Essays in Higher Education

by

Aizat Nurshatayeva

BA Applied Linguistics,
Kazakh University of International Relations and World Languages, 2004

MA Linguistics, Kazakh University of International Relations and World Languages, 2008

Submitted to the Graduate Faculty of the

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Aizat Nurshatayeva

It was defended on

February 28, 2020

and approved by

Osea Giuntella, Assistant Professor, Department of Economics

Sean Kelly, Professor, Department of Administrative and Policy Studies

Maureen McClure, Associate Professor, Department of Administrative and Policy Studies

Francis Alvin Pearman, Assistant Professor, Graduate School of Education, Stanford University

Thesis Advisor/Dissertation Director: Lindsay C. Page, Associate Professor, Department of Psychology in Education

Essays in Higher Education

Aizat Nurshatayeva, PhD

University of Pittsburgh, 2020

This doctoral dissertation consists of three independent studies of higher education. The first study uses a natural experiment at a selective university in Central Asia and a difference-in-differences strategy to estimate the causal effect of switching to English-only instruction on students' college outcomes. The findings suggest that the introduction of English-only instruction led to a decrease of GPAs and probability of graduation and an increase in the number of failed course credits. Although negative, the effects were short-lived. The second study examines whether the chaperone effect, whereby authors who have first published as non-leading authors have better chances to publish later as leading authors, is present in academic publishing in economics, education, political science, psychology, and sociology by analyzing bibliometric data of more than 600,000 articles. The analyses suggest a limited presence of the chaperone effect in social science publications. The third study was a randomized field experiment testing the implementation of the AdmitHub artificially intelligent text-message based virtual assistant to reduce summer melt and improve first-year college enrollment at East Carolina University (ECU) and Lenoir Community College (LCC). At ECU, the positive effects of the virtual assistant were concentrated among first-generation students. For LCC, where the randomized experiment could not be robustly implemented due to having access to cell phone information for only a limited number of students, a qualitative analysis of readiness for chatbot implementation is presented.

Table of Contents

List of Tables

List of Figures

Preface

I am extremely grateful to my phenomenal, extraordinary, and fantastic advisor Dr. Lindsay C. Page. Thank you very much for your unwavering support and encouragement, wise guidance, charismatic and transformative teaching, critical review, creative and inspiring feedback, and inclusion in exciting research projects!

I am very grateful to my wonderful dissertation committee members Osea Giuntella, Sean Kelly, Maureen McClure, and Alvin Pearman. Thank you for teaching and advising me, for encouragement and critique, and for being great role models.

I am grateful to professors Linda DeAngelo, Michael Gunzenhauser, Lindsay Matsumura, Maureen Porter, Najeeb Shafiq, John Weidman, Feifei Ye in the School of Education who taught, mentored, and supported me throughout my Ph.D. studies at the University of Pittsburgh.

I am grateful to the people of Kazakhstan for the Bolashak scholarship, which brought me to the University of Pittsburgh. I am thankful to the Administrative and Policy Studies Department and the Department of Psychology in Education for supporting my studies through teaching assistantships.

I would like to thank the U.S. Department of Education for generous funding of the study on the effects of artificially intelligent chatbots on summer melt. I am grateful to the East Carolina University, Lenoir Community College, and AdmitHub staff, who led the implementation of the chatbot. I thank Nathasha Rodriguez and Dr. John Paul Black for their continued support of the study.

I am grateful to terrific researchers in Dr. Lindsay Page's research lab Dr. Aaron Anthony, Alberto Guzman-Alvarez, Dr. Stacy Kehoe, and Danielle Lowry for willingly mentoring me, collaborating with me, and supporting me with advice, kindness, and optimism.

I thank graduate students in the quantitative research methods course series at the School of Education at the University of Pittsburgh, whom I had the honor of teaching and learning from between 2017 and 2020.

I am grateful to Rebecca Price, Erika Abarca Millan, Daniel Jacobson, Emily Howe, Eben Witherspoon, and the entire cohort of the School of Education Ph.D. students for sharing this academic journey with me and making it not only rigorous but also fun.

I thank the Kazakh community here in Pittsburgh for their support during my studies. Thank you, Shalkar Adambekov and Ainur Mukhambetova, Arai Kairalapova, Fatima Kaupbayeva, Meirman Syzdykbayev, Serzhan Sakipov and Dana Sakipova, and Aigerim Baimoldina for your kindness and warmth.

I thank my parents Alipa and Bakhytbek and brother Azamat, for being the best family in the world and for loving and supporting me all these years. I thank my grandmother Zhamal for giving me the best childhood anyone could ever dream of.

Thank you, Galymzhan, Suleimen, and Aizere, for your love, patience, and support. I was blessed to have you with me in the U.S. all these years, and I admire how much you have grown and learned. You are the real academics in our family.

1.0 Introduction

Higher education has remained on top of the research agenda across a range of disciplines including economics and sociology for at least two reasons. First, investment in higher education yields substantial private and social returns (Angrist & Krueger, 1992; Autor, 2019; Becker, 1993; Card, 1999; Dee, 2004). Therefore, policies aiming to improve college access and to support enrolled students to successfully complete their studies need to be informed by rigorous research. Second, higher education is an important and growing sector of the global economy. The number of college students in the world grew from 32.6 million in 1970 to 182.2 million in 2011 (UNESCO Institute for Statistics, 2014). Such dramatic expansion of higher education has followed demographic trends, wider access to K-12 education, and economic growth (Asian Development Bank, 2011).

This doctoral dissertation examines several aspects of higher education and consists of three independent studies. The first study examines English-only college education in non-English speaking countries, a rapidly growing phenomenon that has been dubbed as the most important trend in higher education internationalization. Despite worldwide popularity, there is little empirical evidence about how the transition to English-only instruction affects students' academic outcomes. Using a natural experiment at a selective university in Central Asia and a difference-in-differences strategy, the study estimates the causal effect of switching to English-only instruction on students' college outcomes. The second study studies the chaperone effect in the social sciences. In the hard sciences, there is evidence of the chaperone effect whereby authors who have first published as non-leading authors have better chances to publish later as leading authors. In other words, to publish as a leading author, one increasingly needs to be chaperoned through prior publication experience. This study examines whether such chaperone effect is present in academic publishing in economics, education, political science, psychology, and sociology by analyzing bibliometric data of more than 600,000 articles. The third study presents the findings of a randomized field experiment testing the implementation of the AdmitHub artificially intelligent text-message based virtual assistant to reduce summer melt and improve first-year college enrollment at East Carolina University (ECU) and Lenoir Community College (LCC).

2.0 Effects of the shift to English-only instruction on college outcomes: Evidence from Central Asia

This is an Author's Original Manuscript of an article written by Nurshatayeva and Page (2019) published by Taylor & Francis in the Journal of Research on Educational Effectiveness on October 17, 2019, available online at the Taylor & Francis Ltd web site: www.tandfonline.com.

The link to the article: https://www.tandfonline.com/doi/full/10.1080/19345747.2019.1652711

2.1 Introduction

In recent decades, many countries where English is not the native language have introduced English as a language of instruction in some or all of their universities. The switch to English as the language of college instruction is considered the most significant trend in higher education internationalization (Parr, 2014). Colleges hope to achieve highly ambitious goals by switching to English. For example, they aim to be more competitive in rankings, produce high quality research, and contribute to economic development (Dearden, 2014; Drljača Margić & Vodopija-Krstanović, 2017; Macaro, Curle, Pun, An, & Dearden, 2018; Salmi, 2009; Wilkinson, 2017). The geography of such countries is notable and includes almost all European, former Soviet, and Asian countries, among others (Ackerley, Guarda, & Helm, 2017; Doiz, Lasagabaster, & Sierra, 2011; Goodman, 2014; Zhao & Dixon, 2017). Finally, the speed with which countries adopt English as the language of instruction in colleges is equally remarkable; in Europe alone, the number of college programs taught in English grew from 725 in 2001 to 8,029 in 2014 (Wächter & Maiworm, 2014).

Despite such international popularity, little is known about the effect of switching to English-only instruction on students. Using English as the language of instruction may have negative effects on students' academic outcomes because instructors and students may be less than proficient in English (Doiz, Lasagabaster, & Sierra, 2013; Macaro et al., 2018). Studies from a variety of contexts consistently have documented the challenges students and instructors experience when instruction is delivered in English (Bolton, Botha, & Bacon-Shone, 2017; Bradford, 2016; Hu & Lei, 2014; Nguyen, Hamid, & Moni, 2016). Yet, we lack reliable evidence about whether these self-reported challenges translate into actual negative effects on academic outcomes (Macaro et al., 2018).

We contribute to filling this gap in the literature on English-only instruction by examining the impact of an institutional switch to English as the language of instruction on student-level academic outcomes. We make use of an arguably exogenous policy shift at a selective university in Central Asia, which introduced English-only instruction at one of its schools while making no changes in the language of instruction in its other schools. We refer to this university as Anon U. We employ a difference-in-differences (DID) design and administrative data on cohorts of students from before and after the policy change to estimate the causal effect of the transition to English-only instruction on a range of academic outcomes. In addition, we examine how the effect of switching to English-only instruction changed over time and explore the mechanisms through which the effect of the language reform may have occurred.

To preview our findings, in the short run the programmatic switch to English has negative effects on student academic outcomes. Specifically, in the first post-treatment cohort, students' GPAs declined by 0.32 points (effects size of 0.36 standard deviations), the number of failed course credits increased by about 6 (0.47 SD), and the probability of graduation decreased by about nine percentage points (0.21 SD). In the following post-treatment cohorts, effects of the language shift on the GPAs and probability of graduation are not statistically significant, and the negative effect on the failed course credits persist in only one subsequent cohort. In other words, the negative effects were driven by a decline in the academic performance of the first cohort of students exposed to English-only instruction, with little evidence of detrimental effects for subsequent cohorts. This rapid fade out of negative academic consequences implies that concerns about potential sustained negative impacts of English-only higher education on students' academic outcomes may not always be warranted. As we discuss below, we find no evidence that a particular, dominant mechanism drove the short-run negative effects.

Our study makes several contributions to the existing literature on English-only instruction. First, we contribute to filling the gap about the effect of English-only instruction on students' academic performance by quantifying this effect using a rigorous quasi-experimental approach. To our knowledge, our paper is the first study of English-only instruction conducted in the causal inference framework. We estimate the impact of shifting to English-only instruction on a set of measurable and policy-relevant academic outcomes. Collectively, students' GPAs, number of failed course credits, and probability of graduation provide a more complete picture of the language reform impacts that can inform institutional policy and further research. Although we recognize this investigation as likely having limited generalizability, it nevertheless represents a research approach that may be applicable in many other settings where English-only instruction is being newly implemented. Second, we extend the debate about English-only instruction by considering how and why the impact changes over time. The short-lived nature of the negative effects we identified and the mechanisms we examined provide the basis for studying the factors that hinder or contribute to the successful implementation of English-only instruction. Third, we use data that is often hard to obtain. Colleges that have implemented English-only instruction may be reluctant to share their data in order to avoid publicizing any negative impacts of their reforms. We show in our study that it is possible to access to such institution-level administrative data and conduct policy-relevant research while protecting the privacy of the data source.

We structure the remainder of the paper as follows: Section 2.2 presents the literature review; Section 2.3 describes the data, background, and analytic approach; Section 2.4 presents estimation results and discusses the mechanisms driving the observed effects; and Section 2.5 concludes.

2.2 Literature Review

2.2.1 English-only instruction: Clarification of the term

We use the term *English-only instruction* and define it as the teaching of academic disciplines in English in non-English speaking countries. There are many other terms that researchers and practitioners use to refer to the use of English as the language of instruction to educate students whose native language is other than English. Common terms include “immersion”, “content and language integrated learning”, and “English-medium instruction”. See Macaro et al. (2018) for a comprehensive review of the phenomenon and terms used to describe it.

2.2.2 English-only instruction may negatively affect students' academic outcomes

Numerous studies suggest that English-only instruction is likely to have adverse effects on students' academic outcomes. For students whose native language is not English, writing, reading, listening to and speaking about academic material in English present additional challenges to the learning process. For example, Bolton et al. (2017) report that college students in Singapore experience difficulties in academic communication even though English-only instruction has been the core national higher education language policy for several decades. Specifically, according to Bolton et al. (2017), nearly 20-30% of engineering students reported difficulty speaking, listening to, and writing in English. Postgraduate students in engineering reported even higher levels of difficulty communicating in English (Bolton et al., 2017).

Similar challenges were documented in South Korea, where English-only instruction has been one of the most salient developments in higher education since the early 2000s. Byun et al. (2011) surveyed students at a university in South Korea and found that 9% of students reported difficulty understanding courses taught in English and that 25% of students requested instructors to use the Korean language during a class taught in English. Similarly, studying implementation of English-only instruction in a Chinese university, Hu and Lei (2014) found that lack of English proficiency hindered most students from understanding large parts of the course content and prevented them from effectively examining abstract disciplinary concepts and from discussing challenging cases in detail. S. Evans and Morrison (2011) found that first-year college students in English-only programs in Hong Kong struggle to understand technical vocabulary, to comprehend lectures, and to write in an appropriately academic style. Combined with difficulties with adjusting to institutional and disciplinary requirements, these challenges posed threats to the academic outcomes of the students (S. Evans & Morrison, 2011). Similarly, studying schools of economics in three Turkish universities, Sert (2008) concluded that the majority of students did not have sufficient English skills to master the material taught in English.

Furthermore, instructors' limited English proficiency may prevent them from helping students to master complex concepts and ideas. Previous studies show that due to a lack of English skills, professors tend to reduce the amount of academic content covered in lectures and slow down lecture pace. Specifically, several studies report that professors cover at least 20-30% less material when teaching in English compared to when they used to teach in their native language. Covering limited amount of course content due to language limitations has been documented in Japan (Bradford, 2016), South Korea (Byun et al., 2011), China (Hu & Lei, 2014), Hong Kong (S. Evans & Morrison, 2011), Turkey (Sert, 2008), and other contexts (Doiz et al., 2013).

Despite these substantial challenges, both students and instructors also report overall positive attitudes towards English-only instruction. Moreover, both students and faculty report using various coping strategies to address the challenges they experience. For example, students in Hong Kong stated that their goals of better academic and labor market opportunities motivated them to counter the difficulties of English-only instruction by working hard and using peer support (S. Evans & Morrison, 2011). Studies in Turkey and China found that students coped with the linguistic challenges of English-only instruction by additionally studying course content in their native language (Hu & Lei, 2014; Sert, 2008). Some authors concluded that better professional opportunities and the desire to help their students may motivate instructors to improve their own English proficiency to become better at teaching through English (Bradford, 2016; Sert, 2008; Zhao & Dixon, 2017).

Our review of the literature highlighted two notable gaps in the literature on English-only instruction. First, although research has documented substantial challenges experienced by students and faculty, it is unclear whether these challenges translate into negative effects on students' academic outcomes. A related question is to what extent both instructors and students whose native language is not English can adjust to college-level coursework in English. It is unclear whether the potential negative effects of switching the language of instruction to English could dissipate over time as students' and professors' positive attitudes and coping strategies help them adapt to the change (Macaro et al., 2018).

These issues deserve careful study. If the documented challenges of English-only instruction convert into deficiencies in content learning and persist over time, then the ambitious goals set by policy-makers implementing English-only instruction may not be attainable, and resources may be wasted. If academic performance suffers when students are instructed in English, then it is unlikely that universities implementing English-only instruction will be able to attain their ambitious goals, such as improving their positions in international rankings (Drljača Margić & Vodopija-Krstanović, 2017), enhancing their research prestige by attracting faculty publishing in top journals (Doiz et al., 2013), enrolling more international students (Macaro et al., 2018; Wilkinson, 2017), and contributing to economic development by building an internationally competitive labor force (Dearden, 2014).

We contribute to filling these gaps in the literature by answering the following research questions:

1) How does the switch to English-only instruction within a postsecondary institution affect student-level academic outcomes?
2) How do the effects of switching to English-only instruction on college outcomes change over time, and what drives these patterns?

2.3 Data, background, and research method

2.3.1 Dataset

We use administrative data from Anon U (pseudonym), a university in Central Asia, for six cohorts of students (N = 2,884) who entered the university between 2007 and 2012. The data come from students' administrative records and include information on their demographic characteristics, high school achievement measures and language of instruction, university achievement measures and language of instruction, and university financial aid status.

2.3.2 Setting

The data from Anon U is well-suited to answer our research questions for three main reasons. First, Anon U is similar in most respects to a typical university that is likely to implement English-only instruction in a non-English speaking country. Specifically, it is selective, western-oriented, and well resourced. It is a small university, enrolling an annual cohort of approximately 350 students, selected from among the top students in the country. As shown in Table 1, the average school-leaving test scores range from 85 to 95 points (out of 100), placing Anon U students in the top quartile of the national distribution for these tests. Other indicators reflect the relatively advantaged background of the average Anon U student. For example, about 85% of Anon U students are from urban areas which are better off economically than rural areas, more than half of the students come from Russian language high schools where academic achievement tends to be higher,[1] and about 90% of Anon U students are of the local ethnicity. In short, the student body is fairly homogeneous and continued to be throughout the time period of our examination. Almost all Anon U graduates find employment within three months of graduation in the national labor market or enroll in competitive graduate programs. Furthermore, Anon U was created in the post-USSR era and modeled after western universities. Expatriates from the U.S. and European countries comprise a considerable share of Anon U's administrators and faculty. Anon U's academic process, grading, and instructional practices are structured similarly to U.S. universities. Anon U is well-funded by the government and enjoys financial support from transnational and local businesses.

Second, document analysis showed that the external and internal pressures surrounding Anon U were similar to the pressures experienced by institutions switching to English-only instruction in other contexts (Doiz et al., 2013).[2] External to the university, the government encouraged the university to pursue international accreditation and impact-factor publications (most commonly written in English). In addition, Anon U experienced competition from newer universities that offered instruction in English. Internally, the university felt the need to diversify its faculty and to improve the quality of education overall.

Third, Anon U's administrative data is of high quality. Modeled after western universities and in order to prevent corruption, Anon U has maintained a comprehensive electronic database of student files and transcripts since its founding. In addition, upon the authors' request, Anon U

[1] Given that this Central Asian country has been a part of the USSR where the Russian language was the dominant language in education and other spheres of life, graduates of the Russian language high schools tend to perform better academically.
[2] Specifically, we examined the country's government plans for strategic development, Ministry of Education plans and reports, and Anon U's senior administrators' interviews in mass media. We do not include sources of these documents so as to protect the identity of the university.

administrators checked the key variables in the electronic database against Anon U's internal documents. For example, given that only financial aid status at entry was recorded in the database, university administrators checked whether students' financial aid status changed during their studies and updated the dataset accordingly.

2.3.3 Language of instruction shift at Anon U

We make use of a natural experiment that took place in 2010 at Anon U to examine the impact of shifting to English as the language of instruction. Before 2010, Anon U allowed its incoming students to choose their language of instruction, with Russian and the local language as options. In the 2010-2011 academic year, Anon U's School of Computer Science changed its language of instruction policy such that from 2010 onward, all students would be able to study in Russian or the local language in the first year only and were required to switch to English starting in the second year. We exploit the fact that only the School of Computer Science switched to English-only instruction in 2010 while all other schools continued to teach in Russian and the local language. Therefore, we use all other schools within the university as the comparison group.

Two features of the language policy shift at the School of Computer Science make identification of causal effects possible. First, the decision was made by the senior management of the university and was imposed on the School of Computer Science. Expecting strong resistance from the faculty and administration of the School of Computer Science, Anon U's senior administration conducted the reform in a very abrupt manner. All possible measures were taken so that faculty and mid-level administrators could not delay implementation of the reform. Within a few months, several non-English speaking faculty members were replaced with lecturers who could teach in English. In short, the policy shift could not have been easily anticipated by faculty and students, making the change an arguably exogenous shock.

Second, course schedules at Anon U are rigidly structured by cohort which enables us to identify the effect of the cohort-based language of instruction change. Specifically, Anon U does not allow students from multiple entry cohorts in a single course. For instance, a computer science course for second-year students would not have any first-, third- or fourth-year students in it. Given this cohort-based curriculum and schedule structure, the 2009 entering cohort experienced no effects of the switch to English and graduated from Anon U taking classes in their originally chosen languages. Such characteristics of Anon U's educational offerings (customary in the country) ensures that the language of instruction switch impacted only the cohorts enrolled in 2010 and onwards.

2.3.4 Research method

We estimate the effects of the language shift on academic outcomes using a difference-in-differences (DID) framework (Imbens & Rubin, 2015; Murnane & Willett, 2011). The School of Computer Science is the treatment group and other schools of Anon U serve as a comparison group. In essence, we compare changes in student outcomes before and after the shift to English instruction to analogous changes in the comparison group to identify the causal effect of the treatment.

The model specification is shown in equation 1.

$$Y_{it} = \alpha + \beta_1 Post_{it} + \beta_2 English_{it} + \beta_3 Post_{it} * English_{it} + \beta_4 \boldsymbol{X}_{it} + \Lambda_t + \Theta_m + \epsilon_{it} \quad (1)$$

In equation 1, $Post_{it}$ is a dummy variable indicating whether student *i* in cohort *t* is from a pre- or post-treatment cohort. $Post_{it}$ is equal to 1 if student *i* entered Anon U in 2010 or later and is equal to 0 if a student entered Anon U before 2010. The pre-treatment cohorts in the data are 2007, 2008, and 2009, whereas the post-treatment cohorts are 2010, 2011, and 2012. $English_{it}$ is a dummy specifying whether a student was enrolled in the school that implemented the shift to English-only instruction (i.e., the School of Computer Science) or not (i.e., the student was enrolled in another school of Anon U). The coefficient β_3 on $Post_{it} * English_{it}$ is the DID estimator of the effect of the language switch on the outcome variable. $\boldsymbol{X}_{it}$ denotes a vector of student characteristics including gender, ethnicity, home locality (urban or rural), language of instruction in high school, school-leaving exam score, financial aid status, and a dummy for whether a student transferred in from another college. When fitting this model, we additionally include fixed effects for cohort (Λ_t) and major (Θ_m).

Y_{it} represents each of our outcomes of interest. We consider three outcomes. The first outcome is GPA at graduation, a continuous variable on a 4-point scale calculated the same way as GPA in U.S. universities. For those who did not graduate, we use GPA at the end of their studies. The second outcome is the number of failed course credits, a continuous variable. To put this variable in context, one credit at Anon U is similar to one credit unit at four-year U.S. colleges, and most courses at Anon U are three-credit courses. The third outcome is degree completion, represented by a dummy variable equal to 1 if a student graduates within five years.[3]

We explore whether the effect of the shift to English-only instruction is stable over time by re-estimating equation 1 using each one of the post-treatment years alone as a post-treatment period. That is, we estimate equation 1 using 2010 as the only post-implementation cohort and

[3] We use five-year graduation to capture both on-time graduation within four years and delayed graduation. Estimates are similar when we use graduation within four years. In the context of Anon U, four years is a typical time to degree for more than 95% of students; longer timeframes are rare and mostly relate to cases in which students take extended leaves due to health reasons, personal circumstances, or, in rare cases, low academic performance.

then do the same for 2011 and 2012. Comparing cohort-specific estimates of β_3 allows us to examine whether the effect of the language switch is stable over time.

2.3.5 Internal validity of the research design

The key assumption underlying causal inference using a DID analytic strategy is the assumption of parallel trends (Imbens & Rubin, 2015; Murnane & Willett, 2011). This assumption, in essence, states that trends in outcomes in the comparison group serve as a valid counterfactual for how students would have performed in the School of Computer Science had the shift to English-only instruction not occurred. We take several steps to consider the reasonableness of this assumption in the Anon U context.

First, we plotted each independent variable across time for the treatment and comparison schools (Figures 1-9). The trends appear parallel across treatment status in terms of gender (Figure 1), locality from which students come (Figure 2), ethnicity (Figure 3), school-leaving exam scores (Figure 5), financial aid (Figures 6-8), and transfer status (Figure 9). The only concern is regarding the divergent trends in the proportions of students from local language high schools (Figure 4). Figure 4 suggests that the School of Computer Science historically attracted more students from local language high schools except for a dip in 2009 and a jump in 2012. In general, this might imply changes in student quality as graduates of the Russian language schools tend to perform better academically due to the dominant role of the Russian language in formal instruction during the USSR times. In addition, Figure 6 and Figure 8 show modest departures. Namely, Figure 6 suggests that trends in the proportion of state grant recipients were not completely identical in the treatment and comparison schools. Similarly, Figure 8 implies that trends in the share of self-supported students were not completely parallel. We discuss whether these patterns threaten causal inference below.

Next, we estimate how each covariate changed across the pre-/post-treatment periods in the treatment and comparison groups. The descriptive statistics shown in columns 1-6 of Table 1 align with the graphical analysis. There were no statistically significant differences in both the treatment and comparison schools after the language shift in terms of gender, locality, ethnicity, and proportions of Anon U students who received grants[4] or were self-supporting. These estimates suggest that although the visual inspection of parallel trends in the share of grant recipients and self-supported students was far from perfectly parallel, those visible divergences do not threaten the parallel trends assumption for these variables.

However, we do observe statistically significant differences for the proportions of students who were educated in the local language (versus the Russian language) at the secondary school level (Table 1 and Figure 4). After the policy shift, the share of students educated in the local language was 11 percentage points higher in the School of Computer Science and slightly lower in the other schools. Third, school-leaving exam scores were significantly lower in the post-treatment period in both groups (Table 1 and Figure 5). In addition, the proportion of state grant recipients was 5 percentage points lower and the proportion of students transferring in from other colleges was 4 percentage points higher in the comparison group while no differences are observed in the treatment group.

We further explore the covariate changes in the DID framework. In Table 1, column 7, we show that four covariates changed considerably in the treatment group after the new policy relative to the comparison group. Specifically, the share of students from local language high

[4] State grants and Anon U grants are mainly merit-based and are awarded to the students whose high school academic performance and school-leaving exam scores.

schools increased, the share of students receiving state grants increased, the share of students receiving Anon U grants decreased, and the share of students who transferred in from other colleges decreased.

We consider whether these shifts in covariates threaten our ability to draw causal conclusions using a DID analytic strategy. As we show in section 4.1, impacts on student outcomes of interest are concentrated in the 2010 cohort. Therefore, we check whether the changes observed in some of the covariates took place simultaneously with the treatment impacts on the outcome variables. We re-fit the DID models for every covariate using only 2010 as the post-treatment period. These DID estimates presented in column 8 of Table 1 show that none of the covariates changed significantly at the time the policy impacts on students were concentrated. In other words, these overall differences in covariates are driven by the cohorts further away from the policy shift. Although we do observe some changes in these baseline measures over time, they are not in obvious alignment with the patterns of impacts of the policy on student outcomes. Furthermore, the shifts in covariates for the later cohorts do not imply these cohorts were comparatively stronger than the 2010 cohort. Specifically, as we show in section 4.2, the post-treatment cohorts were similar to the pre-treatment cohorts in terms of overall academic preparation and English language proficiency. In other words, the post-treatment cohorts were not better able to manage English-only instruction. Taken together, we judge the parallel trends assumption to be reasonably well met. In addition, we control for the covariates discussed here in all of our preferred model specifications.

We now turn to estimating the impact of the policy shift on student-level academic outcomes.

2.4 Results

2.4.1 The effect of switching to English-only instruction on academic outcomes

In Table 2, we present results from estimating equation 1 to examine impacts on student GPA, course failure and college completion. The coefficients associated with the Post*English interaction term denote the estimate of the language switch effect. For each outcome, we present four specifications: the model without any additional covariates, the model with student characteristics, the model with student characteristics and fixed effects for students' majors, and finally, the model with student characteristics and fixed effects for majors and cohorts. In our discussion here, we focus on results from the fourth specification.

As shown in Table 2, the switch to English-only instruction had a negative impact on academic outcomes. First, GPA fell by 0.13 points (0.15 standard deviations). We note that when students at Anon U retake a course they failed, their new grade overrides the one previously earned. In other words, the estimated decline in GPA is likely smaller than the actual decline in GPA before adjusting for retaken courses. Second, the course failure rate increased by about three course credits (0.22 standard deviations). A typical course at Anon U is worth three credits, therefore, students failed one more course on average after English-only instruction was implemented. Third, the probability of graduation dropped by about 7 percentage points (0.17 standard deviations). This impact is considerable, given the high academic capability of Anon U students. As a point of comparison, the typical 5-year graduation rate is around 80% at Anon U.

To test the robustness of our estimates to the choice of comparison, we reran the models using each of the non-treatment schools as a comparison group. We present results in Table 3. The coefficients change slightly depending on which non-treatment school serves as the comparison.

For example, the estimates are larger in magnitude when comparison schools are those where average academic performance has historically been higher (e.g., the School of Chemistry and the School of Business). In contrast, estimates are somewhat smaller when the comparison school is limited to the School of Engineering, in which students have a relatively higher number of failed courses and lower probability of graduation. Substantively, however, estimates remain similar to those using all other schools together as the comparison group.

Next, we explored whether the effect of the shift to English-only instruction is stable over time. First, we graphed raw outcomes across cohorts. Figures 10-12 suggest the fading out of the negative effect of the policy shift, with a "bump" for the 2010 cohort followed by a recovery for the 2011 and 2012 cohorts. The overall pattern in the outcome variables suggests a short-run disruptive effect of the shift to English-only instruction to which the School of Computer Science was better able to adjust in subsequent years. Second, we re-estimated impacts for each of the post-treatment cohorts separately. We present results in Table 4. The policy shift had the most consistently negative effects for the first cohort of students to experience English-only instruction. Specifically, the 0.12 point drop in GPA estimated using the full sample (column 1 of Table 4) was driven by the 0.32 point GPA decline of the 2010 cohort that was the first to experience the switch to English-only instruction (column 2 of Table 4). Similarly, the seven-percentage point decrease in the probability of graduation estimated using the full sample (column 1 of Table 4) was concentrated in the 2010 cohort, which experienced a nine percentage-point decline (column 2 of Table 4). In terms of failed course credits, the increase of 3.03 course credits failed estimated for the pooled sample (column 1 of Table 4) is driven mostly by the increase of failed course credits by 5.7 in the 2010 cohort (column 2 of Table 4). Unlike the other two outcome variables, however, we do observe a negative effect on course failure for the 2012 cohort (column 4 of Table

4). Nevertheless, the negative effect of 3.34 more course credits failed of the 2012 cohort is somewhat smaller in magnitude than for the 2010 cohort (5.66 course credits failed).

These results for failed course credits suggest that the 2012 cohort found it hard to learn in English and failed more course credits than the pre-treatment cohorts. However, it is necessary to view these effects in conjunction with the stable fading out pattern in the GPA and graduation outcomes. The academic guidelines at Anon U (and other universities in this country) indicate that when a failed course is re-taken, the “Fail” grade is replaced by the new grade. Thus, if a failed course is re-taken with a better grade, the “Fail” grade does not affect GPAs and probability of graduation. In sum, the patterns for the 2012 cohort suggest that although students may have still struggled with course instruction in English (leading to more course failures), but retaking courses, they were able to recover in terms of GPA performance and still graduate on time. Given that a typical course equates to three course credits, the typical student would have needed to re-take about one course as a result of course failure.

2.4.2 Why was the negative effect of the language shift temporary?

Here, we consider potential drivers of the short-lived decline in students’ academic performance following the switch to English-only instruction.

Did the School of Computer Science start to enroll students who are academically stronger?

One possibility is that academic ability was higher in the cohorts following the first treated cohort because Anon U’s School of Computer Science started to attract students who were more capable academically. To assess this hypothesis, we use two proxies for the academic ability of Anon U’s incoming students. The first proxy is the national standardized school-leaving exam scores. As shown in Table 1, these school-leaving exam scores fell by about 5 points in the School of Computer Science after it switched to English-only instruction. Exam scores fell by about 6 points in the comparison schools. The DID estimates using the full sample (Table 1, column 7) and using only 2010 as the post-treatment period (Table 1, column 8) were both statistically insignificant suggesting that the treatment school did not experience changes in academic ability of its incoming students relative to its comparison schools. Figure 5 shows that the drop in school-leaving exam scores took place in both the treatment and comparison schools in 2008 and that after 2008 the tests cores remained relatively stable through 2012. These patterns suggest that there was no improvement in student quality as measured by these national school-leaving exam scores.

The second proxy we use for academic ability is grades in the first semester at Anon U. There are two advantages of using these first-semester grades as proxies for academic ability. First, the courses in the first semester were taught in either Russian or the local language corresponding to the language in which students studied in high school. Therefore, we can observe student performance free from the effect of English as the language of instruction. Second, several of these first-semester courses at Anon U are not chosen by students but are mandated in the curriculum. Incoming freshmen are automatically enrolled in a sequence of courses including math, physics, history, and languages. The contents of these courses is fairly standard and has not changed over the period under study. Given such general content and mandated enrollment, grades in these courses serve as an institution-level standardized measure of students' ability to do college-level coursework.

We plot averages of grades in the first semester courses for each cohort in the treatment group in Figure 13. Grades in these first semester courses remained quite stable over the period under study suggesting that there were no substantial improvements in student body. If anything, the 2011 incoming cohort struggled somewhat more in math compared to other cohorts. We confirm that no improvements in students' academic ability took place at Anon U by comparing average grades in the pre-treatment period to the average grades in the post-treatment period. As Table 5 shows, there were no significant differences in the physics, history, and language performance, whereas math grades were somewhat lower in the post-treatment cohorts. Math and physics grades were lower in the comparison schools as well, although history grades in the comparison schools were higher in the post-treatment period. The DID estimates using the pooled sample in column 7 of Table 5 show that the treatment school grades in first semester math, physics, and languages did not change while grades in history decreased relative to the comparison schools. As shown in column 8 of Table 5, a similar pattern of first semester grades is observed when only 2010 is used as the post-treatment period with the exception of slightly better grades in physics. In other words, both the descriptive statistics and the DID estimates indicate that there was no consistent improvement in the academic preparation or first semester performance of students in the School of Computer Science. In sum, examination of national school-leaving exam scores and grades in first semester courses suggests that it is unlikely that the fadeout of the negative effect of switching to English-only instruction happened due to improved academic ability across the cohorts.

Did the English language proficiency of the incoming students improve from 2011 onwards?

Even though overall academic ability didn't change, it is possible that the improvement in academic outcomes was due to an increased level of English proficiency of the students entering the School of Computer Science after 2010. Anon U tests all incoming students' English proficiency to place them into a relevant English language course at the elementary / pre-intermediate, intermediate, or upper-intermediate / advanced levels. We were able to obtain English language course placement for 2010, 2011, and 2012, but not for early cohorts. A limitation is that placement test scores for the 2010 cohort were incomplete, namely, 22% of observations were missing. To handle this, we imputed values of the missing scores.[5]

As shown in Panel A of Table 6, there were no drastic improvements in terms of English proficiency in the treatment school between 2010 and 2012. Although the share of upper-intermediate and advanced students increased from 8% in 2010 to 16% in 2012, the percentage of elementary and pre-intermediate level students remained relatively stable at about 46% in 2010 and 2011 and reached 68% in 2012. In addition, the proportion of intermediate level students declined every year from 43% in 2010 to 16% in 2012. Thus, it is unlikely that improved English proficiency explains the general recovery in the 2011 and 2012 cohorts.

Further, as Panel B of Table 6 shows, the trends in the English language proficiency in the comparison schools were similar to those of the treatment school. Specifically, the proportion of elementary and pre-intermediate level students remained stable at about 60%, the share of intermediate students declined gradually from 30% in 2010 to 20% in 2012, while the fraction of upper-intermediate and advanced level students increased from 8% in 2010 to 23% in 2012. Overall, the proportion of top English proficiency level students didn’t change dramatically while the share of lower English proficiency level students increased in the treatment school and remained stable in the comparison schools. In other words, we do not observe improvements in

[5] We impute English language placement test scores because they are important for examining whether Anon U started to recruit students with better English language skills. We used multiple imputations with chained equations based on the following variables: gender, ethnicity, urbanicity, language of instruction in high school, school-leaving exam score, GPA at graduation, number of fails by graduation, graduation status, grade in first semester English language course, cohort, major, and financial aid status. We estimated the proportions of students by English proficiency level using 20 multiply imputed samples.

English proficiency of the incoming students and thus have no affirmative support for this mechanism driving our main estimates.

Did the School of Computer Science change its criteria for course passage and graduation?

The negative effect of switching to English-only instruction may have been temporary because the School of Computer Science changed its criteria for course passage and graduation after seeing negative effects for the first cohort of students to experience English-only instruction. We rule out this mechanism based on our understanding of the Anon U context. Although the USSR-era total state control over the education system has been relaxed, state standards and regulations still govern all important aspects of college instruction in the context that we study. More than 50 regulatory statutes direct the day-to-day academic process. The skills and themes to be mastered in every course are prescribed by the state curriculum and faculty members have limited space to substantially alter the content covered in their courses. State standards regulate student admissions, student assessment, passage from year to year based on academic performance, and graduation requirements. Every five years the Ministry of Education audits all universities for compliance with the state standards. The Ministry of Education auditing teams examine the university administrative records (including course syllabi and department meeting minutes) for the preceding 5 years, interview faculty and students, and conduct teacher evaluations by attending several classes selected at random. Noncompliance with state regulations uncovered during these audits leads to the recall of the university's license to operate, and either the university or the non-compliant school within it closes down. The Ministry of Education has closed several universities and in some cases rescinded licenses partially so that colleges could not offer certain majors that did not comply with the regulations. The most recent Ministry of Education audits of

Anon U took place in 2010 and 2015 and concluded that Anon U was fully compliant with the state standards. Given the high stakes of noncompliance with the state standards and given the results of the most recent audits of Anon U, it is unlikely that Anon U would risk its license by allowing the School of Computer Science to change its course passage and graduation requirements.

Did a change in instructor characteristics drive the temporary dip in academic outcomes?

The introduction of English-only instruction required hiring faculty who could teach in English. Thus, it is possible that the dip in academic outcomes in the first treatment cohort is a function of changes in instructor characteristics. To explore this potential mechanism, we studied how the proportion of new instructors hired at Anon U changed between 2007 and 2012. As Figure 14 shows, the proportion of new professors has historically been quite high between 30 and 40% in both the treatment and comparison schools. In addition, in some years, the proportion of new professors was 50% or higher. In other words, overall, faculty turnover appears to be quite high at Anon U. Importantly, there are no significant fluctuations in academic outcomes in the pre-treatment period where substantial variation in the proportion of new instructors took place. This suggests that student performance is not linked to whether or not instructors are newly hired at Anon U.

To further test this mechanism, we examined the relationship between instructors' years of experience at Anon U and students' grades.[6] Table 7 presents estimates from regressing each student's grade in each course on the years of experience at Anon U of the instructor who taught every given course. The unit of analysis is each student's grade in each course. According to these

[6] We were not able to obtain data on the average overall years of experience or other instructor characteristics.

OLS estimates in Table 7, instructors' years of experience at Anon U are not correlated with students' performance. The OLS estimate is not statistically significant suggesting that students performed similarly in courses taught by more experienced instructors and in courses taught by more recently hired instructors. Quantile regression estimates presented in Table 7 confirm that instructors' years of experience at Anon U are not systematically related to students' performance. Overall, the OLS and quantile regression results suggest that instructors' years of experience have little to do with students' academic performance at Anon U's School of Computer Science.

We also examined the average years of experience at Anon U among instructors who taught courses that were the most challenging for the students. Table 8 shows the courses in which the majority of "Fail" grades are concentrated. As the rightmost column of Table 8 indicates, these "top three" most challenging courses were taught by experienced instructors, with years of experience at Anon U ranging between 2 to 6 years, on average. For the 2010 cohort which had the biggest decline in academic performance, none of the most difficult courses was taught by the newly hired instructors. In contrast, students struggled the most in the courses taught by professors who previously taught at Anon U and who themselves had to adapt to teaching their courses in English starting from 2010. Collectively, these descriptive statistics and the OLS and quantile regression analyses indicate that a change in instructor characteristics was not the mechanism through which the decline and subsequent rebound of academic performance occurred. The estimates suggest that both students and faculty needed to adapt to the English-only instruction.

Another important insight from Table 8 is that the lists of the most difficult courses for each cohort are represented predominantly by the mathematics courses and elective courses taught by departments outside of the School of Computer Science. Table 8 shows that students found mathematics courses most difficult both before and after the switch to English-only instruction in

2010. The “Calculus 2” and “Probability theory and statistics” courses were challenging for students to master prior to the language reform as well as after the reform had been implemented. In the 2011 and 2012 cohorts, where we observe recovery in academic outcomes according to the DID estimates, students appear to have figured out how to avoid failing “Calculus 2”. However, the courses in “Probability theory and statistics” and “Differential equations” were still challenging for some of the post-treatment cohorts. Overall, the courses that posed the biggest challenges both before and after the language reform were mathematics courses and elective courses such as “Principles of economics” and “English for professional purposes” taught by departments outside of the treatment school.

It is sensible that the computer science courses did not make it to the top of the lists of the most challenging courses in Table 8. Computer coding is done in English even when lectures and tutorials are in the native language, so students did not experience drastic shocks in studying their core computer science courses due to overall familiarity with computer technologies and due to having introductory-level coding skills. In contrast, studying advanced mathematics in English requires abstract thinking using the English language and involves the use of terms that should be learned in English. Elective courses outside of the School of Computer Science also require intensive learning of English terms and expose students to discourses, types of assignments and academic activities (essays, reading literature not related to computer science directly, etc.) that are not as common in the core computer science courses. At the same time, both mathematics and non-core electives taken outside of one’s home school were challenging for students prior to the language reform suggesting that the switch to English-only instruction made the traditionally difficult courses even more difficult for the students. This highlights provision of academic support in challenging courses as an important policy lever that could potentially help universities to successfully implement English-only instruction. Specifically, while implementing English-only instruction, colleges could mitigate the negative effects of the reform on students' academic outcomes by providing more support to instructors and/or offering academic support services to students in courses requiring acquisition of vast new terminology and in courses exposing students to the academic work less common in the core courses.

2.5 Conclusion

We estimated the effects of switching to English-only instruction on college outcomes using a DID framework and administrative data from Central Asia. Effects on the outcome variables are in the hypothesized direction. Students failed more courses and their GPA decreased when English-only instruction was introduced. In addition, more students did not graduate within five years. Our results suggest that the challenges of English-only instruction discussed in the literature (Bolton et al., 2017; Doiz et al., 2013) may indeed translate into lower academic achievement. Despite having the top students of the country and despite giving them one year within Anon U to prepare for the language transition, the short-term negative effects of the switch to English-only instruction on students' academic performance are considerable in magnitude.

Nevertheless, the negative effects of the policy shift fade out rapidly. Our estimates suggest that there was a dip in academic performance in the year of transition from which the students (and the system) subsequently recovered. In other words, the switch to English-only instruction posed challenges for the first cohort of students who experienced the language of instruction transition but did not bring about longer-run sustained challenges. This finding, consistent with Macaro et al. (2018), suggests that at least in contexts similar to the one we studied, it is possible that the students and faculty can adapt quickly to the language of instruction change.

We examined several possible mechanisms for the pattern of effects we observe. Our analyses show that the decline and subsequent improvement in academic outcomes did not happen because the treatment school started to enroll students who were more capable academically. Neither did the English proficiency of incoming students improve. The course passage and graduation requirements did not change either. Finally, the data showed that changes in instructor characteristics did not drive the academic outcomes of students. Our interpretation of the lack of a single mechanism through which the change occurred is that both students and faculty likely adapted to the change.

The main policy implication from our findings is that selective universities aiming to switch to English-only instruction may be successful in doing so in the long run. Nevertheless, they should be prepared to proactively mitigate potential declines in academic performance of the first cohort exposed to the language transition. The negative effects of the language shift may be alleviated by providing more support to the instructors teaching English-only courses requiring mastery of extensive new terminology and in courses expecting students to engage in academic activities less common in their core courses. Offering academic support services to students taking such courses may be another strategy for reducing the shock effect on the transition cohorts and contributing to the success of the language policy. Overall, that short-term losses in student outcomes dissipate over time may justify the significant financial and training investment required for implementing English-only instruction.

There are limits to inference in our study arising from looking at a single discipline at a single institution. Indeed, the impact of the transition to English-only instruction is likely subject to a number of factors that vary from country to country as well as within countries. Due to the nature of English-only instruction in non-English speaking countries, future research is likely to remain restricted to single institutions. To address this limitation, future work on this topic should involve more case studies from various geographic and institutional contexts. Universities in general and universities implementing English-only policies in particular are sitting on vast amounts of rich student-level microdata. The key variables used in our study are available in any institutional database, so we welcome replications and further conceptual and methodological extensions of our study. In addition, future work examining the implementation of English-only instruction should focus on policies that could effectively support both students and instructors before, during, and after the language of instruction transition.

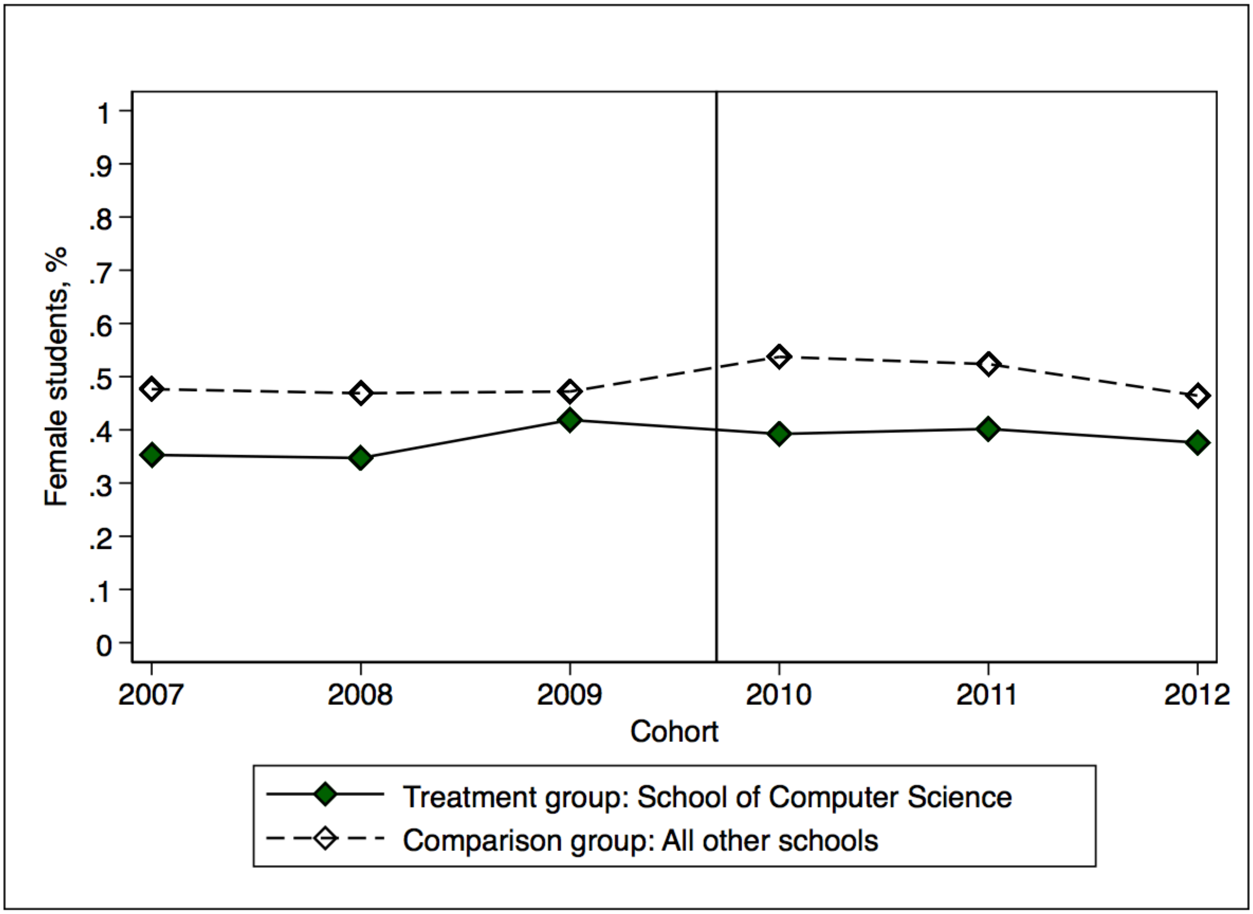

Figure 1 Testing parallel trends assumption: proportion of female students in the treatment and comparison groups by cohort

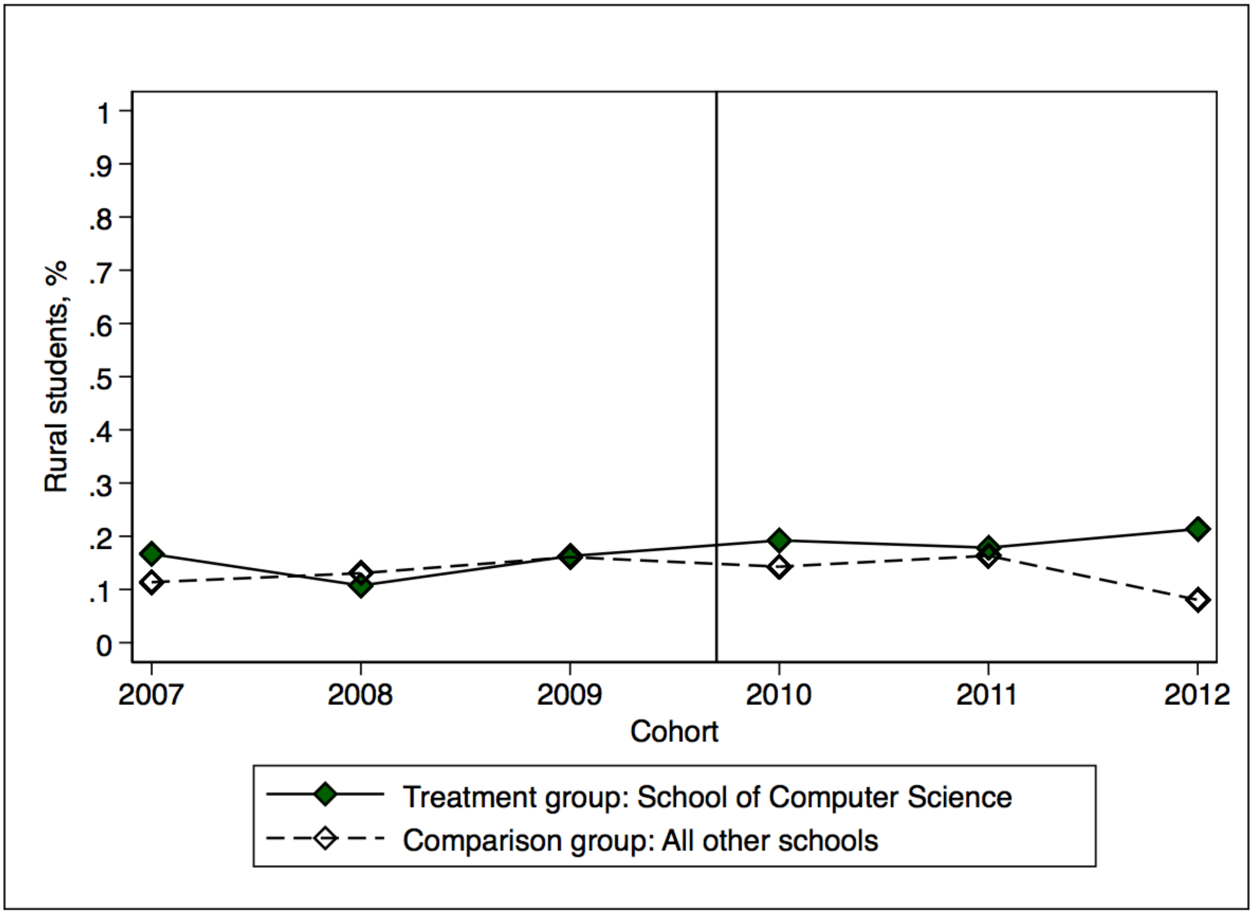

Figure 2 Testing parallel trends assumption: proportion of rural students in the treatment and comparison groups by cohort

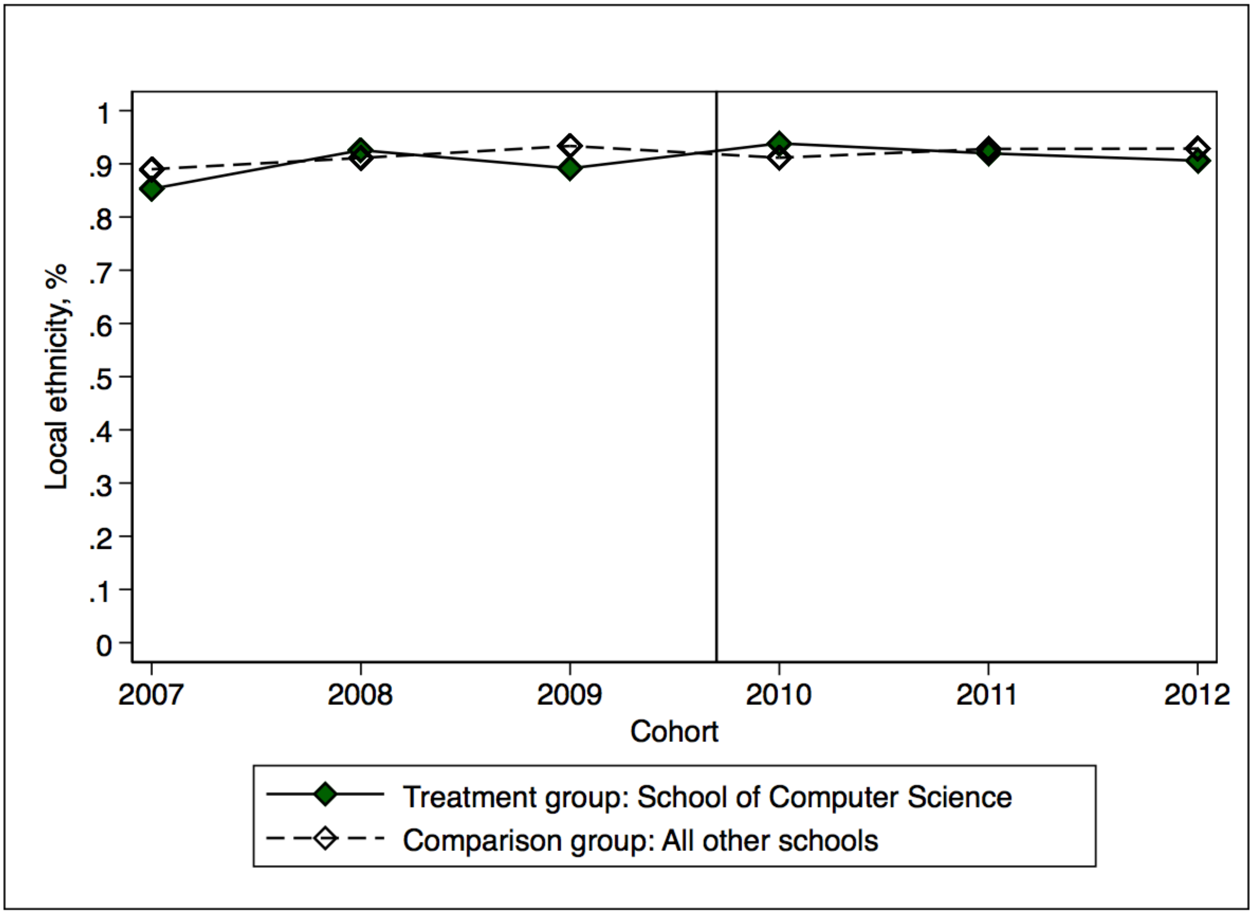

Figure 3 Testing parallel trends assumption: proportion of local ethnicity students in the treatment and comparison groups by cohort

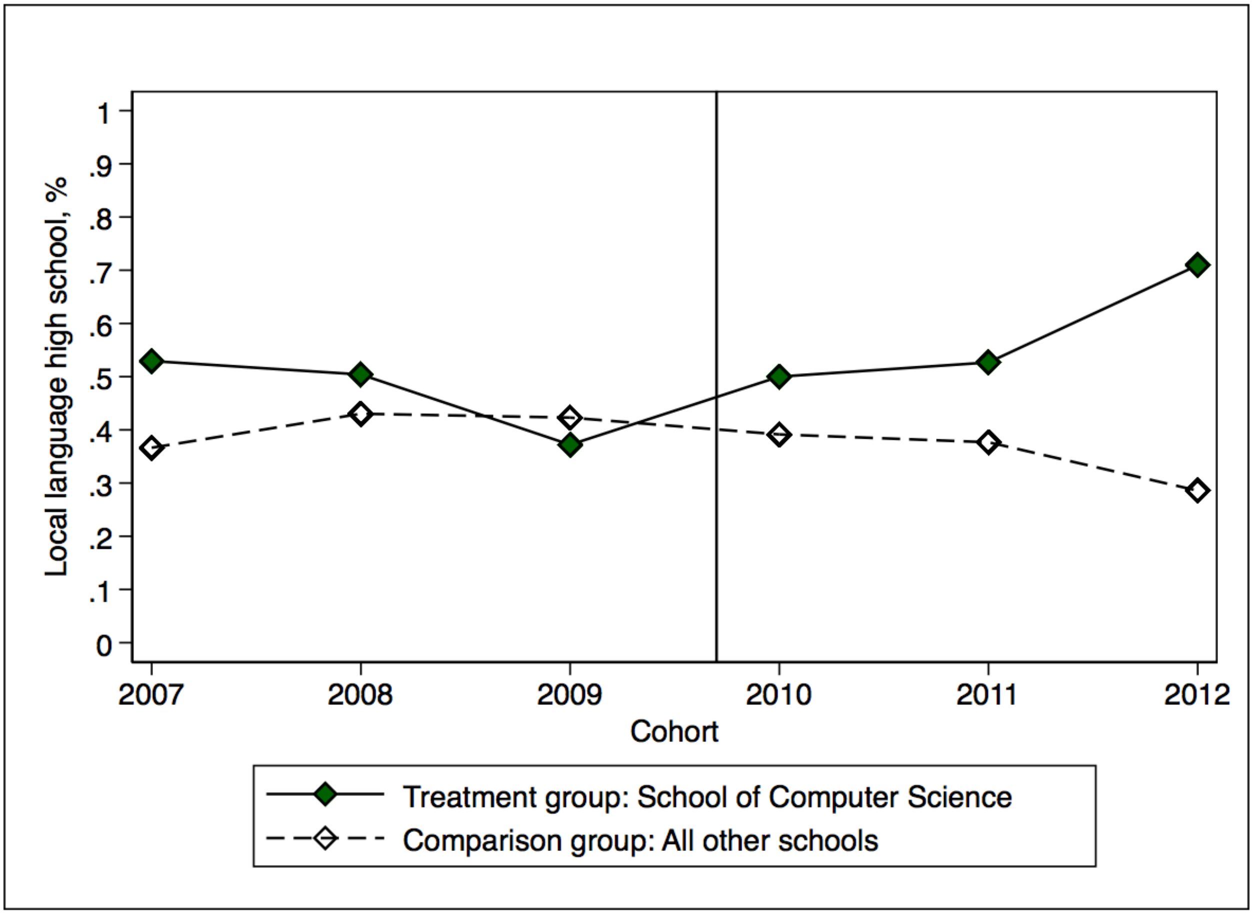

Figure 4 Testing parallel trends assumption: proportion of students from local language high schools in the treatment and comparison groups by cohort

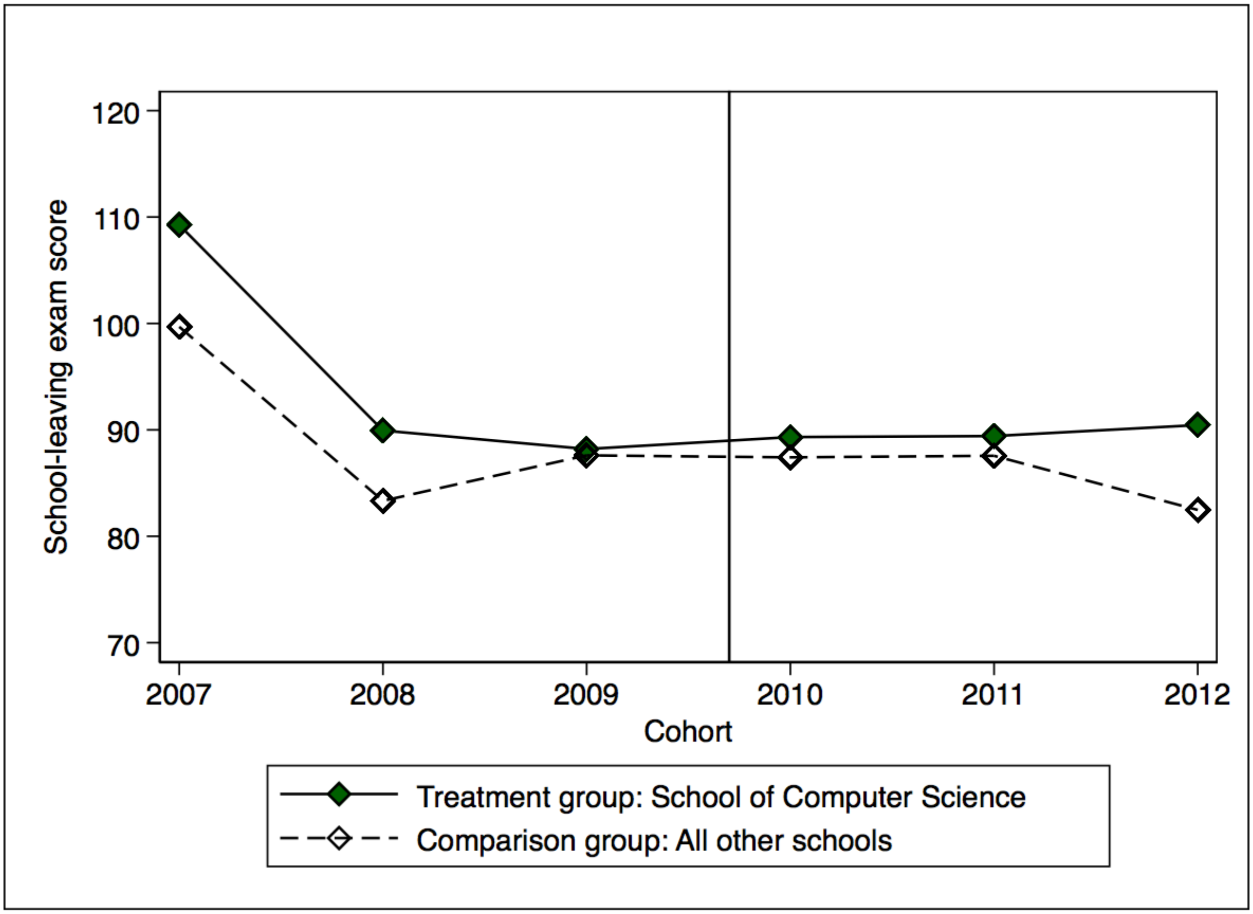

Figure 5 Testing parallel trends assumption: school-leaving exam score in the treatment and comparison groups by cohort

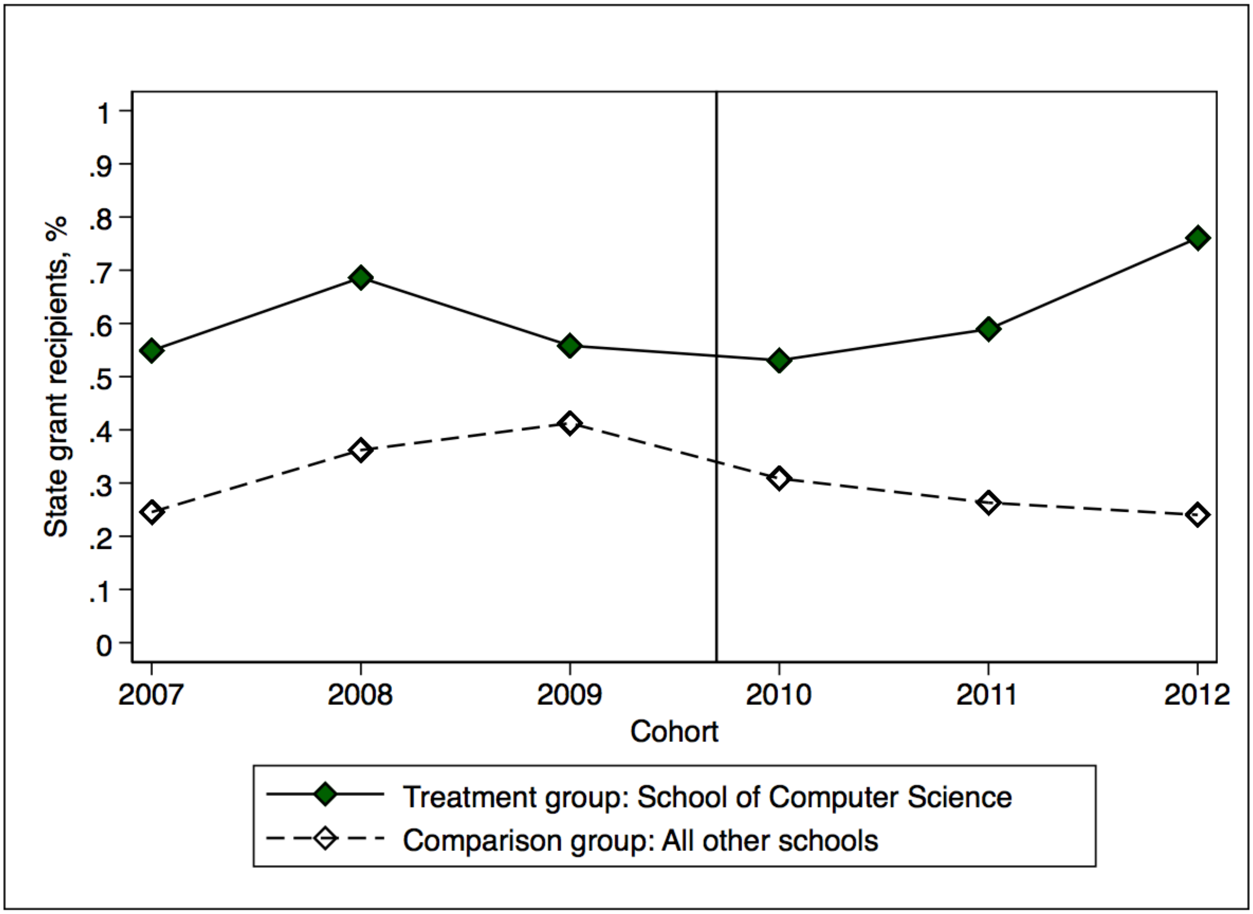

Figure 6 Testing parallel trends assumption: proportion of state grant recipients in the treatment and comparison groups by cohort

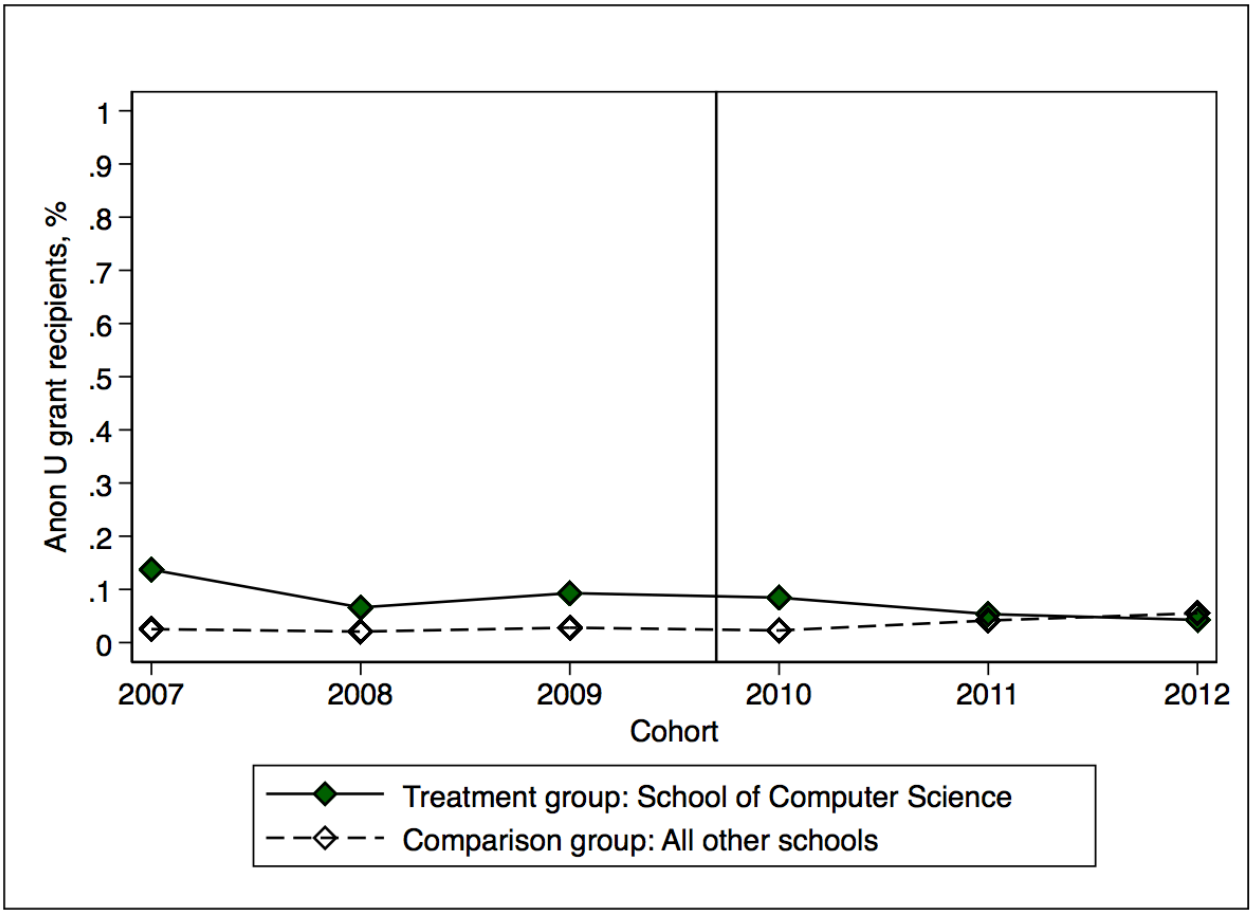

Figure 7 Testing parallel trends assumption: proportion of Anon U grant recipients in the treatment and comparison groups by cohort

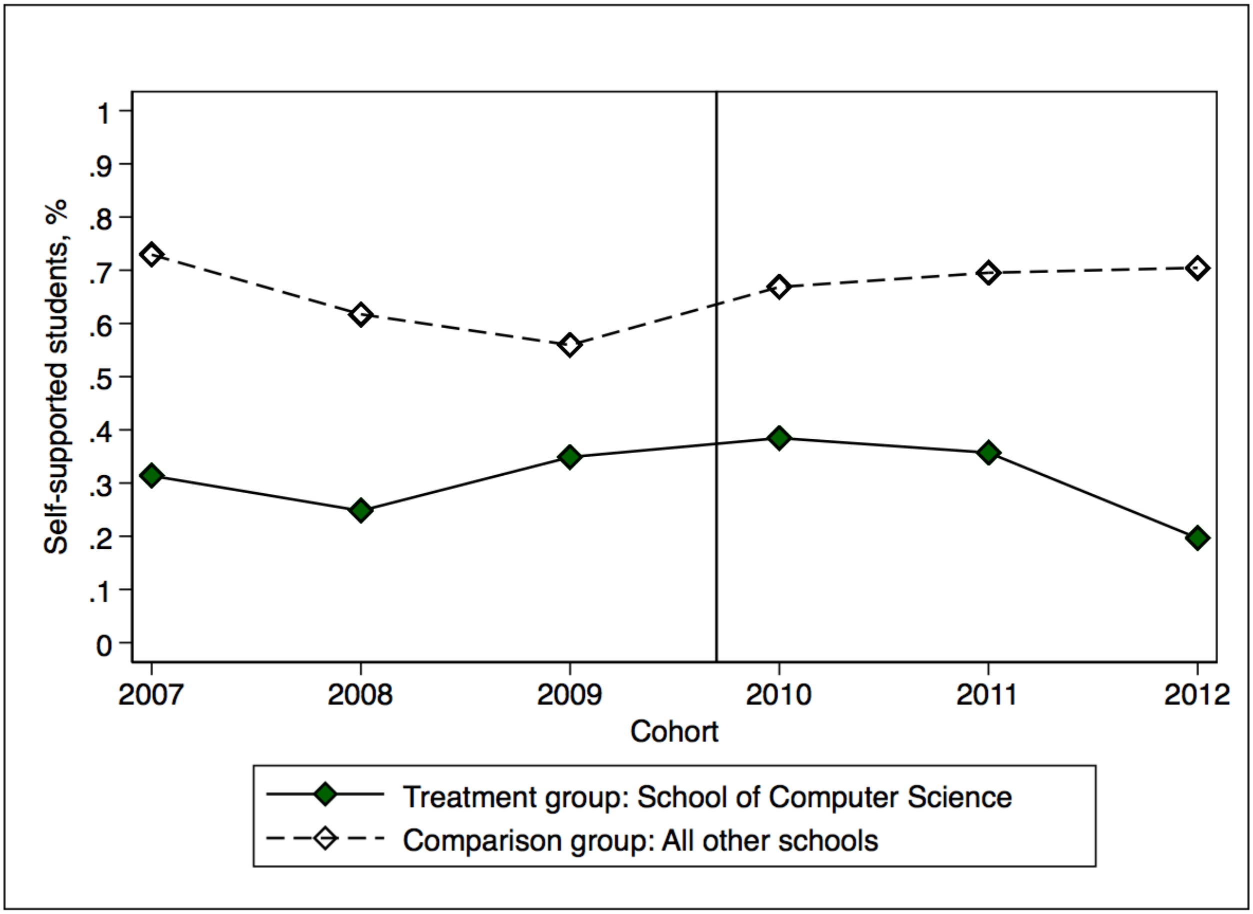

Figure 8 Testing parallel trends assumption: proportion of self-supported students in the treatment and comparison groups by cohort

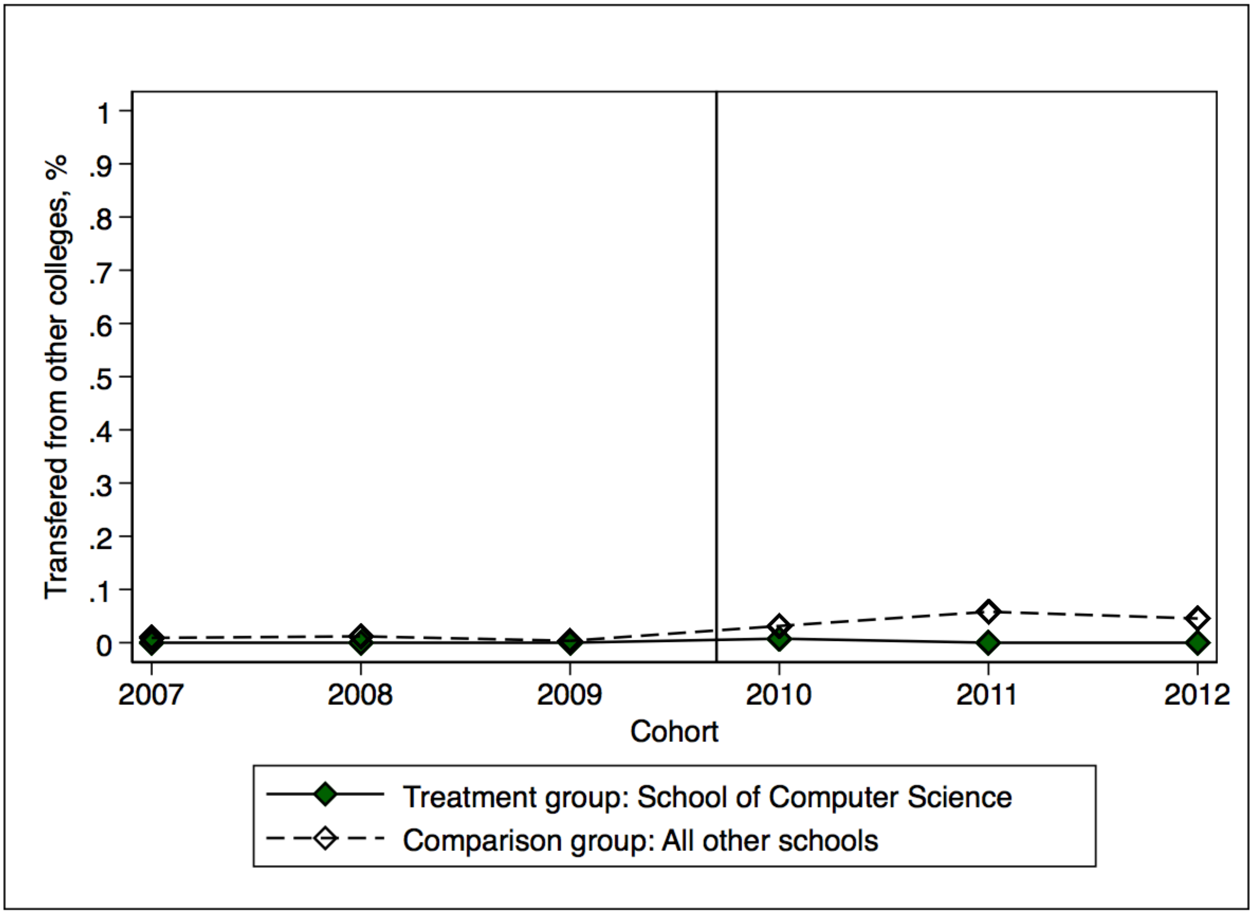

Figure 9 Testing parallel trends assumption: proportion of students who transferred in from other colleges in the treatment and comparison groups by cohort

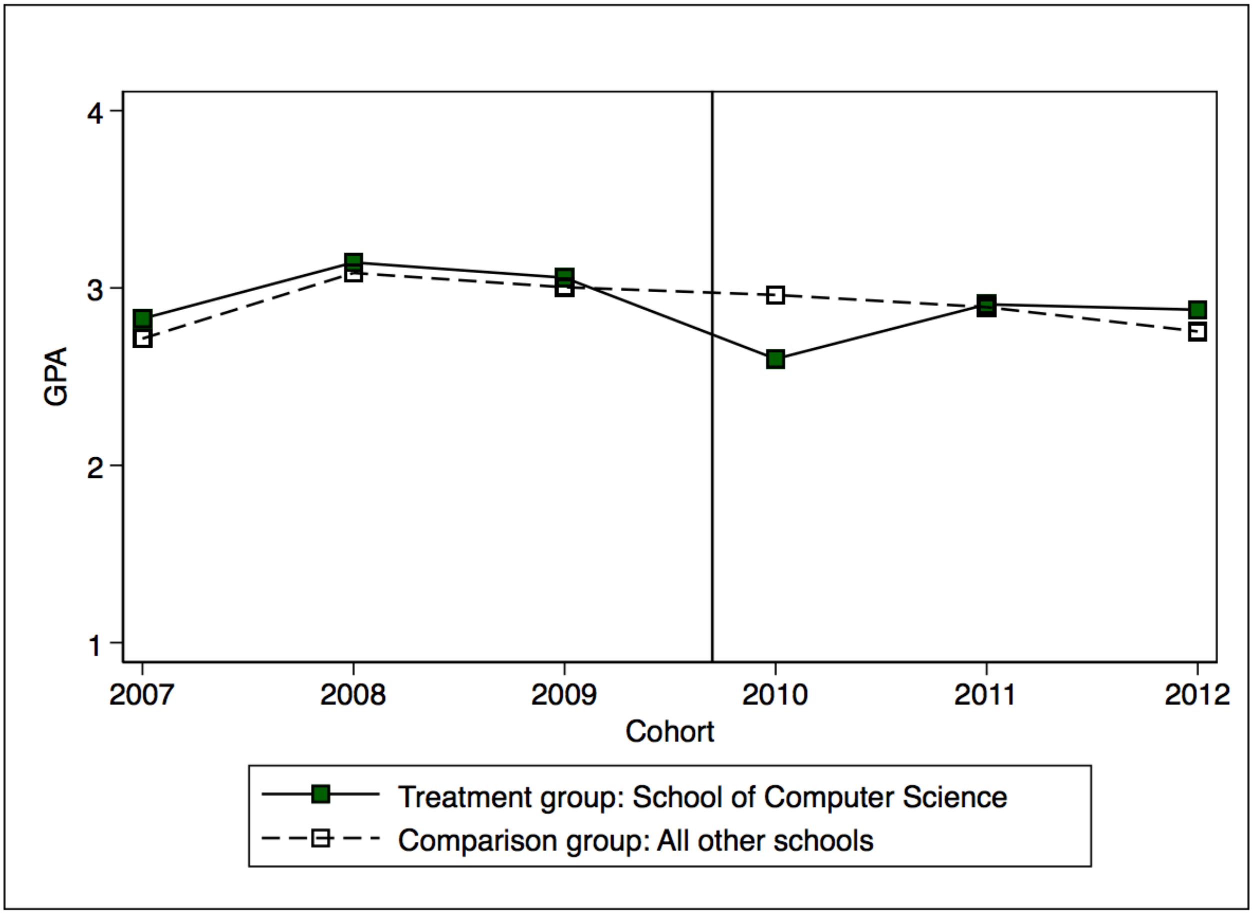

Figure 10 GPA of the treatment and comparison groups by cohort

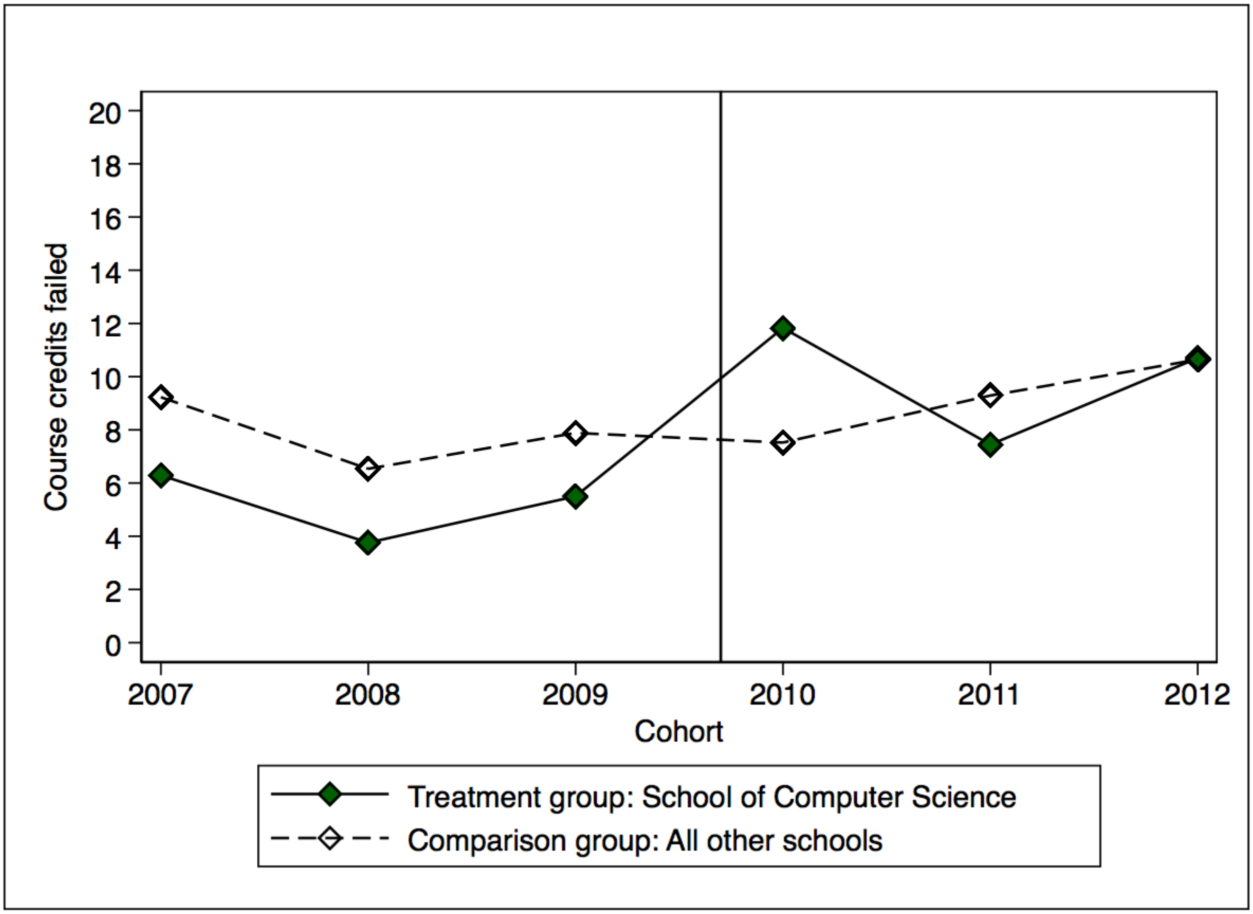

Figure 11 Course credits failed in the treatment and comparison groups by cohort

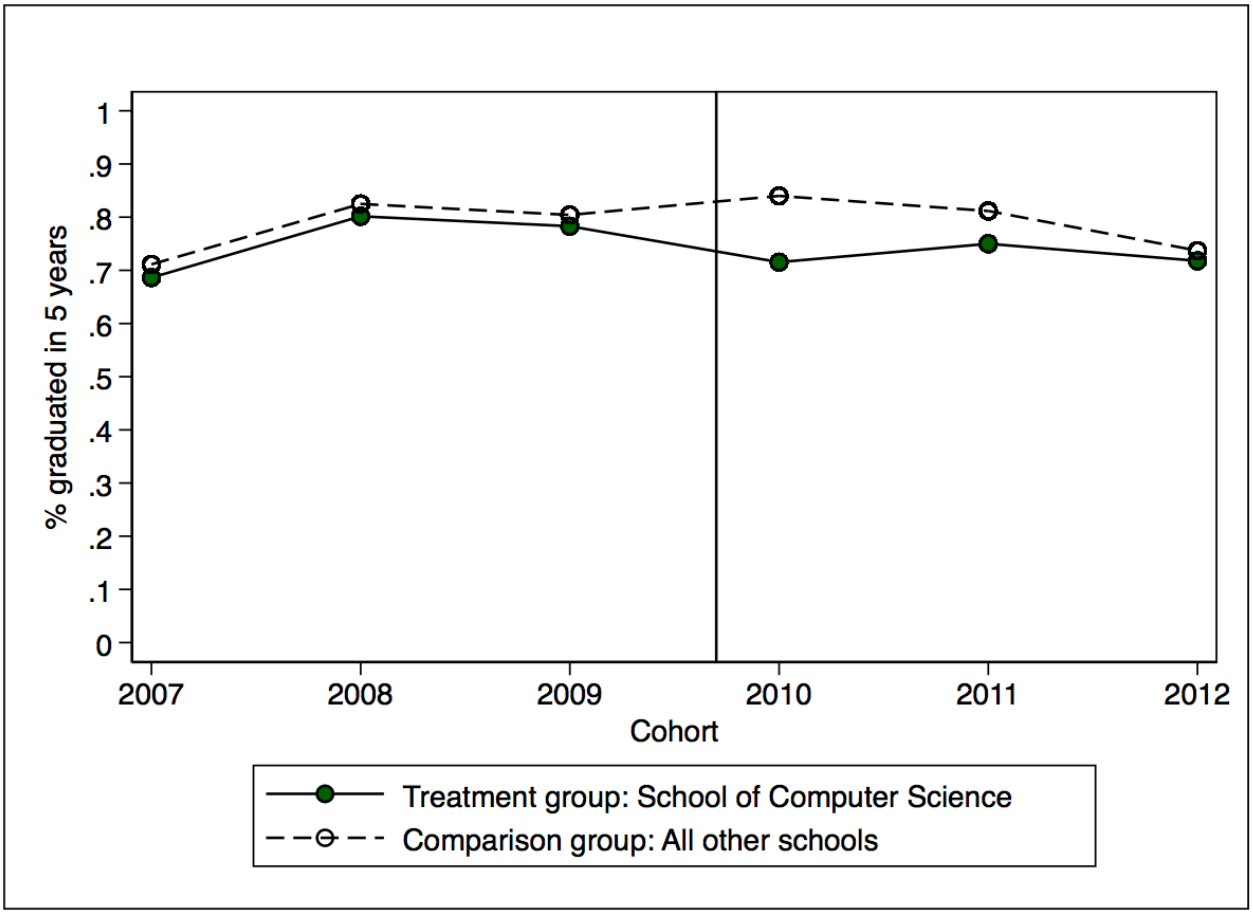

Figure 12 Graduation rate of the treatment and comparison groups by cohort

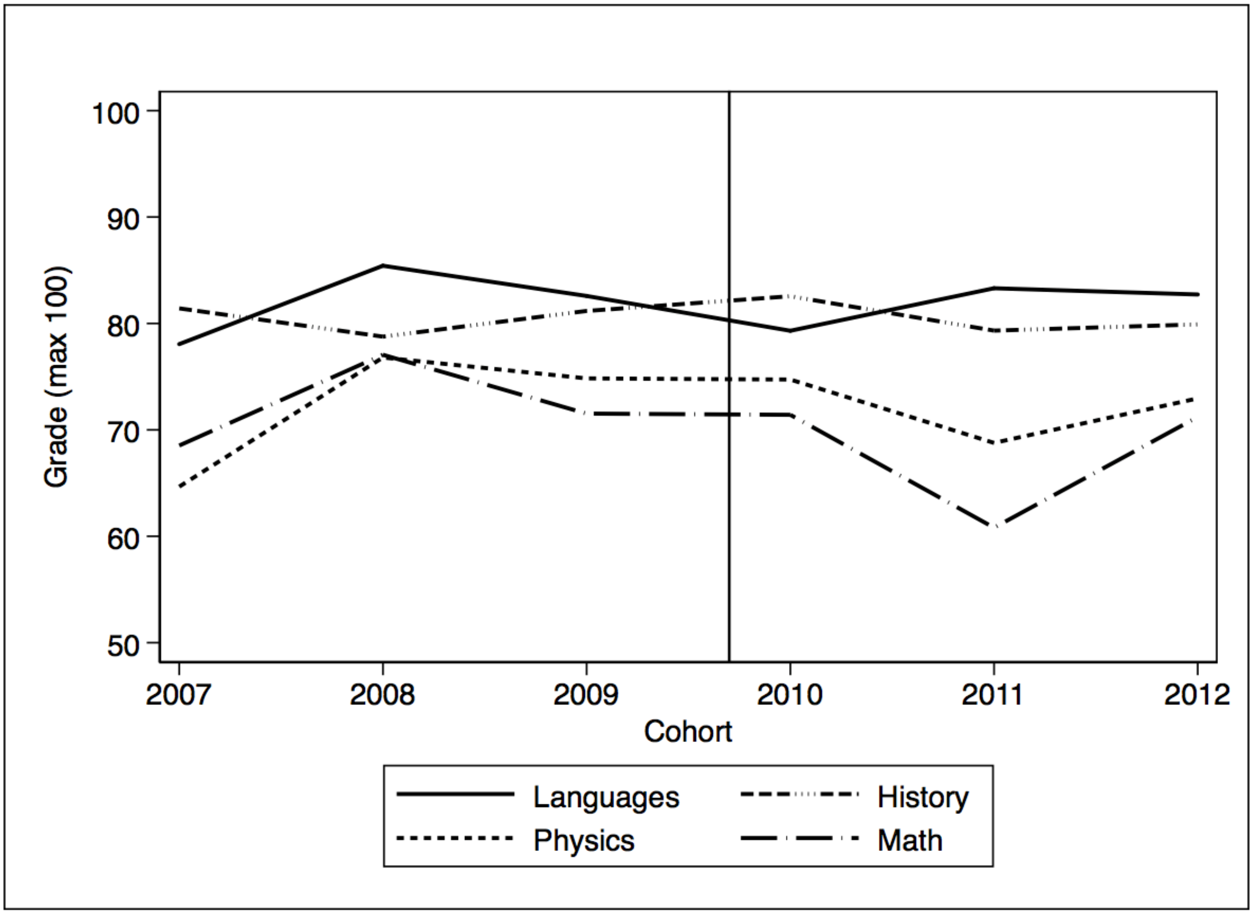

Figure 13 Examining shifts in incoming students' academic ability: grades in first semester courses in the treatment group by cohort

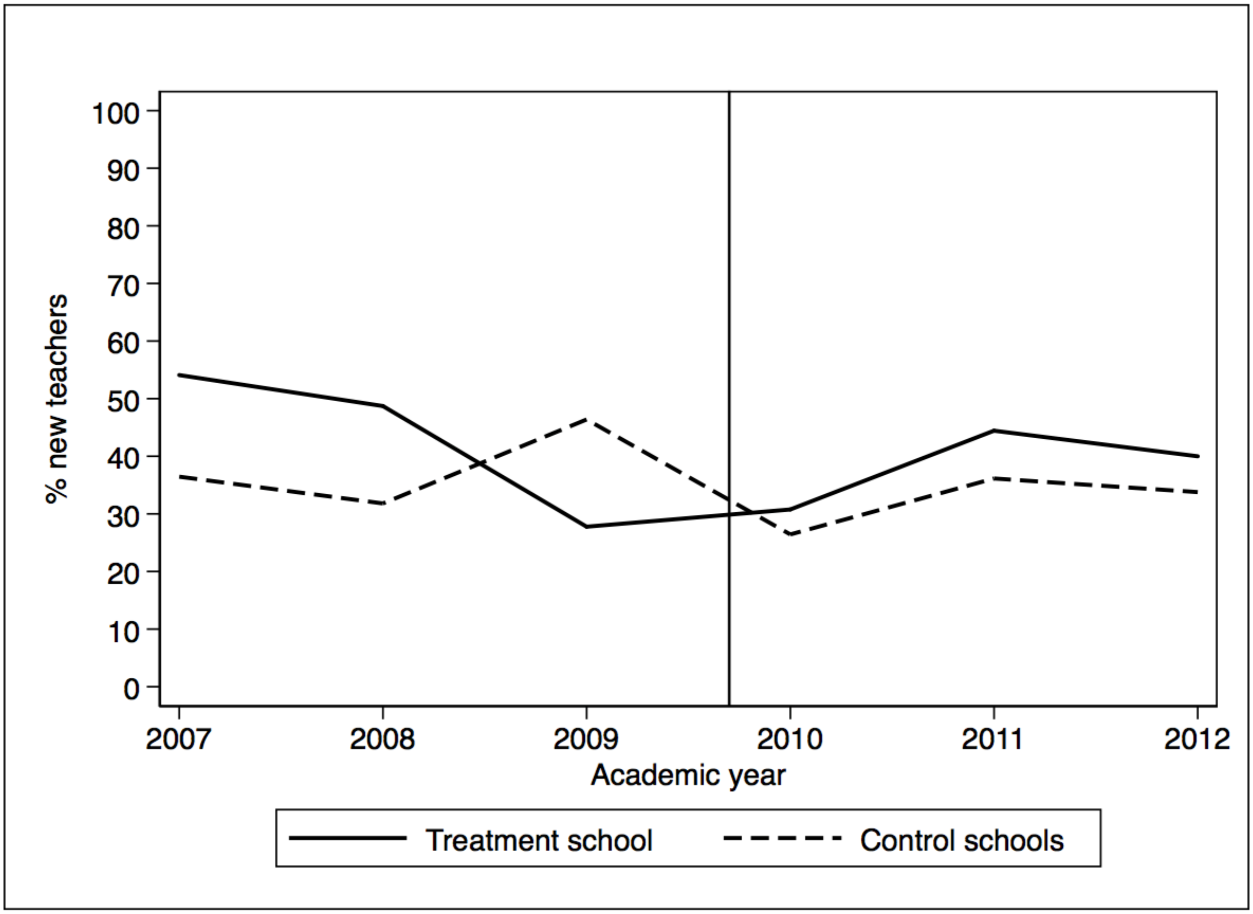

Figure 14 Proportion of new instructors at Anon U by treatment status and academic year

Table 1 Descriptive statistics (2007-2012 cohorts)

	Treatment group: Mean (SD)			Comparison group: Mean (SD)			DiD	
	Before	After	diff.	Before	After	diff.	Pooled	2010
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GPA	3.02	2.79	-0.23***	2.89	2.87	-0.02	-0.22**	-0.49***
	(0.81)	(0.76)	(0.06)	(0.93)	(0.81)	(0.04)	(0.07)	(0.11)
Failed course credits	5.13	10.09	4.96***	8.13	9.09	0.96	4.00***	7.31***
	(9.71)	(12.72)	(0.85)	(13.78)	(13.19)	(0.58)	(1.12)	(1.54)
Graduated	0.76	0.73	-0.03	0.77	0.80	0.03	-0.07~	-0.12*
	(0.43)	(0.45)	(0.03)	(0.42)	(0.40)	(0.02)	(0.04)	(0.05)
Female	0.38	0.39	0.01	0.47	0.51	0.04	-0.02	-0.05
	(0.48)	(0.48)	(0.04)	(0.50)	(0.50)	(0.02)	(0.04)	(0.06)
Rural	0.14	0.19	0.05	0.13	0.13	0.002	0.05	0.03
	(0.35)	(0.40)	(0.03)	(0.34)	(0.34)	(0.01)	(0.03)	(0.04)
Local ethnicity	0.89	0.92	0.03	0.91	0.92	0.01	0.01	0.04
	(0.31)	(0.27)	(0.02)	(0.29)	(0.27)	(0.01)	(0.02)	(0.03)
Local language high school	0.46	0.58	0.11*	0.40	0.35	-0.05*	0.16***	0.04
	(0.50)	(0.49)	(0.04)	(0.49)	(0.48)	(0.02)	(0.04)	(0.06)
School-leaving exam score	94.90	89.72	-5.18***	92.06	85.98	-6.09***	0.91	-0.93
	(13.67)	(10.02)	(0.89)	(15.59)	(12.54)	(0.61)	(1.18)	(1.71)
State grant recipient	0.60	0.62	0.02	0.32	0.27	-0.05*	0.07~	-0.06
	(0.49)	(0.49)	(0.04)	(0.47)	(0.45)	(0.02)	(0.04)	(0.06)
Anon U grant recipient	0.10	0.06	-0.03	0.02	0.04	0.02	-0.05*	-0.01
	(0.30)	(0.24)	(0.02)	(0.16)	(0.19)	(0.01)	(0.02)	(0.02)
Self-supported student	0.30	0.31	0.01	0.66	0.69	0.03	-0.02	0.07
	(0.46)	(0.47)	(0.03)	(0.48)	(0.46)	(0.02)	(0.04)	(0.06)
Transferred in	0	0.003	0.003	0.01	0.05	0.04***	-0.03**	-0.02
	(0)	(0.5)	(0.003)	(0.09)	(0.21)	(0.01)	(0.01)	(0.01)
Observations	352	359		1,177	997		2,884	2,009

˜p<0.10, *p<0.05, **p<0.01, ***p<0.001

Note. Column 7 presents the difference-in-differences estimates on each of the baseline characteristics using the full sample. Column 8 presents difference-in-differences estimates using only 2010 as the post-treatment period. The difference-in-differences estimates are based on the equation 1 specification but do not include other covariates or fixed effects.

Table 2 Impacts of the switch to English-only instruction on GPA, number of failed course credits, and probability of graduation at Anon U (2007-2012 cohorts)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	GPA				Number of failed course credits				Probability of graduation			
Post * English	-0.22**	-0.26***	-0.25***	-0.13*	4.00***	4.28***	4.19***	3.03**	-0.07~	-0.09*	-0.10**	-0.07~
	(0.07)	(0.07)	(0.07)	(0.07)	(1.12)	(1.05)	(1.06)	(1.05)	(0.04)	(0.04)	(0.04)	(0.04)
Post	-0.02	0.09**	0.07~	0.46***	0.96~	-0.28	0.04	-2.46**	0.03~	0.06**	0.07***	0.14***
	(0.04)	(0.03)	(0.04)	(0.06)	(0.56)	(0.53)	(0.55)	(0.92)	(0.02)	(0.02)	(0.02)	(0.03)
English	0.13*	-0.01	-0.13	-0.22	-3.00***	-1.25~	1.87	2.82	-0.00	-0.03	0.13	0.10
	(0.05)	(0.05)	(0.17)	(0.16)	(0.79)	(0.76)	(2.59)	(2.57)	(0.03)	(0.03)	(0.09)	(0.09)
Observations	2,884	2,884	2,884	2,884	2,884	2,884	2,884	2,884	2,884	2,884	2,884	2,884
R-squared	0.00	0.19	0.20	0.25	0.01	0.16	0.17	0.20	0.00	0.06	0.07	0.09
Student characteristics	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Major FE	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Cohort FE	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes

~p<0.10, *p<0.05, **p<0.01, ***p<0.001

Note. The parameter estimates for Post * English show the impact of the switch to English-only instruction on the outcomes. Student characteristics include student gender, locality students come from, ethnicity, language of instruction at high school, school-leaving exam score, financial aid status at entry to Anon U, indicator for whether a student took a year off during studies at Anon U, indicator for whether a student transferred in to Anon U from another institution, and number of years at Anon U. Standard errors are clustered at major-cohort level.

Table 3 Sensitivity of results to choice of comparison schools (2007-2012 cohorts)

Outcome	All other schools	School of Engineering	School of Chemistry	School of Business
GPA	-0.13*	-0.11~	-0.11	-0.13˜
	(0.07)	(0.06)	(0.10)	(0.07)
Number of failed course credits	3.03**	1.49~	3.30*	3.64**
	(1.05)	(0.86)	(1.93)	(1.19)
Probability of graduation	-0.07~	-0.02	-0.17***	-0.08*
	(0.04)	(0.02)	(0.04)	(0.03)
Observations	2,884	1,499	900	1,907

˜p<0.10, *p<0.05, **p<0.01, ***p<0.001

Note. All regressions include student characteristics and fixed effects for college major and cohort. Student characteristics include student gender, locality students come from, ethnicity, language of instruction at high school, school-leaving exam score, financial aid status at entry to Anon U, indicator for whether a student took a year off during studies at Anon U, indicator for whether a student transferred in to Anon U from another institution, and number of years at Anon U. Standard errors are clustered at major-cohort level.

Table 4 Fading out of the treatment effect across the post-treatment period

	Post-treatment period used in estimating the models			
	All post-treatment years	First post-treatment year only	Second post-treatment year only	Third post-treatment year only
Outcome	2010-2012	2010	2011	2012
GPA	-0.13* (0.07)	-0.32*** (0.06)	0.02 (0.06)	-0.10 (0.08)
Number of failed course credits	3.03** (1.05)	5.66*** (1.33)	-0.20 (1.10)	3.34* (1.49)
Probability of graduation	-0.07~ (0.04)	-0.09* (0.03)	-0.05 (0.03)	-0.10 (0.06)
Observations	2,884	2,008	2,001	1,931

˜p<0.10, *p<0.05, **p<0.01, ***p<0.001

Note. All regressions include student characteristics and fixed effects for college major and cohort. Student characteristics include student gender, locality students come from, ethnicity, language of instruction at high school, financial aid status at entry to Anon U, indicator for whether a student took a year off during studies at Anon U, indicator for whether a student transferred in to Anon U from another institution, and number of years at Anon U. Standard errors are clustered at major-cohort level.

Table 5 First semester grades before and after the switch to English-only instruction in the School of Computer Science

	Treatment group: Mean (SD)			Comparison group: Mean (SD)			DiD	
	Before	After	diff.	Before	After	diff.	Pooled	2010
Course:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math	72.14	67.35	-4.79**	58.69	55.40	-3.29**	-1.50	4.17
	(1.02)	(1.12)	(1.51)	(0.69)	(0.78)	(1.04)	(2.07)	(2.89)
Physics	73.23	72.10	-1.13	68.93	65.82	-3.10*	1.98	6.95*
	(0.96)	(0.95)	(1.36)	(0.91)	(0.92)	(1.30)	(1.95)	(2.77)
Language	82.34	81.62	-0.72	79.57	80.17	0.59	-1.31	-2.67
	(0.81)	(0.82)	(1.15)	(0.56)	(0.50)	(0.75)	(1.54)	(2.21)
History	80.49	80.61	0.12	73.10	79.64	6.54***	-6.43**	-5.79~
	(1.07)	(0.88)	(1.39)	(0.79)	(0.55)	(1.01)	(1.94)	(2.98)
Observations	352	359		1,177	997		2,884	2,009

~p<0.10, *p<0.05, **p<0.01, ***p<0.001

Note. All difference-in-differences regressions include student characteristics and fixed effects for college major and cohort. Student characteristics include student gender, locality students come from, ethnicity, language of instruction at high school, financial aid status at entry to Anon U, indicator for whether a student took a year off during studies at Anon U, indicator for whether a student transferred in to Anon U from another institution, and number of years at Anon U. Standard errors of difference-in-differences estimates are clustered at major-cohort level.

Table 6 Proportion of students by English proficiency at entry by cohort

	Cohort		
English proficiency at entry:	2010	2011	2012
	A. Treatment school		
Elementary / pre-intermediate	46%	46%	68%
Intermediate	45%	43%	16%
Upper-intermediate / advanced	8%	11%	16%
Observations	120	125	118
	B. Comparison schools		
Elementary / pre-intermediate	62%	64%	57%
Intermediate	30%	27%	20%
Upper-intermediate / advanced	8%	9%	23%
Observations	371	387	268

Table 7 The relationship between instructors' years of experience at Anon U and grades of students in the School of Computer Science

	OLS	Quantile regression estimates				
		.10	.25	.50	.75	.90
Years of experience at Anon U	0.04	-0.59~	-0.11	0.22~	0.23**	0.01
	(0.18)	(0.30)	(0.19)	(0.12)	(0.09)	(0.06)
Constant	52.32***	26.07***	33.07***	56.09***	77.62***	90.88***
	(8.41)	(4.71)	(2.91)	(1.92)	(1.36)	(0.99)
Observations	12,426	12,426	12,426	12,426	12,426	12,426
R-squared	0.08					

˜p<0.10, *p<0.05, **p<0.01, ***p<0.001

Note. All regressions include student characteristics and fixed effects for college major and cohort. Student characteristics include student gender, locality students come from, ethnicity, language of instruction at high school, financial aid status at entry to Anon U, indicator for whether a student took a year off during studies at Anon U, indicator for whether a student transferred in to Anon U from another institution, and number of years at Anon U. OLS standard errors are clustered at major-cohort level. Unit of analysis is a grade received by a student in the courses taken during the studies at Anon U. The scale for course grades in 0-100.

Table 8 Instructors' years of experience at Anon U in courses with largest proportions of "Fail" grades

Cohort	Course (language of instruction)	% of "Fail" grades among students who took the course	Instructors' average years at Anon U
2009	Calculus 2 (Russian)	17%	2
	Probability theory and statistics (Russian)	15%	2
	English for Professional Purposes (English)	9%	4
2010	Calculus 2 (English)	24%	2
	Probability theory and statistics (English)	16%	3.5
	Basics of circuit theory (English)	10%	2
2011	Algorithms and data structures (English)	21%	2.5
	Databases (English)	20%	4
	Probability theory and statistics (English)	12%	3.5
2012	Differential equations (English)	17%	5
	Principles of economics (English)	15%	6
	Algorithms and data structures (English)	13%	3

3.0 The chaperone effect in social science publications

3.1 Introduction

The term “chaperone effect” first appeared in a paper about academic publishing in the life and physical sciences written by Sekara et al. (2018). The chaperone effect denotes the phenomenon when early-career researchers who publish as non-leading co-authors with an experienced author tend to be more successful later in their careers. Compared to authors with no prior publication experience, such chaperoned authors are more likely to publish later as leading authors and their papers receive more peer recognition in the form of citations. Furthermore, Sekara et al. (2018) document that the chaperone effect is strong and rising over time in the fields of physics, chemistry, medicine, biology, and in leading interdisciplinary science journals.

The concept of the chaperone effect in academic publishing is particularly relevant for the social sciences. Quantifying the chaperone effect in the social sciences reveals access barriers in the publishing markets for early-career researchers. Examining the chaperone effect will also inform the popular and policy debates about how lack of access to academic publishing is linked to such persistent issues in social sciences research training as notoriously long completion times for Ph.D. programs (National Science Foundation, 2018; Stock, Siegfried, & Finegan, 2011), high levels of stress and depression reported by graduate students and early-career scholars (Barreira, Basilico, & Bolotnyy, 2018; T. M. Evans, Bira, Gastelum, Weiss, & Vanderford, 2018; Levecque, Anseel, de Beuckelaer, van der Heyden, & Gisle, 2017), and concerns about the job market worth of the social sciences PhD degrees (Cassuto, 2015; Pannapacker, 2013; Weisbuch & Cassuto, 2016; Zahneis, 2019).

Therefore, this study investigates and quantifies the chaperone effect in the social sciences academic publications. The research utilizes Python's text parsing tools to the Web of Science data on 603,468 articles published in 628 journals between 1945 and 2018 in economics, education, political science, psychology, and sociology. Findings show that there is only a limited presence of the chaperone effect in these social science disciplines. Specifically, new first authors with no prior publication experience substantially outnumber chaperoned first authors who have previously co-published with senior scholars. At the same time, chaperoned first authors' papers receive more citations than new first authors' papers, especially in higher-ranked journals in education, political science, and psychology. Encouragingly, the results show no gender gaps in the chaperone effect in all five disciplines examined. That is, findings demonstrate that social sciences academic publishing overall is accessible for early-career scholars with no previous co-authored publication experience and irrespective of gender.

The study contributes to the growing body of bibliometric literature investigating how research production works and exploring the distinct role of collaborations in new knowledge creation. It provides a comprehensive analysis of the chaperone effect in the social sciences academic publishing by applying novel text analysis techniques to a large bibliometric dataset covering major social science disciplines. The findings may be encouraging news to aspiring and early-career social scientists. Further, given that the findings indicate the accessibility of academic publishing for junior social scientists, aspects of research training other than publishing research in journals might account for the heatedly debated long Ph.D. completion times, high levels of stress, and dire labor market prospects of junior social science researchers.

The rest of the paper is structured as follows: section 3.2 presents the background and context of the study, section 3.3 explains the data, construction of key variables, and methods, section 3.4 describes the results, section 3.5 discusses the findings and concludes.

3.2 Background and context

In the life and physical sciences, the chaperone effect varies by discipline but is substantial and increasing in most fields of study. Sekara et al. (2018) showed that early-career researchers who publish as non-leading authors in a team led by an experienced author tend to be more successful later in their careers. Using publication data for 1965-2013 from the Web of Science database, Sekara et al. (2018) document that the role of publication experience through co-authoring with senior scholars is strong and rising over time in the fields of physics, chemistry, medicine, biology, and in leading interdisciplinary journals. The field of mathematics was found to be an exception to this trend. The findings of Sekara et al. (2018) highlight that, except for mathematics, the hard sciences increasingly expect early-career scholars to have prior publication experience. Perhaps the chaperone effect reflects that much of the research is expensive to conduct, and funding has gotten tighter over time in life and physical sciences (Carter, Berndt, DiMasi, & Trusheim, 2016; Collier, 2009; Wu, 2020).

Nevertheless, it is not clear whether the chaperone effect generalizes to other fields of inquiry, specifically to the social sciences. First, the magnitude of the chaperone effect may be less pronounced in the social sciences academic publishing due to a somewhat lower prevalence of equipment-intensive or grant-funded research. Second, it is possible that social sciences are doing better than expected in striving to be meritocratic and in applying their evaluative criteria as rationally and objectively as possible, in line with the description of academia in Cole and Cole (1974).

At the same time, the chaperone effect may be similarly present in the social sciences because many important trends in modern science generalize to most scientific disciplines. For example, across all disciplines, the number of research papers written by multiple authors exceeds the number of single-authored papers and keeps rising (Waltman, 2012; Wuchty, Jones, & Uzzi, 2007), research teams are growing in size (Milojević, 2014) and increasingly include members from more than one university (Jones, Wuchty, & Uzzi, 2008) or geographic region (Shrum, Genuth, & Chompalov, 2007). If prior co-publication experience helps junior scholars through additional mentoring in the process of research publishing, a sizeable literature has demonstrated that mentoring appear to work according to similar underlying principles irrespective of the profession (Chao, Walz, & Gardner, 1992; Higgins & Kram, 2001).

Further, the chaperone effect may be salient in the social sciences due to the uneven availability of mentoring in academia's highly stratified social system. Co-authorship opportunities are an essential factor for future research success (van Dijk, Manor, & Carey, 2014; Zuckerman, 1967), and senior co-authors' supervision is considerably beneficial for innovative research (Packalen & Bhattacharya, 2015). However, junior scholars have differential access to informal mentoring in the form of co-authorship opportunities, and academic pedigree remains strongly predictive of research productivity (Way, Morgan, Clauset, & Larremore, 2017).

This paper investigates the chaperone effect in the social sciences and seeks to answer the following research questions:

1) How prevalent is the chaperone effect in the social sciences?

2) How does the chaperone effect vary by journal rank and authors' gender in the social sciences?

Answering these research questions will help reveal access barriers in the publishing markets for early-career researchers and show whether access to publishing is heterogeneous across journals of varying rank and by gender. Examining the chaperone effect and its heterogeneity will inform the debates about academic publishing and state of affairs in academia more broadly.

3.3 Data, construction of key variables, and methods

Data

The source of the data is the Web of Science database (www.webofknowledge.com). The database covers research publications across all major scientific domains since 1945. For every publication, the Web of Science database provides detailed data including the title of the manuscript, author names in the order listed in the publication, journal name, volume, issue and page numbers, date of publication, and abstract.

The dataset is based on the lists of top 200 journals in economics, education, political science, psychology, and sociology as ranked by the Scimago (https://www.scimagojr.com) journal ranking. An examination of these top 200 lists showed that sometimes, a list of journals in one discipline might contain journals that more closely align with another discipline. Therefore, we cleaned these top 200 lists using Web of Science disciplinary categorization of journals. Tables A1-A5 of the online appendix present the final list of 628 journals in the study sample including

120 economics journals, 165 education journals, 133 political science journals, 166 psychology journals, 44 sociology journals.

From the publication data for these journals in the Web of Science database, we retained only research articles and literature reviews for analysis. We excluded editorial materials, letters, news items, book reviews, corrections, biographical items, notes, meeting abstracts, retracted publications, discussions, software and hardware reviews, conference proceedings, comments on previously published articles, replies to comments, and announcements. The final analytic sample includes 633,669 publications.

Construction of key variables

We parsed the Web of Science data using text analysis and data science tools of Python version 3.7 to obtain key variables for the study (detailed codebook is in Table A6 of the online appendix). We considered publications in all journals of one discipline as a single publishing pool, and conducted data parsing at the discipline level[7].

Construction of key variables: First author categories

First, following Sekara et al. (2018), we used the ordering of authors and the timing of publications to construct first author category indicators. In case papers were written by a single author, we counted such papers as written by first-authors. For each paper in the dataset, we categorized the first author as *new* if they have never published before, as *established* if they have published earlier as a first-author, and as *chaperoned* if they have published earlier as a non-first

[7] This is the key difference of this study from Sekara et al. (2018) who conduct all analyses at journal level. Due to small number of publications per journal in the social sciences, this study had to be conducted at a discipline level.

author. For example, as shown in Table 1, in 2005, T. Domina was the first author of his first paper in the journal *Sociology of Education*, so we categorized him in this paper as *new*. In his subsequent first-author publications, we classified T. Domina as *established*, as in his 2014 paper in *Teachers College Record*. In 2016, his non-first co-author E.K. Penner published a first-author paper in the *Journal of Research for Educational Effectiveness* and we categorized him as *chaperoned*.

The analysis could potentially be affected if the author lists do not reflect seniority roles, like, for example, in economics, where the field convention prefers alphabetically-ordered author lists (Waltman, 2012). Therefore in our calculations, we account for the differences in authorship ordering conventions across fields of study by creating two versions of the dataset. We re-arrange each paper's authors in the alphabetic order in the first version of the dataset and a randomly generated order in the second version of the dataset. Then we perform calculations on these two amended versions of the data and compare estimates against those obtained using the actual data.

Construction of key variables: Gender indicator

To examine heterogeneity in the role of publication experience across gender, we generate an indicator for first authors' gender using Python's name-to-gender inference service *gender-guesser*. Python's *gender-guesser* has open-source code and was developed using a publicly available high-quality dataset that has been manually checked by native speakers from multiple countries (see Michael (2007)). Further, it is the only freely available name-to-gender inference package. A recent comparative analysis of currently available gender inference services has found Python's *gender-guesser* to have the lowest misclassification rate and the smallest software-introduced gender bias without parameter-tuning for the entire dataset (Santamaría & Mihaljević, 2018).

There are three main limitations of utilizing gender inference software packages with the Web of Science data. First, up to 2007, the Web of Science data provides initials instead of full first names for most journals. In other words, one can only infer gender from first names for papers published in 2007 and later with the exception of a few journals that have full first names data before 2007. Second, overall, assigning gender categories based on first names is an error-prone procedure because cultural, geographic, and historical traditions influence the attachment of first names across gender (Santamaría & Mihaljević, 2018; Torvik & Smalheiser, 2009). Third, gender inference services typically categorize authors non-inclusively as either male or female, thus marginalizing people who do not identify as either (Matias, 2014; Santamaría & Mihaljević, 2018). However, obtaining authors' full first names for articles published prior to 2007 or deploying more precise approaches to determining gender, such as collecting self-identification data, are not possible when conducting bibliometric studies. Therefore, this study uses Python's *gender-guesser* while acknowledging the limitations of automatically inferring gender from first names in the Web of Science data and offering to view the estimates across gender as imperfect and exploratory.

Methods

The analysis includes three steps. First, to compare the proportions of papers written by established authors to the proportions of papers written by authors with limited or no prior experience in the social sciences, we calculate discipline-level shares of articles written by new, chaperoned, and established first authors. To account for time-variant changes in the share of people "vying" for space in journals, we conduct graphical analysis of proportions of papers written by first author types over time.

Second, to examine the relationship between authors' publication experience and the number of times their papers get cited, we estimate OLS regressions for each discipline using equation 1 below:

$$citation5_i = \beta_0 + \beta_1 established_i + \beta_2 chaperoned_i + \beta \boldsymbol{X}_i + \varepsilon_i \quad (1)$$

In equation 1, $citation5_i$ is the number of times publication *i* in has been cited within five years since it was published, $established_i$ is a dummy equal to 1 if the publications' first author is an established author, and $chaperoned_i$ is a dummy equal to 1 if the publication's first author is a chaperoned one, i.e. has previously published as a non-first author. The omitted category is new_i, a dummy equal to 1 if the publication's first author is a new author. A vector of covariates **X** includes the journal's Scimago rank position and number of authors for each paper. The coefficients associated with $chaperoned_i$ and $established_i$ indicate how the five-year citation count differs for publications written by established and chaperoned authors relative to publications written by new authors within a given field.

Third, to examine heterogeneity in the role of publication experience, we perform regression analyses across journal rank and gender categories. Specifically, we examine heterogeneity across journal rank by estimating the equation 1 above in the subsamples of journals by their rank (journals ranked 1-10, 11-50, 51-100, and 101-200).

We explore heterogeneity across gender by estimating equation 2:

$$citation5_i = \beta_0 + \beta_1 established_i * gender_i + \beta_2 chaperoned_i * gender_i$$
$$+ \beta_3 gender_i + \beta_4 established_i + \beta_5 chaperoned_i + \beta \boldsymbol{X}_i + \varepsilon_i \quad (2)$$

which is a version of equation 1 that includes indicators for gender and author types as well as interaction terms of the gender and author type indicators for each published paper.

3.4 Results

Overview of the social sciences publications: output, team sizes, and citation count

The descriptive statistics presented in Table 2 show appreciable variation in the volume of each discipline's research output. The largest field is psychology with 271,604 articles in 166 journals published starting from 1945. The smallest field in the analytic sample is sociology with 44 journals and 30,849 papers published since 1966.

As column 4 of Table 2 shows, psychology also leads in average author team size with approximately 3.4 authors per paper. Education is the second field in terms of author team size with about 2.6 authors per paper, economics is the third with approximately 2 people per paper, sociology is the fourth with about 1.8 authors per paper, and political science has the fewest 1.5 people per paper. Notably, the longest lists of authors for a single paper range from 25 in sociology to 163 in psychology and indicate that researchers undertake complex, large-scale studies requiring collaboration in mega-sized teams.

The average five-year citation count in psychology and economics roughly corresponds to the overall volume of the discipline's research output. Papers in psychology, the largest field, enjoy an average of 17.9 citations and papers in economics, the second-largest field, receive about 11.8 citations within a five year period. In the field of education, papers get approximately 9.4 citations on average, which is relatively low given that education is the third largest field by its research output. Papers in two smallest fields get a similar or a higher number of citations compared to education research papers, namely, 10.7 citations in sociology and 7.9 citations in political science. Such high citation count relative to the research output volume likely reflects the interdisciplinary nature of education research. In other words, education scholars tend to include both within-discipline and across-discipline references in their papers which likely drives down the average number of citations for education scholars and drives up citations for other social sciences.

Notably, in all fields under study, considerable variation exists in the number of citations a published article typically receives. The ranges of five-year citations start at zero in all disciplines and reach 368 in sociology, 536 in education, 737 in political science, 1278 in economics, and as much as 3010 in psychology.

Shares of papers written by new versus chaperoned first authors

Estimates in column 1 of Table 3 show that the number of papers written by chaperoned first authors is overall low and lower than the number of papers by new first authors in all five disciplines indicating that the chaperone effect is practically non-existent in the social sciences. The share of articles by new first authors is greater than the share of articles by chaperoned authors in all the disciplines. Further, the share of papers by chaperoned first authors is lowest at 4% in political science and highest at 12% in psychology. In between these extremes, publications by chaperoned first authors comprise 6% of all publications in sociology, 7% in economics, and 10% in education.

Two other columns in Table 3 present calculations of shares of published articles by the author type after we have re-arranged the author lists for each paper in the alphabetical order (column 1) and a random order (column 2). The main conclusion made from these estimates is that the shares of published papers written by new, established, and chaperoned first authors remain similar after author lists are re-arranged. This exercise implies that the prevalence of the chaperone effect in a field does not change if either alphabetical or random author ordering conventions are adopted.

Figures 2-6 display the dynamics of author type representation over time and reveal that since the early 1990s, all disciplines have retained stable low shares of published articles written by chaperoned authors. The steady low shares of publications by chaperoned first authors over time across the five fields confirm that the chaperone effect is not strengthening over time in contrast to life and physical sciences, as shown by Sekara et al. (2018). It appears that such stability in proportions of papers written by new, chaperoned, and established authors over time reflects the stable low prevalence of equipment-intensive or grant-funded research in the social sciences, which in turn allows new authors to enter the publishing field with fewer cost-related barriers and prevents the growth of a need to be chaperoned.

Relationship between published papers' first author type and citation count

The OLS regression results in Table 4 demonstrate the existence of the chaperone effect in all five disciplines in contrast to what the low share of chaperoned authors imply. Estimates in Table 4 show that papers written by chaperoned first authors have higher citation count than papers written by new authors in all five disciplines[8]. In other words, in the social sciences, a new first author can craft a publishable paper, but these papers tend to be less impactful citation-count-wise than papers written by chaperoned first authors.

Heterogeneity across journal rank

The descriptive analysis of shares of published papers written by new, chaperoned, and established first authors across journal positions in the Scimago rank showed no statistically

[8] Given that the citation count variable is highly skewed, the OLS analysis was repeated using its log-transformed version. Estimates remain substantively similar, as shown in Table 21 in Appendix A.

significant heterogeneity by journal rank. New, established, and chaperoned authors appear capable of getting their research published in journals of all ranks, showing that it is not necessary to be chaperoned in the social sciences.

However, some variation exists in the relationship between articles' citation count and first author type depending on journal ranking. As Table 5 displays, in education and political science, the chaperoned first author premiums are the largest in the top journals and smaller in lower-ranked journals. In these two fields, chaperoned first authors' prior publication experience enables them to produce impactful research publishable in journals of varying rank. In psychology, the chaperone premium is largest in highest-ranked journals and smaller in lower-ranked journals. Finally, in sociology, the chaperone premium exists only in medium- or low-ranked journals. In other words, in education, political science, and psychology, the chaperone citation premia are greater in higher-ranked journals. These findings imply that the association of article citation counts with their first authors' prior publication experience co-authored with senior scholars is uneven across disciplines and journal ranks. That the readership of published research more actively cites papers written by chaperoned first authors in top-tier journals in education, political science, and psychology reveals the importance of social capital and networks in these fields. Given the crucial role of citation impacts for academic careers, researchers might account for their fields' inclination to notice and further promote chaperoned authors' work in planning their publications.

Heterogeneity across gender

As Table 6 shows, we were able to assign either "female" or "male" categories to about 80% of 378,680 first authors whose full names were available in the Web of Science database.

There are more men than women in economics, political science, and sociology; however, women outnumber men in education research while both genders are equally represented in psychology.

Estimates in Table 7 show that despite such variation in gender representation, laudably, all disciplines, except for economics, have similar proportions of published papers written by men and women within each author category. In other words, there are no gender gaps in proportions of papers grouped by first author type in the fields of economics, education, political science, psychology, and sociology.

OLS estimates of the relationship between published articles' first author categories and citation count presented in Table 8 support the gender balance conclusion. In all the disciplines, there are no statistically significant differences across gender in the role publication experience plays in the citation count of published research. Remarkably, discipline-level gender imbalances in representation do not translate into differential citation count for publications written by new, established, and chaperoned first authors across gender. Citations-by-gender-wise, there are no signs of gender gaps in terms of peer recognition based on prior publication experience in any of the five disciplines.

3.5 Discussion and conclusions

This study showed a limited presence of the chaperone effect in economics, education, political science, psychology, and sociology. The shares of publications by chaperoned first authors are low and much lower than the shares of publications by new authors in all disciplines. At the same time, estimates detected a statistically significant citation count premium for papers written by chaperoned authors in the social sciences, similar to the "chaperone effect" detected in the life and physical sciences by Sekara et al. (2018). Analyses of heterogeneity across journal rank showed that the chaperoned citation premia are greater in higher-ranked journals in education, political science, and psychology.

Reassuringly for aspiring and junior researchers, findings demonstrate that publication experience obtained in prior research with other leading authors is not very common among first authors of published papers though prior publication experience is associated with higher number of citations. Even though the broader intended audience of published research appears to prefer to cite papers written by junior researchers who are associated with experienced scholars through prior co-authored work, it appears that social sciences are accessible and meritocratic enough as argued by Cole and Cole (1974).

Encouragingly, estimates show no gender gaps in the chaperone effect in the social sciences. The shares of papers authored by women and men are similar within each category of papers grouped by first author type in all disciplines except economics, where the gender ratio tilts more strongly towards men in papers written by established authors compared to the papers by new and chaperoned authors. Further, articles' citation count by first author type is the same across gender in all fields. It appears that irrespective of the overall gender ratio in the discipline, the gender of a published manuscript's first author does not matter much for peer recognition. This finding does not support earlier research about the lack of support from the scientific community of female researchers' work (Zuckerman, Cole, & Bruer, 1991). Notably, however, this study only examined published papers, and if there are differences by author's gender in what manuscripts get accepted for publication, we are not able to capture such variation.

There are two main limitations of this study. The first limitation is that the analysis does not account for the fact that some manuscripts never make it to publication. It is possible to arrive at different conclusions if one had data on the written, submitted for review, and rejected manuscripts. However, such data are not available at present. For future research, we suggest examining this selection into the published authors' pool by collecting data on all dissertations defended and made publicly available and check whether these authors appear in bibliometric databases. Another interesting follow-up study would be to examine publication frequency and citation count of all papers' first and non-first authors whose names ever appear in bibliometric databases. The second limitation is that the research design does not address the possible assortative matching of co-authors, whereby more talented authors may be more likely to co-publish with colleagues or established scholars. However, the limited presence of the chaperone effect in the estimates implies that assortative matching does not pose severe risks for the findings.

The paper makes three main contributions. First, through a comprehensive analysis of academic publishing markets in the social sciences, the study demonstrates the accessibility of academic publishing for authors with different prior publication experiences. Second, the study is comparative and examines the chaperone effect across major social science disciplines. Third, it applies novel text parsing techniques to a large bibliometric dataset.

The findings have implications for the policies aimed at training of Ph.D. students and early-career researchers in the social sciences. First, the findings will hopefully serve as encouraging news and motivate aspiring and early-career social scientists to press on with their research endeavors. Debates and policies aiming to reduce the social sciences Ph.D. completion times, improving the mental health of early-career scholars, and increasing the job market worth of social sciences Ph.D. may focus on aspects of research training other than the extreme competitiveness and dependence on senior scholars in academic publishing. Second, given the presence of the chaperone effect citation-wise, the scholarly communities in the social sciences should considered policies to make academic publishing more inclusive. One possible strategy could be including aspiring and junior scholars as co-authors in research projects. In some fields, it is common to not list graduate student assistants as co-authors but rather thank them in a footnote for the assistantship. The results of this study underscore that it might be beneficial for graduate students' research careers if they were included as co-authors.

Table 9 Illustration of first author categories

First author category	Publication details
New	Domina, T. (2005). Leveling the home advantage: Assessing the effectiveness of parental involvement in elementary school. *Sociology of education*, *78*(3), 233-249.
Established	Domina, T., Penner, A. M., Penner, E. K., & Conley, A. (2014). Algebra for All: California's eighth-grade Algebra initiative as constrained curricula. *Teachers College record (1970)*, *116*(8), 1.
Chaperoned	Penner, E. K. (2016). Teaching for all? Teach For America's effects across the distribution of student achievement. *Journal of research on educational effectiveness*, *9*(3), 259-282.

Table 10 The description of the study sample

Field of study	(1) Total number of journals[1]	(2) Total number of articles	(3) Years of publication	(4) Authors per paper	(5) Five-year citations[2]
			range	mean (range)	mean (range)
Economics	120	132,509	1945-2018	2.09 (1-29)	11.82 (0-1278)
Education	165	112,669	1945-2018	2.65 (1-84)	9.38 (0-536)
Political science	133	55,837	1966-2018	1.45 (1-31)	7.86 (0-737)
Psychology	166	271,604	1945-2018	3.43 (1-163)	17.85 (0-3010)
Sociology	44	30,849	1966-2018	1.77 (1-25)	10.71 (0-368)

Notes.
[1] In total, the dataset includes 628 journals with 603,468 articles.
[2] The five-year citations are calculated only for papers published before 2014.

Table 11 The proportions of new, chaperoned, and established first authors by discipline

Discipline / author type	(1) Actual data Mean (SD)	(2) Data with authors ordered alphabetically Mean (SD)	(3) Data with authors ordered randomly Mean (SD)
Economics			
new	0.31 (0.07)	0.31 (0.08)	0.31 (0.08)
chaperoned	0.07 (0.01)	0.06 (0.01)	0.10 (0.01)
established	0.62 (0.07)	0.62 (0.07)	0.59 (0.07)
Education			
new	0.45 (0.08)	0.46 (0.09)	0.46 (0.09)
chaperoned	0.10 (0.01)	0.08 (0.01)	0.13 (0.02)
established	0.45 (0.07)	0.46 (0.08)	0.41 (0.07)
Political science			
new	0.48 (0.09)	0.49 (0.10)	0.49 (0.09)
chaperoned	0.04 (0.01)	0.04 (0.01)	0.05 (0.01)
established	0.48 (0.08)	0.47 (0.08)	0.46 (0.08)
Psychology			
new	0.29 (0.04)	0.28 (0.06)	0.28(0.06)
chaperoned	0.12 (0.01)	0.09 (0.00)	0.16 (0.01)
established	0.59 (0.04)	0.63 (0.06)	0.56 (0.05)
Sociology			
new	0.55 (0.08)	0.57 (0.08)	0.57 (0.08)
chaperoned	0.06 (0.01)	0.06 (0.01)	0.07 (0.01)
established	0.39 (0.07)	0.38 (0.07)	0.36 (0.07)

Table 12 OLS estimates of the relationship between five-year citations and first author type

	(1) Economics	(2) Education	(3) Political science	(4) Psychology	(5) Sociology
Chaperoned	1.34***	2.23***	2.14***	2.84***	2.04***
	(0.27)	(0.20)	(0.40)	(0.23)	(0.48)
Established	1.88***	2.11***	2.68***	2.46***	2.90***
	(0.14)	(0.11)	(0.15)	(0.14)	(0.22)
Constant	11.63***	9.01***	7.21***	18.26***	10.59***
	(0.19)	(0.14)	(0.20)	(0.18)	(0.32)
Observations	95,552	75,489	34,894	198,303	21,896
R-squared	0.08	0.05	0.08	0.04	0.07

Notes. All regressions include each paper's number of authors and journal's Scimago rank as controls. The dataset contains only papers published before 2014. Standard errors in parentheses. Statistical significance is presented using Bonferroni-adjusted p-values. *** p<0.01, ** p<0.05, * p<0.1

Table 13 Journal rank heterogeneity of OLS estimates of the relationship between five-year citations and first author type (discipline-level analysis)

	(1)	(2)	(3)	(4)	(5)
	Pooled sample	Journals ranked 1-10	Journals ranked 11-50	Journals ranked 51-100	Journals ranked 101-200
			A. Economics		
Chaperoned	1.34***	3.50	1.30*	1.59***	0.87***
	(0.27)	(1.43)	(0.76)	(0.39)	(0.27)
Established	1.88***	3.79	3.91***	0.45**	1.16***
	(0.14)	(0.74)	(0.39)	(0.19)	(0.14)
Constant	11.63***	7.13***	16.24***	7.74***	10.55***
	(0.19)	(1.03)	(0.69)	(0.55)	(0.37)
Observations	95,552	12,337	18,006	28,272	36,937
R-squared	0.08	0.04	0.05	0.03	0.05
			B. Education		
Chaperoned	2.23***	4.75**	3.14***	2.11***	1.52***
	(0.20)	(2.21)	(0.49)	(0.35)	(0.21)
Established	2.11***	4.38***	2.95***	1.93***	1.45***
	(0.11)	(1.27)	(0.27)	(0.19)	(0.12)
Constant	9.01***	23.82***	12.92***	9.05***	6.45***
	(0.14)	(1.82)	(0.45)	(0.45)	(0.3)
Observations	75,489	2,057	19,641	22,211	31,580
R-squared	0.05	0.06	0.05	0.03	0.04
			C. Political science		
Chaperoned	2.14***	4.46***	2.28***	0.77	1.14***
	(0.40)	(1.63)	(0.6)	(0.5)	(0.41)
Established	2.68***	4.43***	3.11***	1.87	0.75***
	(0.15)	(0.69)	(0.25)	(0.17)	(0.15)
Constant	7.21***	12.45***	7.24***	1.90***	-0.54
	(0.20)	(1.02)	(0.38)	(0.49)	(0.38)
Observations	34,894	5,570	9,502	13,287	6,535
R-squared	0.08	0.04	0.05	0.02	0.05
			D. Psychology		
Chaperoned	2.84***	10.57***	4.13***	2.78***	1.83***
	(0.23)	(3.16)	(0.49)	(0.43)	(0.23)
Established	2.46***	1.19	2.67***	2.45***	1.89***
	(0.14)	(1.9)	(0.3)	(0.28)	(0.15)
Constant	18.26***	29.89***	18.53***	5.56***	17.85***
	(0.18)	(2.22)	(0.41)	(0.74)	(0.38)
Observations	198,303	8,067	55,284	49,740	85,212
R-squared	0.04	0.04	0.04	0.01	0.02

			E. Sociology		

Chaperoned	2.04***	4.24	2.14**	0.3	1.87
	(0.48)	(2.82)	(1.09)	(1.08)	(0.5)
Established	2.90***	3.84***	1.66***	3.40***	2.80
	(0.22)	(1.35)	(0.53)	(0.46)	(0.23)
Constant	10.59***	-23.84***	22.43***	9.10***	2.52***
	(0.32)	(6.68)	(1.18)	(1.82)	(0.59)
Observations	21,896	1,441	5,796	4,209	10,450
R-squared	0.07	0.05	0.06	0.05	0.04

Notes. All regressions include each paper's number of authors and journal's Scimago rank as controls. The dataset contains only papers published before 2014. Standard errors in parentheses. Statistical significance is presented using Bonferroni-adjusted p-values. *** p<0.01, ** p<0.05, * p<0.1

Table 14 Percentage of first authors whose gender was identified based on their first name

	(1)	(2)	(3)	(4)
	Female	Male	Unknown	Total
Economics	13,241 (16%)	50,376 (62%)	17,808 (22%)	81,425
Education	33,110 (42%)	27,699 (35%)	17,396 (22%)	78,205
Political science	9,523 (23%)	25,853 (62%)	6,131 (15%)	41,507
Psychology	65,381 (41%)	65,370 (41%)	27,456 (17%)	158,207
Sociology	6,837 (35%)	9,343 (48%)	3,165 (16%)	19,345
Total	128,092 (34%)	178,641 (47%)	71,956 (19%)	378,689

Notes. Gender categories were assigned to each author using Python 3.7. The “Unknown” category in column 3 includes authors whose names were classified as androgynous, mostly female, mostly male, or unknown.

Table 15 The proportions of men and women among new, chaperoned, and established first authors

	(1)	(2)
Discipline / author type	Men	Women
Economics		
new	74%	26%
chaperoned	77%	23%
established	82%	18%
Education		
new	47%	53%
chaperoned	48%	52%
established	47%	52%
Political science		
new	73%	27%
chaperoned	72%	28%
established	72%	28%
Psychology		
new	54%	46%
chaperoned	54%	46%
established	54%	46%
Sociology		
new	59%	41%
chaperoned	59%	41%
established	59%	41%

Notes. Each row adds up to 100%. Calculations are presented for first authors whose gender could be identified based on the first name (see table 6 for details.)

Table 16 OLS estimates of the relationship between five-year citations and first author's gender

	(1)	(2)	(3)	(4)	(5)
	Economics	Education	Political science	Psychology	Sociology
Chaperoned * female	-1.42	-0.84	0.29	-1.05	2.17
	(1.34)	(0.73)	(1.36)	(0.88)	(1.78)
Established * female	-0.55	0.26	-0.77	-0.35	0.47
	(0.69)	(0.44)	(0.57)	(0.59)	(0.87)
Female	0.95*	-0.51	0.21	-2.03***	-1.56***
	(0.52)	(0.31)	(0.41)	(0.49)	(0.58)
Chaperoned	0.82	0.04	0.18	0.54	-1.06
	(0.63)	(0.53)	(0.69)	(0.61)	(1.13)
Established	2.15***	-0.23	0.71**	0.47	-0.78
	(0.33)	(0.32)	(0.28)	(0.41)	(0.56)
Constant	17.39***	10.70***	4.74***	21.07***	10.70***
	(0.45)	(0.33)	(0.33)	(0.46)	(0.68)
Observations	34,646	28,932	17,059	67,591	7,989
R-squared	0.08	0.01	0.03	0.00	0.02

Notes. All regressions include each paper's number of authors and journal's Scimago rank as controls. The dataset contains only papers published before 2014 by authors whose first full names are available in the Web of Science database. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

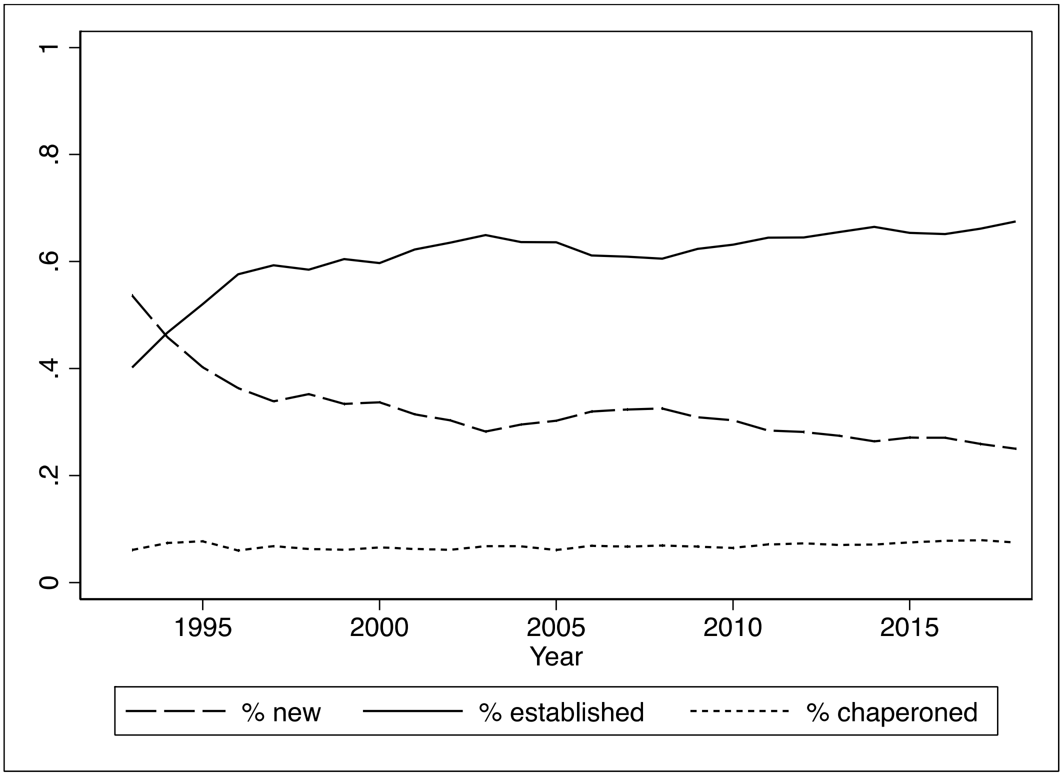

Figure 15 Proportions of new, established, and chaperoned first authors over time in top 120 economics research journals according to Scimago ranking

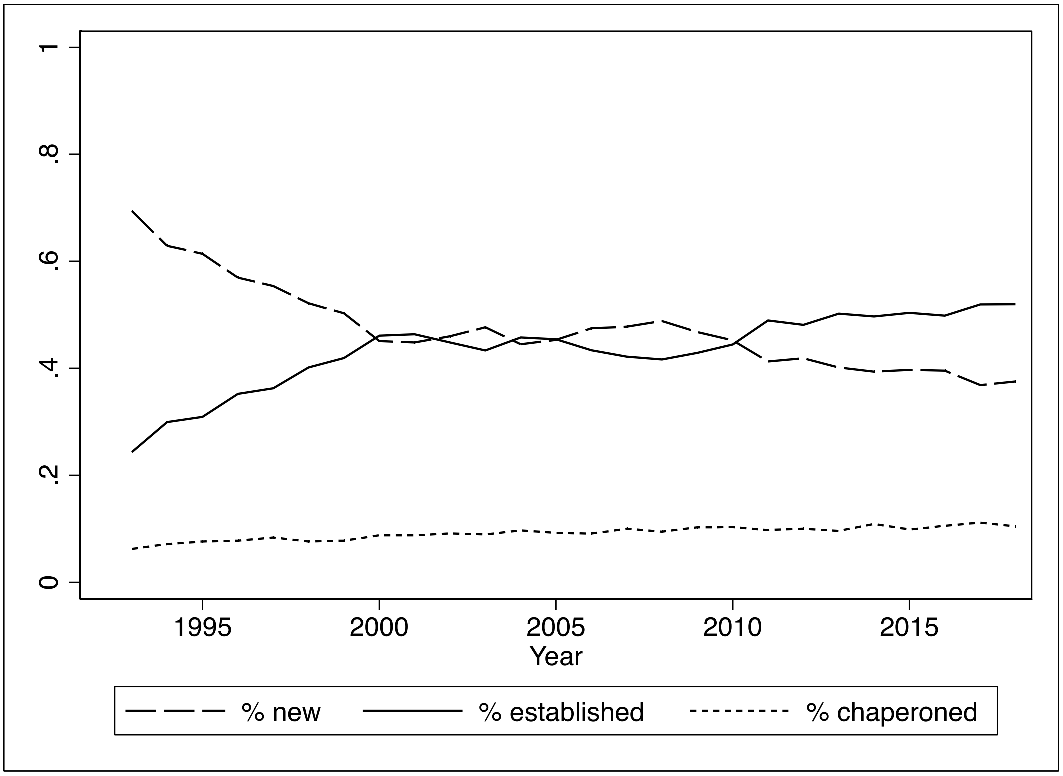

Figure 16 Proportions of new, established, and chaperoned first authors over time in top 165 education research journals according to Scimago ranking

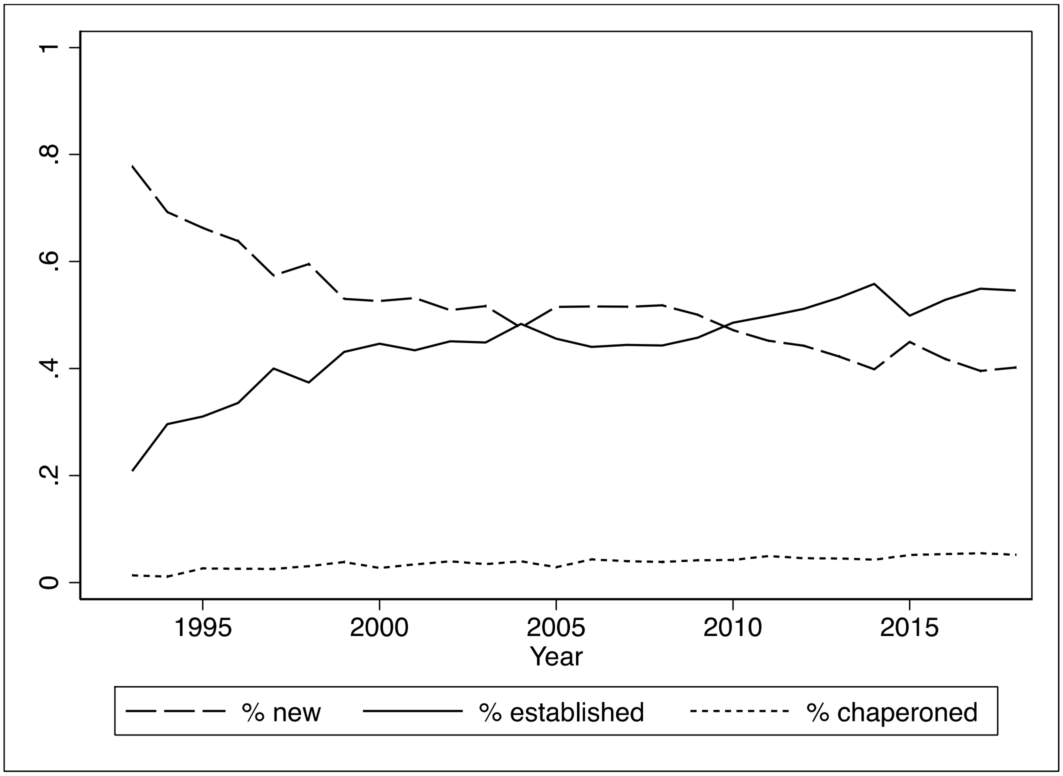

Figure 17 Proportions of new, established, and chaperoned first authors over time in top 133 political science research journals according to Scimago ranking

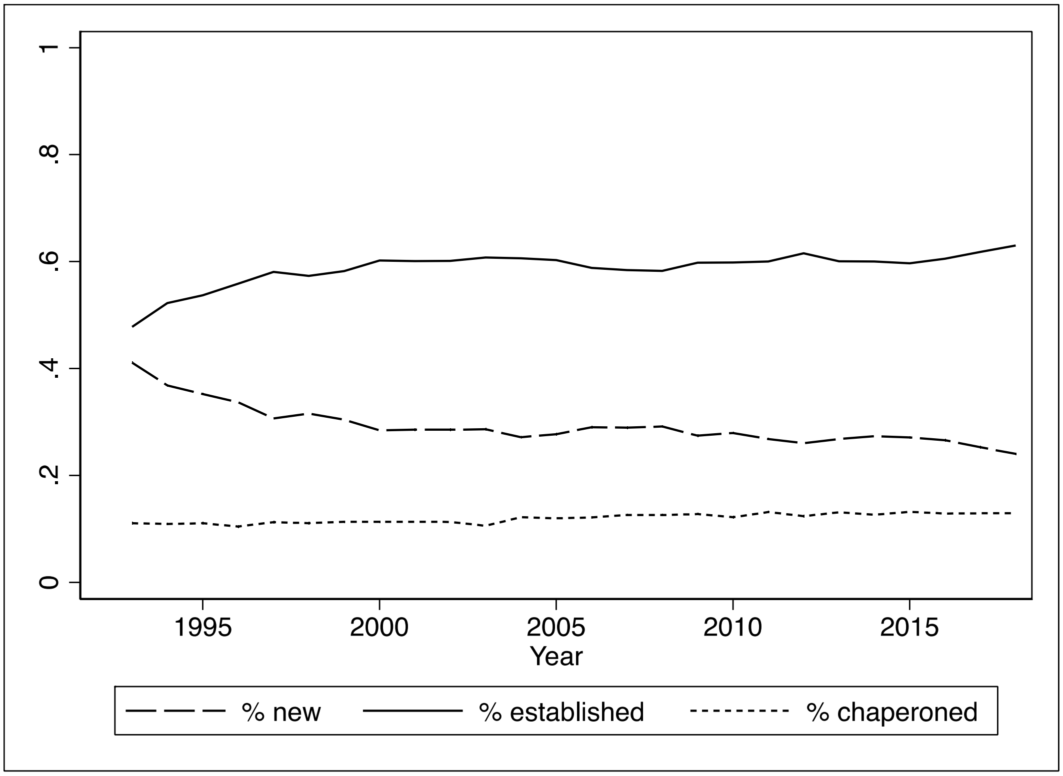

Figure 18 Proportions of new, established, and chaperoned first authors over time in top 166 psychology research journals according to Scimago ranking

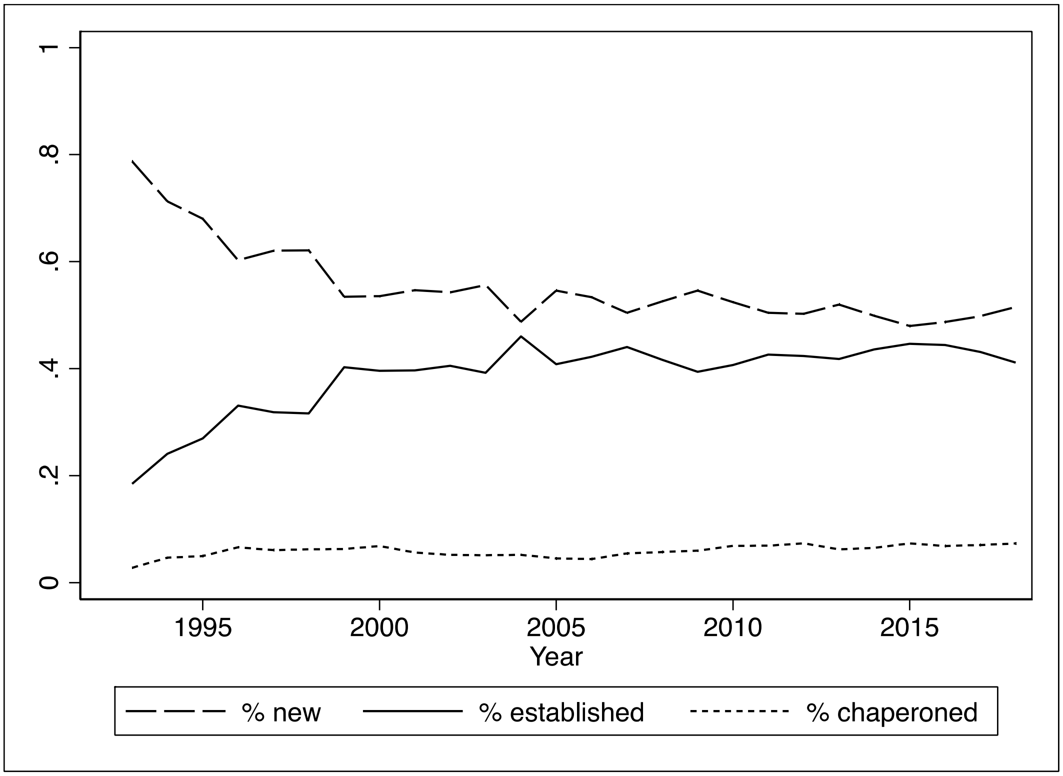

Figure 19 Proportions of new, established, and chaperoned first authors over time in top 44 sociology research journals according to Scimago ranking

4.0 Scaling summer melt prevention with artificially intelligent conversational chatbots: Impacts and lessons for implementation

4.1 Introduction

Every spring, approximately 2.5 million US high school seniors are admitted to college. By September, approximately fourteen percent of those - 350,000 students - who intend to enroll succumb to "summer melt" and fail to matriculate (Castleman & Page, 2014a, 2014b; Castleman, Page, & Schooley, 2014). Students who "melt" over the summer disproportionately come from underserved communities that frequently lack the supportive resources to help students navigate challenging financial, academic, and social situations. Nudging college-intending students by sending them text messages with information and reminders has been an effective strategy for reducing summer melt (Castleman & Page, 2015; Castleman et al., 2014; Page & Scott-Clayton, 2016). Recent advancements in machine learning led to the development of artificial intelligence (AI) enabled chatbots for sending tailored and timely text messages to support college-intending students. These artificially intelligent chatbots hold much potential for scaling student outreach and support effectively and at a low cost.

Page and Gehlbach (2017) demonstrated that artificially intelligent chatbots can be an effective intervention to reduce summer melt. Specifically, the authors experimentally tested the effect of a AI-enabled chatbot designed by AdmitHub for Georgia State University (GSU) and found that chatbot outreach increased GSU-intending students' success with a variety of pre-matriculation tasks and on-time enrollment. These results parallel prior research on summer melt (Castleman, Owen, & Page, 2015; Castleman & Page, 2015; Castleman et al., 2014). However, because the AI allows for the system to manage the majority of student communication without ongoing staff input, it allows college admissions and financial aid staff to redeploy their time on issues that only experienced counselors can solve, making the chatbot intervention highly promising for scaling up.

Although Page and Gehlbach (2017) utilize an experimental design, the gold standard in impact evaluation, more studies are needed to investigate whether the effectiveness of a communication tool such as this generalizes to contexts beyond GSU. Therefore, the goal of this study was to test, through a field experiment, the implementation of an artificially intelligent chatbot to reduce summer melt and improve first-year college enrollment at another four-year university as well as at a community college.

In partnership with East Carolina University (ECU) and Lenoir Community College (LCC), we tested the capacity of AI-enabled chatbots to improve students' success with completing required pre-matriculation tasks and enrolling in college in the fall of 2018. We evaluated chatbots, each named for the mascot of the campuses on which they were implemented. "PeeDee," was tailored to the ECU context, and the other, "Lance," was an analogous system designed for LCC. In both settings, the chatbot communicated with students via text messages to provide them with reminders and follow up support regarding the logistical and administrative tasks that students must complete to successfully matriculate to college.

To preview our findings, at ECU, we observed that the rate of summer melt was relatively low for students overall. Nevertheless, the text-based outreach did increase students' success with navigating the financial aid process and gaining access to student loans by four percentage points. Further, the outreach improved several outcomes for a critical subgroup of students at ECU, namely, those who are first in their family to attend college. For these students, the outreach increased the probability of accepting a loan by eight percentage points, the probability of registering for classes by three percentage points, and the probability of enrolling at ECU by three percentage points.

At LCC, robust chatbot implementation as well as the associated experiment were hindered by a lack of cell phone information for a substantial share of students who were otherwise eligible for outreach at the beginning of the summer of 2018. Given the limited nature of the LCC implementation, we instead turned our focus at LCC to lessons regarding chatbot implementation readiness. In-depth, semi-structured interviews with the LCC staff involved with using the chatbot indicate that, in addition to technical requirements such as valid cell phone number information for students, the implementation of a chatbot tool such as the one we explore required considerable learning and adaptation from both the college staff and the chatbot developers. At the same time, LCC was able to overcome these implementation challenges, and staff report that the use of the chatbot helped LCC to reach and engage students more proactively. Further, staff report that the implementation process served as a driver of organizational learning, as it helped the staff to better understand and articulate their admissions and enrollment processes.

We structure the remainder of the paper as follows: section 4.2 reviews the relevant literature and provides background and context for the study, section 4.3 presents the details and results of the randomized controlled trial at ECU, section 4.4 reports on the qualitative analysis of the chatbot implementation readiness at LCC, and section 4.5 discusses the findings and concludes.

4.2 Background and context

Summer melt is the phenomenon whereby college-intending high school graduates "melt away" during the summer and fail to matriculate to college in the fall semester due to challenges in navigating required pre-enrollment tasks and processes (Castleman & Page, 2014a, 2014b). Lack of family support, financial resources, and knowledge of the higher education system are key correlates of college-intending high school graduates faltering in realizing their college-going plans (Arnold, Fleming, DeAnda, Castleman, & Wartman, 2009). For students succumbing to summer melt, it is challenging to complete all necessary enrollment-related steps, including evaluating financial aid offers, meeting unanticipated costs, filing paperwork, making housing arrangements, taking placement tests, etc. Summer melt affects an estimated 10% to 20% of college-intending students each year, with higher rates among low-income and first-generation college students (Castleman & Page, 2014b).

Several studies have examined whether proactively communicating with and supporting college-intending students via text messages may reduce summer melt. Text messaging campaigns leverage behavioral economics research on how nudges, or timely prompts, increase completion of tasks that benefit both the individual and the society, such as getting flu vaccines or saving money for retirement (Thaler, 2016; Thaler & Sunstein, 2008; Tomer, 2017). Summer melt text messaging nudges are typically designed to (1) inform students about college enrollment-related tasks they might be unaware of and (2) remind students about deadlines for those tasks (Castleman & Page, 2014a). In addition to being theoretically informed, summer melt text messaging campaigns are low cost and use the mode of communication which targeted students tend to use most frequently.

In recent years, advances in computer science have enabled the development of artificially intelligent (AI) chatbots, which offer new opportunities to provide summer melt preventive text messaging support to college-intending high school graduates. In addition to providing information and reminders about completing enrollment-related steps, chatbots offer students an opportunity to interact by asking questions and receiving real-time responses drawn from a knowledge base. In cases where the AI cannot answer a student's question, the communication system is designed to transmit the question to a designated campus staff member. The system is able to learn over time and become more efficient in responding to questions and requests.

Following Page and Gehlbach (2017) who studied the effect of AI-enabled chatbots on summer melt at Georgia State University, we investigate whether the promise of this tool is replicable in other contexts. Specifically, in this study we sought to examine the effect of student outreach and support via an AI-enabled chatbot on student completion of pre-matriculation tasks and timely enrollment at two higher education institutions, East Carolina University (ECU) and Lenoir Community College (LCC). During the study implementation process, we quickly learned that logistical challenges would hinder implementation of the experiment at LCC. Therefore, at LCC, we turned to a qualitative exploration to inform our understanding of factors that support and hinder successful chatbot implementation.

4.3 The randomized controlled trial at East Carolina University

4.3.1 Setting: East Carolina University (ECU)

East Carolina University is a public, four-year research university located in Greenville, North Carolina. Every year, ECU enrolls a freshman class of approximately 5,500 students. Among undergraduate students, approximately 84% are in-state, and 34% receive Pell grants (National Center for Education Statistics, 2018). ECU has a mission to serve the rural eastern North Carolina and currently serves more students from the state's lower-income counties than most colleges in the University of North Carolina System.

4.3.2 Randomization procedure at ECU

For the study, ECU first identified students who qualified for outreach. These were students whom the university classified as intending to enroll at ECU in Fall 2018. A total of 4,442 prospective ECU students participated in the study. We randomly assigned 2,221 students to the treatment group and 2,221 students to the control group. Students assigned to the active treatment condition were targeted for chatbot outreach in addition to all other business-as-usual ECU communication. Students assigned to the control condition received business-as-usual communication but did not receive outreach from the chatbot during the intervention period.

As illustrated in Table 17, the average baseline characteristics of the ECU treatment and control groups are similar, and the differences between groups in these characteristics are not statistically significant. Thus, the randomization procedure worked well. As a result, we are confident that any treatment/control differences that we detect will *not* be due to differences in the types of students that did versus did not receive the opportunity to receive outreach and engage with PeeDee. Participating students are about 18 years old, on average. Female students comprise slightly more than half of the sample. Among participating students, 66% are white, 15% are Black, 8% are Hispanic, 3% are Asian, 1% are Native American, and about 6% are multiracial. About 19% of the participants are first-generation students, and approximately 87% are in-state students. The average study participate has a math-verbal combined SAT score of 1,106, roughly the 58th percentile of the national distribution.

4.3.3 Intervention at ECU: PeeDee

In collaboration with AdmitHub, ECU introduced an artificially intelligent (AI) chatbot (named PeeDee for the university's mascot) into the university's enrollment process for prospective students assigned to the treatment group. The chatbot was designed to: nudge students with reminders relevant to their individual required enrollment and matriculation processes and provide them with timely answers to their questions and, in turn, alleviate staff members from devoting time to answering common questions and allow them to focus on students who needed support that the chatbot could not provide.

University staff designed primary outreach messages, also referred to as text messaging "campaigns", focusing on the following eight categories:

1. Introduction to the chatbot: introducing students to PeeDee's functionality and offering an opportunity to opt out of using PeeDee's assistance
2. Orientation: reminding students to register for orientation and providing the details such as dates and registration links

3. Class registration: reminding students to register for courses and asking whether any information or help was needed
4. Housing: reminding and providing information about actions necessary for timely moving into residence halls
5. Social involvement: invitations to join the campus's official social media groups and to participate in events targeted for freshmen
6. Academic exploration and general enrollment: providing information and offering assistance with degree programs, etc.
7. Relationship building: less-serious messages, such as fun facts about the campus, ECU trivia, and congratulations on the first day of classes.

The university employed the chatbot to text intending ECU students assigned to the treatment group throughout the summer of 2018. Text messaging campaigns sent out to students were either nudges containing reminders to complete matriculation-related actions or interactive messages that invited students to respond to the chatbot. As noted, students were offered the opportunity to opt out at the beginning of the intervention and could opt out (via text) at any time during the course of the outreach. Where possible, text messaging campaigns were tailored to students' specific needs, based on administrative records held by the university. For example, if students had already submitted required paperwork to verify their in-state residency, they did not receive outreach about doing so.

In Table 18, we provide descriptive information on the timing and distribution of the ECU chatbot messaging campaigns. We report on the main categories of student-bound messages but not the contents of additional messages that resulted from the students' interaction with the chatbot. The intervention ran from the beginning of July to nearly the end of August. Throughout the summer, the share of treatment students to receive the outreach varied according to message intention and need. For example, nearly all treatment group students received the introductory message, but only about 10% of students received a reminder to use ECU's internal system (PIER) to register for courses. The remaining 90% had likely already registered for courses and therefore did not need the message. Such variation is indicative of how the chatbot tailored the outreach to each student's needs.

4.3.4 Data and analysis at ECU

We rely on ECU's administrative data to examine impacts of the outreach on completion of required pre-matriculation tasks (e.g., attending orientation or registering for courses). In addition, we used ECU's administrative data linked to records from the National Student Clearinghouse (NSC) to examine whether students who did not enroll in ECU opted to enroll elsewhere.

For all outcomes, we estimated treatment effects using a linear probability model, as follows:

$$Y_i = \alpha + \beta_1 TREATMENT_i + \boldsymbol{X}\gamma + \epsilon_i \tag{1}$$

where $TREATMENT_i$ is an indicator for assignment to the treatment group, $\boldsymbol{X}$ is a vector of student-level covariates included to improve the precision of our treatment impact estimates; and ϵ_i is a residual error term. Our estimates of the β_1 coefficient indicate whether targeting students for outreach served to improve student success on the outcome measures considered. Specifically, we test whether students from the treatment group are more likely to complete the following matriculation-related outcomes: (1) enroll at ECU in the fall of 2018, (2) accept a loan, (3) attend orientation, (4) register for classes, (5) register for more course credits, (6) register for more courses, and (7) enroll at any four year college.

4.3.5 ECU-intending students' interactions with the AI chatbot

In Table 19, we present descriptive statistics on the interactions between ECU-intending students and PeeDee for the sample overall as well as separately by first-generation status. Approximately six percent of the students in the entire treatment group opted out PeeDee outreach during the course of the intervention (Table 19, column 1). Such low level of opt-out is in line with other implementations of text-based summer outreach (Castleman & Page, 2014a). The typical treatment group student received approximately 26 messages from PeeDee throughout the summer (including both the targeted campaigns and the chatbot's responses to students' queries). The messages were tailored to considerably wide-ranging needs and varied by students' responsiveness to the system, so some students received as few as three messages while other students received as many as 97.

In response to PeeDee's outreach, the typical treatment group student sent three messages, on average. Some students did not send any messages, whereas others interacted more frequently. The most engaged student sent PeeDee 52 messages during the course of the summer intervention. The details on the number of days on which students sent messages tell a similar story of engagement heterogeneity. Specifically, some students never sent any messages while other students interacted with the chatbot on many more days, up to 22 days at most. These numbers suggest that some but not necessarily all students have a variety of needs during the summer before starting college. These descriptive statistics suggest that the chatbot is a useful tool for proactively prompting questions and then handling this variation in needs.

The remainder of Table 19 presents summary statistics on the interactions with the chatbot of first-generation students (columns 3 and 4) compared to non-first-generation students (columns 5 and 6). The level of opt-out and average engagement were similar for first-generation and non-first-generation students. However, the maximum values of the engagement metrics differ according to first-generation status. For example, the maximum number of text messages sent by the chatbot to first-generation students (97) was much higher compared to the non-first-generation students (69). The maximum number of messages sent to PeeDee was higher among first-generation students (52) compared to non-first-generation students (34). The maximum number of days (22) interaction with PeeDee occurred was also among the first-generation group. Collectively, these figures suggest that the students who used the system most intensively were first-generation college-goers, even though average system engagement was similar by first-generation status.

4.3.6 Treatment effects at ECU

In Table 20, panel A, we present treatment effect estimates for the full sample. Of the outcomes we consider, the chatbot outreach improved only one outcome: whether ECU-intending students accepted a loan. Specifically, students in the treatment group were approximately four percentage points more likely to accept a loan as a result of being targeted for outreach (column 2). At the same time, the outreach had no effect overall on course registration, orientation attendance or ECU enrollment. To note is that the control group mean values of the full sample suggest that baseline values of each outcome variable, except for loan acceptance, were quite high. In other words, there was little room for improvement in terms of outcomes of interest for the full sample.

By contrast, for first-generation students assigned to the control condition, rates of success with pre-matriculation tasks and timely enrollment are lower than for the sample overall, so first-generation students had comparatively more room for improvement on these outcomes. Treatment effect estimates for these students (Table 20, panel B) indicate positive effects of the chatbot outreach. Specifically, first-generation students in the treatment group were nearly eight percentage points more likely to accept a loan, three percentage points more likely to enroll at ECU, and three percentage points more likely to register for classes compared to the first-generation students in the control group. Further, our estimates using subsamples of data by first-generation student status suggest that the effect of the treatment on loan acceptance in the pooled sample is predominantly driven by the effects on the first-generation students. That more positive effects were concentrated in the subsample of first-generation students at ECU is consistent with the overall prevalence of summer melt among the less advantaged youth.

4.4 Chatbot implementation readiness: Lessons learned at Lenoir Community College

We were unable to implement our initial plan to estimate the effect of chatbot outreach on summer melt at Lenoir Community College because upon receipt of data for randomization, we learned that the college had cell phone contact information for only a small fraction of potentially incoming students. Therefore, to inform future efforts to use text-based communication such as the type we describe in this paper, at LCC we turned to a qualitative examination of factors that may hinder or support chatbot implementation in a community college context. To note is that even though LCC was unable to launch the chatbot in a manner conducive to an experimental study, over a longer time horizon, they did work with and launch the tool in a manner that campus staff considered highly fruitful. We explore the campus's progression to implementation here.

4.4.1 Setting: Lenoir Community College (LCC)

Lenoir Community College is a public, two-year open-admission community college located in Kinston, North Carolina. LCC's student population comprises about 2,700 undergraduate students enrolled in associate's degree programs or certificate programs offered both in-person and online. About 61% of the full-time undergraduate students at LCC receive Pell grants (National Center for Education Statistics, 2018). LCC's mission is to provide accessible higher education for the development of the students and the community.

4.4.2 Lessons learned from chatbot implementation at LCC

We conducted semi-structured interviews with LCC staff involved with chatbot implementation. The analysis of the interview data highlighted the following themes regarding implementation.

Adapting the chatbot for the community college context

LCC was the first community college to implement the AdmitHub chatbot to address summer melt. In the process of implementing the chatbot, it became clear that the design of the chatbot was comparatively better suited to the four-year college context. For a small-scale community college like LCC, where the student body is very fluid and changeable, the chatbot could not be utilized as straightforwardly as in a typical four-year college. As one LCC staff member explained:

> One of the disadvantages that we have is that unlike a university that probably has some hard deadlines for admissions and registration, we're really fluid here. It's possible that somebody is going to apply for this current semester on December 2nd and register for a class that starts December 2nd. I can tell you who our currently enrolled students are today, but that could very well change tomorrow, because we could register somebody tomorrow for a class that's going to start very soon.

Both the LCC staff and AdmitHub developers had to invest considerable time and effort for the chatbot to function as intended. LCC staff had to figure out how to utilize the chatbot in their context. Specifically, they had to prepare information for the chatbot's knowledge base, learn the admin interface for operating chatbot's functionality, study what the chatbot is capable of doing in their specific context, etc. At the same time, the chatbot developers had to learn about the community college context in order to assist LCC staff who administered the chatbot implementation.

Another example of how the fluid nature of a community college operations required the chatbot to be used differently is related to tuition payment deadlines. Typically, four-year colleges have a single, institution-wide tuition payment deadline each term. In contrast, in the community college context, different deadlines are relevant for different students, as explained in the following:

> I've had to send one [campaign] to remind students that tuition was due, and our tuition is due five days before the class starts, but we have classes that start on the first day of the semester, we have classes that start two weeks later, and we have some classes that are getting ready to start October 15th, we even have classes that start December 2.

Because of this more variable course-start time structure, sending a text message to remind students about a seemingly straightforward detail as the tuition payment deadline involves multiple, different outreach campaigns distributed at different times. As a result, implementation of the chatbot within the LCC context required far more manual customization of the messaging campaigns so that students received the messages that were appropriate in terms of timing and content.

Implied by the variation in course start time is the fact that LCC also had a more fluid and ongoing admissions process which required that the chatbot be used far beyond the summer. Therefore, the summer melt campaigns had to be reframed and redeployed for a system of more constant communication with students. The chatbot implementation experience at LCC suggests that the concept of summer melt should likely be redefined in the community college context as an “on-going pre-enrollment melt” to capture how students intending to attend a two-year college might fail to enroll at any time of the year and not only during the summer. Melt prevention appears to be necessary beyond the summer in community colleges. Because community colleges differ considerably in size and scale of operation, the successful design and implementation of chatbots for community colleges should likely begin with identifying pre-enrollment melt patterns throughout the academic year using a given community college’s institutional data.

Limited functionality due to a lack of chatbot’s integration with LCC databases

A strength of the chatbot tool is the ability to integrate it with an institutional student information system to use administrative records to inform the targeting of outreach. For example, a message about FAFSA refiling can be sent only to those students who haven’t refiled by a particular deadline. At LCC, however, this integration was not initially possible.

Therefore, in the LCC context, sending tailored messages to students required more manual work, as one staff member described:

> The communication developed in campaigns was not necessarily dynamically reaching the students who needed it. We had to run queries to identify students, and so the lack of data integration between the bot and our enrollment system is the major challenge. If I had to send a message to 400 students, but really 100 of those students have already resolved what I needed them to resolve or that message is not relevant for them any longer … if the chatbot could extract that information and develop a campaign based on that, that would be the type of dynamic communication that would really take advantage of the chatbot.

Instead, to implement a targeted outreach campaign, LCC staff had to manually query the college's database and import the list of relevant students into the chatbot platform to then send a targeted text campaign. For example, after an initial text message inviting students to the orientation event, an LCC staff member would have to obtain from the institutional data analyst a list of those students who had not yet registered for orientation and enter that list into the chatbot system. Then it was possible to send a reminder text message about the orientation event to these students. The staff member would have to repeat this process for each follow up message focused on orientation. Otherwise stated, LCC was not able to employ the tool's automatic text message tailoring functionality due to a lack of system integration. Thus, chatbot implementation was labor-intensive for LCC staff. Rough estimates suggest that an LCC staff member could spend three to five hours sending a single text messaging campaign, from preparing lists of students to sending the intended text message.

This lack of automation occurred because a robust link between the chatbot system and LCC's database system could not be established. Therefore, campuses considering the implementation of this type of tool should involve their technology and data teams to ensure system integration.

Chatbot as a communication tool and a driver of institutional learning

Even though LCC initially faced several challenges to robust and efficient implementation, after working through these challenges, LCC found the chatbot very useful and chose to continue to use the tool beyond the timeframe of the grant-supported implementation. In short, it took a longer timeframe (approximately six months) for LCC to adapt the tool and integrate it to its campus context. By the second year of implementation, LCC staff confidence in using the chatbot tool was substantially higher, as one staff member describes:

> … I think probably six months into the process we really started to understand the idea of the chatbot. We started to see the benefits of the chatbot. The second year has been very good for us. We've been able to use it in very creative ways and very strategically, as well as tactically when contacting students. … once we were able to understand the concept of the chatbot, then the creativity of an admissions director, a registrar, financial aid director, we could really take advantage of their ideas.

That is, after the initial period of adaptation and learning, LCC staff were able to harness the opportunities offered by the immediacy and interactivity of chatbot's text messaging campaigns.

LCC staff expressed appreciation for the speed with which the chatbot enabled them to reach out to and interact with large numbers of students. As a member of LCC staff put it:

> What we really learned is that students respond much more effectively or efficiently when we have direct communication with them via text [messages] and when they can respond and ask their question through that thing [the chatbot]. We tried postcards, we tried email, we tried other sources of communication, and what we found is that with the chatbot we could really engage students very quickly and get them to respond through those requests very quickly.

In other words, the mode of communication through text messages and the interactivity of the chatbot enabled LCC to connect and engage with the students more effectively compared to other modes previously employed. This improved communication translated to faster resolution to issues students faced. For example, LCC typically observed that about 10% of associate's degree students would not have paid their tuition in the days prior to the start of a semester. After successfully implementing chatbot campaigns focused on bill payment, this rate dropped substantially, with the text outreach spurring increased student activity in terms of calling the campus cashier's office, traveling to the office to pay in person, and paying one's tuition bill online. While by design the chatbot was "learning" and getting better at responding to students' questions, engaging in the chatbot implementation process also sparked new learning among LCC staff themselves. For example, LCC staff consolidated a comprehensive knowledge base from which the chatbot could draw and learned how to update it. Further, they established more regular communication channels among the admissions, registrar, institutional research office, and the staff members administering the chatbot campaigns. The LCC team also reported getting better over time at planning and targeting their chatbot campaigns. For example, they reported gaining proficiency in crafting text message content so that the language is clear but brief enough to fit the character limitation of a text message. Finally, the process pushed staff to attend more carefully to alignment in information via various channels of communication (e.g., whether information on the college website aligned with information they were communicating to students via text). In sum, although data and process barriers hindered quick ramp up to implementation in the context of an experimental study, LCC successfully integrated the chatbot tool into their student communication strategy over a longer time horizon that allowed for institutional learning and necessary adaptation.

4.5 Discussion and conclusions

Chatbot tools such as those investigated here are typically framed as relatively low-cost, however, this framing relates only to the technology itself. As our learnings from LCC reveal, successful implementation and ongoing tool usage can necessitate the input, collaboration and communication among various staff members and offices on campus. If these entities are more accustomed to working in a siloed fashion, successful implementation of a centralized communication tool, such as a student-facing chatbot, can require substantially different business routines. These disparate entities will need to communicate and coordinate more regularly, and new systems of data access and sharing may also be required. Our findings from LCC indicate that in contexts where staff are able to dedicate time and effort to this type of required system-level change, it is possible to successfully implement a new communication strategy, such as a chatbot. However, as we observed in the LCC context, the potential efficiencies that a chatbot tool can offer may take longer to realize, as the tool is adapted to the particular context and the organization systems and procedures are adapted to support implementation.

At ECU, we observe evidence in alignment with that from other campuses (e.g., Georgia State University; Page & Gehlbach, 2017) as well as the broader research literature on summer melt (Castleman et al., 2015; Castleman & Page, 2015; Castleman et al., 2014). That is, when systems are in place to robustly launch chatbot communication focused on summer transition tasks, it can lead to improvement in student success with pre-matriculation requirements as well as with successful matriculation. In the context of ECU, positive impacts of the chatbot tool were realized primarily by students who would be the first in their family to attend college. For these students, the outreach improved students' success with accessing financial aid for college, matriculating on time and registering for courses. That impacts are realized primarily by this subset of students is not surprising, as they may experience less household knowledge about college-going processes compared to their non-first generation peers, and the baseline rates of success with various college-going processes were lower for the first-generation students in our sample.

We considered possible explanations for the limited treatment effects except for the loan acceptance outcome in the pooled ECU sample. First, we explored whether the AI chatbot service was relevant for the ECU students. As the baseline control group means in Table 19 illustrate, 93% of the students had enrolled and 89% registered for classes without outreach or support from the chatbot. Further, the control group students, on average, registered for practically a full-time semester course load of about 5.4 courses or approximately 14 course credits for fall 2018. These control group figures suggest that, for the sample as a whole, there was little room left to improve upon most of the outcomes examined in this study. For colleges and universities considering use of a tool such as this, these control group rates point to the importance of a needs analysis to understand whether or the extent to which implementation is meeting challenges that are being faced in a particular context.

The analysis of the messages the students received reveals that students completed most of the enrollment-related steps prior to the treatment implementation. For example, only 158 students (7% of treatment group) received a message related to course registration. In other words, some of the chatbot messages were focused on actions that many ECU-intending students completed on time, and, in some cases, even before the treatment implementation, as Figure 20 illustrates. Loan acceptance, in contrast, was completed by most students during the summer, as shown in Figure 21. Such temporal concentration of the loan-related decision-making helps to explain why we find positive treatment effects on loan acceptance.

Further, we compared the ECU student body characteristics to the characteristics of students at Georgia State University (GSU) where our first study showed positive effects of the AI chatbot assistant for GSU-intending students overall (Page & Gehlbach, 2017). A detailed comparison of ECU and GSU indicates that the differences in the two institutions' student bodies possibly explain the differences in the chatbot effects. As Table 20 reveals, the student bodies at the two campuses are substantially different. For example, the share of first-generation college students is about 18% at ECU and approximately 32% at GSU. As we might expect, the rate of summer melt at GSU was also substantially higher. In sum, ECU students tend to be overall more advantaged socioeconomically and less prone to summer melt than the GSU students.

Consistent with existing literature, students who "melt" over the summer disproportionately come from underserved communities that frequently lack the supportive resources to help students navigate challenging financial, academic, and social situations related to college matriculation (Castleman & Page, 2015). It appears that, on average, ECU students may be more advantaged socioeconomically than students who attend GSU and, as a result, may be less susceptible to summer challenges that can lead to melt. That more positive effects were concentrated in the subsample of first-generation students at ECU is also consistent with the overall prevalence of summer melt among the less advantaged youth.

In sum, the findings across these two sites help to bolster the evidence that an artificially intelligent chatbot tool has the potential to improve college access in a variety of contexts. We contribute to the growing body of literature examining the potential of technology-supported behavioral interventions in education to scale up (Bird et al., 2019; Gurantz et al., 2020; Page & Nurshatayeva, 2020; Page, Sacerdote, Goldrick-Rab, & Castleman, 2019). We show that scaling behavioral interventions using artificially intelligent chatbots cannot be totally centralized at a massive scale, and instead, should be carefully implemented institution-by-institution. The lessons learned from LCC serve to highlight the data and communication systems and routines that need to be firmly in place in order to support successful implementation. These learnings may inform the types of feasibility assessments in which institutions ought to engage to understand the institutional resources and commitment that are required for a tool such as this to be incorporated fruitfully into regular student communications. At the same time, even though it took LCC time to change their practices, as they did so, they came to see how the chatbot was useful in their working with students in a more student-centered way.

Table 17 Demographic characteristics of ECU participant students

	(1)		(2)		(3)	
	Pooled sample		Treatment group		Control group	
	mean	(sd)	mean	(sd)	mean	(sd)
Age	17.95	(0.50)	17.94	(0.48)	17.96	(0.52)
Female	0.58	(0.49)	0.59	(0.49)	0.57	(0.49)
White	0.66	(0.47)	0.65	(0.48)	0.67	(0.47)
Black	0.15	(0.35)	0.15	(0.36)	0.14	(0.35)
Hispanic	0.08	(0.27)	0.08	(0.27)	0.08	(0.27)
Asian	0.03	(0.16)	0.03	(0.16)	0.03	(0.17)
Native American	0.01	(0.08)	0.00	(0.06)	0.01	(0.09)
Multiracial	0.06	(0.23)	0.07	(0.25)	0.05	(0.22)
First-generation student	0.18	(0.38)	0.17	(0.38)	0.19	(0.39)
In-state	0.87	(0.33)	0.87	(0.34)	0.88	(0.33)
SAT	1,105.89	(120.22)	1,108.14	(119.63)	1,103.63	(120.79)
Observations	4,442		2,221		2,221	

Note. None of the differences between the treatment and control groups are statistically significant.

Table 18 Overview of the chatbot text messaging campaigns at ECU (07/02/2018-08/24/2018)

Chatbot messages to students	Date	% of treatment group recipients
1. Introductory message and offer to opt-out	07/02	99
2. Reminder to register for orientation	07/02	99
3. Happy 4th of July message	07/04	96
4. Reminder to use the internal registration support system	07/05	10
5. Reminder & link to info about how to prepare for moving into residence halls	07/09	84
6. Invite to join the Facebook group of class 2022 with a link	07/13	95
7. Reminder to explore degree programs	07/18	95
8. Reminder to check the admitted students guide	07/20	95
9. College fun facts trivia	07/24	95
10. Reminder to complete NC residency determination for out-of-state students	07/26	1
11. Reminder about the move in day	07/31	84
12. Ask students if they need help registering for classes	08/11	7
13. Reminder to register for classes	08/14	6
14. Song trivia	08/16	95
15. Invitation to the fun event for freshmen	08/17	94
16. Wishing good luck on the first day	08/20	94
17. Reminder about the last day of add/drop	08/24	93
18. Survey on what students think about the chatbot	Variable	84

Table 19 Statistics on the interactions between PeeDee and the treatment ECU-intending students

	(1)	(2)	(3)	(4)	(5)	(6)
	Pooled sample		First-generation students		Non-first-generation students	
	mean (sd)	range	mean (sd)	range	mean (sd)	range
Opt out	0.06 (0.23)	0-1	0.05 (0.21)	0-1	0.06 (0.24)	0-1
Messages sent by PeeDee to students	26.37 (7.51)	3-97	26.52 (7.68)	3-97	26.32 (7.45)	3-69
Messages sent by students to PeeDee	3.36 (4.53)	0-52	3.24 (3.98)	0-52	3.40 (4.37)	0-34
Number of days during which students sent messages to PeeDee	1.42 (1.71)	0-22	1.38 (1.94)	0-22	1.43 (1.63)	0-13
Observations	2,205		548		1,659	

Table 20 Average treatment effects

	(1) Enrolled	(2) Accepted loan	(3) Attended orientation	(4) Registered for classes	(5) Number of credits registered for	(6) Number of courses registered for	(7) Enrolled at another four-year college
A. Full sample estimates							
Treatment	-0.007	0.036**	-0.005	-0.007	-0.108	-0.046	-0.005
	(0.008)	(0.015)	(0.008)	(0.008)	(0.125)	(0.053)	(0.006)
R-squared	0.02	0.05	0.01	0.02	0.01	0.02	0.01
Control group mean	0.93	0.55	0.92	0.93	13.81	5.44	0.96
	(0.01)	(0.01)	(0.01)	(0.01)	(0.09)	(0.04)	(0.00)
Observations	4,442						
B. First-generation sub-sample estimates							
Treatment	0.034*	0.078**	0.002	0.034*	0.378	0.100	0.025
	(0.020)	(0.032)	(0.019)	(0.020)	(0.315)	(0.132)	(0.017)
R-squared	0.01	0.06	0.03	0.01	0.02	0.02	0.02
Control group mean	0.89	0.63	0.92	0.89	13.18	5.18	0.92
	(0.01)	(0.02)	(0.01)	(0.01)	(0.24)	(0.10)	(0.01)
Observations	849						

Robust standard errors in parentheses. Results based on regression models that control for all covariates presented in Table 2.
*** p<0.01, ** p<0.05, * p<0.1

Table 21 Comparing ECU and GSU characteristics

	(1) East Carolina University	(2) Georgia State University
A. Data from IPEDS 2015		
% admitted	70	59
Admission yield	36	41
% receiving undergraduate degree within 4 years	34	23
% receiving undergraduate degree within 6 years	61	53
In-state tuition	$ 4,365	$ 6,846
Out-of-state tuition	$ 20,323	$ 21,414
B. Data from College Scorecard		
Minority-serving	No	1. Asian-American & Native American Pacific Islander- serving institution 2. Predominantly black institution
Average annual net price for federal financial aid recipients	$ 15,203	$ 14,773
Median salary of federal financial aid recipients 10 years after attending	$ 40,500	$ 43,300
Students receiving federal loans	55%	56%
Students who return after their first year	81%	81%
Full-time students	88%	78%
Pell grant recipients	33%	51%
White	68%	25%
Black	16%	42%
Hispanic	6%	10%
Asian	3%	13%
SAT verbal 75th percentile	560	590
SAT math 75th percentile	570	590
Most popular programs	1. Health and related 19% 2. Business, management, marketing, and related 18% 3. Education 8%	1. Business, management, marketing, and related 25% 2. Social Sciences 10% 3. Psychology 9%

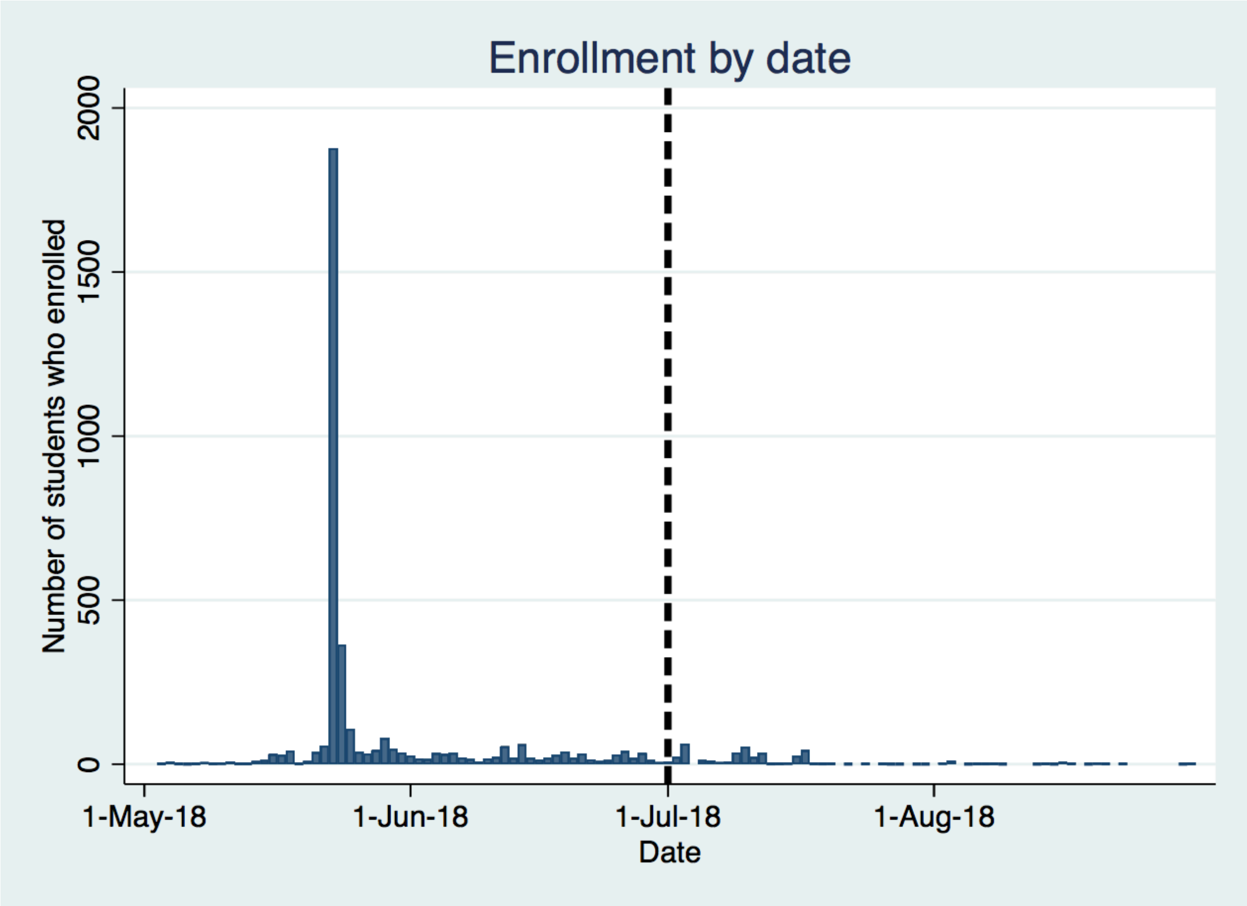

Figure 20 Enrollment at ECU before and after the randomization

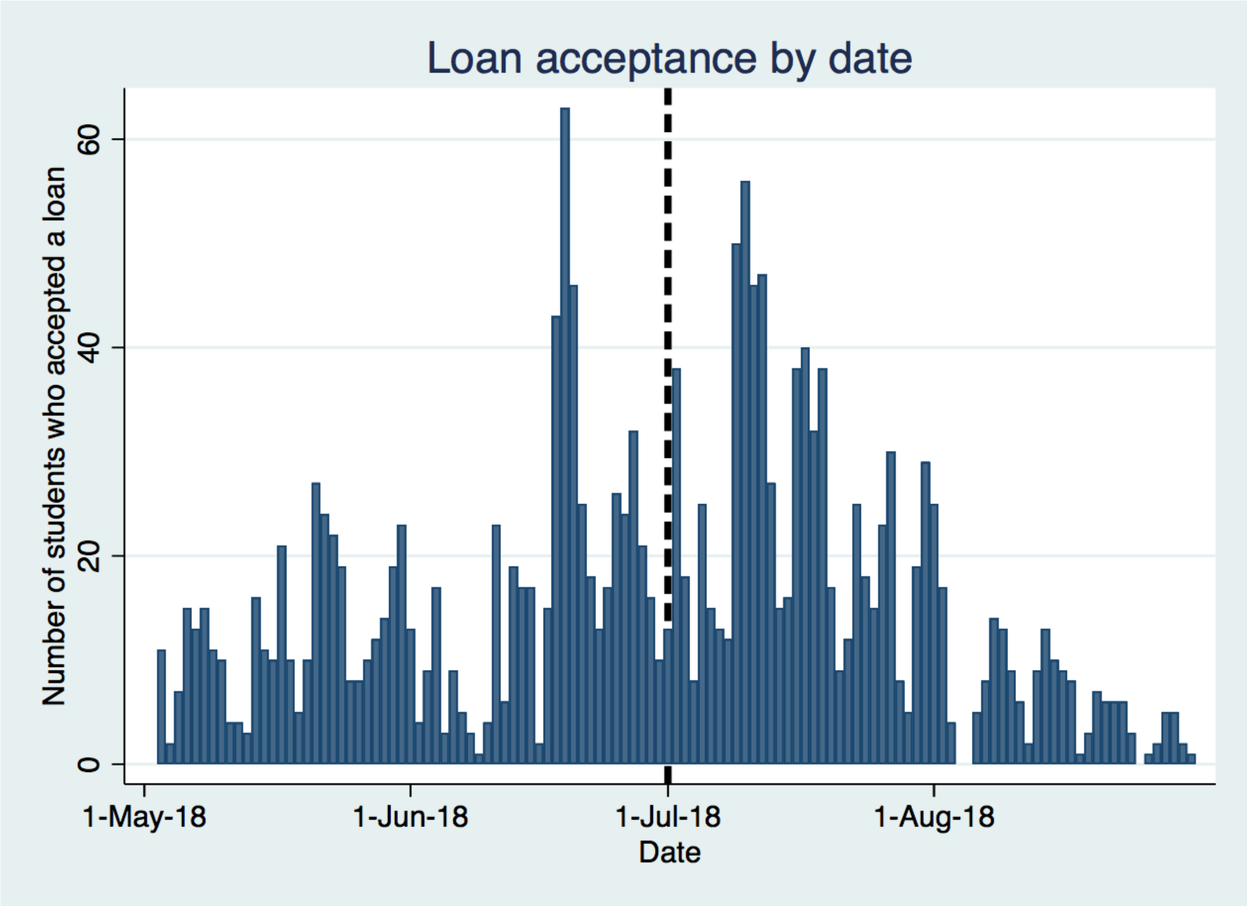

Figure 21 Loan acceptance at ECU before and after the randomization

5.0 Conclusion

This interdisciplinary doctoral dissertation examined several policy-relevant aspects of higher education. The findings suggest the following conclusions. First, at least in selective universities in non-English speaking countries, the switch to English-only instruction may affect college outcomes negatively at the time of transition but may not necessarily imply longer-run negative effects. Second, there is only a limited presence of the chaperone effect in social science publications. New first authors with no prior publication experience outnumber chaperoned first authors who have previous co-publication experience with senior scholars in economics, education, political science, psychology, and sociology. Third, artificially intelligent virtual assistants do reduce summer melt, and positive effects are concentrated among first-generation college students. Community colleges intending to adopt artificially intelligent chatbots should consider adapting the technology to their fluid enrollment processes and to their data systems as well as anticipate embarking on a path of institutional learning and innovation. I hope the studies conducted in this dissertation will inform education policies and spark new research in higher education.

Appendix A Additional calculations for the study of the chaperone effect in the social science publications

Table 22 OLS estimates of the relationship between ln-transformed five-year citations and first author type

	(1) Economics	(2) Education	(3) Political science	(4) Psychology	(5) Sociology
Chaperoned	0.16***	0.24***	0.20***	0.15***	0.20***
	(0.02)	(0.02)	(0.04)	(0.01)	(0.04)
Established	0.01	-0.02*	0.09***	-0.18***	0.03
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Constant	1.95***	1.82***	1.73***	2.50***	1.98***
	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)
Observations	86,829	65,920	28,749	188,675	19,821
R-squared	0.11	0.06	0.11	0.08	0.08

Notes. The outcome variable is ln-transformed five-year citation count. All regressions include each paper's number of authors and journal's Scimago rank as controls. The dataset contains only papers published before 2014. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Bibliography

Ackerley, K., Guarda, M., & Helm, F. (2017). *Sharing perspectives on English-medium instruction*. Bern, Switzerland: Peter Lang.

Angrist, J., & Krueger, A. (1992). *Estimating the payoff to schooling using the Vietnam-era draft lottery*. Cambridge, MA: NBER Working Paper #4067.

Arnold, K., Fleming, S., DeAnda, M., Castleman, B. L., & Wartman, K. L. (2009). The summer flood: The invisible gap among low-income students. *Thought & Action*, *25*, 23-34.

Asian Development Bank. (2011). *Higher education across Asia: An overview of issues and strategies*. Manila, Philippines: Asian Development Bank.

Autor, D. H. (2019). *Work of the past, work of the future*. Cambridge, MA: MIT Department of Economics working paper.

Barreira, P., Basilico, M., & Bolotnyy, V. (2018). *Graduate student mental health: Lessons from American economics departments*. Cambridge, MA: Harvard University.

Becker, G. S. (1993). The economic way of looking at behavior: The Nobel lecture. *Journal of Political Economy*, *101*(3), 385-409.

Bird, K. A., Castleman, B. L., Denning, J. T., Goodman, J., Lamberton, C., & Rosinger, K. O. (2019). *Nudging at scale: Experimental evidence from FAFSA completion campaigns*. Cambridge, MA: National Bureau of Economic Research.

Bolton, K., Botha, W., & Bacon-Shone, J. (2017). English-medium instruction in Singapore higher education: Policy, realities and challenges. *Journal of Multilingual and Multicultural Development*, 1-18.

Bradford, A. (2016). Toward a typology of implementation challenges facing English-medium instruction in higher education: Evidence from Japan. *Journal of Studies in International Education, 20*(4), 339-356.

Byun, K., Chu, H., Kim, M., Park, I., Kim, S., & Jung, J. (2011). English-medium teaching in Korean higher education: Policy debates and reality. *Higher Education, 62*(4), 431-449.

Card, D. (1999). The causal effect of education on earnings. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Vol. 3, pp. 1801-1863). Amsterdam: Elsevier.

Carter, P. H., Berndt, E. R., DiMasi, J. A., & Trusheim, M. (2016). Investigating investment in biopharmaceutical R&D. *Nature Reviews. Drug Discovery, 15*(10), 673-674.

Cassuto, L. (2015). *The graduate school mess: What caused it and how we can fix it*. Cambrdige, MA: Harvard University Press.

Castleman, B. L., Owen, L., & Page, L. C. (2015). Stay late or start early? Experimental evidence on the benefits of college matriculation support from high schools versus colleges. *Economics of Education Review, 47*, 168-179.

Castleman, B. L., & Page, L. C. (2014a). *Summer Melt: Supporting Low-Income Students through the Transition to College*. Cambridge, MA: Harvard Education Press.

Castleman, B. L., & Page, L. C. (2014b). A Trickle or a Torrent? Understanding the Extent of Summer “Melt” Among College-Intending High School Graduates. *Social Science Quarterly, 95*(1), 202-220.

Castleman, B. L., & Page, L. C. (2015). Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates? *Journal of Economic Behavior & Organization, 115*, 144-160.

Castleman, B. L., Page, L. C., & Schooley, K. (2014). The forgotten summer: Mitigating summer attrition among college-intending, low-income high school graduates. *Journal of Policy Analysis and Management, 33*(2), 320-344.

Chao, G. T., Walz, P. M., & Gardner, P. D. (1992). Formal and informal mentorships: A comparison on mentoring functions and contrast with nonmentored counterparts. *Personnel Psychology, 45*(3), 619-636.

Cole, J. R., & Cole, S. (1974). Social stratification in science. *American Journal of Physics, 42*(10), 923-924.

Collier, R. (2009). Rapidly rising clinical trial costs worry researchers. *Canadian Medical Association Journal, 180*(3), 277-278.

Dearden, J. (2014). *English as a medium of instruction–a growing global phenomenon*. London, UK: British Council.

Dee, T. S. (2004). Are there civic returns to education? *Journal of Public Economics, 88*(9), 1697-1720.

Doiz, A., Lasagabaster, D., & Sierra, J. (2011). Internationalisation, multilingualism and English-medium instruction. *World Englishes, 30*(3), 345-359.

Doiz, A., Lasagabaster, D., & Sierra, J. M. (2013). *English-medium instruction at universities: Global challenges*. Bristol, UK: Multilingual matters.

Drljača Margić, B., & Vodopija-Krstanović, I. (2017). *Uncovering English-medium instruction: Glocal issues in higher education*. Oxford, UK: Peter Lang.

Evans, S., & Morrison, B. (2011). Meeting the challenges of English-medium higher education: The first-year experience in Hong Kong. *English for Specific Purposes, 30*(3), 198-208.

Evans, T. M., Bira, L., Gastelum, J. B., Weiss, L. T., & Vanderford, N. L. (2018). Evidence for a mental health crisis in graduate education. *Nature Biotechnology, 36*(3), 282-284.

Goodman, B. A. (2014). Implementing English as a medium of instruction in a Ukrainian University: Challenges, adjustments, and opportunities. *International Journal of Pedagogies and Learning, 9*(2), 130-141.

Gurantz, O., Howell, J., Hurwitz, M., Larson, C., Pender, M., & White, B. (2020). *Realizing Your College Potential? Impacts of College Board's RYCP Campaign on Postsecondary Enrollment*. Annenberg Institute at Brown University: EdWorkingPaper No.19-40.

Higgins, M. C., & Kram, K. E. (2001). Reconceptualizing mentoring at work: A developmental network perspective. *Academy of Management Review, 26*(2), 264-288.

Hu, G., & Lei, J. (2014). English-medium instruction in Chinese higher education: A case study. *Higher Education, 67*(5), 551-567.

Imbens, G., & Rubin, D. B. (2015). Causal inference for statistics, social, and behavioral sciences. New York, NY: Cambridge University Press.

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science, 322*(5905), 1259-1262.

Levecque, K., Anseel, F., de Beuckelaer, A., van der Heyden, J., & Gisle, L. (2017). Work organization and mental health problems in PhD students. *Research Policy, 46*(4), 868-879.

Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English medium instruction in higher education. *Language teaching, 51*(1), 36-76.

Matias, J. (2014). How to identify gender in datasets at large scales, ethically and responsibly. Retrieved from http://mediashift.org/2014/11/how-to-ethically-and-responsibly-identify-gender-in-large-datasets/

Michael, J. (2007). 40 000 Namen, Anredebestimmung anhand des Vornamens. *c't*(17), 182–183.

Milojević, S. (2014). Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(11), 3984-3989.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. New York, NY: Oxford University Press.

National Center for Education Statistics. (2018). *College navigator*. Washington, D.C.: U.S. Dept. of Education, Institute of Education Sciences, National Center for Education Statistics.

National Science Foundation. (2018). *Doctorate Recipients from U.S. Universities: 2016. Special Report NSF 18-304*. Alexandria, VA: National Science Foundation, National Center for Science and Engineering Statistics.

Nguyen, H. T., Hamid, M. O., & Moni, K. (2016). English-medium instruction and self-governance in higher education: The journey of a Vietnamese university through the institutional autonomy regime. *Higher Education*, *72*(5), 669-683.

Nurshatayeva, A., & Page, L. C. (2019). Effects of the shift to English-only instruction on college outcomes: Evidence from Central Asia. *Journal of Research on Educational Effectiveness*, 1-29.

Packalen, M., & Bhattacharya, J. (2015). *Age and the trying out of new ideas*. Cambridge, MA: NBER.

Page, L. C., & Gehlbach, H. (2017). How an artificially intelligent virtual assistant helps students navigate the road to college. *AERA Open*, *3*(4), 1-12.

Page, L. C., & Nurshatayeva, A. (2020). *Behavioral economics of higher education: Theory, evidence, and implications for policy and practice*. Pittsburgh, PA: University of Pittsburgh.

Page, L. C., Sacerdote, B., Goldrick-Rab, S., & Castleman, B. (2019). *Financial aid nudges: A national experiment with informational interventions*. Philadelphia, PA: The Hope Center.

Page, L. C., & Scott-Clayton, J. (2016). Improving college access in the United States: Barriers and policy responses. *Economics of Education Review, 51*, 4-22.

Pannapacker, W. A. (2013). Just look at the data, if you can find any: We need more openness about the job placements of Ph.D.'s. *The Chronicle of Higher Education*.

Parr, C. (2014). English language use 'most significant internationalisation trend for HE'. *Times Higher Education*, Retrieved from: https://www.timeshighereducation.com/news/english-language-use-most-significant-internationalisation-trend-for-he/2013009.article.

Salmi, J. (2009). *The challenge of establishing world-class universities*. Washington, D.C.: The World Bank Publications.

Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*(4), e156.

Sekara, V., Deville, P., Ahnert, S. E., Barabási, A.-L., Sinatra, R., & Lehmann, S. (2018). The chaperone effect in scientific publishing. *Proceedings of the National Academy of Sciences, 115*(50), 12603-12607.

Sert, N. (2008). The language of instruction dilemma in the Turkish context. *System, 36*(2), 156-171.

Shrum, W., Genuth, J., & Chompalov, I. (2007). *Structures of scientific collaboration*. Cambridge, MA: MIT Press.

Stock, W. A., Siegfried, J. J., & Finegan, T. A. (2011). Completion rates and time-to-degree in economics PhD programs (with comments by David Colander, N. Gregory Mankiw, Melissa P. McInerney, James M. Poterba). *American Economic Review, 101*(3), 176-188.

Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *The American Economic Review, 106*(7), 1577-1600.

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. New Haven, CT: Yale University Press.

Tomer, J. F. (2017). *Advanced introduction to behavioral economics*. Northampton, MA: Edward Elgar Publishing.

Torvik, V., & Smalheiser, N. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data, 3*(3), 1-29.

UNESCO Institute for Statistics. (2014). Higher education in Asia: Expanding out, expanding up. Montreal, Canada: The Unesco Institute for Statistics

van Dijk, D., Manor, O., & Carey, L. B. (2014). Publication metrics and success on the academic job market. *Current Biology, 24*(11), R516-R517.

Wächter, B., & Maiworm, F. (2014). *English-taught programmes in European higher education: The state of play in 2014*. Bonn, Germany: Lemmens Medien GmbH.

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics, 6*(4), 700-711.

Way, S. F., Morgan, A. C., Clauset, A., & Larremore, D. B. (2017). The misleading narrative of the canonical faculty productivity trajectory. *Proceedings of the National Academy of Sciences of the United States of America, 114*(44), E9216-E9223.

Weisbuch, R., & Cassuto, L. (2016). *Reforming doctoral education, 1990 to 2015: Recent initiatives and future prospects*. New York, NY: Andrew W. Mellon Foundation.

Wilkinson, R. (2017). Trends and issues in English-medium instruction in Europe. In K. Ackerley, M. Guarda, & F. Helm (Eds.), *Sharing perspectives on English-medium instruction*. Bern, Switzerland: Peter Lang.

Wu, Y. (2020). The funding of academic research in the US. In Y. Wu (Ed.), *America's leaning ivory tower: The measurement of and response to concentration of federal funding for academic research* (pp. 1-10). Switzerland: Springer.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036-1039.

Zahneis, M. (2019). Doctoral education Is flawed. What's one solution from the U. of Chicago? Capping the size of Ph.D. programs. *The Chronicle of Higher Education*.

Zhao, J., & Dixon, L. Q. (2017). *English-medium instruction in Chinese universities: Perspectives, discourse and evaluation*. New York, NY: Routledge.

Zuckerman, H. (1967). Nobel laureates in science: Patterns of productivity, collaboration, and authorship. *American Sociological Review, 32*(4), 391.

Zuckerman, H., Cole, J. R., & Bruer, J. T. (1991). *The outer circle: Women in the scientific community*. New York, NY: WW Norton & Co.