nonresident tuition. Specifically, the study aims to:

1) Analyze the short-term revenue implications of nonresident tuition fees, considering the price elasticity of international and domestic nonresident enrollment.

2) Investigate the long-term effects of nonresident tuition on immigration patterns and citizenship rates.

3) Estimate the external costs and benefits of nonresident tuition and assess the potential impact of price discrimination on the overall calculus of nonresident tuition policies.

4) Identify spillovers from nonresident students to resident students' GPAs, major choices, and degree completion.

3. Data Collection

This is a secondary analysis of existing data rather than the implementation of a new RCT. Existing data will be collected in phases by the principal investigator and any research assistants, which allows for the pre-registration of this analysis plan. Data collection and receipt will proceed in three phases: a “test-run dataset” has been used to validate the research design, collection and linkage of publicly available data into an “initial dataset” will take place in the coming weeks and months, and linkage of publicly available data with academic records followed by de-identification will yield the “finalized dataset” which will be used for analysis.

In the first phase, the de-identified “test-run dataset” on a sample of 1,333 first-time, full-time, non-resident college applicants to an undergraduate program at a large American research university were provided to the principal investigator. Students’ treatment status, a binary indicator for whether or not they enrolled at the research university, and demographic characteristics of the students were included in the dataset. This enabled balance tests to assess whether or not treatment was actually randomly assigned per the tuition lottery’s protocol. I implemented balance tests using both predicted enrollment at the research university and the individual demographics of students. Both types of test confirmed that the computer-randomized lottery was successful in generating treatment that was uncorrelated with predicted outcomes and sample demographics.

In the intermediate phase, the research university will provide publicly available directory information as part of an “initial dataset” on the sample of 1,333 students that the principal investigator will link to other publicly available datasets. Specifically the PI and any research assistants will be provided the following publicly available directory information: Name, Date of birth, Date of college attendance, Major field of study, and Degree received. Under FERPA § 99.37, these are considered directory information that may be disclosed publicly. The principal investigator and any research assistants will then link these directory data to LinkedIn employment records and US voter registration files using names and dates

(See Methods Section for details). Importantly the “initial dataset” will not include information on the treatment arm or treatment status of the individuals, preventing the principal investigator and research assistants from either intentionally or unintentionally tabulating outcomes differently based on treatment arm. This will mark the completion of the “initial dataset”.

In the final phase, the researchers will return the initial dataset to the research university, which will link it to academic records on the 1,333 in-sample students along with another set of contemporaneously enrolled resident students. These academic records will include a full list of variables specified in the Methods Section, will be de-identified, and will then be returned to the principal investigator for analysis. The PI will then follow the methods outlined in this PAP to generate and report results.

4. Methods

4.a General Methods

I will use data from a randomized tuition lottery implemented in 2012 at an anonymous major American research university. 1,333 international and domestic non-resident students from 44 countries and the United States who were admitted to the university were randomly assigned to receive tuition waivers that would reduce nonresident supplemental tuition by $20,000 (444 students), $30,000 (444 students), or $40,000 (445 students) in nominal terms over the course of four academic years. The research university admitted an additional 1,000 to 2,000 international students who were at random selected not to participate in the lottery, but did not retain records on this set of students. The number of students subject to the lottery was chosen to ensure that total tuition waiver offers summed to 10 million dollars. Approximately 83 percent of lottery students were neither American citizens nor legal permanent residents, including 44 percent of all “out-of-state” nonresident students. Because nonresident tuition waivers were randomly assigned, the causal effect of nonresident tuition is equal to the OLS estimate of the association between outcome variables and the amount of nonresident tuition requested (or waived).

Using the test-run dataset, I was able to validate that treatment was randomly assigned. I did so by testing for balance on observable characteristics and by regressing predicted enrollment at the research university on treatment. The observable characteristics I use both independently and to generate predicted outcomes are: (1) an indicator for being raised by a single parent, (2) a head count of a student's household size, (3) an indicator for no data on household size, (4) self-reported family income, (5) an indicator for having reported family income, (6) an indicator for being a first generation college student, (7) estimated age at college entry, (8) an indicator for no age data, (9) a student's best SAT or ACT equivalent score, (10) an indicator for no standardized test score, (11) a student's unweighted high school GPA, (12) a student's overall admission score, (13) a FAFSA filing indicator, (14) an indicator for honors admission, (15) an indicator for receiving a separate merit scholarship offer, (16) an indicator for self-identifying as female, (17) a categorical variable for father's education, (18) a categorical variable for mother's education, (19) a categorical variable for the school/department of the major the student listed as their first preference, (20) a categorical variable for home country or region of a student's mailing address, (21) an overall academic rating of the student from the research university, (22) a holistic score, (23) a student's expected family contribution (EFC) value, (24) an indicator for no EFC data, and (25) a categorical variable for ethnic identity. I plan to reproduce these balance checks in my final analysis by performing the predicted outcome test for each full-sample outcome specified in this section as well as a summary table of the balance checks on each observable characteristic.

Beyond tests for balance, differential attrition is a plausible cause for concern. Considering the risk of attrition, I note that students cannot attrit from binary decisions over whether or not to enroll in colleges, whether or not to graduate from colleges, and whether or not to live in the United States. The more serious attrition risk comes from students being unobservable in my outcome data, but I note that the research university's enrollment data, voter registration records, and National Student Clearinghouse data on college enrollment and graduation are near complete (>95%) records. Some measures of long-run immigration and

earnings (like the LinkedIn outcomes) will miss a minority of students residing in America for whom the data is not available, but there is no obvious reason why, other than changes in immigration patterns, attrition from observed outcomes should be correlated with the magnitude of nonresidential tuition fees other than through immigration and assimilation induced by tuition fees.

Turning to Hawthorne or placebo effects and John Henry effects, I note that this tuition lottery was effectively a single-blind RCT. All 1,333 in-sample students knew they had been offered a tuition waiver, but they were not aware that the value of the tuition waiver had been randomly assigned or that other students had been offered nonresident tuition waivers of different values than their own. Essentially none nonresident students originated from the same high school, making it unlikely that they would have been able to identify one another or communicate with one another. The fact that all students were treated and students didn't know they were part of a random lottery means that comparison of outcomes between treatment arms effectively eliminates behavioral responses that would appear along the extensive margin of treatment but are constant across the intensive margin of treatment. This also means that spillover effects between students subject to the lottery are unlikely and the Stable Unit Treatment Value Assumption (SUTVA) is likely to be satisfied.

To estimate causal effects, I will use the following generalized specification:

$$Y_i = \alpha + \beta Waiver_i + \mathbf{X}_i'\Gamma + \varepsilon_i$$

where Y_i is an outcome of interest for student i, $Waiver_i$ is the net present value (in 2012) in thousands of dollars of the tuition waiver assuming an annual discount rate of 5%, $\mathbf{X}_i$ is a vector of covariates listed in the preceding section, and ε_i is an idiosyncratic error term. In this context, $\hat{\beta}$ is our estimate of the average treatment effect of the tuition waiver on the outcome of interest. I plan to vary the inclusion of covariates for each outcome of interest and to use linear probability models for ease of interpretability with binary outcomes.

I note that all of the estimated impacts of nonresident tuition on various outcomes are likely to be a lower bound (biased toward zero) because the sample of students is observed conditional on admission to the research university at its current posted sticker price of nonresident tuition. If nonresident prospective students were to see lower sticker prices, that could increase application rates and admission rates making all of my outcomes of interest more sensitive to nonresident tuition prices. In the case of estimating tuition recovery, this will mean that higher enrollment elasticity to prices should be observed, biasing the short-run estimated tuition recovery benefits of nonresident tuition upward and the long-run costs of non-resident tuition downward.

4.b Short-Run Outcome Measures

The first set of outcomes I intend to study are short-run outcomes related to college enrollment decisions in the year of the tuition lottery, 2012. Specifically I will estimate the impact of tuition waivers on the following outcomes:

1) A binary indicator for enrollment at the research university in 2012 from the university's own records. This is the only outcome variable that was available to the PI and linked to treatment status prior to the writing of the pre-analysis plan.
2) A binary indicator for enrollment at any college campus, public or private, in the target state that is home to the research university in 2012 from the National Student Clearinghouse, which is retained by the research university.
3) A binary indicator for enrollment at any college campus in the United States in 2012 from the National Student Clearinghouse, which is retained by the research university.

I intend to estimate the impact of waiver size on each of these outcomes in a single table. Each outcome will have two specifications, one with covariates from section 4.a and one without covariates. The minimum detectable effects on these outcomes at a 95 percent confidence interval is at most 0.37 percentage points per $1,000 in tuition before adjustment for multiple hypothesis testing.

I expect that feedback from referees, seminar participants, and other researchers will include the suggestion that I link the specific colleges at which students enroll to other datasets, like IPEDS or Opportunity Insights. Ex ante, I do not have a strong prior on which characteristics to use and do not intend to pre-specify any such analyses. I will note any deviations from this in an appendix of the final paper that is submitted to a peer-reviewed journal if a compelling suggestion arises through the process of sharing the manuscript.

4.c Medium-Run Outcome Measures

The second set of outcomes I intend to study are medium-run outcomes related to two topics: college and postgraduate degree attainment among within-sample nonresident students and spillover outcomes to resident students. With respect to the first set of outcomes, I will estimate the impact of tuition waivers on the following:

1) A binary indicator for bachelor's degree attainment after 2012 in the United States from the National Student Clearinghouse, which is retained by the research university.
2) A binary indicator for STEM bachelor's degree attainment after 2012 in the United States from the National Student Clearinghouse, which is retained by the research university.
3) A binary indicator for postgraduate degree attainment after 2012 in the United States from the National Student Clearinghouse, which is retained by the research university.
4) A binary indicator for STEM postgraduate degree attainment after 2012 in the United States from the National Student Clearinghouse, which is retained by the research university.

I intend to estimate the impact of waiver size on each of these outcomes in a single table (subject to space constraints). Each outcome will have two specifications, one with covariates from section 4.a and one without covariates.

Because I found positive impacts of the tuition waiver on enrollment at the research university, null impacts on these outcomes would be interesting because they would suggest that tuition can be recovered from nonresident students without tangibly reducing the rate at which they consume and complete American postsecondary education. The minimum

detectable effects on these outcomes at a 95 percent confidence interval is at most 0.37 percentage points per $1,000 in tuition before adjustment for multiple hypothesis testing.

With respect to the second set of outcomes, I intend to estimate the impact of the share of nonresident students within a given major on the outcomes of domestic resident students within that major. Specifically, I will instrument for the share of nonresident students enrolled within a major at the research university using the following first-stage equation:

$$Nonres_m = \phi_0 + \phi_1 \overline{Waiver}_m + \phi_2 (NonresApp_m \times \overline{Waiver}_m) + \phi_3 NonresApp_m + u_m$$

where $Nonres_m$ is the share of enrolled students who are nonresident tuition lottery participants within a given major indexed by m, $\overline{Waiver}_m$ is the average tuition waiver offered to nonresident lottery participants within a major and is one of the excluded instruments, $NonresApp_m$ is a control for the share of admitted students who are nonresident tuition lottery participants within a given major, the interaction term $NonresApp_m \times \overline{Waiver}_m$ is another excluded instrument, and u_m is an idiosyncratic error term. Using two stage least squares I will estimate the outcomes of resident students from the 2012 cohort using the following specification:

$$Y_{i,m} = \alpha + \theta \widehat{Nonres}_m + \mathbf{X}'_{i,m}\Omega + \varepsilon_{i,m}$$

where $Y_{i,m}$ is an outcome of interest for student i in major m, $\widehat{Nonres}_{i,m}$ is the predicted share of nonresident tuition lottery participant enrollment in major m, $\mathbf{X}_{i,m}$ is a vector of student i's pre-treatment demographic characteristics from section 4.a (although it is possible that fewer covariates will be available for this sample), and $\varepsilon_{i,m}$ is an error term. The estimate $\hat{\theta}$ will identify the impact of increasing the share of nonresident students within a major on the

outcomes of resident students. I plan to cluster standard errors on major and only report results if the first stage F-statistic exceeds a value of 100.

Using this approach, I will estimate the impact of nonresident students on the following outcomes:

1) 2012 cohort resident student's first year cumulative GPA from the research university's records.
2) A binary indicator for a 2012 cohort resident student's extensive margin of ever graduating with a bachelor's degree from the research university's records.
3) A binary indicator for a 2012 cohort resident student's extensive margin of ever graduating with a STEM bachelor's degree from the research university's records.

I intend to estimate the impact of waiver size on each of these outcomes in a single table. Each outcome will have two specifications, one with covariates from section 4.a and one without covariates. Null outcomes would be of interest here because they would suggest that nonresident students do not meaningfully crowd out human capital attainment by resident students within the same field of study.

4.d Long-Run Outcome Measures

The final set of outcomes I intend to study are long-run outcomes related to three topics: citizenship and assimilation, immigration and impacts on the labor market within the target state, and immigration and impacts on the labor market within the United States. With respect to the first set of outcomes, I will estimate the impact of tuition waivers on the following:

1) A binary indicator variable for naturalization and assimilation measured through voter registration in the target state from L2 Inc. voter file records from the target state. Voter registration records are among the most comprehensive registries over American citizens. I will drop this outcome if less than 5% of the sample is registered to vote in the target state.
2) A binary indicator for naturalization and assimilation measured through voter registration anywhere in the United States from L2 Inc. records on every American state and territory's

voter files. I will drop this outcome if less than 5% of the sample is registered to vote in the United States.

I intend to estimate the impact of waiver size on each of these outcomes in a single table. Each outcome will have two specifications, one with covariates from section 4.a and one without covariates. Null outcomes would be of interest here because they would suggest that nonresident tuition does not reduce future citizenship rates or assimilation by nonresident students. It is worth noting that less than 10 percent of students in this sample were citizens at the time of college application, per the research university's records, meaning that registration rates in excess of this number imply a high rate of naturalization and assimilation among these students.

With respect to the next set of long-run outcomes, I will estimate the impact of tuition waivers on the following:

1) A binary indicator for entrepreneurship in the target state from LinkedIn, Revelio, and L2 voter file records. This indicator will be generated by interacting the LinkedIn indicator for entrepreneurship (defined later in this section) with an indicator for residing in the target state (defined as either being registered to vote in the target state in 2024 or listing a metropolitan area in the target state as a place of residence on LinkedIn).
2) A binary indicator for innovation in the target state from LinkedIn, Revelio, and L2 voter file records. This indicator will be generated by interacting a LinkedIn indicator for innovation with an indicator for residing in the target state (defined as either being registered to vote in the target state in 2024 or listing a metropolitan area in the target state as a place of residence on LinkedIn).
3) A binary indicator for executive leadership in the target state from LinkedIn, Revelio and L2 voter file records. This indicator will be generated by interacting the sum of a LinkedIn indicator for executive leadership (defined later in this section) with an indicator for residing in the target state (defined as either being registered to vote in the target state in 2024 or listing a metropolitan area in the target state as a place of residence on LinkedIn).

4) Estimated earnings in the target state, which will be generated by interacting estimated earnings from LinkedIn occupation titles (from Revelio Labs or BLS data if Revelio Labs data is unavailable) with an indicator for residing in the target state (defined as either being registered to vote in the target state in 2024 or listing a metropolitan area in the target state as a place of residence on LinkedIn).

I intend to estimate the impact of waiver size on each of these outcomes in a single table (if space allows). Each outcome will have two specifications, one with covariates from section 4.a and one without covariates. Null outcomes would be of interest here because they would suggest that nonresident tuition does not reduce the supply of high skill labor or labor demand generated by entrepreneurial or innovative immigrants. If any of these outcomes are positive for less than 2 percent of the total sample, I will drop them as an outcome.

With respect to the last set of long-run outcomes, I will estimate the impact of tuition waivers on the following:

1) A binary indicator for entrepreneurship in the United States from LinkedIn, Revelio, and L2 voter file records. This indicator will be generated by interacting the LinkedIn indicator for entrepreneurship (defined later in this section) with an indicator for residing in the US (defined as either being registered to vote in the US in 2024 or listing a metropolitan area in the US as a place of residence on LinkedIn).
2) A binary indicator for innovation in the United States from LinkedIn, Revelio, and L2 voter file records. This indicator will be generated by interacting the a LinkedIn indicator for innovation (defined later in this section) with an indicator for residing in the US (defined as either being registered to vote in the US in 2024 or listing a metropolitan area in the US as a place of residence on LinkedIn).
3) A binary indicator for executive leadership in the United States from LinkedIn, Revelio, and L2 voter file records. This indicator will be generated by interacting the sum of a LinkedIn indicator for executive leadership (defined later in this section) with an indicator for residing

in the US (defined as either being registered to vote in the US in 2024 or listing a metropolitan area in the US as a place of residence on LinkedIn).

4) Estimated earnings in the United States, which will be generated by interacting estimated earnings from LinkedIn occupation titles with an indicator for residing in the target state (defined as either being registered to vote in the target state in 2024 or listing a metropolitan area in the target state as a place of residence on LinkedIn).

I intend to estimate the impact of waiver size on each of these outcomes in a single table (if space allows). Each outcome will have two specifications, one with covariates from section 4.a and one without covariates. Null outcomes would be of interest here because they would suggest that nonresident tuition does not reduce the supply of high skill labor or labor demand generated by entrepreneurial or innovative immigrants. If any of these outcomes are positive for less than 5 percent of the total sample, I will drop them as an outcome.

I intend to define the variables from LinkedIn by collecting the following data and using the following definitions:

1) LinkedIn Location: defined as the metropolitan area listed on LinkedIn profiles. This variable will be set to the country if the metropolitan area is missing. These data will come from Revelio Labs and manually collected records.
2) LinkedIn Earnings: defined as the imputed earnings based on work history and job title and from Revelio Labs' individual dataset. In cases where this is absent, we will link job titles to the most similar BLS occupation code and its annual mean wages. These data will come from Revelio Labs and manually collected records. I will assume students work for 20 years at a constant level of earnings in their recorded place of residence beginning 8 years after college application to be conservative. Earnings will be imputed for people without LinkedIn job titles by assuming an annual mean earnings level equal to the sample average estimated mean annual wage for students whose occupational titles I observe.

3) LinkedIn Entrepreneurship: defined as having a relevant term in *any* part of the LinkedIn profile. The relevant terms are: Entrepreneur, Founder, Co-founder, Creator, Startup, Owner, CEO, Venture, Investor, or Strategist.
4) LinkedIn Innovation: defined as having a relevant term in *any* part of the LinkedIn profile. The relevant terms are: Inventor, Patent, Innovation/Innovator, Developer/Development, Research, Scientist, Engineer, Technology/Technologist, Design, Data, Idea, or Lab/ Laboratory.
5) LinkedIn Executive Experience: defined as having a relevant term in *any* part of the LinkedIn profile. The relevant terms are: Chief, Officer, President, Director, Board, Executive, Chair/Chairman, Manager/Management/Managing, Partner, Head, Lead, or Senior.

4.e Marginal External Benefits and Costs

To limit researcher discretion, this subsection pre-specifies how I intend to estimate and calculate the social costs and benefits of nonresident tuition.

The primary benefit of nonresident tuition is the short-run recovery of "profit" (defined as net tuition less instructional expenditures) from nonresident students that may be used to cross subsidize resident undergraduate students. To calculate profit, I will begin by calculating total revenue. I will assume that total revenue equals the net present value (assuming a 5 percent annual discount rate) of the total sticker price of tuition less the randomly assigned tuition waiver for four years for nonresident students[1]. By tuition, I refer specifically to the sum of mandatory charges for nonresidents plus official nonresident supplemental tuition fees, excluding other costs of attendance like housing, room and board, and books and excluding non-instructional fees for other student services. I will then calculate profit as the difference between total revenue and the net present value (again assuming a 5 percent annual discount rate) of instructional expenditures per capita from IPEDS for four years of instruction. Finally, I will interact profit with an indicator for enrolling at the research university and use this measure

[1] I note that nonresidents receive essentially no other financial aid at the research university and fewer than 2 percent receive merit scholarships.

as an outcome variable of interest to estimate the number of dollars recovered per 1,000 dollars of posted nonresident tuition. This is mathematically equivalent to interacting the net present value of tuition payments with the indicator used as the first outcome in Section 4.b of this PAP.

There are two reasons why this method is likely to overestimate the social benefits of nonresident tuition. First, this method will overestimate tuition recovery because the tuition lottery occurred among students who were already admitted and therefore recovers a lower bound on elasticity of enrollment to prices by missing out on the elasticity of application rates to prices. Second, it overestimates tuition recovery because it assumes students pay net tuition fees over a four year time period, rather than assuming that some minority of students drop out.

The primary external cost of nonresident tuition would be the loss of high-skill immigrants along with their attendant taxable earnings and externalities. To address the former, we will repeat our estimates for taxable earnings from Section 4.d and then calculate tax revenue by multiplying a students' LinkedIn earnings by an effective tax rate of 25 percent, which should be a lower bound estimate. I believe this is a lower bound estimate, because these students are likely to be high income earners facing higher effective tax rates than the mean American and the 25 percent rate is slightly lower than the United States' total revenue to GDP ratio in recent years. Because there is not an obvious way to calculate the fiscal externalities from having a single innovator in the US, I will not include adjustments for the fiscal externalities of immigration by innovators, entrepreneurs, and executives in my estimates. I will defer to suggestions from referees/seminar participants if I receive any and will note any deviations from this PAP in any future manuscripts. Since I am omitting externalities from innovation and because I am missing out on the elasticity of college application rates to prices, the results I generate will be a lower bound on the external costs of non-resident tuition.

The estimated external costs of non-resident tuition will be generated for both the target state and the United States and will be calculated both with and without covariates.

These results should be a strict lower bound on the cost to benefit ratio due to the assumptions I impose.

4.f Heterogeneous Treatment Effects and Pre-registered Subgroup Analyses

I plan to examine heterogenous treatment effects along two dimensions: country of home address (which does not necessarily coincide with residence) and STEM major intent. I will estimate the marginal external costs and benefits of nonresident tuition for three groups for the former:

1) All nonresident students with a primary home address in China, Taiwan, or Hong Kong. These students are close to half the full sample (45 percent).
2) All nonresident students with a primary home address outside of the United States (including those with a primary residence in China). These students are close to two thirds of the sample (66 percent).
3) All nonresident students with a primary home address in the United States. These students are close to a third of the sample (33 percent).

I note that primary home address does not align perfectly with citizenship which does not align perfectly with specific type of nonresidency (domestic out of state or foreign status). I find this specific dimension of heterogeneity to be more compelling than the two alternatives because it gives a better sense of the region that students would designate as their home absent financial incentives to misreport.

With regard to STEM major intent, I will split the sample into binary groups of STEM and non-STEM major based on whether their first preference major at the research university is a CIP-designated STEM major by the US Department of Homeland Security. This is a useful measure of STEM status because the ease of immigration laws in the United States is relaxed for students completing a CIP designated STEM Major.

I expect referees and other researchers providing feedback on this paper may ask for heterogeneity by socioeconomic status (specifically reported family income) or academic

characteristics (like Best SAT or ACT equivalent). I do not intend to estimate these ex ante because there are multiple measures of each, but I will note any deviation from this plan in an appendix of any manuscript submitted to a peer-reviewed journal.

4.g Multiple Hypothesis Testing Correction

It is worth noting that the pre-registration of regression specifications and outcome variables should allay concerns about specification searching, but nonetheless I plan to correct for multiple hypothesis testing because of the number of outcomes being tested. Specifically, I plan to control for the false discovery rate using the Simes procedure and apply this correction within each specification for each set of outcomes, which are denoted by organized sections with numbered lists in the text of this pre-analysis plan. One interpretation of this method is that for the estimated coefficients with q-values of less than 0.050, I can reject the null hypothesis that all of the coefficients are null effects at a 5 percent level (or a 95 percent confidence interval). Since the hypotheses I am testing are broad rather than specific to a single outcome (e.g. inflows of some types of high skill immigrants is an important outcome rather than immigration of those with executive skills specifically), the Simes procedure is a sensible approach relative to more conservative methods of multiple hypothesis testing.