Spring 3-1-2023

# Essays in Education Policy Analytics: Prediction of At-Risk Students, International Mobility, Cognitive Trade-Offs

Matteo Zullo

# ESSAYS IN EDUCATION POLICY ANALYTICS: PREDICTION OF AT-RISK STUDENTS, INTERNATIONAL MOBILITY, COGNITIVE TRADE-OFFS

A Dissertation
Presented to
The Academic Faculty

By

Matteo Zullo

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Public Policy
Ivan Allen College of Liberal Arts

Georgia Institute of Technology
Georgia State University

May 2023

**ESSAYS IN EDUCATION POLICY ANALYTICS: PREDICTION OF AT-RISK STUDENTS, INTERNATIONAL MOBILITY, COGNITIVE TRADE-OFFS**

Thesis committee:


Dr. Ross Rubenstein
Andrew Young School of Policy Studies
*Georgia State University*

Dr. Diana Hicks
School of Public Policy
*Georgia Institute of Technology*

Dr. Tim R. Sass
Andrew Young School of Policy Studies
*Georgia State University*

Dr. Juan Rogers
School of Public Policy
*Georgia Institute of Technology*

Dr. Audrey J. Leroux
College of Education & Human Development
*Georgia State University*


Date approved: March 1, 2023

## TABLE OF CONTENTS

**List of Tables** . . . . . vi

**List of Figures** . . . . . viii

**List of Acronyms** . . . . . ix

**Summary** . . . . . x

**Chapter 1: Georgia State University's Graduation and Progression Success Advising: Academic Fit Effects from Learning Analytics** . . . . . 1

1.1 Introduction . . . . . 1

1.2 Literature Review . . . . . 4

1.2.1 Learning analytics and advising . . . . . 4

1.2.2 The information value of academic signals . . . . . 6

1.3 Methodology . . . . . 7

1.3.1 Matching design . . . . . 8

1.3.2 Relative academic strength . . . . . 11

1.3.3 Accounting for GPA inflation . . . . . 12

1.4 Empirical Section . . . . . 15

1.4.1 Theory of action . . . . . 16

1.4.2 Effects of algorithmic student advising . . . . . 18

1.5 Discussion . . . 27

1.5.1 Conclusions . . . 27

1.5.2 Limitations . . . 29

**Chapter 2: The Role of Higher Education Agencies in Promoting International Student Mobility: The Expansion of German DAAD's Outbound Network . . . 31**

2.1 Introduction . . . 31

2.2 Background . . . 33

2.2.1 The student migration decision . . . 35

2.2.2 Gravity equations of international student mobility . . . 37

2.3 Methodology . . . 39

2.3.1 Data . . . 39

2.3.2 Standard two-way fixed effects estimation . . . 41

2.3.3 Synthetic difference-in-differences . . . 43

2.4 Results . . . 49

2.4.1 Main estimates . . . 49

2.4.2 Estimates by World Bank income group . . . 52

2.5 Conclusions . . . 54

**Chapter 3: (No) Trade-Off Between Numeracy and Verbal Reasoning Development: Programme for International Student Assessment (PISA) Evidence from Italy's Academic Tracking . . . 56**

3.1 Introduction . . . 56

3.2 Background . . . 58

3.2.1 Academic Strength and Skill Formation . . . 58

3.2.2 The "STEM Advantage" . . . 61
3.2.3 International Education Production Functions . . . 62
3.2.4 PISA: Testing Functional Knowledge . . . 64
3.2.5 Current Study . . . 65
3.3 Methods . . . 66
3.3.1 Data . . . 66
3.3.2 Analytical Model . . . 68
3.4 Results . . . 73
3.4.1 Matching . . . 73
3.4.2 Achievement Decomposition . . . 75
3.4.3 Altonji Decomposition . . . 78
3.5 Robustness Checks . . . 79
3.6 Conclusions . . . 83

**Appendices** . . . 85
Appendix A: Essay II . . . 86
Appendix B: Essay III . . . 89

**References** . . . 93

## LIST OF TABLES


1.1 Sample descriptives of matching features used for inverse probability weighting . . . . . 9

1.2 Relative Academic Strength (RAS) value for mock students with varying aptitude . . . . . 13

1.3 Odds ratios of college graduation for Graduation and Progression System (GPS) advised students . . . . . 22

1.4 Odds ratios of survival in college for students who dropped out . . . . . 24

1.5 Grade inflation-adjusted GPA at graduation . . . . . 26

1.6 Credit hours taken in the student's respective major area . . . . . 26

1.7 Distance between students and their graduation major . . . . . 27

2.1 DAAD office openings by country 1996-2016 . . . . . 34

2.2 Gravity model estimates from selected single-destination studies . . . . . 40

2.3 Average office numbers effect from standard two-way fixed effects . . . . . 51

2.4 Grand-average office foundations effect from synthetic difference-in-differences 52

2.5 Averag estimated effects by World Bank income group . . . . . 54

3.1 Instructional units in year I and II of high school . . . . . 58

3.2 Instructional units in year I and II of high school . . . . . 58

3.3 Individual and parental characteristics . . . . . 68

3.4 School characteristics . . . . . 69

3.5 Mean PISA test scores before and after inverse probability matching . . . . 74

A.1 Selection of model specification using Weight of Evidence test . . . . . 87

A.2 Country-average office foundations effect from synthetic difference-in-differences 88

B.1 Standardized mean differences before and after matching . . . . . 89

B.2 Predictors of PISA math score . . . . . 90

B.3 Predictors of PISA reading score . . . . . 91

B.4 Predictors of PISA science score . . . . . 92

## LIST OF FIGURES

1.1 Relative Academic Strength (RAS) by major in descending order of STEM fit . . . . . 14

1.2 Odds ratios of college graduation for GPS students who updated their majors and who stayed in their intended majors . . . . . 21

3.1 Standardized mean differences before and after inverse probability matching 74

3.2 Decomposition of the STEM-Liberal Arts PISA score gap . . . . . 76

3.3 Projected math and reading proficiency progression . . . . . 81

3.4 Projected math and reading proficiency progression internationally . . . . . 81

A.1 Residuals versus fitted values plots across model specifications . . . . . 86

## LIST OF ACRONYMS

**AICc** Akaike Information Criterion corrected

**DAAD** Deutscher Akademischer Austauschdienst

**GPS** Graduation and Progression System

**GSU** Georgia State University

**LA** Learning Analytics

**MAAPS** Monitoring Advising Analytics to Promote Success

**PIRLS** Progress in International Reading Literacy Study

**PISA** Programme for International Student Assessment

**RAS** Relative Academic Strength

**RUM** Random Utility Maximization

**TIMSS** Trends in International Mathematics and Science Study

**WoE** Weight of Evidence

## SUMMARY

The dissertation includes three essays contributing to our understanding of human capital development and student talent allocation. The first essay provides insights into the impact of algorithmic student advising programs, while the second essay highlights the role of higher education agencies in promoting international student mobility. The third essay evaluates the cognitive development trade-offs entailed by technical coursework. The first essay discusses the Graduation and Progression (GPS) program, which is an algorithmic student advising platform implemented by Georgia State University. The study analyzes the impact of this program on student course-taking by comparing GPS-advised students with those who did not receive advising. The study failed to credit the program to have increased graduation rates by improving academic fit but found that marginal students tended to leave college earlier. Also, the study provides evidence of assortative matching between students and course selection, albeit only for STEM Computational majors. The second essay examines the relationship between the 1996-2016 expansion of the German agency DAAD's outbound offices and international student enrollment in Germany. The findings suggest that an increase in the number of DAAD offices has a positive impact on international student enrollment in Germany, and that the first office foundation has the largest effect. The study concludes by discussing the policy implications of these findings for countries competing in the global race for talent. The third essay evaluates cognitive development trade-offs between numeracy and literacy skills. The study uses PISA data and analyzes the educational and financial gains from technical education versus the potential underdevelopment of verbal skills. The study finds that the technical track outperforms the Liberal Arts track due to greater educational production efficiency, which overcompensates for worse educational production inputs. The findings suggest that the STEM advantage is linked to the four additional instructional units in math and physics, and that there are no secondary effects due to differences in preexisting levels of student skills.

# CHAPTER 1
# GEORGIA STATE UNIVERSITY'S GRADUATION AND PROGRESSION SUCCESS ADVISING: ACADEMIC FIT EFFECTS FROM LEARNING ANALYTICS

## 1.1 Introduction

The growing academic interest in learning analytics (Goldstein and Katz 2005; Siemens 2013; Papamitsiou and Economides 2014; Macfadyen and Dawson 2010; Hlosta, Zdrahal, and Zendulka 2017) matches the expanding investment in Learning Management Systems which has driven a three-fold increase in the financing of educational technology startups during the COVID-19 pandemic (Singer 2021).

Learning Analytics (LA) collected in academic databases can track, predict, and influence student performance in classes, similar to how consumer analytics are used in the entertainment industry to tailor product offerings and improve customer satisfaction (Goldstein and Katz 2005; Siemens 2013; Papamitsiou and Economides 2014; Macfadyen and Dawson 2010; Hlosta, Zdrahal, and Zendulka 2017). These LA systems often involve a large amount of data, can process information quickly, include a variety of data types, and are reliable. Companies like Amazon, Netflix, and YouTube have successfully used recommendation algorithms to reach a diverse range of customer preferences. This approach, known as the "long tail" '(Anderson 2006), allows for the inclusion of niche interests that may not be addressed by mainstream content creation.

The use of algorithmic advising in higher education has significant implications for educational policy (Arcidiacono, Aucejo, and Spenner 2012). In the past, access to different college paths was determined by a variety of factors, including student preferences, abilities, financial resources, and randomness. However, AI-powered recommendation platforms are

designed to minimize randomness in student behavior and increase predictability of coursework choices based on legacy student data such as demographics, test scores, and previous academic performance. This potentially leads to a "predictability revolution" similar to what has occurred in the consumer media industry (Cinelli et al. 2021), where algorithmic tracking of user accounts increases ability to predict consumer behavior.

Using legacy student data to influence coursework choices may compromise our traditional definition of "educational opportunity" (Reardon 2018), which allows for a degree of randomness. The Duke survey (Arcidiacono, Aucejo, and Spenner 2012), a longitudinal study of undergraduate student major intentions at Duke University, highlights the potential unintended consequence of advising students such that their chances of obtaining a degree are maximized, where students may prioritize easier, less marketable degrees over more challenging but potentially more marketable ones. On the other hand, more reliable and standardized advising may improve student outcomes. For example, early advising can decrease the negative impacts of enrolling students who are likely to drop out of college, minimizing the financial burden of tuition payments and lost income. Additionally, AI-powered recommendations may improve the match between students and courses, leading to more specialized skillsets.

The article reports the results of the Graduation and Progression System (GPS) advising program, implemented by Georgia State University (GSU), which is the first and largest application of individualized LA to intrusive student advising (GSU 2019). Intrusive advising is a proactive approach to advising that involves regularly checking in with students and providing guidance on their academic and career goals, and differs from a more reactive approach, in which students only receive assistance when they seek it out. Bill Gates has even commented on the GPS program's reach, stating that "no other institution has accomplished what GSU has over the past decade" in terms of expanding the scope of LA. Using ten years of student data from all of GSU's campuses, departments, and programs, the program has identified nearly 1,000 "academic signals" indicating when students may

not be well-suited to their current academic curriculum. These signals, which can be specific to certain programs or classes or more comprehensive, are used to alert advisors to students who may be struggling and suggest course adjustments. Poorly fitting students are then met with their assigned advisors – a ratio of 1-in-700 has been reported by GSU (GSU 2019) – who use the wealth of students' records to customize advice to their specific needs. The GPS advising program thus combines the personalization of feedback with mass production of advising services, a combination previously thought to be unattainable.

Results are based on unique GSU administrative data from 2006 to 2014 and divided by four discipline categories: STEM Computational, STEM Life Sciences, Social Sciences, and Humanities (Le, Robbins, and Westrick 2014). Quasi-experimental techniques are used to control for bias from observable student characteristics, and validity checks are performed to minimize the influence of concurrent university policies such as microgrants going to students in need, history effects such as grade inflation, and changes in tuition regimes brought about by the reform of the HOPE scholarship system.

A theory of action based on the option value of college (Stange 2012) was used to generate research hypotheses relating the availability of information about academic fit to graduation rates, persistence, and course-taking. Specifically, academic fit was operationalized as the distance between a student's relative academic aptitude and their major. A Relative Academic Strength index was created using academic tilt literature (Coyle and Pillow 2008; Coyle 2018) and ranks students and majors along a STEM fit continuum based on the distance between their verbal and numerical scores on standardized tests.

The findings of the study are mixed and should be considered within the limitations of the research design. A difference-in-differences experiment comparing students who switched majors and those who stayed in their intended majors failed to demonstrate a relationship between improving student-major fit and increased graduation rates, although the dosage effects of the policy are consistent with this interpretation. There was also evidence that the length of time spent in college decreased among students who eventually dropped out, but

only among those with partial or no financial aid. Findings related to course-taking showed an increase in coursework in relevant major areas and a decrease in the absolute distance between students' relative academic aptitude and their majors. Specifically, STEM Computational students took, on average, over 15 more credits in their major area compared to pre-GPS program implementation cohorts.

## 1.2 Literature Review

### 1.2.1 Learning analytics and advising

Intrusive advising (Angrist, Lang, and Oreopoulos 2009; Bettinger and Baker 2014; Oreopoulos, Brown, and Lavecchia 2017; Oreopoulos and Petronijevic 2018; Dobronyi, Oreopoulos, and Petronijevic 2019; Page et al. 2019) consists of text messages, e-mails, phone calls, and other forms of outreach (e.g., chatbots, prerecorded videos and audios, online and offline mentoring, etc.) stimulating student proactive behavior. Advancements in the technology have allowed universities to scale intrusive advising interventions, and their coupling with high-frequency LA (Goldstein and Katz 2005; Siemens 2013; Papamitsiou and Economides 2014; Macfadyen and Dawson 2010; Hlosta, Zdrahal, and Zendulka 2017), including course-specific information from Learning Management Systems (Jovanovic et al. 2019), promises to further customize interventions to the needs and backgrounds of individual students.

The focus on retention in individual classes typifies the first generation of interventions headlined by the Open Academic Analytics Initiative (OAAI). This generation includes other interventions conducted at Purdue (Arnold and Pistilli 2012), San Diego State University (Dodge, Whitmer, and Frazee 2015), and the Open University of Hong Kong (Choi et al. 2018). The Bill & Melinda Gates Foundation has taken the lead on student analytics sponsoring the OAAI (Jayaprakash et al. 2014) and the later expansion of Georgia State University's GPS advising system, with the US Department of Education following suit with generous funding to the Monitoring Advising Analytics to Promote Success

(MAAPS) initiative involving 11 institutions inclusive of GSU (Rossman et al. 2021).

In contrast to the initiative evaluated in the current study, the OAAI used logistic regression to identify and alert students at risk of failing certain classes. The two OAAI treatment groups, one receiving text messages if classified as at-risk and the other receiving both texts and mentoring services, exhibited marginally higher grades but also higher attrition compared to the control group in the classes covered by the initiative. The Purdue experiment (2007-2009) implemented course signals through a traffic light system blinking red, yellow, or green to indicate a student's status in the class and decreased class attrition (Arnold and Pistilli 2012), similar to the San Diego (Dodge, Whitmer, and Frazee 2015) and Hong Kong (Choi et al. 2018) experiments.

A previous evaluation of the MAAPS initiative (Rossman et al. 2021) differs from the current study. The evaluation implements a randomized controlled trial design, in which the treatment group received degree-planning and proactive outreach activities based on self-regulated learning, in addition to the advising services offered to the placebo group. Therefore, the control group in the MAAPS study effectively corresponds to the treatment group in the current study. Additionally, the MAAPS study is a longitudinal analysis of the undergraduate careers of the 2016-17 freshman class at several universities, including approximately 2,000 students from GSU, and focuses mainly on cumulative GPA and credit hours without a specific emphasis on course selection and student-major fit.

The MAAPS study found that Black students had the largest gains, with a 0.22 point increase in GPA, 12 additional credits, 8 percentage point increase in graduation rates, and a 10 percentage point decrease in dropout rates when advised through the MAAPS program. In the discussion, Rossman et al. (2021) raise concerns about a shift towards easier classes and changes in incentives to choose certain classes, but do not further explore or develop a theoretical framework for algorithmic advising at the system level.

### 1.2.2 The information value of academic signals

A substantial effort to understand the determinants of college enrollment and persistence has been pledged across education research. Educational psychology has linked academic performance to self-efficacy (Schunk 1991) developing the concept of self-regulated learning (Zimmerman 2000) while the economist perspective has taken from the incentive-based analysis of obtaining a degree (Mincer 1958; Schultz 1961; Becker 1962; Heckman 1976). The perspective of educational psychology recognizes the iterative nature of course-taking and the impact of early classes on later learning, leading to the need to update the human capital model to account for the informational value of early classes (Stange 2012). This informational value is a key aspect of studies on intrusive advising (Angrist, Lang, and Oreopoulos 2009; Bettinger and Baker 2014; Oreopoulos, Brown, and Lavecchia 2017; Oreopoulos and Petronijevic 2018; Dobronyi, Oreopoulos, and Petronijevic 2019; Page et al. 2019) and is integral to the development of a model for algorithmic student advising.

The iterative enrollment model developed by Stange (2012) decomposes the value $V_{ijt}$ of a degree $j$ to student $i$ at time $t$ into two components: the net present value of the degree $NPV_{ijt}(\cdot)$ and the option or information value $I_{i,t+1}(\cdot)$ of reassessing one's chances to attain a degree at checkpoint $t+1$. The net present value of the degree captures the lifetime returns of the degree ($W_j$) minus its costs ($C_j$); conversely, the information value captures the value of the reduction in uncertainty ($\eta_{ijt}$) concerning degree-specific relative academic strength ($RAS_{ij}$) discounted at the student's self-confidence level ($\delta_{it}$). Because of the indexing of $RAS_{ij}$ to both $i$ and $j$, the index must be understood as relative aptitude rather than general student aptitude or $g$ (Coyle and Pillow 2008; Coyle 2018). Relative academic strength is partially known when entering college ($0 < \eta_{i,j,t=0} << 1$) through high school grades, test score results, and teacher and peer assessments and fully revealed when graduating from college at time $T$ ($\eta_{i,j,t=T} \approx 0$). Note the indexing of $\delta_{it}$ to time $t$ to account for the plasticity of self-confidence (i.e., $0 < \delta_{it} < 1$). The option value of college gets larger when residual uncertainty is greater (i.e., $\partial I_{t+1}/\partial \eta_{ijt} > 0$), and when student

self-confidence is lower (i.e., $\partial I_{t+1}/\partial \delta_{it} < 0$).

The total value of degree $j$ to student $i$ at time $t$ resolves to:

$$V_{ijt} = NPV_{ijt}(W_j, C_j) + I((1-\delta_{it})^t \eta_{ijt} RAS_{ij})$$

Student $i$ would persist in college when the value of their degree $V_{ijt}$ is greater than the non-college option $V_i = NPV(W_k)$, where $k$ is the most lucrative occupation which does not require a degree. The iterative college investment model improves upon the standard model by taking into account the option value of college, which can make the decision to enroll in college seem more rational for students who may have a low ex ante probability of graduation and high costs of attendance. Additionally, this model can help explain why some college students may choose to drop out at later stages without having updated their degrees and absent any material changes in the cost-benefit ratio of their educational investments, as they may be "cashing in" on the new information value they have gained at the end of an academic year.

### 1.3 Methodology

This section hosts discussions of the methods used in the study, including the development of the index of relative academic strength to measure student-major fit and the GPA deflation procedure to account for potential bias due to increases in average GPA over time.

The goal of the methodology is to control for observable differences between students who received advising services and those who did not. However, due to the non-experimental nature of the treatment and the comprehensive rollout of the program across undergraduate education at GSU, assignment to treatment is non-random and there is neither a clean control group available nor a presumably exogenous set of placebo students. In the study, different student subgroups were used to test different aspects of the model and each subgroup had to be matched independently to ensure that observational bias was minimized. Details of this matching procedure are provided in the following subsection.

The matching process used in this study identified two groups of students that were comparable on key characteristics reported in Table 1.1. This table presents the matching features and basic descriptives for the treated group of first-year students from A.Y.s 2012-2013 to 2014-2015 and the control group of first-year students from academic years 2006-2007 to 2008-2009. The $t$-tests and $\chi^2$-tests for numerical and categorical variables respectively showed significant differences in all features except for out-of-state and first-generation status, age, and SAT math test score.

### 1.3.1 Matching design

To address observational bias, outcomes of treatments and controls were weighted using inverse probability weights from optimal matching with replacement (Rosenbaum 1989; Hansen 2004). Propensity scores were calculated on observable student characteristics to balance covariates across students who received GPS advising and students who did not get advised independent of the outcomes. Next, conditional outcomes were calculated using the appropriate method for the outcome, linear regression (continuous outcomes), logistic regression (binary outcomes), or ordinal logistic regression (ordinal outcomes) requiring robust standard errors.

In general, one wants to finds a "control" student (i.e., $GPS_{i'} = 0$) defined by a vector of matching covariates $X_{i,GPS=0}$ about equal to the covariate space $X_{i',GPS=1}$ of a treated student (i.e., $GPS_i = 1$). Propensity score matching matches the covariate spaces indirectly through asymptotic convergence of the conditional features (Rosenbaum and Rubin 1983). Under standard identifying assumptions, inverse probability weights generated from logistic regression of the binary treatment variable on the conditioning factors identify the Average Treatment Effect on the Treated (ATT) in the second-stage difference-of-means or difference-of-proportions. The ATT estimator is the inverse probability weighted difference in graduation rates for the $1, \ldots, N$ treatments and the $1, \ldots, N'$ controls:

Table 1.1: Sample descriptives of matching features used for inverse probability weighting

| **Variable** | **GPS**[1] | **Non-GPS**[2] |
|---|---|---|
| Female*** | 59.9% | 56.6% |
| Race*** | | |
| *American Indian or Alaska Native* | 0.4% | 0.2% |
| *Asian* | 14.7% | 17.5% |
| *Black* | 33.2% | 38.0% |
| *Native Hawaiian or Pacific Islander* | 0.5% | 0.1% |
| *Not Reported* | 5.5% | 4.0% |
| *Two or More Races* | 3.6% | 6.9% |
| *White* | 42.1% | 33.3% |
| Hope Recipient*** | 81.3% | 71.0% |
| Out-of-State | 3.4% | 3.6% |
| Pell Eligible*** | 37.5% | 43.5% |
| First Generation | 20.6% | 21.4% |
| Unmet Need*** | 54.4% | 72.2% |
| Age (yrs) | 18.4 (0.6) | 18.4 (0.6) |
| High School GPA*** | 3.33 (0.32) | 3.36 (0.34) |
| SAT Math | 540.3 (71.9) | 538.4 (77.7) |
| SAT Verbal** | 540.1 (71.4) | 537.4 (73.1) |
| SAT Writing*** | 526.5 (70.9) | 520.6 (73.8) |
| Graduation*** | | |
| *Yes, >4 yrs* | 45.0% | 34.2% |
| *Yes, <=4 yrs* | 17.2% | 21.9% |
| *No, <=4 yrs* | 37.7% | 43.9% |

*Note*. The table reports sample descriptives and the statistical significance of their differences between treatments and controls ($^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$). Standard deviations and $t$-tests are provided for numerical features, and sample proportions and $\chi^2$ tests for categorical features.

[1] The treatment group comprises students who entered their first year anywhere between A.Y. 2012-2013 and A.Y. 2014-15.

[2] The control group comprises students who entered their first year anywhere between A.Y. 2006-2007 and A.Y. 2008-09.

$$\Delta \hat{Graduation} = N \sum_{i=1}^{N} \frac{Graduation_i}{\hat{p}_{GPS_i=1}(X_i)} - N' \sum_{i'=1}^{N'} \frac{Graduation_{i'}}{1 - \hat{p}_{GPS_{i'}=0}(X_{i'})}$$

where the numerator is the observed graduation outcome and the denominator is the inverse probability weight calculated from the individual propensity of treatment $\hat{p}(\cdot)$ conditional on the matching features $X = x_1, \ldots, x_k$. The study uses a difference-in-differences

model to compare the graduation rates of two groups of students: those who switched majors (switchers) and those who did not (non-switchers) before and after the implementation of the program. The difference-in-differences estimator calculates the ratio of graduation rates between these two groups, with the upper and lower bounds defined by the most extreme outcomes. The research design and measurement of other outcomes are discussed in more detail in subsection 1.4.1.

The bias reduction measures the extent to which the differences between the treated group and the control group, in terms of their covariates, are reduced. It is calculated as the percentage reduction in total bias, which is the sum of the absolute standardized mean differences between the two groups:

$$\Delta Bias = 1 - \sum_{k=1}^{K} \frac{\mid \overline{X}_{k,GPS=1}^{after} - \overline{X}_{k,GPS=0}^{after} \mid}{\sqrt{(S_{k,GPS=1}^{2\,after} - S_{k,GPS=0}^{2\,after})/2}} \Bigg/ \frac{\mid \overline{X}_{k,GPS=1}^{before} - \overline{X}_{k,GPS=0}^{before} \mid}{\sqrt{(S_{k,GPS=1}^{2\,before} - S_{k,GPS=0}^{2\,before})/2}}$$

where $\overline{X_k}$ is the average of feature $k$ and $S_k$ is its variance calculated before or after matching. The Conditional Independence Assumption (CIA) requires that: 1) the first-stage binary model fully determines the data generating process; 2) good conditioning factor balance is achieved after matching; and 3) there is no unobservable heterogeneity that differentially affects the outcomes of treatments and controls. Assumption 1) is satisfied via complete specification of an education production function that determines student outcomes in college (Sass, Semykina, and Harris 2014). Assumption 2) is satisfied by detecting minimal violations of optimal post-matching balance at the standard significance threshold of 0.1 standardized mean differences. Assumption 3) might only be violated with particular outcomes subject to unobservable longitudinal shocks (see subsection 1.5.2 for details).

### 1.3.2 Relative academic strength

An index of discipline fit was created using relative academic strength (Coyle and Pillow 2008; Coyle 2018). This index is a continuous value between $-1$ and $+1$ that represents the maximum fit for qualitative and quantitative disciplines, respectively. The index is calculated using the relative percentile ranks of a student's math and verbal scores from the SAT or ACT test. Since students at GSU can choose to submit either the SAT or ACT, the distributions for both tests were averaged and scaled separately. The formula for calculating the Relative Academic Strength (RAS) is as follows:

$$RAS_i = \frac{\nu_i - \lambda_i}{\mid \nu_i - \lambda_i \mid} \cdot (\mid \nu_i - \lambda_i \mid)^{\max(\nu,\lambda)}$$

where $\nu_i$ is the math rank and $\lambda_i$ is the verbal rank measured in percentiles.

A preliminary check of face validity was conducted via simulation (Table 1.2). A student with a math score at the 80th percentile and a verbal score at the 20th percentile would have a RAS value of $-(0.8 - 0.2)^{0.8} = 0.66$. A student with the same percentile rank in verbal and math would be positioned at the center of the RAS range, though this is unlikely to happen. The index takes into account diminishing marginal returns of relative ability, and moves students towards the middle of the range as their composite percentile rank increases. This means that a student with the same math or verbal tilt will be pushed further to the right or left of the range, respectively, if their scores are closer to the median score. For example, a student with a 20-percentile math tilt at the 90th percentile in math and 70th percentile in verbal will have a RAS of 0.23, but a student with the same math tilt at the 70th percentile in math and 50th percentile in reading will have a RAS of 0.32.

RAS values were computed for both students and majors to calculate the distance between students' academic aptitude and their terminal majors. To minimize the potential confounding effect of the program on the index, the entry SAT and ACT test scores of students were collected from the GSU web repository IPORT using Python's Selenium. While the data

is aggregated by major, IPORT provides more detailed information than the information available from administrative data. For instance, IPORT includes the average SAT scores of applicants, admitted, and enrolled students. The data on admitted students was used as it is believed to provide a more accurate representation of the skill requirements for a particular program compared to data on applicants or enrollments, which may be influenced by market demand and other non-academic factors. The entry-level test scores of the freshman cohorts from 2009-2010, 2010-2011, and 2011-2012, which predate the program, were not included in the treatment effect analysis as they were considered to be uncontaminated data. Additionally, using entry test scores helps to ensure that factors endogenous to course-taking at GSU do not influence the construction of the index. For majors with a sufficient number of students, the most recent test data was used, while the average test scores from 2009-2011 were weighted by cohort size and used for smaller majors.

The RAS values of majors, plotted in Figure 1.1, reveal good face validity. The three majors with the highest verbal tilt are Women's and Gender Studies, French, and Journalism, while the most quantitative-leaning majors are Finance, Physics, and Math. Some large programs at GSU, like Exercise Science and Nutrition, show a surprising STEM tilt while the social sciences (e.g., Criminal Justice, Economics, Social Work) tend to be balanced with a slight verbal tilt.

### 1.3.3 Accounting for GPA inflation

There is evidence that grades in higher education have increased since the 1970s (Pattison, Grodsky, and Muller 2013), a phenomenon known as grade inflation, and that students who are close to receiving scholarships or waivers may improve their college performance in response (Henry and Rubenstein 2002). As a result, there is a possibility that comparisons of GPA outcomes over time may be biased due to history and behavioral effects.

To control for the potential bias of grade inflation, the study used a method similar to that of Henry and Rubenstein (2002) by assuming that the relationship between SAT scores

Table 1.2: Relative Academic Strength (RAS) value for mock students with varying aptitude

| **Academic aptitude** | **Math pct.**[1] | **Verbal pct.**[2] | **RAS** |
|---|---|---|---|
| Very high math, very low verbal | 0.9 | 0.1 | 0.82 |
| High math, low verbal | 0.8 | 0.2 | 0.66 |
| Middle-to-high math, middle-to-low verbal | 0.7 | 0.3 | 0.53 |
| Middle-to-high math, middle-to-low verbal | 0.6 | 0.4 | 0.38 |
| Middle math, middle verbal | 0.5 | 0.5 | 0.00 |
| Middle-to-low math, middle-to-high verbal | 0.4 | 0.6 | -0.38 |
| Middle-to-low math, middle-to-high verbal | 0.3 | 0.7 | -0.53 |
| Low math, high verbal | 0.2 | 0.8 | -0.66 |
| Very low math, very high verbal | 0.1 | 0.9 | -0.82 |

*Notes*. The math and verbal percentile scores reported in columns (1) and (2) indicate varying levels of academic strength. The corresponding RAS is reported in column (3).

[1] A student's math percentile score is their percentile rank in the score distribution of SAT/ACT tests in their freshman year.

[2] A student's verbal percentile score is their percentile rank in the score distribution of SAT/ACT tests in their freshman year.

and GPA should remain constant when GPA increases. A GPA discount factor $I_{jt}$ was calculated for each major $j$ and year $t$ by using 2010 as a baseline and considering the relative changes in the average GPA and SAT scores of students graduating from the major:

$$I_{jt} = \frac{SAT_{jt}}{GPA_{jt}} \Big/ \frac{SAT_{j2010}}{GPA_{j2010}}$$

If this ratio remains constant over time, it indicates that the major grades consistently across years. On the other hand, a changing ratio suggests that similar students are graded differently over time. To adjust for this, a major-year-specific GPA discount factor $I_{jt}$ was calculated for each major $j$ and year $t$ using 2010 as the baseline. This discount factor was used to convert nominal GPA into real GPA denominated in 2010 GPA points, so that $I_{jt} = I_{j2010}$:

$$\text{Real GPA} = I_{jt} \times \text{Nominal GPA}$$

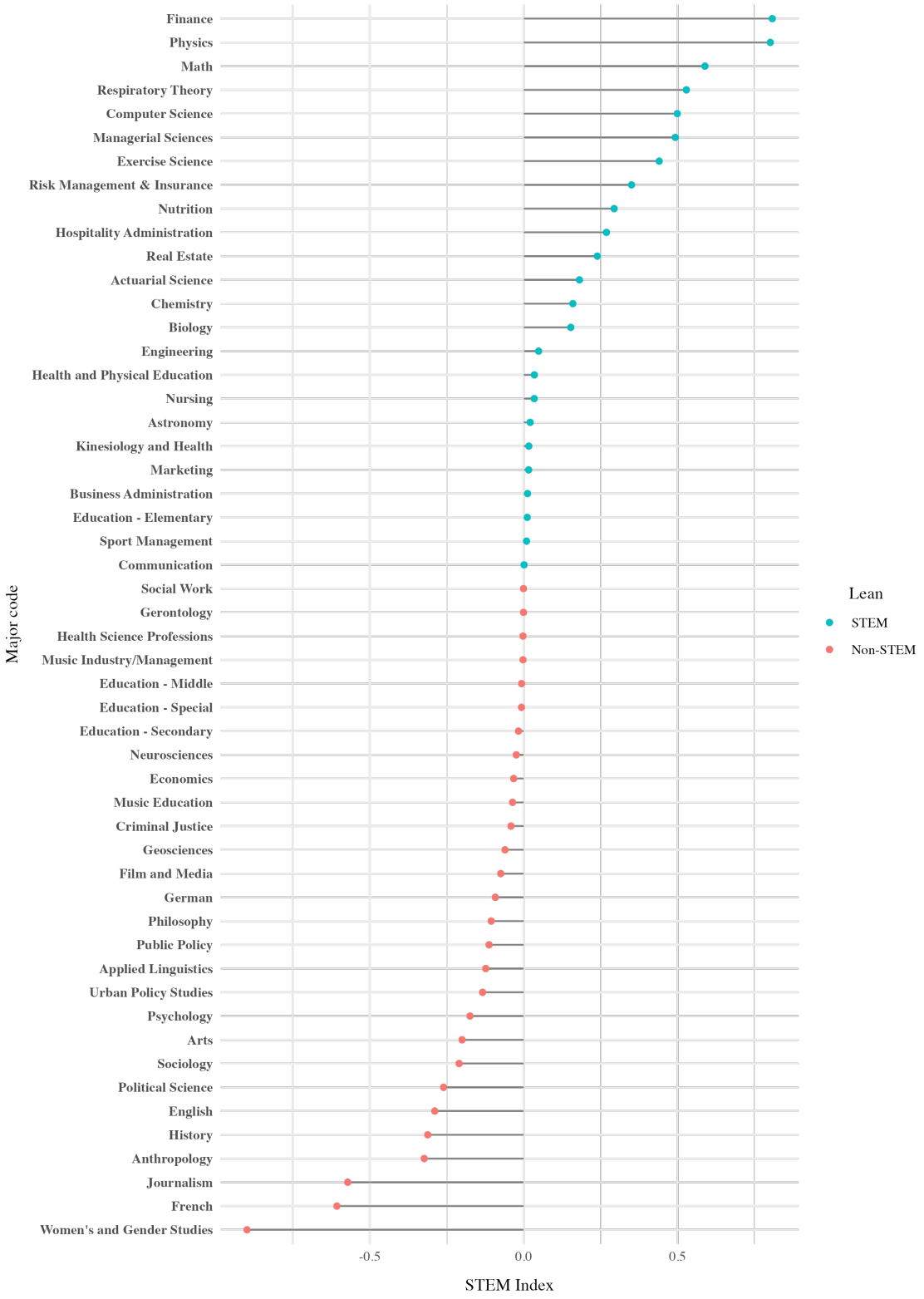


Figure 1.1: Relative Academic Strength (RAS) by major in descending order of STEM fit

After adjusting for grade inflation, the average GPA for all students across all majors was found to be lower than the nominal GPA, with the extent of the decrease differing by major.

Other methods for accounting for changes in GPA over time were considered, but ultimately not used in this study. A common approach is regression-based adjustment of nominal GPA (Brookhart et al. 2016): the assumption here is that the same student features should afford similar grades year over year, less grading has become more or less lenient. However, this approach has the potential flaw of assuming that the education production function remains constant, even when the institutional and environmental context changes. In contrast, the correction method used in this study only assumes the stability of the relationship between GPA and test scores. While there is ongoing debate about the stability of measures of cognition, there is a general consensus that these measures are more stable over time than the relationship between factors of the educational production process and student achievement.

## 1.4 Empirical Section

In this dataset, there are records for undergraduate students at GSU from 2006 to 2014. The time-invariant features are characteristics of the student that remain constant over time, such as ethnicity and high school GPA, and are used to generate propensity weights. The time-variant features, on the other hand, vary on an annual basis and include information such as major declaration and tuition status, primarily used as outcome measures and for inclusion and exclusion criteria in validity checks. The dataset also includes information on credit hours taken by each student in a given academic year (e.g., MATH, CS, ENG, etc.), with credits for each subject area being aggregated to calculate cumulative GPA at the end of each year. Descriptive information for treated and control students is provided in Table 1.1.

### 1.4.1 Theory of action

Central to the option value of college model (Stange 2012) is that students follow an iterative decision-making process, and reassess the value of the degree at each checkpoint. On the other hand, in the standard human capital model (Mincer 1958; Schultz 1961; Becker 1962; Heckman 1976), students make a one-time enrollment decision based on the expected lifetime returns of a degree minus its opportunity costs. Information made available by GPS advisors shapes student decision-making by benchmarking student performance to historical data.

This high-frequency feedback can encourage students to move towards majors that are more closely aligned with their inclinations and provide emotional support for learning. Studies have shown that access to better information about the value of a degree can increase retention rates, particularly in the case of intrusive advising interventions (Angrist, Lang, and Oreopoulos 2009; Bettinger and Baker 2014; Oreopoulos, Brown, and Lavecchia 2017; Oreopoulos and Petronijevic 2018; Dobronyi, Oreopoulos, and Petronijevic 2019; Page et al. 2019). Additionally, students may make suboptimal college choice and enrollment decisions due to informational constraints, self-concepts formed by early performance in college classes (Porter and Swing 2006; Lizzio and Wilson 2013) such as fixed beliefs about their abilities (Dweck 2013), and perceptions carrying over from prior formation (Schunk 1991; Lent, Brown, and Hackett 1994; Zimmerman 2000).

Overall, algorithmic student advising is bound to affect student outcomes through three distinct pathways:

1. *Self-confidence pathway*. Changes in student self-confidence, now effectively becoming a dynamic, time-dependent, construct (i.e., $\delta_{it}$);

2. *Ability revelation pathway*. Reductions in uncertainty about degree-specific ability (i.e., $\eta_{ijt}$);

3. *Academic fit pathway*. Changing net present value of a degree $NPV_{ijt}(\cdot)$ after a

switch to a better-fit major $j$.

The effectiveness of academic signals may vary depending on the individual student's characteristics and circumstances. Stange (2012) notes that "moderate-ability students, who have the most uncertainty about the desirability of schooling, derive even more value" (p.81) from continuous and flexible information provision. Another potential reason for underprivileged students overresponding to GPS advising is the greater plasticity of their self-confidence and perception of ability (Tinto 1987). However, the value of the new information may be more accessible to high-achieving students who have more resources to make better use of it.

There are two main concerns that have been raised in the literature regarding algorithmic advising: the potential for students to be restricted by their pre-college records, and the possibility of shifting marginal students towards less marketable majors. This second concern is supported by the Campus Life and Learning Project (Arcidiacono, Aucejo, and Spenner 2012), which found that a reduction in the first-year Black-White GPA gap from 0.5 to 0.3 points at the end of the fourth year at Duke University resulted in about half of Black females and a third of Black males switching from their intended STEM majors to humanities majors, which had an average of 10% higher grades. The advising intervention, while effective at improving graduation rates, had the unintended consequence of diluting the value of the extra degrees attained.

The first part of the study employs what is essentially a difference-in-difference design to identify transmission mechanisms for graduation rates. The first difference is the difference in the graduation rate of students who did not switch majors before and after treatment and the second difference is the difference in the graduation rate of students who switched majors before and after treatment. The academic fit pathway, which focuses on the effects of algorithmic advising on coursework selection, is only applicable to students who switch their majors. This allows the research design to differentiate the impact of the third pathway from the first two pathways, which are not dependent on major switches. Therefore,

the difference-in-differences is a significance test for academic fit effects from the GPS program.

The research further aims to examine whether algorithmic advising affected persistence of students who eventually dropped out. This is based on the idea that students who are on the fence about continuing their studies may be more likely to drop out if they receive information about their low chances of completing their degree. By providing these students with more accurate and timely information, algorithmic advising may help them exit college early. The information value thus potentially explains why marginal students enroll in college despite their low college readiness. Algorithmic advising might correct suboptimal decision-making avoiding losses of time and income resources.

Thirdly, tailored information about learning paths might lead to greater course-taking in areas of better academic fit. Using the RAS index described in section 1.3, potential reductions in student-major distance resulting from algorithmic advising were evaluated. The impacts of algorithmic advising on GPA and total credit hours were also analyzed to determine any potential trade-offs between the quantity and quality of instruction.

### 1.4.2 Effects of algorithmic student advising

*Graduation rates*

The first research question is whether retention changed as a result of switching to majors that are a better fit for students.

The experimental design uses a difference-in-differences approach to identify transmission mechanisms of the policy. The difference-in-differences estimate results from two differences: the first difference is the graduation rate change for students who did not switch majors before and after the GPS program implementation (i.e., non-switchers), and the second difference is the graduation rate change for students who switched majors (i.e., switchers). If the positive effects of algorithmic advising on student outcomes are due to improvements in academic fit, it is necessary for these effects to be more pronounced in

students who changed majors. It is possible that other confounding factors may prevent a definitive causal claim from being made, but this condition appears to be a minimum requirement for the viability of the transmission mechanism.

To further tease out transmission mechanisms, the study sets up a subexperiment and a visual test. The subexperiment divides the group of students who switched majors into two sets: the first set consists of students who switched to majors that are a better fit for their relative academic aptitude, while the second set consists of students who switched to majors that are a worse fit. If the effects of the policy are primarily due to improved student-major fit, the first group of students should have benefited more from the policy than the second group. The visual test (see Figure 1.2) examines the effects of increasing exposure to the recommendation system (i.e., "dosage effects") to identify discontinuities in the effect of the policy: if effects increased not dissimilarly for switchers and non-switchers, this would suggest that the policy's effects are not primarily due to the academic fit pathway.

The results in Table 1.3 show the graduation odds ratio for non-switchers and two sets of switchers, with confidence intervals for each set of coefficients. The table presents coefficient estimates as graduation odds ratios, i.e., before-after ratios of graduation odds. The difference-in-differences estimate is the net change in graduation odds due to the major switch, taking into account the change in graduation odds for non-switchers. Examining the confidence intervals can reveal whether the association between algorithmic advising services and student graduation rates is statistically significant, or if it is challenging to establish noteworthy and distinctive impacts between major switchers and non-switchers.

The results of the experiment indicate that students who switched majors under GPS advising had an increased chance of graduating, although this difference was not statistically significant (OR = 1.257, 95% CI 0.867-1.823) when compared to students who did not switch majors. The subexperiment on students who switched to majors that were closer matches to their academic aptitude also did not yield statistically significant results (OR = 1.254, 95% CI 0.846-1.859). Hence, one cannot assert that the GPS advising program led

to enhancements in student graduation rates, or at the very least, it is plausible to eliminate the possibility that it achieved this outcome through the major switching pathway, which aligns with the most effective transmission mechanism according to theory.

The visual "dosage effects" provides more moderate evidence in favor of switching effects. Switchers who were exposed to the full four years of the program had a higher positive impact, with the non-switcher confidence falling within a similar range. Furthermore, switchers who were exposed to the GPS program for three or two years had a decrease in the impact on their graduation rates, and there was no significant difference in the effects between those with one and two years of exposure. Conversely, non-switchers did not consistently show an increase in graduation rates with increased exposure to the GPS program, and the trendline for their effects was always within the margin of error for the trendline for switchers. Therefore, while the dosage effects suggest moderate support for the academic fit pathway, the results do not provide strong evidence against the null hypothesis that the policy did not channel any effects through improvements in academic fit.

*Persistence of dropout students*

The second research question is whether the persistence of students who eventually dropped out changed due to the GPS program. To answer this question, ordinal logistic regression was used to analyze the persistence of students who had dropped out. The outcome variable is ordinal because students who left at any point during their college careers stayed for one, two, three, or four years before leaving. The goal was to determine if the GPS advising system influenced the timing of students leaving college.

One potential estimation issue is the change to Georgia merit-based financial aid in the A.Y. 2011-2012. Specifically, the requirement to maintain a full tuition waiver was raised from a 3.0 GPA to a 3.3 GPA, breaking up the previous scholarship program that awarded full tuition waivers to students maintaining a 3.0 GPA into two tiers: one that covered the full tuition costs for students with a 3.3 GPA and another that provided a partial tuition waiver

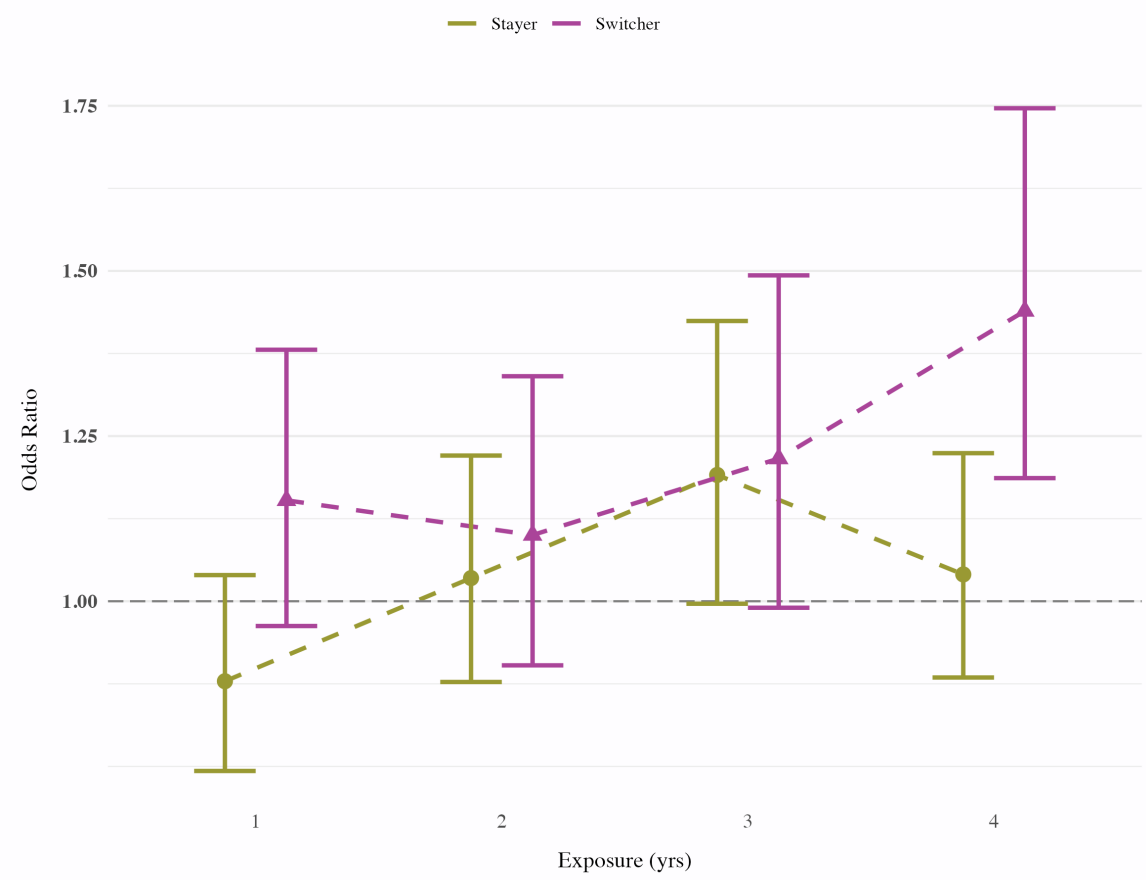


Figure 1.2: Odds ratios of college graduation for GPS students who updated their majors and who stayed in their intended majors

*Notes*. This figure plots the graduation odds of GPS-advised students who did not change their majors (Stayer) and GPS-advised students who switched their majors at any point during their academic career (Switcher) relative to the graduation odds for control students. The x-axis indicates years of exposure to the policy or "dosage effects" from a minimum of one year (2009-10 freshman cohort) to a maximum of four years (2012-13 freshman cohort).

Table 1.3: Odds ratios of college graduation for GPS advised students

| | **Odds Ratio (95% C.I.)** | | | **Matching** | |
|---|---|---|---|---|---|
| | **Coef.** | **Lwr.** | **Upr.** | **n** | **$\Delta$Bias (%)** |
| Non-Switchers[1] | 1.115 | 0.940 | 1.323 | 6,282 | 86.6 |
| *All majors*[2] | | | | | |
| Switchers | 1.402*** | 1.147 | 1.713 | 5,947 | 82.8 |
| Difference-in-Differences | 1.257 | 0.867 | 1.823 | | |
| *Better majors*[3] | | | | | |
| Switchers | 1.399*** | 1.120 | 1.747 | 2,765 | 86.5 |
| Difference-in-Differences | 1.254 | 0.846 | 1.859 | 2,765 | |

*Notes*. Odds ratios are calculated from logistic regression of the binary graduation outcome. Coefficients capture the ceteris paribus likelihood of graduation of GPS-advised students relative to those not. The Difference-in-Difference term is the ratio of the switcher to non-switcher coefficients, with 95% confidence region ranging from the mildest outcome (ratio of the lower bound switcher and upper bound non-switcher estimates) to the most extreme outcome (ratio of the upper bound switcher and lower bound non-switcher estimates). Estimates are inverse probability weighted (IPW) by the most important outcome predictors and use robust standard errors for calculation of the 95% confidence intervals ($^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$).

[1] The non-switcher set is students who graduated from the same majors they declared before college entry.

[2] The first set of switchers graduated in a major different from the majors that they declared before college.

[3] The second set of switchers is a more exclusive set that only includes students switching to majors that are closer fits to their academic aptitude.

for those with a GPA between 3.0 and 3.3.

The effect of the change in financial aid requirements on college persistence was evaluated by Jones et al. (2022) using a regression discontinuity design. The study found that the change, which affected students with a cumulative GPA below 3.3 at the beginning of Fall 2011-2012, did not lead to significant drops in college persistence, likely because the main targets of the change were students from upper socio-economic backgrounds who were less sensitive to changes in price.

To ensure that the results were not biased by the changes to financial aid, a check was performed by excluding the cohort of students who were fully exempt from tuition in the A.Y. 2010-2011 and experienced financial losses from not meeting the required 3.3 GPA. Although the research design does not use freshman cohorts from A.Y.s 2009, 2010, and 2011, multiple treatments bias might stem from the 2008 freshman cohort. If any of the A.Y. 2008-2009 freshman students, who were in their fourth year in Fall 2011-2012, had a GPA below 3.3, their full tuition waiver would have been terminated.

While multiple treatments bias might be a limited threat, financial aid changes modified the playing field. The change was anticipated and non-exogenous for students in high school at the time of the regime swap, but these students faced a tighter funding situation when starting college. Therefore, a second validity check looks at the three categories of funding recipients: students with no tuition waivers, students with partial tuition waivers, and students with full tuition waivers when starting college.

The results of the analysis (see Table 1.4) suggest that the GPS advising system may lead to earlier dropout among students who are struggling academically (OR = 0.662, 95% CI 0.559-0.784). This is seen as a positive outcome, as it may allow these students to leave college and enter the workforce earlier, potentially avoiding the loss of time and resources that may come with continuing their education. The results also show that there are minor differences in estimates when excluding students who experienced changes in financial aid regimes, suggesting that these changes may not have had a significant impact on the results.

It appears as though the advising system is effective at accelerating the decisions to leave college among students who were marginal in the first place.

Students who received full tuition waivers did not drop out of college earlier when they received GPS advising, according to the results of further subset regressions with the three categories of financial aid (i.e., none, partial, and full). While students who received partial tuition waivers or no waivers behaved similarly to the main group, students with full funding did not leave college sooner when advised by GPS. It is possible that the full tuition waivers acted as a "sticky" factor, causing students to remain enrolled even if they received advising signals to leave. However, this result is largely consistent with expectations, as students with full financial support may be less likely to respond to advising signals.

Table 1.4: Odds ratios of survival in college for students who dropped out

| | **Odds Ratio (95% C.I.)** | | | **Matching** | |
|---|---|---|---|---|---|
| | **coef.** | **lwr.** | **upr.** | **treat (%)** | **ΔBias (%)** |
| Overall set | 0.600*** | 0.506 | 0.713 | 5,016 | 79.6 |
| No tuition losers [1] | 0.593*** | 0.497 | 0.706 | 4,988 | 76.2 |
| Financial Aid = None | 0.684*** | 0.470 | 0.997 | 555 | 77.1 |
| Financial Aid = Partial | 0.728*** | 0.585 | 0.905 | 2,406 | 82.3 |
| Financial Aid = Full | 0.914 | 0.691 | 1.210 | 1,649 | 64.0 |

*Notes*. Odds ratios are calculated from ordinal logistic regression of the year of dropout (1,2,3,4). Coefficients capture the ceteris paribus likelihood of surviving one more year in college of GPS-advised students relative to those not. Estimates are inverse probability weighted (IPW) by the most important outcome predictors and use robust standard errors for calculation of the 95% confidence intervals ($^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$).

[1] These are students that suddenly lost part of the their tuition coverage in Fall 2011-2012 from the changing requirements of Georgia's state aid.

*Graduation outcomes*

The last set of findings accounts for course-taking outcomes of students: graduation GPA, student-major fit, and credit hours in their reference disciplinary area. High-ability students in STEM computational disciplines were found to benefit the most from GPS advising in

terms of increased GPA points (0.222, 95% CI $\pm$ 0.099) and credit hours taken in their major area (16.05, 95% CI $\pm$ 4.91) as well as decreased distance between their intended major and their actual graduation major each measured in terms of RAS (-0.075, 95% CI $\pm$ 0.054). These results were seen for three out of the four major denominations, except for Humanities. Importantly, the four labels are defined by the intended major, not by graduation major, to avoid reversing the order of events.

The distance between students and their terminal majors declined the most among students who intended to major in STEM Computational disciplines, followed by STEM Life Sciences and the Social Sciences. Conversely, prospective Humanities students did not significantly reduce the distance with their graduation majors. The fit between a student's academic aptitude and their major is indicated by the absolute distance between the academic tilts of the student and their majors (Coyle and Pillow 2008; Coyle 2018). If the distance between a student's aptitude and their major decreases, it means they have moved closer to a discipline that is a good fit for them. Ability tilt expresses a continuum of outcomes from high verbal proficiency to high math proficiency: if a student or major has high math score compared to their verbal score, they are more heavily STEM leaning.

Students who received algorithmic advising showed significant changes in their GPA, with the exception of those in the Humanities. Students in STEM Computational and Social Sciences disciplines had an increase in their graduation GPA, while students in STEM Life Sciences had a decrease in their GPA. Further analysis of the course structure in different fields may be needed to understand the specific reasons for the changes in GPA. However, it is noteworthy that the three disciplinary labels that reduced the distance between their students' entry academic aptitude and terminal majors also changed their inflation-adjusted GPA.

Table 1.5: Grade inflation-adjusted GPA at graduation

| | **WLS estimates (95% C.I.)** | | | **Matching** | |
|---|---|---|---|---|---|
| | **Coef.** | **Lwr.** | **Upr.** | **n** | **ΔBias (%)** |
| STEM Computational | 0.222*** | 0.123 | 0.321 | 679 | |
| STEM Life Sciences | -0.191*** | -0.328 | -0.054 | 726 | |
| Social Sciences | 0.236*** | 0.152 | 0.320 | 3,309 | |
| Humanities | -0.215 | -0.631 | 0.201 | 1,610 | 87.9 |

*Notes*. Coefficients are calculated from ordinary least squares regression of four-year graduation GPA. Coefficients capture the ceteris paribus change in GPA of GPS-advised students relative to those not. Estimates are inverse probability weighted (IPW) by the most important outcome predictors and use robust standard errors for calculation of the 95% confidence intervals ($^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$).

Table 1.6: Credit hours taken in the student's respective major area

| | **WLS estimates (95% C.I.)** | | | **Matching** | |
|---|---|---|---|---|---|
| | **Coef.** | **Lwr.** | **Upr.** | **n** | **ΔBias (%)** |
| STEM Computational | 16.05*** | 11.14 | 20.96 | 679 | |
| STEM Life Sciences | -1.22 | -4.37 | 1.91 | 726 | |
| Social Sciences | 3.62*** | 1.19 | 6.05 | 3,309 | |
| Humanities | 2.91 | -1.32 | 7.13 | 1,610 | 87.9 |

*Notes*. Coefficients are calculated from ordinary least squares regressions of credit hours in the relevant subject area. Coefficients capture the ceteris paribus change in credit hours taken by GPS-advised students relative to those not. Estimates are inverse probability weighted (IPW) by the most important outcome predictors and use robust standard errors for calculation of the 95% confidence intervals ($^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$).

Table 1.7: Distance between students and their graduation major

| | WLS estimates (95% C.I.) | | | Matching | |
|---|---|---|---|---|---|
| | **Coef.** | **Lwr.** | **Upr.** | **n** | **ΔBias (%)** |
| STEM Computational | -0.075*** | -0.129 | -0.021 | 679 | |
| STEM Life Sciences | -0.052** | -0.102 | -0.003 | 726 | |
| Social Sciences | -0.045** | -0.084 | -0.006 | 3,309 | |
| Humanities | 0.006 | -0.046 | 0.058 | 1,610 | 87.9 |

*Notes*. Coefficients are calculated from ordinary least squares regressions of the distance between the relative academic strength of students and their majors. Coefficients capture the ceteris paribus change in the fit between a student-major pair of GPS-advised students relative to those not. Negative (positive) coefficients indicate a narrower (looser) fit between students and their terminal majors. Estimates are inverse probability weighted (IPW) by the most important outcome predictors and use robust standard errors for calculation of the 95% confidence intervals ($^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$).

## 1.5 Discussion

### 1.5.1 Conclusions

This study examined the effects of largest algorithmic student advising implementation, Georgia State University's Graduation and Progression System (GPS). The program aims to improve student graduation rates by better aligning students with majors that fit their abilities, rather than just improving their chances of success in individual classes as previous programs have done. The research found that the program had an impact on course-taking, particularly among high-ability students in STEM fields, but did not demonstrate a link between academic fit and improved graduation rates. The program may also have led to faster dropouts among students on the margin, albeit not among those receiving full tuition waivers.

The first set of findings used a difference-in-differences design to control for changes in graduation rates over time and identify the potential mechanisms through which the GPS program affects student outcomes. The results showed that both students who switched majors and those who did not had increased chances of graduating under the GPS program,

and the rate of change did not differ significantly between the two groups. This suggests that the program did not have a disproportionately greater impact on students who switched majors. A subexperiment conducted on students who switched to majors that were a better fit for them also yielded insignificant results, further indicating that the student-major fit pathway was not an active factor in improving graduation rates. The dosage effects test also showed similar trends for students who switched majors and those who did not, depending on the length of their exposure to the program. Overall, the results do not support the hypothesis that the student-major fit pathway contributes to improved graduation rates.

The second set of results showed that the GPS program reduced the time it took for students to drop out of college. This effect was not consistent across all students, with those who received full tuition waivers showing no decrease in persistence under the program. These findings suggest that the GPS program may be effective at helping students on the margin make quicker decisions about their college careers, and that students who are fully tuition exempt may not respond as much to the program because they are less sensitive to price. It is important for educational policy to avoid unnecessarily causing students to leave college and to focus on helping those who would be better off seeking employment in the workforce. Therefore, the algorithm should be used with caution to ensure that it is not causing unnecessary harm to students.

The third set of results showed that the GPS program had a significant impact on both credit hours and GPA for students in STEM Computational, STEM Life Sciences, and Social Science disciplines, with no significant effects on students in the Humanities. The greatest beneficiaries seemed to be STEM Computational students, who took more extensive coursework in their field of study and achieved better GPAs under the GPS system. However, it is important to consider that these positive effects may disproportionately benefit students who are already doing well and are best positioned to reap the rewards of their education. Therefore, the scaling of advising interventions should be carefully considered in light of the potential risks of further exacerbating divergence in student outcomes.

### 1.5.2 Limitations

The main limitation of the study is the campus-wide rollout of the program in A.Y. 2012-13 by GSU, which means that synchronous control units were not available. This could lead to bias due to changes in the unobservable composition of the student body. Unobservable bias may affect outcomes that are influenced by history effects during the study period, such as GPA. However, the procedure for adjusting GPA for grade inflation based on the SAT-GPA ratio can mitigate this bias (see subsection 1.3.3). It is also possible that there are other unobservable sources of bias that cannot be properly accounted for and may impact graduation and persistence outcomes, such as changes in the economy that alter the cost-benefit ratio of attending and remaining in college. As a precautionary measure, the time-frame of the study was limited to three years before and after the program rollout.

The second set of validity threats involve concurrent programs that could affect the outcomes being examined. One of these initiatives, the Summer Success Academy, is designed to increase retention for less academically prepared students by providing various forms of support and mentoring before college begins. However, this initiative was deemed to have too limited a scope to potentially influence the results and no action was taken to control for its effects. Another initiative, the Panther Retention Grants, could have affected length of stay, course-taking, and graduation. The program disbursed microgrants of an average size of $1,000 (GSU 2019) to unmet need students who are in good academic standing and have met tuition payment deadlines in the past. To minimize the potential for multiple treatment bias, students who received these grants were excluded from the analysis.

A third and more substantive threat comes from the changing requirements of financial aid in the state of Georgia (Jones et al. 2022). The students who were exogenously stripped of some of their financing in A.Y. 2011-2012 were identified and excluded from the analysis. However, the changes persisted and affected subsequent cohorts of students who began college after the reform, even though it was no longer an exogenous shock. To address that, subset analysis of students in different financial aid bands was conducted, revealing

that GPS-advised dropouts who received partial tuition waivers or no waivers were solely responsible for the decline in persistence.

Minimal threats come from instrumentation as the instrument used to measure student-major fit, based on relative academic strength, demonstrated good face validity (see Figure 1.1). Despite the limitations detailed in this section, the results suggest that effects from scaling LA to offer student advising may be substantial, and they could optimize course-taking and the decision to leave college of marginally enrolled students. However, more research effort is needed to further understand how potential increases in achievement inequality can be tempered and what specific adjustments might be needed for different groups of students with varying socioeconomics and backgrounds. The article also suggests that effects on graduation rates might at times be overstated, and, at the very least, might not depend on improved class selection.

# CHAPTER 2

# THE ROLE OF HIGHER EDUCATION AGENCIES IN PROMOTING INTERNATIONAL STUDENT MOBILITY: THE EXPANSION OF GERMAN DAAD'S OUTBOUND NETWORK

## 2.1 Introduction

The international mobility of students has seen significant growth in recent decades, with the number of internationally mobile students more than doubling between 1998 and 2016, from 1.9 to 5 million (UNESCO 2022). Some countries, such as the UK, have been particularly successful in recruiting a significant portion of their tertiary students from abroad, relying heavily on the tuition fees paid by international students to finance their higher education systems. Other English-speaking countries have also benefited from offering education in the global lingua franca, experiencing increased tuition revenues and a greater availability of high-skilled workers.

Previous research on international student mobility, using gravity models (McMahon 1992; Rosenzweig 2006; Naidoo 2007; Soo and Elliott 2010; Bessey 2012; Jena and Reilly 2013; Beine, Noël, and Ragot 2014; Zheng 2014), has had some success in understanding the factors that influence this phenomenon. Student mobility is driven by international wage differentials rather than domestic schooling capacity constraints (Rosenzweig 2006) and strongly and negatively predicted by distance between countries. However, these studies have primarily focused on factors outside of the scope of educational policy (e.g., wage rate, population, geography, etc.), offering limited guidance to governments and educational institutions. While some governments have implemented recruitment policies (Soo and Elliott 2010; Jena and Reilly 2013; Zheng 2014), the response of students to these types of interventions is largely unknown.

This article seeks to bring the institutional perspective to the literature on international student mobility by examining the relationship between the expansion of outbound offices of Germany's higher education agency, the Deutscher Akademischer Austauschdienst (DAAD) alias "German Academic Exchange Service", and foreign student enrollment in Germany. The DAAD is a higher education agency that competes with other organizations, such as the British Council and Campus France, for the acquisition of international student talent. Higher education agencies are government-funded organizations that partner with schools, universities, and other institutions to provide educational resources, organize outreach activities, and supply information to international students who are considering studying in the country, including information on funding and scholarships, accommodation, and visa requirements. The study uses panel data covering two decades (1996-2016) and 208 countries, during which the DAAD added 53 offices to its network, bringing the total number of offices to 64 at the end of the period (see Table 2.1).

The article examines the impact of hosting a DAAD office on foreign enrollment in a country. It looks at both the level effects from hosting an increasing number of offices and the change in enrollment that occurs when a country becomes a host for a DAAD office. To do that, it implements two treatment effect estimators: a standard two-way fixed effects regression model, where the focal variable is an interval-level variable coding the number of DAAD offices in a country at a time; a synthetic difference-in-differences model, where the focal variable is a binary indicator turning on when a country starts hosting a DAAD office.

Notably, synthetic difference-in-differences cannot capture "increasing doses" of the treatment (D'Haultfœuille, Hoderlein, and Sasaki 2022) nor return estimates for countries that added more than one office. This is because the estimator relies upon the canonical difference-in-difference setup with a dichotomous, not polytomous treatment variable. However, synthetic difference-in-differences are robust to concerns that the relationship between migration and office foundations could work in both directions (i.e., simultaneous or reverse

causality), and that the establishment of offices did not occur at the same time for all countries (i.e., staggered treatment adoption).

The results of the study show that hosting DAAD offices leads to an increase in outbound mobility. In particular, the level effect of each additional office is an 11-point increase (p-value $< 0.001$) relative to the case rate for the same country with one less office. Similarly, the difference-in-differences model estimates a grand average effect from office openings of 19 points (p-value = $0.08$) as compared to the case rate of the same country without any offices. It appears that the majority of the impact is seen in low-income countries, which could indicate that moving into these countries early, before they send large outbound cohorts, may provide a competitive advantage.

This study highlights the potential for countries to influence international student mobility through the funding and maintenance of higher education agencies, rather than simply relying on the relaxation of immigration laws. The findings of the study have direct implications for educational policy-making in the context of intense international competition for foreign talent.

## 2.2 Background

Models of international student mobility have roots in the Random Utility Maximization (RUM) framework (Rosenzweig 2006; Beine, Noël, and Ragot 2014) and the gravity equation framework of human capital mobility (McMahon 1992; Tinbergen 1962; Karemera, Oguledo, and Davis 2000). The RUM framework offers a micro-level understanding of the cost-benefit analysis that individual students undertake, while the gravity equation framework uses quantitative methods to examine the macro-level patterns of international student flows. These two frameworks can be used together in empirical research to provide an understanding of the factors influencing international student mobility.

The decision to study abroad is a multi-stage process that involves choosing to attend college, applying to higher education institutions, and accepting an admission offer. When

Table 2.1: DAAD office openings by country 1996-2016

| Country | Foundation(s) | Classification[1] |
|---|---|---|
| China | 1994, 2001, 2002, 2003 | LM |
| Russia | 1993, 2001, 2002, 2003 | L |
| India | 1960, 2001, 2001, 2009 | L |
| Israel | 2004, 2014 | H |
| Viet Nam | 2001, 2003 | L |
| Turkey | 2000, 2000 | LM |
| Brazil | 1972, 2001 | UM |
| United States | 1971, 2002 | H |
| Lebanon | 2015 | LM |
| Peru | 2015 | LM |
| Ethiopia | 2014 | LM |
| Tunisia | 2014 | L |
| Afghanistan | 2012 | H |
| Belgium | 2012 | L |
| Jordan | 2012 | LM |
| Cameroon | 2009 | L |
| Pakistan | 2008 | L |
| Serbia | 2006 | H |
| United Arab Emirates | 2006 | LM |
| Colombia | 2005 | LM |
| Costa Rica | 2005 | LM |
| Syria | 2005-2011 | LM |
| Armenia | 2004 | L |
| Azerbaijan | 2004 | L |
| Georgia | 2004 | LM |
| Italy | 2004 | H |
| Kazakhstan | 2004 | LM |
| Kyrgyzstan | 2004 | L |
| Latvia | 2004 | LM |
| South Africa | 2004 | UM |
| Spain | 2004 | H |
| Tajikistan | 2004 | L |
| Uzbekistan | 2004 | LM |
| Belarus | 2003 | LM |
| Cuba | 2003 | LM |
| Greece | 2003 | H |
| Iran | 2003 | LM |
| Sudan | 2003-2008 | L |
| Ukraine | 2003 | LM |
| Hungary | 2002 | UM |
| Romania | 2002 | LM |
| Canada | 2002 | H |
| Australia | 2001 | H |
| Ghana | 2001 | L |
| Hong Kong | 2001 | H |
| Singapore | 2001 | H |
| Venezuela | 2001 | LM |
| Viet Nam | 2001 | L |
| Czech Republic | 2000 | UM |
| Mexico | 2000 | UM |
| South Korea | 2000 | H |
| Taiwan | 2000 | H |
| Argentina | 1999 | UM |
| Chile | 1999 | UM |
| Malaysia | 1999 | UM |
| Thailand | 1999 | LM |
| Poland | 1997 | UM |

[1] The World Bank's 1996 classification of countries defines four income groups based on their gross national income per capita: low income (L) if below $786, lower middle income (LM) if in the $786-3,115 range, upper middle income (UM) if in the $3,116-9,645 range, and high income (H) if above $9,645.

students consider moving to a new location, they weigh the potential costs and benefits of the move. These costs and benefits are represented in their decision-making process as a utility function, which is a summary measure of the monetary and non-monetary choice dimensions. If the expected utility of moving to a new location is greater than the expected utility of staying at their current location, then they would decide to make the move. In this decision-making process, the potential destinations are considered as part of the individual's "awareness space" (Wolpert 1965), which refers to the range of locations that they are aware of and considering as potential destinations.

In physics, the gravity equation is a mathematical relationship that describes the strength of the gravitational force between two objects as a function of their masses and the distance between them. In the social sciences, the gravity equation has been used to model the flow of various quantities between locations, such as the flow of goods, services, capital, or people. In the context of international student migration, the gravity equation has been applied to understand the factors that influence the decision of students to move abroad for their studies.

### 2.2.1 The student migration decision

There is broad agreement that the "awareness space" of students, or the institutions and destinations they are aware of and consider, plays a role in their decision, similar to how it influences worker relocations (Sjaastad 1962; Bartel 1979). However, there is a lack of understanding in the literature on the specific timing and considerations involved in this process (Carlson 2013). Do students move conditionally on a positive decision to pursue a tertiary degree? Have they turned down all the potential home destinations by the time they search for destinations abroad?

The research hypothesis being explored here is that higher education agencies may drive in-migrations by acting as information brokers and increasing the awareness space of students. Vrontis, Thrassou, and Melanthiou (2007) describe the consumption of higher education

as a three-part journey: prospective students must perceive a purchase as "urgent"; then, they must become aware of differentiating characteristics setting it apart from the competition; lastly, the perceived utility of the purchase must exceed its monetary and transaction costs. Higher education agencies intervene at all stages of the student consumer journey by creating awareness, highlighting the selling points of national educational institutions, and helping new recruits with paperwork.

The RUM framework (Rosenzweig 2006; Beine, Noël, and Ragot 2014), widely applied in the study of worker migration (Sjaastad 1962; Bartel 1979), can be used to understand the decision-making process of internationally mobile students. Much like relocating workers, students weigh the utility provided by outbound alternatives one at a time against the utility from pursuing the home alternative. To model the migratory flows resulting from individual utility-maximization, scholars have implemented the gravity equation model (McMahon 1992) formerly applied to the study flows of goods, services, and capital as well as migrations (Tinbergen 1962) and which produced "some of the clearest empirical results in international economics and business" (Karemera, Oguledo, and Davis 2000, p. 1746).

The utility of choosing a specific outbound destination $j$, represented by $U_{ij}$, is central to the analytical framework. The utility function accounts for the fact that students migrate at a certain time $M$, complete their studies at a later time $S$, and then enter the workforce at time $S + 1$ until the age of retirement $R$, receiving wage payments during $[S + 1, R]$ based on the market demand for their skills $w_{it}(s_i)$. The utility function also considers the costs of relocation and studies abroad $C_{ijt}$, as well as the one-time psychological burden of migration $P_{ijM}$, and normalizes and omits any non-monetary benefits or "amenities" $A_{ijt}$ that may accrue as a result of the migration (i.e.., lifestyle, exit from warzones, milder climate, etc.). The utility functions for choosing an outbound destination $j$ and the home destination $h$ are as follows:

$$U_{ij} = \left( \int_{t=S+1}^{R} e^{\delta t} w_{it}(s_i) dt + \sum_{t=M}^{R} A_{ijt} \right) - \left( \sum_{t=M}^{S} C_{ijt} - P_{ijM} \right) \tag{2.1a}$$

$$U_{ih} = \left( \int_{t=S+1}^{R} e^{\delta t} w_{iht}(s_i) dt \right) - \left( \sum_{t=M}^{S} C_{ijt} \right) \tag{2.1b}$$

According to the RUM decision rule, a student will only migrate if the utility gained from at least one host destination is greater than the utility gained from remaining at home. The RUM model assumes that students have unbounded knowledge of the benefits and costs of migration, constant non-monetary benefits, and equal living expenses in the host and origin countries. However, it does not assume that students will necessarily join the workforce in the host country after completing their studies. Instead, the model uses the world price of skills $w_{it}(s_i)$, rather than the host country's price, to capture the wages received by the student. It is also assumed that the wage rate in the country of origin $w_{iht}(s_i)$ is different from the world rate and that students have already made the decision to pursue a degree and are not considering workforce employment.

### 2.2.2 Gravity equations of international student mobility

The gravity equation expresses the number of student moves between any two countries $(j, h)$ as proportional to the $K$ dyadic economic, social, and political pull factors $X_{jhk}$ and inversely proportional to their distance $D_{jh}$.

$$Flow_{jh} = \frac{\prod_{k=1}^{K} X_{jhk}^{\beta_k}}{D_{jh}^{\alpha}} \tag{2.2}$$

While some of these factors have shown clear relationships with student migrations, the effects of others have been less consistent and more difficult to discern (see Table 2.2). The table summarizes gravity equation studies with one inflow destination and multiple outflow destinations. Most of the studies use total enrollments as instruments for first-year student

flows exploiting the strong correlation between the two quantities. The time-invariant characteristics of the host country are included as fixed effects and omitted from the output.

Distance is the most significant and negative predictor of student migrations. Other predictors considered are language, borders, population trade, and exchange rate. Sharing a common language and border positively predicts student moves while being landlocked did not have a statistically discernible impact on student inflows into Germany (Bessey 2012). Population is also a driver for out-migrations and so is involvement in international trade (McMahon 1992; Naidoo 2007; Jena and Reilly 2013; Zheng 2014). Contrary to expectations (Jena and Reilly 2013), appreciating exchange rates do not significantly hamper out-migrations (Naidoo 2007; Zheng 2014).

There is ongoing debate regarding the influence of economic factors on international student migrations. Some research suggests that GDP, or a country's overall capacity for economic production, may not significantly impact the number of students enrolling in institutions of higher education. For example, a study by Bessey (2012) found that higher per capita GDP did not lead to increased enrollment in German universities. This may be due to the fact that poorer students who may benefit from increased economic capacity may already be receiving grants and are instead constrained by other non-monetary factors. On the other hand, research by Rosenzweig (2006) suggests that wage differentials play the most significant role in determining enrollment patterns in higher education. Specifically, the study found that international students are more likely to enroll in US higher education institutions in pursuit of higher wages, rather than being crowded out by educational systems at capacity. This supports the "brain drain" model, which posits that students will follow economic opportunities rather than eschewing the constraints of overpopulated domestic educational systems.

There is mixed evidence on the impact of specific features of the education system, such as tuition fees and university rankings. Some studies have found that tuition fees may not significantly impact migration decisions, with students often choosing to attend a particular

institution regardless of the cost. This may be due to the fact that once students have made the decision to pursue higher education, the difference in fees between institutions may not be a major factor in their decision-making process (Soo and Elliott 2010). Similarly, university rankings have been found to be largely uncorrelated with student mobility on a global scale (Rosenzweig, 2006), albeit rankings may serve as a status signal and may impact élite migrations among high-income students.

The current study speaks to a gap in the international student migration literature by bringing attention to the role of institutions. To further encourage the study of this topic, it is important to note that there is strong theoretical support for the effects of higher education agencies, but there is a lack of empirical evidence to back up these claims. Few studies (Soo and Elliott 2010; Jena and Reilly 2013; Zheng 2014) have investigated specific initiatives, such as the two phases of the UK Prime Minister's Initiative (PMI) in 1999 and 2006 which expanded the British Council's capabilities and streamlined the visa application process. Prime Minister Tony Blair's goal was to reassert the British presence in the worldwide market for talent dominated by the United States by reducing the overhead of the migration decision and creating a clear pathway for UK employment. Soo and Elliott (2010) found that the number of British Council exhibitions attended by universities in 2006-2007 predicted international enrollment, but the relationship was non-linear and the effect faded when considering the quality of the institution. Other studies (Jena and Reilly 2013; Zheng 2014) have produced mixed results, potentially due to differences in the operationalization of the focal variable.

## 2.3 Methodology

### 2.3.1 Data

This research study analyzed the universe of international student inflows into German universities from 208 countries in the years between 1996 and 2016 (inclusive). This period witnessed the most rapid expansion of DAAD's outbound network, which grew from 11

Table 2.2: Gravity model estimates from selected single-destination studies

| | Author(s) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Agarwal & Winkler** | **McMahon** | **Rosenzweig et al.** | **Naidoo** | **Bessey** | **Jena & Reilly** | **Zheng** | **Levatino** |
| *Year* | (1985) | (1992) | (2006) | (2007) | (2012) | (2013) | (2014) | (2017) |
| *Period* | 1954-1973 | 1960-1975 | 2003-2004 | 1985-2003 | 1997-2002 | 2001-2008 | 1994-2007 | 2002-2011 |
| *Model* | OLS | OLS | OLS | OLS | RE | OLS | GLS | OLS |
| *R2* | 0.6 | 0.47 | 0.71 | 0.9 | 0.69 | 0.51 | N/A | N/A |
| *Host* | USA | USA | USA | GBR | DEU | GBR | GBR | AUS |
| *n* | 25 | 18 | 125 | 171 | 144 | 89 | 42 | 153 |
| *Outcome* | Enrollment | Mobility rate | Enrollment | Enrollment | Enrollment | VISA issues | Enrollment | Enrollment |
| *GDP* | (+) | (+) | (+) | (+) | N/A | (nss) | (-) | (+) |
| *Forex* | N/A | N/A | N/A | (nss) | N/A | (-) | (-) | N/A |
| *Trade* | N/A | (+) | N/A | N/A | N/A | (+) | (+) | N/A |
| *Population* | N/A | N/A | (+) | N/A | (+) | (nss) | (+) | (+) |
| *Language* | (nss) | N/A | N/A | N/A | N/A | (+) | (+) | (+) |

The table presents the characteristics of the studies and the significance of their gravity parameter estimates, which are coded as follows: negative (-), positive (+), not statistically significant (nss), and not available (N/A).

offices at the beginning of the time-frame to 64 offices. Variables reflecting time-variant heterogeneity at the country-level were collected from the World Bank, the United Nations, and the CIA Factbook and included in the analysis dataset.

The analysis used two different methods to determine how the establishment of DAAD offices impacted foreign enrollment: a standard two-way fixed effects model, which examined how changes in the number of DAAD offices affected foreign enrollment; and a synthetic difference-in-differences model, which compared changes in foreign enrollment before and after the establishment of the first DAAD office in a country. The treatment variable, or the variable being analyzed, was ordinal (i.e., office count) in the first method and binary (i.e., hosting an office or not) in the second method. The treatment group used for the synthetic difference-in-differences model was limited to countries that had established one office.

The dependent variable in the analysis was the count of foreign students enrolled in German universities. This count variable may exhibit overdispersion (Silva and Tenreyro 2006), which can lead to deviations from the residual normality assumption that is required for

multiple linear regression. In addition, the treatment (i.e., establishment of DAAD offices) may be endogenous to pre-treatment factors such as sending capacity, and the treatment was implemented in a staggered manner (Goodman-Bacon 2021; Baker, Larcker, and Wang 2022; Callaway and Sant'Anna 2021). Endogeneity bias may occur if the establishment of offices follows migration flows rather than causing them. In this case, the increase in migration flows may not be a result of the offices themselves, but rather due to other factors that influence both the establishment of offices and migration flows. The staggered implementation of the treatment invalidates the canonical difference-in-differences setup, which assumes a common treatment period across all treated units. This may also lead to a difference-in-differences estimator identifying spurious variation, rather than causal variation, reflecting the timing of treatment and size of the treated group in each cohort (Goodman-Bacon 2021).

### 2.3.2 Standard two-way fixed effects estimation

The expanded gravity equation models used to estimate the effect of adding one or more DAAD offices on international students outflows into Germany was:

$$Flow_{jt} = \exp(\gamma_j + \phi_t + \sum_{k=1}^{K} \beta_k X_{jtk} + \delta DAAD_{jt} + \epsilon_{jt})$$

where:

- $Flow_{jt}$ is the count of foreign students of country $j$ hosted in Germany at time $t$;
- $\gamma_j$ and $\phi_t$ are the country and year fixed effects;
- $[\beta_1, ..., \beta_K]$ is a vector of $K$ coefficients storing $j$'s time-varying effects;
- $\delta$ is the target coefficient storing the level effects from increasing $j$'s DAAD office count;

- $\epsilon_{jt}$ is an individual error clustered at the country-year level.

By virtue of fixed effects estimation, coefficient estimates are calculated using variation within a country over time (i.e., "within-variation"). Resultantly, coefficients capture ceteris paribus differences in the number of student contingents by adding one office rather than differences between countries with a different office count (i.e., "between-variation"). Four panel data estimators were tested: 1) ordinary least squares; 2) log-transformed ordinary least squares; 3) Poisson; and 4) negative binomial. Residual diagnostics (see Figure A.1 in the Appendix) and goodness-of-fit tests were conducted leading to the choice of the negative binomial model as the most appropriate for the data.

First, a baseline linear model was estimated using student enrollment as the dependent variable and population size, GDP per capita, imports, exports, and year and country fixed effects as predictors. The second model considers a log-transformation of the dependent variable as suggested by the log-likelihood profiles for the Box-Cox parameter (i.e., $\lambda = 0$) using the same base predictors and the log of student counts as the outcome variable with a one-unit shift to avoid undefined terms. Next, the Poisson and the more general negative binomial models were tested. The negative binomial distribution generalizes Poisson allowing for separate adjustment of the variance thanks to the shape parameter $\alpha$ found through optimization.[1]

Adoption of negative binomial was upheld by failures of the assumptions required by the other models. Residual diagnostics of the ordinary least squares and log-transformed models revealed heavy tails in the distribution of the residual variance and a violation of the normality assumption. The Poisson model also showed a poor fit to the data as per the quantile-quantile plot and the deviance test (p-value $< 0.001$); furthermore, the conditional mean trailed several orders of magnitude behind the conditional variance (dispersion parameter = 82.3) violating the main Poisson assumption. No need was found to separately account for the zero-counts, which are in the single digits ($< 5\%$).

[1] R's *glm* package comes with a built-in optimizer printing the reciprocal of the dispersion parameter $\theta = \frac{1}{\alpha}$ to the output (Hilbe 2014)

The Hausman (1978) specification test was used to evaluate the appropriateness of using fixed effect estimation. The significant test statistic (p-value $< 0.001$) indicated a correlation between the time-varying predictors and the fixed effects, leading to the conclusion that the fixed effect estimator is more appropriate and efficient in this case. The fixed estimator purges time-invariant heterogeneity, observed and unobserved, by "demeaning" the dependent variable by the group- and time-averaged values of student flows for each country.

In the analysis, serial correlation and heteroskedasticity were potential concerns due to the two levels of clustering in the data by year and country (Arellano et al. 1987), and were detected by the Breusch-Godfrey (Breusch 1978; Godfrey 1978) and Breusch-Pagan (Breusch and Pagan 1979) tests. These types of correlation, longitudinal and cross-sectional, can cause standard error estimates not to fully capture predictive uncertainty. However, the decision was made not to cluster the standard errors following the argument that clustering may not be necessary when the dataset is the universe of an event set rather than a sample, as there is no uncertainty due to sampling that needs to be accounted for through error clustering (Abadie et al. 2022).

### 2.3.3 Synthetic difference-in-differences

The use of synthetic difference-in-differences is motivated by endogeneity concerns due to the potential endogeneity of office foundations, and staggered treatment adoption. In this context, endogeneity concerns refer to the potential for the treatment (i.e., establishment of DAAD offices) to be influenced by third factors which explain the outcome (i.e., foreign enrollment in German universities) rather than the other way around. This can lead to biased estimates of the treatment effect.

Synthetic difference-in-differences is a method that aims to address endogeneity concerns in difference-in-differences analysis by using synthetic control methods. The synthetic control group is designed to closely match the treated unit on the pre-treatment trends and

covariates, in order to control for confounding factors that may affect the outcome. The donor pool is a set of units that are used as a reference group (i.e., placebos) to construct the synthetic control group for each treated unit in synthetic difference-in-differences analysis. The synthetic control group for each treated unit is constructed by combining units from the donor pool in a way that minimizes the difference between the treated unit and the synthetic control group on the pre-treatment trends and covariates.

A growing literature (Goodman-Bacon 2021; Baker, Larcker, and Wang 2022; Callaway and Sant'Anna 2021) attests to the failures of two-way fixed effects regression to identify the causal difference-in-differences parameter when the rollout of the treatment is staggered. Due to asynchronous office foundations, the canonical difference-in-differences treatment matrix with $\{1^+, \ldots, J^+\}$ treatment countries, $\{1^-, \ldots, J^-\}$ placebos, and $\{1, \ldots, T^-, T^+, \ldots,$ time-periods shown below is unavailable:

$$\begin{bmatrix} \text{DAAD}_{1^+,0} & \ldots & \text{DAAD}_{J^+,0} & \text{DAAD}_{1^-,0} & \ldots & \text{DAAD}_{J^-,0} \\ \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ \text{DAAD}_{1^+,T^-} & \ldots & \text{DAAD}_{J^+,T^-} & \text{DAAD}_{1^-,T^-} & \ldots & \text{DAAD}_{J^-,T^-} \\ \text{DAAD}_{1^+,T^+} & \ldots & \text{DAAD}_{J^+,T^+} & \text{DAAD}_{1^-,T^+} & \ldots & \text{DAAD}_{J^-,T^+} \\ \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ \text{DAAD}_{1^+,T} & \ldots & \text{DAAD}_{J^+,T} & \text{DAAD}_{1^-,T} & \ldots & \text{DAAD}_{J^-,T} \end{bmatrix} = \begin{bmatrix} 0 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ 0 & \ldots & 0 & 0 & \ldots & 0 \\ 1 & \ldots & 1 & 0 & \ldots & 0 \\ \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ 1 & \ldots & 1 & 0 & \ldots & 0 \end{bmatrix} \tag{2.3}$$

Conversely, to be defined here is a block matrix with multiple treatment periods $t^{j+}$ like the one in equation (Equation 2.4). Forcing a two-way fixed effects estimator on a similar block matrix might introduce bias in the causal parameter of interest (Goodman-Bacon 2021). The estimator would use a combination of "allowed" and "prohibited" counterfactuals to estimate the difference-in-differences parameter: country $1^+$ hosting an office starting at time $T^{1+}$ would factor into the control group for country $J^+$ hosting an office starting at $T^{J+}$ ("prohibited comparison"), thus joining countries $j^+$ not yet hosting offices

by time $T^{1+}$ and countries $j^{-}$ which never hosted offices ("allowed comparisons"). Vice versa, country $J^{+}$ might be legitimately used as a comparison for treated country $1^{+}$ at time $T^{1+}$ because not yet treated at that time.

$$\begin{bmatrix} \text{DAAD}_{1^+,0} & \dots & \text{DAAD}_{J^+,0} & \text{DAAD}_{1^-,0} & \dots & \text{DAAD}_{J^-,0} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ \text{DAAD}_{1^+,T^-} & \dots & \text{DAAD}_{J^+,T^-} & \text{DAAD}_{1^-,T^-} & \dots & \text{DAAD}_{J^-,T^-} \\ \text{DAAD}_{1^+,T^{1+}} & \dots & \text{DAAD}_{J^+,T^{1+}} & \text{DAAD}_{1^-,T^{1+}} & \dots & \text{DAAD}_{J^-,T^{1+}} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ \text{DAAD}_{1^+,T^{J+}} & \dots & \text{DAAD}_{J^+,T^{J+}} & \text{DAAD}_{1^-,T^{J+}} & \dots & \text{DAAD}_{J^-,T^{J+}} \end{bmatrix} = \begin{bmatrix} 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 1 & \dots & 1 & 0 & \dots & 0 \end{bmatrix} \tag{2.4}$$

One approach to addressing the staggered implementation of the treatment is to reduce the treatment matrix to valid submatrices that conform to the canonical difference-in-differences block design (Goodman-Bacon 2021; Baker, Larcker, and Wang 2022; Callaway and Sant'Anna 2021). This can be achieved through the use of estimators such as the Callaway-Sant'Anna estimator (Callaway and Sant'Anna 2021), which is a two-stage procedure that involves disaggregating the treatment matrix into valid submatrices and reaggregating the effect sizes calculated from individual regressions. These regressions use units treated at one time and controls that are either never-treated or not-yet-treated, reducing the treatment matrix to a canonical difference-in-differences matrix. The use of not-yet-treated controls can increase the power and decrease self-selection bias in the control group, but at the cost of losing a clean, time-invariant control group.

As a case in point we consider the case of Turkey, Italy, and Belgium. Their treatment turns on in 2001, 2005, and 2013 through the end of the panel respectively. In this context, the identification of the causal parameter for Italy in 2005 using Turkey and Belgium as controls would be problematic because Turkey was already treated in 2001, making it an invalid placebo for Italy in 2005. In contrast, Belgium in 2005 would be a valid control

because it had not yet received the treatment. Removing Turkey from the treatment matrix would allow the analysis to conform to canonical difference-in-differences.

$$\begin{bmatrix} \text{TUR}_{1996} & \text{ITA}_{1996} & \text{BEL}_{1996} \\ \vdots & \dots & \vdots \\ \text{TUR}_{2001} & \text{ITA}_{2001} & \text{BEL}_{2001} \\ \vdots & \dots & \vdots \\ \text{TUR}_{2005} & \text{ITA}_{2005} & \text{BEL}_{2005} \\ \vdots & \dots & \vdots \\ \text{TUR}_{2012} & \text{ITA}_{2012} & \text{BEL}_{2012} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ \vdots & \dots & \vdots \\ 1 & 0 & 0 \\ \vdots & \dots & \vdots \\ 1 & 1 & 0 \\ \vdots & \dots & \vdots \\ 1 & 1 & 0 \end{bmatrix} \tag{2.5}$$

$$\begin{bmatrix} \text{ITA}_{1996} & \text{BEL}_{1996} \\ \vdots & \vdots \\ \text{ITA}_{2005} & \text{BEL}_{2005} \\ \vdots & \vdots \\ \text{ITA}_{2012} & \text{BEL}_{2012} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix} \tag{2.6}$$

A synthetic difference-in-differences estimate (Arkhangelsky et al. 2021) is obtained by slicing the treatment matrix down to a form that conforms to the canonical difference-in-differences design (see Equation 2.3), and using the pre-treatment periods from the control countries to generate unit weights and time weights. Unit weights define a linear combination of untreated outcomes that best match the pre-treatment outcomes of the treated and control countries, while time weights represent a linear combination of time periods that best fit the average post-treatment outcome of each control unit with its pre-treatment outcomes. The time weights are the distinctive addition of synthetic difference-in-differences to the standard synthetic control method (Abadie, Diamond, and Hainmueller 2010, 2015) and are pseudo time series parameters weighting more heavily time-periods which are predictive of post-treatment enrollment. In this case, time weights are useful because stu-

dent outflows are being predicted by two factors: enrollment in other countries (cross-sectionally) and lagged enrollment (longitudinally).

A weighted least square regression can be calculated using unit and time weights to estimate the average treatment effect. The ATT parameter for country $j^+$ results from a two-way fixed effects regression of the outcome on the submatrix including the treated country $j^+$ and its donor pool observed throughout the panel:

$$Flow_{j^+t} = exp(w_{j^+} \cdot w_t(\gamma_j + \phi_t + \delta_{j^+} DAAD_{j^+t} + \epsilon_{j^+t})) \tag{2.7}$$

where unit weights $w_j$ are assigned to the treated country and the control countries, while time weights $w_t$ are assigned to each time-period. The undefined unit weight for the treated unit defaults to the relative size of the donor pool, and the undefined time weights for the post-treatment period default to the average post-treatment indicator.

The remainder of the section discusses parameter interpretation and effects aggregation. The synthetic difference-in-differences framework does not impose restrictions on the functional form, therefore fixed effects negative binomial regression can be used to return the causal parameters $[\hat{\delta}_{1^+}, \ldots, \hat{\delta}_{J^+}]$. However, the prior ordinal interpretation of the causal parameter does not transfer to synthetic difference-in-differences. Now, each $\hat{\delta}_{j^+}$ is a binary indicator reflecting the average change in student enrollment determined by an office foundation. Using the potential outcome notation:

$$\hat{\delta}_{j^+} = \frac{1}{T - T^{j+}} \sum_{t=T^{j+}}^{T} Flow_{j^+t}(1) - Flow_{j^+t}(0) \tag{2.8}$$

Synthetic difference-in-differences estimates allow for the analysis of the impact of one office foundation on foreign enrollment in German universities, but only for countries that have hosted one and only one office over the observation period. This is because countries that have opened more than one office are not considered valid candidates for estimation.

Potentially, one could truncate the observation period until before the opening of the second office, but none of the countries that opened a second office started the panel as not-yet-treated and all hosted at least one office from the beginning.

Similarly, countries that started and completed the observation period with one office (e.g., France) are also excluded from the estimation set. The control group for each treatment unit $j^+$ is composed of countries that have not opened any offices by the end of the observation period. Synthetic counterfactuals can only be constructed using untreated variation, so any countries that receive treatment during $j^+$'s post-treatment period $(T^{j+}, T)$ are not included in its synthetic control. This allows the block matrix for each treated country to conform to a valid 2x2 block matrix with one treated country-treatment period combination and a group of not-yet-treated units.

The final analysis aggregates the individual treatment effects using two methods: inverse-variance weighting (IVW) and exposure and inverse-variance weighting (EIVW). IVW is a method of weighting the individual treatment effects by the inverse of their variance, in order to give more weight to estimates with lower variance and higher precision. The final estimate is calculated as the weighted average of the individual treatment effects, with the weights being the inverse of the variance of each estimate.

EIVW is a variant of IVW that takes into account the length of the exposure window (i.e., the time period during which the treatment was in effect) in the calculation of the weights. The EIVW estimate is calculated as a weighted average of the individual treatment effects, with the weights being the inverse of the variance of each estimate multiplied by the length of the exposure window. This method gives more weight to countries that opened offices earlier in the panel and provided more post-treatment data points, as these countries have longer and potentially less noisy exposure periods. The EIVW-ATT estimate is calculated as follows:

$$\overline{\hat{\Delta}} = \frac{\sum_{j^+=1^+}^{J^+} \delta_{j^+}(1/\hat{\sigma}_{j^+})(T - T^{j+})}{\sum_{j^+=1^+}^{J^+}(1/\hat{\sigma}_{j^+})(T - T^{j+})} \tag{2.9}$$

where $(1/\hat{\sigma_{j^+}})$ is the inverse variance weight and $(T - T^{j+})$ is the exposure weight. The IVW aggregate removes the exposure weight $(T - T^{j+})$ and only considers the variance of the estimate.

## 2.4 Results

### 2.4.1 Main estimates

The analysis in this study found that the opening of German Academic Exchange Service (DAAD) offices is strongly related to an increase in the number of students coming to study in Germany. This relationship was examined using two-way fixed effect estimation and synthetic difference-in-differences, and was found to be robust across multiple model specifications (see Table 2.3). The negative binomial model was determined to be the best fit for the data according to a Weight of Evidence (WoE) test, which compares the goodness of fit of different regression models. The results from this model align with estimates obtained through log-transformed ordinary least squares regression, but contrast with estimates from ordinary least squares and Poisson regression.

The model specification was validated using the Akaike Information Criterion corrected (AICc), which is a measure of the relative fit of different models (see Table Table A.1 in the Appendix). The WoE test is a method used to evaluate and compare the goodness of fit of different regression models ranking the models based on their relative distance from the best-performing model, which is set to zero. The WoE statistic is calculated by subtracting the best model's AICc value from the AICc value of each model in the set to obtain the $\Delta_{AICc}$, and then normalizing this difference to a value between 0 and 1. Larger WoE values indicate a better relative fit of the model.

According to the WoE test, the model that includes GDP per capita as a predictor and specifies the other quantitative predictors (population, GDP per capita, imports, exports, and urban population) as linear is the most appropriate for the data. This is indicated by the WoE statistic in column 5 of Table Table A.1. The models that include GDP as a predictor rank second and third, followed by the GDP per capita model with log specification of the quantitative predictors. It is also notable that the log-likelihood (column 6) is stable across the highest-ranked models, indicating that the estimates are not sensitive to model specification. This suggests that the chosen model is a reliable fit for the data.

Table 2.3 presents the main estimates from the model selected using the WoE test. These estimates include those from the negative binomial model, as well as ordinary least squares regression, log-transformed linear regression, and Poisson regression. Table 2.3 also provides information about the interpretation of the coefficients (column 2), the parameter estimates (column 3), and their significance (column 4). The linear regression coefficient can be interpreted as the ceteris paribus increase in student moves to Germany with one more office, while the coefficient from the log-transformed model represents the percentage increase. The Poisson and negative binomial coefficients have the same interpretation as incidence rate ratios, or the sending rate that countries adding one office exhibit relative to the baseline model with one less office and the same specification otherwise. All of the coefficient estimates capture within-variation or longitudinal changes in the target measure, rather than cross-sectional variation across countries with different office counts.

The negative binomial model estimates that the opening of a German DAAD office leads to a 11-point increase in the sending rate, or the rate at which students are sent to study in Germany (p-value $< 0.001$). This means that the sending rate for a country with one additional DAAD office is estimated to be 1.11 times as high as the rate for the same country without the office. The log-transformed model also estimates a similar percent increase in students sent to be about 11%. In contrast, the Poisson and linear regression models did not fit the data well, as indicated by the residuals against fitted values plots, and returned

similar and insignificant effect estimates.

Table 2.3: Average office numbers effect from standard two-way fixed effects

| **Functional Form** | | **95% C.I.** | | |
|---|---|---|---|---|
| **Link** | **Interpretation** | **Coef.** | **Lwr.** | **Upr.** |
| *Linear* | Students sent (number) | 106 | -36 | 248 |
| *Log* | Students sent (% increase) | $0.11^{***}$ | 0.04 | 0.19 |
| *Poisson* | Sending rate (ratio) | 1.00 | 0.99 | 1.01 |
| *Negative Binomial* | Sending rate (ratio) | $1.11^{***}$ | 1.03 | 1.19 |

*Note*. This table presents the standard two-way fixed effects estimates and their significance for each of the four functional forms ($^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$).

Synthetic difference-in-differences estimates (Table 2.4) validate the effects estimated from standard two-fixed effects. Grand average treatment effects (column 2) and p-values (column 3) are reported for both inverse-variance weighting (IVW) and Length of Exposure and Inverse-Variance Weighting (EIVW), and are stable with and without the inclusion of covariates in weights calculations.

The grand average EIVW estimate, which is a measure of the relationship between the opening of DAAD offices and student inflows to Germany, amounts to 1.19 incidence rate ratios. This estimate is larger than the IVW estimate of 1.13 and is significant at the 10% confidence level. The EIVW estimate is more directly comparable to standard two-way fixed effects estimates because it takes into account the fact that countries with longer exposure to DAAD offices are given more weight in the analysis (Goodman-Bacon 2021). In contrast, the IVW estimate treats countries with longer and shorter exposures equally and only accounts for the uncertainty of short exposure windows through inverse-variance weights. However, these weights are based on the standard errors of the synthetic difference-in-differences estimates, which are largely determined by the size of the placebo groups. Therefore, the EIVW estimate is the preferred alternative.

Average treatment effects from individual countries (see Table A.2) provide further di-

rectional learning although intensity of treatment might not be properly evaluated in this environment. Of the 48 individual effects, 38 underscore positive impacts from office foundations. Furthermore, of the ten negative effects, only a handful of extreme cases grab attention. The most extreme case is Sudan (case ratio = 0.47), which briefly hosted a DAAD office in 2003 at the same time that it was experiencing the Darfur conflict and the eventual secession of South Sudan in 2011. The second most extreme case is Hungary (case ratio = 0.68), which was undergoing a tightening of its repressive regime under the leadership of Victor Orbán during the period of analysis.

Table 2.4: Grand-average office foundations effect from synthetic difference-in-differences

| | **Raw**[1] | | **Covariate**[2] | |
|---|---|---|---|---|
| **Weighting** | **ATT** | **P-value** | **ATT** | **P-value** |
| Inverse-Variance (IVW) | 1.13 | 0.50 | 1.13 | 0.50 |
| Length of Exposure Inverse-Variance (EIVW) | 1.19* | 0.09 | 1.20 | 0.08 |

*Note*. This table presents the synthetic difference-in-differences estimates and their significance using inverse-variance and length of exposure inverse-variance weighting ($^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$). The coefficients reflect sending rate ratios or ratios of the migration case rate with and without the office.

[1] The weights are determined by matching on the pre-foundation trends.

[2] The weights are determined by matching on the pre-foundation trends and the covariates included in the standard two-way fixed effects regression.

### 2.4.2 Estimates by World Bank income group

Analyzing the results by country allows for the identification of potential transmission mechanisms of the impact of DAAD office openings on student inflows to Germany. As shown in Table 2.5, the overall findings of this study contrast with those from previous studies conducted in the UK, but align with a prior study on Germany's international student migrant flows (Bessey 2012).

The countries were divided into three groups based on World Bank income classification, with lower middle-income and upper middle-income countries combined to ensure suf-

ficient sample sizes for estimation. The results showed that only the low-income group returned significant and positive estimates, while the middle- and high-income groups largely returned insignificant or negative estimates. The negative synthetic difference-in-differences estimate for the high-income group (incidence rate ratio = 0.91) warrants further investigation.

The study found that the low-income group was the only group to show a positive response to the opening of offices. This contradicts the expectation that the middle-income group, which has both the financial resources to support migrations and often underdeveloped educational opportunities for their best and brightest, would show the most significant response (McMahon 1992). The finding also conflicts with the findings of a previous study (Zheng 2014) that observed little response from poorer non-OECD countries to the British PMI.

In the low-income group, the standard two-way fixed effects estimate (incidence rate ratio = 1.15) has the same sign, but is smaller in magnitude than the synthetic difference-in-differences estimate (incidence rate ratio = 1.32). It must be reiterated that standard two-way fixed effects estimates capture the marginal effects of adding offices, while the synthetic difference-in-differences estimates capture the factor effects of starting to host offices. The standard two-way fixed effects estimate shows that, holding other factors constant, the addition of one more office leads to 15-point increase in the sending rate. The synthetic difference-in-differences estimate, on the other hand, suggests that the first office leads to a 32-point increase after foundation.

All in all, the analysis suggests that the impact of opening DAAD offices on student inflows to Germany is most pronounced in low-income countries, possibly due to the ability to establish informational rents and dominate the "awareness space" of student migrants in these countries. In contrast, the impact of opening offices in middle-income and high-income countries is generally insignificant, possibly due to the presence of established outbound contingents and more entrenched "awareness space". The results thus suggest that the financial value of opening offices in developed countries is low in terms of student

flow.

Table 2.5: Averag estimated effects by World Bank income group

| | Two-Way Fixed Effects | | Synthetic DiD | |
|---|---|---|---|---|
| **Countries** | **Coef.** | **P-value** | **ATT** | **P-value** |
| Low-Income | 1.15* | 0.050 | 1.32*** | <0.001 |
| Middle-Income | 0.98 | 0.702 | 0.98 | 0.238 |
| High-Income | 0.97 | 0.616 | 0.91* | 0.063 |

*Note*. This table presents the standard two-way fixed effects and synthetic difference-in-differences estimates ($^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$). The coefficients reflect sending rate ratios or ratios of the migration case rate with and without the office.

## 2.5 Conclusions

This study found that the establishment of a higher education agency, such as the DAAD office, is associated with an increase in student migration. This section summarizes the policy insights from the study while indicating transmission mechanisms from theory and highlighting the limitations of the study.

On average, the addition of one office leads to a migration incidence that is 1.11 times as high as the pre-expansion incidence, and the establishment of the first office leads to a migration incidence ratio of 1.19. This finding is in contrast to the limited impact of most factors that influence international student mobility, which are often beyond the control of governments.

The largest opportunity for this type of policy is in low-income countries, where informational asymmetries and first mover advantage may play a role in attracting international students. This includes building a reputation through the establishment of outbound offices, which can put destinations on the map for potential student recruits and establish the reputation of educational systems over time.

Higher education agencies can profit from informational asymmetries and the path-dependent

nature of migration flows. First of all, the decision-making process of international students tends to resemble Herbert Simon's concept of satisficing rather than utility maximization as students maximize their utility from migration subject to the informational constraints of their awareness space. Second of all, the process is path-dependent, as international students tend to gravitate towards destinations with established expatriate communities.

There are several limitations to the findings of this study that should be considered. Firstly, it is important to recognize that only a subset of college-age students becomes internationally mobile at any given time, so it is difficult to determine how much of the potential migrant pool is being influenced by the presence of DAAD offices. Additionally, it is unclear whether the presence of offices leads to an increase in international mobility among students who may not have previously migration-seeking. Finally, it is possible that the causal relationship between office foundations and student migration may be reversed, with office foundations following rather than causing migration patterns. However, the use of synthetic difference-in-differences helps to temper this potential concern.

This study emphasizes the important role that institutions, such as higher education agencies, can play in increasing international student mobility. These agencies can help to build the reputation of educational systems in countries that are still developing, particularly for recruiters that do not have the advantage of offering instruction in the English language. Findings are therefore of particular significance for countries outside of the Anglophone block, which may need to rely on the first mover advantage to establish a presence in the higher education market.

# CHAPTER 3

# (NO) TRADE-OFF BETWEEN NUMERACY AND VERBAL REASONING DEVELOPMENT: EVIDENCE FROM ITALY'S ACADEMIC TRACKING

## 3.1 Introduction

A growing literature validates the effect of math and science coursework on students' academic outcomes (Rose and Betts 2004; Long, Iatarola, and Conger 2009; Long, Conger, and Iatarola 2012) and earnings (Joensen and Nielsen 2009; Cortes, Goodman, and Nomi 2015; Goodman 2019). However, the existence of learning trade-offs between numerical and verbal reasoning skills has not been evaluated in the process of skill formation (Heckman 2006) and educational investment (Cattell 1987; Park, Lubinski, and Benbow 2007; Coyle et al. 2015).

The failure to observe any trade-offs between types of cognition counters Nobel Prize recipient Umberto Eco's argument that technical and critical thinking skills are substitutes in the education production function. Famously, Eco defended the teaching of humanities and the classical heritage of the country as a conduit to foster critical thinking in a public debate featuring economist Andrea Ichino (Schiavazzi 2014). Eco's central claim was that critical thinking skills might not be fostered vicariously through technical instruction and must be fed directly into the high school curriculum. Eco thus downplayed Ichino's suggestion to downsize the teaching of the humanities. Findings of this study, which will be argued to be lower bound estimates of the "STEM advantage", dispute the alleged trade-off at the core of Eco's argument.

The study elicits academic tracking in Italy's secondary education using OECD's Programme for International Student Assessment (PISA) test scores as a testbed. PISA is credited to be a test of "functional knowledge" in math, reading, and science (OECD 2016)

or adult-life preparedness test (Reardon 2018) which relates to but is neither a general intelligence nor a domain knowledge test (Pokropek et al. 2022). Suppose a trade-off exists between types of cognitive skills students gained in different tracks. In that case, gains in the math and science sections of the test afforded by a stronger STEM curriculum must happen at the expense of achievement in verbal literacy.

No evidence of trade-offs between numerical and verbal reasoning ability is found. After matching on the factors determining track selection, Liberal Arts students exhibit lower achievement than students in the STEM track. Decomposition of the gap further shows that differences in resources – i.e., the individual, parental, and school inputs of the education production function – are not responsible for the STEM advantage, but differences in returns to resources – i.e., the "production function" – are.

The relevance of the findings is established by the prevalence of pull-out and elective classes (OECD 2021). Italy tracks its students at the end of middle school at the age of 13, sorting them into a secondary institution from one of the tracks in the three main pathways, vocational (*Professionale*), technical (*Tecnico*), and general education (*Liceo*). This study compares the achievement of students in the Liberal Arts and STEM track on the math, reading, and science section of OECD's PISA

Table 3.1 and Table 3.2 summarize the instructional units for year I and II at each track. The track envisions physics requirements for year I and II, absent from the Liberal Arts curriculum, it has less Latin credits hours (3) compared to Liberal Arts (5) and none of the Greek instructional units (5). Liberal Arts graduates most of the administrative, political, and intellectual upper class, similar to the British Politics, Philosophy, and Economics (PPE) curriculum. Notably, all serving Italian Prime Ministers for the first two decades of the 21st century were Liberal Arts graduates. Conversely, the technical track attracts a much larger enrollment at about 25% of total secondary students, nearly three times that of Liberal Arts and 45% of general education enrollment.

Table 3.1: Instructional units in year I and II of high school

| **Liberal Arts** | **STEM** |
|---|---|
| Latin (5) | Math (5) |
| Greek (5) | Italian (4) |
| Italian (4) | Latin (3) |
| Foreign language (3) | Foreign language (3) |
| History & Geography (3) | History & Geography (3) |
| Math (3) | Physics (2) |
| Science (3) | Science (2) |
| Physical Education (2) | Technical Drawing (2) |
| | Physical Education (2) |

*Notes*. Credit requirements are measured in hours/week (in parentheses) and apply nationwide in schools offering the relevant curricula.

Table 3.2: Instructional units in year I and II of high school

| | **Liberal Arts** | **STEM** |
|---|---|---|
| *Literacy* | Grammar, Epic Poetry | Grammar, Epic Poetry |
| *Math* | Algebra, Geometry, IT | Algebra, Geometry, IT |
| *Science* | Earth Science, Biology, Chemistry | Earth Sciences, Biology, Chemistry |
| *Physics* | – | Optics, Dynamics, Thermodynamics |

*Notes*. Course requirements apply nationwide in schools offering the relevant curricula.

## 3.2 Background

### 3.2.1 Academic Strength and Skill Formation

There are two ways to look at skills: as a generative process or statically as a skill bundle. This article considers both, placing itself at the crossroads of four bodies of literature: academic tilt and investment theories of niche-picking (Cattell 1987; Park, Lubinski, and Benbow 2007; Coyle et al. 2014; Coyle et al. 2015; Coyle 2018; Becker et al. 2022), human capital accumulation and skill development (Heckman 2006; Cunha and Heckman 2007; Cunha, Heckman, and Schennach 2010; Doyle et al. 2017), gifted youth (Achter et al. 1999; Lubinski et al. 2001; Park, Lubinski, and Benbow 2007; Wai et al. 2010;

Wang, Eccles, and Kenny 2013; Bernstein, Lubinski, and Benbow 2021), and returns from technical coursework (Altonji 1995; Levine and Zimmerman 1995; Rose and Betts 2004; Long, Conger, and Iatarola 2012; Cortes, Goodman, and Nomi 2015; Long, Iatarola, and Conger 2009; Goodman 2019).

Ability tilt is proxied by relative differences between a student's percentile score in the verbal and math domains. The construct has shown to be predictive of major and career choices (Coyle et al. 2014; Coyle et al. 2015; Coyle 2018), other academic and industry outcomes such as patent issues (Park, Lubinski, and Benbow 2007), and has been linked to national economic output using PISA data (Hunt and Wittmann 2008). Consistency over time of ability tilt has been investigated across genders (Coyle, Snyder, and Richmond 2015) and geographies (Becker et al. 2022), and the higher predictiveness of math tilt over verbal tilt has been noted along with field-specific differences in predictive power (Coyle et al. 2014).

Studies of precocious children intertwine with academic tilt literature pinpointing early life divergence of educational investment based on relative academic aptitude. Two normative and not necessarily related conclusions from this literature are of interest:

- the earlier the tilt manifests, the larger the expected size of tilt in adulthood (Lubinski et al. 2001; Bernstein, Lubinski, and Benbow 2021)
- pre-collegiate experience in technical (creative) domains is key to later life accumulation of technical (creative) human capital (Wai et al. 2010).

Therefore, coursework policy-induced convergence might be harmful as divergence is quintessential to manifesting talent (Hunt and Wittmann 2008). Reducing the ability tilt by increasing math scores while keeping verbal scores unchanged – and therefore reducing the absolute size of the tilt – might result in what Wang, Eccles, and Kenny (2013) have described as the "paradox of choice". Individuals with similar math and reading scores, especially as found among females, might base their career choices on other factors such

as personal interests and potentially forgo the economic gains from seeking STEM careers. The "non-interventionist" argument likely applies to the far right tail of the ability distribution. An extension of the argument to the bulk of the distribution is dubious though. More of any competencies appears to be the better choice than the tilt-maximizing option for multiple reasons: socioeconomic determinacy of early life school choices (Reardon 2018); supobtimal student decision-making, particularly in earlier grades and in heavy-tracking environments (Checchi and Flabbi 2007; Tamm 2008; Pekkarinen, Uusitalo, and Kerr 2009; Piopiunik 2014); and bearing of pre-collegiate skill accumulations over future career choices (Wai et al. 2010).

The bridge between academic tilt and skill formation literature is human development. In the context of gender-based academic tilt differences, Coyle, Snyder, and Richmond (2015) have hypothesized that "males may show an early math tilt bias, which produces later STEM preferences" (p. 217). Coyle (2018) has made similar comments concerning the link between early ability tilts and later life skill accumulation in analyzing the relation of SAT non-*g* residuals to niche-picking. Early life differences thus possibly reflect in subsequent skill formation by governing the direction and gradient of learning. Likewise, skill formation literature centers around early life differences in skills, albeit focusing on absolute levels of achievements or "human capital".

In the Heckman life-cycle model of skill accumulation (Heckman 1976, 2006), the quantity and quality of human capital at any time – a distinction later added to account for non-cognitive skills (Cunha, Heckman, and Schennach 2010) – is path-dependent upon prior skill stocks through an exponential link function. Therefore, the same amount of educational input reaps the largest yield, the earlier its deployment. A wealth of empirical evidence supports the life-cycle model validating the early opening of achievement gaps before kindergarten entry (Reardon 2018) and as early as the first 18 months of life (Doyle et al. 2017). A staggering figure comes from the Early Childhood Longitudinal Study in which of the 1.25 standard deviations that the achievement gap is worth at the end of grade

8, about 90% (1.15 SD) is instantiated before kindergarten entry.

The following section presents evidence of the "STEM advantage". To date, potential adverse effects from decreased verbal reasoning competence have not been evaluated nor quantified. One must acknowledge the opportunity cost associated with underinvestment in developing verbal reasoning skills in order to flesh out the case behind the "STEM advantage".

### 3.2.2 The "STEM Advantage"

The primary research question addresses math and verbal reasoning development trade-offs. Of all the inputs of educational production, coursework is one of the most understudied due to the lack of cross-sectional variation in coursework policies, the self-selection of high-ability students into advanced courses (Levine and Zimmerman 1995; Joensen and Nielsen 2009), and the low predictive power of credit hours over student outcomes (Altonji 1995). Regarding the relationship between credit hours and student outcomes, Altonji (1995) has famously noted that the effect of one additional year of schooling on earnings is larger than the cumulative effect of individual credit hours. Else known as the "curriculum puzzle", Altonji's is an important reason why "most research on the relationship between education and wages has examined the effects of years of schooling" (p. 410) while overlooking coursework.

Unsurprisingly, early literature has not found significant effects of coursework on student outcomes (Levine and Zimmerman 1995). By contrast, more recent non-experimental (Rose and Betts 2004; Long, Iatarola, and Conger 2009) and quasi-experimental studies (Joensen and Nielsen 2009; Cortes, Goodman, and Nomi 2015; Goodman 2019) point to a more significant relationship. Rose and Betts (2004) and Long, Iatarola, and Conger (2009) have found that advanced high school math classes increase ACT scores and earnings in adulthood. In a follow-up study, Long, Conger, and Iatarola (2012) pinned down a positive impact on math scores and GPA although they could not find any effects on later life out-

comes, be high school graduation or college attendance.

Quasi-experimental studies (Joensen and Nielsen 2009; Cortes, Goodman, and Nomi 2015; Goodman 2019) exploit exogenous changes in coursework requirements to estimate "local treatment effects" from increasing math and science requirements. Joensen and Nielsen (2009) looked at a 1998's Danish pilot program that took out the physics requirements necessary to take advanced math classes. The reform made advanced math coursework accessible to a much broader student population, as post-reform cohorts could enroll without taking the physics classes formerly required. Their difference-in-differences estimate nets an increase in the yearly rate of return for impacted cohorts nearing 20% of their future income. Similarly, Cortes, Goodman, and Nomi (2015) have used a regression discontinuity design to evaluate a Chicago Public School policy administering a double dose of algebra to students falling below a specific test score cutoff. Double-dosed students achieved better GPA and ACT scores while exhibiting greater chances of graduating from high school and attending college than students just above the cutoff. Lastly, Goodman (2019) has exploited the timing of reforms that changed math requirements for cohorts graduating in the years 1982-1994 in 40 US states. Impacted cohorts increased their earnings between 5% and 9%, primarily by majoring in higher-paying college fields. Crucially, the reforms targeted students with little prior training in technical subjects and therefore had a sizeable impact at the extensive margin. That is, the reforms extended access to student populations previously excluded from quantitative training rather than increasing access for already served populations.

### 3.2.3 International Education Production Functions

Human capital production is a form of production generating knowledge through transforming educational inputs. Education production functions (Boardman and Murnane 1979; Hanushek 1979; Todd and Wolpin 2003; Sass, Semykina, and Harris 2014) express knowledge as a function of individual, parental, and school characteristics. Educational

attainments at a point in time are instrumented by a measure of achievement such as test scores and depend on cumulative inputs from starting grade until the current grade.

Because of the complex data generating process of educational production (Sass, Semykina, and Harris 2014), researchers invoke simplifying assumptions to model student achievement. A reduced-form education production function is expressed by:

$$A_{ist} = A\left(I_{it}, P_{it}, S_{st}, \lambda A_{i,s,t-1}, g_i, \epsilon_{ist}\right) \tag{3.1}$$

where $I_{it}$, $P_{it}$, and $S_{st}$ are individual, parental, and school characteristics governing the attainments of student $i$ in school $s$ at time $t$ ($A_{ist}$). Baseline attainments one year prior ($A_{i,s,t-1}$) are discounted at the standard geometric rate of decay $\lambda$ and proxy for the history of educational inputs (Sass, Semykina, and Harris 2014). The coefficient on each characteristic captures the marginal product of the educational factors of production. Hence, coefficients suggest the optimal allocation of educational resources to educational policymakers facing real-world budget constraints.

The unavailability of general ability measures ($g_i$) might cause omitted variables bias and must be addressed by researchers. Also, international standardized testing does not take repeated measurements of student achievement, thus added-value models may not be estimated. Furthermore, lack of prior educational inputs records calls for a further assumption about the invariance of individual and parental inputs over time. What is assumed is that richer parents would have afforded greater access to kindergarten and better K-5 schooling to their children in earlier grades. Thus, much of the variation in test scores attributable to pre-primary and primary education would be stored in the parental variables at time $t$, withal consistently with the life-cycle model of human capital accumulation (Heckman 1976, 2006).

The PISA variables most predictive of student achievement internationally, parental socioeconomic status and sex (Hanushek and Wössmann 2011), channel most of the effects of the

individual-level factors responsible for student achievement, including *g*, conscientiousness, and family factors (Hanushek et al. 2022). Notably, female test-takers outperform male test-takers on the reading section of the PISA test in all 77 countries tested by PISA 2018 while differences in math and science are more erratic. These results stand among the most consequential for comparative educational policy analysis adding to evidence on the persisting math male advantage (Stoet and Geary 2018; Wai, Hodges, and Makel 2018), especially at the far right tail of the distribution (Wai, Hodges, and Makel 2018). The universality of the inverse reading gender gap revealed by PISA is of particular relevance to the study. Of more limited interest is the debate on the plasticity of the math gender gap with respect to a country's degree of gender parity, enlivened by some evidence that points towards convergence (Guiso et al. 2008; Makel et al. 2016).

The lack of school expenditures information in the PISA dataset might not be seen as an issue for estimation (Hanushek and Wössmann 2011; Hanushek and Woessmann 2017). While school expenditures famously divide the field into proponents of the "money doesn't matter" (Hanushek 1989, 1997) and the "money matters" (Card and Krueger 1996; Jackson, Johnson, and Persico 2015) camps, scholars agree that failing schools are typically bound to inefficient resource allocations (Hanushek 2003). Internationally, the inclusion of all the factors of production brings down the predictive power of educational expenditures to marginal (Hanushek and Woessmann 2017).

### 3.2.4 PISA: Testing Functional Knowledge

The data used in this study come from three rounds of the PISA examination administered in 2012, 2015, and 2018. PISA is a low-stake test assessing the functional proficiency in math, reading, and science of 15-years olds independently of grade. The test measures "functional skills" rather than domain knowledge (OECD 2016) and provides an adult-life preparedness checkpoint close to a measure of equality of opportunity (Reardon 2018).

OECD's stated intent to test functional knowledge is one of the bedrocks of PISA's presence

in the public discourse. Although some concerns have been raised about the discriminant validity of the different sections of the PISA test (Pokropek et al. 2022), the clear bearing of *g* on test score achievements (Geary et al. 2017) is mostly a confounder in the context of within-country studies and one mired in the different degrees of participation in secondary education and other differences in educational systems (Rindermann 2007). PISA performance has been shown to correlate with *g* across countries (Rindermann 2007; Burhan et al. 2014; Jones and Potrafke 2014), a result which is likely second order to differences in levels of economic development and dovetails with "the cumulate effects of the social and economic context children encounter as well as the overall quality of the education systems in which they grow and develop" (Pokropek et al. 2022, p. 11). The covariation between national GDP and *g* (Lynn and Vanhanen 2012) as well as generational differences in national IQs (Roivainen 2012) also point to a connection between economic growth and *g*, and so do the path-dependencies in student achievement stressed by the input-dependent model of cognitive development (Heckman 2006).

Overall, the correlation between national IQs and PISA test scores leaves PISA's mission statement to test functional knowledge largely unscathed. In addition to that, PISA tests differently from general ability testing. The test is content-neutral and packaged with domain-specific language. Therefore, students must be familiar with the mathematical terminology to decode math and science tasks. Likewise, verbal tasks require breadth and depth of active and passive vocabulary and grammar structures.

### 3.2.5 Current Study

The study tackles the following research questions:

1. Are there any trade-offs between numerical and verbal skills?
2. Are differences in achievement attributable to differences in coursework?
3. Are differences in achievement attributable to instructional unit requirements?

The analytical model addresses the research questions using *i*) inverse probability weighting; *ii*) Blinder-Oaxaca achievement decomposition into coefficients, endowments, and endowments by coefficients effect (Blinder 1973; Oaxaca 1973); *iii*) Altonji decomposition of achievement into instructional unit requirements (Altonji 1995). For a trade-off to exist between numerical and verbal skills, it must be that the STEM advantage reverts upon conditioning on the inputs of educational production. Suppose there is no section of the PISA test, which the Liberal Arts curriculum adds more to. That would rule out cognitive trade-offs in learning underpinning a dominant STEM education production function.

## 3.3 Methods

### 3.3.1 Data

The analysis uses three public use datafiles from PISA 2012, 2015, and 2018 obtained from Italy's INVALSI. The use of three waves of the PISA test is instrumental to having a sample size large enough for analysis. Prior waves of the test could not be included because tested cohorts were schooled under a regime different from that of law 1/2009, making track identification impossible. Using instructional requirements in the different subjects, 841 observations of Liberal Arts and 1,968 technical students were identified. Among the variables included in the national version of the data files which are not featured in the OECD datasets, are the geographic and education pathways variables.

National test-takers are sampled following a two-stage sampling procedure that collects extensive background data from questionnaires administered to students, teachers, parents, and school principals. Through two-stage sampling, participants are randomly chosen from non-randomly selected schools. The sampling design thus balances representativeness and unbiasedness of selection. National sampling weights are post-processed and added to the dataset by OECD to further increase the representativeness of samples.

At a very high level, PISA scores of students reflect their percentile rank in the international preparedness or adult-life readiness distribution in math, reading, and science. The scores

are ability parameters from a two-parameter Item Response Theory model without a guessing parameter (Birnbaum 1968).[1] Scores are scaled internationally to a normal distribution with a mean of 500 and a standard deviation of 100 achievement points.

The calibration of item-response parameters adds to the substantial labor demands of the two-staged sampling procedure. This is responsible for the lag between test administration and results dissemination. Adding to the lead time, national statistical offices augment the PISA datasets with country-specific variables before release. Italy's National Institute for the Evaluation of the Education System (INVALSI) adds geographic information coding the five Italian macroareas and an indicator coding the different high school pathways.

Table 3.3 reports descriptive statistics. Liberal Arts students are more likely to be female (75.1% to 52.4%), grade repeater (4.4% to 3.0%), and about as likely as STEM students to be immigrants (2.4% to 1.8%). Liberal Arts students come from a more advantaged socioeconomic background, as attested by home book possessions. More than half of Liberal Arts declare to possess more than 200 books at home relative to about one-third of STEM. A similar proportion of students declare home book possessions in the 101-200 books range while STEM are about twice as likely to be represented in the 26-100 and in the 11-25 books categories. Likewise, for each STEM household declaring a Master's degree or higher educational attainments, there are two such Liberal Arts households. Much smaller shares of Liberal Arts students have parents who completed a high school diploma (14.3% to 39.5%) or lower educational level (14.5% to $< 1\%$).

Differences in school institutions are more marginal (see Table 3.4). Schools offering the Liberal Arts and STEM tracks are about equally balanced in the proportion of full-time staff, student-teacher ratio, school size, and ownership status. Schools staff about 15% of part-time teachers, have a student-teacher slightly exceeding 10-to-1, and host an average of 800+ students with huge degrees of variation across individual institutions. Publicly-owned schools are prevalent and private institutions more likely to be found in Liberal Arts

[1]The 2012 version of the test used a one-parameter model (Rasch 1960). Since PISA 2015, OECD transitioned to the two-parameter model.

Table 3.3: Individual and parental characteristics

| | **Liberal Arts** | **STEM** |
|---|---|---|
| Female | 52.4 | 75.1 |
| Grade repeater | 3.0 | 4.4 |
| Immigrant | 1.8 | 2.4 |
| Area | | |
| North | 19.1 | 27.0 |
| Central | 24.9 | 21.9 |
| South & Islands | 56.0 | 51.1 |
| Highest parental education | | |
| Less than high school | 14.5 | $< 1$ |
| High school | 39.5 | 14.4 |
| Bachelor's | 5.7 | 5.9 |
| Master's or higher | 40.3 | 79.1 |
| Books at home | | |
| 0-10 | 4.7 | 1.8 |
| 11-25 | 11.6 | 4.8 |
| 26-100 | 30.3 | 17.8 |
| 101-200 | 22.9 | 24.4 |
| 200+ | 30.5 | 51.2 |

*Notes*. Features of Liberal Arts ($n$ = 841) and STEM ($n$ = 1,968) students from PISA 2012-2018 background surveys. Sample proportions are reported as percentages (%).

than STEM (2.5% to 0.6%).

Students are similarly represented across locations and geographies. The majority of schools is located in large towns hosting anywhere between $15,000$ and $100,000$ inhabitants, followed by suburban ($100,000$-$1,000,000$), small-town ($3,000$-$15,000$), urban ($> 1,000,000$), and village ($< 3,000$) settings. About equal proportions of samples were drawn from the two most affluent areas of the countries, the Northern and Southern regions, as compared to the underdeveloped South and Islands.

### 3.3.2 Analytical Model

To exclude that observable confounders are responsible for observed gaps, difference-of-mean tests were performed using inverse probability weighting. Characteristics of students that predate high school entry were used to create propensity scores for the chances of

Table 3.4: School characteristics

| | **Liberal Arts** | **STEM** |
|---|---|---|
| Location size | | |
| Rural | 0.6% | 1.3% |
| Small town | 25.1% | 13.3% |
| Large town | 48.9% | 53.4% |
| Suburban | 18.1% | 24.3% |
| Urban | 7.4% | 7.7% |
| School ESCS | 0.30 | 0.32 |
| School size | 870.5 | 841.1 |
| Student-teacher ratio | 12.3 | 11.2 |
| Full-time teachers | 86.8% | 84.6% |
| Private | 0.6% | 2.5% |

*Notes*. Features of schools offering Liberal Arts ($n$ = 300) and STEM ($n$ = 317) curricula from PISA 2012-2018 background surveys. Sample proportions are reported as percentages (%).

entering the Liberal Arts track. The propensity score equation for a student attending the Liberal Arts track is given by:

$$p(Liberal\ Arts_i = 1|I_i, P_i) = \sigma(\delta_0 + \delta_1 I_i + \delta_2 P_i + \eta_{is}) \tag{3.2}$$

where $\sigma(\cdot)$ is the logistic function evaluated at the individual and parental characteristics influencing the high school choice of Italian middle school students. Inverse probability weights were derived from the probability of treatment estimated through logistic regression and used to calculate weighted differences of means.[2] A weighted difference-of-means captures the average treatment effect (ATE) from attending the Liberal Arts track relative to attending the technical track.

Next, indication about the causes of the gaps was obtained from achievement decomposition. A regression decomposition à la Blinder-Oaxaca (Blinder 1973; Oaxaca 1973) de-

[2]The features predating high school entry that are included in the model are the same reported in Table Table 3.3: *female*, *grade repeater*, *immigration status*, *books at home*, *region*. The *location* variable, although collected at the school-level, was included as a proxy for place of residence.

composes the estimated achievement gaps into the endowments, coefficients, and endowments by coefficients effects. Three counterfactuals are quantified by the decomposition:

- Would Liberal Arts students have scored as well as technical students had they had the same characteristics?
- Would Liberal Arts students have scored as well as technical students had they had the same returns on characteristics?
- Would Liberal Arts students have scored as well as technical students had they had the same characteristics and returns on characteristics?

The decomposition is conducted into three steps. Firstly, differences of mean characteristics are calculated for each factor of educational production (i.e., the same features included in the regressions in Table B.2, Table B.3, and Table B.4). Secondly, two separate sets of parameters are estimated via regression of PISA test scores on educational production features of STEM and Liberal Arts students. Thirdly, using the estimated coefficients and mean characteristics, three quantities are calculated, the endowments effect, coefficients effect, and endowments by coefficients effect:

$$
\begin{aligned}
Endowments &= \beta_{LibArts}(\overline{X}_{STEM} - \overline{X}_{LibArts}) \\
Coefficients &= (\beta_{STEM} - \beta_{LibArts})\overline{X}_{LibArts} \\
Endowments \times Coefficients &= (\beta_{STEM} - \beta_{LibArts})(\overline{X}_{STEM} - \overline{X}_{LibArts})
\end{aligned}
$$

The vectors of coefficients $\beta_{LibArts}$ and $\beta_{STEM}$ include the parameters from the two regressions while $\overline{X}_{LibArts}$ and $\overline{X}_{STEM}$ are the mean characteristics of students at the two schools. A threefold achievement decomposition defines the difference in predicted mean achievement as the sum of the three effects:

$$\hat{A}_{STEM} - \hat{A}_{LibArts} = Endowments + Coefficients + Endowments \times Coefficients$$

Decomposition methods are particularly helpful to direct policy action and found wide application in the analysis of the education and healthcare sector performance. For example, a school or hospital might be performing below the national average because it operates on low levels of resources. In that case, the Blinder-Oaxaca decomposition would return a significant and positive effect on the endowments component, indicating that redistribution of resources to the school would be performance-enhancing. The contribution of each component is measured on the scale of the output variable and indicates how much convergence is expected under full input equalization. Suppose the achievement gap between two schools is 50 points on the scale of a particular test. In that case, an endowments effect of 50 will project a near-zero gap under complete resource equalization (e.g., the two schools would look the same).

The coefficients effect indicates how much of the achievement gap depends upon the differential rate of return on one unit of resource. In the previous example, a school or hospital might be underperforming despite operating on the national-average level of resources. Coefficients effects are referred to as "production function", in that they capture the production rate of individual inputs. Equalization of coefficients will reduce the achievement gap to zero if the component contributes 50 points to the gap. Highly productive teachers and doctors, that is, teachers and doctors who outproduce the production rate expected from their measurable characteristics, contribute to positive coefficients effects.

The interaction of endowments and characteristics covers simultaneous increases of resources and production rate. While this component is largely irrelevant for this assessment, it informs on nonlinear effects of redistributive policy. A school or hospital might gain more from adding extra resources when its production rate is higher (positive interaction effect); or, marginal gains might taper off at higher production rates (negative interaction

effect). Very productive workers might gain little from more technology; or, they might gain more than less productive workers would.

The larger the endowments effect estimated through Blinder-Oaxaca, the closer the education production function of the Liberal Arts and STEM tracks would be. Conversely, a large coefficients effect would suggest that a larger share of the gap rests within differences in educational inputs.

A third and final research question relates differences in achievement to differences in instructional units. Are differences in achievement attributable to instructional unit requirements? To give this question an answer, the achievement model was re-estimated by specifying the instructional units taken in math, reading, and science. The new model was then compared to the baseline model. If the independent effects of credits in the different classes add up to the total difference in achievement, it is possible to attribute achievement gaps to credit hours requirements. For PISA math score, the target equality is:

$$\hat{PISA}_{STEM} - \hat{PISA}_{LibArts} = (m_{STEM} - m_{LibArts}) \times \beta_{Math}$$

The expression requires that the difference in predicted math achievement between the STEM and the Liberal Arts track is about equal to the differences in instructional units in math $m_{STEM} - m_{LibArts}$ (in minutes) multiplied by the coefficients from the new model with instructional units in math and science. The reading and science achievement gaps were evaluated in a similar fashion.

Robustness checks conclude the discussion of results. Validity was assessed by: *i*) employing evidence from Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) to pin down student learning progression and dynamic human capital; *ii*) discussing qualitative findings from PISA 2012's *Mathematics, Reading, Science, Problem Solving and Financial Literacy* special focus survey (OECD 2013) on math-related beliefs (e.g., intrinsic motivation to learn),

self-beliefs (e.g., self-efficacy), and intentions (e.g., desire to major in the STEM field); *iii*) considering prior data points (Checchi and Zollino 2001), achievement regression residuals (Coyle 2018), and other qualitative arguments.

## 3.4 Results

The section reports results from inverse probability weighting (see Table 3.5 and Figure 3.1, and Table B.1 in the Appendix), Blinder-Oaxaca decomposition (see Figure 3.2), and Altonji decomposition. Empirical findings indicate no reading loss experienced by Liberal Arts students. Furthermore, the cognitive gains afforded by the STEM track could be linearly linked to math requirements (via Altonji decomposition) and shown not to depend on observable characteristics of students or endowments effects (via achievement decomposition). Potential omitted variable bias due to self-selection of more motivated and academically prepared students into the STEM track is addressed in the Robustness Checks section.

### 3.4.1 Matching

Using inverse probability weighting, difference-of-means tests were conducted for PISA test score achievements of Liberal Arts and STEM students (see Table 3.5). Propensity scores were obtained from a logistic regression of the binary treatment (0 = attending STEM, 1 = Liberal Arts) on the individual characteristics of students and their parents. Further conditional differences were estimated regressing PISA test scores on school characteristics and the same student covariates used in the propensity scores calculations and weighting observations by the inverse probability weights. Estimates were negligibly different at standard statistical level of certainty and are therefore omitted from the presentation.

Standardized mean differences before and after adjustment validate the matching strat-

Table 3.5: Mean PISA test scores before and after inverse probability matching

| | | STEM | Liberal Arts | difference | t-stat |
|---|---|---|---|---|---|
| | Unmatched | 552.6 | 508.4 | -44.2*** | -14.4 |
| *Math* | Matched | 553.6 | 503.5 | -50.1*** | -16.2 |
| | Unmatched | 549.5 | 541.1 | -8.4*** | -2.8 |
| *Reading* | Matched | 552.7 | 530.7 | -22.0*** | -7.4 |
| | Unmatched | 544.7 | 519.9 | -24.8*** | -8.0 |
| *Science* | Matched | 546.2 | 513.4 | -32.9*** | -10.6 |

*Notes*. Score averages are calculated from *a*) the observed sample ("Unmatched"); and *b*) the weighted sample after inverse probability matching on observables ("Matched"). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

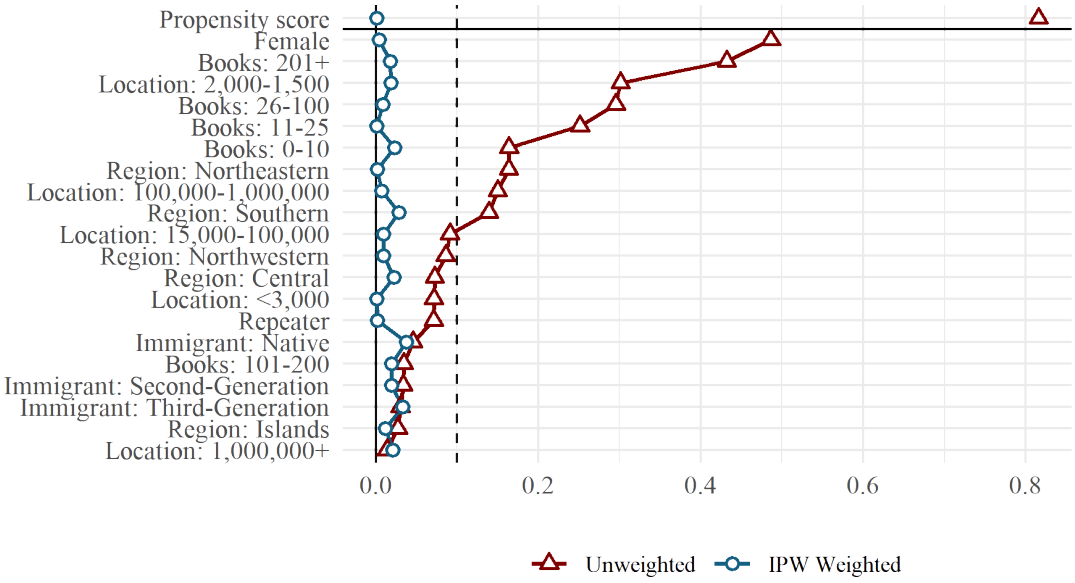


Figure 3.1: Standardized mean differences before and after inverse probability matching

The figure shows substantial bias reduction upon inverse probability weighting. Each red triangle is a standardized mean difference between Liberal Arts and STEM before matching and each blue circle is a difference after matching. The dashed line indicates the significance threshold of a 0.1-difference.

egy (see Figure 3.1 as well as Table B.1 in the Appendix for numeric results).[3] Of the matching variables *female*, *grade repeater*, *immigration status*, *book at home*, *location*, and *region*, all achieved substantial bias reduction upon weighting. The standardized mean

[3]A standardized mean difference is a unit-less measure that normalizes the average difference by a measure of spread. It is calculated as the mean difference in the level of the matching covariate between treated and controls divided by the square root of the mean of the variances, or: $\frac{\overline{X}_{LibArts} - \overline{X}_{STEM}}{\sqrt{(S^2_{LibArts} - S^2_{STEM})/2}}$.

difference in the female predictor dropped down from 0.486 standardized mean difference to near-zero. Comparable bias reduction was achieved for home book possessions bringing down the 0.432 difference in the 200+ books category to a difference of 0.018. None of the variables were unbalanced in the matched data beyond the standard significance threshold of one-tenth of a standardized mean difference.

After matching on educational features, PISApoint gaps in math (50.1) and science (32.9) did not change much relative to the unmatched differences in the same subjects of 44.2 and 24.8. Remarkably, reading differences were more than twice as large in the matched sample (22.0) relative to the unmatched sample (8.4), all differences being significant as per Welch's two-sample t-tests. Results thus fail to support cognitive trade-offs from acquisition of numerical skills. On the contrary, there appear to be positive learning spillovers from the technical curriculum into verbal competence. Furthermore, the size of the spillover appears to be larger when STEM and Liberal Arts students are more evenly matched on their most important observable characteristics. Upon matching, STEM students do even better in the reading domain: thus, the "STEM advantage" would be larger if STEM and Liberal Arts students were more alike.

### 3.4.2 Achievement Decomposition

Results from the threefold achievement decomposition (Figure 3.2) exclude that test score differences between STEM and Liberal Arts depend on endowments effects (i.e., superior educational inputs for STEM). The decompositions use technical students as the reference, therefore:

- The endowments component calculates the sample-mean difference between STEM and Liberal Arts multiplied by the coefficient estimates for Liberal Arts;
- The coefficients component is the product of the endowments of Liberal Arts and the difference in coefficient estimates;

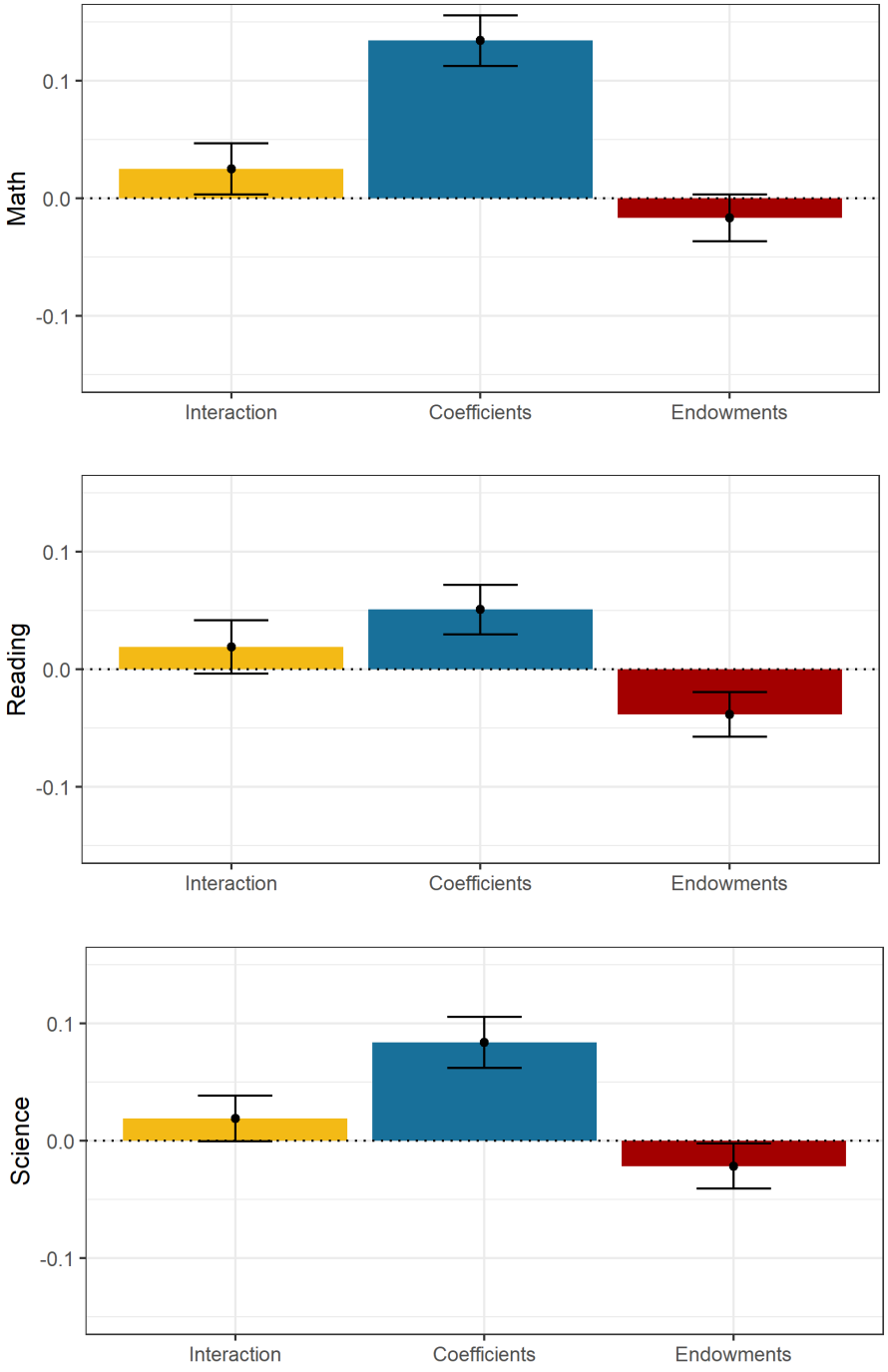


Figure 3.2: Decomposition of the STEM-Liberal Arts PISA score gap

Three-factor decomposition of percentile score differentials: *a*) endowments, or contributions of educational resources (i.e., mean levels of inputs such as sex, parental resources, location, etc.); *b*) coefficients, or unitary contributions of resources as measured by regression coefficients; and *c*) their interactions. Positive (negative) y-axis values positively (inversely) explain the gap.

- The endowments by coefficients component is the product of the two differences.

The decomposition of math scores returns a coefficient effect (41.9) which is about one order of magnitude larger than the endowments effect (4.0). Similarly, the science decomposition indicates that coefficients (25.5) are mainly responsible for differences in PISA attainments, while differences in endowments marginally and negatively predict the gap (-5.7).

Notably, the reading decomposition returns a negative score of -11.7 PISA points on endowments and a positive 14.4 PISApoint effect on coefficients. Therefore, the reading score of STEM students would be larger if they were endowed with the same characteristics as Liberal Arts students and smaller if they shared the same marginal products. The negative reading endowments effect points to negative selection of students into the STEM track while the endowments effect echoes the math and science decompositions. Overall, the STEM advantage would decrease if the STEM production technology was common across the two tracks and it would increase if the characteristics of Liberal Arts students were shared across both tracks.

Achievement decompositions were calculated using coefficient estimates from the two subgroup regressions that use the two subsamples of Liberal Arts and STEM students (see Table B.2, Table B.3, and Table B.4 in the Appendix). Notably, results are largely consistent with prior estimates of international PISA production functions (Hanushek and Wössmann 2011). The predictive model includes the variables *female*, *grade repeater*, *immigration status*, *books at home*, *location*, *region*, *proportion of full-time teachers*, *student-teacher ratio*, *school size*, and *school ownership*. The most important predictor of achievement is parental socioeconomic status (i.e., *books at home*) while school characteristics are less important and often insignificant. Consistently with international literature, female-male differences are negative in math and science and positive in reading. Thus, a gender gap in technical subjects and an inverse gender gap in the verbal domain must be noted.

Models (2) and (3) in the regressions add controls for the other PISA test scores and the

imputed TIMSS and PIRLS scores (more on that in the next section). Notably, coefficient estimates on reading test scores are indistinguishable in the STEM and Liberal Arts math regressions in Table Table B.2 (33.0 PISApoint for a one-SD increase in test scores) and so are the math coefficients in the reading regressions in Table B.3 (32.0). Returns are therefore the same, inconsistently with underlying differences in *g* and differential effects by ability (Coyle et al. 2014). The point will become more apparent with subsequent regression decomposition and validity checks attesting to superior pre-high school human capital stocks among Liberal Arts and lack of patterns in the residuals.

### 3.4.3 Altonji Decomposition

The Altonji decomposition convincingly links the STEM advantage to differences in instructional unit requirements. What was left to explain from the Blinder-Oaxaca decomposition is that the STEM advantage depends on curricular requirements. In order to recommend the STEM curriculum for policy adoption, the attribution must be unmistakable.

The Altonji decompositions neatly attribute differences across test domains to differences in math and physics credits (i.e., 240 minutes) multiplied by their marginal product. The decompositions come from independent re-estimation of the achievement model with the inclusion of instructional unit requirements in math, science, and readings (in minutes) without subsetting the total sample into independent Liberal Arts and STEM samples. Almost all of the 44.2 PISApoint difference in math is explained by curricular differences (43.3), all of the 24.8 science difference (26.2), and the explained reading difference (15.1) is actually larger than the predicted difference of 8.4 points.

Results reveal that, all else being equal, equalization of math coursework would reduce most of the gap in the technical domains, math and science, and revert the reading gap. The evidence provided by the Altonji decomposition complements evidence from Blinder-Oaxaca indicating that at least some of the Liberal Arts requirements could be taken out without trading away much verbal reasoning competence. As suggested by evidence pre-

sented in the next section, the STEM advantage manifests itself notwithstanding ability sorting driving more academically prepared students into the Liberal Arts track.

### 3.5 Robustness Checks

This section addresses potential bias due to unobserved heterogeneity. Although general ability and prior test scores information is not available in the PISA dataset, the internal and external validity of the findings might be affirmed.

It must be noted that elementary and middle school is comprehensive in Italy. Therefore, the prestige of the Liberal Arts track entices students who had access to better instructional resources in earlier comprehensive grades. Older data shows that Liberal Arts students were more likely to graduate with the highest grade of "Outstanding" from middle school and that more than half of Liberal Arts entrants achieved that grade (Checchi, 2001). That led labor economist Ichino (Ichino 2000) to note that "the favorable later life outcomes of Italy's Liberal Arts graduates might solely depend on the positive selection of students with greater motivation and proclivity to learn". This jives with the descriptive statistics presented in the Results section flattering the socioeconomics of Liberal Arts students.

Yet, it might be argued that more motivated students are more likely to enter the STEM track. One immediate counterargument to that is that the STEM track does not require prerequisite knowledge while students entering the Liberal Arts track are often expected to have taken the extracurricular Latin classes offered in middle school. These classes pull middle schoolers out of their main classes and impart the basics of Latin conjugation and vocabulary. A second counterargument comes from the PISA 2012 special focus questionnaire on math (OECD 2016). Students in the Liberal Arts track declared more hours of self-study outside of the classroom (20.3 hrs) relative to STEM students (18.5 hrs). Also, Liberal Arts students appear to be largely driven by intrinsic motivation, which has been linked to performance on standardized tests including PISA (Lee 2020) and TIMSS (Mullis et al. 2020). Conversely, extrinsic or instrumental motivation is the leading motivator for

STEM students who scored more than one-half of a standard deviation higher on the *INST-MOT* index (i.e., self-reported instrumental motivation) constructed by PISA At face value, these statistics uphold greater intrinsic motivation to learn among students sorting into the Liberal Arts track and present no threat to the validity of conclusions.

One additional avenue for self-selection is academic preparedness or readiness. Academic preparedness was imputed using test scores from TIMSS and PIRLS, two "sister tests" assessing the same latent constructs tested by PISA, verbal and numerical ability. Namely, PIRLS tests fourth-graders on verbal competence while TIMSS tests fourth and eighth-graders on numerical competence. The *books at home* and *location* variables are also collected by TIMSS and PIRLS and coded consistently with PISA Thus, tracking the proficiency progression is possible by linking PISA records with its sister tests through exact covariate matching. Average PIRLS and TIMSS scores were calculated for the subgroups defined by the *books at home*, *location*, and *female* variables. Implied numerical scores were obtained for grades 4 and 8 from TIMSS, and implied verbal scores for grade 4 from PIRLS. The imputed verbal and numerical scores were used in two different ways. First, inverse probability weighting and decomposition analyses were re-estimated a second time using imputed scores. Results from this second iteration are within error from the main presentation with no statistical nor substantive departures. The second, visual approach, was to plot the implied proficiency progression for Liberal Arts and STEM students (see Figure 3.3).

The trendlines in Figure 3.3 indicate better selection of students into the Liberal Arts track (in red) as compared to STEM (in blue). Liberal Arts students average out at a higher percentile of implied verbal and numerical scores in comprehensive grades. Namely, they are implied to enter at the 63rd pct. in math in grade 4 and to progress to the 67th pct. in math in grade 8. By contrast, STEM students move down from the 61st pct. in grade 4 to the 57th pct. in grade 8 in math. In reading, Liberal Arts students start off at the 78th pct. in the verbal score distribution at grade 4, much higher than STEM who enter at the 61st

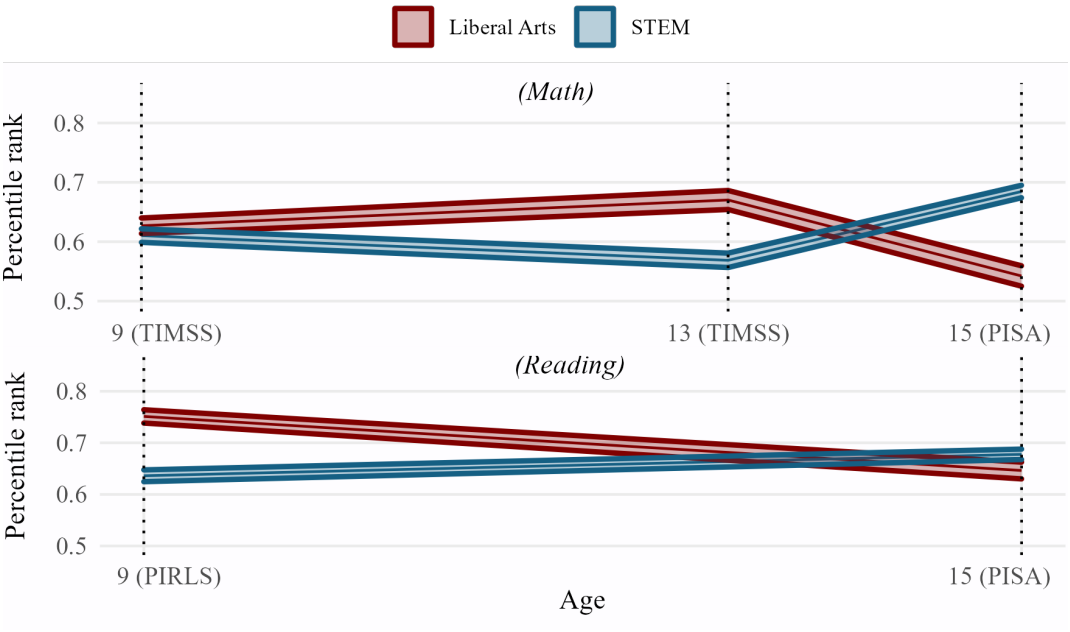


Figure 3.3: Projected math and reading proficiency progression

The graph plots the average projected national percentile rank with 95% confidence interval of STEM and Liberal Arts students based on exact covariate matching with TIMSS (grades 4 and 8) and PIRLS (grade 4) test-takers. Stronger academic readiness among Liberal Arts students is evidenced in either of the two domains.

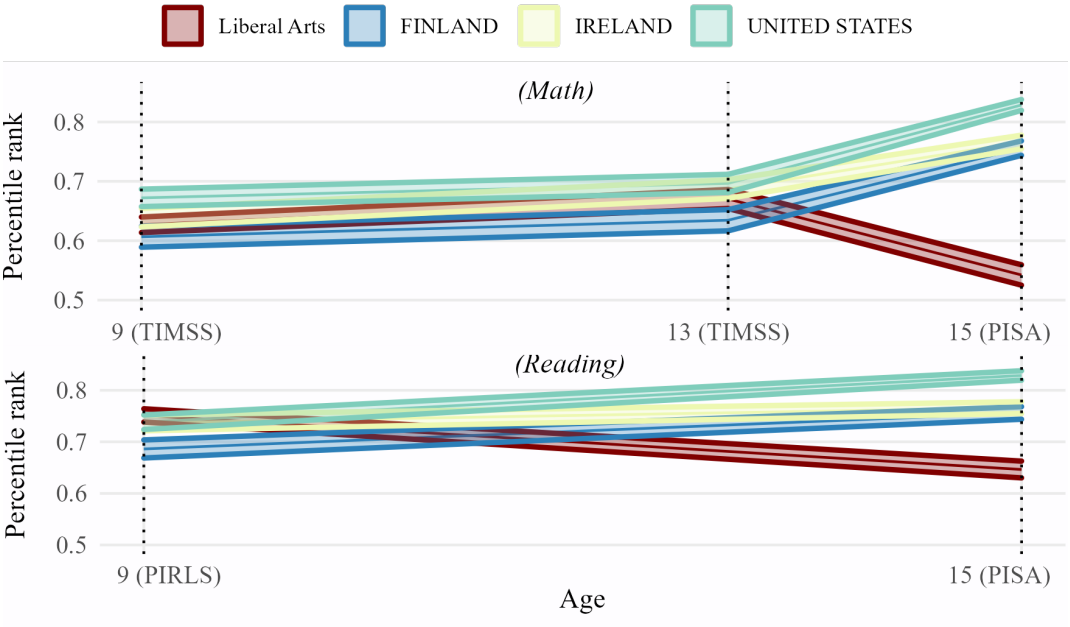


Figure 3.4: Projected math and reading proficiency progression internationally

The graph plots the average projected national percentile rank with 95% confidence interval of Liberal Arts students based on exact covariate matching with TIMSS (grades 4 and 8) and PIRLS (grade 4) test-takers of comparison countries. Academic readiness of Liberal Arts is competitive internationally before fading with PISA.

pct. in PIRLS. Implied scores thus indicate greater academic preparedness in the verbal and numerical domains for Liberal Arts. Their performance drops dramatically after tracking, in comport with negative effect from the Liberal Arts coursework.

One more sanity check involves counterfactual proficiency progression for the US and two other European countries, Finland and Ireland. Countries were chosen based on availability of data for TIMSS and PIRLS. Unfortunately, TIMSS records were incomplete for Germany, the European country whose high school education is the most comparable to Italy's. Implied test scores for students with the same characteristics as Liberal Arts students were calculated using the three variables *books at home*, *location*, and *female*. The interpretation of the plot is straightforward: the Liberal Arts trend is consistent with the international trend before tracking and inconsistent thereafter. Foreign counterfactuals follow a similar progression from grade 4 (67th, 61st, and 64th) through grade 8 in math (70th, 63th, 69th), and start from a similar baseline in reading (74th, 69th, 74th). However, they would have achieved at much higher levels than Italy's Liberal Arts on the math and reading sections of PISA. That is, had students with the same characteristics as Italy's Liberal Arts entered secondary education in the US, Finland, or Ireland, their implied progression would have placed them much higher in PISA.

One last piece of information comes from analysis of the residuals from the three achievement models. If residuals were heavily patterned, some unobservable feature disproportionally found among one set of students might drive student achievement. The correlations between residuals and observed values were similar for STEM and Liberal Arts in either subject, math (0.52 and 0.54), reading (0.59 and 0.55), or science (0.51 and 0.50). All in all, the different quantitative and qualitative arguments dismiss omitted variable bias concerns. If anything, qualitative evidence is consistent with positive self-selection into the Liberal Arts track. Therefore, results might, at worse, be intended as valid and, at best, as conservative estimates of the positive spillover of numerical training into the verbal reasoning domain.

## 3.6 Conclusions

The article investigates trade-offs in cognition in the numerical and verbal domains. The statistical analysis conducted on a sample of Italian students in the two élite tracks of secondary education shows that the Liberal Arts curriculum decreases math, science, and reading achievement on the PISA test relative to expectations based on prior imputed achievement levels.

Results suggest that standard arguments supporting Liberal Arts instruction such as Umberto Eco's might overstate the benefits of a heavy Liberal Arts curriculum. Realistically, a comprehensive high school system with strong technical foundations and few elective classes to observe student choice would build up Italy's STEM workforce and fuel long-term economic growth (Hunt and Wittmann 2008; Coyle et al. 2018). The term technical here may not be intended as stricto sensu "quantitative" and encompasses quantitative and qualitative training which builds socially desirable and economically viable skillsets. In the best-case scenario, Italy might be able to add as much as one-quarter of an international standard deviation to its mean PISA achievement (or about as much as Liberal Arts's math and literacy preparedness advantage is worth at high school entry) by placing its "best and brightest" into the STEM track. The "deadweight loss" of Liberal Arts education might explain the laggard performance of Italy on PISA despite the quality of its primary education, the wide availability of social welfare, and its scientific heritage.

The claim might be tempered by unintended scale effects of extending the STEM curriculum, such as recruitment of a marginal teacher workforce. Furthermore, PISA reading might not completely overlap with a broader definition of critical thinking skills held by advocates for Liberal Arts education. To that, it must be countered that critical thinking skills must crystallize into one or more assessment dimensions to be material. Lastly, students might be selected into tracks based on unobserved academic interest, not academic strength and readiness. The latter concern, however, is partially assuaged by the prevalence

and success of comprehensive schooling worldwide. Tracking might only increase parental wealth's bearing on student achievement (Pekkarinen, Uusitalo, and Kerr 2009), and elective classes might just be enough to differentiate the academic offer.

The evidence provided speaks against learning trade-offs in creating technical knowledge. Standard arguments favoring non-technical education as soul-enriching and fostering critical thinking might be largely hackneyed. Thus, the article contributes to the slowly expanding body of knowledge on coursework policy, under-explored due to the low degree of variation of coursework policies nationally and lacking comparability of educational systems internationally (Levine and Zimmerman 1995; Joensen and Nielsen 2009). Findings also extend policy-relevant knowledge about the non-monetary and monetary gains from technical coursework (Long, Iatarola, and Conger 2009; Goodman 2019), and add one key dimension to the skill formation (Heckman 2006) and academic investment (Cattell 1987; Park, Lubinski, and Benbow 2007; Coyle et al. 2015) bodies of literature.

What remains to be discussed are deeper transmission mechanisms. The minor institutional differences between schools offering the Liberal Arts and STEM curricula are unlikely to be strong candidates. A more likely candidate is the use of verbal reasoning as vehicular to analyzing relatable real-world problems as it happens in the technical track. The indirect teaching of second language skills as a medium of instruction in bilingual education programs (Cummins 2014) might be appropriate proof of concept. Much like language is better learned through contextual applications, indirect nurturing of verbal reasoning skills through technical applications could be the better approach.

At the moment, these hypotheses are speculative. However, concerns about the loss of verbal reasoning skills in technical education are largely unfounded. Mastering a technical coursework increases student cognitive scores in either domain, numerical or verbal, with no measurable loss.

# Appendices

# APPENDIX A

# ESSAY II

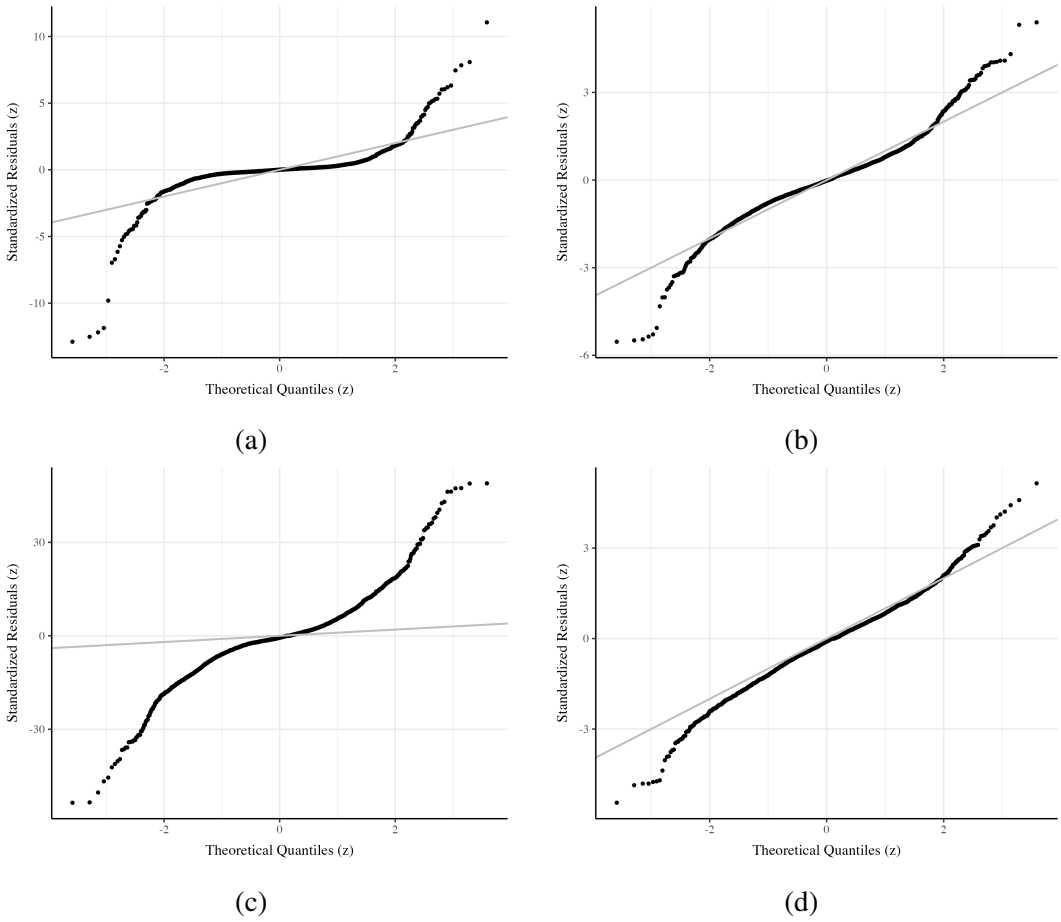


Figure A.1: Residuals versus fitted values plots across model specifications

The figure shows residual versus fitted values plots for four different regression models: a) ordinary least squares, b) log-transformed ordinary least squares, c) Poisson, and d) negative binomial. The 45-degree line indicates an ideal fit.

Table A.1: Selection of model specification using Weight of Evidence test

| **Model** | ***k*** | ***AICc*** | **$\Delta$*AICc*[1]** | ***WoE*[2]** | ***LL*** |
|---|---|---|---|---|---|
| Full - GDP per capita (linear) | 190 | 33142.8 | 0.0 | 99.5 | -16367.9 |
| Full (linear) | 190 | 33157.4 | 14.6 | 0.5 | -16375.2 |
| Full - No GDP (linear) | 190 | 33164.1 | 21.3 | <0.1 | -16378.6 |
| Full (log) | 190 | 33218.5 | 75.7 | <0.1 | -16405.8 |
| Full - GDP per capita (log) | 190 | 33219.4 | 76.6 | <0.1 | -16406.3 |
| Full - No GDP (log) | 190 | 33220.8 | 78.0 | <0.1 | -16406.9 |
| Full - No Import-Export (linear) | 198 | 38039.4 | 4896.6 | <0.1 | -18809.4 |
| Full - No Import-Export (log) | 198 | 38096.3 | 4953.5 | <0.1 | -18837.9 |
| No Covariates - Fixed Effects | 202 | 39374.1 | 6231.3 | <0.1 | -19472.8 |
| No Covariates - Gravity (log) | 5 | 46953.1 | 13810.3 | <0.1 | -23471.5 |
| No Covariates - Gravity (linear) | 5 | 48926.6 | 15783.8 | <0.1 | -24458.3 |

*Note*. The table presents the number of parameters ($k$), Akaike information criterion corrected ($AICc$), delta-AICc ($\Delta_{AIC}$), Weight of Evidence ($WoE$), and Log-likelihood ($LL$) for each model. Quantitative predictors are specified as either linear or logarithmic (in parentheses).

[1] The $\Delta_{AIC}$ measures the distance of each model $m$ to the best model in the set $m^*$ by measure of their $AICc$ values (i.e., $\Delta_m = AIcc_m - AICc_{m^*}$).

[2] The $WoE$ statistic is a scaled measure of relative goodness of fit for each model $m$ calculated as the ratio: $e^{-\frac{\Delta_m}{2}} / \sum_{m=1}^{M} -e^{\frac{\Delta_m}{2}}$.

Table A.2: Country-average office foundations effect from synthetic difference-in-differences

| Country | DAAD foundation | ATT |
|---|---|---|
| Malaysia | 2000 | 3.58 |
| Viet Nam | 2002 | 2.91 |
| Mexico | 2001 | 2.50 |
| Azerbaijan | 2005 | 2.10 |
| Kyrgyzstan | 2005 | 1.93 |
| Colombia | 2006 | 1.92 |
| Thailand | 2000 | 1.75 |
| Iran | 2004 | 1.70 |
| Singapore | 2002 | 1.70 |
| United Arab Emirates | 2007 | 1.67 |
| Pakistan | 2009 | 1.56 |
| Australia | 2002 | 1.46 |
| Afghanistan | 2013 | 1.45 |
| Israel | 2005 | 1.39 |
| Ukraine | 2004 | 1.37 |
| Armenia | 2005 | 1.31 |
| Chile | 2000 | 1.27 |
| Turkey | 2001 | 1.27 |
| Tunisia | 2015 | 1.26 |
| Costa Rica | 2006 | 1.26 |
| South Korea | 2001 | 1.25 |
| Belarus | 2004 | 1.23 |
| Canada | 2003 | 1.21 |
| Italy | 2005 | 1.19 |
| Kazakhstan | 2005 | 1.16 |
| South Africa | 2005 | 1.13 |
| Venezuela | 2002 | 1.10 |
| Spain | 2005 | 1.08 |
| Lebanon | 2016 | 1.05 |
| Greece | 2004 | 1.02 |
| Peru | 2016 | 1.01 |
| Jordan | 2013 | 1.01 |
| Romania | 2003 | 0.99 |
| Argentina | 2000 | 0.97 |
| Georgia | 2005 | 0.94 |
| Ethiopia | 2015 | 0.94 |
| Cuba | 2004 | 0.93 |
| Belgium | 2013 | 0.89 |
| Latvia | 2005 | 0.87 |
| Czech Republic | 2001 | 0.86 |
| Ghana | 2002 | 0.79 |
| Hungary | 2003 | 0.68 |
| Sudan | 2004 | 0.47 |

*Note*. This table presents the synthetic difference-in-differences estimates and their significance for individual office foundations using length of exposure inverse-variance weighting ($^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$). The coefficients reflect sending rate ratios or ratios of the migration case rate with and without the office and are sorted by their magnitude.

## APPENDIX B

## ESSAY III

Table B.1: Standardized mean differences before and after matching

| | **Raw differences** | **Weighted differences** |
|---|---|---|
| Propensity score | 0.816 | 0.001 |
| Female | 0.486 | 0.004 |
| Books: 201+ | 0.432 | 0.018 |
| Location: 3,000-15,000 | 0.295 | 0.008 |
| Books: 26-100 | 0.301 | 0.018 |
| Books: 11-25 | 0.251 | 0.001 |
| Books: 0-10 | 0.164 | 0.024 |
| Region: Northeastern | 0.164 | 0.002 |
| Location: 100,000-1,000,000 | 0.150 | 0.007 |
| Region: Southern | 0.139 | 0.028 |
| Location: 15,000-100,000 | 0.091 | 0.009 |
| Region: Northwestern | 0.086 | 0.010 |
| Region: Central | 0.073 | 0.022 |
| Location: ¡3,000 | 0.072 | 0.001 |
| Repeater | 0.071 | 0.002 |
| Immigrant: Native | 0.046 | 0.037 |
| Books: 101-200 | 0.034 | 0.019 |
| Immigrant: Second-Generation | 0.034 | 0.019 |
| Immigrant: Third-Generation | 0.030 | 0.033 |
| Region: Islands | 0.027 | 0.012 |
| Location: 1,000,000+ | 0.014 | 0.021 |

*Notes*. Bias reduction after inverse probability matching. Matching weights factor the probability of assignment into the tracks ("Weighted differences") to mitigate observational biases in sample means ("Raw differences").

Table B.2: Predictors of PISA math score

| | STEM | | | Liberal Arts | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (1) | (2) | (3) |
| Constant | 582.0*** | 81.0*** | 94.0 | 508.0*** | 90.0*** | 145.0 |
| | (21.0) | (16.0) | (105.0) | (29.0) | (23.0) | (274.0) |
| Female | -26.0*** | -20.0*** | -37.0 | -28.0*** | -22.0*** | -51.0 |
| | (3.1) | (2.0) | (86.0) | (5.4) | (3.7) | (228.0) |
| Grade repeater | -62.0*** | -9.9 | -9.8 | -50.0*** | -16.0 | -16.0 |
| | (8.6) | (6.0) | (6.1) | (12.0) | (8.5) | (8.5) |
| *Immigration status (base: Native)* | | | | | | |
| Second-generation | -11.0 | -7.2 | -7.2 | -39.0 | 15.0 | 16.0 |
| | (16.0) | (8.4) | (8.4) | (33.0) | (20.0) | (20.0) |
| Third-generation | -29.0 | 2.4 | 2.4 | -55.0* | -18.0 | -18.0 |
| | (16.0) | (8.4) | (8.5) | (25.0) | (17.0) | (17.0) |
| *Books at home (base: 0-10)* | | | | | | |
| 11-25 | -12.0 | -10.0* | -4.7 | 0.3 | -12.0 | 55.0 |
| | (9.0) | (5.0) | (108.0) | (21.0) | (13.0) | (282.0) |
| 26-100 | 3.9 | -3.5 | 21.0 | 27.0 | -5.7 | 269.0 |
| | (8.1) | (4.6) | (411.0) | (19.0) | (12.0) | (1,083.0) |
| 101-200 | 19.0* | -1.2 | 33.0 | 53.0** | 5.6 | 490.0 |
| | (8.4) | (4.8) | (692.0) | (19.0) | (12.0) | (1,825.0) |
| 200+ | 22.0** | -1.5 | 40.0 | 64.0*** | 5.8 | 544.0 |
| | (8.2) | (4.7) | (749.0) | (19.0) | (12.0) | (1,980.0) |
| *School random effects* | | | | | | |
| School ESCS | 15.0* | 3.1 | 3.0 | 42.0*** | 3.6 | 3.7 |
| | (6.3) | (3.6) | (3.6) | (6.9) | (4.4) | (4.4) |
| Student-teacher ratio | 2.1*** | 0.6 | 0.6 | 0.8 | -0.8 | -0.8 |
| | (0.6) | (0.4) | (0.4) | (0.9) | (0.6) | (0.6) |
| Number of students | 0.7 | 0.2 | 0.2 | 0.8 | 0.9* | 0.9* |
| | (0.5) | (0.3) | (0.3) | (0.7) | (0.4) | (0.4) |
| Private | -60.0** | -19.0 | -19.0 | -53.0** | -3.8 | -4.4 |
| | (23.0) | (14.0) | (14.0) | (18.0) | (10.0) | (10.0) |
| Full-time ratio | 18.0 | 5.8 | 5.9 | -9.3 | 7.1 | 6.7 |
| | (11.0) | (5.9) | (5.9) | (17.0) | (12.0) | (12.0) |
| *PISA test score (unit: 1 SD)* | | | | | | |
| Reading | | 32.0*** | 32.0*** | | 32.0*** | 32.0*** |
| | | (2.4) | (2.4) | | (3.3) | (3.3) |
| Science | | 56.0*** | 56.0*** | | 50.0*** | 50.0*** |
| | | (2.4) | (2.4) | | (3.3) | (3.3) |
| *Imputed test score (unit: 1 pct.)* | | | | | | |
| TIMSS – Grade 4 | | | -0.4 | | | 0.03 |
| | | | (1.9) | | | (4.8) |
| TIMSS – Grade 8 | | | -0.4 | | | -4.5 |
| | | | (5.3) | | | (14.0) |
| PIRLS | | | 0.2 | | | -2.4 |
| | | | (2.6) | | | (6.9) |
| Region FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Location FE | Yes | Yes | Yes | Yes | Yes | Yes |
| N | 1,968 | 1,968 | 1,968 | 841 | 841 | 841 |
| Adjusted $R^2$ | 0.21 | 0.73 | 0.73 | 0.24 | 0.69 | 0.69 |

*Notes*. Models for PISA math test score with sequential inclusion of regressors: *a*) student features; *b*) PISA test scores on the domains of reading and science; and *c*) imputed TIMSS and PIRLS test scores. T-statistics are reported in parentheses: $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$.

Table B.3: Predictors of PISA reading score

| | **STEM** | | | **Liberal Arts** | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (1) | (2) | (3) |
| Constant | 528.0*** | 70.0*** | 133.0 | 512.0*** | 76.0** | 52.0 |
| | (18.0) | (15.0) | (104.0) | (35.0) | (24.0) | (269.0) |
| Female | 13.0*** | 30.0*** | -23.0 | 14.0* | 34.0*** | 50.0 |
| | (3.0) | (1.9) | (86.0) | (5.5) | (3.3) | (221.0) |
| Grade repeater | -67.0*** | -22.0*** | -22.0*** | -56.0*** | -23.0** | -23.0** |
| | (8.4) | (5.7) | (5.7) | (12.0) | (7.3) | (7.3) |
| *Immigration status (base: Native)* | | | | | | |
| Second-generation | 5.2 | 13.0 | 13.0 | -73.0 | -28.0 | -28.0 |
| | (16.0) | (8.1) | (8.1) | (39.0) | (15.0) | (15.0) |
| Third-generation | -32.0* | -5.6 | -5.4 | -38.0 | 6.2 | 6.2 |
| | (14.0) | (9.5) | (9.6) | (26.0) | (18.0) | (18.0) |
| *Books at home (base: 0-10)* | | | | | | |
| 11-25 | -5.1 | -1.5 | 66.0 | 0.1 | -13.0 | -46.0 |
| | (7.8) | (5.0) | (109.0) | (20.0) | (11.0) | (270.0) |
| 26-100 | 3.4 | -2.9 | 249.0 | 27.0 | -6.9 | -126.0 |
| | (7.1) | (4.6) | (413.0) | (19.0) | (9.5) | (1,044.0) |
| 101-200 | 16.0* | -2.4 | 421.0 | 44.0* | -8.8 | -211.0 |
| | (7.3) | (4.7) | (695.0) | (18.0) | (9.2) | (1,760.0) |
| 200+ | 22.0** | 0.9 | 458.0 | 57.0** | -6.4 | -218.0 |
| | (7.2) | (4.7) | (753.0) | (18.0) | (9.1) | (1,908.0) |
| *School random effects* | | | | | | |
| School ESCS | 2.1 | -12.0** | -12.0** | 40.0*** | 0.2 | 0.2 |
| | (6.1) | (3.9) | (3.9) | (7.4) | (4.6) | (4.7) |
| Student-teacher ratio | 1.5** | -0.1 | -0.1 | 1.6 | 0.3 | 0.2 |
| | (0.6) | (0.4) | (0.4) | (1.1) | (0.7) | (0.7) |
| Number of students | 1.3** | 1.0** | 1.0** | 0.1 | 0.1 | 0.1 |
| | (0.5) | (0.3) | (0.3) | (0.8) | (0.5) | (0.5) |
| Private | -25.0 | 21.0 | 22.0 | -59.0** | -8.6 | -8.6 |
| | (22.0) | (16.0) | (16.0) | (22.0) | (13.0) | (13.0) |
| Full-time ratio | 27.0* | 18.0** | 18.0** | -20.0 | -6.3 | -6.0 |
| | (11.0) | (6.3) | (6.2) | (19.0) | (18.0) | (18.0) |
| *PISA test score (unit: 1 SD)* | | | | | | |
| Math | | 33.0*** | 33.0*** | | 33.0*** | 33.0*** |
| | | (2.4) | (2.4) | | (3.4) | (3.4) |
| Science | | 45.0*** | 45.0*** | | 52.0*** | 52.0*** |
| | | (2.4) | (2.4) | | (3.6) | (3.6) |
| *Imputed test score (unit: 1 pct.)* | | | | | | |
| TIMSS – Grade 4 | | | -1.2 | | | 0.4 |
| | | | (1.9) | | | (4.7) |
| TIMSS – Grade 8 | | | -3.2 | | | 1.3 |
| | | | (5.4) | | | (14.0) |
| PIRLS | | | -1.6 | | | 1.1 |
| | | | (2.7) | | | (6.7) |
| Region FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Location FE | Yes | Yes | Yes | Yes | Yes | Yes |
| N | 1,968 | 1,968 | 1,968 | 841 | 841 | 841 |
| Adjusted $R^2$ | 0.14 | 0.65 | 0.65 | 0.20 | 0.68 | 0.68 |

*Notes*. Models for PISA math test score with sequential inclusion of regressors: *a*) student features; *b*) PISA test scores on the domains of math and science; and *c*) imputed TIMSS and PIRLS test scores. T-statistics are reported in parentheses: $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$.

Table B.4: Predictors of PISA science score

| | STEM | | | Liberal Arts | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (1) | (2) | (3) |
| Constant | 592.0*** | 88.0*** | -101.0 | 512.0*** | 70.0** | 158.0 |
| | (21.0) | (15.0) | (97.0) | (33.0) | (24.0) | (222.0) |
| Female | -18.0*** | -9.5*** | 148.0 | -21.0*** | -15.0*** | -93.0 |
| | (3.1) | (1.9) | (80.0) | (5.3) | (3.4) | (182.0) |
| Grade repeater | -54.0*** | 3.3 | 3.0 | -32.0* | 14.0 | 14.0 |
| | (8.2) | (5.6) | (5.7) | (13.0) | (7.6) | (7.6) |
| *Immigration status (base: Native)* | | | | | | |
| Second-generation | -9.0 | -5.6 | -5.8 | -62.0 | -13.0 | -13.0 |
| | (17.0) | (7.3) | (7.4) | (37.0) | (16.0) | (16.0) |
| Third-generation | -38.0* | -10.0 | -11.0 | -50.0* | -9.4 | -9.3 |
| | (18.0) | (9.2) | (9.2) | (23.0) | (11.0) | (11.0) |
| *Books at home (base: 0-10)* | | | | | | |
| 11-25 | 0.6 | 8.6 | -189.0 | 25.0 | 25.0* | 114.0 |
| | (8.7) | (4.7) | (100.0) | (19.0) | (10.0) | (230.0) |
| 26-100 | 11.0 | 7.9 | -739.0 | 48.0** | 24.0** | 354.0 |
| | (8.0) | (4.4) | (378.0) | (17.0) | (9.1) | (876.0) |
| 101-200 | 26.0** | 11.0* | -1,245.0 | 67.0*** | 25.0** | 570.0 |
| | (8.2) | (4.5) | (637.0) | (17.0) | (9.0) | (1,477.0) |
| 200+ | 30.0*** | 10.0* | -1,349.0 | 81.0*** | 28.0** | 606.0 |
| | (8.1) | (4.4) | (690.0) | (17.0) | (8.9) | (1,600.0) |
| *School random effects* | | | | | | |
| School ESCS | 20.0** | 11.0*** | 12.0*** | 51.0*** | 15.0*** | 15.0*** |
| | (6.2) | (3.4) | (3.3) | (7.1) | (4.0) | (4.0) |
| Student-teacher ratio | 1.9** | 0.2 | 0.2 | 2.2* | 1.1** | 1.1** |
| | (0.6) | (0.3) | (0.3) | (1.0) | (0.4) | (0.4) |
| Number of students | 0.2 | -0.7* | -0.7** | -0.4 | -0.8 | -0.8 |
| | (0.5) | (0.3) | (0.3) | (0.7) | (0.5) | (0.5) |
| Private | -59.0** | -18.0* | -19.0** | -62.0** | -13.0 | -13.0 |
| | (19.0) | (7.8) | (7.4) | (20.0) | (12.0) | (12.0) |
| Full-time ratio | 6.3 | -13.0* | -13.0* | -20.0 | -7.4 | -7.3 |
| | (11.0) | (6.1) | (6.1) | (20.0) | (20.0) | (20.0) |
| *PISA test score (unit: 1 SD)* | | | | | | |
| Math | | 51.0*** | 51.0*** | | 43.0*** | 43.0*** |
| | | (1.8) | (1.8) | | (3.0) | (3.0) |
| Reading | | 39.0*** | 39.0*** | | 44.0*** | 44.0*** |
| | | (1.9) | (1.9) | | (2.8) | (2.8) |
| *Imputed test score (unit: 1 pct.)* | | | | | | |
| TIMSS – Grade 4 | | | 3.4* | | | -2.0 |
| | | | (1.7) | | | (3.9) |
| TIMSS – Grade 8 | | | 9.6* | | | -3.6 |
| | | | (4.9) | | | (11.0) |
| PIRLS | | | 4.7 | | | -2.0 |
| | | | (2.4) | | | (5.7) |
| Region FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Location FE | Yes | Yes | Yes | Yes | Yes | Yes |
| N | 1,968 | 1,968 | 1,968 | 841 | 841 | 841 |
| Adjusted $R^2$ | 0.19 | 0.74 | 0.74 | 0.25 | 0.74 | 0.74 |

*Notes*. Models for PISA science test score with sequential inclusion of regressors: *a*) student features; *b*) PISA test scores on the domains of math and reading; and *c*) imputed TIMSS and PIRLS test scores. T-statistics are reported in parentheses: $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$.

## REFERENCES


Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge. 2022. "When Should You Adjust Standard Errors for Clustering?" *Quarterly Journal of Economics.*

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490): 493–505.

———. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59 (2): 495–510.

Achter, John A, David Lubinski, Camilla P Benbow, and Hossain Eftekhari-Sanjani. 1999. "Assessing Vocational Preferences Among Gifted Adolescents Adds Incremental Validity to Abilities: A Discriminant Analysis of Educational Outcomes Over a 10-year Interval." *Journal of Educational Psychology* 91 (4): 777.

Altonji, Joseph. 1995. "The Effects of High School Curriculum on Education and Labor Market Outcomes." *Journal of Human Resources* 30 (3): 409–438.

Anderson, Chris. 2006. *The Long Tail: Why the Future of Business is Selling Less of More.* First. London, UK: Hachette.

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2009. "Incentives and Services for College Achievement: Evidence from a Randomized Trial." *American Economic Journal: Applied Economics* 1 (1): 136–63.

Arcidiacono, Peter, Esteban M Aucejo, and Ken Spenner. 2012. "What Happens After Enrollment? An Analysis of the Time Path of Racial Differences in GPA and Major Choice." *IZA Journal of Labor Economics* 1 (1): 1–24.

Arellano, Manuel, et al. 1987. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics* 49 (4): 431–434.

Arkhangelsky, Dmitry, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. 2021. "Synthetic Difference-in-Differences." *American Economic Review* 111 (12): 4088–4118.

Arnold, Kimberly E, and Matthew D Pistilli. 2012. "Course Signals at Purdue: Using Learning Analytics to Increase Student Success." In *Proceedings of the 2nd international conference on learning analytics and knowledge,* 267–270. Vancouver, Canada: Association for Computing Machinery, April.

Baker, Andrew C, David F Larcker, and Charles CY Wang. 2022. "How Much Should We Trust Staggered Difference-in-Differences Estimates?" *Journal of Financial Economics* 144 (2): 370–395.

Bartel, Ann P. 1979. "The Migration Decision: What Role Does Job Mobility Play?" *The American Economic Review* 69 (5): 775–786.

Becker, David, Thomas R Coyle, Tyler L Minnigh, and Heiner Rindermann. 2022. "International Differences in Math and Science Tilts: The Stability, Geography, and Predictive Power of Tilt for Economic Criteria." *Intelligence* 92:101646.

Becker, Gary S. 1962. "Investment in Human Capital: A Theoretical Analysis." *Journal of Political Economy* 70 (5, Part 2): 9–49.

Beine, Michel, Romain Noël, and Lionel Ragot. 2014. "Determinants of the International Mobility of Students." *Economics of Education Review* 41:40–54.

Bernstein, Brian O, David Lubinski, and Camilla P Benbow. 2021. "Academic Acceleration in Gifted Youth and Fruitless Concerns Regarding Psychological Well-Being: A 35-year Longitudinal Study." *Journal of Educational Psychology* 113 (4): 830.

Bessey, Donata. 2012. "International Student Migration to Germany." *Empirical Economics* 42 (1): 345–361.

Bettinger, Eric P, and Rachel B Baker. 2014. "The Effects of Student Coaching: An Evaluation of a Randomized Experiment in Student Advising." *Educational Evaluation and Policy Analysis* 36 (1): 3–19.

Birnbaum, A Lord. 1968. "Some Latent Trait Models and their Use in Inferring an Examinee's Ability." In *Statistical Theories of Mental Test Scores,* 395–479. Reading, MA: Addison-Wesley.

Blinder, Alan S. 1973. "Wage Discrimination: Reduced form and Structural Estimates." *Journal of Human Resources,* 436–455.

Boardman, Anthony E, and Richard J Murnane. 1979. "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement." *Sociology of Education,* 113–121.

Breusch, Trevor S. 1978. "Testing for Autocorrelation in Dynamic Linear Models." *Australian Economic Papers* 17 (31): 334–355.

Breusch, Trevor S, and Adrian R Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica: Journal of the Econometric Society* 47 (5): 1287–1294.

Brookhart, Susan M, Thomas R Guskey, Alex J Bowers, James H McMillan, Jeffrey K Smith, Lisa F Smith, Michael T Stevens, and Megan E Welsh. 2016. "A Century of Grading Research: Meaning and Value in the Most Common Educational Measure." *Review of Educational Research* 86 (4): 803–848.

Burhan, Nik Ahmad Sufian, Mohd Rosli Mohamad, Yohan Kurniawan, and Abdul Halim Sidek. 2014. "The Impact of Low, Average, and High IQ on Economic Growth and Technological Progress: Do All Individuals Contribute Equally?" *Intelligence* 46:1–8.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2): 200–230.

Card, David, and Alan B Krueger. 1996. "School Resources and Student Outcomes: an Overview of the Literature and New Evidence from North and South Carolina." *Journal of Economic Perspectives* 10 (4): 31–50.

Carlson, Sören. 2013. "Becoming a Mobile Student – A Processual Perspective on German Degree Student Mobility." *Population, Space and Place* 19 (2): 168–180.

Cattell, Raymond Bernard. 1987. *Intelligence: Its Structure, Growth and Action.* Amsterdam, The Netherlands: North-Holland.

Checchi, Daniele, and Luca Flabbi. 2007. *Intergenerational Mobility and Schooling Decisions in Germany and Italy: The Impact of Secondary School Tracks.* Technical report.

Checchi, Daniele, and Francesco Zollino. 2001. "Struttura del Sistema Scolastico e Selezione Sociale." *Rivista di Politica Economica* 8:43–84.

Choi, Samuel PM, Sze Sing Lam, Kam Cheong Li, and Billy TM Wong. 2018. "Learning Analytics at Low Cost: At-Risk Student Prediction with Clicker Data and Systematic Proactive Interventions." *Journal of Educational Technology & Society* 21 (2): 273–290.

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. "The Echo Chamber Effect on Social Media." *Proceedings of the National Academy of Sciences* 118 (9): e2023301118.

Cortes, Kalena E, Joshua S Goodman, and Takako Nomi. 2015. "Intensive Math Instruction and Educational Attainment Long-Run Impacts of Double-Dose Algebra." *Journal of Human Resources* 50 (1): 108–158.

Coyle, Thomas R. 2018. "Non-*g* Residuals of Group Factors Predict Ability Tilt, College Majors, and Jobs: A Non-*g* Nexus." *Intelligence* 67:19–25.

Coyle, Thomas R, and David R Pillow. 2008. "SAT and ACT Predict College GPA After Removing g." *Intelligence* 36 (6): 719–729.

Coyle, Thomas R, Jason M Purcell, Anissa C Snyder, and Miranda C Richmond. 2014. "Ability Tilt on the SAT and ACT Predicts Specific Abilities and College Majors." *Intelligence* 46:18–24.

Coyle, Thomas R, Heiner Rindermann, Dale Hancock, and Jacob Freeman. 2018. "Nonlinear Effects of Cognitive Ability on Economic Productivity: A Country-Level Analysis." *Journal of Individual Differences* 39 (1): 39.

Coyle, Thomas R, Anissa C Snyder, and Miranda C Richmond. 2015. "Sex Differences in Ability Tilt: Support for Investment Theory." *Intelligence* 50:209–220.

Coyle, Thomas R, Anissa C Snyder, Miranda C Richmond, and Michelle Little. 2015. "SAT Non-*g* Residuals Predict Course Specific GPAs: Support for Investment Theory." *Intelligence* 51:57–66.

Cummins, Jim. 2014. "Rethinking Pedagogical Assumptions in Canadian French Immersion Programs." *Journal of Immersion and Content-Based Language Education* 2 (1): 3–22.

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97 (2): 31–47.

Cunha, Flavio, James J Heckman, and Susanne M Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931.

D'Haultfœuille, Xavier, Stefan Hoderlein, and Yuya Sasaki. 2022. "Nonparametric Difference-in-Differences in Repeated Cross-Sections with Continuous Treatments." *Journal of Econometrics.*

Dobronyi, Christopher R, Philip Oreopoulos, and Uros Petronijevic. 2019. "Goal Setting, Academic Reminders, and College Success: A Large-Scale Field Experiment." *Journal of Research on Educational Effectiveness* 12 (1): 38–66.

Dodge, Bernie, John Whitmer, and James P Frazee. 2015. "Improving undergraduate student achievement in large blended courses through data-driven interventions." In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge,* 412–413. Poughkeepsie, NY: Association for Computing Machinery, March.

Doyle, Orla, Colm Harmon, James J Heckman, Caitriona Logue, and Seong Hyeok Moon. 2017. "Early Skill Formation and the Efficiency of Parental Investment: A Randomized Controlled Trial of Home Visiting." *Labour Economics* 45:40–58.

Dweck, Carol S. 2013. *Self-Theories: Their Role in Motivation, Personality, and Development.* New York, NY: Psychology Press.

Geary, David C, Alan Nicholas, Yaoran Li, and Jianguo Sun. 2017. "Developmental Change in the Influence of Domain-General Abilities and Domain-Specific Knowledge on Mathematics Achievement: An Eight-Year Longitudinal Study." *Journal of Educational Psychology* 109 (5): 680.

Godfrey, Leslie G. 1978. "Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables." *Econometrica: Journal of the Econometric Society* 46 (6): 1293–1301.

Goldstein, Philip J, and Richard N Katz. 2005. *Academic Analytics: The Uses of Management Information and Technology in Higher Education.* Vol. 8. 1. EDUCAUSE.

Goodman, Joshua. 2019. "The Labor of Division: Returns to Compulsory High School Math Coursework." *Journal of Labor Economics* 37 (4): 1141–1182.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2): 254–277.

GSU. 2019. *2019 Complete College Georgia Report – Georgia State University.* Technical report. Atlanta, GA: Georgia State University.

Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales. 2008. "Culture, Gender, and Math." *Science* 320 (5880): 1164.

Hansen, Ben B. 2004. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association* 99 (467): 609–618.

Hanushek, Eric A. 1979. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions." *Journal of Human Resources,* 351–388.

———. 1989. "The Impact of Differential Expenditures on School Performance." *Educational Researcher* 18 (4): 45–62.

———. 1997. "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19 (2): 141–164.

———. 2003. "The Failure of Input-Based Schooling Policies." *The Economic Journal* 113 (485): F64–F98.

Hanushek, Eric A, Lavinia Kinne, Philipp Lergetporer, and Ludger Woessmann. 2022. "Patience, Risk-taking, and Human Capital Investment across Countries." *The Economic Journal* 132 (646): 2290–2307.

Hanushek, Eric A, and Ludger Woessmann. 2017. "School Resources and Student Achievement: A Review of Cross-Country Economic Research." *Cognitive Abilities and Educational Outcomes,* 149–171.

Hanushek, Eric A, and Ludger Wössmann. 2011. "The Economics of International Differences in Educational Achievement." In *Handbook of the Economics of Education,* 3:89–200. Elsevier.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica: Journal of the Econometric Society* 46 (6): 1251–1271.

Heckman, James J. 1976. "A Life-Cycle Model of Earnings, Learning, and Consumption." *Journal of Political Economy* 84 (4, Part 2): S9–S44.

———. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science* 312 (5782): 1900–1902.

Henry, Gary T, and Ross Rubenstein. 2002. "Paying for Grades: Impact of Merit-Based Financial Aid on Educational Quality." *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* 21 (1): 93–109.

Hilbe, Joseph M. 2014. *Modeling Count Data.* Cambridge, UK: Cambridge University Press.

Hlosta, Martin, Zdenek Zdrahal, and Jaroslav Zendulka. 2017. "Ouroboros: Early Identification of At-Risk Students without Models Based on Legacy Data." In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference,* 6–15. Vancouver, Canada: Association for Computing Machinery, March.

Hunt, Earl, and Werner Wittmann. 2008. "National Intelligence and National Prosperity." *Intelligence* 36 (1): 1–9.

Ichino, Andrea. 2000. "L'Istruzione e il futuro: Il liceo "à la carte" Meglio del Menu Fisso [The Future of Secondary Education: The "à La Carte Menu" Beats the Set Menu]." *Il Sole 24 Ore,* March.

Jackson, C Kirabo, Rucker C Johnson, and Claudia Persico. 2015. "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms." *The Quarterly Journal of Economics* 131 (1): 157–218.

Jayaprakash, Sandeep M, Erik W Moody, Eitel JM Lauría, James R Regan, and Joshua D Baron. 2014. "Early Alert of Academically at-Risk Students: An Open Source Analytics Initiative." *Journal of Learning Analytics* 1 (1): 6–47.

Jena, Farai, and Barry Reilly. 2013. "The Determinants of United Kingdom Student Visa Demand from Developing Countries." *IZA Journal of Labor & Development* 2 (1): 1–22.

Joensen, Juanna Schrøter, and Helena Skyt Nielsen. 2009. "Is There a Causal Effect of High School Math on Labor Market Outcomes?" *Journal of Human Resources* 44 (1): 171–198.

Jones, Garett, and Niklas Potrafke. 2014. "Human Capital and National Institutional Quality: Are TIMSS, PISA, and National Average IQ Robust Predictors?" *Intelligence* 46:148–155.

Jones, Todd R, Daniel Kreisman, Ross Rubenstein, Cynthia Searcy, and Rachana Bhatt. 2022. "The Effects of Financial Aid Loss on Persistence and Graduation: A Multi-Dimensional Regression Discontinuity Approach." *Education Finance and Policy* 17 (2): 206–231.

Jovanovic, Jelena, Negin Mirriahi, Dragan Gašević, Shane Dawson, and Abelardo Pardo. 2019. "Predictive Power of Regularity of Pre-Class Activities in a Flipped Classroom." *Computers & Education* 134:156–168.

Karemera, David, Victor Iwuagwu Oguledo, and Bobby Davis. 2000. "A Gravity Model Analysis of International Migration to North America." *Applied Economics* 32 (13): 1745–1755.

Le, Huy, Steven B Robbins, and Paul Westrick. 2014. "Predicting Student Enrollment and Persistence in College STEM Fields Using an Expanded PE Fit Framework: A Large-Scale Multilevel Study." *Journal of Applied Psychology* 99 (5): 915.

Lee, Jihyun. 2020. *Non-Cognitive Characteristics and Academic Achievement in Southeast Asian Countries Based on PISA 2009, 2012, and 2015.*, OECD Education Working Papers 850. Paris: Secretary-General of the OECD: Organisation for Economic Co-operation and Development.

Lent, Robert W, Steven D Brown, and Gail Hackett. 1994. "Toward a Unifying Social Cognitive Theory of Career and Academic Interest, Choice, and Performance." *Journal of Vocational Behavior* 45 (1): 79–122.

Levine, Phillip B, and David J Zimmerman. 1995. "The Benefit of Additional High-School Math and Science Classes for Young Men and Women." *Journal of Business & Economic Statistics* 13 (2): 137–149.

Lizzio, Alf, and Keithia Wilson. 2013. "Early Intervention to Support the Academic Recovery of First-Year Students at Risk of Non-Continuation." *Innovations in Education and Teaching International* 50 (2): 109–120.

Long, Mark C, Dylan Conger, and Patrice Iatarola. 2012. "Effects of High School Course-Taking on Secondary and Postsecondary Success." *American Educational Research Journal* 49 (2): 285–322.

Long, Mark C, Patrice Iatarola, and Dylan Conger. 2009. "Explaining Gaps in Readiness for College-Level Math: The Role of High School Courses." *Education Finance and Policy* 4 (1): 1–33.

Lubinski, David, Rose Mary Webb, Martha J Morelock, and Camilla Persson Benbow. 2001. "Top 1 in 10,000: A 10-year Follow-Up of the Profoundly Gifted." *Journal of Applied Psychology* 86 (4): 718.

Lynn, Richard, and Tatu Vanhanen. 2012. "National IQs: A Review of their Educational, Cognitive, Economic, Political, Demographic, Sociological, Epidemiological, Geographic and Climatic Correlates." *Intelligence* 40 (2): 226–234.

Macfadyen, Leah P, and Shane Dawson. 2010. "Mining LMS Data to Develop an "Early Warning System" for Educators: A Proof of Concept." *Computers & Education* 54 (2): 588–599.

Makel, Matthew C, Jonathan Wai, Kristen Peairs, and Martha Putallaz. 2016. "Sex Differences in the Right Tail of Cognitive Abilities: An Update and Cross Cultural Extension." *Intelligence* 59:8–15.

McMahon, Mary E. 1992. "Higher Education in a World Market." *Higher education* 24 (4): 465–482.

Mincer, Jacob. 1958. "Investment in human capital and personal income distribution." *Journal of political economy* 66 (4): 281–302.

Mullis, Ina VS, Michael O Martin, Pierre Foy, Dana L Kelly, and Bethany Fishbein. 2020. *TIMSS 2019 International Results in Mathematics and Science.* Boston, MA: Boston College, TIMSS & PIRLS International Study Center.

Naidoo, Vikash. 2007. "Research on the Flow of International Students to UK Universities: Determinants and Implications." *Journal of Research in International Education* 6 (3): 287–307.

Oaxaca, Ronald. 1973. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review,* 693–709.

OECD. 2013. *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy.* Paris: Secretary-General of the OECD: Organisation for Economic Co-operation and Development.

OECD. 2016. *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy.* Paris: Secretary-General of the OECD: Organisation for Economic Co-operation and Development.

———. 2021. *Education at a Glance 2021.* Paris: Secretary-General of the OECD: Organisation for Economic Co-operation and Development.

Oreopoulos, Philip, Robert S Brown, and Adam M Lavecchia. 2017. "Pathways to Education: An Integrated Approach to Helping at-Risk High School Students." *Journal of Political Economy* 125 (4): 947–984.

Oreopoulos, Philip, and Uros Petronijevic. 2018. "Student Coaching: How Far can Technology Go?" *Journal of Human Resources* 53 (2): 299–329.

Page, Lindsay C, Stacy S Kehoe, Benjamin L Castleman, and Gumilang Aryo Sahadewo. 2019. "More than Dollars for Scholars the Impact of the Dell Scholars Program on College Access, Persistence, and Degree Attainment." *Journal of Human Resources* 54 (3): 683–725.

Papamitsiou, Zacharoula, and Anastasios A Economides. 2014. "Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence." *Journal of Educational Technology & Society* 17 (4): 49–64.

Park, Gregory, David Lubinski, and Camilla P Benbow. 2007. "Contrasting Intellectual Patterns Predict Creativity in the Arts and Sciences: Tracking Intellectually Precocious Youth Over 25 Years." *Psychological Science* 18 (11): 948–952.

Pattison, Evangeleen, Eric Grodsky, and Chandra Muller. 2013. "Is the Sky Falling? Grade Inflation and the Signaling Power of Grades." *Educational Researcher* 42 (5): 259–265.

Pekkarinen, Tuomas, Roope Uusitalo, and Sari Kerr. 2009. "School Tracking and Intergenerational Income Mobility: Evidence from the Finnish Comprehensive School Reform." *Journal of Public Economics* 93 (7-8): 965–973.

Piopiunik, Marc. 2014. "The Effects of Early Tracking on Student Performance: Evidence from a School Reform in Bavaria." *Economics of Education Review* 42:12–33.

Pokropek, Artur, Gary N Marks, Francesca Borgonovi, Piotr Koc, and Samuel Greiff. 2022. "General or Specific Abilities? Evidence from 33 Countries Participating in the PISA Assessments." *Intelligence* 92:101653.

Porter, Stephen R, and Randy L Swing. 2006. "Understanding how First-Year Seminars Affect Persistence." *Research in Higher Education* 47 (1): 89–109.

Rasch, Georg. 1960. *Probabilistic models for Some Intelligence and Attainment Tests.* Copenhagen, Denmark: Danish Institute of Educational Research.

Reardon, Sean F. 2018. "The Widening Academic Achievement Gap between the Rich and the Poor." In *Inequality in the 21st Century,* edited by David B Grusky and Jasmine Hill, 177–189. New York, NY: Routledge.

Rindermann, Heiner. 2007. "The *g*-Factor of International Cognitive Ability Comparisons: The Homogeneity of Results in PISA, TIMSS, PIRLS and IQ-Tests across Nations." *European Journal of Personality* 21 (5): 667–706.

Roivainen, Eka. 2012. "Economic, Educational, and IQ Gains in Eastern Germany 1990–2006." *Intelligence* 40 (6): 571–575.

Rose, Heather, and Julian R Betts. 2004. "The Effect of High School Courses on Earnings." *Review of Economics and Statistics* 86 (2): 497–513.

Rosenbaum, Paul R. 1989. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association* 84 (408): 1024–1032.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rosenzweig, Mark R. 2006. "Global Wage Differences and International Student Flows." In *Brookings Trade Forum: Global Labor Markets,* edited by Susan Collins and Carol Graham, 57–86. Washington, DC: Brookings Institution Press.

Rossman, Daniel, Rayane Alamuddin, Martin Kurzweil, and Julia Karo. 2021. *MAAPS Advising Experiment: Evaluation Findings after Four Years.* New York, NY: Ithaka S+R.

Sass, Tim R, Anastasia Semykina, and Douglas N Harris. 2014. "Value-Added Models and the Measurement of Teacher Productivity." *Economics of Education Review* 38:9–23.

Schiavazzi, Vera. 2014. "Il liceo classico? Assolviamolo, ma va riformato [Liberal Arts Education? We shall Forgive, but we Must Reform]." *La Repubblica,* November.

Schultz, Theodore W. 1961. "Investment in Human Capital." *The American Economic Review* 51 (1): 1–17.

Schunk, Dale H. 1991. "Self-Efficacy and Academic Motivation." *Educational Psychologist* 26 (3-4): 207–231.

Siemens, George. 2013. "Learning Analytics: The Emergence of a Discipline." *American Behavioral Scientist* 57 (10): 1380–1400.

Silva, JMC Santos, and Silvana Tenreyro. 2006. "The Log of Gravity." *The Review of Economics and statistics* 88 (4): 641–658.

Singer, Natasha. 2021. "Learning Apps Have Boomed in the Pandemic. Now Comes the Real Test." *The New York Times* (March 17, 2021).

Sjaastad, Larry A. 1962. "The Costs and Returns of Human Migration." *Journal of Political Economy* 70 (5, Part 2): 80–93.

Soo, Kwok Tong, and Caroline Elliott. 2010. "Does Price Matter? Overseas Students in UK Higher Education." *Economics of Education Review* 29 (4): 553–565.

Stange, Kevin M. 2012. "An Empirical Investigation of the Option Value of College Enrollment." *American Economic Journal: Applied Economics* 4 (1): 49–84.

Stoet, Gijsbert, and David C Geary. 2018. "The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education." *Psychological Science* 29 (4): 581–593.

Tamm, Marcus. 2008. "Does Money Buy Higher Schooling?: Evidence from Secondary School Track Choice in Germany." *Economics of Education Review* 27 (5): 536–545.

Tinbergen, Jan. 1962. "An Analysis of World Trade Flows." In *Shaping the World Economy: Suggestions for an International Economic Policy,* edited by David B Grusky and Jasmine Hill, 262–293. New York, NY: Twentieth Century Fund.

Tinto, Vincent. 1987. *Leaving College: Rethinking the Causes and Cures of Student Attrition.* Second. Chicago, IL: University of Chicago Press.

Todd, Petra E, and Kenneth I Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal* 113 (485): F3–F33.

UNESCO. 2022. *UNESCO Institute for Statistics (UIS).*

Vrontis, Demetris, Alkis Thrassou, and Yioula Melanthiou. 2007. "A Contemporary Higher Education Student-Choice Model for Developed Countries." *Journal of Business Research* 60 (9): 979–989.

Wai, Jonathan, Jaret Hodges, and Matthew C Makel. 2018. "Sex Differences in Ability Tilt in the Right Tail of Cognitive Abilities: A 35-year Examination." *Intelligence* 67:76–83.

Wai, Jonathan, David Lubinski, Camilla P Benbow, and James H Steiger. 2010. "Accomplishment in Science, Technology, Engineering, and Mathematics (STEM) and its Re-

lation to STEM Educational Dose: A 25-Year Longitudinal Study." *Journal of Educational Psychology* 102 (4): 860.

Wang, Ming-Te, Jacquelynne S Eccles, and Sarah Kenny. 2013. "Not Lack of Ability but More Choice: Individual and Gender Differences in Choice of Careers in Science, Technology, Engineering, and Mathematics." *Psychological Science* 24 (5): 770–775.

Wolpert, Julian. 1965. "Behavioral Aspects of the Decision to Migrate." *Papers of the Regional Science Association* 15 (1): 159–169.

Zheng, Ping. 2014. "Antecedents to International Student Inflows to UK Higher Education: A Comparative Analysis." *Journal of Business Research* 67 (2): 136–143.

Zimmerman, Barry J. 2000. "Attaining Self-Regulation: A Social Cognitive Perspective." In *Handbook of Self-Regulation,* 13–39. Elsevier.