APPLYING DATA SCIENCE TECHNIQUES TO PROMOTE EQUITY AND MOBILITY IN EDUCATION AND PUBLIC POLICY

A Dissertation

Presented to

The Faculty of the School of Education and Human Development

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Brian Heseung Kim

May 2022

Educational Policy Studies
School of Education and Human Development
University of Virginia
Charlottesville, Virginia

APPROVAL OF THE DISSERTATION

This dissertation, ("Applying Data Science Techniques to Promote Equity and Mobility in Education and Public Policy"), has been approved by the Graduate Faculty of the School of Education and Human Development in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Benjamin L. Castleman (Chair)

_____
Daphna Bassok

_____
Vivian C. Wong

_____
Brian Wright

_____ Date

DEDICATION

To the many mentors and teachers – formal and informal – who have shared with me their wisdom, kindness, and energy throughout my life. May we all find the strength to continue passing these treasures forward.

ACKNOWLEDGMENTS

I am deeply grateful for the support of so many in my life as I conclude my doctoral studies. Countless family members, friends, mentors, colleagues, teachers, and mentors have allowed me to grow into the researcher and thinker I am today, and there is no part of this dissertation that stands untouched by their influence.

Several mentors and colleagues have also supported this work directly: through their sage guidance, research partnership, mentorship, and more. I am especially grateful for the fantastic and generous mentorship of Benjamin L. Castleman throughout my doctoral studies; the early and crucial guidance of Kelli A. Bird; the tireless support and partnership of Preston Magouirk, Mark Freeman, Don Yu, Trent Kajikawa, Dave Tiss, and other colleagues at the Common Application; the phenomenal long-term research partnership with Catherine Finnegan and others at the Virginia Community College System; the phenomenal co-authorship of Daniel Rodriguez-Segura, Katharine Meyer, and Alice Choe; the constructive feedback from Daphna Bassok, Brian Wright, Vivian C. Wong, James Wyckoff, Benjamin T. Skinner, Walter Herring, Amanda Lu, Kylie Anglin, Julie Park, AJ Alvero, Brendan Bartanen, Michal Kurleander, OiYan Poon, Katharine Sadowski, Vandeen Campbell, and Erich Purpur; the encouragement of my colleagues in the Center on Education Policy and Workforce Competitiveness (EdPolicyWorks) and the Nudge4 Solutions Lab; and research assistance from Danielle Peacock, Emily Arizaga, Hannah McBride, Kelly Owczarski, Jordan Murphy, Sarah Mackey, and Riley Devol.

This work as crucially been made possible with support from the National Academy of Education and Spencer Foundation Dissertation Fellowship Program, and the Institute of Education Sciences under grants #R305B140026 (Virginia Education Science Training Pre-Doctoral Fellowship program) and #R305N160026 (Nudges to the Finish Line).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DISSERTATION OVERVIEW

Data science continues to percolate in high-profile ways across academic disciplines, so much so that it has almost become a contrivance to ask, "How can data science be applied to my field?" That being said, this optimism is not without merit given the wide variety of promising possibilities data science has presented as of late. Methodological advances marrying the flexibility of data science with the rigor of econometrics have produced tools that enhance our ability to identify treatment heterogeneity – even at the *individual* level (Athey et al., 2019) – and synthesize data-driven theoretical models of causality across studies (Hünermund & Bareinboim, 2019). Applied data science frameworks such as predictive analytics can be usefully deployed in many policy contexts to allocate scarce resources more efficiently (Kleinberg et al., 2018) and identify individuals most likely in need of intensive supports (Liang et al., 2020). And rapid advances in data processing subfields like text mining, machine vision, and machine hearing can now make tractable an enormous wealth of data and insight previously not amenable to traditional data collection and statistical analysis techniques (Anglin, 2019; Fesler et al., 2019; Tiecke et al., 2017). All said, data science offers a compelling toolset that can serve to expand upon and enhance the methods we as policy analysts already employ.

While part of data science's recent rise in prominence is indeed due to its increasing utility, another noteworthy factor is how quickly the barriers to entry are diminishing at the same time. Open-source libraries developed by the likes of Google and Microsoft now make

it possible to implement state-of-the-art algorithms "off the shelf" in fewer than ten lines of code and at no cost to the researcher (e.g. HuggingFace for natural language processing; Wolf et al., 2020). Though these tools are typically quite computationally demanding, growing access to high-performance computing environments (e.g. via institutional clusters or private cloud computing services) enables the execution of these analyses within the bounds of even modest financial and logistical constraints.[1] Lastly, an enormous array of high-quality training resources, support communities, user-friendly toolsets, and thoughtful exemplars are widely available and freely accessible thanks to the ethos of democratization and open-sourcing pervasive throughout data science (Wickham et al., 2019).

The potential applications of data science for education policy research, more specifically, abound. And though there exist several noteworthy examples of educational researchers applying these methods fruitfully, this work remains largely nascent. For instance, massive stores of text data are increasingly prevalent throughout education: from online discussion boards (Bettinger, Liu, & Loeb, 2016), to teachers' job application essays (Penner et al., 2019), and school-level policy documents (Anglin, 2019). Particularly as higher education researchers gain greater facility at extracting data from large-scale learning management systems (e.g. Canvas, Blackboard; Baker et al., 2020) and broad swathes of K–12 classrooms accelerate their digitization in response to COVID-19 (De Vynck & Bergen, 2020), the availability of comprehensive text datasets seems likely to only increase. Where before these data were largely restricted to analysis using more focused qualitative approaches, text mining methodologies now allow researchers to access broader, albeit less

[1] For example, University of Virginia Research Computing makes 100,000 compute-hours of their supercomputing cluster available to any and all faculty, free of charge, with several no-cost options for additional resources as needed. Google Colab provides free cloud access to state-of-the-art machine learning hardware with only minor limitations, and these limitations can be lifted (for most use-cases) for only $10/mo.

nuanced, distillations of valuable insights from these pervasive text data. Similarly, there exists a large number of beneficial, but costly, interventions throughout K12 and higher education that must be carefully distributed to best ameliorate learning disparities and improve student success given scarce resources; when the business-as-usual for targeting often involves arbitrary eligibility cut-offs, predictive analytics techniques have enormous potential to add value by more rigorously grounding resource allocation in available empirical evidence (Herring, 2021). Thus, just as widespread access to longitudinal administrative databases facilitated a rapid expansion in what policy questions were viable for study, widespread access to data science methodologies may well be poised to do the same.

While these pieces together may read to many as a story of optimism, the rapid proliferation of new and increasingly powerful methodologies, many enticing applications, *and* diminishing barriers to entry should also give us pause as a potentially toxic combination. As with any analytic method, these data science approaches carry with them important assumptions and limitations that must be carefully considered and evaluated in the context of each individual application. Yet the potency and opaqueness of newer data science methods, in particular, serve often to obscure this fact and almost encourage their uncritical use.[2] Paired with lower barriers to entry that support their deployment by individuals without formal statistical or scientific training, it becomes clear that the democratization of data science may not necessarily come without unintended consequences. For example, prediction errors often assumed to be idiosyncratically distributed can have catastrophic

[2] To illustrate: the aforementioned HuggingFace library allows users to deploy cutting-edge natural language processing pipelines in less than 10 lines of code, e.g. to translate entire passages of text into multiple languages. However, it does not, at least at time of writing, offer any tools to interpret nor evaluate their output by default.

consequences for algorithmic discrimination if left unexamined prior to implementation (Kleinberg et al., 2016; Obermeyer et al., 2019). Similarly, the presence of names and gendered pronouns can systematically skew the output of modern natural language processing algorithms, as the underlying algorithms are "trained" on datasets that contain the biases of society more broadly (Sun et al., 2019). And even as these tools offer us more ways by which to understand and quantify inequities in our society (e.g. Adukia et al., 2021), the very collection of these data for analysis itself can replicate and reinforce power hierarchies and systems of oppression (Eubanks, 2018; Zuboff, 2019). Education researchers and policy analysts thus have a growing opportunity to enhance the measurement, analysis, and advancement of equity and mobility in our society with these methods, but also a heavy responsibility to do so carefully, thoughtfully, and with a keen awareness for when the benefits of these methods may ultimately be outweighed by their unintended consequences.

In this dissertation, I hope to offer three concrete examples for how education and policy researchers may navigate this precise tension: exemplifying how these new tools from data science can ultimately be harnessed for greater social good, while also demonstrating the requisite diligence, methodological rigor, and critical awareness necessary for that to happen. Importantly, I frame my contribution through these papers in four parts. First, while I do not seek to advance the frontiers of data science by developing *new* methodologies (in my reckoning, a realm likely better left to statisticians and computer scientists), I refine the procedures and checks necessary to usefully apply *existing* data science approaches for the policy analysis context. Second, I strive to make publicly available and open-source my methodologies, code, and documentation, such that this dissertation may serve as a resource for future analysts similarly concerned with the precarious tensions we navigate in these research endeavors. Third, I use the narrative of each paper to make more accessible these

methodologies through translational and intuitive explanations; when paired with technical appendices, footnotes, and my open-source codebases, I believe I am able to accomplish this pedagogical endeavor without sacrificing technical precision. Lastly, I moreover argue that each of the proposed papers also makes a substantive contribution to their respective literatures, offering timely new insights into dynamics previously difficult or impossible to ascertain.

In the first chapter, I present a peer-reviewed paper published at Development Engineering and co-authored with fellow graduate student, Daniel Rodriguez-Segura. With recent advances in high-resolution satellite imagery and machine vision algorithms, fine-grain geospatial data on population are now widely available: kilometer-by-kilometer, worldwide. In this paper, we showcase how researchers and policymakers in developing countries can leverage these novel data to precisely identify "education deserts" – localized areas where families lack physical access to education – at unprecedented scale, detail, and cost-effectiveness. We demonstrate how these analyses could valuably inform educational access initiatives like school construction and transportation investments, and outline a variety of analytic extensions to gain deeper insight into the state of school access across a given country. Throughout the paper, we make explicit the many decisions, considerations, and potential issues our analytic framework presents, signposting for researchers various ways to diagnose and overcome these issues (when possible) in the process. Our intention is that this work can be usefully deployed by a wide array of researchers, policy analysts, and policymakers using our open-source codebase to easily replicate our proposed analyses for other countries, educational levels, and public goods more generally.

In the second chapter, I present a paper with Katharine Meyer and Alice Choe. Interactive, text message-based advising programs have become an increasingly common

strategy to support college access and success for underrepresented student populations – yet we currently know little about how students engage in these text-based advising opportunities and whether that relates to stronger student outcomes – factors that could help explain why we've seen relatively mixed evidence about their efficacy to date. In this paper, we use data from a large-scale, two-way text advising experiment focused on improving college completion to explore variation in student engagement using nuanced interaction metrics and natural language processing. We then explore whether student engagement patterns are associated with key outcomes including persistence, GPA, credit accumulation, and degree completion. Our results reveal substantial variation in engagement measures across students, indicating the importance of analyzing engagement as a multi-dimensional construct. Especially as advising interventions of this kind proliferate across higher education institutions, we show the value of applying a more codified, comprehensive lens for examining student engagement in these programs and illustrate how researchers might usefully approach the application of these scalable data mining techniques going forward.

In the third chapter, I present a sole-authored paper slated for submission to journals as a next step. While scholars have already uncovered many ways that inequities can manifest across the postsecondary application portfolio – from standardized tests to advanced course-taking opportunities – we know almost nothing about whether teacher letters of recommendation also present differential barriers to students' college aspirations. In this paper, I conduct the first system-wide, large-scale text analysis of teacher recommendation letters in U.S. postsecondary applications using data from 1.6 million students, 540,000 teachers, and 800 postsecondary institutions. I use sophisticated natural language processing methods to examine the prevalence of "linguistic biases" within these letters: whether

students are described by teachers in systematically different ways across race and gender groups, even after accounting for salient confounding factors like student academic and extracurricular qualifications, teacher fixed effects, and institution fixed effects. I find evidence of salient linguistic biases across gender, but less evidence for linguistic biases across race. Moreover, these biases are generally most meaningful in terms of the *topical content* of letters; differences in terms of the *positivity* of letters are far smaller in relative magnitudes and thus are less likely to be perceptible in the actual reading of letters. Altogether, these findings have broad implications for the use of recommendation letters in selective admissions, affirmative action policies, and gender diversity in STEM fields. Building off of the work I present in my second dissertation chapter, I further refine my frameworks for applying natural language processing and text mining methodologies and offer a means of quantitatively examining linguistic biases within the educational research context.

Taken together, I hope this dissertation offers a useful base on top of which other analysts might build in the pursuit of rigorous applications of data science methodologies that support equity and mobility in our society.

CHAPTER 1

The Last Mile in School Access: Mapping Education Deserts in Developing Countries

Daniel Rodriguez-Segura, Brian Heseung Kim

Abstract

With recent advances in high-resolution satellite imagery and machine vision algorithms, fine-grain geospatial data on population are now widely available: kilometer-by-kilometer, worldwide. In this paper, we showcase how researchers and policymakers in developing countries can leverage these novel data to precisely identify "education deserts" – localized areas where families lack physical access to education – at unprecedented scale, detail, and cost-effectiveness. We demonstrate how these analyses could valuably inform educational access initiatives like school construction and transportation investments, and outline a variety of analytic extensions to gain deeper insight into the state of school access across a given country. We conduct a proof-of-concept analysis in the context of Guatemala, which has historically struggled with educational access, as a demonstration of the utility, viability, and flexibility of our proposed approach. We find that the vast majority of Guatemalan population lives within 3 km of a public primary school, indicating a generally low incidence of distance as a barrier to education in that context. However, we still identify concentrated pockets of population for whom the distance to school remains prohibitive, revealing important geographic variation within the strong country-wide average. Finally, we show how even a small number of optimally-placed schools in these areas, using a simple algorithm we develop, could substantially reduce the incidence of education deserts in this context. We make our entire codebase available to the public – fully free, open-source, heavily documented, and designed for broad use – allowing analysts across contexts to easily replicate our proposed analyses for other countries, educational levels, and public goods more generally.

I. Introduction

Developing countries have recently made significant strides in improving fundamental educational outcomes like literacy rates and primary school enrollment. For instance, net enrollment in primary school worldwide went from 72% in 1970 to 89% in 2018, thanks to widespread efforts and strategic investments from governments and international agencies (World Bank, 2017a). These encouraging advances have motivated a corresponding change in the policy priorities of development organizations and policy institutions from getting students *into* school, to improving the learning outcomes of students while attending school (World Bank, 2017a). However, despite this meaningful progress in terms of enrollment, much of the developing world is *still* far from achieving universal education. For instance, 1 of every 6 age-appropriate children for primary and secondary school in low-income countries remained out of school by 2018 – a total of 258 million children around the world (UNESCO, 2019).

While the particular reasons students remain unenrolled in school varies by context and individual, available evidence shows that actually having a school physically nearby is *the* first-order necessity for attending school and improving human capital. As Evans and Mendez-Acosta put it, "ultimately, construction is likely a necessary condition for other interventions to work when there are insufficient schools" (Evans and Mendez-Acosta, 2021). As such, ensuring that the full population of a region has reasonable physical access to a school is a critical first step in this pursuit of universal school enrollment. Adequately addressing this need requires that policymakers and researchers identify highly localized areas in which populations lack physical access to school. Yet to date, fine-grain analyses of this kind for developing countries have been logistically and financially prohibitive due to the

costs of conducting local surveys and standing up the extensive analytic infrastructure required.

In this paper, we develop an open-source analytic framework to precisely identify areas of lower physical access to schools (i.e., "education deserts", per Hillman, 2016) using recently available estimates of the distribution of population across nearly every square kilometer on the planet (WorldPop, 2018). By cross-referencing these publicly-accessible data with administrative records on school locations within a given country – data that are also broadly available and accessible to the public across many contexts – we can empirically quantify the extent to which distance to school is a problem within a given country, and further identify the exact areas, if any, where people do not have access to schools nearby. Prior analyses of educational access, particularly in developing countries, were typically limited to characterizing broad regional tracts, such as counties or departments (e.g., Lehman et al., 2013), or local areas with extensive data collection resources, such as larger urban centers. By comparison, our framework can identify education deserts across nearly every country in the world down to the 1 km^2 level – a resolution substantially more amenable to targeted policy interventions like school construction when paired with the contextual expertise of local policymakers. To provide a demonstration of this analytic framework in the present paper, we exemplify our approach in the Guatemalan context, a country which has historically struggled with educational access and equity.

Ultimately, our analytic framework offers a multitude of actionable insights for policymakers and researchers. First, it allows us to estimate how far individuals in every square kilometer of a country must travel to reach a school – analyzable separately by primary/secondary/postsecondary schools, public/private, or other categories of interest. We further visualize these results using a variety of figures and maps to make the wealth of

output easily parsed by policymakers. Second, we can re-contextualize these results by setting a baseline "threshold" norm of what should constitute reasonable physical access to school and thus identify education deserts. For example, if policymakers wish to ensure that every child lives within three kilometers of a school (a commonly used international benchmark),[3] our framework can quickly identify what proportion of the population lacks this access, and precisely where those populations are located. Such insights allow for a more nuanced understanding of regional-level enrollment rates and potential barriers to greater enrollment, as well as changes in physical access over time. Third, using this same threshold definition for an education desert, our framework can algorithmically identify school construction sites that would most reduce the share of population living in an education desert and thus maximize the efficiency of school construction as a lever for improving educational access. To illustrate the potential value of this algorithmic optimization, we conduct a simulation analysis in Guatemala and find that building a mere 350 optimally-placed schools based on the algorithm's recommendations from 2008 data would have had the same impact on the share of population living in a public primary school desert as the 7000 schools that were actually opened in the ensuing decade. Finally, we provide guidance for analysts who wish to further refine these analyses to account for geographic factors like elevation, impassable terrain, and similar considerations.

Most importantly, we deliver all of these analytic components in an extensively documented open-source codebase alongside this manuscript designed around the goal of "plug-and-play" utility; assuming an analyst can obtain, at minimum, school location data for a given country context, the entirety of our main analysis can be replicated with minimal

[3] Theunynck (2009) notes that this norm is in line with the recommendations of the International Institute for Education Planning (IIEP) in Paris and the World Bank (Gould, 1978).

effort, zero cost (all requisite software and packages used in our analysis are also free and open-source), and only modest computational resources.[4] This code base is publicly available at: https://github.com/brhkim/mapping-education-deserts, from which the code can be downloaded, and adapted by other analysts. Indeed, while we focus on Guatemala for the body of this manuscript, we include in the appendix parallel analyses for Peru, Costa Rica, Tanzania, Kenya, Rwanda, and South Africa, as a testament to the portable nature of our analysis. Aligning our codebase to analyze each additional country takes as little as ten minutes, excluding time for the computation itself. And while our analysis is geared towards assessing the accessibility of schools, our codebase requires only clerical adjustments to instead analyze the physical accessibility of any other statically-located public good (e.g., vaccination sites, water sources, libraries, hospitals, etc.).

While our findings ultimately show that physical access to public primary schools is not a prominent barrier to universal school enrollment in Guatemala, we observe meaningful variation in the extent to which this is true across the country. Moreover, this analysis then offers empirical evidence to suggest that low regional enrollment rates in Guatemala are more likely the result of barriers besides physical access – insights that could prove invaluable for policymakers moving forward. In sum, we argue that as policymakers seek to traverse the last mile in school access and enrollment, fine-grain geolocated data infrastructure and identification algorithms like the one we propose here can offer enormous utility by ensuring that school investments are made in areas where they would have the highest returns in terms of educational access.

[4] We were able to replicate our analysis on a single consumer-grade laptop, which took approximately one hour to complete our main analytic components. Analytic extensions will take substantially longer depending on country size but should impose no additional hardware constraints.

The rest of the paper is structured as follows. Section II describes the background and conceptual framework for this paper. Section III describes the data sources and Guatemalan context we focus on to demonstrate our analysis. Section IV describes our main methodology, Section V reviews our main results for Guatemala, and Section VI describes how the main methodology can be expanded and adapted to produce additional analytic insights. Finally, Section VII explores the implications and possible applications of this analysis.

II. Background

IIa. School Proximity and Educational Outcomes

Previous research is clear in highlighting the educational benefits of policies that target school construction in areas which are underserved by educational institutions. In a meta-analysis of the effect of physical inputs on educational outcomes from 1990-2010, Glewwe et al. (2014) find that there are five high-quality studies on building new schools in developing countries, which all find consistently positive effects on enrollment and the time the students spend in school. More recently, Evans and Mendez-Acosta (2021) review 6 new studies on school construction in Africa since 2014, finding general increases in enrollment and learning across contexts, and highlighting that these programs seemed most effective when physical access to schools was indeed the binding constraint to school enrollment (e.g., in rural areas with few or no schools nearby). Similarly, in experimental work in Afghanistan, Burde and Linden (2013) find that the construction of community schools that decreased students' physical distance to school increased enrollment by 47 p.p., raised test scores by 0.59 standard deviations, and helped girls more than boys, nearly eliminating the gender gap in enrollment. Duflo (2001) also shows that school construction in places in Indonesia

where there were no or few schools led to returns to education of 6.8 to 10.6 percent in Indonesia – which in turn translated into long-run and intergenerational effects (Akresh et al., 2021) – and Koppensteiner and Matheson (2019) demonstrate that secondary school construction in Brazilian regions previously without schools led to a substantial decrease in teen pregnancy.

Not only is there evidence for the benefits of school construction on educational outcomes, but parents themselves seem to also favor school proximity. For instance, Solomon and Zeitlin (2019) run a discrete-choice experiment with Tanzanian parents, in which they find that parents indeed value outcomes (i.e., school test scores) and school proximity more than other inputs such as pupil-teacher ratios and desk availability. They find that the average travel distance to school in Tanzania is about 5 km, but that parents are willing to trade off more positive reported outcomes for proximity. For instance, parents are willing to send their children an additional 1.16 km for a school that scores about 8% higher over the mean on average on a primary exit exam. Conversely, similar work in Kenya by Ngware and Mutisya (2021) found that poor households often sent children to low-fee private schools because of physical convenience, as opposed to other factors like educational quality.

In all, these studies elucidate the idea that enrollment in these contexts is often negatively related to distance to education (i.e. that the distance elasticity of enrollment demand is negative) due to the logistical constraints and costs that greater distance imposes, a dynamic also well-studied in the contexts of U.S. higher education (see Alm & Winters, 2009, for a helpful review) and K-12 school choice markets (He & Giuliano, 2018). Thus, targeted school construction in areas where there are few or no schools seems to be, perhaps expectedly, a powerful way to improve school enrollment, as well as other important

indicators along the lines of learning, gender parity, and equality of opportunities more broadly.

IIb. School Proximity as One Barrier to Access of Many

In spite of the strong evidence in favor of building schools in remote areas with low physical access to schools, little is known about how researchers and policymakers can best understand the extent to which distance, specifically, may be a barrier to enrollment for certain sub-populations and geographic areas on a comprehensive scale. For example, while local school enrollment rates are often referenced as a primary metric of school accessibility, these measures could be driven by a variety of context-specific issues ranging from family finance, motivation, cultural priorities, as well as *physical access* – each of which require drastically different policy interventions in circumstances where resources for such interventions are scarce. Relying on enrollment rates to guide intervention in this manner then masks to a large degree the potential heterogeneity in physical access to schools by region, locality, or settlement pattern. In order to maximize the effectiveness and impact of any investments made in educational access across developing nations, policymakers would ideally be able to differentiate between the previously described scenarios using a data-driven, empirical approach.

As an illustration of this quandary, the World Bank reported in 2016 that 84% of all age-appropriate children in Tanzania were enrolled in primary schools (World Bank, 2016). It is nevertheless unclear what the barriers to access look like for the remaining 16%. One can imagine a scenario where these students *would* attend school if one were available, but currently lack access; conversely, it could be that they currently have physical access, but choose not to enroll for other reasons like fees or high opportunity costs. Both stories would

be consistent with the overall aggregate statistic, but they would require drastically different policy recommendations. In the case of the first scenario, policymakers might consider policies like investment in school construction and infrastructure, whereas investment in outreach campaigns or scholarships could likely be a higher priority in the second scenario. In short, without more fine-grain data than aggregate enrollment statistics, it is infeasible to systematically assess the varying educational needs in terms of increasing access to and enrollment in school.

In order to conceptualize the policy issue described here, we borrow the term "education deserts" in the spirit of Hillman (2016). Hillman's study uses data on the location of higher education institutions within commuting zones in the United States – defined by the U.S. Department of Agriculture as clusters of counties that form discrete labor market regions using detailed journey-to-work data (USDA ERS, 2019) – to identify communities that do not have reasonable access to higher education. While we focus on calculating actual distance to primary education in developing countries in the present analysis, the core of Hillman's analysis is the same as ours: the systematic identification of areas without physical access to education given a particular definition for access. More broadly, the international education literature refers to this type of rule regarding optimal school construction and placement as a "norm" (Theunynck, 2009; Lehman et al., 2013), and categorizes distance under the norm of "accessibility and efficiency." Previous policy and research efforts to establish these accessibility and efficiency norms have generally focused on selecting a maximum acceptable distance that children would be expected to travel to school, thus defining the "catchment area" for schools. For example, a commonly applied distance norm is to locate schools within a radius of 3 km from students' homes (Gould, 1978; Theunynck, 2009), though these numbers are often context-specific and can be sensitive to factors like

mountainous areas where the effort of traveling such distances can vary greatly. Another example is Lehman et al. (2013), who report that in rural Mali, the distance norm in 2004 was set at 5 km.[5]

While these norms have been pervasive in the theory underpinning school construction, it has long been difficult to actually implement them at scale into decision-making frameworks given the costly and time-consuming nature of collecting such data for any given locality. For instance, Lehman et al. (2013) set out to do this in Mali, across 12 of the country's 70 educational administration districts. Ultimately, only 8 of these 12 intended districts were successfully georeferenced by surveyors, identifying all the schools, villages, and hamlets within them. While the Lehman et al. (2013) report is an extensive and valuable effort to quantify physical access to schools, the dependence on in-person surveying of schools, villages, and population makes the marginal costs of including new areas using this methodology prohibitively high for many. This is true in terms of financial costs, as well as logistical difficulty for areas that may be too remote or afflicted by conflict.

III. Data and Study Context

IIIa. Data Specifications

Our main methodology, by contrast, requires only two critical data components: the locations of schools across a country (through pairs of latitude and longitude coordinates), and the geographic distribution of population across a country. For the methodological extensions that we articulate in this paper, we further incorporate data on elevation geography to examine the repercussions of alternate "pathing" algorithms to school, a

[5] If a reasonable estimate for the average walking speed of a 12-year-old is 5 km/hour (which is faster than for younger children), this would imply a two-hour, daily journey to school (Cavagna et al., 1983).

second wave of historical schools and population data to examine trends over time, and regional enrollment rates to facilitate comparisons across traditional and geographic measures of access.

School location data is perhaps the least standardized across contexts of our data requirements in terms of how countries report it, and stands as the primary barrier to replicating our analysis broadly. Still, this information is commonly obtainable through administrative records in many countries, either as latitude-longitude coordinates, or as physical addresses that are easily translated into coordinates through "geocoding." Recent grassroots efforts using commonly available modern technology have also shown that school locations can be "crowdsourced" in contexts where the government has not actively located where all the educational institutions are. For instance, Mulaku and Nyadimo (2013) describe the "Kenyan School Mapping Project," where the researchers identified and geolocated over 70,000 institutions across the Kenyan territory.

As is the case with any secondary data analyses, the exact process and scope of data collection for these administrative datasets will have meaningful repercussions for the robustness and interpretation of applications of our geospatial analysis. Therefore, researchers should be careful to interrogate these data accordingly before applying the algorithm we propose. For example, what are the formal conditions for a school to be included in the data? Are there relevant institutions likely to be excluded, such as private or parochial schools?[6] And how might such details affect specific areas, contexts, or

[6] Note that enrollment in private schools can vary widely by context. As an example, private school enrollment amounted to 82% of all primary school students in Belize (World Bank, 2019b). In such contexts, policymakers are faced with the additional choice of first reducing the number of people without access to *any* school, or prioritizing potentially more populated areas with access to only private schools where parents are burdened by higher private school fees.

populations differentially? Moreover, the concept of *location* should itself be interrogated. For example, if studying a context in which schools commonly have several linked campuses, or typically large campuses relative to the resolution of population data used for analysis, using a singular set of coordinates per school could understate access or imply unwarranted precision.[7]

For the purposes of this paper, we use government administrative data that focus exclusively on locating publicly-run primary schools in Guatemala in 2017 (Ministerio de Educación, 2020) and 2008 (SEGEPLAN, n.d.). We expect that other types of schooling in this context are valuable to consider when characterizing the broader landscape of education, but these publicly run schools as tracked by the government are likely the most policy-relevant sample to consider when analyzing, and intervening upon, the public's broad access to educational services. This is particularly true in the context of Guatemala, as primary enrollment in private schools was only 13% of the total primary enrollment in the country (World Bank, 2019).

Our geolocated, fine-grain population data come from the freely available "Global High-Resolution Population Denominators Project" datasets (WorldPop, 2018).[8] These layers provide estimates of human population distribution at a resolution of approximately

---

[7] Our provided code can account for multiple campuses so long as each are recorded as a separate observation in the school data. Importantly, though, note that recording data in this way assumes access to any one campus is equivalent to having access to any other campus (which would not be the case if a school has geographically separated academic and athletic facilities, for example). As accounting for large campuses would require a substantially different approach to our calculations, and we leave this task to future work where these features are a more critical factor in analysis. Anecdotally, such abnormalities are nearly unheard of in the context of Guatemala.

[8] The specific version of the data used for this analysis is known as the "Top-Down Unconstrained Individual Countries 2000-2020 (1 km^2 Resolution)" dataset. No changes to the algorithm would be required if the data used was the version with resolution at the 100 m resolution. However, this does increase computational time substantially. Analysts focusing on only one country context at a single point in time may opt to use the "Bottom-Up" datasets instead; we encourage all those interested to examine the trade-offs of these datasets closely before use.

100 or 1000 meters[2] for all years between 2000-2020[9]. The unusually fine-grain data comes from a combination of census and satellite imagery data, as well as careful application of machine learning algorithms (Stevens et al., 2015), developed through a partnership between School of Geography and Environmental Science at University of Southampton; the Department of Geography and Geosciences, at the University of Louisville; the Departement de Geographie, Universite de Namur, and the Center for International Earth Science Information Network (CIESIN), Columbia University. Discussion of their exact methodology is outside the scope of this paper, but the end result is that these data are highly standardized and available for nearly every country in the world at time of writing. In other words, the need to obtain these fine-grain population data to implement our proposed methodology should not pose a constraint for nearly any application.

One noteworthy feature of the Global High-Resolution Population Denominators Project is that they estimate both *overall* population within each gridded square, as well as *disaggregated* age-sex groupings, for each country. For our present analysis, this means that we are also able to isolate the population estimates to children of school-going age in this context, potentially avoiding some mismatch if relevant children are distributed distinctly from the overall population estimates. This feature will also be of use to researchers interested in other age demographics for certain school contexts (e.g., university-going age) or sex-specific policy margins (e.g., access to school specifically for female students).

That said, we still opt in the main body of this analysis to focus only on overall population estimates. This is because the methodology used to estimate these disaggregated figures impose substantially more functional form assumptions with respect to population

[9] WorldPop has not released a schedule of data releases for additional years going forward, but our best understanding is that these data are intended to be maintained over time.

growth and change over time (see Pezzulo et al., 2017). For example, if *migration* into and out of the various geographic units is heterogeneous with respect to age groups, or if such patterns are heterogeneous over time (as they use a singular base year to extrapolate population age pyramid ratios over time), it will be more difficult to ascertain how consistently accurate those population estimates are across a geographic context. For simplicity, and to make more transparent the limitations of the present analysis, we focus on the overall population estimates in the main body.[10] We conduct a sensitivity analysis in the Appendix to examine whether our estimates for Guatemala meaningfully change in response to using the age-specific data (children ages 5-14), finding that this distinction is completely immaterial for this particular context. We still urge analysts to consider and weigh this decision carefully for their own use-cases, however.

IIIb. The Guatemalan Context

While our main methodology should be broadly applicable given these relatively modest data requirements, we focus the current paper on Guatemala to showcase our approach for two primary reasons. First, Guatemala is a country which has historically struggled with an array of social challenges, and educational outcomes in Guatemala are particularly weak. For example, in terms of net school enrollment, 86% of school-age children were enrolled in primary school as of 2017 (compared to 94% in Latin America in 2017), and down from 94% in 2008 (World Bank, 2008; World Bank; 2017b). In terms of

[10] That said, our codebase can accommodate analysts interested in utilizing these disaggregated data simply by pointing the scripts to the disaggregated population dataset, instead. Note that the disaggregated data may require additional preprocessing if multiple demographics are desired (i.e. by adding the raster files together) and a resolution other than $100m^2$ is desired (as the disaggregated datasets are not provided "off-the-shelf" at $1km^2$).

learning, the World Bank estimates that 2 in 3 Guatemala children experience "learning poverty", meaning that they are not proficient in reading, even by the time they get to grade 6 (World Bank, 2019a). These challenges are typically worsened by the large inequities along ethnic and geographic lines within Guatemala (McEwan, 2007), given a very diverse geographic landscape with mountain ranges, lakes, and volcanos throughout the southern regions, and deep tropical jungle in more northern areas. Taken together, these challenges in terms of educational inequalities and physical characteristics make Guatemala an appropriate case study to pilot our methodology.

The second reason why we chose Guatemala is because of the public availability of all the needed data sets required for our main analysis and extensions. While our main analysis requires only a single year's worth of school and population data, additional data (such as multi-year school data) offer a useful opportunity to test the methodology's robustness and to assess the extent to which it offers new insight versus traditional measures. As such, this paper is best served by selecting a context that facilitates these valuable comparisons, as these additional data requirements do impose meaningful constraints to the exclusion of many otherwise viable contexts. Finally, note that we further test the "portability" of our method by conducting our main analyses in the contexts of six other developing countries in Sub-Saharan Africa and Latin America for which we could easily find data. We include this analysis in the appendix, and remark on individual data sources there.

Given our present focus on the Guatemalan context, we now move to describe the existing policies that relate to school construction norms to better understand the current business-as-usual. Unlike the distance norms we describe above, the current Guatemalan policy legislating school construction instead mandates where schools *can* be built, not where

they *must* be built (Acuerdo Ministerial, 2012). This policy imposes a dual norm: that schools cannot be built within 2km of one another, and they must serve a minimum number of potential students within their catchment area, which varies by educational level. For our case, primary schools must serve on average 25 potential students per grade in schools with separated grade levels, or 30 potential students per grade in schools with mixed grade levels. The policy moreover allows for a "deficit" of up to 5 potential students in total within a potential area for school construction.

The framing for Guatemala's school construction policy thus does not impose an automatic trigger policy on school construction, and instead places the burden of starting the process for construction on local governments and communities. Underserved communites must compile and submit comprehensive requests to the Ministry of Education with technical details on why the school is needed and how it meets the requirements set out in the aforementioned policy (see for instance, Municipalidad de San José, Chacayá, n.d., or Municipalidad de San José, Pinula, n.d.). We were unable to ascertain the exact process by which communities are mobilized from the ground up to submit these proposals, and by which these requests are ultimately approved in any public sources, academic literature, "grey" literature, news articles, or even anecdotal evidence. It may be the case that these processes are purposefully informal so as to provide the most flexibility for local policymakers to exercise their judgment and contextual knowledge. More cynically, we have evidence in the context of other developing countries that government inefficiencies (Batabyal & Nijikamp, 2004), lack of political representation, ethnic favoritism (Ejdemyr et al., 2018; Burgess et al. 2015), information asymmetries, and coordination problems may each ultimately play a role in the provision of public goods. In either case, it remains likely

that our proposed approaches for assessing and meeting school access needs in a data-driven manner provide novel insight against the current counterfactual in the Guatemalan context.[11]

IV. Main Methodology

The goal of our framework is to systematically identify areas of low physical access to educational facilities in a scalable and reproducible way. Our main methodology consists of a conceptually-straightforward algorithm which estimates the nearest distance from each population pocket to a public primary school, and then analyzes these distances in different ways to compute interpretable statistics and output. Specifically, the method follows these basic steps:

1. Load the fine-grain population raster data from the "Global High-Resolution Population Denominators Project," publicly available for all countries, discretized at either the 100x100m or 1x1 km plot level. Each discrete geographic unit will be treated as the basic unit of analysis, and each such observation contains an estimate of the number of people that live inside this unit.

[11] To provide a rough illustration, we conducted a supplementary analysis related to section VIc and examined empirically how many Guatemalan schools could be built that meet the stated policy requirements. In this exercise, we require that schools be built at least 2km away from one another and serve an age-relevant population of 175 (taking 30 students per grade, minus the allowed deficit of 5 students per grade, times 6 grades). In brief, we find that there are currently 1087 potential areas, with no overlap among them, where a school could be built while abiding by stated requirements. We estimate that if schools were constructed at all 1087 sites, these schools would reach 376,316 age-appropriate students total, or an average of 346 students each. Remarkably, we estimate from administrative data that the average *existing* public primary school in Guatemala in 2017 had 124 students, so many of these potential new schools would not be considered "small" in this context. Note that we opt to use the age-specific population datasets from WorldPop, ages 5-14, for only this analysis. While an imperfect alignment with the true primary age demographic, this seemed the most appropriate data to use for the exercise.

2. Load the school location data describing the latitude and longitude of each school.
3. Estimate the straight-line distance ("as the crow flies") between the center of each population unit and its nearest public school.

The output we obtain is a geolocated set of land plots with two key attributes: a) the estimated population living in each plot area, and b) the minimum distance from that plot to a public primary school. From this dataset, we can create several outputs to understand where the areas of low physical access, or "education deserts," are. Since these high-resolution population grids are much more disaggregated than even localized aggregate statistics on school access, we can pinpoint the specific areas where the distance to schools is prohibitively far. For our geospatial analysis, we use the excellent open-source R packages "sf" (Pebesma et al., 2021) and "raster" (Hijmans et al., 2020).

Our approach has three key advantages. First, it is very straightforward to implement and to understand conceptually, facilitating its broad use and easy interpretation by analysts and policymakers. Second, and relatedly, this analysis requires nothing more than a consumer-grade laptop and access to the internet, as all software involved (at least in the implementation we provide alongside this paper) are free and open-source. Third, the data it requires are readily available for many contexts. The fine-grain population data we use is available for virtually all countries in the world, at a resolution of 100 m^2, or 1 km^2 for faster computation. There are moreover other sources that take a different approach to estimating overall and subgroup population data for which our algorithm is also compatible.[12] And as

[12] For example, the High-Resolution Settlement Layer (HRSL) datasets, which are the product of a long-term collaboration between Columbia University and the Facebook Connectivity Lab (CIESIN, 2016; Tiecke et al., 2017). Their approach combines intensive survey work with advanced machine learning to estimate the population of every 30 x 30m block in a country, for almost every country worldwide. The disadvantage of this,

mentioned earlier, many governments already maintain administrative databases tracking the location of schools (such as Education Management Information Systems, or "EMIS"), which are often publicly available, either by default or on request.

The simplicity of our proposed methodology is an intentional decision to offer greater flexibility, allowing it to be adapted and responsive to specific contexts as necessary, but it also makes three important methodological choices that should be stated explicitly. First, the choice of population pockets at the 1 km^2 resolution clearly defines how granular and precise our analysis is. Although the population data that we use is also available at the level of 100 m^2 resolution, we observe similar results when this population layer is used, but with the important drawback of much higher computational times and memory limits that could put the analysis beyond the computational resources of many users. Ultimately, this decision should be for the user of the algorithm to determine given their context-specific knowledge and the policy action being considered.

Second, and relatedly, we assume that population is dispersed evenly within each geographic unit of 1 km^2 when we calculate distance from the center of each plot to each school. This is because if population is distributed evenly across a 1 km^2 plot, their average distance to school will be equivalent to the distance from the center of that plot, which is what we seek to estimate. That said, this assumption is obviously untenable and may serve to cause some measurement error in our process, but is done so for conceptual and computational ease as before Importantly, this issue becomes negligible when the resolution

admittedly more disaggregated dataset, is that the current data for most countries is for a singular year, meaning that if the school data does not match this year, there might be some meaningful mismatch in the analysis. In addition, the WorldPop has open-sourced all of their estimation procedure, code, and underlying data, making their population estimates imminently replicable. This exceptional transparency felt important to privilege and endorse given the nature of our work here, and the likely desire for future users of our code to conduct more rigorous population data diagnostics depending on their specific use-case.

is sufficiently small (as with the 100 m^2 resolution), and it is actually possible to use the finer-grain population data to "weigh" population within coarser-grain population data. Given what we observed when running our analysis at the 100 m^2 resolution, this assumption is unlikely to be consequential except in very specific cases.

Third, we choose to calculate distance using an "as-the-crow-flies" approach (i.e., a straight line connecting each population pocket to the nearest school). We recognize that this approach is most certainly an under-estimate as it may ignore geographic constraints such as swift elevation changes or lack of a clearly marked path or road. We discuss how to incorporate some of these features into our methodology in the extensions later. However, we decide to use to "as-the-crow-flies" as our baseline measure for several reasons. Much like in the discussion about resolution of the population data, computation time increases substantially by including these factors. Moreover, as we show in the extension later, we find that at least in the case of Guatemala, including elevation changes as a factor does not significantly change the results. Lastly, we believe that the inclusion of other constraints in the landscape should be context-dependent, as a mountainous country with a relatively low number of roads such as Bhutan may need different adjustments compared to a flat country composed of many islands such as the Maldives. As such, we default to the as-the-crow-flies approach and leave it to users to modify this base-level algorithm to their specific needs.

V. Main Results

We begin our proof-of-concept analysis by running our main algorithm using the Guatemalan population and primary schools data from 2017. Using the resulting data set, we create several outputs to better understand the nature of physical access to primary schools throughout the country. First, we examine the distribution of distances to school across the

whole Guatemalan population. We display this distribution in Figure 1.1 (Panel A). The median Guatemalan person lives 0.8 km from a public primary school, and the person at the 95th percentile lives 2.9 km from the nearest school. For comparison, this is lower than the median distance of 2.2 km in Tanzania, the same as in Kenya, and higher than the median distance of 0.5 km in Costa Rica (see the appendix for more details and contexts). This continuous measure can be dichomitized into the share of the population that lives further than a specific distance away from a school, and those that do not, to define the population living in an "education desert." This threshold distance for living in an education desert, effectively a distance norm, can be varied to explore the sensitivity of the dichotomous measure to different definitions/norms. We show this in Figure 1.1 (Panel B), where we calculate the proportion of Guatemalan population living in an education desert on the y-axis, at varying distance thresholds along the x-axis. For example, at a distance threshold of 1 km, 36% of the population lives in a primary school desert. Conversely, at a distance threshold of 5 km, only 1% lives in a primary school desert. For the most commonly used international distance norm of 3 km, only 5% of the population lives in a public primary school desert. Broadly speaking, Figure 1.1 suggests that prohibitive physical distances to school in Guatemala only affect a small share of the population, and that a relatively small but targeted school construction initiative might be effective at closing these access gaps.

Beyond quantifying the distribution of physical access to schools as an aggregated metric, our algorithm can also map out these distances to the nearest school for every square kilometer in the country. This type of figure serves as a visual primer on areas with greater and lesser physical access to school across the country, providing valuable insight on geographic heterogeneity in the aggregated measures we described above. In our map of Guatemala in Figure 1.2, we see that areas of low physical access (i.e., long distances to

school) are concentrated mostly in the northern region (Petén region), and in the southwestern region (around the Escuintla and Santa Rosa departments). We argue that such visualizations allow for far more contextual interpretation of these distance-to-school measures.

VI. Extensions to the Methodology

As mentioned earlier, the main algorithm we propose in the previous section is relatively straightforward by design to allow enough flexibility in its adaptation across contexts and educational levels. In other words, it could be extended in several ways to yield a more nuanced and tailored analysis for different policy questions in other contexts. In this section, we demonstrate four ways in which our methodology could be modified or refined accordingly. The replication files for all four extensions are likewise publicly available in our included codebase.

VIa. Before and After Comparisons

One of the simplest extensions that can be made in our framework is the analysis of physical access trends over time, a task we facilitate in our codebase and demonstrate here. In the Guatemalan context, we were able to obtain paired schools and population data for 2008 and 2017, allowing us to compare how physical access in the country has changed over the course of about a decade. Our data shows that between 2008 and 2017, the net number of public primary schools in Guatemala increased by 2,077, or approximately 15%. However, the Guatemalan population between the same period grew from 13.7 to 16.1 million people (18%). Therefore, at its face, the effect of the increase in the number of schools is ambiguous in terms of changes to the aggregate level of physical access to schools. Our

methodology can be used to compare two points in time, as we show in Figure 1.3. Figure 1.3 shows that even though population growth outpaced school construction, the distribution of peoples' distance to their nearest school shifted leftward, i.e., that physical access to school improved over time. That said, this fact should not necessarily be taken as a straightforward endorsement of school placement policy in that period, given that many factors may be contributing to this shift besides targeted school construction. For instance, in the extreme case where population growth was exclusively concentrated in high-density areas with existing schools nearby, the share of the population living far from schools would fall *mechanically* given that the relative share of people living near schools is rising relative to the pre-existing share of people living far from schools, even if no schools were constructed at all. Therefore, instead of being a standalone evaluation of the optimality of school placement over time, this method simply provides one measure for how physical access changed over time in aggregate.

VIb. Choosing a Distance Norm

Policymakers have typically relied on fixed distance norms or thresholds to determine whether a certain population pocket is within a school's catchment area (Theunynck, 2009; Lehman et al., 2013). This threshold is highly context-dependent, and should be chosen, if at all, by agents with rich knowledge of the specific geographical, infrastructural, social, and budgetary landscape. As such, our main algorithm does not take an ex-ante stance on what this threshold should be, or what constitutes an "education desert." However, the algorithm can be easily modified to accommodate a given distance norm for more in-depth analysis. This dichotomization has two main advantages. First, it most closely resembles the previous work on identifying areas as "education deserts," with

the added advantage that this task can now be done at scale in many contexts with minimal data and no surveying costs using our algorithmic approach. Second, it allows for quick identification of the most problematic areas given a certain threshold, offering a clear and interpretable "target" for policy intervention. For example, policymakers and their constituents may find it meaningful to ensure that all students in a given context live no further than X km from school.[13]

To showcase this extension to our main methodology, we choose a tentative threshold of 3 km in the Guatemalan context. Besides this being a common international distance norm, we estimate that just the cost of gas to cover even 3 km to school every day back and forth would lead to an expenditure of 4.4% (USD 7.40) of the average individual income per month in rural Guatemala, not taking into account school fees, books, bike maintenance, or other materials.[14] If instead students take the bus, the monthly transportation cost could be USD 5.20 or 3% of the monthly rural income.[15] These household expenses can start to look prohibitively high, especially for disadvantaged populations, further supporting the use of 3 km as a distance norm. This choice mirrors the spirit of Hillman (2016), where the author examines the distribution of postsecondary institutions across commuting zones in the United States as a proxy for access within a reasonable commuting distance.

Figure 1.4 shows the resulting geolocated "education deserts," as defined by a distance norm of 3 km, by plotting only those population points further than 3km from the

---

[13] For example, the Virginia Community College System advertises that, "If you are in Virginia, you are 30 miles from a community college" (Rorem, 2015).
[14] Assuming an efficiency of 45 km per gallon, an average cost of 2.75 USD per gallon, and an average income in rural Guatemala of 168 USD per month (Voorend, et al., 2017).
[15] Assuming a cost of 2 quetzales (0.13 USD) per ride (Cueva, 2020).

nearest primary school. The first panel pinpoints these areas on the map of Guatemala using color to moreover represent the density of population in each of these areas (white representing areas not in an education desert), while the second panel uses the additional dimension of height to more clearly display the relative populations of these deserts. These two panels taken together highlight an important distinction: while most of the land that constitutes "education deserts" is located in the northern regions (Panel A), the real concentration of the population in education deserts is generally localized in the southern regions (Panel B). These figures, much like Figure 1.2, can provide an important perspective for policymakers to decide where to strategically locate schools to increase physical access to education.

We can moreover examine the geographic distribution of population in a 3 km education desert and compare these insights against the information provided in traditional regional enrollment rates (defined in this case as the percent of age-appropriate students enrolled in primary school). Figure 1.5 (Panel A) displays the same information as Figure 1.4 (Panel A) except with regional enrollment rates underlaid in blue. What we observe immediately is that while some regions have high regional enrollment rates, they nonetheless contain several areas, of non-trivial population size, in education deserts. For example, the southern department of Escuintla (annotated with a red "A") has a fairly high enrollment rate relative to other departments, yet still has many pockets of education deserts. Conversely, Totonicapán (annotated with a red "B") has some of the lowest enrollment rates in the country, yet has no incidence of education deserts by our measure. We examine this relationship more explicitly using the basic scatterplot in Figure 1.5 (Panel B), where each region is plotted as a single point according to its population, proportion of population in an education desert, and proportion of age-appropriate children enrolled in primary school. If

enrollment rates were solely driven by whether people lived in an education desert, we would expect a perfectly negative relationship between regional enrollment rate and their share of population living in a 3 km education desert. Yet what we observe is only a weak relationship; running a simple population-weighted regression of the proportion of population in a desert on the proportion of age-appropriate population enrolled at the department-level, we estimate a coefficient on proportion enrolled of -0.32 (p-value of 0.04 and R-squared of 0.15).[16] This indicates to us, at least on a conceptual level, that our measure of physical access is providing novel information compared with enrollment rates alone, and that the picture remains complex and multi-faceted even after analyzing physical access as we do here.

VIc. Prioritization of School Construction Sites Based on Population

A natural extension of the identification of education deserts for a given distance norm is determining how to prioritize these areas given their relative population sizes. In other words, if policymakers were to invest in school construction, what construction locations would most reduce the share of population in an education desert? To do so, we propose an additional algorithm that extends our main methodology. After the main algorithm is applied, we use the previously discussed extension to identify the areas that fall outside of a given distance norm (i.e., the "education deserts"). Then, the new algorithm examines where a school could be constructed (within a 1 square kilometer area) to maximize new population reached given the distance norm. It is able to do this iteratively for

[16] Interestingly, this again implies a negative distance elasticity of enrollment demand per our literature review – albeit calculated using less direct proxy measures for both distance and demand. That said, an unweighted regression produces a non-significant coefficient of -0.24 instead.

any set number of schools to be constructed (i.e., it can produce any number of optimally-placed schools, always taking into account any previously placed schools for the next school). This process can be reiterated until the desired number of schools is reached (e.g., as determined by some budget constraint), or a minimum target of population reached by schools is reached (e.g., "for a school to be built, it needs to have at least X population within its catchment area"). Therefore, this approach is especially helpful to policymakers under constraint conditions: if the budget constraint only allows the government to build a given number of schools, and the goal is to maximize the number of people reached, then this approach can ensure a more efficient placement of schools. Similarly, this approach could be helpful if governments have tiered proposals to address issues of physical access to education. In other words, a government might require a minimum number of people served for a school to be built, and locations that fall below this minimum might be prescribed other policies like remote instruction (such as "telesecundarias" in Mexico).[17] In this case, this extension could help to quickly categorize localities at a large scale.

We test the efficiency of this algorithm at minimizing the share of population in a 3 km education desert by leveraging Guatemala data from 2008 and 2017 to conduct a simple simulation exercise: how different would the share of population in education deserts in 2017 look if Guatemala had used our algorithm in 2008 to determine new school placements instead of its business-as-usual procedure? To begin, we first conduct our main and distance

[17] Telesecundarias are "are a type of junior secondary school that delivers all lessons through television broadcasts in a classroom setting, with a single support teacher per grade" (Navarro-Sola, 2019). Although these schools were initially introduced to deal with issues of delivering education in remote areas, they are also used now in urban areas to deal with issues of poor teacher quality. While these schools do require certain personnel and a physical building, these requirements are less stringent in terms of teacher training and building size. For instance, Navarro-Sola (2019) mentions that the administrative cost per student of telesecundarias is half the cost of brick-and-mortar schools. As such, our algorithm can support the identification of areas that may be best served with a full-fledged school (with the logistical, staffing, and administrative requirements this might pose) versus a lighter investment like a telesecundaria.

norm analysis on Guatemala using population and primary school data from 2008, and a distance threshold of 3 km. Then, we run our school placement algorithm as described above given these data.

Once that analysis is complete, we determine how many schools Guatemala would have constructed in the time period between 2008 and 2017. Our dataset shows that Guatemala had a total of 14,033 public primary schools in 2008, but of these, only 9,040 remained open by 2017. Given that 16,110 schools were on record by 2017, we infer approximately 7,070 new schools were constructed by 2017.[18] To be realistic, we assume that policymakers in this exercise would not have known which schools in 2008 were going to close over the next decade, nor how the distribution of population would change by 2017. In other words, they choose to construct and place new schools based only on the "snapshot" of population in an education desert using 2008 data.

We find that if policymakers had placed *all* 7,070 new schools using our school placement algorithm and given these parameters, there would not be a single person living in an education desert by 2017; indeed, this feat would have been accomplished after constructing only 3,167 optimally-placed schools. That said, we recognize that there exist many other factors determining how new schools are placed, making this scenario fairly unrealistic. For example, Panel A of Figure 1.6 shows the cumulative new population reached per new school constructed, demonstrating the quickly diminishing returns to each additional optimally-placed school. This panel also highlights the important caveat that each

[18] Note that these numbers come from the presence of schools by their unique administrative ID in either data set (2008 or 2017). However, if schools simply had their unique IDs changed over this period (e.g., if they merged with another school, took on an additional level, etc.), we would still consider this as a school closing, and another one opening by this tallying method. That said, the precise number of new schools we estimate here is not hugely consequential, given the nature of the results we describe later.

additional new school would likely lack the requisite student body to justify new school construction well before this benchmark was reached (because building a school to serve a single person would not actually happen).

To explore a more realistic scenario, we proceed to ask the following question: given that the proportion of Guatemalan population in an education desert actually did decline from 2008 to 2017 after the 7,070 schools were constructed (see Section VIa above), how few optimally-placed schools would it take to produce this same reduction? Panel B of Figure 1.6 displays the results of this thought experiment. The blue line shows the share of Guatemalan population in an education desert across varying distance thresholds, for the actual schools that existed in Guatemala in 2017 – essentially, our target to meet. The red line shows this same dynamic, but under the hypothetical circumstance that Guatemala had constructed *no* new schools at all between 2008 and 2017 – serving as our reference baseline. We find that it would take only 350 new optimally-placed schools to match the actual reduction of population living in a 3 km education desert by 2017, the hypothetical circumstance represented by the green line. Put another way: *350 optimally-placed schools had the same impact on the share of population in an education desert as the 7,070 schools actually built between 2008 and 2017*. We take this finding as especially hopeful and actionable for policymakers because it roughly indicates that – at least in the Guatemalan context – substantial strides in physical access can be made even if only *one in 20* schools are constructed with physical access in mind. Conversely, it also makes clear that even a large amount of school construction may not necessarily increase physical access to school across the country by default (e.g., new schools are built in locations already being served by other schools). Policymakers are the best suited to determining when and to what extent physical access

should be a consideration for new school construction, but so long as it remains even a minute priority, progress can be made with the help of these proposed algorithms.

VId. Elevation and Geographic Features

Our main algorithm relies on estimating distance "as-the-crow-flies", or a completely linear trajectory between the population pockets and school locations. This approach has three key advantages. First, it is a simple and straightforward measurement choice that allows for easy conceptualization of the way in which distance was measured and minimizes the number of contextually-dependent assumptions made about travel patterns, infrastructure, etc. Second, it makes computation vastly faster than other approaches (like the extension we will discuss here). Third, it does not require additional data layers besides what we have described before: solely population data and school locations. Still, all of these advantages come at the expense of ignoring potential barriers like geographic features or lack of roads connecting two places in a fairly linear fashion.[19]

Therefore, we showcase an extension of our main algorithm where we consider elevation changes and compute the "path of least resistance" between a population pocket and a school.[20] Put simply, we first obtain elevation data across Guatemala from ArcGIS's online servers (gspeedAIST, 2019) – though note that robust elevation data are universally available for all regions of the world from a variety of sources. Using these data, we can then calculate how elevation changes when moving from each geographic cell to each adjacent

[19] While roadways are an attractive feature to consider, geographically heterogeneous data availability and reliability, as well as computational complexity and costs, make such analysis infeasible and potentially biased for certain contexts (e.g., if roadway data is more complete and accurate in regions of higher income). Given our intention to provide a broadly applicable and easily accessible toolset in this paper, as well as the methodological concerns such analyses present, we opt not to explore this style of analysis ourselves.
[20] We leverage the implementation offered by van Etten & Sousa (2020) in the R package "gdistance."

cell.[21] As in our main algorithm, we calculate distances between each population point and each nearby school; however, we instead calculate the distance of the route that minimizes *walking time* after accounting for the fact that speed is inversely related to the steepness of the terrain's gradient (per Tobler's Hiking Function; Tobler, 1993). Any sufficiently steep gradient is considered impassable and avoided for any routing entirely. In practice, this might take different forms. If there is a very large mountain between a school and a population pocket, the "path of least resistance" is likely around the mountain. If instead there is a very small hill between these two areas, the path of least resistance might still be a straight line over the hill (depending on the elevation of the hill and its circumference), instead of going all the way around it.

After incorporating this extension to our algorithm, we compare the results to our main results using the as-the-crow-flies methodology for Guatemala in 2017. Figure 1.7 (Panel A) plots, for each population pocket, the estimated distance to school using the as-the-crow-flies methodology (x-axis) against the estimated distance to school consider the path of least resistance (y-axis). For visual clarity, we bin observations and scale color according to the sum of population in that bin. The vast majority of population indeed cluster close to the 45-degree line in red, meaning that for nearly all cases, the difference in distance between the two methodologies is small.[22] In fact, Figure 1.7 (Panel B) displays the distribution of the difference in estimated distances between the two methodologies. The vast majority of the observations fall below a 20% difference between the two

[21] For computational tractability, we use elevation data at a resolution of 500m^2. Finer-grain data allow for more nuanced pathing, but also drastically increase computational time and the likelihood of hitting software memory storage constraints.

[22] Note that in all cases, the distance for the algorithm that takes into account elevation is equal or larger than for the main algorithm, since the main algorithm computes a straight line connecting two points.

methodologies. Therefore, in the case of Guatemala, accounting for elevation does not make much of a difference in the identification of where education deserts are, and may come at the expense of increased barriers to analysis (e.g., data requirements, computational costs). However, this extension might be particularly valuable for other hilly or rugged contexts like Rwanda. Importantly, the estimation of the path of least resistance can also accommodate further geographical barriers such as accounting for internal bodies of water or impassable national parks.[23] In this sense, this extension provides the most flexibility to further adapt our main algorithm to local conditions, at admittedly much longer computation times.[24]

VII. Discussion

In this paper, we propose a framework to identify populated areas that are not served by public primary schools in developing countries, where surveying costs may be prohibitively high and other types of administrative data may be lacking. We use Guatemalan data as a proof-of-concept to identify geographic areas within the country where individuals lack physical access to primary schooling, as well as to showcase some of the useful extensions we propose to our main methodology. We find that education deserts, defined as pockets of population outside of a school's catchment area, are somewhat rare in Guatemala, and that a relatively few but strategically placed schools could significantly universalize physical access to education.

---

[23] These could be incorporated in two ways. The first option would be to clip "holes" in the population and elevation raster data files using layers that signal where the national park or water bodies are. The second option would be to change the elevation of these impassable areas to an unrealistically high number. This way, the algorithm will never consider these as viable routes while searching for the path of least resistance.

[24] Conducting this analysis for Guatemala took our workstation computer approximately 9 hours, compared with only 30 minutes for the main analysis. Moreover, we expect the computational time of this extension to increase exponentially with country area.

This type of disaggregated, fine-grain analyses can be especially valuable as policymakers and investors around the world attempt to guarantee universal access to education. If indeed a country has pockets of population in remote areas where there are no schools, and information is not readily available on where new schools could be more impactful, then it is not clear how to make these investments in a way that creates as much social welfare as possible. Unfortunately, the regions where it is most important to identify education deserts are often the same regions where traditional, aggregate administrative data is typically most lacking. In such circumstances, policymakers would need to resort to either costly surveying endeavors, or fall back on analyses aggregated in larger regions that could critically mask meaningful heterogeneity within those aggregations. By strategically locating educational institutions using these finer-grain analyses and their own contextual expertise, policymakers can indeed ensure that all populations are served by such reforms, at least in terms of physical access to a school.

That said, primary school access is far from the only frontier in which physical access is a relevant consideration for equity and social welfare, and most of the data required to replicate this style of analysis in similar circumstances is publicly available or is of easy access to researchers and policymakers. We thus create and make available highly documented and portable code as a public good for others to recreate and extend our analysis to other contexts. Applying our codebase to analyzing primary school access in additional countries (as we show in our Appendix) can take as little as ten minutes, excluding time for data acquisition and computational processing. Similarly, applying our codebase to analyzing the parallel issues of secondary school access – an increasingly prominent goal for many development organizations and governments (Cosentino, 2017) – or postsecondary institution access should be equally straightforward. While outside of our expertise, we also

ensured that the codebase should be fully capable of applications to other statically located public goods, for example libraries, health institutions, vaccination facilities, water wells, and so on. In short, if a good can be meaningfully characterized by a coordinate, one can apply our code to better understand a population's physical access to it.

But in closing, we will caution that while applying the code to said contexts should be nearly costless from a logistical perspective, any such analyses should still attend to the many important contextual and data quality considerations we have outlined in this article. For instance, it remains an important critique of our approach that we assume the costs associated with traveling a kilometer in one geographic area is equal to the costs of traveling a kilometer in another geographic area. On its face, this assumption can be entirely untenable – whether comparing within the same country, same region, same city, or even same neighborhood – even after accounting for elevation as we do in the extension analysis above. Analysts must then be cognizant of how their own context and data constraints relate to the value of such analysis in spite of this assumption. To put it concisely, we subscribe to an adapted version of the old adage: if all you have is an education desert mapping tool, everything may look like an education desert problem. We thus ultimately hope that analyses stemming from our methodology provide an *additional* source of insight for researchers and policymakers, to be understood and contextualized in concert with many other sources of evidence, to better serve the public and their well-being more broadly.

Figure 1.1 (Panel A). Distribution of distance to nearest school across Guatemalan population

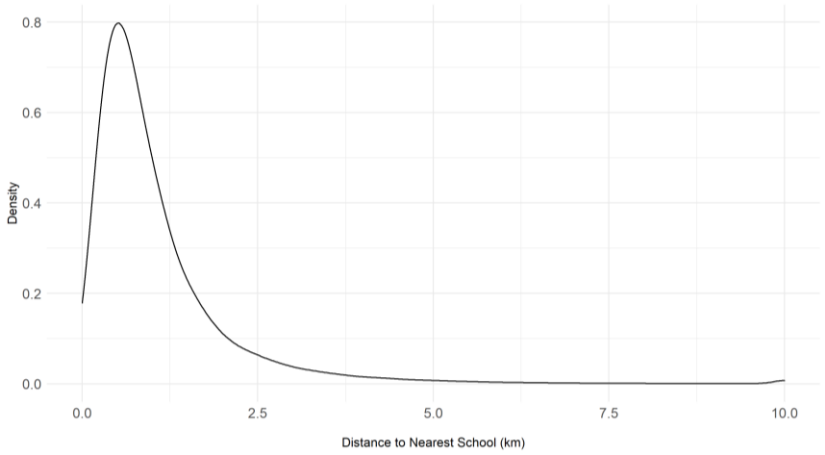
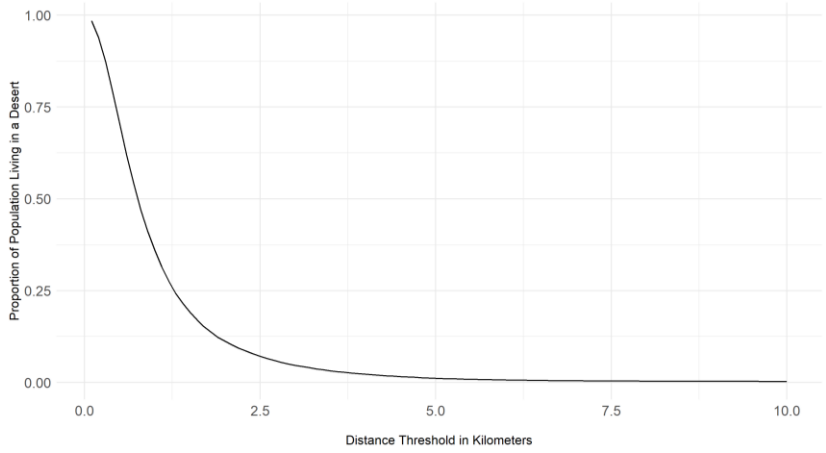
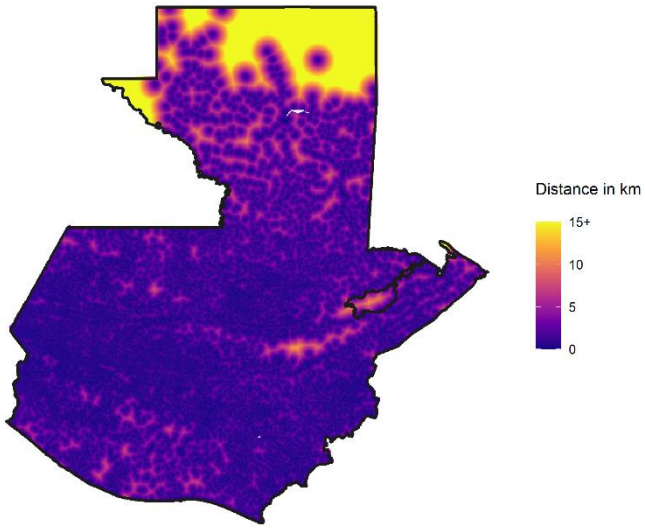


Figure 1.1 (Panel B). Proportion of Guatemalan population living in education desert at varying distance norms



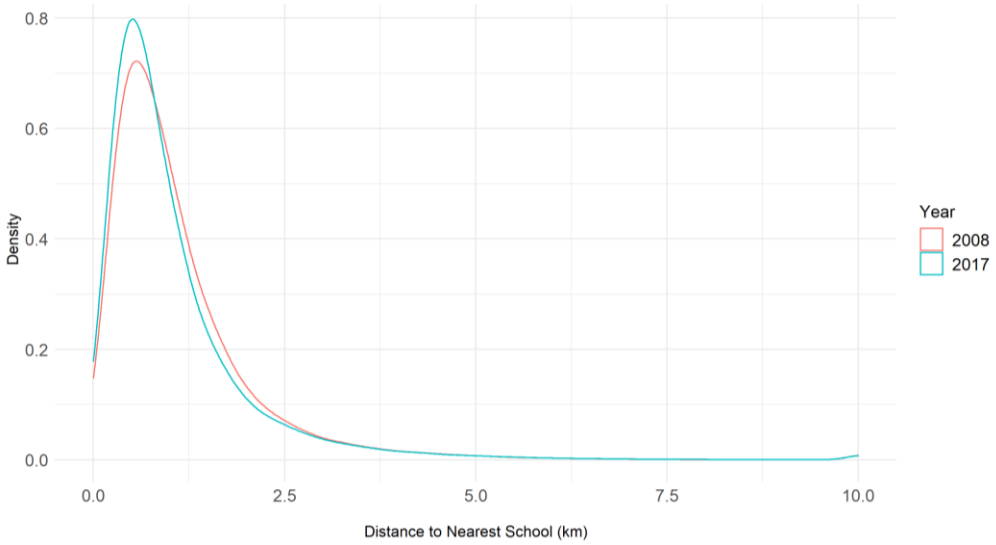
Note: Sample subsets to only public primary schools in 2017 in Guatemala.

Figure 1.2. Heatmap of distance to nearest public primary school by population



Note: Primary school and population data from 2017. Distance is measured as-the-crow-flies from the center of each population plot to the nearest primary school.

Figure 1.3. Comparison of the distributions of distance to nearest school across Guatemalan population in 2008 and 2017



Note: Analysis limited to public primary schools in Guatemala for the years shown.

Figure 1.4 (Panel A). Geographic distribution of Guatemalan population, including only those population points at least 3 km away from a school

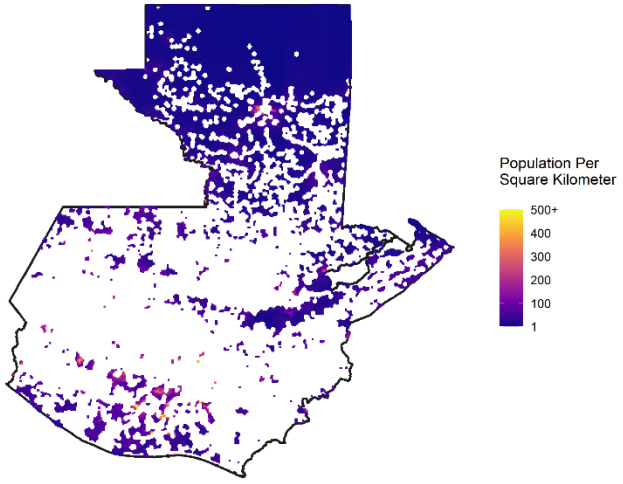
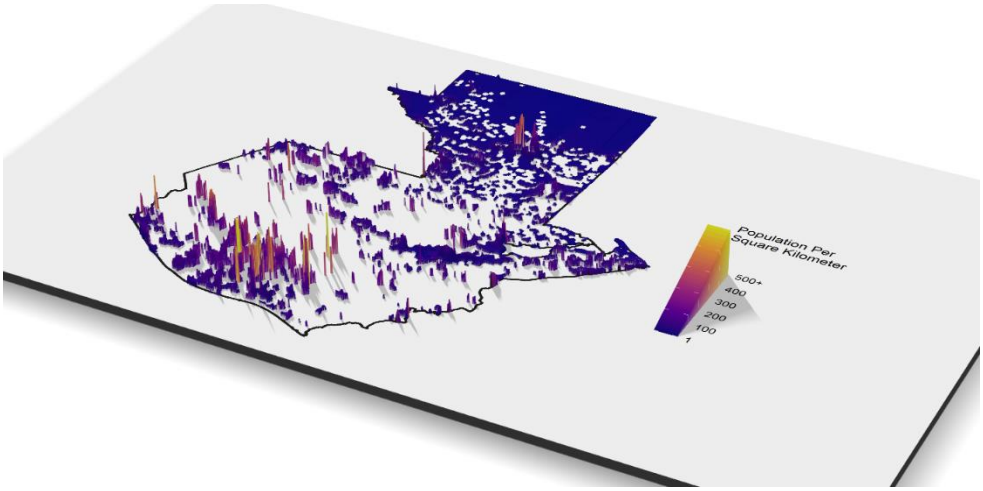


Figure 1.4 (Panel B). 3-dimensional geographic distribution of Guatemalan population at least 3 km away from a school



Note: Sample subsets to only public primary schools in 2017 in Guatemala. Panel B represents the exact same information as Panel A, but plots both color and height to population count to better visualize differences in population than color alone. 3D visualization made possible by the "rayshader" package from Morgan-Wall (2021).

Figure 1.5 (Panel A). Geographic distribution of Guatemalan population at least 3 km away from a school against regional enrollment rates

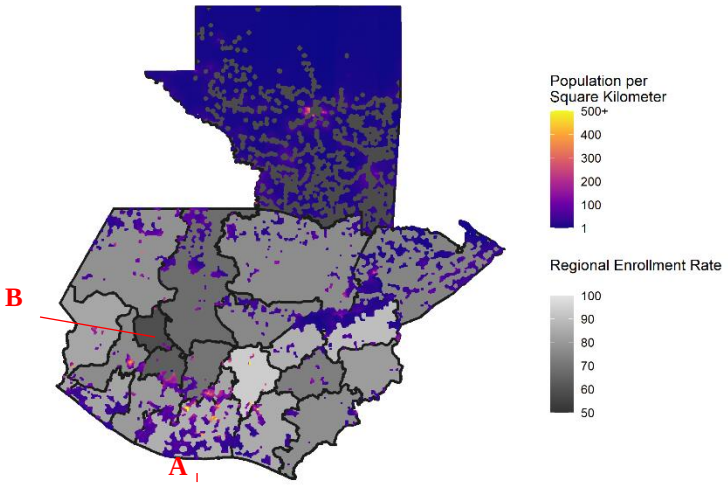
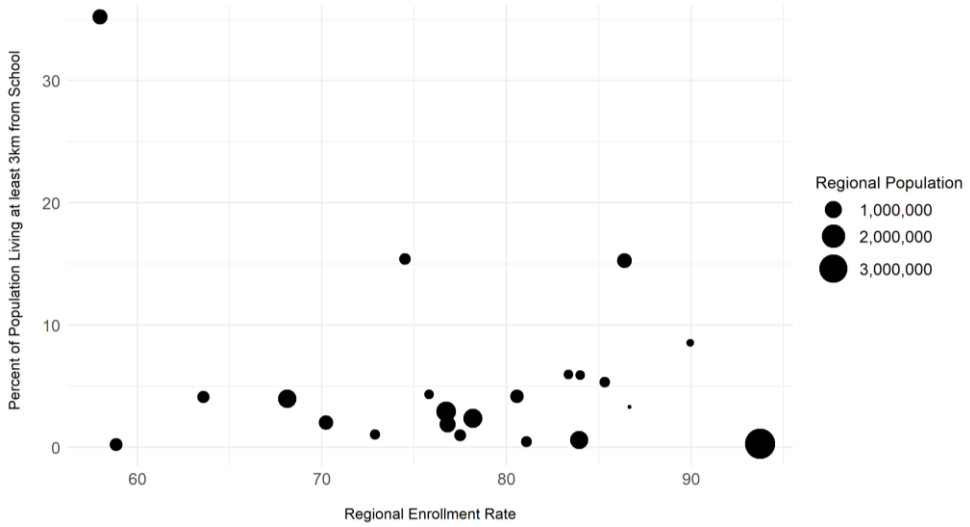
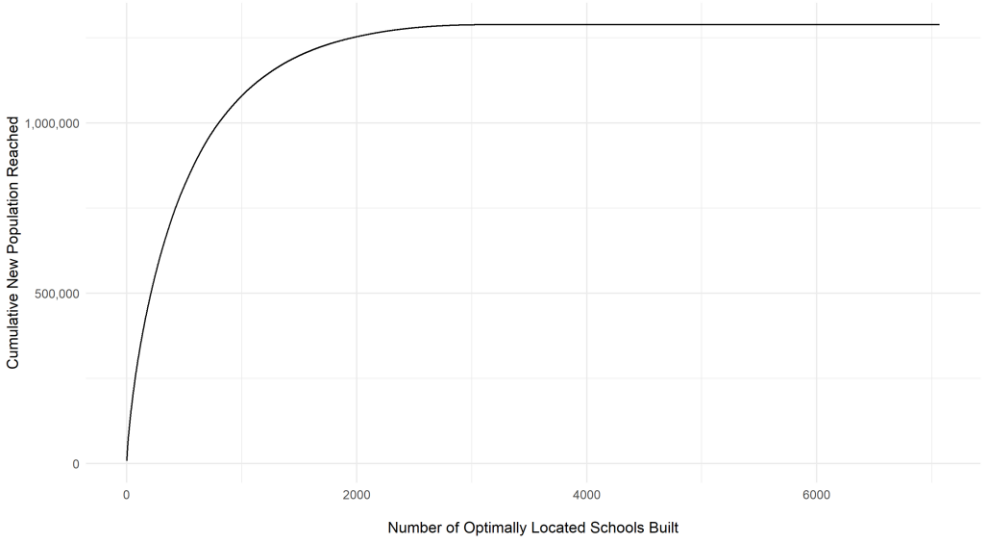


Figure 1.5 (Panel B). Scatterplot of regional enrollment rates against percent of regional population living at least 3 km away from a school



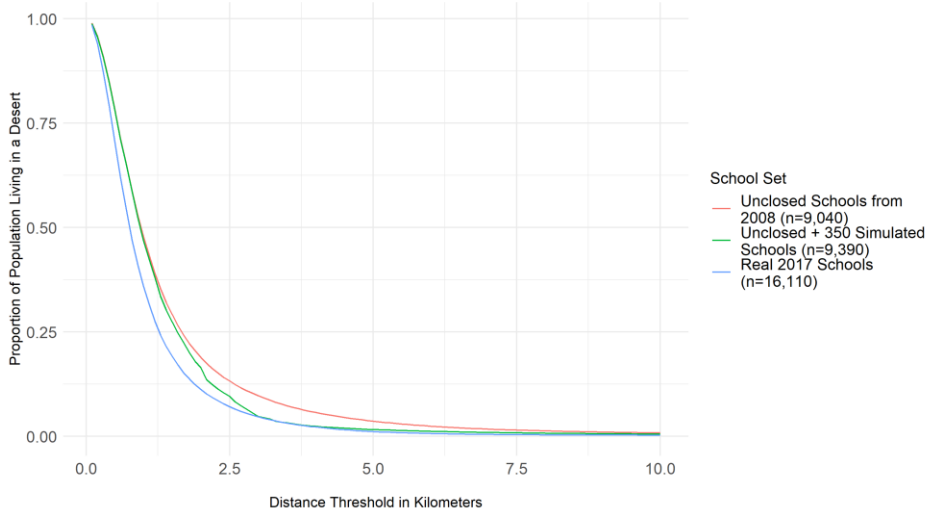
Note: Sample focuses on only public primary schools in 2017 in Guatemala. Enrollment data were collected in 2016. Panel B displays each region of Guatemala as a point, plotting its population (in point size), proportion of population in desert (on the y-axis), and proportion of age-appropriate children enrolled in primary school (on the x-axis). When running a simple population-weighted regression of the proportion of population in a desert on the proportion of age-appropriate population enrolled at the department-level, we estimate a coefficient on proportion enrolled of -0.32 (p-value of 0.04 and R-squared of 0.15).

Figure 1.6 (Panel A). New population reached per optimally located school



Note: For simulated public primary schools in 2017.

Figure 1.6 (Panel B). Comparison of the distribution of the Guatemalan population living in an education desert in 2017, across several real and simulated school construction scenarios



Note: Population data used are from 2017 regardless of school construction scenario.

Figure 1.7 (Panel A). Comparison of "as-the-crow-flies" distances with distances calculated using the "path of least resistance" through elevation changes

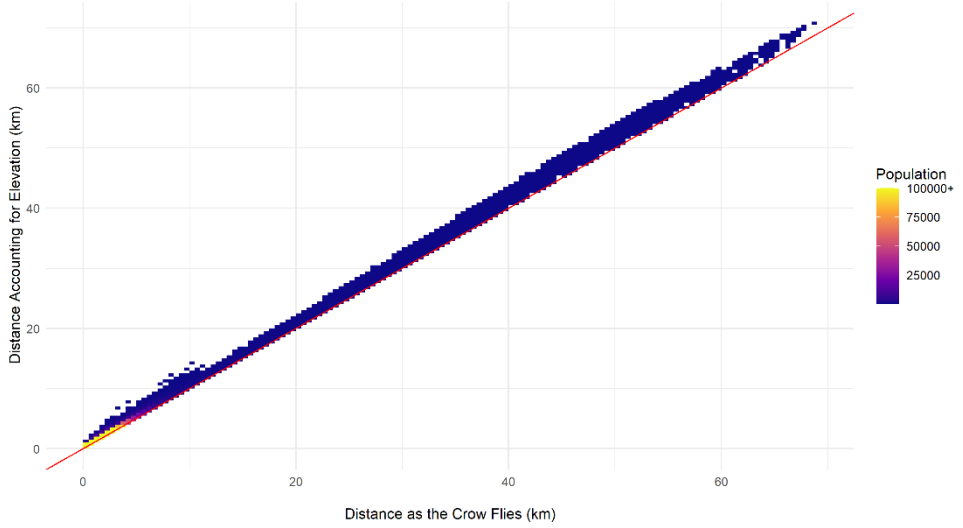
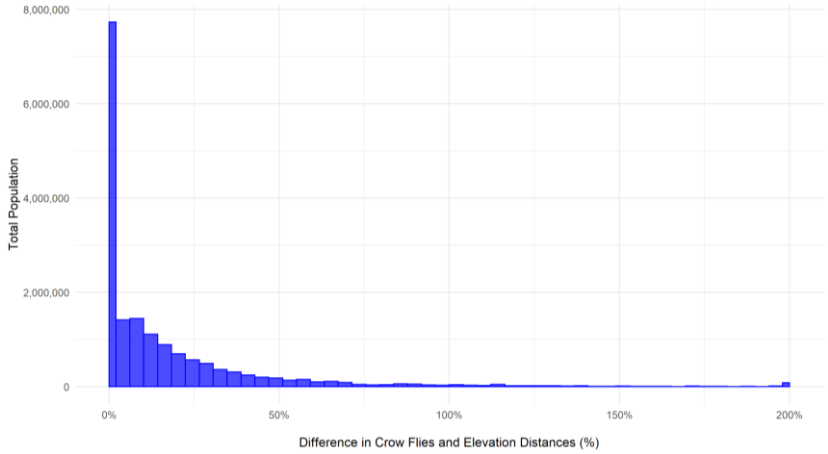


Figure 1.7 (Panel B). Histogram displaying the distribution in the difference between "as-the-crow-flies" distances with distances calculated using the "path of least resistance" through elevation changes



Note: Sample subsets to only public primary schools in 2017 in Guatemala.

CHAPTER 2

Gauging Engagement: Measuring Student Response to a Large-Scale College Advising Field Experiment

Brian Heseung Kim, Katharine Meyer, Alice Choe

Abstract

Interactive, text message-based advising programs have become an increasingly common strategy to support college access and success for underrepresented student populations. Despite the proliferation of these programs, we know relatively little about *how* students engage in these text-based advising opportunities and whether that relates to stronger student outcomes – factors that could help explain why we've seen relatively mixed evidence about their efficacy to date. In this paper, we use data from a large-scale, two-way text advising experiment focused on improving college completion to explore variation in student engagement using nuanced interaction metrics and automated text analysis techniques (i.e., natural language processing). We then explore whether student engagement patterns are associated with key outcomes including persistence, GPA, credit accumulation, and degree completion. Our results reveal substantial variation in engagement measures across students, indicating the importance of analyzing engagement as a multi-dimensional construct. We moreover find that many of these nuanced engagement measures have strong correlations with student outcomes, even after controlling for student baseline characteristics and academic performance. Especially as virtual advising interventions proliferate across higher education institutions, we show the value of applying a more codified, comprehensive lens for examining student engagement in these programs and chart a path to potentially improving the efficacy of these programs in the future.

I. Introduction

Despite the high economic returns to college completion (Avery & Turner, 2012; Carnevale, Jayasundera, & Gulish, 2016), just over half of students who enroll at college have attained a bachelor's degree (Bound, Lovenheim, & Turner, 2010; Denning, Eide, & Warnick, 2019; Shapiro et al., 2016). Colleges and non-profits have invested in various strategies to improve college completion, ranging from resource-intensive advising (Scrivener et al., 2015) to light-touch messaging campaigns (Castleman & Page, 2016). More recently, organizations have implemented hybrid text-based advising models that provide light-touch supports *as well as* personalized advising to students after matriculation (Gurantz et al., 2020; Oreopoulos & Petronijevic, 2018; Page & Gehlbach, 2017; Sullivan et al., 2019).

Given that we have increasing evidence for the limited effectiveness of *one*-way texting campaigns (a la many nudging interventions) when scaled to broader contexts and populations (Bird et al., 2021), we might expect the majority of benefits from these text-based advising campaigns to then come from the *two*-way interactions between students and their advisors. Despite this intuition, past evaluations have found that the effectiveness of two-way text-based advising in improving college persistence and graduation *also* varies across contexts and students (Avery et al., 2021). It remains unclear exactly what factors contribute to this variation in program effectiveness, but one potential explanation that remains underexplored is the variability in student engagement both within and across interventions. If the two-way interactions between students and advisors are the key mechanism for program effectiveness, the extent of student engagement in any text-based advising program can be thought of as a form of endogenous intensity for the intervention, where some students choose to engage in and "receive" the intervention at greater intensities than others.

While some scholars have begun attempting to document variation in student engagement in such text-based advising campaigns (e.g., Fesler, 2020; Arnold et al., 2020), how best to operationalize and measure student engagement remains uncodified, and the relationship of any such engagement with student outcomes is still ambiguous. And though there exist some consistent measures of engagement used in interactive text campaigns in other contexts (e.g., smoking cessation texting programs), these tend to be simplistic and overly focused on response rates. A deeper understanding of appreciable differences in student behaviors can help scholars generate and test hypotheses around which behaviors are potentially malleable through an intervention and would lead to improved student outcomes. For example, if we observe a strong positive correlation between academic persistence and students who more frequently solicit assistance from their advisors, this insight could motivate future research into the causality of this relationship through thoughtful experimental design (i.e., conducting an interactive messaging experiment where one treatment wing more explicitly solicits student requests for assistance in its prompts). Our descriptive exploration is then intended to highlight future venues for research that could, eventually, refine the design of two-way text advising programs and improve their efficacy. Such design insight has become increasingly important to gather as an increasing number of higher education institutions have turned to remote advising practices and campus-wide text messaging campaigns to support their students.

In this study, we seek to refine our understanding of student engagement in text-based advising using a variety of data- and text-mining (i.e., natural language processing, or NLP) techniques to analyze student-advisor interactions from the Nudges to the Finish Line (N2FL) text-advising intervention (Bettinger, et al., 2021). N2FL was randomly delivered to students approaching degree completion at over twenty colleges and universities in five

states across the U.S. Treated students in the intervention received pre-scheduled and pre-written messages (what we herein refer to as "scheduled messages") that provided information about important deadlines and encouraged use of campus resources like academic tutoring centers and financial aid. Because N2FL was designed as a two-way interactive campaign, the scheduled messages encouraged students to write back and engage in impromptu conversations with advisors at their campus (e.g., "...Do you need help with applying for financial aid?").

To quantify variation in how students engaged with the intervention, we employ a wide array of measures related to the intensity, duration, response speed, and content of student-written messages (what we herein refer to as "student replies"). For example, we examine the proportion of scheduled messages students responded to, the number of requests for help students made, the positivity or negativity of student texts (known as "sentiment analysis" in NLP), and the extent to which students discussed various topics of conversation (known as "topic modeling" in NLP). We rely on approaches from prior literature to form these measures, while also constructing several novel measures to explore. In general, we find wide variation in nearly every measure we construct, emphasizing the variation in students' interaction with text-based advising interventions. We moreover interpret measures with wide variation as more likely to be malleable (i.e., when compared with measures that have extremely narrow distributions across students), though we are unable to explore this explicitly in this study. We also find that this array of engagement measures tend to be uncorrelated with one another, indicating that patterns and intensities of student engagement are multi-faceted and are unlikely to be well-captured using simple response rates as is the current business-as-usual in the field.

After describing the construction of these measures and examining the extent of variation in student engagement along these measures in N2FL, we proceed to descriptively examine how these measures correlate with the outcomes of interest in the intervention: re-enrollment, credit accumulation, GPA, and degree receipt in each term following the start of the intervention. We find that even after controlling for student baseline characteristics and academic performance, there exist large and persistent relationships between engagement measures and student outcomes. For example, greater frequency of student engagement was strongly related to credit accumulation, GPA, and eventual degree receipt. We also observe higher GPA and credit accumulation among students whose responses were more positive in tone, and more mixed relationships with outcomes based on the topics of discussion students brought up in their replies. While not causal, these results suggest several pathways for future research to more explicitly test the malleability and impact of these more promising engagement measures.

Taken together, our findings offer two main contributions to the field. First, this analysis demonstrates the value of examining text interaction data and student engagement in the context of two-way text-based interventions. We find meaningful variation in engagement behaviors. That engagement varies implies the opportunity to change engagement behaviors, highlighting opportunities to test different strategies to affect student engagement. While this is a descriptive study and student engagement patterns in this context are endogenous, we nonetheless view this as an important initial step in identifying the engagement patterns correlated with college persistence that future interventions might target through careful programmatic design. An increasing number of higher education institutions are turning to text-based advising as a cost-effective tool to reach students at scale. Without a stronger understanding of the potential mechanisms underlying text-based

support programs' efficacy, institutions run the risk of designing interventions that do not meaningfully engage students and therefore do not meaningfully improve student outcomes. This paper thus draws important attention to student engagement as a valuable and multi-dimensional mechanism demanding more consistent study in program evaluation and design.

Second, we add to the growing literature showcasing robust text-as-data/NLP methods to enhance our understanding of large-scale educational text data and field experiments. The measures we deploy here are imminently usable across any two-way texting intervention given the common collection of text interaction data via large-scale messaging platforms (e.g., Signal Vine), and we "open-source" our code and methodology to facilitate continued iteration on these engagement measures more broadly. Versions of this framework are also imminently applicable to many two-way educational interactions captured in text, such as discussion board posts, email exchanges, virtual tutoring, and other texting-based interventions. Our hope is to advance the field's ability to apply more comprehensive and codified engagement measures more broadly. Through a combination of simple (e.g., keyword-based) and sophisticated (e.g., neural network-based) NLP methodologies, we are able to provide nuanced insights about how students engage in the two-way advising program – both *at scale* and *in real-time*. While close qualitative reading of the text conversations in this intervention would be instructive in its own right, the scale of analysis and rapidity of insights afforded by these automated text analysis systems (e.g., using a single analyst to examine thousands of text messages almost immediately after their receipt) allow us to gain a far greater understanding of interventions both after the fact and as they happen – unlocking immense potential in program design and evaluation going forward.

In Section II we summarize insights from other studies of engagement in text-based outreach, and in Section III we describe the intervention context and student sample. We

outline our methodology in Section IV and share results in Section V. We conclude with a discussion of our results and the implications for future research and practice in Section VI.

II. Prior Literature on Measuring Engagement

Despite a proliferation of text message intervention studies across domains (e.g., education, healthcare, finance, political campaigns), the literature is sparse and disconnected around how to define participant engagement as a construct. Most commonly, studies adopt simplified definitions that rely on aggregate response rates, such as overall responsiveness rate (e.g., "high" is ≥90%, "good" is ≥70–<90%, and "low" is <70%, per Zhang et al., 2018), absolute thresholds for response counts (e.g., 10 or fewer responses, 10-20 responses, or more than 20 responses, per Irvine et al., 2017), and responsiveness by message category (e.g., prompted versus unprompted text messages, per Psihogios et al., 2019).

Other studies have also chosen to define engagement in more contextually-specific ways, further complicating the process of developing a more unified definition. For example, Nelson (2020) initially presented participants with a choice about how frequently they wanted to receive text messages; accordingly, they defined engagement using a joint measure of participants' stated preference for messaging and their ensuing response rates. Another two-way text messaging study designed to reduce binge drinking defined engagement, in part, by manually coding participant responses for whether the informational texts were understood correctly (e.g., participants responded with personal and specific details that demonstrated they cognitively processed how the messages related to their own lives) and whether responses were related to specific components of the behavior change theory that the researchers were testing (e.g., increasing the salience of the perception of harm, encouraging goal setting, subjective norming) (Irvine et al., 2017). In a texting program for

smoking cessation, researchers measured the frequency, length (i.e., more than 30 characters), common themes (e.g., well-being, self-efficacy, and reasons to quit smoking), and use of keywords (e.g., "stress" or "alcohol") in recipient responses throughout the intervention (Cartujano-Barrera et al., 2019). Research on broader virtual engagement contexts (e.g., online shopping, video games, etc.) has also sought to measure the emotional affect (e.g., positive versus negative) and "window" of interactions (i.e., point of engagement, period of sustained engagement, disengagement, and reengagement; O'Brien and Toms, 2008).

Even focusing on programs conducted in the education arena, definitions of text message engagement vary. Following text-based outreach to parents that promoted literacy activities for their children, York, Leob, and Doss (2018) conducted a separate survey of participants to determine whether they read and/or used the text messages, found them helpful, and shared them with other parents. In another two-way interactive text campaign designed to provide parents with personalized information about their children's school attendance, Smythe-Leistico and Page (2018) generally scanned parents' inbound text messages to a single staff member and found that most messages were questions about school schedules or requests to relay information to teachers. Castleman et al. (2017) used coarse measures of punctuation to identify the frequency of students asking questions in a financial aid filing messaging campaign. Furthermore, in a two-way texting intervention designed to reduce summer "melt," Castleman and Page (2015) looked at response rates broken out by experimental groups and sites, such as the share of students who replied to at least one text message and the share of students who replied to at least one message to request an advising meeting.

Overall, the literature to date suggests relatively little consensus around how engagement in text-based contexts is measured besides the most generic measurement of response rates. Especially as technology-enabled interventions (e.g., those delivered via SMS, email, Zoom, etc.) and the collection of these data proliferate, future research would greatly benefit from greater consistency to understand engagement as a construct across contexts. In addition, the relationships between engagement behaviors and outcomes remain understudied - i.e., whether a particular behavior is associated with or leads to a desirable outcome. This gap is surprising given participants' behavioral engagement and response are central to the success of any intervention or treatment program. This paper seeks to start bridging this gap in knowledge, specifically by identifying any correlations between specific engagement behaviors and academic outcomes like college persistence.

III. Context and Data

IIIa. Intervention Context

Nudges to the Finish Line (N2FL) was a field experiment that investigated the use of text-based nudge strategies to increase degree completion among students who had accumulated substantial credits but were at risk of withdrawal before finishing their program of study. The goal of the intervention was to increase rates of college persistence and completion. The N2FL intervention spanned several academic years, including a pilot phase during the 2016-17 academic year, followed by full-scale implementation during the 2017-2018 and 2018-2019 academic years. The Nudge[4] Solutions Lab at the University of Virginia partnered with 20 broad-access public two- and four-year institutions across Virginia, New York, Texas, Ohio, and Washington. Experimental analyses reveal the intervention did not

improve student persistence or graduation rates, either in the full sample or within student subgroups (Bettinger et al., 2020).

The intervention targeted students who had **(a)** registered for classes during the first term of implementation by their institution's census date, **(b)** had a valid phone number on file, and **(c)** had previously completed at least 50% of the credits required to graduate from their program of study. All students meeting those criteria were randomly assigned to an experimental condition, and those assigned to treatment were automatically enrolled into the interactive texting campaign. Institutional partners provided student-level administrative data (e.g., cell phone numbers and first names) necessary for delivering personalized messages.

Treated students were enrolled in their campus's texting campaign for an average of 2-3 academic semesters (excluding summer terms). They received approximately one scheduled text message per week. The messages prompted students to complete important tasks (e.g., submit the FAFSA), encouraged them to use campus resources to advance toward their degree (e.g., academic tutors, financial aid officers), and addressed feelings of stress and anxiety (e.g., financial hardships, balancing family and work). The messages leveraged behavioral insights (i.e., planning prompts, descriptive social norms, loss aversion) to increase follow-through for intended actions, and some embedded infographics (that appeared as images on students' phones) to further reinforce the call to action and increase the salience of relevant information. The research team worked with each partner campus to develop and tailor scheduled message content to their institutional context, such as inserting the specific name of a tutoring center or adjusting the tone and language for an older student population, while maintaining general consistency in content and intention across sites.

An important feature of this intervention was the ability for students to write back to the scheduled campaign messages and ask questions. The majority of the scheduled

messages invited responses from students (e.g., "Registration for Spring semester starts 10/1. Want to work together to check which courses you still need to complete your degree?"), but students were able to write to their advisor at any point to initiate conversations as well.[25] This opportunity for student input and engagement is the focus of the present study. Each institution adopted a different advising model depending on the individual(s) they identified as responsible for monitoring and responding to inbound text messages from students. Campuses further adjusted the language and timing of scheduled messages to reflect the scope of those individuals' role (e.g., financial aid advisor vs. general staff assistant) and availability (e.g., to respond to students who texted in). More detailed information on these advising models can be found in Table A2.1.

We note the N2FL intervention was largely restricted to text interactions between an individual student and their assigned advisor. Students may have met with an advisor or used campus resources like a tutoring center as a result of those conversations, but the core intervention consisted of two-way interactions via text message between students and a designated advisor or staff, and we only have access to message logs (and not, for example, in-person advisor visit logs) for measuring engagement in this analysis.

IIIb. Study Sample

Our analysis focuses on student-advisor interactions during the Scale Phase of N2FL that took place during the 2018-19 and 2019-20 academic years at the City University of

[25] While students were reading and texting from their phones, designated campus staff or advisors were reading and texting from their computers through a web-based portal called Signal Vine. Signal Vine's interface is very similar to that of an email client like Gmail or Microsoft Outlook. A couple benefits of the Signal Vine system in this project included filtering which inbound messages from students were unread and required follow-up, and scheduling messages for future delivery (e.g., if the advisor or staff wanted to schedule a reminder message about an upcoming scholarship deadline that was still a couple weeks away).

New York (CUNY), Virginia Community College System (VCCS) and the Texas Higher Education Coordinating Board (THECB). Colleges in this phase typically implemented the intervention for 2-3 academic terms, not including summer terms (e.g., sending messages during spring 2018, fall 2018, and spring 2019 semesters for a three-term schedule) across multiple cohorts (e.g., at one college, the first cohort's messages spanned spring 2018 through spring 2019 semesters, while the second cohort's messages spanned fall 2018 through spring 2019).

For the purposes of this study on patterns of student engagement, we focus explicitly on the texts sent by students. The full message log transcripts include roughly 327,000 texts sent or received throughout the intervention, and about 34,000 (~10%) were sent by students.[26] We moreover focus on students who responded to at least one scheduled message and did not request to opt-out of texting at any point during the intervention. This is to focus on patterns of engagement among those students who *actually* engaged in the study.[27] This results in a total of 4,914 students in our sample and 33,177 student texts.

Table 2.1 reports the demographic and academic baseline characteristics of each respondent group: our analytic sample, students who never responded to any scheduled message, and students who requested to opt-out of texts. First, we note that about half of treated students either opted-out (8.5%) or never responded (42.9%), indicating that the actual take-up of any text-based advising was fairly low (48.6%). In general, we see that the

[26] If treated students texted their advisors after the full set of scheduled messages were sent, they received an automated response notifying them that the campaign had concluded. To keep the timeframe of interest in the present study consistent across cohorts (and to avoid including any texts sent during the COVID-19 pandemic), we exclude any texts sent to or from students more than 14 days after the last scheduled message in the intervention was delivered. This excludes 104 texts by students and 60 texts by advisors.

[27] Subsequent analyses will examine the extent to which design features of the intervention affected likelihood of engagement as well as patterns of engagement.

three samples are largely similar to one another with only minor differences. Students in the analytic sample were only slightly older than the other groups. About 40% of the analytic sample of responders were male relative to 45% of non-responders and 39% of opt-outs. Our study sample was also more likely to be Black (18%) than non-responders (14%) or opt-out students (15%). Analytic sample students seemed to be slightly higher in academic performance than other groups, with more credits earned at baseline, slightly higher cumulative GPA, and more terms enrolled. Study students were more likely to have had a stopout (unenrolled from the college without graduating) before the intervention than non-responders, but less likely to have had a stopout relative to the opt-outs. Sample students were otherwise about as likely to change their major or transfer prior to the intervention start date as the other groups. Finally, we see that VCCS and TX students were slightly more likely to be a non-responder or opt-out than included in the sample, whereas CUNY students were substantially more likely to be included in the sample.

IV. Methodology

IVa. Analytic Framework

The broader N2FL evaluation examined the effect of enrolling students in the text-based advising, and we can think of the overall treatment experience as the bundle of scheduled messages (e.g., helpful reminders about upcoming deadlines) and the availability of text-based advising (i.e., personalized support that students could access via text message). This latter aspect of two-way interaction was one of the most distinctive features of the N2FL campaign, setting it apart from one-way informational texting campaigns.

We might expect the availability of this text-based advising to impact student outcomes through two separate, but related, mechanisms. The first mechanism is what we

herein refer to as "active engagement," where students benefit from the personalized support and information they receive specifically by way of actively engaging with the text-based advising. We then consider variation in students' active engagement a meaningful source of *treatment intensity* in the broader intervention. That is, students who engaged heavily with advisors via texting would have received greater levels of this text-based advising "treatment" in the intervention than students who did not, and variance in these greater levels of engagement may then relate to variance in student outcomes.

Text-based advising may also affect student outcomes through a second mechanism we herein refer to as "passive engagement," in which students benefit specifically from the mere knowledge that they are being supported in the abstract. In other words, just knowing that an advisor is at their fingertips may confer some psychological benefits to students, such as a heightened sense of belonging in the college community that helps them perform better in their classes and engage in adaptive behaviors (Gopalan & Brady, 2019).

In contrast with the N2FL evaluation that examines the causal effect of the combined bundle of scheduled messaging, passive engagement, and active engagement, our analysis focuses on the characteristics, predictors, and correlates specifically of *active* engagement in the N2FL intervention to the extent possible.

Directly assessing the causal impact of active engagement on student outcomes in this context is complicated for two reasons. First, active engagement is not randomly assigned across students. Variance in active engagement rates is primarily driven by students' decisions to engage with the advising platform (and, to a lesser extent, the responsiveness of individual advisors), and active engagement rates may then be endogenous to observable and unobservable student characteristics. We thus cannot estimate a causal relationship between engagement and student outcomes using these observational data without our results being

contaminated by omitted variables bias from unobservable student characteristics (e.g., students who are more conscientious are both more likely to respond to any texts they receive *and* perform well in their classes). Moreover, because we cannot know the engagement levels of students in the control group *had they been sent texts*, we cannot leverage the original intervention's RCT design to estimate the impact of varying levels of engagement, either.

Second is the measurement issue. Engagement (as we described earlier in Section II) is a multi-dimensional and complex construct, and it is not clear which measurable proxies best stand-in for active engagement of text-based advising itself. For example, the N2FL texting data did not capture whether a scheduled message was actually read by students, and so we cannot know how many messages a student read. While we can look at the number of responses a student sent, a single-word text response from a student (e.g., "Yeah") would still be conceptually different from a more involved, inquisitive response (e.g., "Yeah, and I was hoping you could tell me more about…") when thinking about engagement and treatment intensity.

These considerations set the stage for our study in two parts. We begin by exploring a variety of possible engagement measures to document variation across students and any correlation with one another, as guided by the literature when possible. Our driving motivation is to generate novel insight into student engagement patterns that are broadly applicable across intervention contexts; to this end, we choose to focus on measures that are observable and conceptually relevant to the construct of engagement (i.e., behavioral dimensions like length of engagement periods), that can be easily communicated to and applied by other researchers, that document meaningful levels of variation across students, and that are generally uncorrelated with other measures. We interpret those measures with

large levels of variation as those more likely to be malleable, though this dynamic will need to be explored in more detail in future work. We then examine whether any of these engagement measures are also correlated (descriptively) with the outcomes of interest. Ultimately, we are attempting to identify promising proxies for student engagement in these text-based advising campaigns that are associated with better student outcomes, thus generating testable hypotheses about which engagement measures are most malleable and impactful. These hypotheses would then open pathways for future work to explore the manipulation of such engagement measures in experimental contexts as we seek to improve the efficacy of text-based advising programs.

IVb. Defining Engagement

We break our student-level engagement measures into four main categories as guided by the literature and the data we have at our disposal. Rather than attempting to derive a single measure of engagement, we attempt instead to create an array of measures that captures the various nuances of engagement across several dimensions.

In the first category, we are interested in examining the **Frequency and Intensity of Student Replies**. This category most closely mirrors how response rates have been measured in related literature. Because some sites for the N2FL intervention chose to send their scheduled messages at differing intervals and frequencies (e.g., due to different dates for financial aid filing, course registration, etc.), we examine the *percent of scheduled messages that a student responded to at least once*. This can be thought of as the relative frequency with which a student had *any* apparent engagement in the scheduled messages they received, meaning it does not consider circumstances when students have sustained engagement after a given scheduled message. To address that shortcoming of the measure, we also examine a

student's *average replies per scheduled message* (inclusive of any messages students send in response to their advisor afterward), to assess how frequently a student engaged in a more sustained way with the scheduled messages. To examine engagement intensity, we also look at a student's *average reply length in words* under the assumption that longer messages are indicative of a more intense level of engagement with their advisor.[28] Similarly, we examine a student's *proportion of substantive replies*, or the share of replies at least 5 words in length.[29] This is to distinguish quick, gestural answers (e.g., "Yes", or "No") from more deliberate responses.

We also are interested in examining **Engagement Duration** as a separate category. For example, consider two students who responded to exactly four scheduled messages: Student A responded to the first four scheduled messages of the campaign and then disengaged completely, and Student B responded to four scheduled messages of the campaign throughout a period of three academic semesters. We thus calculate the *percent of scheduled messages sent before a student engaged* (i.e., sent their first reply) and the *percent of scheduled messages sent before a student disengaged* (i.e., sent their last reply). We then also calculate the share of scheduled messages sent within this window, which we interpret as the *percent of scheduled messages a student was engaged*.

Another category of interest for student engagement behaviors that is measurable across most texting intervention contexts is their **Response Speed**. Perhaps students who are quick responders paid closer attention to the intervention than those who spent several hours or days before replying. Conversely, it is possible that students who responded later

---

[28] We also examine reply length in *characters* as an alternative specification and find no substantive difference to our results. We focus on words in this manuscript for its interpretability and concision.

[29] This threshold for 5 words is based largely on our informal examination of short texts in the data, where we find that responses below 5 words tend to be gestural or confirmatory in nature and without other substance.

spent more time thinking about and internalizing the scheduled messages before responding. To measure this dynamic empirically, we first calculate a student's *average response time to scheduled messages* in hours (conditional on responding). We also calculate a student's *average response time to any advisor-generated messages* in hours (conditional on responding) to also capture a student's reply speed to non-scheduled advisor-generated messages (e.g., messages their advisor directly wrote to them as part of the conversation).

Finally, we are not just interested in the patterns of when and how often a student responds, but also *what they discuss* when they do. To that end, we employ a range of text analysis techniques to better understand and measure **Response Content**. The first technique is relatively straightforward where we scan each student's text for the prevalence of a given language category. For example, this approach would scan messages for "help-asking" language – messages with question marks, as well as messages including phrases like "how do I…," how do you…," "can you…," "I need…," and so on. We can then calculate the *proportion of student messages asking for help* as a specific type of engagement relevant to the text-based advising context of the N2FL intervention.[30]

We go on to create two more complex text content measures using NLP techniques known as sentiment analysis and structural topic modeling. For concision, we provide only a brief and intuitive explanation of these two methods in the next sections of the main paper; greater detail on the techniques themselves, their implicit assumptions, our text cleaning decisions, the robustness checks we ran, and the validity exercises we deployed, can be found in Appendices A2.1-A2.5.

---

[30] This approach also enables us to identify students who opted out from the intervention, searching for phrases such as "stop messaging" or singular responses consisting only of "sotp."

IVc. Analyzing Message Content with Sentiment Analysis

Sentiment Analysis is a common NLP task in which analysts use an algorithm to "read" a given text string and rate the extent to which the string contains/expresses a positive, negative, or neutral/factual sentiment (Pang & Lee, 2008). One can think of this process as generating output similar to hand-coded qualitative analysis, but in an automated and highly scalable way that facilitates quantitative analysis. Though matching human judgment *perfectly* remains out of reach for the current state-of-the-art, modern algorithms are nonetheless exceptionally nuanced, flexible, and accurate[31] at this task (Vaswani et al., 2017). Using vast volumes of text data to first "learn" a general understanding of language syntax and word relationships, modern sentiment analysis algorithms are now able to account for the complexities of word context (e.g. that "I wish I were happy" actually indicates sadness), multiple word meanings (e.g. that "bank" has two separate meanings in "The river bank was wet" and "I went to the bank this morning"), and informalities (e.g. "that was sick, dude;" Ambartsoumian & Popowich, 2018) far better than early sentiment analysis approaches (e.g., dictionary-based methods).

We operationalize the definition of sentiment for the present study as the *perceived positivity of emotions and ideas present in a given text*. This definition then is a conglomeration of the speaker's stated emotions ("I feel sad" v. "I am excited"), communicated intention ("I hope you die" v. "I wish you the best!"), and, at least to some extent, topical content

[31] Importantly, it is often the case that a given string of text has no one "right" answer for its sentiment, and so expecting an algorithm to perfectly match human judgment may be an impossible bar to set to begin with. In our own validation exercises, for example, our team of five human coders often disagreed about a given text's sentiment due to individual interpretations of subtext and implications. We ultimately find that the inter-rater reliability of a team of human coders is not significantly different from the inter-rater reliability of a team of human coders *plus the algorithm*, indicating that it does not seem to disagree with a human's judgment about a given text any more than humans disagree with one another.

("financial hardship" v. "vacation time"). Note that this definition is complicated for longer strings of text, in which multiple emotions/implications may be present and the overall sentiment becomes more ambiguous.[32] We argue this definition remains appropriate for our context because of the nature of the N2FL text data: texts were generally only one or two sentences long (making the ambiguity of sentiment in long text strings less problematic), students are sending these texts "as-is" (i.e., they are not transcribed spoken words with greater context than what we can observe in the text data), and we need not perfectly describe the student's *intended* sentiment for this to nonetheless be a useful typology for classifying distinct modes of engagement. For demonstration purposes, we provide a list of real N2FL texts from students alongside their algorithmically-generated sentiment score in Table 2.2.

We ultimately measure the *average positivity of emotions and ideas present in a student's replies* (average "sentiment") on a scale from very negative (-2) to very positive (2), where sentiment with a score of 0 can be thought of as more factual in nature (e.g., "I enrolled in my courses" rather than "I was so relieved to enroll in my courses"). In other words, what is the general tone across a student's responses?

IVd. Analyzing Message Content with Topic Modeling

Topic Modeling is another common task in NLP in which analysts use an algorithm to "learn" what discrete topics of discussion exist across a series of text documents (in our

[32] While we would like to lean on a more standardized definition, we were unable to find a detailed and widely-accepted definition for sentiment in transformer-based models in the literature. Interestingly, sentiment as a construct across modern data science (i.e. neural network-based models rather than dictionary models) is almost entirely dependent on the SST's definition due to the strong incentive for data scientists to optimize their algorithm's SST performance for benchmarking purposes. That said, the SST intentionally encouraged their human coders to view sentiment as a flexible and subjective notion, making a formal definition elusive.

case, texting conversations) and then measure how prevalent each topic is within each of the provided documents. In brief, the algorithm does this by examining how words are used in conjunction with one another across documents, under the assumption that words about the same general topic of conversation will often appear together in the same documents (Blei, 2003). For example, "financial," "aid," "deadline," and "FAFSA" might often appear in the same documents, thus indicating to the algorithm that they are used to discuss the same topic of conversation. By then constructing several sets of words that often appear together in this way, the algorithm will have identified the word groups that it thinks represent each distinct topic of conversation within the text data; analysts then interpret these word groupings for meaning (such as, "FAFSA filing") and, in our case, make "supertopics" that combine multiple word groups together under a single broader category of conversation (such as, "financial aid" that encompasses topics about FAFSA filing, tuition payments, etc.).

Ultimately, we are interested in whether there exists variation in the prevalence of these supertopics across students' replies. Such variation would reflect substantively different engagement behaviors, and thus different patterns of *how* students navigate their responses to the advising intervention as a result. Once the supertopics are identified, the algorithm can determine the prevalence of each supertopic across a given student's texts based on the combination of keywords they used. For example, the algorithm can tell us how many keywords in a given student's responses are spent discussing the "financial aid" as a supertopic, versus "course planning." We thus construct measures to describe the *percent of student replies* about each of the following topics: course planning (e.g., course registration, registration deadlines, etc.), financial aid (e.g., applying for financial aid, paying tuition, etc.), academic planning (e.g., graduation deadlines, transfer requirements, career planning, etc.), general academic support (e.g., tutoring services, study skills, etc.), and meeting logistics (e.g.,

scheduling an in-person advising meeting, getting the right contact email, etc.). In other words, what are the more prevalent topics of conversation in a student's responses? A partial list of the most influential keywords that fall under each supertopic is displayed in Table 2.3. A complete list can be found in Table A2.6.

IVe. Regression Analysis

Besides examining how students vary in terms of their behavior across the engagement measures, we are especially interested in the extent to which these engagement measures are related to the outcomes of interest for the intervention: re-enrollment term-to-term (binary), credits earned (number of credits), term GPA (raw GPA units), and degree receipt (binary). In other words, do we see any relationship between how students engage and their ensuing academic performance? Because this is a descriptive analysis, we again cannot take any of these relationships as causal; instead, we think of this as an exploratory analysis meant to generate testable hypotheses for future research (e.g., experiments) around what engagement measures might be both malleable and also impactful on outcomes. For example, if it is the case that students who ask for help more frequently via text tend to perform significantly better in terms of desired student outcomes, researchers might explicitly explore this relationship further in future work by designing a text-based advising intervention with prompts greater or fewer student questions across treatment wings to see if such intervention features enhance the effectiveness of said intervention.

We examine the relationship between each engagement measure and each outcome of interest (in the first, second, and third term immediately following the start of the N2FL intervention) in the context of a regression analysis where we also control for salient student demographics and baseline academic characteristics. For controls, we include all of the

variables explored in Table 2.1, as well as the randomization block of students during the initial study randomization process.[33] More formally, we iteratively estimate the following equation:

$$(1) \quad Y_i = \lambda_t + X_i + A_i + \beta_1 Engagement_i + \varepsilon_i$$

where Y_i represents any one of the student outcomes of interest, λ_t represents the vector of fixed effects for student randomization blocks, X_i represents the vector of student demographic characteristics, A_i represents the vector of student baseline academic characteristics, $Engagement_i$ is any one of the engagement measures, and ε_i represents the idiosyncratic error term. The coefficient of interest will thus be beta 1, revealing the controlled relationship between each engagement measure and each outcome. Finally, we cluster our standard errors at that randomization block level.

V. Results

Va. Variation in Engagement Measures Across Students

We first look at the distribution of each measure at the student-level to understand how they vary across students and potentially uncover salient patterns of engagement in the intervention. In all following plots, the X-axis charts out the range of the values for a given engagement measure, the Y-axis shows the density of students at each value along the X-axis, the dotted line shows the mean value of the engagement measure at the student-level (also reported in the subtitle), the solid line shows the median value of the engagement measure at the student-level (also reported in the subtitle), and the number of students

[33] Note that the randomization blocks were separated by school system, meaning school system is completely collinear with randomization block across all students and is thus unnecessary to include separately in this regression.

displayed in each plot (i.e., number of non-missing values) is indicated in the subtitle of each plot.

Figure 1.1 shows the distribution for each of the **Frequency and Intensity of Student Replies** measures. In the top panel, we see the vast majority of students who responded to *any* message still responded to fewer than 25% of the scheduled messages they receive, with a long tail extending beyond that. In the next panel, we see that most students only respond with a single reply after a given scheduled message on average, indicating that they very rarely engage in actual back-and-forth texting with their advisors when they do respond. The long tail here also indicates that a small handful of students regularly engaged in lengthier conversations. These metrics point to the reality that, even in well-designed interventions seeking to elicit two-way student engagement, genuine student engagement may be less common than we might otherwise expect. In the following panel, we note that the student-level median for average reply length is 10 words – the equivalent of a short sentence, which makes sense given the medium of texting. Interestingly, the highest density area of students always responded with a substantive reply (>=5 words), as shown in the bottom-right panel, while a far smaller proportion never did. That the student-level median for the proportion of substantive replies is 0.72 also indicates that insubstantial replies (e.g., "yeah," "okay," "no," "thanks", etc.) were less common than we might have initially anticipated a priori for a texting intervention.

Figure 2.2a shows the distribution for each of the **Engagement Duration** measures. In the top plot we show the percent of messages sent prior to students' first response - the mass near zero indicates that most students engaged for the first time very early on in their scheduled messages, with a student-level median of 0.17. That said, a long tail here also indicates that a meaningful share of students were nonetheless engaging for the first time all

along the sequence of scheduled messages. The next plot reveals also that the median student disengaged (i.e., sent their last reply) about two-thirds of the way through the intervention, though many didn't disengage until the very end given the mass around 1. Finally, in the bottom plot we show the distribution of the percent of messages with which students engaged. The mass of points near zero in the bottom plot reveals that, despite the distributions showing many early engagements and many late disengagements, relatively few students were engaged for the majority of the intervention. The mass near zero indicates that a large share of students were only ever engaged for a brief period of the intervention.

Figure 2.2b is another way of visualizing the same engagement duration data to better differentiate individual students' behaviors and explain these seemingly contradictory results. Along the X-axis is the percent of scheduled messages before students sent their first message, while along the Y-axis is the percent of scheduled messages before students sent their last message. We can then, for an individual student represented as a single point, see when they engaged relative to when they disengaged. As an example, students in the top-left of the plot engaged immediately at the start of the intervention (the percent of scheduled messages that passed before they engaged was nearly zero) and disengaged at the very end of the intervention (the percent of scheduled messages that passed before they disengaged was nearly one), indicating that they were engaged throughout the entirety of the intervention (the percent of scheduled messages they were engaged for was 1). Students along the 45-degree line are students who engaged and disengaged at the same time, and thus must have only ever sent one reply.

Overall, we see that many students could only nominally be considered actively engaged at all under this definition given the mass of points along the 45-degree line. That said, a fair cluster of students in the top-left and along the left side of the plot had

immediately engaged and stayed engaged for a large proportion of the intervention, reflecting the long tail of students in the prior plot for the percent of scheduled messages they were engaged (bottom plot of Figure 2.2a). Lastly, students were most likely to engage early on in the intervention if at all, given the decreasing number of points as we move along the X-axis, reflecting the large mass of points near zero in the distribution of percent of scheduled messages before students engaged (top plot of Figure 2.2a).

Figure 2.3 shows the distribution for each of the **Response Speed** measures.[34] The first plot reveals that the median student responded within 1.41 hours of scheduled messages, though there exists an exceptionally long tail where students replied days, or even weeks, after most scheduled messages; we might interpret this to mean that, even if students did not reply promptly to scheduled messages, they were aware of the availability of text-based advising and turned to this mode of communication days and weeks further out. We do not see that this distribution pattern changes meaningfully when more broadly considering students' responses to *any* advising texts (e.g., ad hoc messages that advisors wrote in response to students' initial replies), though the average response time is slightly reduced from 15.13 hours to 11.74 hours.

Finally, Figures 4.4a and 4.4b show the distribution for each of the **Response Content** measures, beginning with the help-asking and sentiment measures in Figure 4.4a. The first panel shows that there was relatively wide variation in the proportion of messages each student sent asking for help, though a large proportion of students never asked for help given the mass near zero. Paired with the fact that most student replies were about a sentence in length (third panel of Figure 2.1), students seemed to more often be answering

[34] Note that while response times are reported with hours as the unit, all calculations are accurate to the second (i.e., measurements were not rounded to whole hours).

questions in the scheduled messages with declarative statements than with explicit requests for additional help. We also see this dynamic reflected in the sentiment analysis results, with most students showing an average sentiment of replies at 0, suggesting the prevalence of factual statements was greater than emotionally-charged student messages (e.g., those remarking on difficulties, frustration, excitement, etc.). That said, we still see meaningful variation around 0 in both directions, so students were still sending texts with more positive and more negative sentiment.

Turning now to the topical content measures in Figure 4.4b, we see generally wide distributions of topical content prevalence for every supertopic *except* financial aid in the top-right plot. That is, the median student spent between 13% and 20% of their replies focused on each topic of course planning, academic planning, academic support, and meeting logistics, but only 2% of their replies focused on financial aid. The wide distributions for all but financial aid indicate that students seemed to generally vary quite a bit in terms of how much they discussed each topic. This variation perhaps reflects one of the strengths of an advising intervention in that it is responsive to individual students' needs and interests. That relatively few of the replies focused on financial aid is puzzling but could be the result of a few likely dynamics: **(a)** the deadline-dependent nature of FAFSA filing and tuition payments means they might be relevant only during specific timepoints of the year, whereas the other topics could more naturally come up throughout the entirety of the intervention; **(b)** the scheduled messages on the topic instigated responses that were more confirmatory in nature (e.g., "Have you filed your FAFSA yet?") and thus didn't require students to respond using financial aid phrases; **(c)** students may have been less comfortable raising financial aid questions or issues via text message; and/or **(d)** students targeted by the intervention are near-completion, and so may already be familiar and comfortable with financial aid filing

processes by this point in their college trajectories. In any case, the relatively low level of student replies about financial aid is surprising given the substantial proportion of lower-income demographics at the broad-access institutions represented in the sample.

To summarize at a high level the aforementioned results, we generally see the least variation across responsive students in terms of their response times and their average responses per scheduled message. We see the greatest variation in terms of the topical content of their replies, and still meaningful variation in terms of the proportion of their messages asking for help, the sentiment of their replies, when they engaged and disengaged, and the length of their replies in words. While each measure helps reveal useful insights about patterns of student behavior (e.g., that most students respond almost immediately after they receive scheduled messages, if at all), these measures with greater levels of variation are most likely to help us distinguish students' engagement patterns from one another (e.g., versus measures with low variation such as response speed).

Vb. Correlations Between Engagement Measures

While we can learn much about how students engaged in the intervention by examining each of the measures individually, we are also interested in the extent to which these measures correlate with one another. Measures that have high levels of absolute correlation with one another can reveal "bundles" of common engagement patterns in the context of text-based advising interventions, while measures that have low levels of absolute correlation might best be interpreted as measures that capture distinct information from one another. The former might be especially useful to gain deeper insight into how students navigate text-based advising interventions in general (e.g., to inform future program design), while the latter might be especially useful as we seek to create a parsimonious set of

engagement measures we could commonly track and/or encourage across text-based advising interventions of this kind.

Table 2.4 is a correlation matrix that shows the correlation coefficient between each of our engagement measures against one another. Cells are shaded according to their correlation, with red shading indicating stronger negative correlations and blue shading indicating stronger positive correlations. Any coefficients presented in bold are statistically significant at the $p<0.05$ level. Borders are drawn around each of the four groups of measures (Frequency/Intensity, Duration, Response Time, Response Content) for visual clarity.

We call attention to a few surprising and noteworthy dynamics for concision. To begin with one example, we see that the percent of scheduled messages students responded to is at most weakly correlated with every other measure except for the percent of messages before engagement (-0.34) and disengagement (0.45) and the percent of scheduled messages engaged (0.61). Each of these strong relationships make mechanical sense, in that a student must have been engaged for a higher percentage of the messaging duration if they responded to a greater proportion of scheduled messages in general, and thus were more likely to have engaged earlier or disengaged later in the intervention. The general lack of correlation otherwise also indicates that response rates do not tell us much about response *content* at all, re-emphasizing the usefulness of examining engagement beyond just response rates.

Reply length (specified either as average reply length or as the proportion of substantive replies) is positively related to the proportion of messages asking for help (0.36 or 0.34) and negatively related to the average sentiment of messages (-0.21 or -0.24). This makes some intuitive sense, in that students who are asking for assistance with something would likely provide more detailed messages versus a student who has no need for

assistance. That sentiment itself is also negatively related to help-asking (-0.34) also makes intuitive sense given that bids for help are often predicated on students sharing their issues (e.g., course registration portals not working properly) or hardship (e.g., inability to pay for tuition).

Note here that the response time measures seem only weakly correlated with other measures across the board, with no correlation coefficient higher than an absolute value of 0.11 (except for the coefficient between the two response speed measures, which is to be expected mechanically). This indicates that response speed seems to capture a completely different dimension of student engagement behavior than the other measures, though we will examine whether this group of measures seem to provide any worthwhile information with respect to student outcomes in the next section.

Interestingly, the topical content measures only seem related to one another, and this is largely the result of a mechanical relationship whereby the topical content measures must sum to 1 as proportions of the student replies. This again emphasizes the intuition that *what* students engage about is a critical piece of the puzzle in understanding *how* students engage with text-based advising, distinct from response rates. That said, we observe negligible relationships between topical content measures only in the case of the prevalence of financial aid against course planning (-0.02) and academic support (-0.06). This might be due to there being a higher likelihood of students discussing *both* finances and course-taking or finances and academic performance (e.g., maintaining GPAs for scholarships, or needing more academic support if finances are an issue), resulting in less negative correlations than those we observe between other topics. We also observe particularly negative correlations between meeting logistics and both academic planning and course planning. This may reflect the idea that these topics are often sufficiently complex that students and advisors would prefer to

discuss them in-person rather than over text, and so we wouldn't observe the ensuing conversations about academic or course planning in our measures.

In general, we do not see particularly strong correlations in measures *across* categories, and strong correlations *within* categories tend to be mechanical in nature (e.g., topical content proportions). This again indicates that these measures seem to be capturing quite different information from one another, suggesting the potential value of thinking about student engagement along multiple dimensions when possible, beyond simple response rates.

Vc. Relationships Between Engagement Measures and Student Outcomes

While the aforementioned analyses provide excellent insight into how students navigated the texting intervention, we now move to examine the extent to which these engagement measures are actually related to the outcomes of interest for the intervention. Again, we view these descriptive analyses as purely exploratory for the sake of generating hypotheses about what kinds of engagement may relate to better student outcomes, and thus what kinds of engagement future intervention designers may wish to elicit in their construction of similar text-based advising programs. Table 2.5 thus displays the results of many regressions (16 engagement measures by 12 outcomes), where each cell represents a separate regression as specified in Section IVe. Bolded cells indicate relationships significant at the $p<0.05$ level.

Among the **Frequency and Intensity of Student Replies** measures, the percent of scheduled messages students responded to seemed to be the only measure with meaningful relationships to student outcomes: students who responded to a greater share of scheduled messages experienced substantially higher term credits, higher term GPAs, and higher

likelihood of degree receipt in every term, even after controlling for academic baseline covariates. Interestingly, a higher proportion of substantive replies seems negatively correlated to re-enrollment and credit receipt in the later terms. While this could be a result of the dynamic we hypothesized earlier that more substantial messages were reflective of greater individual struggles, we surprisingly do not see the same relationship for help-asking messages.

Looking at the **Engagement Duration** measures, we observe many strong relationships all in the direction of more positive outcomes for students who were engaged for a greater share of the intervention. That is, students who engaged earlier, disengaged later, and were engaged for a larger proportion of the intervention had substantially higher re-enrollment rates, credit accumulation, and GPA. These same students had higher levels of degree receipt, but only at the later time intervals of T+2 and T+3 with increasing magnitude further from the intervention start term.

We observe no significant relationships across the board for the **Response Speed** measures, as well as the proportion of messages asking for help. Turning to the remainder of the **Response Content** measures, higher sentiment levels (i.e., positive sentiment) correlate with slightly higher levels of credits, GPA, and degree receipt, with increasing magnitude in more distal term periods. It should be noted that these magnitudes are smaller than the other measures at least in part because it is one of the only non-proportion measures we constructed with a range of -2 to 2. Thus, a unit change in the average sentiment is more feasible in reality than a unit change in a proportion variable like the percent of scheduled messages responded to (i.e., going from 0% of scheduled messages responded to, to 100%).

For the topical content measures, increased levels of course planning were positively correlated with re-enrollment and credits earned, again with increasing magnitude over time.

It was also negatively correlated with degree receipt in the earlier terms, which makes some intuitive sense given that students focused on course enrollment for the coming term are likely not ready to graduate for the given term. Discussion of academic support and meeting logistics reflected some of these same trends with student outcomes, likely for the same reasons: students looking for either of these types of support from their advisors were unlikely to be immediately ready to graduate, but seemed to benefit in terms of re-enrollment and successful completion of credits in the given term. Greater shares of discussion about financial aid was negatively correlated with both GPA and degree receipt, perhaps reflecting that academic difficulty has a strong relationship in general with student finances. The share of discussion about academic planning has, perhaps intuitively, almost the opposite relationship with outcomes to course planning. That is, higher rates of academic planning discussion was associated with far higher levels of degree receipt and far lower levels of re-enrollment and credit accumulation, likely because academic planning discussion includes topics like job applications, graduation logistics, and so on.

VI. Discussion

Taken altogether, our results suggest several high-level insights about student engagement patterns in the context of text-based advising interventions. First, we see that even interventions designed to elicit strong engagement from students don't necessarily succeed in doing so across the board. Moreover, we see that response rates alone are likely insufficient to characterize the nuance and multi-dimensionality inherent in how students engage in these personalizable interventions. We see broad variation across students in many of the four categories and sixteen measures we constructed, and these measures generally seem to capture quite different information about students' engagement from one another

given low between-measure correlations. In general, we view these widely varying engagement measures as more likely to be malleable through thoughtful program design, but this will need to be explored in greater detail in future study. There is also likely a relationship between the simplicity of the engagement measure and malleability - for example, it may be relatively easy to change the average length of student response by shifting from sending students closed-response prompts (e.g., "Are you planning to submit the FAFSA?") to sending students open-response prompts (e.g., "How can I help you submit the FAFSA?"). In contrast, more complex measures such as the sentiment that students convey in their messages may prove harder (and potentially undesirable) to manipulate.

Second, we see that many of these engagement measures have statistically and substantively significant relationships with academic outcomes of interest like student persistence and degree receipt, even after holding constant students' demographics and baseline academic characteristics. Most notably, among students who texted into the campaign at least once, responding to a greater share of scheduled messages was positively correlated with better academic performance and degree completion. Similarly, *longer* periods of engagement were moreover associated with higher persistence and academic performance. Although not causal, these findings together suggest that we stand to learn much more about improving the efficacy of two-way text advising campaigns, for example by experimentally exploring how more sustained engagements with students could enhance outcomes.

In contrast, we found that longer student messages to advisors are associated with higher rates of help-seeking language and negative sentiment, but lower rates of persistence and credit completion. While it makes intuitive sense that students who seek help are typing

out longer messages and expressing negative affect (e.g., frustration), that we observe worse academic performance among these students merits further exploration. For instance, it is possible that at-risk students who engage in help-seeking behaviors via text are not receiving the support they seek or need. A deeper understanding of the relationships observed here--between help-seeking language, negative affect, subsequent interactions with campus supports, and worse academic performance--could help shed light on ways to design advising programs in a way that delivers enhanced support for students who self-identify as needing help.

Additionally, we found that a greater share of student replies about meeting logistics is positively correlated with re-enrollment and credit accumulation during the first two semesters of the texting intervention. This supports the notion that one of the ways in which text-based outreach could help students is to make scheduling advising appointments easier. Particularly as the challenges that students face become increasingly complex (e.g., financial aid issues, uncertainty about plans for transferring to a four-year university), it stands to reason that text-based engagement will be useful insofar as it allows students to plan when they will meet with an advisor for more in-depth assistance.

Overall, this study calls for a more careful look under the surface of text-based advising programs such as the N2FL intervention. Our exploratory findings confirm our hypothesis that there is meaningful variation across students in terms of how they respond topically and length-wise and for how long they choose to engage in interactive texting campaigns spanning multiple academic semesters. A deeper understanding of the heterogeneity in student engagement behaviors and the identification of specific behaviors that are correlated with academic success could help scholars and practitioners alike design text-based advising programs with greater intentionality, precision, and efficacy.

Usefully, all the measures we construct are imminently scalable and applicable to any similar texting context, meaning that they can serve as more consistent tools to help us better understand and contextualize the results of interventions that have previously taken place, as well as diagnostics to inform program management and implementation *as an intervention is happening*. To push for greater codification of such interaction measures across intervention contexts, we also offer our code open-source for other researchers to build upon and implement in their own studies. This ability to perform real-time diagnostics is especially appealing in that local institutions can glean important insights about their specific student population (that might not be applicable in other contexts) and test approaches to adjust their messaging strategy accordingly. With the combination of these more nuanced standard engagement measures and sophisticated NLP techniques, we can offer researchers and practitioners greater visibility into important dynamics like student uptake going forward.

Table 2.1. Demographic and Academic Baseline Characteristics by Respondent Group

Variable	Study Sample	Non-Responders	Opt-Outs
Sample			
N	4914	4330	857
Age at Entry	21.84	20.23	21.2
Male	0.4	0.45	0.39
Race/Ethnicity			
White	0.32	0.38	0.4
Black	0.18	0.14	0.15
Hispanic	0.24	0.24	0.24
Other Race	0.12	0.13	0.09
Missing Race	0.14	0.12	0.12
Academics at Baseline			
Credits Earned	55.99	51.49	50.48
Cumulative GPA	2.95	2.88	2.9
Terms Enrolled	4.4	4.19	4.3
Any Prior Stopout	0.31	0.28	0.34
Any Prior Change of Major	0.21	0.2	0.23
Any Prior Transfer	0.27	0.23	0.26
System			
VCCS	0.21	0.25	0.27
CUNY	0.49	0.39	0.39
TX	0.3	0.35	0.34

Table 2.2. Sample N2FL Messages and Assigned Sentiment Scores

Text Message (sic)	Sentiment Score
Terrible I have to find a class or two to sign up for. I'm so behind.	Very Negative (-2)
No. Everything has been piling up at school and it's kind of been too stressful to decide what to get done.	Very Negative (-2)
I was trying to drop a class and it doesnt allow me	Negative (-1)
Hi! I applied for graduation and got an update but I do not understand it because it doesn't match the update list on the [institution] graduation page	Negative (-1)
I am in school Tuesdays and Thursdays	Neutral (0)
When does summer classes start?	Neutral (0)
Just tried again and it let me register haha, thank you for your help	Positive (1)
It's fine. Thank you very much for the link. If I have any other questions in the future, can I text this number?	Positive (1)
Ok thanks so much! I finished this semester strong! I got a 100.5% on my Anatomy Final Exam! That grade replaced my lowest test grade of an 85%	Very Positive (2)
It was very helpful thank you so much!	Very Positive (2)

Note: Texts shown here were specifically selected from the set of N2FL texts where the human coders and the algorithm output were in agreement to clearly illustrate what differing levels of sentiment can look like. These examples should not be interpreted as a general demonstration of algorithm accuracy.

Table 2.3. Supertopic Groupings and Sample Subtopics and Words

Academic Planning	math, science, requirement, biology, art, spanish, registrar, language, college
	credit, graduate, course, major, requirement, internship, psychology, minor
	graduate, congratulations, graduating, applied, feel, free, ready, december
	degree, transfer, major, change, associates, plan, audit, transcript, bachelors
Academic Support	hope, information, tokenurl, hey, center, tutoring, located, helpful, office, visit
	question, hey, info, yeah, answer, reaching, nice, assist, specific, study
	im, semester, grade, luck, checking, enrolled, final, exam, planning, lol
	campus, service, counselor, job, support, mind, ahead, provide, care, set
Meeting Logistics	appointment, time, tomorrow, wednesday, thursday, tuesday, monday, meet
	message, office, time, answer, frame, time frame, message time, answer message
	tokenphonenumber, call, phone, person, walk, call tokenphonenumber, monday
	advisor, contact, academic, tokenname, meet, advising, academic advisor, track
	appointment, schedule, schedule appointment, set, advising, advisor, tokenurl
Course Planning	spring, registration, date, winter, spring semester, november, enrollment, session, register
	student, id, drop, time, gpa, student email, check, access, withdraw, student id
	summer, fall, course, taking, summer class, online, fall semester, summer course
	professor, department, told, writing, speak, permission, request, alright, issue
Financial Aid	tokensis, hold, account, plan, payment, pay, bursar, log, check, tokenurl
	financial, aid, financial aid, fafsa, office, aid office, scholarship, loan, tuition, pay

Note: We display the top ~10 words within each sub-topic in terms of its *probability* metric (how much an appearance of that word contributes to the detection of that topic). We display only the first four subtopics under each supertopic for concision – a full list of the subtopics can be found in Table A2.6. "tokenname, "tokenphonenumber," "tokensystemname," and so on, were placeholders used for scrubbed PII words.

Table 2.4. Correlations Between Engagement Measures

	% of Scheduled Messages Responded to	Average Replies Per Scheduled Message	Average Reply Length in Words	Proportion of Substantive Replies	% of Scheduled Messages Before Engagement	% of Scheduled Messages Before Disengagement	% of Scheduled Messages Engaged	Response Time to Scheduled Messages (Hours)	Response Time to Any Advising Messages (Hours)	Proportion of Messages Asking for Help	Average Sentiment of Replies	% of Replies About Course Planning	% of Replies About Financial Aid	% of Replies About Academic Planning	% of Replies About Academic Supports	% of Replies About Meeting Logistics
% of Scheduled Messages Responded to	1.00	**0.12**	**-0.03**	-0.02	**-0.34**	**0.46**	**0.66**	0.03	0.00	**-0.07**	**0.06**	**0.09**	**0.03**	-0.01	0.00	**-0.07**
Average Replies Per Scheduled Message	**0.12**	1.00	0.00	0.01	**-0.11**	**0.07**	**0.14**	0.02	**-0.09**	0.02	-0.02	**0.08**	0.01	**0.08**	**-0.16**	0.01
Average Reply Length in Words	**-0.03**	0.00	1.00	**0.53**	**-0.06**	**-0.11**	**-0.05**	**0.04**	0.02	**0.36**	**-0.21**	**0.06**	0.02	**0.11**	**-0.07**	**-0.08**
Proportion of Substantive Replies	-0.02	0.01	**0.53**	1.00	**-0.06**	**-0.11**	**-0.04**	0.02	0.00	**0.34**	**-0.24**	**0.03**	0.00	**0.03**	**-0.06**	-0.01
% of Scheduled Messages Before Engagement	**-0.34**	**-0.11**	**-0.06**	**-0.06**	1.00	**0.21**	**-0.54**	**0.05**	**0.07**	**0.03**	**-0.11**	**-0.04**	0.01	**-0.03**	-0.01	0.02
% of Scheduled Messages Before Disengagement	**0.46**	**0.07**	**-0.11**	**-0.11**	**0.21**	1.00	**0.71**	**0.11**	**0.07**	**-0.04**	0.00	**0.12**	**0.07**	-0.02	**-0.05**	**-0.06**
% of Scheduled Messages Engaged	**0.66**	**0.14**	**-0.05**	**-0.04**	**-0.54**	**0.71**	1.00	**0.06**	0.01	**-0.06**	**0.07**	**0.13**	**0.05**	0.00	**-0.04**	**-0.07**
Response Time to Scheduled Messages (Hours)	0.03	0.02	**0.04**	0.02	**0.05**	**0.11**	**0.06**	1.00	**0.88**	**0.09**	**-0.10**	**0.04**	-0.01	0.00	**-0.07**	**0.04**
Response Time to Any Advising Messages (Hours)	0.00	**-0.09**	0.02	0.00	**0.07**	**0.07**	0.01	**0.88**	1.00	**0.08**	**-0.11**	0.02	0.02	**-0.03**	**-0.04**	**0.04**
Proportion of Messages Asking for Help	**-0.07**	0.02	**0.36**	**0.34**	**0.03**	**-0.04**	**-0.06**	**0.09**	**0.08**	1.00	**-0.34**	0.00	0.02	**0.12**	**-0.11**	**0.04**
Average Sentiment of Replies	**0.06**	-0.02	**-0.21**	**-0.24**	**-0.11**	0.00	**0.07**	**-0.10**	**-0.11**	**-0.34**	1.00	**-0.05**	**-0.10**	**-0.11**	**0.14**	**0.08**
% of Replies About Course Planning	**0.09**	**0.08**	**0.06**	**0.03**	**-0.04**	**0.12**	**0.13**	**0.04**	0.02	0.00	**-0.05**	1.00	-0.02	**-0.22**	**-0.19**	**-0.34**
% of Replies About Financial Aid	**0.03**	0.01	0.02	0.00	0.01	**0.07**	**0.05**	-0.01	0.02	0.02	**-0.10**	-0.02	1.00	**-0.23**	**-0.06**	**-0.20**
% of Replies About Academic Planning	-0.01	**0.08**	**0.11**	**0.03**	**-0.03**	-0.02	0.00	0.00	**-0.03**	**0.12**	**-0.11**	**-0.22**	**-0.23**	1.00	**-0.17**	**-0.41**
% of Replies About Academic Supports	0.00	**-0.16**	**-0.07**	**-0.06**	-0.01	**-0.05**	**-0.04**	**-0.07**	**-0.04**	**-0.11**	**0.14**	**-0.19**	**-0.06**	**-0.17**	1.00	**-0.27**
% of Replies About Meeting Logistics	**-0.07**	0.01	**-0.08**	-0.01	0.02	**-0.06**	**-0.07**	**0.04**	**0.04**	**0.04**	**0.08**	**-0.34**	**-0.20**	**-0.41**	**-0.27**	1.00

Note: Each cell represents the Pearson's correlation coefficient between the two engagement measures indicated. Bolded numbers are significant at the $p<0.05$ level.

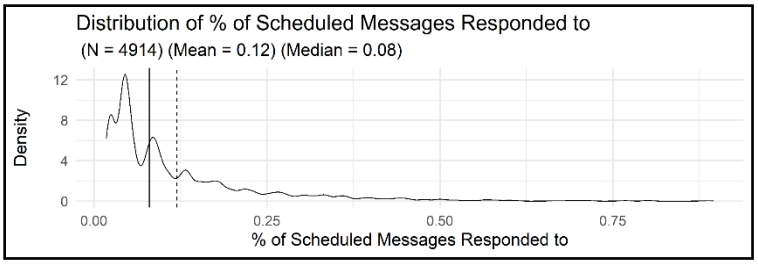
Table 2.5. Relationships Between Engagement Measures and Student Outcomes

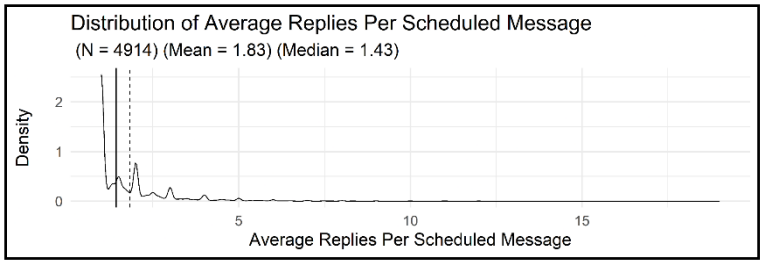
	T+1				T+2			
	Re-Enrolled	Credits Earned	GPA	Degree Receipt	Re-Enrolled	Credits Earned	GPA	Degree Receipt
% of Scheduled Messages Responded to	0.05 (0.05)	**4.95*** (1.20)**	**0.62*** (0.11)**	**0.31*** (0.05)**	0.02 (0.07)	**5.66*** (1.71)**	**0.56*** (0.10)**	**0.31*** (0.06)**
Average Replies Per Scheduled Message	**-0.02*** (0.01)**	-0.09 (0.13)	**-0.03** (0.01)**	0.01 (0.01)	**-0.02** (0.01)**	-0.25 (0.17)	**-0.03** (0.01)**	-0.00 (0.01)
Average Reply Length in Words	-0.00 (0.00)	**-0.03* (0.01)**	0.00 (0.00)	-0.00 (0.00)	-0.00 (0.00)	-0.03. (0.02)	0.00 (0.00)	-0.00 (0.00)
Proportion of Substantive Replies	-0.02 (0.02)	-0.53 (0.42)	0.08. (0.04)	0.02 (0.02)	**-0.07** (0.02)**	-1.03. (0.55)	0.07. (0.04)	0.01 (0.02)
% of Scheduled Messages Before Engagement	-0.04 (0.03)	-0.97. (0.57)	0.00 (0.06)	-0.02 (0.02)	-0.03 (0.03)	**-1.72* (0.76)**	0.01 (0.05)	**-0.07** (0.03)**
% of Scheduled Messages Before Disengagement	**0.13*** (0.02)**	**3.17*** (0.44)**	**0.15** (0.05)**	-0.03. (0.02)	**0.12*** (0.03)**	**4.69*** (0.60)**	**0.15*** (0.04)**	0.02 (0.02)
% of Scheduled Messages Engaged	**0.10*** (0.02)**	**2.76*** (0.38)**	**0.12** (0.04)**	-0.00 (0.02)	**0.09*** (0.02)**	**4.15*** (0.52)**	**0.11** (0.04)**	**0.06** (0.02)**
Response Time to Scheduled Messages (Hours)	0.00. (0.00)	-0.00 (0.00)	0.00 (0.00)	-0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.00 (0.00)	-0.00 (0.00)
Response Time to Any Advising Messages (Hours)	0.00 (0.00)	-0.00 (0.01)	0.00 (0.00)	-0.00 (0.00)	0.00 (0.00)	-0.00 (0.01)	0.00 (0.00)	-0.00 (0.00)
Proportion of Messages Asking for Help	-0.00 (0.02)	0.03 (0.44)	0.03 (0.04)	0.00 (0.02)	-0.00 (0.02)	-0.18 (0.58)	0.04 (0.04)	-0.01 (0.02)
Average Sentiment of Replies	0.02. (0.01)	**0.76*** (0.23)**	**0.04* (0.02)**	0.01 (0.01)	**0.04** (0.01)**	**1.28*** (0.31)**	**0.06** (0.02)**	**0.03** (0.01)**
% of Replies About Course Planning	**0.14** (0.05)**	**2.10* (1.05)**	0.09 (0.10)	**-0.22*** (0.05)**	**0.21*** (0.06)**	**4.24** (1.46)**	0.09 (0.10)	**-0.15** (0.05)**
% of Replies About Financial Aid	0.03 (0.08)	-2.85 (1.76)	**-0.40* (0.17)**	**-0.22** (0.07)**	0.01 (0.10)	-1.78 (2.33)	**-0.37* (0.16)**	**-0.32*** (0.09)**
% of Replies About Academic Planning	**-0.20*** (0.05)**	0.63 (0.89)	**0.26** (0.08)**	**0.42*** (0.05)**	**-0.29*** (0.05)**	-1.74 (1.20)	**0.19* (0.08)**	**0.36*** (0.05)**
% of Replies About Academic Supports	0.00 (0.09)	-2.45 (1.86)	-0.20 (0.18)	-0.14 (0.09)	**0.22* (0.11)**	-1.30 (2.57)	-0.15 (0.17)	-0.19. (0.10)
% of Replies About Meeting Logistics	**0.16*** (0.05)**	**2.10* (0.92)**	-0.06 (0.08)	**-0.16*** (0.04)**	**0.16** (0.05)**	**3.35** (1.27)**	-0.00 (0.08)	-0.04 (0.05)

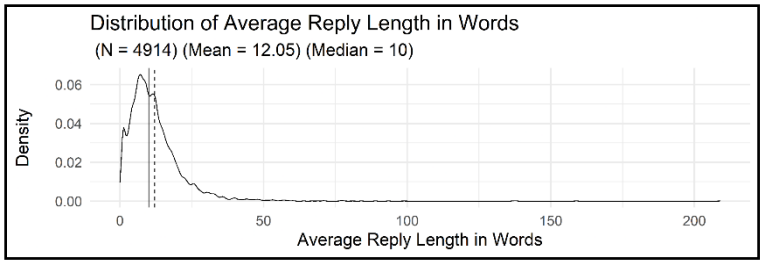
	T+3			
	Re-Enrolled	Credits Earned	GPA	Degree Receipt
% of Scheduled Messages Responded to	-0.04 (0.07)	**4.93* (2.37)**	**0.60*** (0.11)**	**0.29*** (0.07)**
Average Replies Per Scheduled Message	**-0.02* (0.01)**	-0.30 (0.20)	**-0.04** (0.01)**	-0.01 (0.01)
Average Reply Length in Words	-0.00 (0.00)	**-0.05* (0.02)**	0.00 (0.00)	-0.00 (0.00)
Proportion of Substantive Replies	**-0.05* (0.03)**	**-1.66* (0.74)**	0.07 (0.05)	0.01 (0.02)
% of Scheduled Messages Before Engagement	0.04 (0.04)	-1.17 (1.00)	-0.02 (0.06)	-0.02 (0.03)
% of Scheduled Messages Before Disengagement	**0.06* (0.03)**	**5.67*** (0.81)**	**0.13** (0.05)**	**0.08** (0.03)**
% of Scheduled Messages Engaged	0.02 (0.02)	**4.58*** (0.72)**	**0.11** (0.04)**	**0.08*** (0.02)**
Response Time to Scheduled Messages (Hours)	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)	-0.00 (0.00)
Response Time to Any Advising Messages (Hours)	0.00 (0.00)	0.00 (0.01)	0.00 (0.00)	-0.00 (0.00)
Proportion of Messages Asking for Help	0.01 (0.03)	-0.43 (0.80)	0.02 (0.04)	0.02 (0.02)
Average Sentiment of Replies	0.02 (0.01)	**1.57*** (0.40)**	**0.07** (0.02)**	**0.03** (0.01)**
% of Replies About Course Planning	0.12. (0.07)	**6.28** (2.05)**	0.07 (0.11)	-0.06 (0.06)
% of Replies About Financial Aid	0.08 (0.11)	-0.21 (3.07)	-0.32. (0.17)	**-0.21* (0.10)**
% of Replies About Academic Planning	**-0.20*** (0.06)**	**-3.67* (1.59)**	0.16. (0.08)	**0.26*** (0.05)**
% of Replies About Academic Supports	0.06 (0.12)	-0.54 (3.23)	-0.08 (0.19)	-0.20. (0.10)
% of Replies About Meeting Logistics	0.09 (0.06)	**3.90* (1.71)**	-0.01 (0.09)	0.00 (0.05)

Note: All coefficients shown above are the result of a regression as described in Section IVe that includes student academic baseline and demographic characteristics, as well as student randomization block fixed effects. Each engagement measure is then included in the regression with the given outcome of interest *without* any other engagement measure. Thus, each cell represents its own separate regression. Standard errors in parentheses. Bolded coefficients are significant at the p<0.05 level. (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Figure 2.1. Distribution of Engagement Measures: Frequency and Intensity of Student Replies







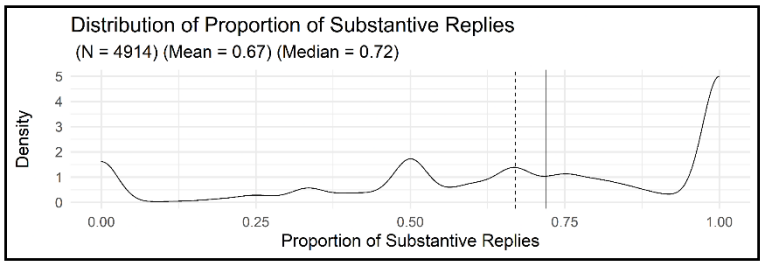
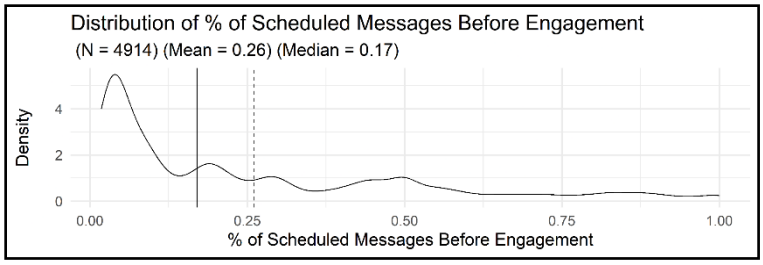
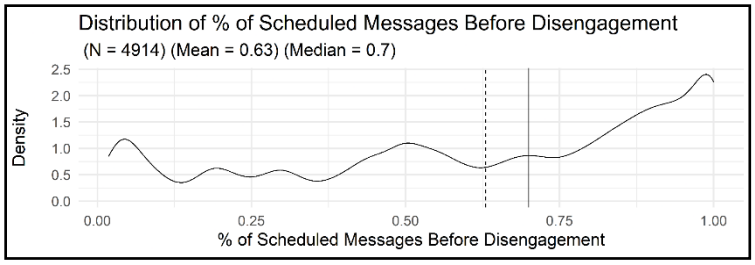


Figure 2.2a. Distribution of Engagement Measures: Engagement Duration





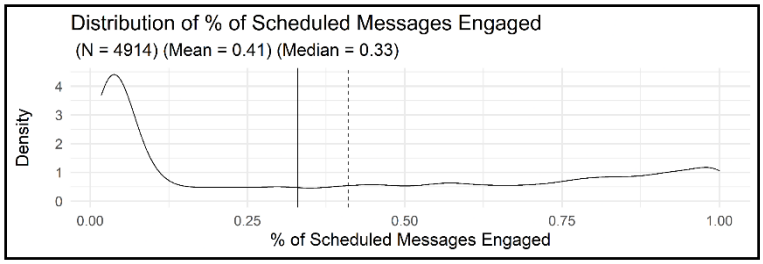


Figure 2.2b. Scatterplot of Engagement Duration Patterns

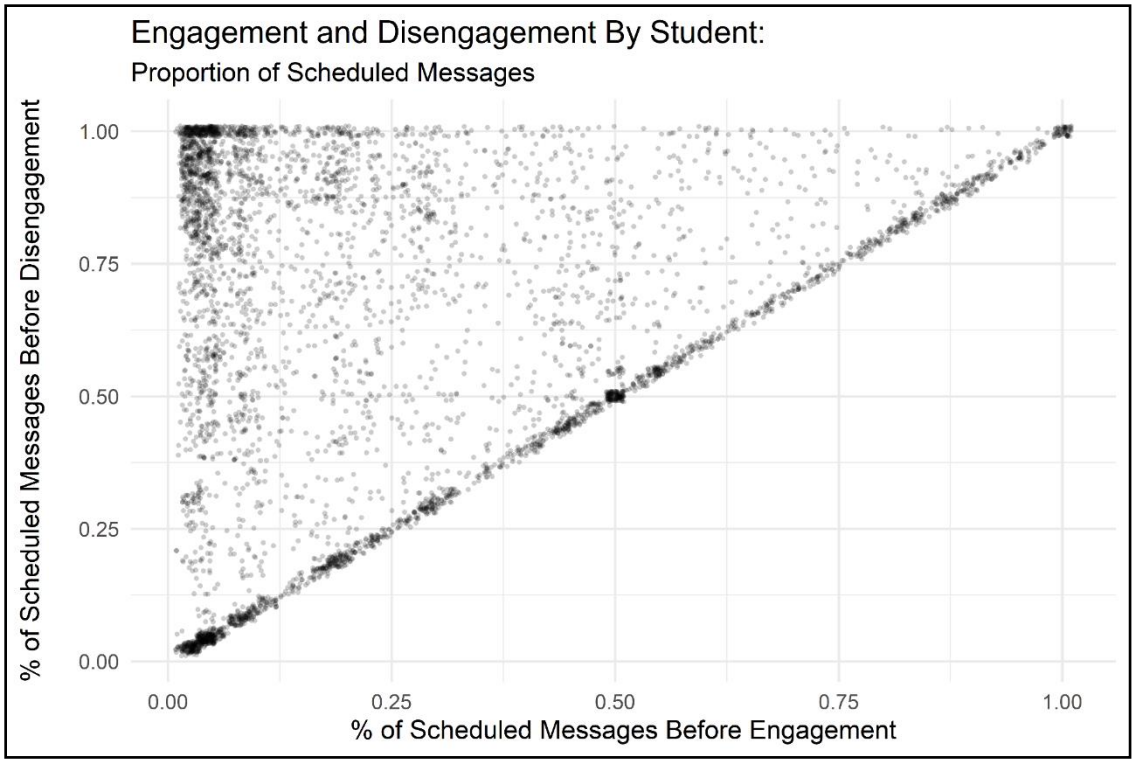


Figure 2.3. Distribution of Engagement Measures: Response Speed

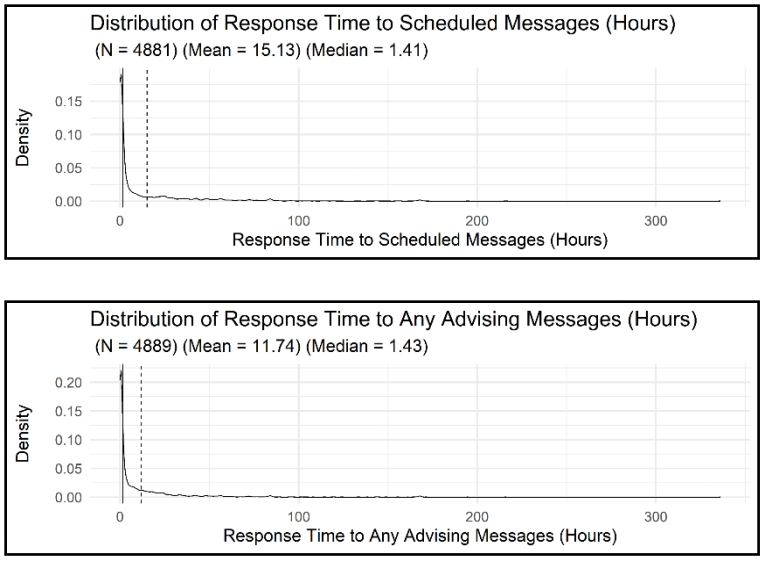


Figure 2.4a. Distribution of Engagement Measures: Help-Asking and Sentiment Response Content

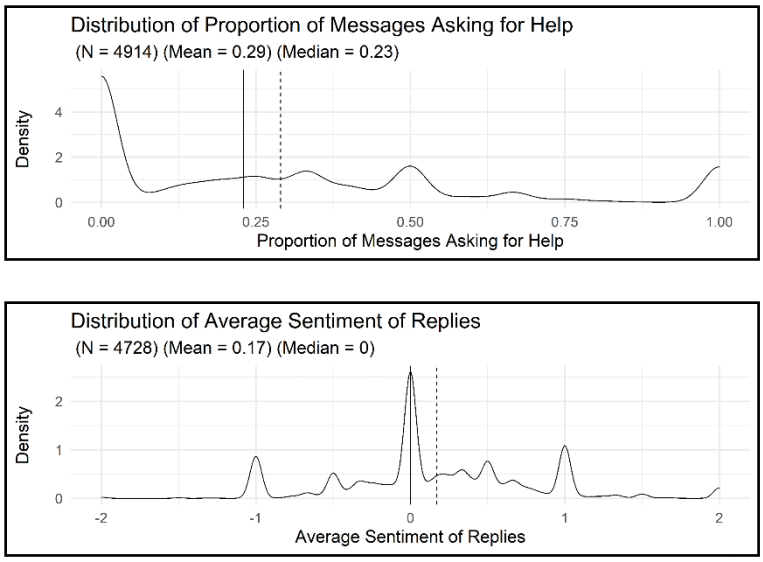
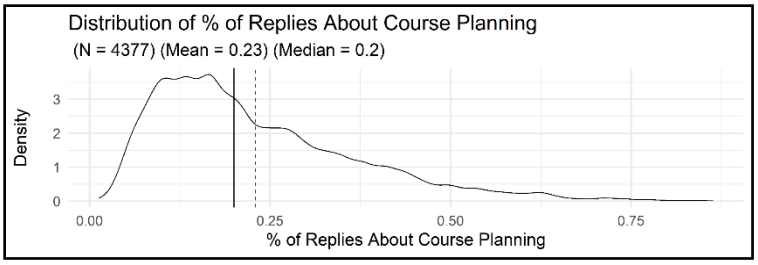
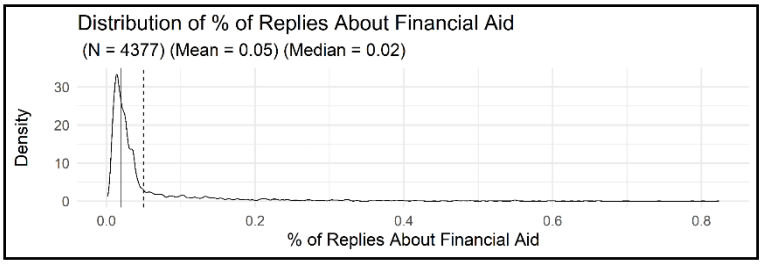
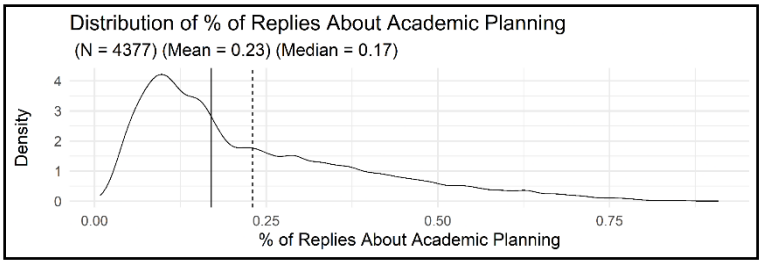
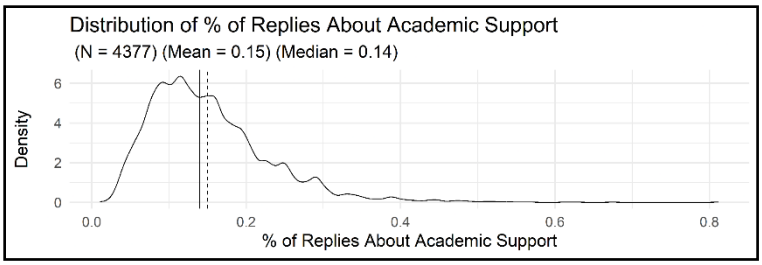


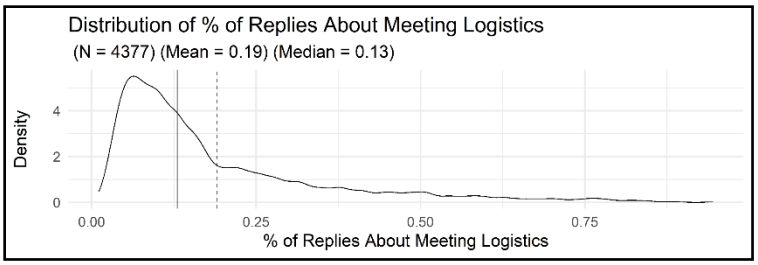
Figure 2.4b. Distribution of Engagement Measures: Topic Modeling Response Content











CHAPTER 3

What's in a Letter? Using Natural Language Processing to Investigate Systematic Differences in Teacher Letters of Recommendation

Brian Heseung Kim

Abstract

While scholars have already uncovered many ways that inequities can manifest across the postsecondary application portfolio – from standardized tests to advanced course-taking opportunities – we know almost nothing about whether teacher letters of recommendation also present differential barriers to students' college aspirations. This blind spot is especially concerning given mounting evidence that recommendation letters in other contexts can contain biased language, that teachers can form biased perceptions of their students' abilities, and that narrative application components more generally may contribute to racial discrimination in selective college admissions. In this paper, I conduct the first system-wide, large-scale text analysis of teacher recommendation letters in U.S. postsecondary applications using data from 1.6 million students, 540,000 teachers, and 800 postsecondary institutions. I use sophisticated natural language processing methods to examine the prevalence of potential inequities within these letters: whether students are described by teachers in systematically different ways across race and gender groups, even after accounting for salient confounding factors like student academic and extracurricular qualifications, teacher fixed effects, and institution fixed effects. I find evidence of salient linguistic differences in letters across gender, but less evidence for differences across race – except in the case of highly competitive admissions, where both Black and Asian students tend to have markedly different letters than White students. Moreover, these differences are generally most meaningful in terms of the *topical content* of letters; differences in terms of the *positivity* of letters are far smaller in relative magnitudes and thus are less likely to be perceptible in the actual reading of letters. Taken together, these findings have broad implications for the use of recommendation letters in selective admissions, affirmative action policies, and gender diversity in STEM fields.

I. Introduction

Researchers have identified numerous sources of systematic disadvantage for low-income, racial/ethnic minority, or first-generation-to-college ("first-gen") students across nearly every stage of the postsecondary application process (Page & Scott-Clayton, 2016; Hoxby & Avery, 2013). One prominent component of the college application process remains largely unexamined for such inequities, however: the teacher recommendation letter. Every year, hundreds of thousands of teachers across the country write millions of letters of recommendation in support of college-aspiring students. To illustrate, 40% of the 904 colleges and universities using the ubiquitous Common Application ("Common App") require at least one teacher recommendation, alongside 40 of the top 50 national universities and 48 of the top 50 national liberal arts colleges (as ranked by the U. S. News & World Report in 2019; author's calculations). Available evidence also suggests that these letters carry heft in admissions decisions: 57% of the admissions officers surveyed by the National Association for College Admission Counseling responded that teacher letters were of "considerable" or "moderate importance" in their decision-making – placing them above other salient factors like student extracurriculars, class rank, interviews, and even advanced placement exam scores (Clinedinst & Koranteng, 2017). Given that many institutions, including 89% of colleges that accept the Common Application, moved to test-optional application policies for the 2021-2022 application cycle, and "holistic" admissions practices continue to grow in popularity, admissions officers are likely to weigh these letters even more heavily going forward (author's calculations; Rosinger et al., 2020).

Despite their rising prevalence and importance, there is very limited evidence on whether teacher recommendation letters offer a more equitable indication of student qualifications for the college application process. On the one hand, letters and other

"holistic" application components may provide admissions officers a more complete picture of students' assets, backgrounds, and potential contributions to college communities, perhaps offsetting reliance on other application components, such as test scores, that may be biased in favor of students from more privileged backgrounds (Bastedo et al., 2019). On the other hand, the content of these letters may actually reflect and deepen existing inequities by presenting biased depictions of student character and ability. Researchers have previously found that K-12 teachers hold implicit racial biases (Starck et al., 2020) and adjust both their expectations and evaluations of students based on race and gender at other stages of the educational pipeline (Dee, 2005; Grissom & Redding, 2016). If the language, content, and tone of recommendation letters are influenced by such biases, the growing influence of letters in college admissions decisions may more specifically disadvantage racial minority and female students in this increasingly competitive process.

In this paper, I conduct the first system-wide, large-scale text analysis of teacher recommendation letters in postsecondary applications to dissect the language of these letters and explore the potential for systematic racial and gender differences within them. In partnership with the Common Application, I analyze the universe of teacher recommendation letters submitted via their platform for the 2018-2019 and 2019-2020 application seasons,[35] as well as the student applications the letters are associated with: a total of approximately 2.5 million unique recommendations and 1.5 million unique students across 800 postsecondary institutions. I combine rigorous econometric frameworks with

[35] While the 2019-2020 application season technically overlaps with the onset of COVID-19 in the U.S. circa March, 2020, the overwhelming majority of recommendations were written and submitted earlier in the season. For reference, 99% of recommendations in my data from the 2019-2020 season were submitted prior to March 1, 2020. As such, changes to the content of recommendation letters, student application behaviors, or student qualifications due to circumstances in the pandemic are extremely unlikely to play a role in my analysis.

sophisticated natural language processing ("NLP") and text mining techniques from the field of data science to facilitate the rapid and consistent coding of linguistic features within the letters themselves. I leverage these methods to investigate the potential for systematic differences in **(a)** the words and phrases teachers use to write about students, **(b)** the topical content of letters, and **(c)** the perceived positivity of letter tone, all while accounting for the rich set of data on student qualifications, teacher letter writing experience, and student-teacher relationship dynamics included in their applications. My data moreover allow me to separately employ regressions with teacher fixed effects, thus estimating differences within the writing of *individual teachers*, and institution fixed effects, thus estimating differences within the pool of applicants to *individual institutions*. Lastly, I further explore whether differences in recommendation letters are more pronounced in particular institutional contexts (e.g., highly selective institutions) and teacher subject areas (e.g., STEM teachers).

Looking across my results, I find evidence of several salient linguistic differences across gender, but linguistic differences across race tend to be substantially smaller and less consistent. Moreover, these differences are generally most meaningful in terms of the topical content of letters; while I detect statistically significant differences in terms of the positivity of letters, most of these differences are far smaller in relative magnitudes and thus are less likely to be perceptible in the actual reading of letters.

In terms of more specific trends, letters for female students tend to be far more positive than letters for male students, holding constant academic and extracurricular background, but also tend to have substantially more discussion about topics often stereotypically associated with women: community engagement (i.e., community service, social good, etc.), extracurriculars (i.e., clubs, student government organizations, etc.), and time/life management (i.e., balancing a busy schedule, family responsibilities).

Commensurately, letters for female students tend to have less discussion about topics like their academic excellence (i.e., high quality of coursework, problem-solving skills), intellectual promise (i.e., curiosity, intellect, ability), sports, and STEM subjects (i.e., chemistry, math, computer science, etc.). These differences generally persist regardless of the specification and appear to be driven primarily by how *individual* teachers write about male and female students.

Letters about Black students tend to be quite similar to those about White students, depending on specification; while they have slightly fewer positive sentences and slightly more negative sentences, the magnitude of these differences are often smaller than 1% in relative terms. Content-wise, letters about Black students generally have greater discussion about community engagement and leadership (in and out of the classroom) and less discussion about sports and time and life management. Looking only at competitive applicants to highly selective institutions, it appears that Black students in this group tended to receive more discussion generally about topics related to their potential campus contributions (e.g., leadership and extracurriculars) and less discussion about certain academic topics (e.g., intellectual promise and humanities) – a dynamic that could potentially be concerning at the margin of these highly selective institutions depending on how these varied portrayals of students are ultimately valued by admissions officers.

Finally, letter trends for Asian students suggest slightly greater discussion about their community engagement, extracurriculars, STEM subjects, and future potential (i.e., likelihood of career success, business acumen, etc.) and slightly greater discussion about their intellectual promise when compared to letters about White students, regardless of specification. Interestingly, there exist no consistent differences in topics of discussion closely related to their personality (e.g., character excellence, which includes traits like

diligence, conscientiousness, commitment, etc.) – even among only those competitive applicants to highly selective institutions. Asian students in this competitive subgroup receive, if anything, slightly more positive letters than White students on average, though again this difference is substantively very small. While I cannot rule out their letters may nonetheless be different on measures I do not observe, my results indicate that letters for equally competitive Asian and White students are broadly similar in content and tone.

Ultimately, this study offers several timely contributions. First, this work advances the interdisciplinary literature on understanding systematic differences in recommendation letter writing and quantifying such differences using robust NLP and text mining frameworks. The data I leverage in my analysis likely represent the largest known repository of recommendation letters linked to individual qualifications and recommender characteristics in *any* evaluation context (e.g., graduation school admissions, job applications), thus offering greater precision and external validity than prior work. Paired with recent advances in the accuracy, complexity, and robustness of NLP techniques, I am also able to better account for more complex linguistic phenomena like sentence structure and context-specific word meanings to explore the phenomena of systematic differences with finer nuance than earlier computation-based attempts (e.g., Akos & Kretchmar, 2016; Schwarz, 2016).

As such, my results provide crucial insight for practitioners, policymakers, and researchers as they consider the potential implications of using recommendation letters in selective admissions processes. The fact that I find somewhat inconsistent and often negligible differences with respect to student race *after* considering other student and teacher characteristics indicates that these letters do not necessarily disadvantage some students versus others in those regards, though the exact truth of this statement hinges on: (a) the

extent to which admissions counselors attempt to "norm" letters for other salient student characteristics (i.e., compare letters only between highly similar applicants), and (b) exactly what characteristics about the letters are ultimately considered for admissions decisions. To the extent that I am able to observe, anecdotal differences in letter content by student race may more likely be driven by systematic differences along other elements of the student portfolio (e.g., extracurricular involvement and advanced coursetaking). Such results have immediate implications for high-profile affirmative action litigation (e.g., *Students for Fair Admissions v. Harvard*), in particular. But they also point to the possibility that letters *can* serve as a fair representation of student characteristics and credentials in admissions processes holding constant other factors (e.g., access to extracurriculars and advanced coursetaking) – an increasingly important consideration for admissions criteria as test scores continue to be de-emphasized.

The important exception here is with respect to letters for female applicants, which is of particular interest given widespread policy and organizational efforts to increase (e.g., the National Science Foundation ADVANCE program) the gender diversity of many STEM fields and professions. That differences in letter content between male and female applicants with respect to discussion of STEM subjects persisted across all specifications, and were in fact larger when examining only those letters written by STEM teachers, suggests that female applications to STEM-centric programs or institutions may be disadvantaged if admissions counselors rely, at least partially, on these letters to contextualize female student STEM capabilities. To the extent that admission to STEM-centric programs then has implications for the composition of female professionals in STEM fields, greater policy attention may need to be paid to how recommendation letters are utilized in these processes, and how

biases in the STEM *teacher* workforce may impact female students' STEM aspirations otherwise.

Finally, this paper contributes to the development of robust methodological frameworks for using NLP and text mining in education policy research. Given the novelty of these methods in education research, the personal nature of the data at hand, and the many subjective analytic decisions inherent to this work, my intention is for this paper to serve as an exemplar for the effective and responsible application of NLP by transparently modeling the procedures, validity checks, and theoretic frameworks necessary for success. To that end, I provide ample documentation of my methodology, decision-making, and processes in this paper, and I further provide all of my code open-source to other researchers intending to replicate or build on my approaches here.

The remainder of this paper proceeds as follows: Section II reviews relevant literature; Section III describes my data and analytic sample; Section IV provides an overview of my analytic approach, both for NLP methods and regression specifications; Section V explores the results of my main and subsample analyses; and Section VI concludes and discusses future directions for this research.

II. Literature Review

IIa. Racial and Gender Bias in Evaluations of Students

Central to my study of systematic differences in recommendation letter writing is the prevalence of implicit bias among teachers. Under the framework of implicit bias, the unconscious associations and stereotypes that an individual holds about particular groups go on to shape their perceptions and judgments of people within those groups (Bertrand et al., 2005; Greenwald & Krieger, 2006); the result is a form of discrimination that can be hard to

identify as those involved are often unaware it is happening (Devine et al., 2012). Empirical work has shown that these biases are difficult to intervene upon even in the best of circumstances (Lai et al., 2014), and they are particularly impactful on judgment and decision-making in circumstances where individuals are rushed (Payne, 2006), fatigued (Ma et al., 2013), or distracted (Danziger et al., 2011).

Classroom environments often align with these conditions, and the literature indicates that racial and gender biases may indeed be prevalent among teachers. First, we have evidence that teachers can hold negative implicit associations of Black individuals (Starck et al., 2020), and also that the degree of negativity varies by geography, school demographic composition, and regional test score disparities (Chin et al., 2020). We also have evidence that such biases can be consequential: math teachers with more negative stereotypes about female students were more likely to encourage female students to pursue vocational tracks instead of scientific/academic ones (Carlana, 2019) and even grade their exams more harshly (Avitzour et al., 2020). Research on demographic matching between teachers and students also offers additional evidence for the prevalence of such biases: white teachers held systematically lower expectations for Black students than Black teachers of the *same* students (Gershenson, Holt & Papageorge, 2016), with similar patterns for gender-matching (Dee, 2005).

IIb. Racial and Gender Bias in Recommendation Letters

While there is consistent evidence for the prevalence of biased attitudes and expectations among educators, research on whether and how these biases ultimately manifest in letters of recommendation is much more mixed. Much of the existing evidence comes from smaller-scale studies of letters at individual institutions, so the lack of

consistency may reflect lack of statistical power, sample idiosyncrasies, and other contextual differences (e.g., institutional selectivity). Moreover, they deploy a range of text analysis methodologies that trade nuance against scalability, from subjective ratings of letters by trained readers to simple word-count analyses.

There is suggestive evidence that female applicants to STEM-related academic positions are described with less exemplary adjectives and phrases (e.g., "good" versus "phenomenal") after controlling for observable qualifications (n=880; Schmader et al., 2007) and less positivity in general (n=1,224; Dutt et al., 2016). I refer to these phenomena as letter "**tone**." Similarly, researchers have found female applicants can be described with more "communal" terms (e.g., "team player," "helpful") and fewer "agentic" terms (e.g., "leader," "pioneer") than male counterparts (n=624, Madera et al., 2009). Letters for female applicants are also more likely to have "doubt-raising" language: language that is outright negative, preceded by hedging, or irrelevant (Madera et al., 2018). I refer to these phenomena as letter "**word choice**."

Importantly, however, these studies all note that letters were broadly more similar than different, and they often failed to find anticipated evidence of bias along other measures. For example, a study of teacher recommendation letters for a single postsecondary institution found that female students had letters that were *more* positive, while racial minority students had letters that were more neutral. But when looking at whether teachers focus on describing different student characteristics across groups (e.g., athletic ability versus intellectual curiosity), the author found very few differences (n=24,000; Schwarz, 2016). I refer to these phenomena as letter "**topical content**." In a similar study at a different institution, researchers found that that female students were more likely, and racial minority students less likely, to have letters using "grindstone" words (e.g., "hardworking," "diligent")

– but found no meaningful differences along any other conceptually-relevant categories like achievement (n=4,792; Akos & Kretchmar, 2016).

These studies provide crucial groundwork to identify the ways implicit bias may manifest as systematic differences in letters (tone positivity, word choice, and topical content), but they also point to the need for studies that can at least partially overcome the steep tradeoff between scalability/generalizability and analytic nuance. A growing literature in education has demonstrated the utility of NLP methods to this end (Anglin, 2019; Fesler et al., 2019), motivating and guiding my application of these approaches for the present study.

IIc. Theoretical Model for Implicit Bias in Teacher Recommendations

To summarize the role that implicit bias likely plays in the teacher recommendation process given prior literature, I create a concise theoretical model in Figure 3.1. The first row across the top is a simplified sequence of events describing how student actions are eventually translated into the recommendation letter that admissions counselors ultimately review. That is, we begin with a bundle of student actions, interactions, and behaviors that exist in reality; these behaviors are then interpreted by the teacher and attributed to the student. Assuming the student asks the teacher for a recommendation and the teacher agrees, the teacher will then transcribe some version of their subjective perception of the student's actions/behaviors/etc. into the text of a recommendation letter. Lastly, the letter is interpreted by the admissions counselor, whose evaluation of the letter is utilized alongside a complex host of additional information and context into an actual admissions decision (Clinedinst & Koranteng, 2017; Schwarz, 2016).

Importantly, teachers' implicit biases can conceptually influence this process at two distinct points: the formation of a teacher's perceptions about the student's actions (i.e., between the first and second boxes), and then the teacher's transcription of those perceptions into written language (i.e., between the second and third boxes). The extent to which bias plays a role in either of those moments would be driven by the teacher's own implicit biases about the student's demographics (e.g., if a teacher has a negative bias against female students in the sciences, per Carlana, 2019), which in itself is influenced by the teacher's self-identity as well (e.g., if a Black teacher is writing for a Black student, per Gershenson, Holt, & Papageorge, 2016). Moreover, constraints on the teacher's time and attention, as well as other stressors, would likely exacerbate the influence of any such biases as well (e.g., per Payne, 2006).

Importantly, this model relies on the teacher's *perceptions* of student demographics, rather than the student's self-identified demographics. In the framework of implicit bias, we should only expect student race/gender to influence teacher writing if the teacher believes the student to be of a specific demographic and holds implicit attitudes about that demographic. For example, if a teacher does not perceive a student as Black, their implicit attitudes about Black individuals should not meaningfully come into play either in observations of the student or while writing about the student; this would be indistinguishable in my data from a teacher who *does* perceive a student as Black, but holds no negative implicit attitudes about Black individuals. In short, teacher misperception or ignorance of student demographics can complicate the interpretation of my results. It is for this reason that I also focus the present study on those demographic characteristics that are

most likely to be (though not guaranteed to be) salient to teachers, such as race and gender,[36] rather than those that likely rely on stronger assumptions about teachers' knowledge of students' backgrounds, like SES and ethnicity.[37]

I also include in this theoretical model an acknowledgement that these letters likely only impact admissions decisions insofar as they impact admissions readers' evaluations of the letters – which are in turn potentially subject to their biases as well, even if I am unable to examine this dynamic in the present study. There is some experimental evidence that untrained readers of letters can actually read the *same letter* and evaluate it differently based on experimentally-manipulated student descriptions the letters are paired with (e.g., race, names; Madera et al., 2018; Morgan, Elder, & King, 2013). Fact-finding from the Students for Fair Admissions (SFFA) v. Harvard University (2019) case made it clear that formal implicit bias training for admissions counselors is rare, even at highly selective, well-resourced institutions like Harvard; admissions counselors may then exhibit behavior similar to the untrained experiment participants as well. While some admissions officers might be aware of bias and attempt to account for it in their reading of letters out of personal motivation (Schwarz, 2016), we do not have evidence as to whether counselors are actually successful in this endeavor. Regardless, research on the mitigation of implicit bias in other settings suggests that bias is quite pernicious, even in the face of thoughtful and well-devised intervention (Devine, Forscher, Austin, & Cox, 2012; Paluck & Green, 2009).

---

[36] To be clear, race and gender identity are deeply complicated topics, and students may not necessarily *present* as the race or gender that they *identify* as. That said, gender in this context is likely known to teachers by virtue of how often teachers interact with students and discuss them in the third person (i.e., a student's preferred pronouns are likely known by the teacher). Race may not be known to teachers, but as racial identity is often conflated in the United States with visible skin color, it may still be salient to their perceptions of students.

[37] Latinx identity is further complicated in my data due to the fact that Latinx status overrides student race (i.e., a student cannot be both Black and Latinx – they are only recorded as Latinx). Thus, students who identify as Latinx may be of any race in my data, conflating which biases are likely to be at play. I intend to study this dynamic in future work where I am able to disaggregate race from ethnicity.

Finally, I note that broader social and socioeconomic factors play an important role in nearly every element of my theoretical model. That is, they influence student behavior/actions, teacher biases, counselor biases, teacher and counselor self-identity, college admissions dynamics, and so on. I exclude these dynamics from the model only for visual clarity (i.e. many overlapping arrows), but recognize that I cannot extricate what I observe in my analyses from this context. For example, I control for student extracurriculars in my models, but must acknowledge that race and gender play a role in how students engage in, and have access to, extracurriculars to begin with. My intention is not to ignore these important disparities from other moments in a student's educational journey, but instead to present the best estimates I can on this specific margin and nuance our understanding of equity in college access in the process.

IId. Teacher Recommendation Letters in Admissions Processes

As I allude to in my theoretical framework, it is critical to note that the importance of any systematic differences in the writing of recommendation letters across student groups is only relevant for equity and fairness in college admissions insofar as admissions counselors actually incorporate letters into admissions decisions. To illustrate: if admissions counselors at a given institution completely disregard any recommendation letters they receive on behalf of students, systematic differences in the content, tone, and word choice of letters across student groups would have no impact on disparities in admissions. Conversely, if admissions counselors at a given institution disregard every application element *except* recommendation letters, systematic differences in the content, tone, and word choice of letters across student groups would drive the totality of disparities in admissions decisions.

Unfortunately, determining where along this spectrum the reality lies is both difficult to determine and highly contextual. Research on "holistic" admissions policies demonstrate that there is fairly broad variation in how different institutions approach the incorporation of subjective application factors (e.g., essays, recommendation letters, interviews; Bastedo et al., 2018) from a policy perspective, with some trends along the lines of institutional selectivity and sector (Hossler et al., 2019). For example, some institutions articulate trying to evaluate the "whole" student via reading their application components together, while other institutions instead use narrative elements and background characteristics together to "contextualize" students' academic achievements. And as Rosinger et al. (2020) synthesize, this picture becomes even more complicated given that how letters are used in decisions may even vary from student-to-student *within* institutions depending on the exact composition of their individual application and the cultural values of the individual admissions officer. Schwarz (2016) also points out that how admissions officers perceive the students' teachers, and the likely closeness of their relationship or their ability to speak to the interests of admissions officers, can also affect their willingness to weigh the letters in their decisions.

These factors, among others, will ultimately make it difficult to ascertain the exact implications of the results I produce in this study without further insight into specific cases of concern. Thus, I can speak only to the broad trends in the data I analyze here, with the primary goal of offering researchers, practitioners, and policymakers a *starting place* to interrogate some of the ways these letters may influence later decision-making in the admissions process with their own expertise.

III. Data and Sample

My primary data consist of all applications submitted through the Common App platform in the 2018 and 2019 application cycles. This robust dataset consolidates de-identified versions of all application materials submitted by students, teachers, and counselors via their platform.[38] For the present study, I limit my sample to first-year applicants (thus excluding transfers) with completed applications who submitted at least one teacher recommendation.[39] About 825,000 such students submit an application through the Common App each year, yielding a total of about 1.65 million students and 2.8 million teacher recommendations in my complete dataset.

These data contain nearly every field self-reported by students in their applications: gender, race/ethnicity, socioeconomic status (parental education and an indicator for receipt of an application fee waiver for low-income students), academic performance (GPA, class rank, SAT/ACT, AP/IB tests), extracurricular involvement (separate open text responses for activity names and positions held), anonymized IDs of colleges applied to, and anonymized high school IDs. Note that the data I have at my disposal are not validated by any official sources (e.g., cross-referenced against their transcripts or counselors); to the extent that students perceive this as a high-stakes process with great consequences for dishonesty, deliberate reporting error is unlikely to be a prevalent factor in this analysis. Missingness is still relatively high for certain academic characteristics given both that many are not required fields on the application, and because inaccurate answers are easily identified

---

[38] In order to filter personally identifiable information (PII) from the text of the recommendation letters, researchers at the Common App utilized Amazon's proprietary Comprehend service to detect and scrub PII. While I am not able to comment on the exact details of how this algorithm functions, random checks and in-depth examinations of algorithm output show exceptionally low incidence of false positives and false negatives.

[39] Approximately 0.6% of students in the raw sample chose to apply in multiple years (e.g., sending a few applications in their junior year but ultimately re-applying in their senior year); for these students, I take only their most recent application and letter data. To calculate letter writing experience by teacher, however, I consider all applications from all students.

(e.g., SAT scores higher than 800 for a given subcategory, or GPAs more than twice as large as the reported GPA scale).

These data also contain every field self-reported by teachers in each of their recommendations -- anonymized school ID, courses taught with the student, and recommendation letter text -- as well as the total count of recommendations submitted by each teacher, each year, to proxy for their letter writing experience within the data timeframe. Importantly, I do not observe teacher demographics; as such, I intend to explore teacher-student demographic match dynamics using supplementary state staffing data in future work.[40] Finally, I observe the sector of students' high schools and the selectivity of institutions (proxied by median combined SAT score of admitted students). I intend to leverage additional high school and institution covariates as they become available in future iterations of this work.[41]

Table 3.1 displays descriptive statistics for my analytic sample of students. First, note that my analytic sample (i.e., the sample used to conduct all hypothesis testing, and as described in the table) encompasses only 90% of the full data available to me per recommendations of Egami et al. (2018) to impose a "training-testing" data split when using

---

[40] While theoretically possible to impute certain teacher characteristics from teacher names, such as gender and race, attempts to algorithmically derive individual demographics in this way carries with it deep ethical and moral concerns (e.g., that advancing such technology could eventually facilitate the systematic discrimination of individuals by their race/ethnicity if utilized by bad actors – or simply uncritically). This concern is especially pressing in my circumstance, as I intend for my work to serve as a model for future researchers, and I plan to open-source my code to the full extent practicable. Even from a straightforward methodological perspective, it is not possible with the provided data to assess the accuracy and potential biases in this prediction process, making it difficult to assess the confidence of any ensuing conclusions. For the totality of these reasons together, I will not attempt imputation analysis of this kind, and will hold this strand of demographic match analyses until supplementary data sources make these analyses possible without imputation.

[41] While the Common App combines information about students' high schools and the higher education institutions they are applying to with publicly available education data sources, I was not provided the majority of these additional covariates (e.g., those available through sources such as the NCES Common Core of Data or the Integrated Postsecondary Education Data System) to prevent the reidentification of partner schools and institutions.

natural language processing measures for hypothesis testing; I describe this decision and process in more detail in Section IVa as I describe my overarching analytic approach for text analysis. My final analytic sample thus encompasses about 1.5 million students and 2.5 million unique letters.

Beginning with the left panel, the students represented in these data are majority female (57%), with a small minority of international students (11%). About a quarter of students in the sample received a Common App fee waiver[42] for low-income students (23%), and 27% of the sample identified as first-generation students. In terms of student race/ethnicity, the sample is majority White (49%) with a meaningful level of missingness (13%) given that the question is optional to students; note also that in these data, Latinx ethnicity supersedes other racial identities. Students are predominantly coming from public school institutions (75%) and are split about evenly between the 2018 and 2019 cohorts.

In the right panel, I show statistics for a handful of key application and academic characteristics, revealing that this sample is a higher-performing subset of the overall high school student population. The majority of the sample sends at least 4 college applications through the Common App (and may send more outside of their platform), with a substantial 29% sending at least 8. Students also tend to have exceptionally high GPAs in the sample, with the majority reporting a scaled GPA of at least 0.9 (i.e., a 90 on a 100 point scale, or a 3.6 on a 4.0 scale). A substantial proportion also report GPAs above their reported scale maximum (26%), likely reflecting differences in school weighting schemes. Note also

[42] Fee waiver eligibility for the Common App is self-reported by students who: (**a**) Received a test fee waiver from the College Board, (**b**) Receive free or reduced price lunch, (**c**) are enrolled in a federal program specifically targeted at low-income students (e.g., Upward Bound), (**d**) receives public assistance (e.g., subsidized housing), and (**e**) are an orphan, ward of the state, in foster care, or experiencing homelessness.

relatively high levels of missingness for GPA (17%).[43] Looking finally at standardized test score percentiles,[44] much of the sample scored at least in the 75th percentile in both math and reading. Reflecting that test score reporting was not mandatory for many institutions in the sample timeframe, about 25% of the sample reported no test scores in their applications.

Table 3.2 examines the sample of teachers and their letter-writing experience. First, note that the sample overall contains letters from nearly 540,000 unique teachers across the two years, with about a third of those teachers writing letters in both years. Interestingly, the vast majority of teachers were asked to write relatively few letters each year, with 44% writing only one in a given year, and 37% writing only 2-5. When looking at a teacher's past letter writing experience, 42% of teachers in the sample had not written a single letter in the two years prior.[45]

Lastly, Table 3.3 examines the relationships between teachers and the students they wrote letters for, as reported by the recommending teachers themselves. First note that teachers predominantly taught students in the "core" subject areas: English, Social Studies, Math, and Sciences. A substantially smaller proportion taught subjects such as World Language and Computer Science, though many teachers indicated Other (15%). About 5% of teachers in the sample remarked on having some coaching relationship with the student (overwhelmingly in the realm of sports, though sometimes debate and similar activities). The

---

[43] Because students self-report their GPA and their school's GPA scale separately, these data needed to be cleaned extensively for use. Students who reported scaled scores above 1.5 were labeled as missing given likely conflict in their reported GPA and the scale to be used.

[44] To harmonize SAT and ACT scores, ACT scores were translated first into SAT scores using official ACT concordance tables for the appropriate test-taking year. SAT scores were then translated into percentiles using official SAT score reports for the appropriate test-taking year. If a student reported both an SAT and an ACT score for a given category (e.g., math), the higher score was taken.

[45] Because I do not observe the full history of past letter writing experience, I restrict my frame of observation to the prior two years of a teacher's history (e.g., the 2016 and 2017 seasons for a teacher writing in 2018).

majority of teachers knew students for a single year at the time of writing (51%), though about 32% knew them for two years. Note that in the "Years Known" question, teachers could also indicate teaching the student in some other capacity than a high school grade (e.g., middle school) – only 3.4% of teachers indicated this on a recommendation.

IV. Analytic Approach

My analysis is divided into two distinct phases: **(1)** deriving quantitative measures of letter characteristics with NLP and text mining, and **(2)** analyzing these letter characteristics in a regression framework to investigate differences in letter characteristics across student demographics while holding constant as much about the students and recommending teachers as possible.

IVa. Text Analysis Framework

In the first phase, I analyze the text of the teacher recommendation letters using NLP and text mining techniques to construct a series of numeric measures that reflect letter characteristics. This ultimately produces output similar to approaches where researchers manually code text for the incidence of certain phenomena or characteristics, but in a mostly automated and highly scalable way suitable for large-scale analyses such as mine.[46]

I leverage a sequence of three separate NLP techniques to characterize each letter along lines that may systematically differ based on prior literature: **(a) Word Frequency Analysis**, which tabulates the specific words used in each letter (word choice); **(b) Topic**

[46] That said, these text-as-data methods are not *substitutes* for rigorous qualitative document analysis – rather, they offer distinct and complementary insights at a scope particularly appropriate for research questions like my own.

Modeling, which quantifies the extent to which each letter discusses various categories of topical content (e.g., coursework, sports, leadership); and **(c) Sentiment Analysis**, which quantifies the estimated tone/positivity of language used in each letter. Because **(a) Word Frequency Analysis** produces high-dimensional output unsuitable for my main regression model, I consider this an exploratory analysis and include it in Appendix A3.1 for concision. In the following two sections, I outline my approach for **(b) Topic Modeling** and **(c) Sentiment Analysis.**

Importantly, note that I follow recommendations from Egami et al. (2018) to first develop and refine my approach for applying each of the aforementioned methods using a random subset of 10% of the overall letter data before applying these methods (without further adjustment) on the remaining 90% of data for use as my actual analytic sample.[47] To explain briefly, Egami et al. (2018) argue that NLP analysis pipelines for social science require a host of subjective decisions about data pre-processing, sample definitions, and model construction, and it is difficult to commit to any of these decisions a priori given the messy and unpredictable nature of text data. Thus, analysts risk "baking in" an anticipated effect no matter how cautious and judicious they are, as they are constantly responding to trends and issues they observe in the data along the way. Egami et al. (2018) thus recommend that analysts pilot, refine, and finalize their NLP pipelines using an entirely separate "training" subsample of their data first. Once the NLP pipelines are finalized, the "training" subsample is discarded and the remainder of the data is processed through the

[47] This split proportion was somewhat arbitrary; the scale of data available in this study makes even 10% of the sample more than sufficient to train the NLP algorithms well and vet edge-cases in the text cleaning pipeline. Using larger proportions quickly becomes intractable from a time and computational perspective, even while using supercomputing clusters, due to the sheer size of the data and the resource-intensivity of model training processes.

established pipeline – without any further adjustment – as the main analytic sample for hypothesis testing. To minimize reductions in the degree of common support in my sample due to this adjustment (an issue with the fixed effects approach I describe in more detail in Section IVd), I randomly sample *teachers* from the overall dataset into the training and analytic sets, stratified by their letter-writing volume across both years of data, and separate all of their letters accordingly.[48] Analyses to examine differences in salient sample characteristics across the training and analytic datasets can be found in Appendix A3.2, but I ultimately find no evidence of meaningful imbalance that would threaten the validity of the NLP pipelines I construct.

IVb. Topic Modeling

Through **(b) Topic Modeling**, I measure differences in the topics that teachers discuss in each letter. In this framework, keywords[49] that frequently appear together in the same paragraphs of letters across the dataset are thought to be associated with the same abstract "topic" of discussion, which analysts then interpret for their unifying substantive meaning (Blei, 2003).[50] For example, if "baseball," "competitive," and "sports" frequently

---

[48] Note that *students* are not exclusive across the two subsamples, as a teacher in the training dataset may have written a letter for a student who also had a letter written about them by a teacher in the analytic dataset.

[49] Note it is often prudent to include n-gram analysis in topic models, e.g., 2-3 word phrases, in addition to individual words. For clarity of explanation, I use "keywords" as a stand-in for n-gram for the rest of this document.

[50] In more technical terms, I am treating a paragraph, not a letter, as a "document" for training the topic model. This is because these letters are all roughly discussing the same broad topic (students and their academic qualifications), and variation at the letter-level would likely be insufficient to identify meaningfully distinct topics. By contrast, paragraphs within the letter are more likely to be varied in topic; e.g., a teacher might write about a student's achievements in a particular class in one paragraph, and then write about a student's involvement in school leadership in the next. A similar adjustment was helpful in improving the interpretability of topical output in Kim et al. (2021). To collapse these down to the letter level, I analyze each paragraph for its prevalence of each topic in terms of estimated keyword counts and sum the keyword counts across all paragraphs in the letter.

appeared together in paragraphs throughout the dataset, the algorithm would identify them as belonging to the same topic; an analyst might subjectively interpret this keyword group as representing the substantive topic of "student sports involvement." Importantly, the algorithm identifies several such topics *based on the provided text data* – this allows for uniquely context-sensitive and flexible output compared to methods using predefined keyword groups (e.g., the commonly-used LIWC), but also means that the substantive topics identified by the algorithm are not knowable before analysis. Once the topics are identified, each recommendation letter can be quantified for the number of keywords spent discussing each topic, based on the combination of keywords within each of its constituent paragraphs.[51]

To improve the tractability of the topic model output for later regression analyses, I transform the ensuing measures in two important ways. I first consolidate closely related topic groupings into a smaller number of substantively-relevant "supertopic" groups using an approach that mirrors the spirit of qualitative thematic analysis, thereby reducing their dimensionality and improving their interpretability (piloted in Kim et al., 2021).[52] For example, a topic describing a student's commitment and dedication might be combined with

---

[51] In this study, I leverage the implementation by Roberts et al. (2019) in R known as Structural Topic Modeling (the "stm" package). This implementation offers a number of important methodological advancements over the standard Latent Dirichlet Allocation (Blei, 2003), most notably the ability to allow the prior distribution of topic prevalence in the data to vary based on document metadata and improve the overall fit of the model. I thus use the array of covariates present in my regression model as "topic prevalence covariates" in the stm model; it is recommended by Roberts et al. (2019) that these models be congenial in terms of covariate inclusion. Intuitively, this risks "baking in" an anticipated relationship (e.g., if student race is allowed to influence the topic measurement process, and we observe a relationship in a certain topic's prevalence and student race later on); I remark on my approach to overcoming this issue in the prior section using the adjustment proposed by Egami et al. (2018).

[52] In brief, I begin with a list of the most frequent-and-exclusive and highest probability words from each topic. I then group topics into conceptually coherent categories based on apparent themes and salient word usages (e.g., by examining snippets of words being used in context). In the present analysis, this process is conducted only by myself; in future iterations of this work, I intend to use an iterative grouping process alongside a team of researchers to align supertopic groupings more robustly.

a topic describing a student's diligence and maturity to produce a supertopic more generally about a student's Character Excellence. This produces a continuous measure indicating the number of keywords in each letter spent discussing each supertopic.[53]

As a second step, I turn these continuous measures into binary variables indicating whether a letter was in the 80th percentile (top quintile) or higher in terms of the number of keywords about a given supertopic.[54] I then interpret these binary variables as indicating whether or not a letter contained *notably large levels of discussion* about a given supertopic (e.g., Character Excellence). This is again to facilitate salient interpretation of results. If I relied on the raw counts, my analysis would measure the differences in the *number of keywords* in a given letter across student demographics – but the inclusion of a single additional keyword about a given supertopic in a letter has an ambiguous conceptual impact in this context and may not actually be indicative of appreciably different letter content. I argue the more relevant margin is in differences across the *primary* or *defining* themes of a letter, those that a human reader would more readily detect as they read. This is especially informative to construct in my data context, as I can leverage my ability to examine the prevalence of supertopics across *all letters* in the dataset and see how letters compare. My topic modeling measures are then relative in nature, but relative to an enormous and meaningful reference dataset.

Two salient limitations of this method are that it is highly sensitive to idiosyncrasies in the data, and topic interpretations are ultimately quite subjective. Standard practice in the

---

[53] Mathematically speaking, the topic modeling algorithm will output the calculated proportion and count of keywords in each letter that come from each topic – derived in a probabilistic manner using Bayesian methods under the hood. Supertopics then represent a simple shorthand for signifying the proportion/count of keywords in each letter that come from constituent topic A, *OR* constituent topic B, *OR* constituent topic C, and so on.

[54] Robustness analyses show that my main results are not sensitive to the decision to use top quartile, top quintile, or top decile. Results of these checks are available upon request. I select the top quintile to balance between sufficient cell-sizes (i.e., enough students measuring both 1 and 0) and conceptual clarity (i.e., high enough at the end of the distribution to be meaningful).

field is to procedurally adjust the number of topics the algorithm identifies in the data to maximize conceptual cohesion or model fit in the resulting topic groups – only then does the analyst interpret the substantive meaning for each topic. This data-driven approach means that it is not possible to state a priori how many topics will be created, nor which of the identified topics will be most theoretically relevant to the focal concept of higher education admissions for this study. In Appendix A3.3, I discuss the specific steps I take to clean and prepare the text data for topic modeling, improve the robustness of my topic modeling approach, and evaluate the construct validity of the final supertopic measures.

Table 3.4 presents each of the supertopics I ultimately constructed through this topic modeling process, alongside up to two of their constituent subtopics and a subset of five of each subtopic's most prominent words, for illustration purposes. The full set of supertopics, their constituent subtopics, and the complete list of words comprising them, are available upon request. As can be seen in the table, the supertopics range from academic topics of discussion like a student's advanced coursetaking, to personality traits like a student's character excellence. Disparities in the prevalence of the former are likely to be of interest in cases where students may be considered academically marginal students for a given college program based on their quantitative application materials (e.g., standardized test scores, GPA, etc.). Disparities in the prevalence of the latter are likely to be of interest in cases where students are applying to highly selective institutions and more abstract character traits are considered pivotal in a student's application competitiveness (e.g., as revealed to often be the case in Harvard University admissions per *SFFA v. Harvard University*, 2019). Note also that I group "stock" recommendation letter language together into a supertopic I refer to as "Formalities" – phrases like "I highly recommend this student…," "Please don't hesitate to reach out with any questions," and "This student will contribute to any campus

community…" – to proxy for a teacher's compliance with the general form of letter writing. I take these phrases to be less informative of the students themselves, though they can arguably still overlap with other supertopic areas (e.g., part of how a teacher describes a student's academic excellence).

IVc. Sentiment Analysis

Through **(c) Sentiment Analysis**, I measure systematic differences in the occurrence of positivity, negativity, and lack thereof, in recommenders' writing (e.g., "He thrived in my class last year" v. "He struggled in my class last year" v. "He was in my class last year"). While sentiment analysis has been conducted using a wide variety of approaches over time, most modern algorithms are now built around the same overarching architecture (Vasawani et al., 2017). First, the algorithm "learns" basic linguistic relationships and structure by ingesting enormous quantities of text data and attempting to statistically derive the common mechanics and word relationships for a language. After establishing this language schema and generating a series of context-dependent "definitions" for each word within its vocabulary,[55] the algorithm is then "fine-tuned" to interpret entire sentences for their perceived positivity, negativity, or neutrality using databases of human-generated crosswalks as its point of reference.[56] These modern approaches to the task allow the

---

[55] The newest methods on this front, "transformer" neural networks (Vaswani et al., 2017), are unique in their ability to incorporate context before *and* after each word it examines, thus allowing it to change its "understanding" of a given word based on the exact sentence in which it appears. This allows the transformer family of algorithms to understand that the word "bank" in "I went to the bank to cash a check" is different in meaning than in "I sat by the bank and watched the water flow by."

[56] The Stanford Sentiment Treebank ("SST"; Socher et al., 2013) is the most common crosswalk for such a purpose. In brief, they "crowd-sourced" human-generated ratings of word, phrase, and sentence positivity/negativity on a five-point scale using Amazon Mechanical Turk. Unfortunately, their definition for positivity and negativity was intentionally vague and left up to scorers' interpretation, and I was unable to find inter-rater reliability metrics for the source data.

algorithm to better interpret more complex linguistic features like negation (e.g., "not bad"), multiple word meanings (e.g., "I *ran* to the mall" versus "I *ran* the code") and the subjunctive tense (e.g., "I wish this were good").

Per Kim et al. (2021), I operationalize the definition of sentiment as the *perceived positivity of emotions and ideas present in a given text*. This definition then is a conglomeration of the writer's stated emotions ("Her lack of effort saddened me..." versus "I was excited to hear..."), communicated intention ("I wish her the best!" v. "I wish she'd try harder"), and, at least to some extent, topical content ("financial hardship" v. "class valedictorian"). Table 3.5 displays a sample of real sentences from the recommendation letter data, alongside their algorithmically-assigned sentiment score, to illustrate how this construct works in practice – though note importantly that I have hand-selected these examples for illustration of the construct only, and this should not be taken as an illustration of algorithm accuracy.

Ultimately, I rely on the sentiment analysis algorithm trained by Barbieri et al. (2020) to measure each sentence of each letter for its positivity on a three-point scale (negative, neutral/factual, and positive), which I then collapse to the total count of sentences at each level of positivity in each letter (e.g., total count of positive sentences). As an additional measure, I also calculate the *average* sentence sentiment across the letter. This specific implementation was initially trained to examine Twitter "tweets" for their sentiment using the "RoBERTa" architecture created by Liu et al. (2019). I selected this particular implementation after directly comparing several models' performance at matching the judgment of a team of human analysts on a random sample of 480 sentences from the teacher recommendation letters (stratified by student race and gender). I find that this algorithm was categorically better than all others tested, matching human judgment about 77% of the time with no appreciable difference in accuracy by gender or student race.

Importantly, human coders matched with one another only 84% of the time, indicating that 100% accuracy is not a feasible benchmark of accuracy due to expected differences in human opinion when faced with ambiguity in a given sentence's sentiment; with this in mind, the algorithm performs *exceptionally* well.

While my approach leverages methods at the state-of-the-art to conduct the most nuanced and capable version of sentiment analysis possible, this strategy trades off in transparency and the potential for algorithmic bias (see Shah et al., 2020, for a helpful conceptual review). Thus, in Appendix A3.4, I describe the process by which I selected the sentiment analysis algorithm from several candidate options, details about each candidate option, and how I assess the potential prevalence of algorithmic bias in sentiment output.

IVd. Regression Analysis

For the second phase of analysis, my conceptual goal is to compare the letter characteristics of two students whose only substantive difference is their race or gender to reveal any systematic differences in letters across these groups. That in mind, I use my NLP-derived measures as the outcomes for a series of regressions to examine whether the measures – and the writing of the letters by proxy – vary systematically by student race and gender, even after controlling for additional student characteristics and qualifications that would likely influence letter characteristics and content and, in turn, college admissions decisions. I examine relationships in three separate models. The first model looks broadly across all submitted letters together (which I herein refer to as the "Landscape" model), revealing the extent of systematic differences in these letters in general. The second model employs teacher fixed effects to look across only those letters written by the *same teacher* (which I herein refer to as the "Teacher Fixed Effects" model), revealing the extent of

systematic differences in these letters stemming from individual teachers' writing patterns and habits. The third model employs institution fixed effects to look across letters received by the *same institution* (which I herein refer to as the "Institution Fixed Effects" model), revealing the extent of systematic differences in these letters in the immediate pool of applicants a given institution might be selecting from.

For the student-level controls included in all of my models, I leverage the rich covariates provided by students in their applications as displayed in Table 3.6. In brief, I include intuitive covariates on their demographics (including non-focal demographics like age and class year), academics and test-taking, extracurricular characteristics, and student application behaviors in the same spirit as the "college application profile" controls employed by Dale and Krueger (2002; 2011) shown to be strong proxies for student academic ability.[57] The intention here is to include as many covariates as possible likely to influence the characteristics and writing of the recommendation letters. For example, we might expect students more heavily involved in service activities to then have letters where the student's community service is discussed at relatively greater length. If it's the case that female students participate in service activities more often than male students, a naive regression of service-related letter content on gender would be conflated by the underlying demographics of who participates in these service activities, rather than revealing a "truer" signal of systematic differences in the letter writing processes per se.

---

[57] I opt to specify continuous covariates as binned factor variables when possible, as there exist many potential nonlinearities in the relationship between these covariates and letter characteristics. For example, we likely wouldn't expect standardized test math percentile to trend linearly with letter characteristics given the important conceptual difference between moving a student from the 95th to 99th percentile, versus from the 55th to the 59th percentile. When possible, I bin students into roughly even-sized groups while maintaining conceptually useful breakpoints (e.g., number of SAT/ACT tests taken, in which students who take only one standardized test are distinct from students who likely retake, or students who retake several times).

For the first model ("Landscape"), I am interested in understanding generally whether there exist systematic differences in letter characteristics across the full universe of letters being sent to institutions, after controlling for the rich set of covariates I describe above. This can be thought of as trying to examine whether systematic differences exist in the broader system of teacher recommendation letters writ large. Equation 1 is my formal regression specification for this model:

$$(1) \quad Y_{it} = \tau_i + X_i + L_{it} + E_t + C_{it} + \beta_1 R_i + \beta_2 F_i + \varepsilon_{it}$$

Y_{it} represents any one of the NLP-derived letter characteristics as described previously (e.g., whether a letter has notably large levels of discussion about Character Excellence) from the recommendation letter for student *i* by teacher *t*. $\tau_i$is an indicator for student cohort-year, X_i represents the vector of student covariates as specified in Table 3.6, in addition to the student's school sector (public/private), L_{it} represents an additional set of student-teacher covariates as described in Table 3.3,[58] E_t represents two teacher-level categorical variables as described in Table 3.2,[59] C_{it}represents the total count of sentences (for sentiment analysis measures) or keywords (for topic modeling measures) in each letter, and ε_{it}represents the idiosyncratic error term.

My coefficients of interest are betas 1 and 2 on R_i and F_i: indicators for each race category (White as reference baseline) and female-identifying students. I interpret these coefficients as the observed difference in letter characteristic *Y* (e.g., whether a letter has

[58] To reiterate for convenience, these covariates are: how many years the teacher has known the student, the teacher's subject area with the student, whether the teacher was a coach for the student, and whether the teacher indicated some other teaching relationship with the student.

[59] To reiterate for convenience, these covariates are: number of letters written in the present year and number of letters written in the past two years.

notably large levels of discussion about advanced coursetaking) on average for each student group after controlling for other student characteristics to the greatest extent possible.

$$(2) \quad Y_{it} = \lambda_t + \tau_i + X_i + L_{it} + C_{it} + \beta_1 R_i + \beta_2 F_i + \varepsilon_{it}$$

For the second model ("Teacher Fixed Effects"), I add teacher fixed effects in Equation 2 to understand whether there exist systematic differences in letter characteristics (after controls) among letters *written by the same teacher*. Applying teacher fixed effects allows me to compare letter characteristics among the group of students for whom a single teacher has written recommendations, thus controlling for any teacher characteristics that are fixed within teachers (e.g., teacher subject area when fixed, school characteristics/culture, writing ability, etc.). This can be thought of as assessing the extent to which systematic differences manifest via teachers' individual writing styles. Equation 2 is thus identical to Equation 1, except with λ_t included to represent the vector of teacher fixed effects and the teacher-level covariates E_t removed since they would be fixed within teachers. Importantly, this fixed effects specification relies on a region of common support within each variable of interest. For example, the coefficient on F_i is estimated only using the set of teachers who have written for both male *and* female students in the data, which fundamentally changes the external validity of these coefficients (i.e., if teachers who write for male and female students are importantly different from teachers who write only for male or only for female students). I discuss the repercussions of this concern in more detail in Appendix A3.5 and examine explicitly the sample sizes for each demographic variable's region of common support. In short, while the teachers included in the effective estimation sample for each variable is reduced (e.g., only 15% of teachers are included in the estimation sample for Black students), the number of letters included in each estimation sample remain very large (e.g., 35.9% of all letters remain in the estimation sample for Black students, or roughly 913,000 letters).

$$(3) \quad Y_{itu} = \kappa_u + \tau_i + X_i + L_{it} + E_t + C_{it} + \beta_1 R_i + \beta_2 F_i + \varepsilon_{itu}$$

Lastly, I leverage institution fixed effects for the third model ("Institution Fixed Effects") in Equation 3 to understand whether there exist systematic differences in letter characteristics (after controls) among letters *received by a given institution*. Applying institution fixed effects allows me to compare letter characteristics among the letters received by a given institution while controlling for any institutional characteristics fixed within institutions (e.g., institutional selectivity, size, region, sector, specialty, etc.). This can be thought of as trying to examine whether systematic differences exist in the letters for the applicant pool of a given institution, which may be a more relevant estimand when considering whether individual institutions may need to be concerned about disparities in letters they, specifically, review in admissions decisions.[60] Equation 3 is identical to Equation 1, except with κ_u included to represent the vector of institution fixed effects (note that teacher fixed effects are not included here). Note here also that the unit of observation changes to the *application* level, such that an observation is identified by a student, the teacher who wrote for them, and the university/college *u* they applied to. Thus, the extent to which an individual student contributes to the coefficients on each demographic variable of interest is mechanically weighted by the number of institutions that student applied to (e.g., a female student who applies to only one institution will have less weight than a female student who applies to eight). While this institution fixed effects approach is also susceptible to issues of common

[60] To expound: the landscape results may not be relevant to an individual institution if they have reason to believe the overall pool of applicants is meaningfully different in composition from their specific pool of applications. The teacher fixed effects results are only relevant to individual institutions if they receive sufficient volume from a *single teacher* to norm recommendations within that teacher. This is almost certainly not the case, and admissions departments tend not to have the statistical capacity nor airtime to conduct such analyses given the pace of application seasons otherwise. Thus, systematic differences that exist within their individual pool are most realistically within their actual capacity for adjustment or consideration.

support like the teacher fixed effects approach, the region of common support is far more forgiving here (e.g., institutions that receive at least one application from both a male and female student), and greater than 99% of all letters are included in the estimation sample of all demographic variables. Regardless, I also report on the region of common support for each demographic variable of interest in Appendix A3.5 for completeness.

In all three specifications, I cluster standard errors at the teacher level to account for the likelihood that letter characteristics in the group of letters written by a single teacher are correlated, and the likelihood that student characteristics in the group of students a teacher writes letters for are also correlated (either due to homogeneity in their student population, or due to homogeneity in their willingness to write letters for certain types of students).

IVe. Subsample Analyses

My regression models as noted above will examine the prevalence of systematic differences in the letters across the whole sample, but it may be the case that there is meaningful heterogeneity across student, teacher, or institution characteristics. For example, it may be the case that STEM teachers specifically write about female students more negatively than they write about male students. As is, my regression models would not be able to account for such dynamics without adding specific interaction terms. But because including interactions in this way can quickly become intractable, and because they do not allow for flexibility in the estimation of coefficients *besides* those being interacted (e.g., STEM teachers also writing differently about students with high GPAs would not be captured by a single interaction term between STEM teacher and female student covariates), I use subsample analyses to examine whether the degree of systematic differences vary meaningfully across student, school, teacher, and institution characteristics of particular

theoretical interest. The results of these regressions then estimate the prevalence of systematic differences among that particular subset of letters.

That said, there are an enormous number of subsample analyses that I could potentially examine; for the purposes of this manuscript, I focus on the issues of racial bias in highly selective college admissions and gender bias in STEM subjects given their current salience in contemporary college admissions.

To first shed light on potential biases in highly selective college admissions, I lean on the institution fixed effects model. Using selectivity data from each institution (median SAT/ACT score of admitted students), I can restrict my sample to only those institutions in the top decile of selectivity (within my sample of institutions receiving letters via the Common App that report selectivity data). This results in 4.1 million applications from 700,000 students to 59 institutions with a median SAT score of 1300 and above among admitted students.

In an additional cut of the data, I further restrict the sample of interest to only those students in the top quintile of application profile selectivity (i.e., applied on average to the most selective institutions by median SAT/ACT score of admitted students) and in at least the 95th percentile for both their verbal and math SAT/ACT scores (among the broader distribution of SAT/ACT test takers). This then filters out those students who were applying to these highly selective institutions as a "far reach" and thus focuses on those students most likely to be competitive for admissions. The remaining subsample then includes about 1.3 million applications and 108,000 students. These specifications are likely to be of immense interest to ongoing litigation related to affirmative action in cases like SFFA v. Harvard (2019). For example, if we observe that Asian students are systematically written about more negatively among this population, it would more empirically substantiate the basis for the

systematically lower "Personal" scores that Asian students received at Harvard. Given that the lower personal scores were then used to, at least partially, justify their lower admissions probability, examining their empirical basis is of strong interest to the case.

To examine potential gender biases among teachers of STEM subjects, I rely on the teacher fixed effects model. Because I observe the subject a given teacher taught with a given student, I can restrict my sample of analysis to only those student-teacher relationships within STEM subjects: mathematics, science, and computer science. This thus removes English, social studies, world languages, and any miscellaneous categories. The resulting sample includes approximately 194,000 teachers, 966,000 letters, and 820,000 students. Of particular interest here are any systematic differences that may detriment female students (given prior literature) and benefit Asian students (given prevalent societal stereotypes) in the context of STEM subjects, specifically, even after accounting for teacher unobservables (due to the teacher fixed effects) and relevant student and student-teacher covariates.

IVf. Limitations

While these data, alongside recent advances in NLP methods, offer me the opportunity to examine this research question with uncommon scope and comprehensiveness, there are still a number of limitations and considerations that should be acknowledged.

First, my results are descriptive in nature. Despite my ability to include many compelling covariates as controls, there will almost always be additional confounders in observational data without a strong quasi-experimental or experimental design. Moreover, we know that several critical features of the data generating process in this context make a true apples-to-apples comparison across student-teacher pairs intractable, as we are

fundamentally trying to control for multi-dimensional, longitudinal relationships between students and their teachers. In other words, the proxies I use to describe students, teachers, and their relationship in these data may inadvertently mask or misestimate relevant dynamics that I assume I have captured.

Second, my analysis focuses entirely on identifying systematic differences as manifested within the content of these teacher recommendation letters. There are importantly many inequities and biases at other points in the educational process that contribute to differences in the control variables I utilize (e.g., GPA, extracurriculars, SAT/ACT scores) and selection into this sample of college-aspiring students. My intention is not to ignore these important disparities, but to present the best estimates I can on this individual margin as my contribution to a broader discourse on gender and race in our society.

Third, my exact sample for estimating differences in letter writing is a large but still *specific* group of students. That is, I am estimating the systematic disparities in letter content across student demographics among the group of students who meet all of the following criteria:

1. Applied to postsecondary institutions in the U.S. using the Common Application[61]
2. Applied to institutions that accepted teacher recommendations[62]

---

[61] Knight & Schiff (2019) find that institutions accepting the Common App tend to be more selective.

[62] While almost all institutions accepted teacher recommendations through the Common App, my initial descriptive analyses show that those *requiring* teacher recommendations are more selective. I argue these more selective institutions remain broadly relevant given concerted efforts across stakeholder groups to improve college access to such selective institutions, especially among low-income and minority students (Hoxby & Turner, 2014; Page & Scott-Clayton, 2016).

3. Successfully obtained at least one teacher recommendation[63]

4. Successfully submitted a completed application to at least one institution

I intend to explore the repercussions of **2** and **3** using these data in separate work, but until then, we should be cognizant about the potentially idiosyncratic sample of student-teacher pairs here. Arguably, my data still represent an exceptionally relevant population of students well, and it remains the best available source of data to analyze in these pursuits.

Fourth, my reliance on algorithmic NLP techniques means I am unlikely to reveal all forms of systematic differences potentially present in the letters. While researchers have made great strides incorporating word context and sentence composition in topic modeling and sentiment analysis algorithms, there will always be exceptions and fringe circumstances in language that cannot be accounted for with these approaches.[64]

Fifth, this analysis assumes that the self-reported race and gender of students are salient and perceptible to teachers. It is likely that many students who self-identify in particular ways may be misperceived by their teachers, resulting in a sort of measurement error in my demographic variables (e.g., student reports being Black, but teacher actually perceives white). This has repercussions not just for attenuating my estimates, but also for my interpretation – I am technically estimating the dynamics of teacher *perceptions of* student demographics, not *actual* student demographics.

---

[63] This criteria actually stands in for two important dynamics. The teacher recommendation process is logistically complex, requiring substantial planning and coordination on behalf of the students. Further, this recommendation process requires a two-sided match: students must select a teacher, and the teacher must accept. The students who can then successfully obtain even one letter are likely to be meaningfully different in many ways from students who cannot.

[64] For example, none of my approaches would be able to *reliably* capture something subtle like group attribution – a teacher diffusing a student's individual successes across a broader group. E.g., "Brian excelled in my math class" versus "The groups Brian worked with excelled in my math class."

Lastly, I again cannot comment on the extent to which any measured disparities in letter language would contribute to disparities in actual admissions outcomes for specific students or at specific institutions. For now, this is a descriptive exploration that may motivate such research by practitioners and field experts in the future.

V. Results

Va. Topic Modeling Main Results

Table 3.7 displays the coefficients for each student demographic variable of interest using Equation 1 ("landscape"), where each column is a separate regression. The outcome of each regression is an indicator for having notably large levels of discussion about each supertopic (e.g., Academic Excellence in the first column). Recall that this indicator is set to 1 if a letter is in the 80th percentile for the number of keywords in their letter about that supertopic, meaning all coefficients are effectively reported as *percentage point* differences; in addition, the sample mean is then mechanically 20% for all of these indicators, facilitating quick calculation of any *percentage* differences as well. The first row of each column also displays the 80th percentile threshold in terms of the number of keywords required to cross that threshold. For example, a letter has notably large levels of discussion about Academic Excellence if it contains 18.04 or more keywords about Academic Excellence.[65] By contrast, a letter has notably large levels of discussion about Extracurriculars if it contains 3.96 or

[65] Because keyword counts are actually counted *probabilistically* across many Bayesian simulations in the structural topic modeling method, it is actually expected that a single letter has *fractional* keywords in each supertopic. This is because a single keyword can actually belong to several supertopics (e.g., "strong" can be in reference to Academic Excellence as "strong student" or in reference to Character Excellence as "strong character"), and so the topic model actually returns a *probability* that each word belongs to each supertopic. Collapsing the expected probabilities for each word within a given letter then produces the rough number of keywords in each supertopic, which is unlikely to be a whole number.

more keywords about Extracurriculars. Note then that these indicators are not necessarily mutually exclusive for a given letter. Finally, these regressions all include: student-level covariates, teacher letter-writing experience covariates, student-teacher relationship covariates, no fixed effects, and clustering at the teacher-level.

To begin interpretation of these results, the second row of each column displays the coefficients on the indicator for being a female student. Looking at the first column, female students are 1.1 percentage points (pp) *less* likely to have notably large levels of discussion about Academic Excellence topics. On a sample mean of 20%, this also means that female students are roughly 5% less likely than male students to have letters with notably large levels of discussion about this supertopic. Interestingly, we see that they are also 1.0pp (5%) less likely to have notable discussion about Intellectual Promise, and 0.9pp (4.5%) less likely to have notable discussion about STEM courses. These findings are all roughly consistent with the concern that female students' academic characteristics may be underplayed by teachers when compared to male students in these letters, especially in STEM.

Interestingly, we see other common cultural narratives at play here as well. Female students are 2.6pp (13%) *more* likely to have notable discussion about Community Engagement, aligning with stereotypes of females being more service- and community-oriented. Female students also are 2.4pp (12%) more likely to have notable discussion about Extracurriculars, which is consistent with being 1.2pp (6%) more likely to have notable discussion about Time Management; greater engagement in extracurriculars would likely result in building the skills necessary to balance these commitments. Finally, female students are 2.8pp (14%) less likely to have notable discussion about Sports, even with indicators in place for student sports involvement and the recommending teacher being a coach – perhaps owing to cultural biases downplaying the value or prominence of female sports

involvement as well. That the aforementioned differences all persist in relatively large magnitudes despite the presence of relatively strong controls suggest letters about female students tend to look quite different from letters about male students, on average.

Differences by race tend to be smaller by comparison – for example, there are three coefficients above 2.0pp for female students, but not a single one larger than 1.6pp for Black or Asian students. For Black students, their letters are 0.5pp (2.5%) *more* likely to have notable discussion about Academic Excellence than White students, 0.9pp more likely to have notable discussion about Advanced Coursetaking, and 1.0pp more likely to have notable discussion about STEM. These results are not particularly large, but still somewhat unintuitive given common concerns about Black students being culturally mischaracterized as "less academic" than White students. However, because these regressions control for advanced coursetaking (via APs and IB) and academic performance, it could be the case that teachers feel more compelled to write about Black students' academics than White students holding constant the same academic profile (e.g., because it may be more "noteworthy" to them that a Black student is performing well). Looking across the more character-based supertopics, there is a less clear through-line to interpret: Black students tend to have letters with greater discussion about Community Engagement (0.7pp) and Leadership (0.9pp), but less discussion about Sports (-1.4pp) and Time and Life Management (-1.2pp). Perhaps teachers are more likely to situate Black students within their community context (i.e., as leaders) than their individual context (i.e., discussing time management).

Turning lastly to Asian students, there is a generally strong narrative of academic supertopics in their letters. For example, Asian students are 0.7pp more likely than white students to have notable discussion of Advanced Coursetaking, 1.4pp (7%) more likely to have notable discussion of STEM topics, and 1.6pp less likely to have notable discussion of

Sports. At least at this level of analysis, Asian students are depicted with greater "campus contribution" qualifications in their letters: they are 0.9pp more likely to have notable discussion of Extracurriculars, 1.1pp more likely to have notable discussion about Community Engagement, and 0.6pp more likely to have notable discussion about their Future Potential. While there are significant differences for Character Excellence (-0.3pp), Humanities (-0.3pp), and Intellectual Promise (-0.3pp), these do not seem to be substantively meaningful differences given their magnitudes.

Table 3.8 is structured in the same way as Table 3.7, but using Equation 2 as the regression specification. To summarize, this model includes student-level covariates, student-teacher covariates, teacher fixed effects, but no teacher experience controls (because they are invariant within teachers, they are subsumed by the teacher fixed effects). I thus interpret these coefficients as the average percentage point difference in the likelihood that a letter contains notably large levels of discussion about a given supertopic, among letters written *by the same teacher*. As an example to set the intuition for the model, coefficients for letters having notable discussion about Formalities are substantially smaller across the board (mostly absolute values of 0.3pp and below) than in the Landscape regressions. This makes sense given that we wouldn't expect a teacher to vary drastically in their use of formal letter language (e.g., "I write to you on behalf of…") across students, especially if they tend to use template language as a skeleton for the rest of the letters.

Looking first at the results for female students, the coefficients are almost identical to those we observed in the Landscape model. That is, female students are still 1.1pp less likely to have notable discussion about Academic Excellence (versus -1.1pp in the Landscape regression), 1.1pp less likely to have notable discussion about Intellectual Promise (versus -1.0pp), 0.7pp less likely to have notable discussion about STEM (versus -0.9pp), and 2.6pp

less likely to have notable discussion about Sports (versus -2.8pp). They are, conversely, 2.3pp more likely to have notable discussion about Community Engagement (versus 2.6pp), 2.0pp more likely to have notable discussion about Extracurriculars (versus 2.4pp), 0.7pp more likely to have notable discussion about Leadership (versus 0.6pp), and 1.3pp more likely to have notable discussion about Time and Life Management (versus 1.2pp). In other words, the differences across student gender surfaced in the Landscape model appear to be driven, at least in part, by how individual teachers tend to write about male versus female students. Or at least, it doesn't seem to be the case that the differences are driven by female and male students systematically asking different teachers who have different writing styles or focuses.

Turning now to letters for Black students, all differences have shrunk considerably versus the Landscape model, with no coefficient above 0.8pp. Interestingly, the coefficients for several academic supertopics have either been reduced substantially or flipped entirely: from 0.5pp to -0.4pp for Academic Excellence, from 0.9pp to -0.5pp for Advanced Coursetaking, and from 1.0pp to 0.1pp for STEM. These negative coefficients are relatively small in magnitude, but the large differences from the Landscape model suggests that sorting may be more of a factor for Black students than female students. That is, Black students may systematically request letters from teachers who are more likely to write about these academic performance topics more often, explaining the positive coefficients in the Landscape model versus the teacher fixed effects model. It could also be that the teachers who write letters for both Black *and* White students are less likely to write about Black students in terms of their academic performance than teachers who only write letters for Black students (who would then be excluded from the effective estimation sample on the coefficient for Black students). The remainder of coefficients that were noteworthy in the

Landscape model remain generally the same here: Black students were 0.7pp more likely to have notable discussion about Community Engagement (versus 0.7pp), 0.5pp more likely to have notable discussion about Leadership (versus 0.9pp), 0.7pp less likely to have notable discussion about Sports (versus -1.4pp), and 0.8pp less likely to have notable discussion about Time and Life Management (versus -1.2pp).

The estimated coefficients in this model look almost entirely the same for Asian students across the board when compared to the Landscape model, though some are slightly reduced in magnitude. As discussed for our results on the coefficients for female students, this again suggests that differential sorting between students and teachers seems unlikely to be the reason for the observed differences in letter content between Asian and White students. Unlike the results on the coefficients for Black students, it also doesn't seem to be the case that the teachers included in the common support in this model are substantially different than the teachers included in the Landscape model; in other words, teachers who only write letters for Asian students are not substantially different in writing style or content from teachers who write letters for both White and Asian students.

Table 3.9 displays the results of Equation 3, the institution fixed effects model. This model is almost identical to the Landscape model (student-level covariates, teacher writing experience covariates, and student-teacher relationship covariates), but now includes institution fixed effects and is structured at the *application* level rather than the letter level. This latter nuance should be kept in mind when interpreting coefficients, in that students who applied to more institutions within a given demographic group are contributing to the coefficients for that demographic variable with greater weight. Unlike the Landscape model, which looked at differences in the full set of letters being sent to higher education

institutions together, this model looks more closely at the differences in letter content among the set of letters the average institution receives in their own applicant pool.

With the exception of only a few mild differences, coefficients across the indicators for female students, Black students, and Asian students in this model remain consistent with those estimated in the Landscape model. Put another way: it doesn't seem to be the case that the pools of letters at individual institutions contain larger or smaller differences in topical content across student demographics than the pool of letters in general.

To summarize the results of the topic modeling analysis, letters about female students are more likely to have notable discussion about Community Engagement, Extracurriculars, and Time Management, but less likely to have notable discussion about Academic Excellence, Sports, Intellectual Promise, and STEM – regardless of the specification. Trends for Asian students are similar in that they are also more likely to have notable discussion about Community Engagement, Extracurriculars, Future Potential, and STEM, less likely to have notable discussion about Intellectual Promise, and not meaningfully less likely to have notable discussion about Academic Excellence, again regardless of specification. That these two groups see fairly consistent estimates from the Landscape to the teacher fixed effects model suggests that the differences we observe may be driven primarily by how individual teachers differentially write about male versus female students, or Asian versus White students. While there were some meaningful differences for Black students in the Landscape model when it comes to notable discussion of Academic Excellence, Advanced Coursetaking, STEM, Community Engagement, and Leadership, these differences *do not* seem entirely driven by how individual teachers write about White versus Black students, given how much smaller the coefficients were in the teacher fixed effects model.

Vb. Sentiment Analysis Main Results

Moving now to the sentiment analysis results, Table 3.10 is structured identically to Tables 3.7-3.9 but with the sentiment analysis measures set as the outcomes in each column. The outcomes of interest here are then the number of sentences in each letter classified by the algorithm as each level of positivity ("Positive," "Neutral," and "Negative") and the average sentiment across all sentences in a given letter ("Mean Sentiment"). Coefficients are reported in native units for each outcome (e.g., differences in the average number of positive sentences between letters for male and female students). The first row of each column also displays the sample mean for that outcome; for example, there are 13.81 positive sentences in each letter, on average. Perhaps as expected, these letters are generally positive in tone, with very few negative sentences in general (given a sample average of 0.48). Like Table 3.7, Table 3.10 uses the Landscape specification, which includes no fixed effects.

Turning first to the coefficients for female students, results indicate that female students have 0.192 more positive sentences in their letters than male students on average, even after accounting for the many salient controls, a roughly 1.4% difference given a sample mean of 13.81 positive sentences. Because sentiment classifications for a given sentence are mutually exclusive, we should mechanically expect female students to then have fewer neutral and negative sentences; this is indeed the case, as we see that female students tend to have 0.174 fewer neutral sentences (-3.2%) and 0.018 fewer negative sentences (-3.8%). This latter difference is larger in relative magnitude and may be the more impactful one in terms of how admissions readers perceive and experience the letters given psychological dynamics like negativity dominance (e.g., presence of negativity is more influential than the absence of positivity), though this remains to be studied empirically in

this context. Finally, the average sentence about female students is 0.009 points more positive than the average sentence about male students, a difference of about 1.3% given the sample mean of 0.7.

For Black students, we see that they have 0.116 (0.8%) fewer positive sentences than White students on average, and instead have 0.110 (2%) more neutral sentences and 0.006 (1.3%) more negative sentences. Put another way, Black students see fewer positive sentences in their letters and slightly more negative letters. This stands in contrast to results for Asian students, who have 0.045 fewer positive sentences on average when compared to White students, but do not have more sentences of any negativity (0.002) on average. This suggests that both Black and Asian students are slightly less likely than White students to have positive statements about them in their letters, but where Asian students just see more *neutral* statements replacing them, Black students actually see some *negativity* replacing them.

Table 3.11 displays sentiment analysis regression results using the teacher fixed effects model (Equation 2), again estimating the average difference in letter positivity between students of varying demographics among letters written by the same teacher. The coefficients for female students on each sentiment analysis outcome are generally the same as in the Landscape model, but only slightly reduced in magnitude across the board. As was the case in the topic modeling results, this seems to indicate that the differences in letter positivity we observed in the Landscape model were not primarily the result of female students sorting to different teachers with different writing styles (e.g., teachers that are simply more or less positive in their writing in general), but rather that these differences seem at least partly driven by how individual teachers write about male versus female students.

The trend for female students is then in stark contrast to the trend we observe for Asian students, where all the differences we observed in the Landscape model are actually *more* pronounced in this teacher fixed effects approach. That is, Asian students now have 0.073 fewer positive sentences (and commensurately more neutral sentences) versus having only 0.045 fewer in the Landscape model. This seems to indicate that the differences in letter positivity we observe for Asian students are *primarily* happening at the individual teacher level, rather than the result of sorting to different teachers. Put another way, Asian students seem to receive less positive letters than White students do from *the same teacher,* even conditional on having the same observable characteristics. Importantly, the magnitude of difference in positivity here is quite small (0.5%), and it is unclear to what extent this difference would be detectable to any human readers of the letters.

Lastly, the dynamics we observed in the Landscape model for Black students are substantially reduced when moving to the teacher fixed effects model. For example, Black students now have only 0.018 fewer positive sentences than White students (0.1% difference), down from 0.116 fewer in the Landscape model. This seems to indicate that if Black and White students with equal observable qualifications were to ask the same teachers for letters, we wouldn't expect to see meaningful differences in the degree of positivity for each student. The fact that we observed larger differences in the degree of positivity in the Landscape specification indicates that there is indeed some sorting that happens, whereby Black students are asking for letters from teachers who tend to write more negative letters in general than White students do.

Finally, Table 3.12 displays the results of our sentiment analysis regressions using the institution fixed effects approach, again looking for differences in letter positivity within the pool of applications sent to individual institutions. In general, we see that the coefficients

remain generally unchanged for female, Black, and Asian students versus the Landscape model. The only notable difference is that the difference between Black and White students in the number of negative sentences loses significance, but more due to decreasing precision than a change in the estimate itself. This consistency in estimates from the Landscape to the institution fixed effects model implies that student sorting to different institutions neither exacerbates nor ameliorates the detectable differences in letter positivity across groups.

To summarize the results of these sentiment analysis regressions, we generally see that female students tend to have more positive sentences in their letters than male students, regardless of the specification. In other words, whether looking across the whole body of letters, letters written by the same teacher, or letters received by a given institution, letters about female students are more positive than male students when holding constant as much as we can observe about their qualifications. While we detect reduced positivity for Black and Asian students versus White students, these differences are comparatively quite small in magnitude. For Black students, the differences seem driven by differential sorting to teachers, given that within-teacher estimates shrink towards zero. For Asian students, it is the opposite: the differences instead seem driven by writing behaviors of teachers themselves, given that the within-teacher estimates were appreciably larger. This seems to go against speculation that Asian students come across as less exceptional in their narrative application materials among students with the same academic qualifications (e.g., as discussed in SFFA v. Harvard, 2019). Indeed, the magnitude of differences are small enough that they seem unlikely to drive any major application disparities, and Black students would appear more disadvantaged than Asian students in this regard (if detectable by humans at all).

Vc. Subsample Analysis Results: Highly Selective Admissions

Tables 3.13 and 3.14 display the results of the first subsample analysis examining only those applications to the top decile of institutions in terms of their selectivity using the institution fixed effects model (Equation 3). As mentioned before, the coefficients here can be interpreted as systematic differences in the letter content within the pool of applications received by these highly selective institutions, holding fixed as much about the student, teacher, and student-teacher relationship characteristics as possible. Beginning with topic modeling results in Table 3.13, nearly all of the coefficients for female across supertopic indicators remain the same in direction and magnitude when compared with the full sample institution fixed effects results. The only difference here appears to be that female students are 0.6pp (3%) less likely to have notable discussion about Advanced Coursetaking, where in the full sample the difference was nearly zero, and are now slightly more likely to have notable discussion about Humanities (0.6pp from 0.2pp). These differences seem relatively immaterial in scale, suggesting that the trends in the whole sample remain broadly true among highly selective institutions as well.

Black students tend to have greater discussion of many campus contribution supertopics: they are more likely to have notable discussion about Community Engagement (1.2pp or 6%), Extracurriculars (0.7pp or 3.5%), Future Potential (0.7pp or 3.5%), and Leadership (1.6pp or 8%) than White students. Conversely, we see that they tend to have less discussion about some of the academic supertopics like Academic Excellence (-0.4pp), Humanities (-1.4pp), and Intellectual Promise (-1.0pp). In other words, among this sample of highly selective institutions, letters about Black students tend to have less discussion about academic topics and more discussion about campus contribution topics relative to White students.

Asian students are even more likely than Black students to have notable discussion about Community Engagement (1.8pp or 9%), Extracurriculars (1.6pp or 8%), Future Potential (1.3pp or 6.5%), and Time and Life Management (-0.4pp or 2%). Magnitudes for the more academic supertopic coefficients look roughly equivalent with those for Black students. This result seems to suggest then that in terms of topical content, Asian students have more discussion about campus contribution topics than both White and Black students – a surprising result given that analyses in SFFA v. Harvard (2019) suggested that the personal rating is where Asian American applicants were *weakest* relative to other students. It may be that Asian students see greater levels of discussion about these topics, but that such discussion is more negative; however, the next set of results seems to reject that hypothesis as well.

Table 3.14 displays the results of this subsample analysis for the sentiment analysis outcomes. Looking first at results for Asian students, there are no notable differences across the board in terms of the positivity or negativity of sentences in their letters versus White students. This suggests that, at least looking broadly among the applications to highly selective institutions similar to Harvard, I don't detect any differences in the recommendation letters themselves that would explain the lower personal ratings among Asian American applicants. Black students have letters with slightly fewer positive sentences (-0.095 or -0.6%), slightly more neutral sentences (0.107 or 1.7%), and slightly fewer negative sentences (-0.012 or -2%). These results are very small in magnitude and suggest, if anything, Black students have letters that are *slightly less* favorable than both Asian and White students among this highly selective institution subset. As with results in the full sample, female students tend to have meaningfully more positive letters than male students.

Drilling down to those most competitive applicants to these highly selective institutions in Tables 3.15 and 3.16, many of the same trends surfaced in the prior tables are magnified in the case of Black students, while trends for female and Asian students remain largely the same. For Black students, we see greater discussion of Community Engagement (2.8pp from 1.2pp when not restricting to competitive applicants to these highly selective institutions), Extracurriculars (3.3pp from 0.7pp), and Leadership (2.4pp from 1.6pp), and even less discussion about Intellectual Promise (-2.7pp from -1.0pp) and Humanities (-2.6pp from -1.4pp). In other words, Black students' letters are even more strongly characterized by these campus contribution topics among this subsample of applicants.

Table 3.16 finally displays sentiment analysis results for competitive applicants to these highly selective institutions. As before, letters for female students are generally more positive than letters for male students. Results for Black students are nearly the same as in the prior results, but lose significance largely due to lack of precision. Interestingly, letters for Asian students among this competitive applicant pool are actually slightly more positive than letters for White students, further emphasizing that teacher recommendation letters may actually be a strength for Asian students in terms of revealing campus contributions, relative to White and Black students.

To summarize the findings of this section, differences between letters for female and male students are largely consistent with what was surfaced in the full sample. Interestingly, both Black and Asian students may be described with slightly more discussion of campus contribution supertopics than academic ones in these letters when compared to White students. Indeed, I show here that Asian students may actually have *more favorable* discussion in their letters than White students, indicating that teacher recommendation letters don't seem to explain the systematically lower personal ratings they receive from Harvard. That

said, it may still be the case that those lower ratings can instead be explained by other narrative elements (e.g., counselor letters, essays, interviews).

Vd. Subsample Analysis Results: STEM Teachers

Turning now to examine the prevalence of systematic differences among only those letters written by STEM teachers, Tables 3.17 and 3.18 display the topic modeling and sentiment analysis results, respectively. Topic modeling results for female, Black, and Asian students are broadly identical to the results I observed in the full sample using teacher fixed effects, with only minor exceptions. First, note that Black students are now about as likely as White students to have notable discussion about Advanced Coursetaking, whereas in the full sample they were 0.5pp less likely than White students. Conversely, female students are much less likely than in the full sample to have notable discussion about Advanced Coursetaking versus male students (-1.1pp from -0.6pp in the full sample). Moreover, female students are substantially less likely to have discussion about STEM than in the full sample versus male students (-1.5pp from -0.7pp). That this difference is about twice as big among the subsample of STEM teachers suggests that the difference I observed in the full sample of teachers is indeed *largely* driven by individual STEM teachers' writing behaviors, especially given that this analysis already accounts for advanced science coursetaking and other observable academic characteristics.

In Table 3.18, sentiment analysis results are nearly identical to results in the full sample. That is, female students tend to have more positive sentences in their letters, while both Black and Asian students tend to have fewer positive and more neutral sentences in their letters – a dynamic that is substantially stronger for Asian students, but still relatively small given the magnitudes at play here. In general, these results suggest that individual

STEM teachers do indeed seem to discuss STEM topics in letters for female students less than in letters for male students on average, holding as much constant about the students and student-teacher relationships as possible. This is likely to be of concern with respect to STEM-specific program admissions looking for evidence of subject expertise, even as letters written about female students by this subsample of teachers are *generally* more positive in tone versus letters written about male students.

VI. Discussion

Taking these results together suggest a handful of broad findings regarding systematic differences in teacher recommendation letters.

First, letters are generally similar in terms of their **tone** and **positivity** regardless of student demographics when holding constant other student qualifications. While female students tend to see more positive letters across all specifications, the magnitude of this difference remains substantively quite small – so small that it would stretch plausibility to suggest this would ultimately alter how a reader perceives the letters. That differences for Black and Asian students were generally much smaller and more mixed in direction suggests that the positivity of letters is not likely to be a vector of concern in terms of systematic differences in the letters. This makes some intuitive sense given the occasion and form of the letters: students must request them from teachers who are under no obligation to write them, and recommendation letters are traditionally highly positive in nature. It may still be the case, however, that this dynamic would change under different circumstances – for example under systems where a teacher *is* obligated to write recommendations for a broader set of students. But at least for now, I find no evidence that systematic differences manifest broadly in terms of *positivity* here.

Second, the topical **content** of letters does indeed seem to differ by student demographics when holding constant other factors, though the exact nature and magnitude of these differences are more nuanced and complicated in interpretation. Female students most consistently had letters with greater discussion about community engagement, time and life management, and extracurriculars, and with less discussion about sports and STEM topics. This finding remained true across specifications and subsample analyses, and the difference in STEM topics was nearly twice as large when looking only at letters written by STEM teachers. That these differences align with some of the more prominent stereotypes about female students should give us pause to the extent that teachers' focus on these topics may mischaracterize female students' actual qualifications in these arenas or distract from their other qualifications and skillsets. In particular, this may be of concern in competitive admissions to STEM-focused institutions or programs, especially given current policy and programmatic focus on gender diversity in STEM-related fields.

Differences in letter content across student race were generally smaller than differences in letter content across student gender. In the full sample, across specifications, Black and White students generally saw highly similar letters, though Black students generally had less discussion about sports and time and life management. Asian students likewise had greater discussion about STEM topics and less discussion about sports. But again, the smaller magnitude of these differences suggest that letters may not be of concern in general when it comes to racial disparities. That said, this narrative is more complicated when looking specifically at the subset of competitive applicants to highly selective institutions. Black and Asian students in this group both tend to see slightly more discussion about "campus contribution" topics (e.g., leadership and community engagement), with fewer differences across most academic topics. The only exceptions here are that Black and Asian

students both saw less discussion about humanities and intellectual promise than White students, and the differences for Black students are substantially larger.

The finding that Asian students tended to see slightly *greater* discussion about these campus contribution topics, with no meaningful difference in letter positivity, is likely to have immediate implications for the U.S. Supreme Court's review of Students for Fair Admissions v. Harvard University (2019) later this year. Initial fact-finding and expert testimony to the U.S. District Court of Massachusetts made clear that lower "personal ratings" for Asian American students at Harvard were a critical element in justifying their lower admissions chances, and the court decision ultimately urged for greater examination into the more narrative components of student applications like essays, interviews, and recommendation letters as a result. My research indicates that the observable differences in teacher recommendation letters for Asian and White students with equally competitive applications to highly selective institutions do not explain the debated difference in admissions probability for Asian applicants. It may be the case that this difference is instead grounded in components like the essay, interviews, counselor letters, or other aspects of their application not analyzed here, and further inquiry is required in these directions. My work thus broaches one prominent element of this conversation, but cannot offer definitive evidence about the presence of racial discrimination per se given the remaining work to be explored.

Whether the difference in letter content among these competitive applicants has implications for racial equity otherwise then also depends on exactly how the letters are used. If they are most instrumentally used to contextualize students' intellectual credentials *beyond* their classroom performance, the fact that both Black and Asian students saw systematically less discussion about their intellectual promise could be detrimental to their college

aspirations. But if they are instead used only to contextualize students' character credentials, Black and Asian students may actually be benefited from greater emphasis on these letters in this way. That said, further study is required to better understand how letter content relates to admissions officers' perceptions of the letters, as well as how those perceptions are ultimately incorporated into admissions decisions, and whether these perceptions are also subject to reader biases. Only then can we know in what contexts, and to what degree, these content differences are likely to matter for college access concerns – but these results establish an imperative for caution and care among practitioners incorporating these letters into high-stakes decision making.

With this robust dataset and thoughtful application of NLP methods, even given my aforementioned limitations, these results offer the most comprehensive evidence regarding systematic differences in teacher recommendation letters to date – examining letters across student demographics, teacher characteristics, and institutional contexts. My hope is that this work not only illuminates greater insight into the dynamics of systematic differences in recommendation letters, but also offers a useful example other researchers might follow to apply these cutting-edge NLP methodologies to answer questions of import in the field of education policy. To that end, I eagerly invite other researchers to review my analytic code for the purposes of replication, additional robustness checks, and future application through the open-source codebase I make available alongside this manuscript.

I intend to expand on this present analysis in a variety of ways. First, there exist a large number of more fundamental equity questions related to the system of recommendation letters that are worth exploring in greater detail, such as the distribution of highly experienced letter writers across schools and students. Also of primary concern is the possibility that the requirement of letters themselves may differentially affect students' ability

to successfully apply to certain institutions – for example, might it be the case that some students simply abandon their attempts to apply to an institution if they find there is a teacher recommendation requirement too close to the deadline?

Second, there are several extensions to this work that could be fruitful for exploration. There are a variety of robustness improvements and alternative specifications I hope to explore for the results presented here. For example, I plan to improve on the robustness of the content analysis by exploring additional algorithmic (e.g., by calculating semantic word distance and clustering that way) and qualitative approaches (e.g., through thematic analysis with a team of analysts) for creating relevant word groupings. such as the examination of letter "archetypes" and template forms via a cluster analysis of letter characteristics to determine whether certain students are more likely to receive certain *styles* of letter than others. Similarly, it remains critical that we better understand how these letter characteristics may be *perceived* by readers – and exploring this in partnership with trained admissions officers will be of utmost importance to that end. I am also interested in examining letter characteristics beyond whole-letter trends – for example, by examining trajectories of sentence positivity over the course of a letter to track stories of persisting through difficulty. Lastly, there remain many subsample analyses and specifications I hope to explore, such as among students more likely to be marginal for admissions to selective institutions, or among less selective institutions.

Table 3.1. Descriptive Statistics for the Analytic Sample: Students

Variable	2018	2019	Pooled
Sample			
Students	744848	751465	1496313
Letters	1266389	1264928	2531317
Student Demographics			
Female	0.565	0.566	0.566
First Generation	0.274	0.267	0.271
International	0.104	0.108	0.106
Fee Waiver Recipient	0.23	0.229	0.229
Student Race/Ethnicity			
White	0.491	0.486	0.488
Black	0.085	0.086	0.086
Latinx	0.139	0.142	0.14
Asian	0.105	0.109	0.107
Other	0.047	0.048	0.047
Missing	0.133	0.13	0.132
Student School Sector			
Public School	0.748	0.753	0.75
Private School	0.246	0.24	0.243
Other School	0.007	0.007	0.007

Variable	2018	2019	Pooled
Applications Sent			
1-3	0.338	0.32	0.329
4-7	0.378	0.379	0.378
>=8	0.284	0.301	0.293
Scaled GPA Group			
Other/Missing	0.206	0.142	0.174
<0.90	0.228	0.247	0.238
0.90-0.99	0.282	0.301	0.292
1.00	0.034	0.037	0.035
>1.00	0.249	0.274	0.262
Math SAT/ACT Percentile Group			
Missing	0.266	0.231	0.249
<75	0.253	0.278	0.266
75-89	0.207	0.22	0.213
90-94	0.102	0.104	0.103
>=95	0.173	0.166	0.169
Verbal SAT/ACT Percentile Group			
Missing	0.265	0.23	0.247
<75	0.248	0.27	0.259
75-89	0.178	0.183	0.181
90-94	0.12	0.12	0.12
>=95	0.189	0.197	0.193

Table 3.2. Descriptive Statistics for the Analytic Sample: Teachers

Variable	2018	2019	Pooled
Sample			
Teachers	358597	361755	537306
Letters Written in Current Year			
1	0.441	0.446	0.443
2-5	0.371	0.371	0.371
6-10	0.116	0.113	0.114
11-25	0.065	0.063	0.064
25+	0.007	0.007	0.007
Letters Written in Past Two Years			
0	0.429	0.406	0.418
1	0.115	0.116	0.116
2-5	0.21	0.218	0.214
6-10	0.111	0.117	0.114
11-25	0.104	0.109	0.106
25+	0.031	0.033	0.032

Table 3.3. Descriptive Statistics for the Analytic Sample: Student-Teacher Relationships

Variable	2018	2019	Pooled
Subject Area			
English	0.229	0.228	0.228
Social Studies	0.174	0.174	0.174
Math	0.165	0.165	0.165
Science	0.201	0.2	0.201
World Language	0.066	0.063	0.064
Computer Science	0.015	0.017	0.016
Other	0.15	0.153	0.152
Student-Coach Relationship			
Coach	0.049	0.049	0.049
Years Known			
0	0.016	0.017	0.017
1	0.513	0.509	0.511
2	0.314	0.316	0.315
3	0.098	0.098	0.098
4	0.058	0.06	0.059
Other Relationship			
Other Relationship	0.034	0.035	0.034

Table 3.4. Example Subtopics and Keywords for Topic Modeling Supertopics

Supertopic	Subtopic Interpretation	Top 5 Keywords
Academic Excellence	High academic performance	top, grades, earned, performance, academic
Academic Excellence	Studying and assessments	study, learned, quickly, solving, test
Advanced Coursetaking	Advanced coursetaking	level, advanced, courses, college, placement
Advanced Coursetaking	Advanced coursetaking	ap, classes, language, composition, computer
Character Excellence	Commitment and dedication	dedicated, takes, social, conscientious, committed
Character Excellence	Consistency and diligence in improving	consistently, effort, improve, quality, rare
Community Engagement	Ability and desire for social impact	life, people, world, potential, environment
Community Engagement	Community service	community, active, school community, service, tokenname active
Extracurriculars	Extracurriculars	activities, involved, extracurricular, extracurricular activities, participated
Formalities	Thrilled to support with this letter	support, application, tokenname tokenname, chance, support tokenname
Formalities	Wonderful asset to your college	college, asset, addition, wonderful, university
Future Success Potential	Likelihood of future success	success, future, career, education, field
Future Success Potential	Future success potential	successful, business, possesses, developed, setting
Humanities	Research project	research, issues, real, project, presentation
Humanities	Social studies	history, art, arts, world, history class
Intellectual Promise	Ability and desire to learn	ability, learn, impressed, desire, succeed
Intellectual Promise	Capacity for intellectual growth	mind, growth, intellectual, curiosity, willingness
Leadership	Leadership in music	leader, leadership, music, band, roles
Leadership	Being a role model	peers, role, maturity, model, lead
Sports	Past coaching anecdote	past, opportunity, track, tokenname past, country
Sports	Sports	team, varsity, soccer, athlete, player
STEM	STEM classes	chemistry, biology, lab, honors, ap chemistry
STEM	Math classes	math, mathematics, calculus, algebra, pre
Time and Life Management	Balancing a challenging schedule	academic, academics, extra, job, schedule
Time and Life Management	Time management	time, time tokenname, management, spent, amount

Table 3.5. Sample Sentences and Assigned Sentiment Scores

Letter Sentence (sic)	Sentiment Score
At one point in the year, this student became extremely ill and missed quite a bit of school.	Negative
When I had her in my class first semester of last year, to put it bluntly, her writing was atrocious.	Negative
Math is not this student's favorite subject, nor does it come without struggle.	Negative
This student struggled a bit with material and concepts covered early in the course, and he seemed resigned to just "get by" without pushing himself.	Negative
This letter is in reference to this student, who is currently a student in my calculus class.	Neutral
In my career, I have taught courses including biology, AP Biology, chemistry, and physics.	Neutral
All assessments are either written or oral presentations.	Neutral
There were numerous group projects required in this class.	Neutral
This student always asked relevant questions in class, demonstrating her desire to understand and improve her knowledge.	Positive
He is an active participant in discussions and is determined to do well.	Positive
This student is an exceptionally talented young man.	Positive
I highly recommend this student be accepted into your college or university because she would make an excellent addition to your student body.	Positive

Note: Texts shown here were specifically selected to clearly illustrate what differing levels of sentiment, per the aforementioned sentiment construct definition, would look like. These examples should not be interpreted as a general demonstration of algorithm accuracy.

Table 3.6. Student-level Covariates in Regression Models

Demographics	Academics
○ Class year (Senior or not) ○ Cohort year (2018 or 2019) ○ Gender (male, female, or missing) ○ Age (above 17, below 17, or 17) ○ International ○ Race/Ethnicity (White, Black, Asian, Latinx, Other, or Missing) ○ First-gen Status ○ Fee Waiver Receipt (low-income proxy) ○ Attended multiple schools ○ School Sector (public, private, other)	○ Class rank quintile (58% miss.) ○ Scaled GPA (<0.9, 0.9-0.99, 1.0, >1.0) (17% miss.) ○ Number of SAT/ACT tests taken (0, 1, 2, >2) ○ SAT/ACT Math Percentile (<75, 75-89, 90-94, >=95) ○ SAT/ACT Verbal Percentile (<75, 75-89, 90-94, >=95) ○ Any SAT subject tests taken ○ Average SAT subject test score (<750, >=750) ○ Number of AP tests taken by subject (English, World Languages, Social Studies, STEM, Arts) ○ Avg AP score by subject (Missing, <4.5, >=4.5) ○ Any IB tests ○ TOEFL Percentile (missing, <90, >=90)
Extracurriculars	**Application Behaviors**
○ Any activity by category (Academic, Career, Arts, Service, Athletic, Other) ○ Count of activities by category ○ Any leadership role by category ○ Any excellence award by category ○ Total leadership roles (across categories) ○ Total excellence awards (across categories) ○ Total mentorship roles (across categories)	○ Applications sent (1-3, 4-7, >=8) ○ Any early deadline applications ○ *Only* early deadline applications (to distinguish students w/ high competitiveness but low application count) ○ Avg. selectivity quintile of institutions applied to

Table 3.7. Topic Modeling Main Results: Landscape Analysis

Variable	Academic Excellence	Advanced Course-taking	Character Excellence	Community Engage-ment	Extra-currics.	Formalities	Future Potential	Humanities	Intellectual Promise	Leadership	Sports	STEM	Time and Life Mgmt
80th Pctile	18.04	8.23	36.34	12.84	3.96	33.72	6.02	23.81	20.72	15.69	6.64	8.67	10.41
Female	-0.011***	-0.000	-0.000	0.026***	0.024***	-0.001	0.002**	0.002**	-0.010***	0.006***	-0.028***	-0.009***	0.012***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Black	0.005***	0.009***	-0.000	0.007***	-0.002	0.014***	-0.003*	-0.005***	0.001	0.009***	-0.014***	0.010***	-0.012***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Asian	-0.001	0.007***	-0.003**	0.011***	0.009***	0.000	0.006***	-0.003**	-0.003**	0.000	-0.016***	0.014***	-0.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317
R^2	0.29564	0.19368	0.32775	0.22409	0.10652	0.15310	0.15356	0.32912	0.31067	0.26354	0.22457	0.37098	0.21611
Adj. R^2	0.29561	0.19364	0.32772	0.22405	0.10648	0.15307	0.15352	0.32909	0.31064	0.26351	0.22454	0.37095	0.21608

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table 3.8. Topic Modeling Main Results: Teacher Fixed Effects Analysis

Variable	Academic Excellence	Advanced Course-taking	Character Excellence	Community Engage-ment	Extra-currics.	Formalities	Future Potential	Humanities	Intellectual Promise	Leadership	Sports	STEM	Time and Life Mgmt
80th Pctile	18.04	8.23	36.34	12.84	3.96	33.72	6.02	23.81	20.72	15.69	6.64	8.67	10.41
Female	-0.011***	-0.006***	0.003***	0.023***	0.020***	0.003***	0.003***	-0.004***	-0.011***	0.007***	-0.026***	-0.007***	0.013***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)
Black	-0.004***	-0.005***	0.005***	0.007***	0.002*	0.003***	0.003***	-0.001	-0.003**	0.005***	-0.007***	0.001	-0.008***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Asian	-0.004***	0.007***	-0.003***	0.008***	0.008***	-0.001	0.005***	-0.001	-0.006***	0.004***	-0.014***	0.010***	-0.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317	2,531,317
R^2	0.59350	0.60489	0.63157	0.56409	0.55629	0.67116	0.56487	0.69866	0.61492	0.58794	0.52681	0.69296	0.51784
Within R^2	0.11937	0.06808	0.16055	0.12280	0.05081	0.08280	0.05796	0.09463	0.12499	0.15930	0.15436	0.06829	0.10442

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table 3.9. Topic Modeling Main Results: Institution Fixed Effects Analysis

Variable	Academic Excellence	Advanced Course-taking	Character Excellence	Community Engage-ment	Extra-currics.	Formalities	Future Potential	Humanities	Intellectual Promise	Leadership	Sports	STEM	Time and Life Mgmt
80th Pctile	18.04	8.23	36.34	12.84	3.96	33.72	6.02	23.81	20.72	15.69	6.64	8.67	10.41
Female	-0.012***	0.000	-0.001	0.024***	0.021***	-0.001	0.001	0.002*	-0.010***	0.005***	-0.028***	-0.006***	0.011***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Black	0.002	0.006***	-0.001	0.005***	-0.007***	0.012***	-0.000	-0.006***	-0.001	0.005***	-0.011***	0.009***	-0.013***
	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Asian	-0.001	0.003.	-0.002	0.014***	0.011***	-0.001	0.008***	-0.007***	-0.006***	0.002.	-0.007***	0.014***	-0.006***
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	12,708,496	12,708,496	12,708,496	12,708,496	12,708,496	12,708,496	12,708,496	12,708,496	12,708,496	12,708,496	12,708,496	12,708,496	12,708,496
R^2	0.30931	0.19756	0.33194	0.22634	0.10453	0.15452	0.15470	0.36041	0.32509	0.26219	0.21498	0.39426	0.22143
Within R^2	0.29061	0.18274	0.32889	0.22364	0.10033	0.15374	0.15297	0.33976	0.30488	0.25664	0.20919	0.37556	0.21724

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table 3.10. Sentiment Analysis Main Results: Landscape Analysis

Variable	Positive	Neutral	Negative	Mean Sent.
Sample Mean	13.81	5.38	0.48	0.7
Female	0.192***	-0.174***	-0.018***	0.009***
	(0.006)	(0.005)	(0.002)	(0.000)
Black	-0.116***	0.110***	0.006*	-0.003***
	(0.010)	(0.009)	(0.003)	(0.001)
Asian	-0.045***	0.046***	-0.001	-0.002***
	(0.011)	(0.009)	(0.003)	(0.001)
Observations	2,531,317	2,531,317	2,531,317	2,531,317
R^2	0.69810	0.53992	0.17462	0.09162
Adj. R^2	0.69809	0.53990	0.17458	0.09158

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table 3.11. Sentiment Analysis Main Results: Teacher Fixed Effects Analysis

Variable	Positive	Neutral	Negative	Mean Sent.
Sample Mean	13.81	5.38	0.48	0.7
Female	0.154*** (0.004)	-0.145*** (0.003)	-0.009*** (0.001)	0.007*** (0.000)
Black	-0.018** (0.006)	0.024*** (0.006)	-0.006* (0.003)	-0.000 (0.000)
Asian	-0.073*** (0.005)	0.073*** (0.005)	0.000 (0.002)	-0.003*** (0.000)
Observations	2,531,317	2,531,317	2,531,317	2,531,317
R^2	0.88920	0.82223	0.53054	0.65550
Within R^2	0.59314	0.43317	0.10552	0.05428

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table 3.12. Sentiment Analysis Main Results: Institution Fixed Effects Analysis

Variable	Positive	Neutral	Negative	Mean Sent.
Sample Mean	14.13	5.61	0.52	0.69
Female	0.190*** (0.007)	-0.172*** (0.006)	-0.017*** (0.002)	0.009*** (0.000)
Black	-0.123*** (0.013)	0.119*** (0.012)	0.005 (0.004)	-0.004*** (0.001)
Asian	-0.052*** (0.013)	0.051*** (0.011)	0.001 (0.004)	-0.002*** (0.001)
Observations	12,708,496	12,708,496	12,708,496	12,708,496
R^2	0.69017	0.55032	0.17897	0.09865
Within R^2	0.68424	0.54092	0.17400	0.09236

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table 3.13. Highly Selective Institution Subsample Results: Topic Modeling

Variable	Academic Excellence	Advanced Course-taking	Character Excellence	Community Engage-ment	Extra-currics.	Formalities	Future Potential	Humanities	Intellectual Promise	Leadership	Sports	STEM	Time and Life Mgmt
80th Pctile	18.04	8.23	36.34	12.84	3.96	33.72	6.02	23.81	20.72	15.69	6.64	8.67	10.41
Female	-0.012***	-0.006***	-0.001	0.021***	0.014***	-0.002	-0.001	0.006***	-0.012***	0.004***	-0.029***	-0.007***	0.010***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Black	-0.004*	0.003	-0.001	0.012***	0.007**	0.011***	0.007**	-0.014***	-0.010***	0.016***	-0.006**	0.009***	-0.012***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Asian	-0.003	-0.001	0.003.	0.018***	0.016***	-0.003	0.013***	-0.014***	-0.009***	0.008***	0.002	0.012***	-0.004*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	4,181,426	4,181,426	4,181,426	4,181,426	4,181,426	4,181,426	4,181,426	4,181,426	4,181,426	4,181,426	4,181,426	4,181,426	4,181,426
R^2	0.32717	0.20150	0.33844	0.24144	0.11067	0.15797	0.16386	0.39485	0.33370	0.26545	0.20233	0.45602	0.23192
Within R2	0.32543	0.19861	0.33755	0.23980	0.10884	0.15764	0.16318	0.39114	0.33174	0.26315	0.19981	0.45099	0.23093

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table 3.14. Highly Selective Institution Subsample Results: Sentiment Analysis

Variable	Positive	Neutral	Negative	Mean Sent.
Sample Mean	14.85	6.19	0.58	0.68
Female	0.192***	-0.182***	-0.010**	0.009***
	(0.011)	(0.009)	(0.003)	(0.000)
Black	-0.095***	0.107***	-0.012*	-0.001
	(0.019)	(0.017)	(0.006)	(0.001)
Asian	0.002	0.004	-0.006	0.001
	(0.017)	(0.015)	(0.005)	(0.001)
Observations	4,181,426	4,181,426	4,181,426	4,181,426
R^2	0.67961	0.56750	0.18008	0.10184
Within R^2	0.67868	0.56655	0.17937	0.10118

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table 3.15. Highly Selective Institution and Competitive Applicant Subsample Results: Topic Modeling

Variable	Academic Excellence	Advanced Course-taking	Character Excellence	Community Engage-ment	Extra-currics.	Formalities	Future Potential	Humanities	Intellectual Promise	Leadership	Sports	STEM	Time and Life Mgmt
80th Pctile	18.04	8.23	36.34	12.84	3.96	33.72	6.02	23.81	20.72	15.69	6.64	8.67	10.41
Female	-0.015*** (0.002)	-0.015*** (0.002)	0.000 (0.002)	0.019*** (0.002)	0.008*** (0.002)	-0.002 (0.002)	-0.002 (0.002)	0.011*** (0.002)	-0.013*** (0.002)	0.002 (0.002)	-0.033*** (0.002)	-0.008*** (0.002)	0.011*** (0.002)
Black	-0.017. (0.009)	0.002 (0.009)	0.004 (0.008)	0.028*** (0.008)	0.033*** (0.009)	0.003 (0.009)	-0.004 (0.008)	-0.026*** (0.008)	-0.027** (0.009)	0.024** (0.008)	0.001 (0.008)	0.009 (0.008)	0.003 (0.009)
Asian	-0.004 (0.003)	-0.006 (0.003)	0.005. (0.003)	0.017*** (0.003)	0.021*** (0.003)	-0.004 (0.003)	0.014*** (0.003)	-0.017*** (0.003)	-0.009** (0.003)	0.012*** (0.003)	0.012*** (0.003)	0.009*** (0.003)	-0.002 (0.003)
Observations	1,281,526	1,281,526	1,281,526	1,281,526	1,281,526	1,281,526	1,281,526	1,281,526	1,281,526	1,281,526	1,281,526	1,281,526	1,281,526
R^2	0.32811	0.20335	0.34388	0.24591	0.11729	0.16159	0.17094	0.42958	0.32689	0.27214	0.20227	0.52270	0.23852
Within R2	0.32771	0.20206	0.34319	0.24448	0.11580	0.16126	0.17027	0.42675	0.32627	0.27062	0.20032	0.52039	0.23798

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table 3.16. Highly Selective Institution and Competitive Applicant Subsample Results: Sentiment Analysis

Variable	Positive	Neutral	Negative	Mean Sent.
Sample Mean	15.65	6.8	0.64	0.67
Female	0.210*** (0.021)	-0.205*** (0.019)	-0.005 (0.006)	0.008*** (0.001)
Black	-0.063 (0.073)	0.082 (0.066)	-0.019 (0.024)	0.000 (0.004)
Asian	0.059* (0.029)	-0.041 (0.026)	-0.018* (0.008)	0.003** (0.001)
Observations	1,281,526	1,281,526	1,281,526	1,281,526
R^2	0.65926	0.57588	0.17939	0.10851
Within R^2	0.65881	0.57552	0.17900	0.10798

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table 3.17. STEM Teacher Subsample Results: Topic Modeling

Variable	Academic Excellence	Advanced Course-taking	Character Excellence	Community Engage-ment	Extra-currics.	Formalities	Future Potential	Humanities	Intellectual Promise	Leadership	Sports	STEM	Time and Life Mgmt
80th Pctile	18.04	8.23	36.34	12.84	3.96	33.72	6.02	23.81	20.72	15.69	6.64	8.67	10.41
Female	-0.012*** (0.001)	-0.011*** (0.001)	0.007*** (0.001)	0.021*** (0.001)	0.023*** (0.001)	0.003** (0.001)	0.006*** (0.001)	-0.004*** (0.000)	-0.014*** (0.001)	0.009*** (0.001)	-0.025*** (0.001)	-0.015*** (0.001)	0.018*** (0.001)
Black	-0.005** (0.002)	-0.000 (0.001)	0.005*** (0.001)	0.002 (0.002)	0.002 (0.002)	0.004** (0.001)	0.002 (0.002)	0.000 (0.001)	-0.002 (0.002)	0.004* (0.002)	-0.008*** (0.002)	-0.003 (0.002)	-0.007*** (0.002)
Asian	-0.004*** (0.001)	0.007*** (0.001)	-0.005*** (0.001)	0.008*** (0.001)	0.008*** (0.001)	-0.001 (0.001)	0.004** (0.001)	0.000 (0.001)	-0.008*** (0.001)	0.005*** (0.001)	-0.013*** (0.001)	0.015*** (0.001)	-0.008*** (0.001)
Observations	966,910	966,910	966,910	966,910	966,910	966,910	966,910	966,910	966,910	966,910	966,910	966,910	966,910
R^2	0.61873	0.59747	0.62646	0.54324	0.54967	0.67044	0.55554	0.57069	0.62083	0.55465	0.51130	0.62123	0.51557
Within R2	0.12784	0.07304	0.15268	0.11090	0.05439	0.08467	0.05936	0.07199	0.12380	0.15162	0.15828	0.11563	0.10341

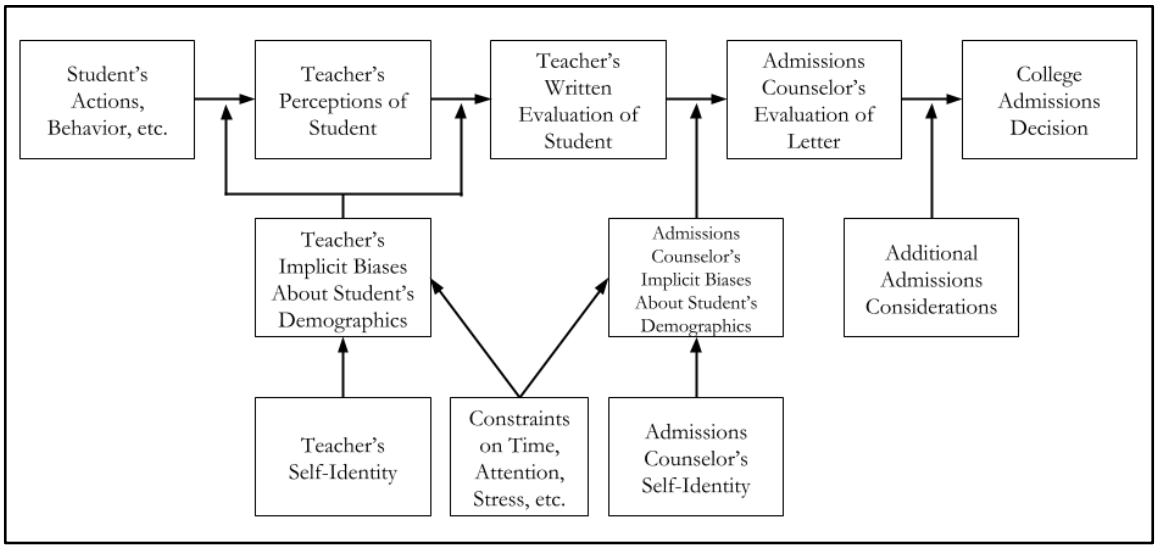
Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table 3.18. STEM Teacher Subsample Results: Sentiment Analysis

Variable	Positive	Neutral	Negative	Mean Sent.
Sample Mean	13.79	5.48	0.47	0.69
Female	0.136*** (0.006)	-0.137*** (0.005)	0.002 (0.002)	0.006*** (0.000)
Black	-0.036*** (0.010)	0.038*** (0.010)	-0.002 (0.004)	-0.001* (0.001)
Asian	-0.075*** (0.008)	0.081*** (0.008)	-0.006. (0.003)	-0.003*** (0.000)
Observations	966,910	966,910	966,910	966,910
R^2	0.89118	0.82004	0.51491	0.64853
Within R2	0.60133	0.42218	0.09453	0.04836

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Figure 3.1. Theoretical Model for Implicit Bias in Teacher Recommendations



REFERENCES

Acuerdo Ministerial. No. 4025-2012 (2012). https://leyes.infile.com/index.php?id=182&id_publicacion=67124

Adukia, A., Eble, A., Harrison, E., Runesha, H. B., & Szasz, T. (2021). What We Teach About Race and Gender: Representation in Images and Text of Children's Books (SSRN Scholarly Paper ID 3825080). Social Science Research Network. https://doi.org/10.2139/ssrn.3825080

Airoldi, E. M., & Bischof, J. M. (2016). Improving and Evaluating Topic Models and Other Models of Text. Journal of the American Statistical Association, 111(516), 1381–1403. https://doi.org/10.1080/01621459.2015.1051182

Akos, P., & Kretchmar, J. (2016). Gender and Ethnic bias in Letters of Recommendation: Considerations for School Counselors. Professional School Counseling, 20(1), 1096-2409-20.1.102. https://doi.org/10.5330/1096-2409-20.1.102

Akresh, R., Halim, D., & Kleemans, M. (2021). Long-Term and Intergenerational Effects of Education : Evidence from School Construction in Indonesia. Policy Research Working Paper;No. 9559. World Bank, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/35208 License: CC BY 3.0 IGO.

Alikaniotis, D., & Raheja, V. (2019, August 7). Under the Hood at Grammarly: Leveraging Transformer Language Models for Grammatical Error Correction |. Grammarly Engineering Blog. https://www.grammarly.com/blog/engineering/under-the-hood-at-grammarly-leveraging-transformer-language-models-for-grammatical-error-correction/

Alm, J., & Winters, J. V. (2009). Distance and intrastate college student migration. Economics of Education Review, 28(6), 728–738. https://doi.org/10.1016/j.econedurev.2009.06.008

Alvero, A., Giebel, S., Gebre-Medhin, B., antonio, anthony lising, Stevens, M. L., & Domingue, B. W. (2021). Essay Content is Strongly Related to Household Income and SAT Scores: Evidence from 60,000 Undergraduate Applications (No. 21–03; CEPA Working Papers). Stanford Center for Education Policy Analysis. https://cepa.stanford.edu/content/essay-content-strongly-related-household-income-and-sat-scores-evidence-60000-undergraduate-applications

Ambartsoumian, A., & Popowich, F. (2018). Self-Attention: A Better Building Block for Sentiment Analysis Neural Network Classifiers. ArXiv:1812.07860 [Cs]. https://doi.org/10.18653/v1/P17

Anglin, K. L. (2019). Gather-Narrow-Extract: A Framework for Studying Local Policy Variation Using Web-Scraping and Natural Language Processing. Journal of Research on Educational Effectiveness, 12(4), 685–706. https://doi.org/10.1080/19345747.2019.1654576

Arnold, K. D., Owen, L., & Lewis, J. (2020). Inside the Black Box of Text-Message College Advising. Journal of College Access, 5(2), 5.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. The Annals of Statistics, 47(2), 1148–1178. https://doi.org/10.1214/18-AOS1709

Avery, C., & Turner, S. (2012). Student Loans: Do College Students Borrow Too Much – Or Not Enough? Journal of Economic Perspectives 26(1), 165-193.

Avery, C., Castleman, B. L., Hurwitz, M., Long, B. T., & Page, L. C. (2021). Digital messaging to improve college enrollment and success. National Bureau of Economic Research Working Paper No 27897. Retrieved from https://www.nber.org/system/files/working_papers/w27897/w27897.pdf

Avitzour, E., Choen, A., Joel, D., & Lavy, V. (2020). On the Origins of Gender-Biased Behavior: The Role of Explicit and Implicit Stereotypes (SSRN Scholarly Paper ID 3692175). Social Science Research Network. https://papers.ssrn.com/abstract=3692175

Baker, R., Xu, D., Park, J., Yu, R., Li, Q., Cung, B., Fischer, C., Rodriguez, F., Warschauer, M., & Smyth, P. (2020). The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: Opening the black box of learning processes. International Journal of Educational Technology in Higher Education, 17(1), 13. https://doi.org/10.1186/s41239-020-00187-1

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. ArXiv:2010.12421 [Cs]. http://arxiv.org/abs/2010.12421

Bastedo, M. N., Bowman, N. A., Glasener, K. M., & Kelly, J. L. (2018). What are We Talking About When We Talk About Holistic Review? Selective College Admissions and its Effects on Low-SES Students. The Journal of Higher Education, 89(5), 782–805. https://doi.org/10.1080/00221546.2018.1442633

Bastedo, M. N., Glasener, K. M., Deane, K. C., & Bowman, N. A. (2019). Contextualizing the SAT: Experimental Evidence on College Admission Recommendations for Low-SES Applicants. Educational Policy, 0895904819874752. https://doi.org/10.1177/0895904819874752

Batabyal, A. A., & Nijkamp, P. (2004). Favoritism in the Public Provision of Goods in Developing Countries. Economics Bulletin, 15(1), 1-9.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. (2018) "quanteda: An R package for the quantitative analysis of textual data". Journal of Open Source Software. 3(30), 774. https://doi.org/10.21105/joss.00774.

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit Discrimination. The American Economic Review, 95(2), 94–98.

Bettinger, E., Liu, J., & Loeb, S. (2016). Connections Matter: How Interactive Peers Affect Students in Online College Courses: Interactive Peers Affect Students in Online Courses. Journal of Policy Analysis and Management, 35(4), 932–954. https://doi.org/10.1002/pam.21932

Bettinger, E. P., Castleman, B. L., Choe, A., & Mabel, Z. (2021). Finishing the Last Lap: Experimental Evidence on Strategies to Increase College Completion for Students At Risk of Late Withdrawal. EdWorkingPaper No. 21-488. Retrieved from https://edworkingpapers.org/sites/default/files/ai21-488.pdf

Bird, K. A., Castleman, B. L., Denning, J. T., Goodman, J., Lamberton, C., & Rosinger, K. O. (2021). Nudging at scale: Experimental evidence from FAFSA completion campaigns. Journal of Economic Behavior & Organization, 183, 105-128.

Blei, D. M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 30.

Bohnet, B., McDonald, R., Simoes, G., Andor, D., Pitler, E., & Maynez, J. (2018). Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings. ArXiv:1805.08237 [Cs]. http://arxiv.org/abs/1805.08237

Bound, J., Lovenheim, M., & Turner, S. (2010). Why have college completion rates declined? An analysis of changing student preparation and collegiate resources. American Economic Journal, 2(3), 129-157.

Burde, D., & Linden, L. L. (2013). Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools. American Economic Journal: Applied Economics, 5(3), 27–40. JSTOR.

Burgess, R., Jedwab, R., Miguel, E., Morjaria, A., & Padró i Miquel, G. (2015). The Value of Democracy: Evidence from Road Building in Kenya. American Economic Review, 105(6), 1817–1851. https://doi.org/10.1257/aer.20131031

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186. https://doi.org/10.1126/science.aal4230

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias. The Quarterly Journal of Economics, 134(3), 1163–1224. https://doi.org/10.1093/qje/qjz008

Carnevale, A. P., Jayasundera, T., & Gulish, A. (2016). America's Divided Recovery: College Haves and Have-Nots. Georgetown University Center on Education and the Workforce.

Cartujano-Barrera, F., Arana-Chicas, E., Ramírez-Mantilla, M., Perales, J., Cox, L. S., Ellerbeck, E. F., ... & Cupertino, A. P. (2019). "Every day I think about your messages": assessing text messaging engagement among Latino smokers in a mobile cessation program. Patient preference and adherence, 13, 1213.

Castleman, B.L. & Page, L. (2016). Freshman year financial aid nudges: An experiment to increase FAFSA renewal and college persistence. Journal of Human Resources, 51(2), 389-415.

Castleman, B. L., Meyer, K. E., Sullivan, Z., Hartog, W. D., Miller, S. (2017). Nudging students beyond the FAFSA: The impact of university outreach on financial aid behaviors and outcomes. Journal of Student Financial Aid, 47(3) 2.

Cavagna, G. A., Franzetti, P., Fuchimoto, T. (1983). The mechanics of walking in children. The Journal of Physiology. 343, 323-339.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. Proceedings of the 22Nd International Conference on Neural Information Processing Systems, 288–296. http://dl.acm.org/citation.cfm?id=2984093.2984126

Chin, M. J., Quinn, D. M., Dhaliwal, T. K., & Lovison, V. S. (2020). Bias in the Air: A Nationwide Exploration of Teachers' Implicit Racial Attitudes, Aggregate Bias, and Student Outcomes: Educational Researcher. https://doi.org/10.3102/0013189X20937240

CIESIN: Facebook Connectivity Lab and Center for International Earth Science Information Network. (2016). High Resolution Settlement Layer (HRSL). Columbia University and DigitalGlobe. https://www.ciesin.columbia.edu/data/hrsl/

Clinedinst, M., & Koranteng, A.-M. (2017). 2017 State of College Admission. National Association for College Admission Counseling. https://www.nacacnet.org/globalassets/documents/publications/research/soca17final.pdf

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. ArXiv:1701.08230 [Cs, Stat]. https://doi.org/10.1145/3097983.309809

Cosentino, C. (2017, September 18). Supporting Secondary Education in Developing Nations. Mathematica. https://www.mathematica.org/commentary/supporting-secondary-education-in-developing-nations

Cueva, D. (2020, July 29). Transportistas anuncian un incremento de hasta el triple del costo del pasaje. Prensa Libre. www.prensalibre.com.

Dale, S. B., & Krueger, A. B. (2002). Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables. The Quarterly Journal of Economics, 117(4), 1491–1527. https://doi.org/10.1162/003355302320935089

Dale, S. B., & Krueger, A. B. (2011). Estimating the Return to College Selectivity over the Career Using Administrative Earnings Data (Working Paper No. 17159). National Bureau of Economic Research. https://doi.org/10.3386/w17159

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences of the United States of America, 108(17), 6889–6892. https://doi.org/10.1073/pnas.1018033108

Dee, T. S. (2005). A Teacher like Me: Does Race, Ethnicity, or Gender Matter? The American Economic Review, 95(2), 158–165.

Dee, T. S. (2007). Teachers and the Gender Gaps in Student Achievement. Journal of Human Resources, XLII(3), 528–554. https://doi.org/10.3368/jhr.XLII.3.528

Denning, J., Eide, E., & Warnick, M. (2019). Why have college completion rates increased? Working Paper No. 12411. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3408309

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. Journal of Experimental Social Psychology, 48(6), 1267–1278. https://doi.org/10.1016/j.jesp.2012.06.003

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [Cs]. http://arxiv.org/abs/1810.04805

De Vynck, G., & Bergen, M. (2020, April 9). Google Classroom Users Doubled as Quarantines Spread. Bloomberg.Com. https://www.bloomberg.com/news/articles/2020-04-09/google-widens-lead-in-education-market-as-students-rush-online

Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. The American Economic Review, 91(4), 795–813. JSTOR.

Duppada, V., Jain, R., & Hiray, S. (2018). SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in Tweets. ArXiv:1804.06137 [Cs]. http://arxiv.org/abs/1804.06137

Dutt, K., Pfaff, D. L., Bernstein, A. F., Dillard, J. S., & Block, C. J. (2016). Gender differences in recommendation letters for postdoctoral fellowships in geoscience. Nature Geoscience, 9(11), 805–808. https://doi.org/10.1038/ngeo2819

Egalite, A. J., Kisida, B., & Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. Economics of Education Review, 45, 44–52. https://doi.org/10.1016/j.econedurev.2015.01.007

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2018). How to Make Causal Inferences Using Texts. ArXiv:1802.02163 [Cs, Stat]. http://arxiv.org/abs/1802.02163

Ejdemyr, S., Kramon, E., & Robinson, A. L. (2018). Segregation, Ethnic Favoritism, and the Strategic Targeting of Local Public Goods. Comparative Political Studies, 51(9), 1111–1143. https://doi.org/10.1177/0010414017730079

Etten, J. van, & Sousa, K. de. (2020). gdistance: Distances and Routes on Geographical Grids (1.3-6) [Computer software]. https://CRAN.R-project.org/package=gdistance

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press.

Evans, D. K., & Mendez Acosta, A. (2021). Education in Africa: What Are We Learning? Journal of African Economies, 30(1), 13–54. https://doi.org/10.1093/jae/ejaa009

Fesler, L. (2020). Opening the Black Box of College Counseling. CEPA Working Paper No. 20-03. Retrieved from https://files.eric.ed.gov/fulltext/ED605975.pdf.

Fesler, L., Dee, T., Baker, R., & Evans, B. (2019). Text as Data Methods for Education Research. Journal of Research on Educational Effectiveness, 12(4), 707–727. https://doi.org/10.1080/19345747.2019.1634168

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. Economics of Education Review, 52, 209–224. https://doi.org/10.1016/j.econedurev.2016.03.002

Glewwe, P., Hanushek, E. A., Humpage Liuzzi, S., & Ravina, R. (2014) School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010. In Glewwe, P. (Ed.), Education policy in developing countries (pp.13-64). The University of Chicago Press. https:// 10.1086/680396

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. Proceedings of the First ACM Conference on Online Social Networks, 27–38. https://doi.org/10.1145/2512938.2512951

Gopalan, M., & Brady, S. (2019). College students' sense of belonging: A national perspective. Educational Researcher, 49(2).

Gould, William T. S. 1978. "Guidelines for School Location Planning." Staff Working Paper 308, World Bank, Washington, DC.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit Bias: Scientific Foundations. California Law Review, 94(4), 945–967. JSTOR. https://doi.org/10.2307/20439056

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. Journal of Personality and Social Psychology, 74(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3), 267–297. https://doi.org/10.1093/pan/mps028

Grissom, J. A., & Redding, C. (2016). Discretion and Disproportionality: Explaining the Underrepresentation of High-Achieving Students of Color in Gifted Programs. AERA Open, 2(1), 2332858415622175. https://doi.org/10.1177/2332858415622175

gspeedAIST. (2019, March 11). Guatemala DEM and Hillshade. ArcGIS. https://www.arcgis.com/home/item.html?id=c9f9b32c4221455ca3600d28c961c642

Gurantz, O., Pender, M., Mabel, Z., Larson, C., & Bettinger, E. (2020). Virtual advising for high-achieving high school students. Economics of Education Review, 75, 101974.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutorials in Quantitative Methods for Psychology, 8(1), 23–34.

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. Communication Methods and Measures, 1(1), 77–89. https://doi.org/10.1080/19312450709336664

He, S. Y., & Giuliano, G. (2018). School choice: Understanding the trade-off between travel distance and school quality. Transportation, 45(5), 1475–1498. https://doi.org/10.1007/s11116-017-9773-3

Herring, W. (2021). Better Data, Better Methods, Better Decisions? Using Statewide Longitudinal Data Systems to Create Early Warning Systems for Literacy Interventions. University of Virginia.

Hijmans, R. J., Etten, J. van, Sumner, M., Cheng, J., Baston, D., Bevan, A., Bivand, R., Busetto, L., Canty, M., Fasoli, B., Forrest, D., Ghosh, A., Golicher, D., Gray, J., Greenberg, J. A., Hiemstra, P., Hingee, K., Geosciences, I. for M. A., Karney, C., … Wueest, R. (2020). raster: Geographic Data Analysis and Modeling (3.4-5) [Computer software]. https://CRAN.R-project.org/package=raster

Hillman, N. W. (2016). Geography of College Opportunity: The Case of Education Deserts. American Educational Research Journal, 53(4), 987–1021. https://doi.org/10.3102/0002831216653204

Hong, M. K. and N. (2018, October 16). Harvard Cites Weaker Teacher Recommendations for Asian-American Applicants. Wall Street Journal. https://www.wsj.com/articles/harvard-cites-weaker-teacher-recommendations-for-asian-american-applicants-1539721051

Hossler, D., Chung, E., Kwon, J., Lucido, J., Bowman, N., & Bastedo, M. (2019). A Study of the Use of Nonacademic Factors in Holistic Undergraduate Admissions Reviews. The Journal of Higher Education, 90(6), 833–859. https://doi.org/10.1080/00221546.2019.1574694

Hoxby, C., & Avery, C. (2013). The Missing "One-Offs": The Hidden Supply of High-Achieving, Low-Income Students. Brookings Papers on Economic Activity, 2013(1), 1–65. https://doi.org/10.1353/eca.2013.0000

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. ICWSM.

Hünermund, P., & Bareinboim, E. (2019). Causal Inference and Data-Fusion in Econometrics. ArXiv:1912.09104 [Econ]. http://arxiv.org/abs/1912.09104

Irvine, L., Melson, A. J., Williams, B., Sniehotta, F. F., McKenzie, A., Jones, C., & Crombie, I. K. (2017). Real time monitoring of engagement with a text message intervention to reduce binge drinking among men living in socially disadvantaged areas of Scotland. International journal of behavioral medicine, 24(5), 713-721.

Kane, T. J. (1998). Racial and Ethnic Preferences in College Admissions. Ohio State Law Journal, 59(3), 971–996.

Kazianga, H., Levy, D., Linden, L. L., & Sloan, M. (2013). The Effects of "Girl-Friendly" Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso. American Economic Journal: Applied Economics, 5(3), 41–62. JSTOR.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. ArXiv:1408.5882 [Cs]. http://arxiv.org/abs/1408.5882

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. The Quarterly Journal of Economics, 133(1), 237–293. https://doi.org/10.1093/qje/qjx032

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. ArXiv:1609.05807 [Cs, Stat]. http://arxiv.org/abs/1609.05807

Koppensteiner, M., & Matheson, J. (n.d.). Secondary Schools and Teenage Childbearing: Evidence from the School Expansion in Brazilian Municipalities. Policy Research Working Paper. No. 9420.

Kraft, M. A., & Dougherty, S. M. (2013). The effect of teacher–family communication on student engagement: Evidence from a randomized field experiment. Journal of Research on Educational Effectiveness, 6(3), 199-222.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., … Nosek, B. A. (2014). Reducing implicit racial preferences: A comparative investigation of 17 interventions. Journal of Experimental Psychology: General, 143(4), 1765–1785. https://doi.org/10.1037/a0036260

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. ArXiv:1909.11942 [Cs]. http://arxiv.org/abs/1909.11942

Lehman, D., Buys, P., Atchina, F., Laroche, L., & Prouty, B. (2013). The Rural Access Initiative: Shortening the Distance to Education for All in the African Sahel. Washington, DC: World Bank.

Liang, W., Liang, H., Ou, L., Chen, B., Chen, A., Li, C., Li, Y., Guan, W., Sang, L., Lu, J., Xu, Y., Chen, G., Guo, H., Guo, J., Chen, Z., Zhao, Y., Li, S., Zhang, N., Zhong, N., … for the China Medical Treatment Expert Group for COVID-19. (2020). Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. JAMA Internal Medicine, 180(8), 1081–1089. https://doi.org/10.1001/jamainternmed.2020.2033

Ma, D. S., Correll, J., Wittenbrink, B., Bar-Anan, Y., Sriram, N., & Nosek, B. A. (2013). When Fatigue Turns Deadly: The Association Between Fatigue and Racial Bias in the Decision to Shoot. Basic and Applied Social Psychology, 35(6), 515–524. https://doi.org/10.1080/01973533.2013.840630

Mabel, Z., Castleman, B., & Bettinger, E. (2019). Finishing the last lap: Experimental evidence on strategies to increase college completion for students at risk of late departure. Working Paper. Retrieved from https://scholar.harvard.edu/zmabel/publications/finishing-last-lap-experimental-evidence-strategies-increase-college-completion

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. Journal of Applied Psychology, 94(6), 1591–1599. https://doi.org/10.1037/a0016539

Madera, J. M., Hebl, M. R., Dial, H., Martin, R., & Valian, V. (2018). Raising Doubt in Letters of Recommendation for Academia: Gender Differences and Their Impact. Journal of Business and Psychology. https://doi.org/10.1007/s10869-018-9541-1

Meet Our Members. (2018). Coalition for College. http://www.coalitionforcollegeaccess.org/members-new.html

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [Cs]. http://arxiv.org/abs/1301.3781

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 262–272. http://dl.acm.org/citation.cfm?id=2145432.2145462

Ministerio de Educación (2020). Establecimiento Educativos. [Data set]. https://datosabiertos.mineduc.gob.gt/dataset/establecimiento-educativos

Ministerio de Educación. (2016). Manual de Criterios Normativos para el Diseño Arquitectónico de Centros Educativos Oficiales. Policy report. ISBN: 978-9929-688-70-4

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. Proceedings of The 12th International Workshop on Semantic Evaluation, 1–17. https://doi.org/10.18653/v1/S18-1001

Morgan, W. B., Elder, K. B., & King, E. B. (2013). The emergence and reduction of bias in letters of recommendation. Journal of Applied Social Psychology, 43(11), 2297–2306. https://doi.org/10.1111/jasp.12179

Morgan-Wall, T. (2021). rayshader: Create Maps and Visualize Data in 2D and 3D (0.24.5) [Computer software]. https://CRAN.R-project.org/package=rayshader

Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. ArXiv:2005.05909 [Cs]. http://arxiv.org/abs/2005.05909

Mulaku, G. C., & Nyadimo, E. (2011). GIS in Education Planning: The Kenyan School Mapping Project. Survey Review, 43(323), 567–578.

Municipalidad de San José, Chacayá, Sololá. (n.d.). Construcción Escuela Primaria Colonia Romec, San José Chacayá, Sololá. http://snip.segeplan.gob.gt/share/SCHE$SINIP/PLANOS_DISENOS/186461-TQBEVQZYMJ.pdf

Municipalidad de San José, Pinula. (n.d.). Construcción Escuela Primaria y Pre-Primaria Colonia Santa Sofía, Municipio De San José, Pinula, Departamento De Guatemala. https://www.guatecompras.gt/concursos/files/1241/6202896%40PERFIL%20DE%20ESCUELA%20SANTA%20SOFIA.pdf

National Liberal Arts College Rankings. (2019). US News & World Report. https://www.usnews.com/best-colleges/rankings/national-universities

National University Rankings. (2019). US News & World Report. https://www.usnews.com/best-colleges/rankings/national-universities

Navarro-Sola, L. (2019). Secondary School Expansion through Televised Lessons: The Labor Market Returns of the Mexican Telesecundaria. Working Paper. https://laianaso.github.io/laianavarrosola.com/Navarro-Sola_JMP.pdf

Nelson, L. A., Spieker, A., Greevy, R., LeStourgeon, L. M., Wallston, K. A., & Mayberry, L. S. (2020). User Engagement Among Diverse Adults in a 12-Month Text Message–Delivered Diabetes Support Intervention: Results from a Randomized Controlled Trial. JMIR mHealth and uHealth, 8(7), e17534.

Ngware, M. W., & Mutisya, M. (2021). Demystifying Privatization of Education in Sub-Saharan Africa: Do Poor Households Utilize Private Schooling because of Perceived

Quality, Distance to School, or Low Fees? Comparative Education Review, 65(1), 124–146. https://doi.org/10.1086/712090

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. International Journal of Qualitative Methods, 16(1), 160940691773384. https://doi.org/10.1177/1609406917733847

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Oreopoulos, P. & Petronijevic, U. (2018). Student coaching: How far can technology go? Journal of Human Resources, 53(2), 299-329.

O'Brien, H.L. & Toms, E.G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. Journal of the American Society for Information Science & Technology, 59(6), 938- 955. DOI: 10.1002/asi.20801.

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., & Scott-Clayton, J. (2016). Improving College Access in the United States: Barriers and Policy Responses. Economics of Education Review, 51, 4–22. https://doi.org/10.1016/j.econedurev.2016.02.009

Paluck, E. L., & Green, D. P. (2009). Prejudice Reduction: What Works? A Review and Assessment of Research and Practice. Annual Review of Psychology, 60(1), 339–367. https://doi.org/10.1146/annurev.psych.60.110707.163607

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

Papageorge, N. W., Gershenson, S., & Kang, K. M. (2018). Teacher Expectations Matter (Working Paper No. 25255). National Bureau of Economic Research. https://doi.org/10.3386/w25255

Park, J. H., Shin, J., & Fung, P. (2018). Reducing Gender Bias in Abusive Language Detection. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2799–2804. https://doi.org/10.18653/v1/D18-1302

Payne, B. K. (2006). Weapon Bias: Split-Second Decisions and Unintended Stereotyping. Current Directions in Psychological Science, 15(6), 287–291. https://doi.org/10.1111/j.1467-8721.2006.00454.x

Pebesma, E., Bivand, R., Racine, E., Sumner, M., Cook, I., Keitt, T., Lovelace, R., Wickham, H., Ooms, J., Müller, K., Pedersen, T. L., & Baston, D. (2021). sf: Simple Features for R (0.9-8) [Computer software]. https://CRAN.R-project.org/package=sf

Penner, E. K., Rochmes, J., Liu, J., Solanki, S. M., & Loeb, S. (2019). Differing Views of Equity: How Prospective Educators Perceive Their Role in Closing Achievement Gaps. The Russell Sage Foundation Journal of the Social Sciences, 5(3), 103–127. https://doi.org/10.7758/RSF.2019.5.3.06

Pezzulo, C., Hornby, G. M., Sorichetta, A., Gaughan, A. E., Linard, C., Bird, T. J., Kerr, D., Lloyd, C. T., & Tatem, A. J. (2017). Sub-national mapping of population pyramids and dependency ratios in Africa and Asia. Scientific Data, 4(1), 170089. https://doi.org/10.1038/sdata.2017.89

Psihogios, A. M., Li, Y., Butler, E., Hamilton, J., Daniel, L. C., Barakat, L. P., ... & Schwartz, L. A. (2019). Text message responsivity in a 2-way short message service pilot intervention with adolescent and young adult survivors of cancer. JMIR mHealth and uHealth, 7(4), e12547.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. American Journal of Political Science, 54(1), 209–228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv:1910.10683 [Cs, Stat]. http://arxiv.org/abs/1910.10683

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., Jang, H., Chavez, B, Buontempo, J., & DiSalvo, R. (2019). Stanford Education Data Archive (Version 3.0). Stanford University. http://purl.stanford.edu/db586ns4974

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. Journal of Statistical Software, 91(2). https://doi.org/10.18637/jss.v091.i02

Rorem, A. (2015, January 20). America's College Promise in Virginia. Stat Chat - A Web Series from the University of Virginia Weldon Cooper Center for Public Service Demographics Research Group. http://statchatva.org/2015/01/20/americas-college-promise-in-virginia/

Rosinger, K. O., Ford, K. S., & Choi, J. (2020). The Role of Selective College Admissions Criteria in Interrupting or Reproducing Racial and Economic Inequities. The Journal of Higher Education, 0(0), 1–25. https://doi.org/10.1080/00221546.2020.1795504

Sang, E. F. T. K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. ArXiv:Cs/0306050. http://arxiv.org/abs/cs/0306050

Schaeffer, B. (2020, August 12). Three-Fifths of Four-Year Colleges and Universities Are Test-Optional for Fall 2021 Admission; Total of Schools Not Requiring ACT/SAT Exceeds 1,450. FairTest: The National Center for Fair and Open Testing. https://www.fairtest.org/threefifths-fouryear-colleges-and-universities-are

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A Linguistic Comparison of Letters of Recommendation for Male and Female Chemistry and Biochemistry Job Applicants. Sex Roles, 57(7), 509–514. https://doi.org/10.1007/s11199-007-9291-4

Schwarz, J. D. (2016). Lost in Translation: Elite College Admission and High School Differences in Letters of Recommendation [Dissertation Manuscript].

Scott-Clayton, J. (2016). Early Labor Market and Debt Outcomes for Bachelor's Degree Recipients: Heterogeneity by Institution Type and Major, and Trends Over Time (CAPSEE Working Papers, p. 38). Center for Analysis of Postsecondary Education and Employment. http://ccrc.tc.columbia.edu/media/k2/attachments/early-labor-market-debt-outcomes-bachelors-recipients.pdf

SEGEPLAN. (n.d.) Descargas SINIT: Escuelas de Guatemala (MINEDUC). [Data set]. http://ide.segeplan.gob.gt/descargas.php

Shapiro, D., Ryu, M., Huie, F., & Liu, Q. (October 2019). Some College, No Degree , a 2019 Snapshot for the Nation and 50 States, Signature Report No. 17, Herdon, VA: National Student Clearinghouse Research Center

Smythe-Leistico, K., & Page, L. C. (2018). Connect-text: Leveraging text-message communication to mitigate chronic absenteeism and improve parental engagement in the earliest years of schooling. Journal of Education for Students Placed at Risk (JESPAR), 23(1-2), 139-152.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 1631–1642. https://www.aclweb.org/anthology/D13-1170

Starck, J. G., Riddle, T., Sinclair, S., & Warikoo, N. (2020). Teachers Are People Too: Examining the Racial Bias of Teachers Compared to Other American Adults: Educational Researcher. https://doi.org/10.3102/0013189X20912758

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. PLOS ONE, 10(2), e0107042. https://doi.org/10.1371/journal.pone.0107042

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over Many Models and Many Topics. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 952–961. http://dl.acm.org/citation.cfm?id=2390948.2391052

Sullivan, Z., Castleman, B., & Bettinger, E. (2019). College advising at a national scale: Experimental evidence from the CollegePoint initiative. EdWorkingPaper No. 19-123. Retrieved from https://edworkingpapers.com/index.php/ai19-123

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. ArXiv:1906.08976 [Cs]. http://arxiv.org/abs/1906.08976

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics, 37(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology, 29(1), 24–54. https://doi.org/10.1177/0261927X09351676

Theunynck, S. (2009). School Construction Strategies for Universal Primary Education in Africa: Should Communities Be Empowered to Build Their Schools? The World Bank. https://doi.org/10.1596/973-0-8213-7720-8

Tiecke, T. G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., Kilic, T., Murray, S., Blankespoor, B., Prydz, E. B., & Dang, H.-A. H. (2017). Mapping the world population one building at a time. ArXiv:1712.05839 [Cs]. http://arxiv.org/abs/1712.05839

Tobler, W. (1993). "Three Presentations on Geographical Analysis and Modeling." URL http://www.ncgia.ucsb.edu/Publications/Tech_Reports/93/93-1.PDF.

United Nations Educational, Scientific and Cultural Organization Institute for Statistics. (2019). New Methodology Shows that 258 Million Children, Adolescents and Youth Are Out of School (UIS/2019/ED/FS/56; Fact Sheet). http://uis.unesco.org/sites/default/files/documents/new-methodology-shows-258-million-children-adolescents-and-youth-are-out-school.pdf

USDA ERS - Commuting Zones and Labor Market Areas. (2019, March 26). USDA Economic Research Service. https://www.ers.usda.gov/data-products/commuting-zones-and-labor-market-areas/documentation/

van Aken, B., Winter, B., Löser, A., & Gers, F. A. (2019). How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations. Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 1823–1832. https://doi.org/10.1145/3357384.3358028

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. ArXiv:1706.03762 [Cs]. http://arxiv.org/abs/1706.03762

Voorend, K., Anker, R., & Anker, M. (2018). Living Wage Report: Guatemala (Series 1, Report 16). Global Living Wage Coalition.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

World Bank, World Development Indicators. (2008). School enrollment, primary (% net) [Data file]. Retrieved from https://data.worldbank.org/indicator/SE.PRM.NENR

World Bank, World Development Indicators. (2016). School enrollment, primary (% net) [Data file]. Retrieved from https://data.worldbank.org/indicator/SE.PRM.NENR

World Bank, World Development Indicators. (2017b). School enrollment, primary (% net) [Data file]. Retrieved from https://data.worldbank.org/indicator/SE.PRM.NENR

World Bank, World Development Indicators. (2019b). School enrollment, primary, private (% of total primary) [Data file]. Retrieved from https://data.worldbank.org/indicator/SE.PRM.PRIV.ZS

World Bank. (2017a). World Development Report 2018: Learning to Realize Education's Promise. The World Bank. https://doi.org/10.1596/978-1-4648-1096-1

World Bank. (2019a). Guatemala: Learning Poverty Brief. http://pubdocs.worldbank.org/en/640231571223409894/LAC-LCC2C-GTM-LPBRIEF.pdf

WorldPop (www.worldpop.org - School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur) and Center for International Earth Science Information Network (CIESIN), Columbia University (2018). Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation (OPP1134076). https://dx.doi.org/10.5258/SOTON/WP00670

Wu, A. (2017). Gender Stereotype in Academia: Evidence from Economics Job Market Rumors Forum (No. 2017–09; Working Papers). Princeton University, Woodrow Wilson School of Public and International Affairs, Center for Health and Wellbeing. https://ideas.repec.org/p/pri/cheawb/2017-09.html

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). XLNet: Generalized Autoregressive Pretraining for Language Understanding. ArXiv:1906.08237 [Cs]. http://arxiv.org/abs/1906.08237

Zhang, S., Hamburger, E., Kahanda, S., Lyttle, M., Williams, R., & Jaser, S. S. (2018). Engagement with a text-messaging intervention improves adherence in adolescents

with type 1 diabetes: brief report. Diabetes technology & therapeutics, 20(5), 386-389.

Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (1st edition). PublicAffairs.

APPENDICES

Appendix A1.1: Other Contexts and Cross-National Comparisons

Our main methodology is primarily designed to identify areas *within* a given region where physical access to education is limited. However, we use this appendix to (1) demonstrate the portability of our analysis, and (2) illustrate some of the considerations when extending the analysis to example multiple countries by adding similar analyses for four Sub-Saharan African countries and two Latin American countries: Tanzania (in 2016), Rwanda (2012), South Africa (2020), Kenya (2018), Peru (2020), and Costa Rica (2020).

We observe two main benefits to cross-country analyses. First, applying this methodology to other contexts allows analysts to create potentially informative benchmarks for a given region of interest. For example, we report in the main narrative that 95% of the population in Guatemala lives within 3 km of a public primary school. In a vacuum, this number is not too informative. But when coupled with distance norms, policy goals, and statistics from peer countries, this can serve as a meaningful data point of comparison. In the case of this metric, Guatemala performs better than all other countries analyzed in Table A1.1 except for Costa Rica. Second, this type of comparison can moreover facilitate a rough classification for countries in terms of the issues they face with enrollment. In an ideal world, countries would have high enrollment rates and a low prevalence of education deserts, like Peru and Costa Rica in the table. Deviations from this categorization can offer a useful shorthand for thinking about extant enrollment barriers. For instance, Guatemala and South Africa can be thought of as having relatively low desert prevalence and low enrollment, while

Rwanda can be thought of as having relatively high enrollment *in spite of* high desert prevalence– indicating countries where distance may not be the primary issue for enrollment. We can moreover examine countries where desert prevalence is high while enrollment is low – perhaps contexts where deserts are more impactful – like Tanzania, with 4 in 10 people living further than 3 km from a public primary school and deserts pervasive throughout the country.

We also want to highlight that there are clear challenges in the cross-country comparison of our analyses. First, while the data-generating process for the *population* data is fairly uniform across countries, the data-generating process for *school* data can vary meaningfully by country. As we allude to in Section III above, what qualifies as a "public" school may vary across contexts (e.g., is it only schools run by governments, or does it also include privately-run government schools?), as well as what qualifies as a "primary" school (e.g., if the grades covered in primary schools differ by location). Similarly, the data collection capabilities of governments may vary, and the degree of missingness for geo-locations can differ as well.

Finally, differences in the actual geographic distribution of a country's population can also affect the usefulness of cross-country comparisons. Costa Rica, where ~45% of the overall population lives in an extended capital area of only about 2000 km^2 (*Gran Área Metropolitana*), is arguably incomparable to a largely rural context like Tanzania, where the most populous metropolitan area (Dar es Salaam) houses only 11% of its population, and the next-largest city only has about a fifth of this number (Mwanza). This non-exhaustive list of contextual factors can lead to shortcomings in cross-country comparisons in results derived from the methodology we proposed, and as such, these comparisons should be made carefully and sparingly, if at all.

Appendix A2.1: Sentiment Analysis Introduction and Methodology

Sentiment analysis is a common NLP task in which analysts use an algorithm to "read" a given text string and rate the string as containing/expressing positive, negative, or neutral sentiment (Pang & Lee, 2008). This task is especially common in commercial applications (e.g. analyzing consumer sentiment towards your product by analyzing tweets) but is becoming more pervasive in the field of education research as well (Fesler et al., 2019). One can think of this process as generating output similar to hand-coded qualitative analysis, but in an automated and highly scalable way that facilitates quantitative analysis.

There are a wide variety of techniques that data scientists use in this pursuit, but the recent NLP literature has coalesced around complicated neural network algorithms known as "transformers" (Vaswani et al., 2017). These transformer algorithms perform substantially better than previous sentiment analysis approaches because the transformer's specific architecture allows it to better account for the complexities of word context (e.g. that "I wish I were happy" actually indicates sadness), multiple word meanings (e.g. that "bank" has two separate meanings in "The river bank was wet" and "I went to the bank this morning"), and informalities (e.g. "that was sick, dude;" Ambartsoumian & Popowich, 2018).

In brief, a transformer neural network is a neural network algorithm that has been fed immense volumes of text data (such as aggregated Wikipedia articles, novels, and news articles) to generate a nuanced statistical model describing how words are put together into sentences - called a "language model." This can be thought of as giving the algorithm a generalized understanding of grammar, syntax, vocabulary, and word relationships by example. For instance, it will have seen thousands of examples of "...it is hot outside…" in its training data, but likely no examples of "...outside it hot is..." nor "...clam hot it

outside...", teaching it what combinations and sequences of words are considered valid. Despite the bluntness of this approach, it is so effective at capturing complex idiosyncrasies within language that it now drives some of the most advanced and widely-used grammar checking engines (e.g. Grammarly; Alikaniotis & Raheja, 2019).

Once that language modeling process is complete, analysts then "fine-tune" the algorithm to perform a more specific task, such as sentiment analysis, using a traditional supervised machine learning framework (i.e. provide the algorithm a set of example texts with ground-truth sentiment scores so that it can optimize for accurate scoring on its own). The motivation behind separating the language modeling task from the classification task is somewhat analogous to the idea that it is easier to teach someone to play a new sport when they already have a good grasp of basic physics, fitness, and competition, versus starting from a completely blank slate. Similarly, because the transformer is well-trained in general language, it can leverage this understanding to better approach new language-based tasks afterward.

The unique contribution of the transformer architecture is a mechanism called "attention" that allows it to more effectively process longer strings of text at once by weighting words according to their functional importance in the text (e.g. using a subject introduced two paragraphs earlier to interpret a referential statement in the sentence at hand). Algorithms using the transformer architecture have literally revolutionized the landscape of NLP, pushing the state-of-the-art for model performance on nearly every single performance task and benchmark, including sentiment analysis (Devlin et al., 2019).

Choosing which transformer to use for a sentiment analysis task is a highly consequential decision. While sharing the same general principles, transformers vary due to different training datasets, different underlying mechanics and optimization processes, and

different end-applications in mind. Ideally, we would be able to find a robust transformer that is trained on text data similar to ours so that we could be more assured of its appropriateness for our context.[66] Lacking that, we have opted to employ an "ensemble" approach that combines several of these transformers together. More concretely, we use five pre-existing transformer algorithms to produce sentiment analysis scores for every text message in our data, and then combine these separate classifications together in a data-driven manner using a random forest algorithm to produce a final sentiment score. By the end of this process, each text sent and received during the intervention is assigned a sentiment score from -2 to 2, corresponding with very negative, negative, neutral, positive, or very positive sentiment.

This approach is attractive because it leverages the unique strengths and insights of each of these separate models while mitigating some of their potentially problematic idiosyncrasies - the intuition here being that if each model weighs different considerations in its individual decision, they can each contribute valuably distinct insights to be incorporated into the final model. For a more detailed discussion of our constituent models, the model construction process, and performance benchmarks, please see Appendix A2.2. In sum, we find that our ensemble model matches current state-of-the-art performance on the most common sentiment analysis benchmark, the Stanford Sentiment Treebank (SST) test (Socher et al., 2013).

We operationalize the definition of sentiment for the present study as the *perceived positivity of emotions and ideas present in a given text*. This definition then is a conglomeration of

---

[66] While it would be conceptually attractive to "fine-tune" our own transformer model to best account for our educational context, training these models is both logistically and computationally complex. The authors are exploring this opportunity for related work going forward.

the speaker's stated emotions ("I feel sad" v. "I am excited"), communicated intention ("I hope you die" v. "I wish you the best!"), and, at least to some extent, topical content ("financial hardship" v. "vacation time"). Note that this definition is complicated for longer strings of text, in which multiple emotions/implications may be present and the overall sentiment becomes ambiguous.[67]

We argue this definition is appropriate for our context because of the nature of the N2FL text data: texts were generally only one or two sentences long (making the ambiguity of sentiment in long text strings less problematic), students encounter these texts "as-is" (e.g. they are not transcribed spoken words with greater context than what we observe), and their perception of a text's sentiment is likely driven by a combination of factors (e.g. stated emotions, communicated intention, topical content).

To provide evidence for the validity of the algorithm output and its concordance with our definition of sentiment, we conduct two validation exercises:

1. Pull a random sample of N2FL texts, have human coders briefed in the construct definition manually classify each text, and then compare the human codes against the algorithm codes using traditional accuracy statistics (with human codes set as the ground-truth).
2. Pull a random sample of N2FL texts (distinct from the sample constructed in exercise 1) alongside their assigned sentiment scores from the algorithm, have human

[67] While we would like to lean on a more standardized definition, we were unable to find a detailed and widely-accepted definition for sentiment in transformer-based models in the literature. Interestingly, sentiment as a construct across modern data science (i.e. neural network-based models rather than dictionary models) is almost entirely dependent on the SST's definition due to the strong incentive for data scientists to optimize their algorithm's SST performance for benchmarking purposes. That said, the SST intentionally encouraged their human coders to view sentiment as a flexible and subjective notion, making a formal definition elusive.

coders briefed in the construct definition approve or disapprove of each pairing's accuracy, and calculate overall and class-by-class approval rates.

A summary of the validation exercise results are displayed in Table A2.2, and more details on each procedure can be found in Appendix A2.3. Note that these exercises were conducted using only one human coder for now (a coauthor on this paper); we plan to expand this process to multiple trained coders in a later draft to improve robustness.

We find that our algorithm performs about as well in terms of perfect accuracy measures on the N2FL data as it does on the benchmark SST5 at 56%. This is high by sentiment analysis standards, but still suggests our analysis will suffer from measurement error. In addition, if we consider the N2FL text data a "test" dataset per supervised machine learning frameworks, the comparable performance of our algorithm on both N2FL data and the SST data may suggest: **(1)** that our algorithm was not reliant on idiosyncrasies specific to the SST data, and it is actually reading some true, generalizable signal within the text data to inform its classifications, and/or **(2)** that the N2FL text data may not be substantially different from the SST text data despite the difference in contexts and sources. This in mind, we now have evidence that using models trained on the SST rather than text data closer to the circumstances of N2FL is appropriate for our purposes.

Because the algorithm outputs a predicted probability of each possible sentiment score classification, we can also run diagnostics on the relative confidence of each of its predictions. For example, we might be more skeptical of the algorithm's sentiment classifications if it's torn between two very likely options (e.g. assigning a score of -2 with 40% probability, and a score of -1 with 38% probability) versus if it's selecting one classification with high certainty (e.g. assigning a score of 2 with 76% probability). We find in general that the algorithm is fairly confident in its classifications, and that there are few

"close calls" among the N2FL texts. We dissect the results of this diagnostic test in more detail in Appendix A2.4.

Finally, we have plentiful evidence from NLP bias research that gendered names and pronouns can systematically skew the results of text classification algorithms because these algorithms are trained on datasets that implicitly contain the biases of societal writing more broadly (Park et al., 2018; Sun et al., 2019). For example, the algorithm may interpret "She is assertive" as negative, but "He is assertive" as positive; similarly, the algorithm may interpret "Jane is assertive" as negative, but "Joe is assertive" as positive. Our text data removed names for de-identification purposes (replaced with a token stand-in, "advisorname"), making gendered names a non-issue. Moreover, because students and advisors most often spoke in the first- and second-person (I/you/we), we find exceptionally low prevalence of gendered pronouns (he/him/his/she/her/hers) in our data: out of 27,942 unique texts, only 552 (2%) contained any gendered pronoun. While we cannot rule out residual gender and racial bias in our algorithm, we have good reason to believe their impact on our analysis is negligible after these processing steps given our data context.

We also directly compare the output of the algorithm before *and* after replacing any gendered pronouns ("masking" the data) for exploratory purposes (Table A2.3) and find that only 3% of texts change their sentiment scores at all. The overwhelming majority that do change classifications vary only slightly from masked to unmasked datasets. We intend to include more in-depth analyses of any classification inconsistencies here in a future draft.

Appendix A2.2: Ensemble Model Construction

As mentioned in the prior section, our algorithm is what we refer to as an "ensemble" method. This approach involves training several models to conduct the same

task, and then utilizing each of those models' output as inputs into a final model that considers each of these models' output in a final classification. This is akin to gathering a panel of experts on an issue and making a decision based on their combined recommendations. While each expert may see the same data and evidence, the variance in their interpretations may lead to importantly different conclusions worth considering.

The majority of our constituent models are built using transformer neural networks (BERT, ALBERT, XLNet, T5, RoBERTa) as described in the main narrative. In Table A2.4, we provide a rough breakdown of each of these algorithms in terms of their language model training data, task training data (i.e. sentiment analysis training data), and benchmark performance. These details are important to keep in mind as we interpret the results of our algorithm - the language model training data tells us what contexts it learned its general understanding from, and the task training data tells us what contexts it learned to classify sentiment from. For example, the BERT model we utilize was trained specifically on Yelp restaurant reviews for sentiment classification - a context where even a "lukewarm" sentence may really correspond with a quite negative sentiment score.

In our process, we have each algorithm classify each text and provide its calculated probabilities for each possible classification (e.g. 52% probability of a very negative sentiment, 30% of negative, 12% of neutral, 6% of positive, and 0% of very positive; these scores will always sum to 100%). We then use these outputs as inputs into the random forest classifier, which is finally trained using the Stanford Sentiment Treebank, 5-class set.

Our algorithm has a **base accuracy score** on the **SST5** of 55% (proportion of perfect classifications). This is tied for the current state-of-the-art across all NLP research to date. In Figure A2.1, I display the accuracy diagnostics of our random forest algorithm on the SST5 test set. The confusion matrix is really the key figure; if the algorithm performed

perfectly, we would see all observations would fall into the diagonal cells. Note that as is common for these fine-grain sentiment analysis classifications, our model performs noticeably less well at detecting neutrality in these data.

Another common way of assessing accuracy on these 5-class sentiment scores is to consider the "one-off accuracy" - or, what proportion of cases the score is only one point off of the true score. This is a good way to gauge how far off the algorithm is when it provides an incorrect classification. In our case, we have a one-off accuracy rate of 96%. In other words, even when our algorithm is wrong for the exact classification, it's not off by much (i.e. it is not seeing a very positive text and calling it very negative, or even neutral).

Yet another way of slicing performance is by thinking only of "valence" (emotional direction) without magnitude (e.g. combine negative and very negative scores into just a single negative category). Using the standard Stanford Sentiment Treebank 3-class test, we achieve a base accuracy score of 76%. This is a less common task in the most recent wave of NLP research, and so it is unclear how this performs relative to the state of the art. For reference, Stanford's Stanza model (which is a constituent model of ours) achieves an accuracy of 70%. Figure A2.2 displays the accuracy diagnostics on the 3-class set.

Appendix A2.3: Validating Algorithm Classification Output

In brief, we conduct our validation exercises according to the following procedure:

1. Comparing human-coded texts to algorithmically-coded texts
    a. Pull a random sample of 5 unique N2FL texts from each crossed group defined by the message type (scheduled text from advisor, personalized text from advisor, text from student) and algorithm sentiment score (e.g. sample from those given a score of -2, then from those given a score of -1, etc.), for

a total of 75 texts. Sample an additional 75 texts completely at random, for a total of 150 texts.

b. Brief human coders in our operationalized definition, debriefing texts shown in Table 2.2 in the main text. For now, we have only one human coder, but intend to expand this group to at least 5 in a future draft.

c. Show human coders the texts in random order, **without** the algorithm sentiment scores attached, and ask them to rate each text from -2 to 2 per our sentiment definition. If they were unsure, they were asked to still provide their best guess.

d. Consolidate human-coded texts by taking the average score, rounded to the nearest integer value.

e. Compare the scores given by the consensus of human-coders against the algorithm's scores overall and by each grouping of algorithm-coded sentiment score.

2. Auditing algorithmically-coded texts for human approval

a. Create a random sample exactly as described for exercise 1, but with a different random seed (such that there may be overlap between the two samples, but they are generated totally independently)

b. Brief human coders in our operationalized definition, debriefing texts shown in Table 2.2 in the main text. For now, we have only one human coder, but intend to expand this group to at least 5 in a future draft.

c. Show human coders the texts in random order, **with** the algorithm sentiment scores attached, and ask them to approve or disapprove (binary) of each

text's score per our sentiment definition. If they were unsure, they were asked to still provide their best guess.

d. Consolidate human-coded texts by taking the raw average of approval ratings across coders (0 is disapprove, 1 is approve)
e. Calculate approval rates overall and by each grouping of algorithm-coded sentiment score.

Appendix A2.4: Assessing the Confidence of Random Forest Sentiment Scores

As mentioned in the narrative above, an important diagnostic to evaluate here is how "certain" the algorithm is when making a decision. Our ensemble outputs how likely it thinks *each* possible sentiment score is for a given sentence. Figure A2.3 shows the associated probability of each of the *final* classifications the algorithm has provided; roughly, how confident it was in each individual classification. We see that the mean sits around 60%, meaning the algorithm is quite certain. We'd be worried if the mean were closer to 20-30% (given that complete uncertainty would produce a 20% probability across each of the five possible classifications).

We can also compare the likelihoods of the algorithm's first- and second-choice sentiment classification to see how close the two are. The closer they are, the harder a time the algorithm is having while picking between its best options. Figure A2.4 plots the *difference* between the probability of the algorithm's first-choice and the second-choice predictions across N2FL texts. The mean and median difference is approximately 34 percentage points, indicating that for the majority of texts, the algorithm is quite certain that its final classification is by far the best one. However, there is still a substantial volume of texts

where the difference is negligible, and we may consider handling those predictions differently than the others in future drafts for this reason.

Appendix A2.5: Topic Modeling Introduction and Methodology

Topic modeling is another common task in NLP in which analysts use an algorithm to "learn" what discrete topics of discussion exist across a series of documents and then measure how prevalent each topic is within each document. Ultimately, we are interested in whether there exists variation in the prevalence of these topics across students. Such variation would potentially reflect substantively different advising interaction content and arguably different "treatments" as a result.

To accomplish this, the standard topic modeling algorithm takes a large body of documents (in our case, text conversations) and attempts to analyze those documents for groupings of words that frequently co-occur together in the same documents (Blei, 2003). Words that frequently co-occur in this framework are thought to belong to the same abstract topic of discussion, and word groups identified by the algorithm in this way can then be interpreted by analysts for its substantive meaning. For example, the words "financial," "aid," and "loans" may all occur together at high frequency across various student-advisor conversations; the topic modeling algorithm would group these words together under one "topic," which could be interpreted by an analyst as the abstract topic of "financial aid information." The algorithm identifies several such topics based on the provided text, which allows for uniquely context-sensitive and flexible output compared to similar content analysis methods that use pre-defined word groups (e.g. the Linguistic Inquiry and Word Count method per Tausczik & Pennebaker, 2010). Once topics are identified, the algorithm can determine the prevalence of each topic in each conversation based on the combination

of words in that conversation. For example, the algorithm can tell us how many words in a given conversation are spent discussing "financial aid information" as a topic, versus another topic that it discovers such as "student sports involvement."

For the present study, we opt to utilize the Structural Topic Model implementation proposed by Roberts et al. (2019). This methodology offers a refinement on the traditional topic modeling approach in a variety of ways, but most importantly for our purposes, it allows us to specify topical prevalence covariates as part of the topic modeling process. Put simply, this feature lets the topic model discover associations between the provided covariates and the prevalence of each topic to inform model fit. For example, if students who are older systematically discuss financial aid more often than students who are younger, the structural topic model can pick up on this and form more accurate expectations of how prevalent that topic is for all older students' conversations.[68] For our model, we use the following student baseline covariates for our topical prevalence covariates: institution, state, sex, race/ethnicity, age (over/under sample median of 23 years old), BA transfer intention, and prior transfer status.

There are three main challenges with regards to robustness and usability in topic modeling output. First, topic modeling is highly sensitive to the structure of the raw data and produces poor results when the length of the text documents are **(a)** too short (a good rule of thumb in practice is to use documents about the length of a paragraph), and **(b)** when the text documents simultaneously cover too many topics. In our case, single messages alone are

---

[68] In a more technical sense, the entirety of the topic modeling process is rooted in Bayesian frameworks, such that a word is really assigned a *probability* of being "about" each topic, and each conversation is assigned a *probability* of discussing each topic. Topical prevalence covariates allows the model to adjust a conversation's *prior* distribution of discussing each topic based on those covariates, which then shapes its *posterior* distribution of discussing each topic given the words within it and that prior distribution.

likely too short a document size, while compiling all messages sent between a student and their advisor together likely covers too many distinct topics at once. As such, we chose to structure the data at the *conversation* level, defined as any messages sent by either the student or their advisor after a scheduled message, but before the next scheduled message. Given that we have good reason to expect the group of messages following a scheduled message would be related in content (e.g., a scheduled message about the FAFSA is likely followed by a student-advisor conversation about financial aid), and that grouping messages in this way would increase the length of each text document, condensing messages to the conversation level nicely addresses both of these concerns. We moreover restricted our training set to only personalized messages to prevent the content of scheduled messages from having undue influence on topic formation.[69] Thus, our topic modeling training set includes only conversations from students who responded to at least one scheduled message, for a total of 16,828 unique conversations with an average length of 19 "keywords."

Second, topic modeling results are highly sensitive to the number of topics the algorithm is asked by the analyst to identify, known as the K parameter. Standard practice in the field is to run the topic model several times while arbitrarily changing K across a wide range of values, and then selecting the final model's K parameter based on a variety of model fit metrics (Roberts et al., 2019). We test every multiple of 5 up to 55 for our K parameter and find that K=30 produces the greatest balance of general model fit (held-out likelihood and model residuals) against an algorithmically-derived measure of word

---

[69] To pre-process the text, we also **(1)** removed any stopwords, numbers, URLs, and other non-English language, **(2)** replaced proper nouns with stand-in tokens (e.g. "tokensystemname" instead of "John Jay") to increase language uniformity across contexts, and **(3)** manually spell-checked and aligned word/verb forms for the 3000 most common words across the dataset. We moreover include both unigrams (single words) and bigrams (common two word pairs) in our model, and exclude any tokens that occur in fewer than 10 (for unigrams) or 20 (for bigrams) documents to reduce the sparsity of the model.

coherence within topics (semantic coherence, per Mimno et al., 2011). The results of our specification test are visualized in Figure A2.5.

Lastly, the topic interpretation process can be highly subjective given that it is up to analysts to determine the substantive meaning (if any) in the topic groupings. To address this directly, we set up a multi-stage, multi-coder process to interpret the topic groupings inspired by Penner et al., 2019. We began by providing three coders (each of the co-authors) with a list of the words that most distinguished each of the 30 topics.[70] Each coder was asked to identify a unifying idea or concept for each topic to the best of their ability. Once each coder completed this process independently, we reviewed any disagreements and discussed how to harmonize these interpretations collaboratively. We found that all three coders had **perfect or near-perfect agreement on 26 out of 30 topic interpretations**, with the remaining 4 showing only minor disagreement. Complete harmonization tables and process documents are available upon request. Table A2.5 displays example topics with perfect agreement, near perfect agreement, and minor disagreement among coders. Moreover, the 4 minor disagreements were easily resolved with brief clarification of the task and our written interpretations, and were ultimately inconsequential given our next step of combining related topics into larger "supertopics."

Following the interpretation of individual topics, each coder then individually combined topics into broader supertopics to reduce the dimensionality of our topic model

---

[70] Coders received both the highest probability and highest frequency/exclusivity (FREX) words for each topic. The highest probability words are those that the algorithm thinks are most likely to belong to a given topic when they appear (i.e. the strongest indicators for a topic being discussed). The highest FREX words are words that are both highly frequent, *and* highly exclusive to that particular topic, in that they don't tend to appear in other topic groupings (Airoldi & Bischof, 2016). Balancing exclusivity with frequency is important to focus on words that matter in the documents; terms with high exclusivity but low frequency tend to have very little impact on the algorithm overall, and tend to be highly noisy.

output and improve our ability to relate our output to substantively relevant advising practices (e.g. share of conversations focusing on financial aid versus course registration detail).[71] In a similar manner to our initial topic interpretation process, we then harmonized the supertopic groupings across each of the coders' proposed schemes. Our harmonized supertopics are displayed in Table A2.6. Note that four of the 30 topics were not ultimately grouped into an actual supertopic for analysis due to their lack of substantively relevant meaning (e.g. pleasantries like "hey, i'll, glad, yeah, awesome, haha, alright" or more basic communication logistics like "email, student id, check, text, stop" etc.).

Figure A2.6 displays the prevalence of each underlying topic, and its corresponding supertopic, in terms of word frequency within the training dataset (personalized messages between students and advisors, collapsed to the conversation level). Note that the process of deriving the number of words in each conversation that come from each topic is probabilistic in nature. That is, because a single word can belong to multiple topics at once (e.g. "deadline" might appear in financial aid and in course enrollment discussions) at varying probabilities (perhaps it is more common in financial aid than course enrollment), the algorithm will use these probabilities to assign it to a topic each time the word appears. The algorithm runs many simulations given these parameters and the input text, and the output topic assignments are the modal value from the distribution of those words to topics across simulations.

To summarize, our topic modeling process allows us to estimate, for each student, the share of their conversation focused on each supertopic of conversation: **(a)** financial aid,

---

[71] Mathematically speaking, we are considering the probability that a given document discusses each supertopic a shorthand for "discussing topic A, OR discussing topic B, OR discussing topic C," etc. Thus, the probabilities that each document discusses each topic are summed to the probabilities that each document discusses each supertopic instead.

(b) advising meeting scheduling logistics, **(c)** course registration and enrollment, **(d)** broader academic planning, and **(e)** academic resources. While there is some overlap and close relationship between these supertopics in concept (e.g. some advising meetings are likely set up to discuss financial aid, or course registration, etc.), we argue that these present clearly delineated characterizations about the content of the text messages themselves and allow us to credibly characterize trends and variation in texting patterns across students as a result.

Appendix A3.1: Word Frequency Analysis Methodology and Analysis

(a) Word Frequency Analysis is a straightforward endeavor where the occurrence of each word in each letter is counted and trends in individual word usage across letters can then be explored. While there are a number of ways to both preprocess the text and analyze the resulting frequencies, I opt for as parsimonious an approach as possible given that this is a fairly exploratory analysis intended only to examine the text of letters in a more "raw" format compared to the other NLP methods I apply here.

First, I use the same text pre-processing steps as I do in the topic modeling analysis. To summarize, I remove "stopwords" that generally do not convey substantive meaning in and of themselves (e.g., "the," "it," "and"). After removing these words from consideration, I also remove numbers and any explicitly gendered references to individual students (e.g., "ms," "mr," "gentleman," "lady," "boy," etc.).[72] From the remaining words, I construct a list of all words *and* two-word phrases (also referred to as "bigrams") to construct a combined "vocabulary" of interest for the word frequency analysis. I also remove what I refer to as

[72] Note that I attempt only to remove words likely to refer to the student themselves. Thus, "Women's" or "Men's" remains in the vocabulary given that they are likely to be referring to things like sports (e.g., "Women's Field Hockey"), as well as other gendered words unlikely to refer directly to students (e.g., teachers do not tend to use "male" or "female" to directly describe students per my investigations).

contextual stopwords – words that are so common across letters as to be uninformative. To do this, I remove words that appear in greater than 20% of all paragraphs in the training text dataset (e.g., "school," "student," "class"). As the only deviation from my topic modeling pipeline, I restrict the vocabulary of consideration to only the 1000 words that appear in the most letter paragraphs. This is to remove the more idiosyncratic words from consideration that may nonetheless still be relevant for topic modeling.

After completing these text pre-processing steps, I replicate the analysis of Wu (2017) to identify the remaining words and phrases that are most *uniquely* used for each demographic. This process involves setting the binary indicators for each race/gender group as the outcome of a LASSO-regularized logistic regression on the vector of word/phrase counts in each paragraph, without additional controls.[73] The ensuing coefficients for each word/phrase then indicate how predictive each word is of each student demographic; put another way, words with the highest coefficients are most predictive of a student being of that demographic when they appear in a given paragraph. I also examine the inverse – which words are most predictive of a student *not* being of that demographic when they appear in a given paragraph. This allows me to directly explore the extent to which students of varying demographics are written about using different words by teachers – a potentially major component of implicit biases as surfaced in my literature review.

As with the main text, I focus on gender and race groups: Female/Male, White, Black, and Asian. Because the results of this analysis are extremely high-dimensional, I

[73] For this process, I use the implementation offered by the R package quanteda (Benoit et al., 2018) specifically designed for this exact style of inquiry. This approach leverages a 10-fold cross validation focused on minimizing deviance to determine an optimal shrinkage coefficient, and this process is conducted separately for each demographic variable.

restrict my reported results to only those 20 most positively and negatively predictive of the given demographic. Full results are available upon request.

Table A3.1 displays the results of this word frequency analysis for female students; as this is a binary indicator, these results are identical for male students but reversed in direction. Looking at the most positively predictive words/phrases, we see that many of these phrases align with prevailing stereotypes about female students like "compassion," "caring," and "shy." On the opposite end, the most negatively predictive words include "humor," "personable," and "respectful." Interestingly, we also see mentions of certain activities and subjects – "dance," "art," "psychology," and "medical" – are more prevalent among female students, while "computer," "engineering," "technology," and "economics" are more prevalent among male students. These activity and subject differences also corroborate my topic modeling results showing that STEM topics are far less common among female students than male students.

Table A3.2 displays the words/phrases that are most positively and negatively predictive of White students. Interestingly, the word use here reveals equally stark differences in the sorts of activities and subjects teachers discuss for White students versus other students. Most notable in these trends is the positive predictive quality of sports-related terms like athlete, athletics, varsity, and coach, as well as subject-related terms like social studies, AP English, honors, and musical. Conversely, subjects and activities like Economics, English Language (i.e., English Language Learners), and tutoring are all negatively predictive of being White.

Table A3.3 shows the word frequency analysis results for Black students. We see a strong positively predictive value for sports like basketball, football, and track, but also a strong negatively predictive value for cross country (here appearing as two separate words),

team, competitive, and coach, as well as courses like calculus, physics, and other AP STEM classes. These results similarly bear out the results of the topic modeling analysis, where I observed lower prevalence of sports and STEM topics. Also of note are some surprising words like church, scholarship, GPA, and "pleasure teaching," all of which suggest that these more specific phrases are brought up disproportionately for Black students.

Lastly, Table A3.4 displays results for Asian students. The top 7 most negatively predictive words are all sports related, again bearing out the substantially lower incidence of Sports topics in the topic modeling results for Asian students. Interestingly, the rest of the most negatively predictive words bear out demographic differences in activities like jobs, musicals, and student council. Most positively predictive words include tutoring, AP coursework, volunteering, and STEM subjects.

In general, the word frequency results offer some additional nuance for interpreting the topic modeling analysis, speaking to specific words and phrases that are disproportionately written in letters for individual demographics of students. These results also call strong attention to differences, most commonly, in the subjects and activities discussed in letters – perhaps reflective of differences in access to those resources and opportunities, as well. Importantly, recall that these results are completely uncontrolled, meaning that they do still stand as primarily exploratory results when compared with my other analyses that account for student differences in qualifications, activities, and student-teacher relationships.

Appendix A3.2: Assessing Covariate Balance Across Training and Analytic Samples

Though I split the complete dataset into training and analytic subsamples via a stratified randomization process, there is always the risk that the randomization nonetheless

resulted in two substantively different samples. While we can never truly assess the extent to which two samples are definitively similar in all ways, we can at least analyze the extent to which they are similar on observable characteristics and assume that this serves as a suitable enough proxy for unobservable characteristics as well. To that end, Tables A3.5-A3.7 display the same host of descriptive statistics I analyzed in Section III, except comparing students/teachers/student-teacher relationships across the training and analytic ("test") samples. In sum, I observe no substantive differences across data subsamples along any of the teacher characteristics measured, nor along any of the student-teacher relationship characteristics measured. I do observe slight differences in the proportion of students who sent greater shares of applications, the proportion of students with lower scaled GPA, and the extent of missingness in student SAT/ACTs. That said, it is exceptionally unlikely that these magnitudes of difference represent a meaningful concern for the applicability of the trained NLP pipeline to analytic data.

Appendix A3.3: Detailed Topic Modeling Methodology

My use of this topic model approach brings up three main questions regarding the robustness of my analysis: how do I clean and prepare the text data, how do I select the optimal number of topics, and how do I ensure that the resulting supertopics ultimately exhibit a reasonable degree of construct validity? Because these are highly subjective decisions relying largely on data at hand, I document my decisions and processes for each question below, in addition to providing my code open-source alongside this manuscript for more detailed explanations along the way.

How do I clean and prepare the text data?

As described in Appendix A3.1 for my word frequency analysis approach, I clean the text data in a series of steps to improve the applicability and "signal" that the final set of words included in the model ultimately captures. The only difference from my word frequency analysis pipeline is that I restrict the vocabulary of consideration to only those words that appear in at least 1 out of every 2000 letter paragraphs on average (rather than restricting to only those top 1000 words in terms of appearance across paragraphs). This is to remove only the most idiosyncratic words from consideration (i.e., misspellings and/or specific places, club names, and so on). More explicit details on the text cleaning process to go from the raw PDF-scanned letters to the processed text can be found directly in my codebase (e.g., how I systematically remove letter header text and running footer text from the data).

How do I select the number of topics that the topic model algorithm looks for?

Prior research has found that topic modeling is often highly sensitive to the number of topics the algorithm searches for, and this number has no "silver bullet" for deriving its optimal value using the data. Lee and Mimno (2014) suggest a particular method to derive such a value, but they caution against relying on it exclusively versus other diagnostic tests that are valuable to run. Using the same process as Kim et al., 2021, I first use the specification of Lee and Mimno (2014) to get an initial number of topics given the training data (again, per the recommendation of Egami et al.), which in this case was 73. From there, I re-ran the topic model training process in increments of 5 in either direction, down to 13 and up to 113, and assessed each model along a common set of model "fit" metrics. Figure A3.1 displays the fit metrics for all models produced in this process.

I ultimately settled on a total of 73 topics for my structural topic model. From the diagnostics, this value nicely balanced the residuals of the model (which had diminishing returns past this point), semantic coherence of the word groupings (which plateau and then decrease continuously past this point), and align with the initial value arrived at through the Lee and Mimno method.

How do I ensure that the resulting supertopics ultimately exhibit a reasonable degree of construct validity?

In this endeavor, I mirror the spirit of the approach that Quinn et al. (2010) utilized in their analysis of U.S. Senate speeches. They first analyzed 118,000 transcribed speeches to derive their topics and then interpreted these topics for substantive meaning. To check the robustness of these interpretations, they looked at whether the prominence of each topic across speeches trended intuitively with related events in U.S. history. For example, they found a topic that seemed to represent the substantive topic of "defense," and then examined whether the occurrence of words in this topic trended alongside major defense-related events like the Kosovo bombing, the Iraq War authorization, and debates around Abu Ghraib (Figure A3.2). While not conclusive, this approach is one way to provide compelling evidence to support the validity of topic interpretations.

In my case, I similarly use some creative correlation analyses to conduct the same conceptual checks for each identified topic. For example, to examine the validity of the "Sports" supertopic that I constructed, I can examine the extent to which higher prevalence of this supertopic in students' letters is more strongly associated in my main regression specifications with students who indicated having greater levels of sports involvement in their extracurriculars. Logically, we should expect that students who actually played sports are more likely to have notably large levels of discussion about sports in their

recommendations. Examining correlations within the main regression results for each other supertopic indicator can follow in this same context-specific way.

Because these regression models include a vast array of additional control variables, I can also conduct two categories of ad-hoc falsification tests to better guard against spurious correlations. In the first set of falsification tests, I benchmark the aforementioned relationships between supertopic variables and their conceptually-related covariates against the relationships between supertopic variables and the student's cohort year – a variable that should, at least intuitively speaking, be far less related to any substantive content in the recommendation letter across the sample given that no major changes occurred to the teacher recommendation form from 2018 to 2019. For example, it should intuitively be the case that the degree to which a student's letter discusses the sports supertopic does not meaningfully relate to the student's cohort-year. We should then expect that the latent time-varying change of prevalence for the sports supertopic can serve as somewhat of a "noise" floor for assessing spurious correlations. That is, if the relationship between the prevalence of the sports supertopic and a student's extracurricular sports involvement is weaker than the relationship with a student's cohort year, this measure demonstrates no meaningful construct validity. In the second set of falsification tests, I can more broadly examine the relationship with the prevalence of the supertopic variable and other activity types, academic performance variables, and so on that should be conceptually unrelated, as well. This serves as yet another benchmark against the conceptually related relationships we observe to begin with.

That said, I focus my construct validity checks on those supertopics that demonstrated the greatest degree of difference in the main narrative for concision: STEM, Humanities, Sports, Community Engagement, and Extracurriculars. I conduct all of these

examinations in the context of the main landscape regression model for simplicity and to ensure the greatest coverage across the sample. Additional validity checks across the other supertopics are available upon request. Moreover, because this style of analysis is highly subjective and ad-hoc in nature, the full results of my regression specifications (revealing coefficients across all covariates) are also available upon request for auditing.

Table A3.8 displays selected coefficients from the landscape regression on a letter having notably large levels of discussion about STEM topics. The interpretation of these coefficients mirrors the interpretation throughout the main topic modeling results; in other words, the coefficients are percentage point differences off a sample mean of 20%. Rows labeled in green are those I anticipate to have a relationship with the indicator for STEM discussion, while rows labeled in red are those I anticipate to have no relationship with the indicator for STEM discussion. We see as an example that students who received an average score greater than 4.5 across all STEM AP tests they took were 3.2 percentage points more likely (16% given a sample mean of 20%) to have notable discussion about STEM in their letters. In addition, they were a very substantial 37.2 percentage points (186%) more likely to have notable discussion about STEM in their letters if their letter writer was a science teacher. Turning to some of the falsification tests, we see as an example that letters were 0.3pp less likely to have notable discussion about STEM if the student was in the 2019 cohort; while statistically significant, this is substantively not meaningful. Likewise, the correlation with whether a student was under 17 years of age (with 17 years old as the reference group) has a coefficient of 0. Finally, we would not necessarily expect to see a strong relationship between notable discussion of STEM and a student's service activities; commensurately, we see no meaningful relationship there, either.

Table A3.9 displays results for the same set of analyses, but with the Humanities supertopic instead. The parallels all align with the same intuition that this supertopic indicator does indeed seem to be picking up meaningful signal. As an example, letters where the letter writer was an English teacher were 23.8pp more likely to have notably large levels of discussion about humanities; letters where the letter writer was a Social Studies teacher were 15.6pp more likely. Again, we see no meaningful relationships with any of the falsification test variables displayed here.

Table A3.10 reviews results for the Sports supertopic. Students having any athletic activity listed in their extracurriculars were 10.5pp more likely to have notably large levels of discussion about Sports in their letters, and students whose letter writers were coaches were also 45.4pp more likely.

The community engagement supertopic validation results are displayed in Table A3.11. Students reporting any service activities in their extracurriculars were 3.0pp more likely to have notably large levels of discussion about community engagement, and students who had received some form of excellence award for service were also 2.8pp more likely. Because this supertopic also covers the concept of broader social impact and social good, it makes some sense that teachers of social studies (e.g., history, sociology, etc.) would be more likely to have notable discussion about community engagement at 5.3pp (26.5%). That said, this supertopic is slightly more difficult to check for validity on due to the fact that not many of my other covariates have a clear conceptual relationship to this variable.

Lastly, Table A3.12 displays results for the Extracurriculars supertopic. For this category, I report coefficients for every category of student activity I measure: Career, Service, Academic, Athletics, Other, and Arts. The only categories that don't have strong positive relationships with having notable discussion about extracurriculars are Career (-

0.7pp) and Arts (0.3pp). The former is likely because career activities (e.g., part-time jobs) load more heavily onto the Time and Life Management supertopic (due to students needing to balance responsibilities), while the arts activities load more heavily onto the Humanities supertopic (as project-based activities were included in that supertopic).

In all, the supertopics generally display a fairly strong degree of construct validity through these tests. In future work, I hope to further validate these outputs by comparing the supertopic codes against human judgment in a manner similar to my sentiment analysis results and per recommendations by Chang et al. (2009).

Appendix A3.4: Sentiment Analysis Model Selection, Accuracy, and Bias

Because there are a variety of sentiment analysis models, and prior research has shown that these models tend to suffer from low inter-model agreement, the decision for which exact model I deploy for this analysis is highly consequential (Gonçalves et al., 2013). Moreover, the field of NLP has quickly advanced beyond simple word counting models and towards black-box neural network models that tend to be more accurate and nuanced in their classification processes, but substantially harder to interpret and more prone to algorithmic bias: changing its interpretation/classification of a given text based on irrelevant demographic features like the presence of female pronouns (e.g., classifying "he is assertive" as positive, but "she is assertive" as negative). This is because these algorithms are trained on massive text datasets that implicitly contain the biases of societal writing more broadly, "teaching" the algorithm these same biases (Caliskan et al., 2017; Park et al., 2018; Sun et al., 2019). Bias of this nature would threaten the validity of the entire analysis by *creating* bias in the coding of the letters, rather than *identifying* biases in the letters themselves.

For these reasons, I conduct the sentiment analysis with a big-tent approach and select a final analytic model using an empirical process that considers human-judged accuracy and algorithmic bias. First, I deploy a range of transformer models on the data with varying architectures, fine-tuning data, and parameters. Table A3.13 describes each of the models I consider in more detail.

Once the sentiment analysis has been conducted using each method described above, I further train an ensemble model that considers all of these various classifications together to produce a joint sentiment classification. In brief, I use the fine-grain output of each constituent algorithm (to include both its final prediction, its confidence in its prediction, and the likelihood of classifications besides its final prediction) as predictors in a random forest model trained on the SST-5 training dataset mirroring Kim et al., 2021. This approach ultimately achieves exceptionally high performance on the SST-5 test dataset at 55%, approaching the state-of-the-art at the time of writing.

However, performance on the standard SST dataset does not necessarily equate to accuracy on the context-specific text data of my study. Thus, with this array of potential algorithms to finally use in my analysis, I employed a team of six research assistants to manually read and classify (5 levels) the same stratified random sample of 480 letter sentences from the actual teacher recommendation data (stratified on student gender and race/ethnicity) without seeing any algorithmic output.[74] Ultimately, I could then calculate how closely each algorithm performed compared with human judgment. To compare the greatest number of fine-grain algorithms together, I collapse both the algorithmic output and

[74] The training slides and information for this process, as well as the raw human coder output (without the sentences themselves, to prevent identifiability of students) are available upon request.

human output to 3 classes (positive, neutral, and negative) when possible.[75] As a simple test of accuracy, I calculate how often the algorithm perfectly matched each human coder, and then average the algorithm's accuracy across coders. Because the sample of sentences the human coders analyzed were stratified by gender and race/ethnicity, I can also separately calculate the specific accuracy of each algorithm by gender and race/ethnicity as well to detect the potential for algorithmic biases in its accuracy.

Table A3.14 displays the results of this exercise. I find that the sentiment analysis algorithm trained by Barbieri et al. (2020), initially designed to examine Twitter "tweets," outperformed all other algorithms by a fairly wide margin across all student subgroups (the "RoBERTa" column). At an overall accuracy of 77%, this model achieves exceptionally high human-judged accuracy for a three-class sentiment analysis task. Calculated the same way, humans only have about 84% accuracy when compared with one another, indicating that 100% accuracy is not a feasible benchmark to expect for this type of ambiguous and subjective task to begin with. In other words, the algorithm performs *exceptionally* well relative to a human alternative. It is for this reason that I ultimately use this algorithm for my main analysis. Male-female bias in sentiment analysis has generally received the most attention with respect to algorithmic bias; the fact that the accuracy of this algorithm is identical for sentences about both male and female students is especially heartening in this context. I note a slightly lower level of accuracy for Asian students at 74.7%. Even so, the magnitude of this difference compared with White students is unlikely to fully explain the very small magnitude of differences we observe in the regression analyses.

---

[75] Thus, note that several of the candidate algorithms capable only of 2-class output (positive and negative) are excluded from this comparison directly; however, they remain included in the training process of the ensemble random forest algorithm.

Appendix A3.5: Issues of Common Support with Fixed Effects Specifications

My teacher and institution fixed effects strategies reduce the effective estimation sample for each coefficient to only those observations with common support. That is, the coefficient on a binary indicator for female students can only be estimated using recommendations from teachers who have written for both male *and* female students. This is unlikely to be a concern for gender given that most teachers and institutions will have written or received letters for both male and female students, but it could present issues for proportionally smaller racial groups within certain schools.

Table A3.15 displays the proportion of teachers and letters included in the region of common support for each demographic variable of interest given the teacher fixed effects specification. This shows that even for the female demographic variable, we lose about 61% of all teachers in the sample – likely driven by the fact that many teachers only ever write one letter, and many who write very few may still write for only female students. The region of common support in terms of *letters* is far higher, again reflective of the fact that teachers in the region of common support necessarily are a sample of teachers writing a greater number of letters each.

That said, the region of common support by racial demographics is substantially lower. For Black students, only 15% of teachers remain in the sample. For Asian students, about 18%. Another concern this brings up is whether the sample of teachers with common support for these students are meaningfully different from other teachers. That is, there may be a form of selection occurring here: for a teacher to have common support, they must be teaching students of multiple demographics and be *asked* and *willing* to write postsecondary recommendations for these students. Teachers who either teach in schools/classes without

substantial college-going populations of students across demographics, or are outright discriminatory (i.e. refuse to write recommendations for certain groups of students), would be excluded here. Unfortunately, I am unable to observe any meaningful variables about teachers or their schools in these data to assess for differences; I intend to partner with state departments of education to explore this dynamic in future work given access to more detailed teacher staffing data. Lastly, this issue of common support among teacher fixed effects is my primary motivation for not considering crossed race/gender categories in the present analysis. While absolutely of interest both theoretically and substantively speaking, my fixed effects strategies make the region of common support for these interacted categories substantially smaller and more likely to be idiosyncratic despite the size of my overall sample.

Table A3.16 displays the proportion of institutions and applications included in the region of common support for each demographic variable of interest given the institution fixed effects specification. We see here that the regions of common support cover the vast majority of the sample, across all demographic variables. This should make some intuitive sense, given that few institutions restrict their sample of applicants so readily besides female-only institutions and historically black colleges and universities.

APPENDIX TABLES AND FIGURES

Table A1.1. Comparison of "Education Desert" Analyses Across Countries

Country (year of analysis)	Median distance to a public primary school (km)	Mean distance to a public primary school (km)	Share of the population that lives further than 3 km from a public primary school	Net primary enrollment rate, according to World Bank Development Indicators (latest year available)	Classification
Guatemala (2017)	***0.8***	***1.1***	***4.7%***	***85.6% (2017)***	***Low desert prevalence, low enrollment***
Tanzania (2016)	2.2	5.9	40.6%	83.5% (2016)	High desert prevalence, low enrollment
Peru (2020)	0.6	1.4	11.5%	95.7% (2018)	Low desert prevalence, high enrollment
Costa Rica (2020)	0.5	0.6	3.0%	97.3% (2018)	Low desert prevalence, high enrollment
Kenya (2018)	0.8	2.0	12.9%	80.0% (2012)	High desert prevalence, low enrollment
Rwanda (2012)	1.4	1.7	11.9%	98.8% (2016)	High desert prevalence, high enrollment
South Africa (2020)	0.7	1.1	5.1%	87.0% (2017)	Low desert prevalence, low enrollment

Table A2.1. Advising Models Used in the N2FL Intervention

Model	Example Advisor Background(s)	Advisor Role	Sample Message	Number of Insts
Professional Advisor	Hired specifically for the N2FL project	Direct assistance with tasks (e.g., registering for courses, financial aid applications)	Hi, it's <Professional Advisor>. With finals coming up, I wanted to check if you've used <Support Center> for help with classes. Can I help you get connected?	Nine
Faculty Advisor	University faculty	Direct assistance with questions in their specialization (e.g., course selection) and recommending campus resources for other questions (e.g., financial aid)	Hey, it's <Faculty Advisor>. As you're planning for spring, think about picking up an extra course. This can help you graduate sooner. Can I help you choose another class?	One
Staff Point Person	Administrative assistant on student engagement team	Direct students to the resource most appropriate for providing assistance	Hi <Student>! Registration for fall and summer starts 4/2. Have you talked to an advisor about the next classes you need to take in your program?	Six
Segmented Advising	Mix of campus staff (e.g., some faculty advisors coupled with a career services counselor)	Leveraged multiple staff depending on question (e.g., student replies to automated questions about course registration went to an Academic Advisor's portfolio)	Hi, it's <Advisor>. Fafsa.gov is now open for the 2018-2019 school year and applying early gets you the most financial aid. Have you started FAFSA yet? [student replies are routed to a Financial Advisor's inbox]	Two

Notes: This table illustrates the four primary advising models that emerged in terms of how institutions staffed the text messaging campaign.

Table A2.2. Validation Exercise Results

Accuracy Metric	Score	Cases Reviewed	Cases in Full Dataset
Perfect Accuracy Across Groups	56%	150	27719
Perfect Accuracy for Group (-2)	67%	3	114
Perfect Accuracy for Group (-1)	53%	15	6627
Perfect Accuracy for Group (0)	52%	87	13537
Perfect Accuracy for Group (1)	58%	31	6250
Perfect Accuracy for Group (2)	79%	14	1191
One-off Accuracy Across Groups	97%	150	27719
One-off Accuracy for Group (-2)	100%	3	114
One-off Accuracy for Group (-1)	93%	15	6627
One-off Accuracy for Group (0)	97%	87	13537
One-off Accuracy for Group (1)	100%	31	6250
One-off Accuracy for Group (2)	100%	14	1191
Perfect 3-Class Accuracy Across Groups	64%	150	27719
Perfect 3-Class Accuracy for Group (-1)	87%	15	6627
Perfect 3-Class Accuracy for Group (0)	52%	87	13537
Perfect 3-Class Accuracy for Group (1)	68%	31	6250
Approval Rate Across Groups	83%	150	27719
Approval Rate for Group (-2)	80%	10	114
Approval Rate for Group (-1)	52%	27	6627
Approval Rate for Group (0)	87%	68	13537
Approval Rate for Group (1)	97%	31	6250
Approval Rate for Group (2)	100%	14	1191

Note: Both "Cases Reviewed" and "Cases in Full Dataset" indicate the number of class cases within the exercise dataset/full dataset as labeled by the *algorithm output*. Our random sample for each exercise was stratified based on algorithm output class and message type (sent by student, scheduled text sent by advisor, personalized text sent by advisor).

Table A2.3. Classification Concordance Table Between Masked and Unmasked Text Data

	Unmasked					
Masked	-2	-1	0	1	2	Total
-2	113	5	0	0	0	118
-1	12	6,582	149	0	0	6,743
0	0	211	13,850	171	0	14,232
1	0	0	199	6,597	51	6,847
2	0	0	1	51	1,196	1,248
Total	125	6,798	14,199	6,819	1,247	29,188

Table A2.4. Constituent Model Characteristics

Model Name	Language Model Training Data	Task Training Data	Sentiment Class Type	Task Score
BERT (Google)	11,000 books from SmashWords ("BookCorpus") English Wikipedia articles	Yelp Restaurant Reviews	5-class (review stars)	**SST5**: 40% Accuracy
BERT (Google)	11,000 books from SmashWords ("BookCorpus") English Wikipedia articles	150,000 product reviews (1-5 stars)	5-class (review stars)	**SST5**: 42% Accuracy
ALBERT (Google)	11,000 books from SmashWords ("BookCorpus") English Wikipedia articles	Movie Reviews (Stanford Sentiment Treebank, 2-class)	2-class (pos/ neg)	**SST2**: 94% Accuracy
XLNet (Carnegie Mellon and Google)	11,000 books from SmashWords ("BookCorpus") English Wikipedia articles News Articles ("Gigaword 5th Edition") Websites ("Common Crawl" and "ClueWeb")	Movie Reviews (Stanford Sentiment Treebank, 2-class)	2-class (pos/ neg)	**SST2**: 94% Accuracy
Stanza (Stanford)	N/A (not pre-trained)	Movie Reviews (Stanford Sentiment Treebank, 3-class) Sitcom Dialogue ("MELD") IMDB, Amazon, and Yelp reviews (UCIrvine "SLSD") TripAdvisor Hotel Reviews ("ArguAna") Tweets re: Airlines ("CrowdFlower")	3-class (pos/ neg/ neu)	**SST3**: 73% Accuracy
RoBERTa (UWash and Facebook AI)	11,000 books from SmashWords ("BookCorpus") English Wikipedia articles News Articles ("Common Crawl News" Web content extracted from websites shared on Reddit ("OpenWebText") Story dataset ("CommonCrawl" Stories) ~1 year of Tweets	Tweets (TweetEval, per Barbieri et. al, 2020)	3-class (pos/ neg/ neu)	**SST3:** 65% Accuracy
T5 (Google)	Websites ("Clean Common Crawl")	Highly polarized IMDB movie reviews	2-class (pos/ neg)	**SST2**: 90% Accuracy

Table A2.5. Sample Topic Interpretations and Word Groups

Distinguishing Words	Coder 1	Coder 2	Coder 3	Coder Agreement
graduation, apply, apply graduation, express, tokensystemname express, application, link, walk, ceremony, cunyfirst	applying to graduation	Graduation application	apply for graduation	Perfect
office, hour, late, friday, monday, response, stop, visit, late response, c107	finding a time to visit a campus office	office hours	[office] contact logistics	Near-Perfect
campus, service, counselor, job, support, mind, ahead, provide, care, set	campus resources to support students	work	Counseling services	Minor Disagreement

Table A2.6. Complete Supertopic Groupings and Sample Words

Academic Planning	math, science, requirement, biology, art, spanish, registrar, language, college, mat
	tokensystemname, program, school, college, website, tokensystemname tokensystemname, nursing, online, application, tokensystemname website
	graduation, apply, apply graduation, express, tokensystemname express, application, link, walk, ceremony, cunyfirst
	credit, graduate, course, major, requirement, internship, psychology, minor, elective, missing
	graduate, congratulations, graduating, applied, feel, free, ready, december, feel free, graduated
	degree, transfer, major, change, associates, plan, audit, transcript, bachelors, finish
Academic Supports	hope, information, tokenurl, hey, center, tutoring, located, helpful, office, visit
	question, hey, info, yeah, answer, reaching, nice, assist, specific, study
	im, semester, grade, luck, checking, enrolled, final, exam, planning, lol
	campus, service, counselor, job, support, mind, ahead, provide, care, set
Meeting Logistics	appointment, time, tomorrow, wednesday, thursday, tuesday, monday, meet, availability, day
	message, office, time, answer, frame, time frame, message time, answer message, frame patience, patience
	tokenphonenumber, call, phone, person, walk, call tokenphonenumber, monday, hour, plan, discuss
	advisor, academic, meet, helpful, hope information, information helpful, appointment, academic advisor, meet academic, hope
	advisor, contact, academic, tokenname, meet, advising, academic advisor, track, meet advisor, reach
	appointment, schedule, schedule appointment, set, advising, advisor, tokenurl, link, met, meet
	email, tokenemailaddress, send, email tokenemailaddress, check, received, information, connect, forward, contact
	office, hour, late, friday, monday, response, stop, visit, late response, c107
	tokensystemname, academic advisement, advisement, academic, advisement center, center, line, reach, call, tokenphonenumber
Course Planning	spring, registration, date, winter, spring semester, november, enrollment, session, register, semester
	student, id, drop, time, gpa, student email, check, access, withdraw, student id
	summer, fall, course, taking, summer class, online, fall semester, summer course, plan, im taking
	professor, department, told, writing, speak, permission, request, alright, issue, morning
	class, register, registered, time, class semester, add, pin, class im, register class, left
Financial Aid	tokensis, hold, account, plan, payment, pay, bursar, log, check, tokenurl
	financial, aid, financial aid, fafsa, office, aid office, scholarship, loan, tuition, pay
	day, happy, wondering, start, break, due, hope, yesterday, deadline, january
	text, message, stop, receive, letting, update, wrong, list, text message, send
	dont, week, ill, youre, ive, havent, sounds, awesome, didnt, fine
	assistance, time, reach, hear, glad, taking, text, respond, hesitate, taking time

Table A3.1. Word Frequency Analysis Results: Most Positively and Negatively Predictive Words for Female Students

	Positively Predictive		Negatively Predictive	
Rank	Keyword	Coefficient	Keyword	Coefficient
1	dance	1.119	football	-1.894
2	art	0.387	computer	-0.646
3	psychology	0.370	engineering	-0.501
4	medical	0.364	humor	-0.448
5	compassion	0.342	technology	-0.324
6	compassionate	0.335	economics	-0.308
7	health	0.328	respectful	-0.308
8	children	0.294	sports	-0.303
9	organized	0.279	business	-0.272
10	caring	0.270	soccer	-0.249
11	voice	0.263	basketball	-0.248
12	smile	0.252	physics	-0.246
13	arts	0.240	athletics	-0.210
14	notes	0.211	polite	-0.205
15	amazing	0.209	fine	-0.199
16	shy	0.209	improved	-0.178
17	joy	0.200	country	-0.175
18	wonderful	0.193	personable	-0.174
19	french	0.187	mathematics	-0.174
20	determined	0.183	solve	-0.171

Table A3.2. Word Frequency Analysis Results: Most Positively and Negatively Predictive Words for White Students

	Positively Predictive		Negatively Predictive	
Rank	Keyword	Coefficient	Keyword	Coefficient
1	social studies	0.381	international	-0.527
2	athlete	0.315	economics	-0.414
3	athletics	0.315	english language	-0.386
4	college university	0.309	tutoring	-0.334
5	athletic	0.283	subjects	-0.324
6	language composition	0.275	results	-0.277
7	ap english	0.268	keen	-0.260
8	coach	0.240	won	-0.252
9	ii	0.229	achievements	-0.238
10	tokenaddress school	0.226	subject	-0.228
11	varsity	0.208	exams	-0.224
12	spring	0.205	progress	-0.214
13	musical	0.200	english	-0.212
14	fine	0.194	mathematics	-0.202
15	conscientious	0.189	office	-0.198
16	special	0.184	medical	-0.195
17	honors	0.182	culture	-0.195
18	honor society	0.182	tutor	-0.187
19	job	0.178	head	-0.186
20	acceptance	0.175	lot	-0.185

Table A3.3. Word Frequency Analysis Results: Most Positively and Negatively Predictive Words for Black Students

	Positively Predictive		Negatively Predictive	
Rank	Keyword	Coefficient	Keyword	Coefficient
1	basketball	0.439	cross	-0.322
2	track	0.428	economics	-0.260
3	roll	0.320	competitive	-0.236
4	church	0.284	tokenname recommendation	-0.209
5	collegiate	0.280	mathematics	-0.197
6	contact tokenphone	0.259	physics	-0.194
7	football	0.254	country	-0.190
8	scholar	0.249	musical	-0.187
9	school tokenaddress	0.223	international	-0.180
10	government	0.192	traits	-0.180
11	scholarship	0.188	calculus	-0.176
12	pleasure teaching	0.164	enthusiastic	-0.173
13	college university	0.157	coach	-0.168
14	gpa	0.156	ap physics	-0.166
15	academy	0.151	student ap	-0.165
16	assist	0.150	schedule	-0.161
17	health	0.146	teams	-0.160
18	maintained	0.141	ap chemistry	-0.157
19	institution	0.139	special	-0.157
20	office	0.139	sports	-0.157

Table A3.4. Word Frequency Analysis Results: Most Positively and Negatively Predictive Words for Asian Students

	Positively Predictive		Negatively Predictive	
Rank	Keyword	Coefficient	Keyword	Coefficient
1	tutoring	0.310	football	-0.625
2	medical	0.283	athlete	-0.470
3	clubs	0.240	soccer	-0.465
4	student ap	0.232	sports	-0.380
5	collaborative	0.226	athletic	-0.378
6	tutor	0.206	athletics	-0.317
7	ap chemistry	0.200	basketball	-0.245
8	debate	0.188	church	-0.216
9	science	0.173	admission school	-0.175
10	volunteering	0.164	roll	-0.165
11	ap biology	0.155	senior	-0.145
12	volunteer	0.151	job	-0.143
13	advanced placement	0.146	ii	-0.139
14	calculus	0.145	secondary	-0.138
15	tokenname ap	0.143	musical	-0.133
16	sciences	0.132	student council	-0.132
17	ap	0.131	special	-0.126
18	volunteered	0.126	government	-0.125
19	enthusiastically	0.125	environmental	-0.121
20	volunteers	0.124	tokenname past	-0.120

Table A3.5. Student-Level Descriptive Statistics Across Training and Analytic Data Subsamples

Variable	Train Data	Test Data
Sample		
Students	266161	1496313
Letters	280887	2531317
Student Demographics		
Female	0.573	0.566
First Generation	0.259	0.271
International	0.112	0.106
Fee Waiver Recipient	0.225	0.229
Student Race/Ethnicity		
White	0.484	0.488
Black	0.081	0.086
Latinx	0.137	0.14
Asian	0.115	0.107
Other	0.046	0.047
Missing	0.137	0.132
Student School Sector		
Public School	0.743	0.75
Private School	0.251	0.243
Other School	0.006	0.007

Variable	Train Data	Test Data
Applications Sent		
1-3	0.295	0.329
4-7	0.384	0.378
>=8	0.321	0.293
Scaled GPA Group		
Other/Missing	0.174	0.174
<0.90	0.218	0.238
0.90-0.99	0.295	0.292
1.00	0.037	0.035
>1.00	0.276	0.262
Math SAT/ACT Percentile Group		
Missing	0.237	0.249
<75	0.248	0.266
75-89	0.216	0.213
90-94	0.11	0.103
>=95	0.189	0.169
Verbal SAT/ACT Percentile Group		
Missing	0.236	0.247
<75	0.244	0.259
75-89	0.182	0.181
90-94	0.125	0.12
>=95	0.214	0.193

Table A3.6. Teacher-Level Descriptive Statistics Across Training and Analytic Data Subsamples

Variable	Train Data	Test Data
Sample		
Teachers	59733	537306
Letters Written in Current Year		
1	0.443	0.443
2-5	0.373	0.371
6-10	0.113	0.114
11-25	0.064	0.064
25+	0.007	0.007
Letters Written in Past Two Years		
0	0.416	0.418
1	0.115	0.116
2-5	0.216	0.214
6-10	0.115	0.114
11-25	0.105	0.106
25+	0.033	0.032

Table A3.7. Student-Teacher Relationship-Level Descriptive Statistics Across Training and Analytic Data Subsamples

Variable	Train Data	Test Data
Subject Area		
English	0.226	0.228
Social Studies	0.171	0.174
Math	0.167	0.165
Science	0.202	0.201
World Language	0.065	0.064
Computer Science	0.015	0.016
Other	0.153	0.152
Student-Coach Relationship		
Coach	0.049	0.049
Years Known		
0	0.015	0.017
1	0.514	0.511
2	0.313	0.315
3	0.097	0.098
4	0.06	0.059
Other Relationship		
Other Relationship	0.034	0.034

Table A3.8. Additional STEM Supertopic Correlations in Landscape Regression

Covariate	Coefficient (Standard Error)
Received Average Score Greater than 4.5 Across All STEM AP Tests	0.032*** (0.001)
Letter Writer was Science Teacher	0.372*** (0.004)
Number of Science AP Tests Taken	0.051*** (0.001)
Student in 75th Percentile or Below in SAT/ACT Math	-0.035*** (0.002)
2019 Cohort Indicator	-0.003*** (0.001)
Student's Age Under 17	0.000 (0.001)
Any Service Activity	-0.002** (0.001)

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table A3.9. Additional Humanities Supertopic Correlations in Landscape Regression

Covariate	Coefficient (Standard Error)
Received Average Score Greater than 4.5 Across All English AP Tests	0.017*** (0.002)
Received Average Score Greater than 4.5 Across All Social Studies AP Tests	0.010*** (0.001)
Letter Writer was English Teacher	0.238*** (0.003)
Letter Writer was Social Studies Teacher	0.156*** (0.004)
2019 Cohort Indicator	-0.000 (0.001)
Student's Age Under 17	-0.004*** (0.001)
Any Service Activity	-0.005*** (0.001)

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table A3.10. Additional Sports Supertopic Correlations in Landscape Regression

Covariate	Coefficient (Standard Error)
Any Athletics Activity	0.105*** (0.001)
Leadership in Athletics Activity	0.036*** (0.001)
Excellence Award in Athletics Activity	0.037*** (0.002)
Student-Teacher Coach Relationship	0.454*** (0.002)
2019 Cohort Indicator	0.000 (0.001)
Student's Age Under 17	0.002* (0.001)
Received Greater than 4.5 Average Across All Social Studies AP Tests	-0.001 (0.001)

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table A3.11. Additional Community Engagement Supertopic Correlations in Landscape Regression

Covariate	Coefficient (Standard Error)
Any Service Activity	0.030*** (0.001)
Excellence Award Received in Service Activity	0.028*** (0.001)
Letter Writer was Social Studies Teacher	0.053*** (0.003)
2019 Cohort Indicator	-0.006*** (0.001)
Student's Age Under 17	0.003*** (0.001)
Number of English AP Tests Taken	0.002. (0.001)

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table A3.12. Additional Extracurriculars Supertopic Correlations in Landscape Regression

Covariate	Coefficient (Standard Error)
Total Leadership Activities (Count)	0.007*** (0.000)
Any Career Activity	-0.007*** (0.001)
Any Service Activity	0.011*** (0.001)
Any Academic Activity	0.017*** (0.001)
Any Athletics Activity	0.024*** (0.001)
Any Other Activity	0.010*** (0.001)
Any Arts Activity	0.003*** (0.001)
Other School Type	-0.070*** (0.004)
2019 Cohort Indicator	-0.005*** (0.001)
Student's Age Under 17	0.001 (0.001)
Number of English AP Tests Taken	0.002 (0.001)

Notes: (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Table A3.13. Sentiment Analysis Model Characteristics

Model Name	Language Model Training Data	Task Training Data	Sentiment Class Type	Task Score
BERT (Google)	11,000 books from SmashWords ("BookCorpus") English Wikipedia articles	Yelp Restaurant Reviews	5-class (review stars)	**SST5**: 40% Accuracy
BERT (Google)	11,000 books from SmashWords ("BookCorpus") English Wikipedia articles	150,000 product reviews (1-5 stars)	5-class (review stars)	**SST5**: 42% Accuracy
ALBERT (Google)	11,000 books from SmashWords ("BookCorpus") English Wikipedia articles	Movie Reviews (Stanford Sentiment Treebank, 2-class)	2-class (pos/ neg)	**SST2**: 94% Accuracy
XLNet (Carnegie Mellon and Google)	11,000 books from SmashWords ("BookCorpus") English Wikipedia articles News Articles ("Gigaword 5th Edition") Websites ("Common Crawl" and "ClueWeb")	Movie Reviews (Stanford Sentiment Treebank, 2-class)	2-class (pos/ neg)	**SST2**: 94% Accuracy
Stanza (Stanford)	N/A (not pre-trained)	Movie Reviews (Stanford Sentiment Treebank, 3-class) Sitcom Dialogue ("MELD") IMDB, Amazon, and Yelp reviews (UCIrvine "SLSD") TripAdvisor Hotel Reviews ("ArguAna") Tweets re: Airlines ("CrowdFlower")	3-class (pos/ neg/ neu)	**SST3**: 73% Accuracy
RoBERTa (University of Washington and Facebook AI)	11,000 books from SmashWords ("BookCorpus") English Wikipedia articles News Articles ("Common Crawl News" Web content extracted from websites shared on Reddit ("OpenWebText") Story dataset ("CommonCrawl" Stories) ~1 year of Tweets	Tweets (TweetEval, per Barbieri et. al, 2020)	3-class (pos/ neg/ neu)	**SST3:** 65% Accuracy
T5 (Google)	Websites ("Clean Common Crawl")	Highly polarized IMDB movie reviews	2-class (pos/ neg)	**SST2**: 90% Accuracy

Table A3.14. Sentiment Analysis Algorithm Accuracy Versus Human Judgment

Subset	Obs.	Random Forest	BERT Yelp	BERT Products	Stanza	RoBERTa
Overall	480	0.717	0.625	0.681	0.664	0.771
Female	240	0.719	0.619	0.687	0.635	0.771
Male	240	0.715	0.631	0.675	0.692	0.772
White	120	0.726	0.676	0.738	0.675	0.779
Black	120	0.715	0.603	0.697	0.614	0.768
Asian	120	0.708	0.606	0.682	0.669	0.747
Latinx	120	0.718	0.615	0.607	0.696	0.790

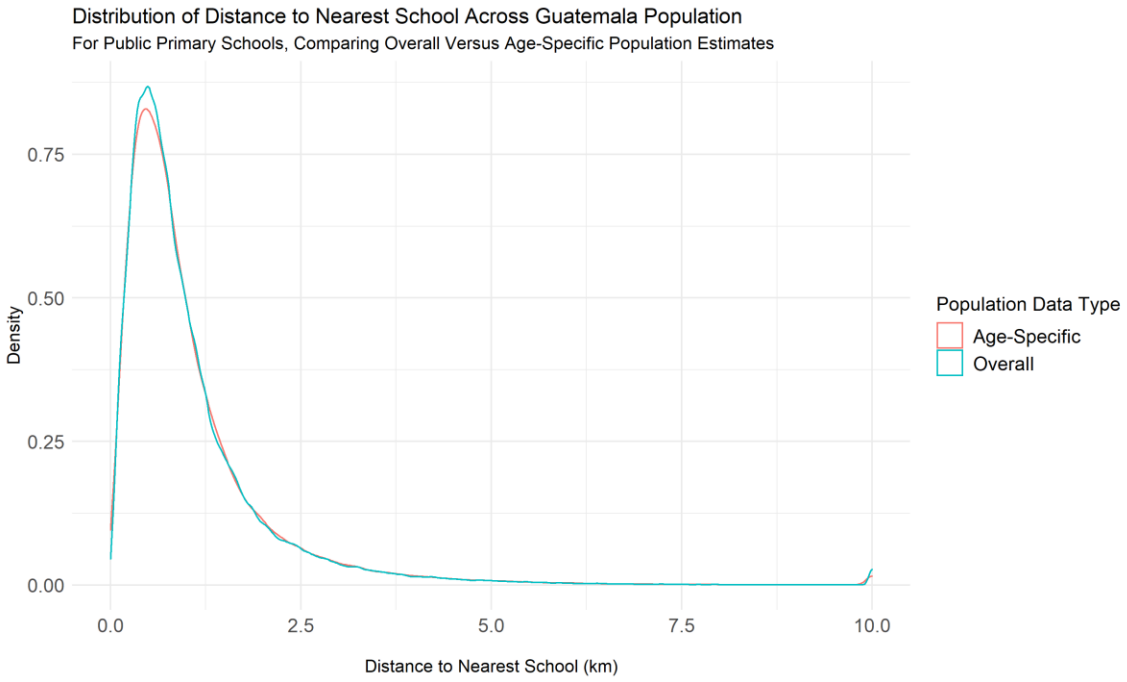
Table A3.15. Proportion of Teachers and Letters in the Common Support Region for Demographic Variables of Interest Using Teacher Fixed Effects

Demographic Variable	% of Teachers	# of Teachers	% of Letters	# of Letters
Female	0.392	218585	0.761	1937606
White	0.331	184729	0.697	1775344
Black	0.149	82944	0.359	913897
Latinx	0.225	125855	0.538	1368432
Asian	0.183	102370	0.478	1215642
Other	0.135	75479	0.382	972834
First-gen	0.328	182829	0.677	1722192
Fee Waiver	0.379	211315	0.757	1926621
International	0.065	36021	0.175	446423

Table A3.16. Proportion of Teachers and Letters in the Common Support Region for Demographic Variables of Interest Using Institution Fixed Effects

Demographic Variable	% of Insts	# of Insts	% of Apps	# of Apps
Female	0.971	775	0.998	12795735
White	0.991	791	1	12815250
Black	0.977	780	1	12813542
Latinx	0.986	787	1	12814886
Asian	0.977	780	1	12814279
Other	0.979	781	1	12814606
First-gen	0.991	791	1	12815250
Fee Waiver	0.991	791	1	12815250
International	0.99	790	1	12815248

Figure A1.1. Comparisons of Results Using Overall and Age-Specific Population Data



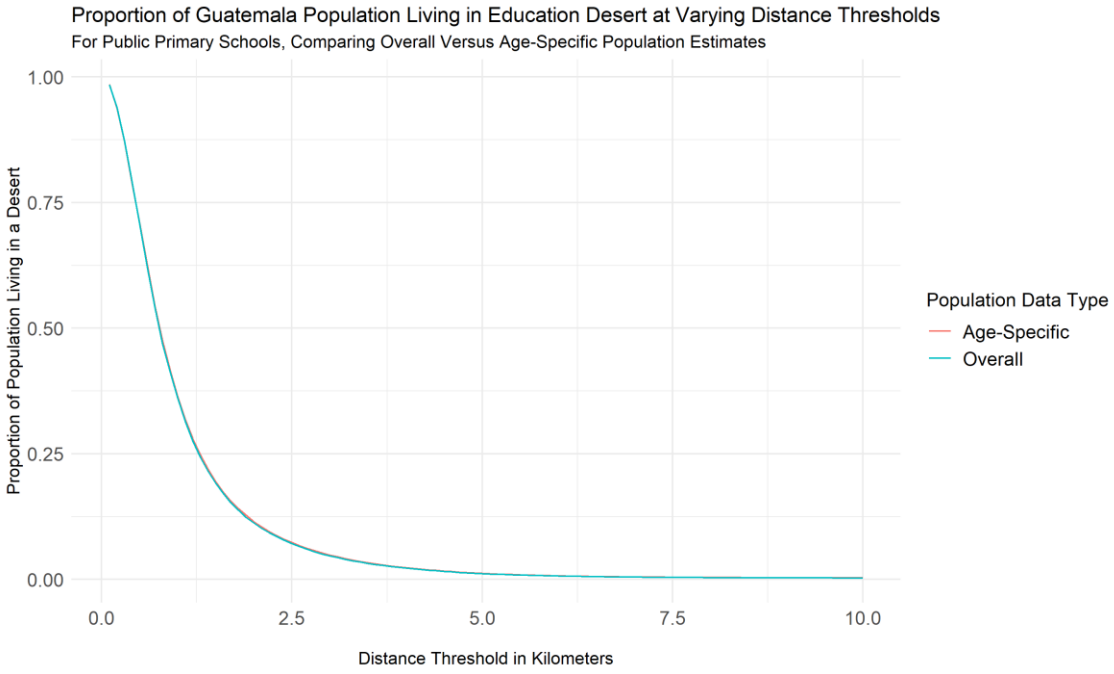


Figure A2.1. Accuracy Diagnostics for Random Forest Classifier on SST5

	precision	recall	f1-score	support
LABEL_0	0.58	0.27	0.37	279
LABEL_1	0.59	0.67	0.63	633
LABEL_2	0.39	0.50	0.44	389
LABEL_3	0.55	0.67	0.60	510
LABEL_4	0.72	0.45	0.55	399
accuracy			0.55	2210
macro avg	0.57	0.51	0.52	2210
weighted avg	0.57	0.55	0.54	2210

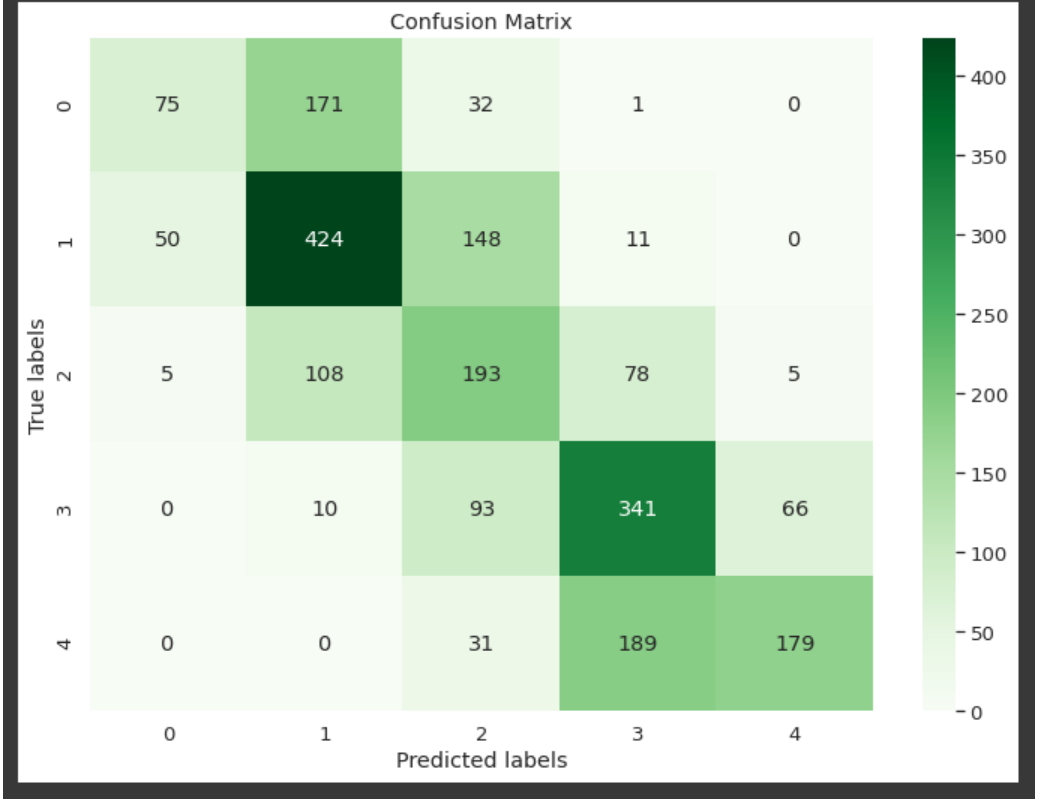


Figure A2.2. Accuracy Diagnostics for Random Forest Classifier on SST3

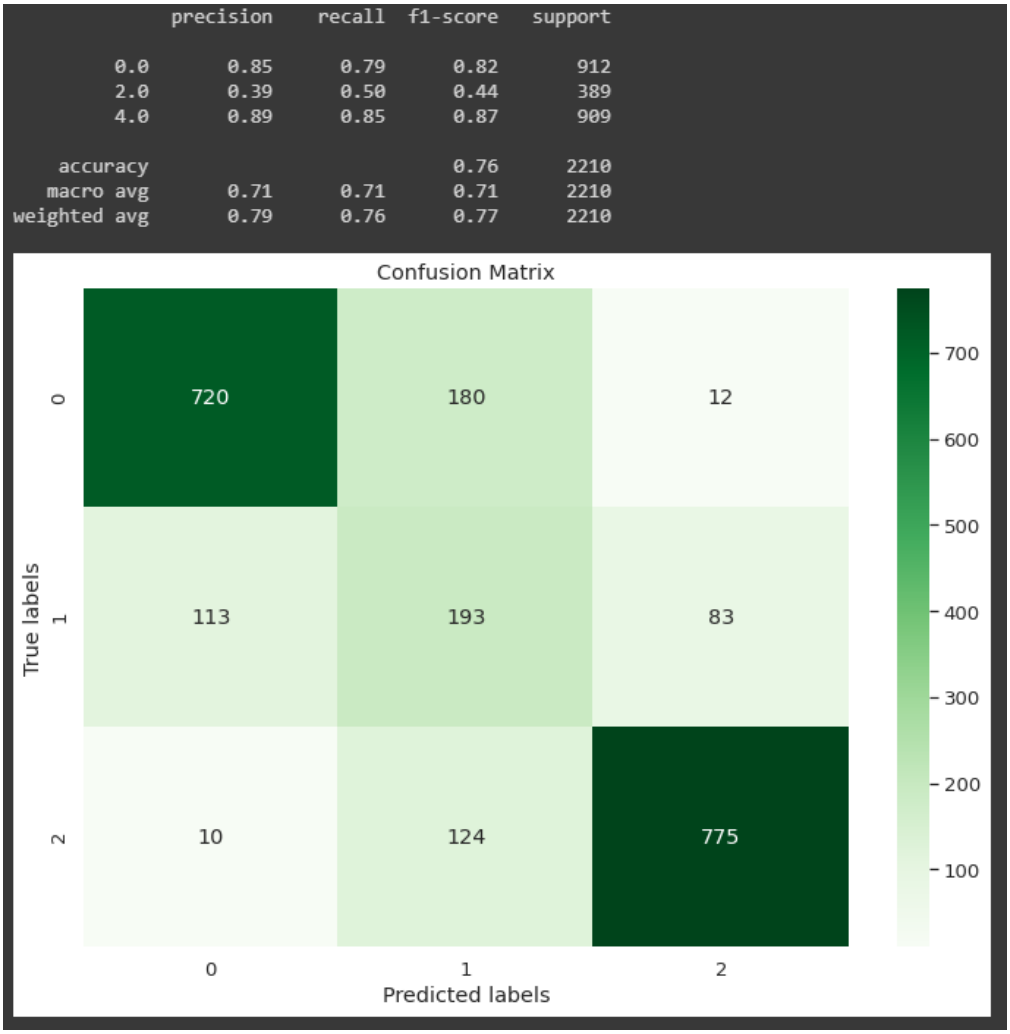


Figure A2.3. Distribution of Probabilities for Assigned Sentiment Scores

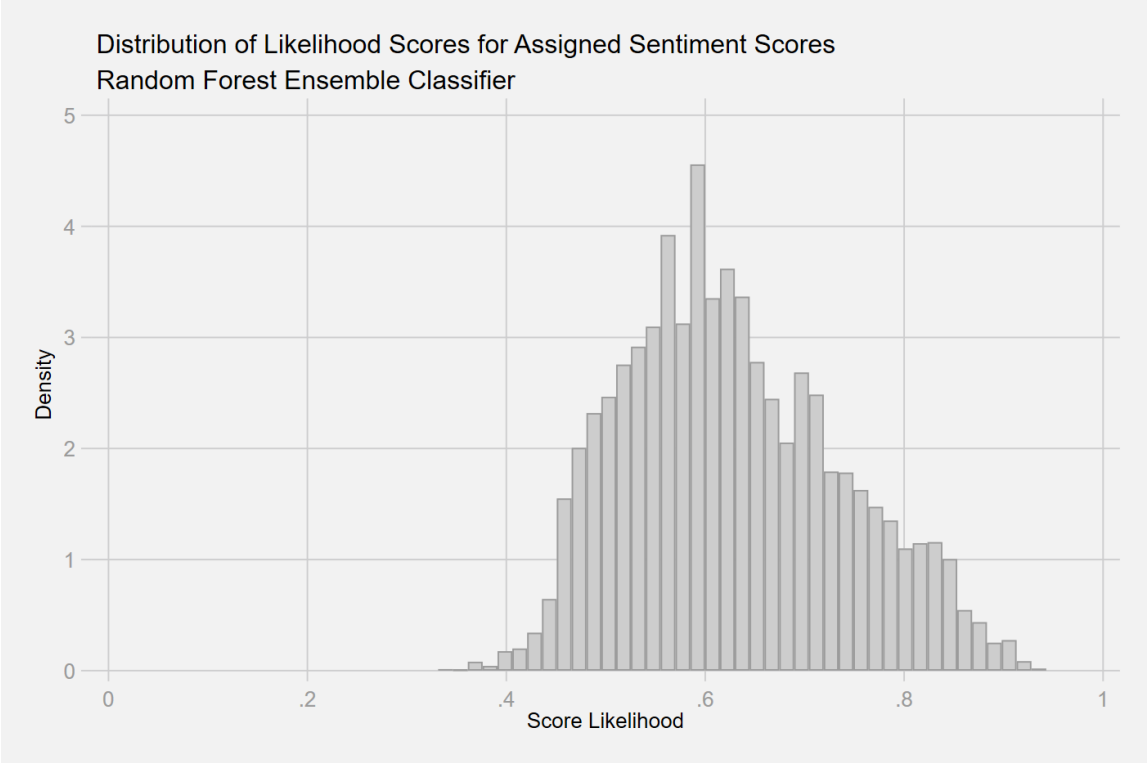


Figure A2.4. Distribution of Differences Between 1st and 2nd Sentiment Classification Probabilities

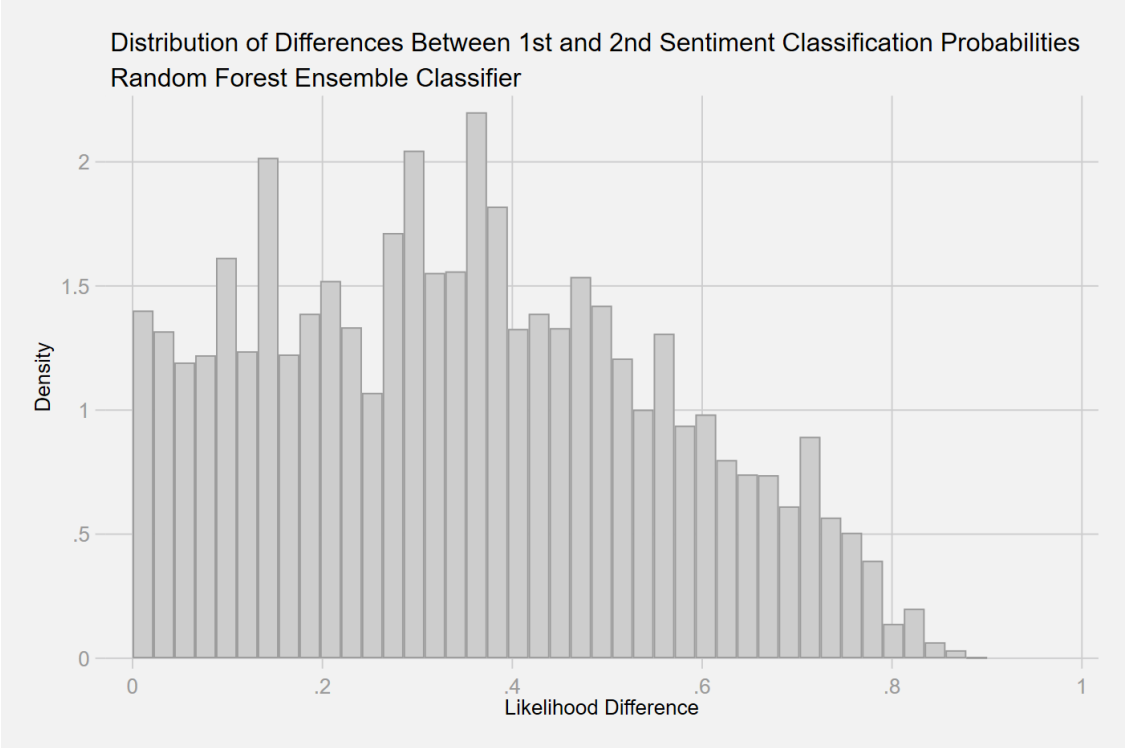


Figure A2.5. Common Model Fit Metrics Across K Parameter Specifications

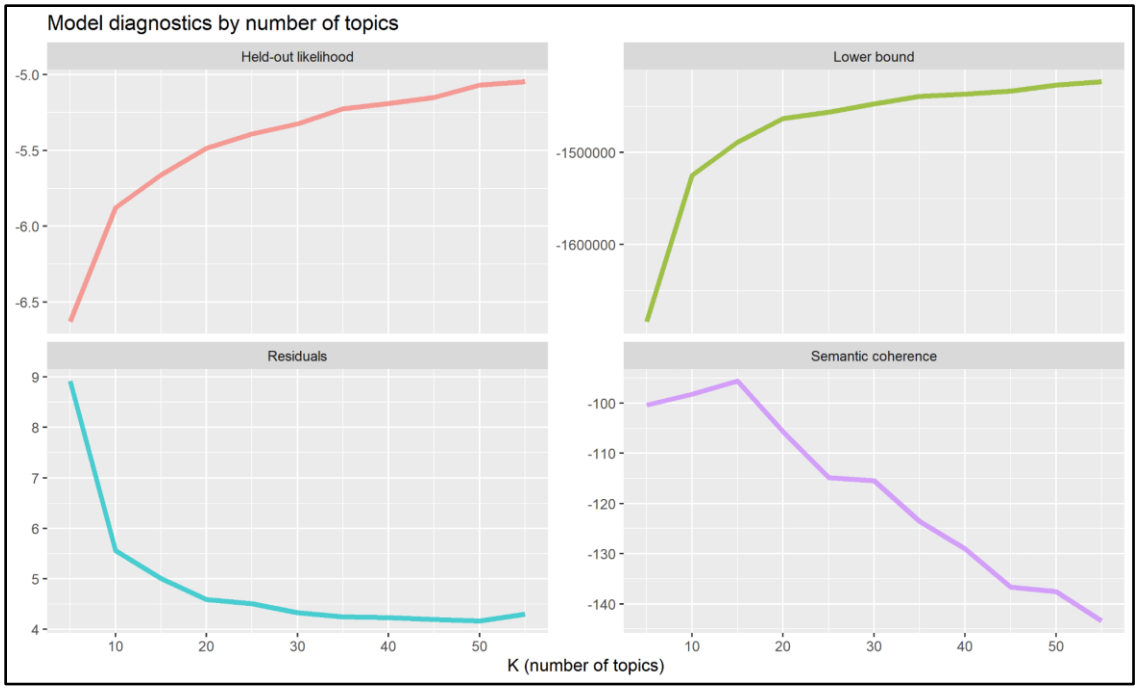


Figure A2.6. Topics and Supertopics by Frequency of Word Occurrences in Topic Model Training Data

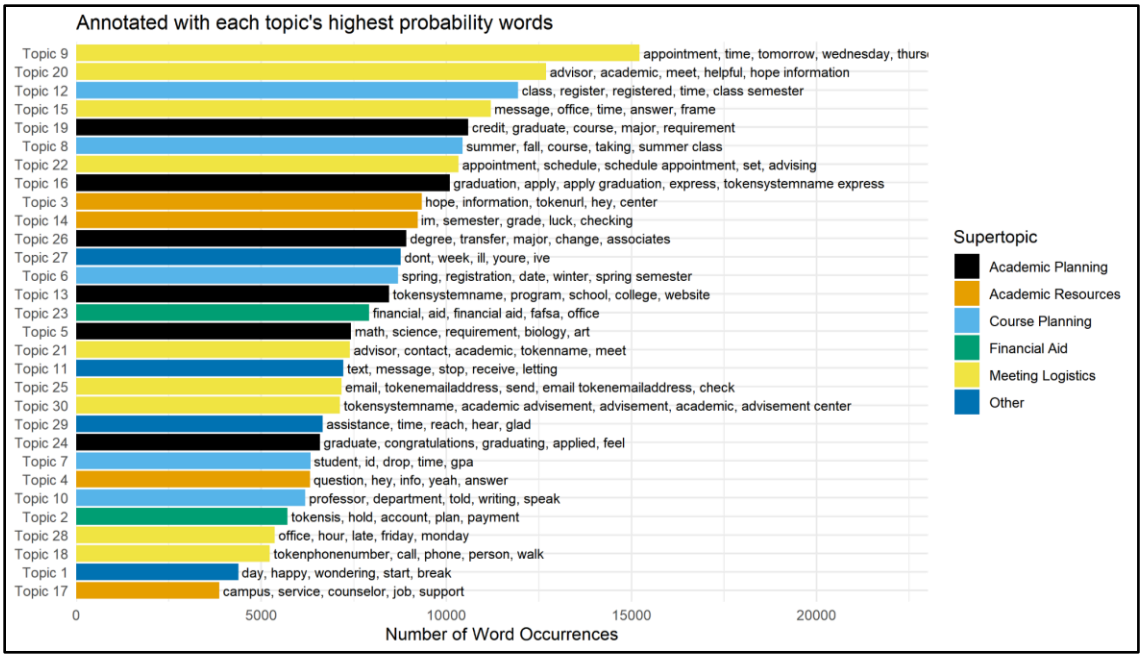


Figure A3.1. Measures of Fit Across Topic Model Search Process

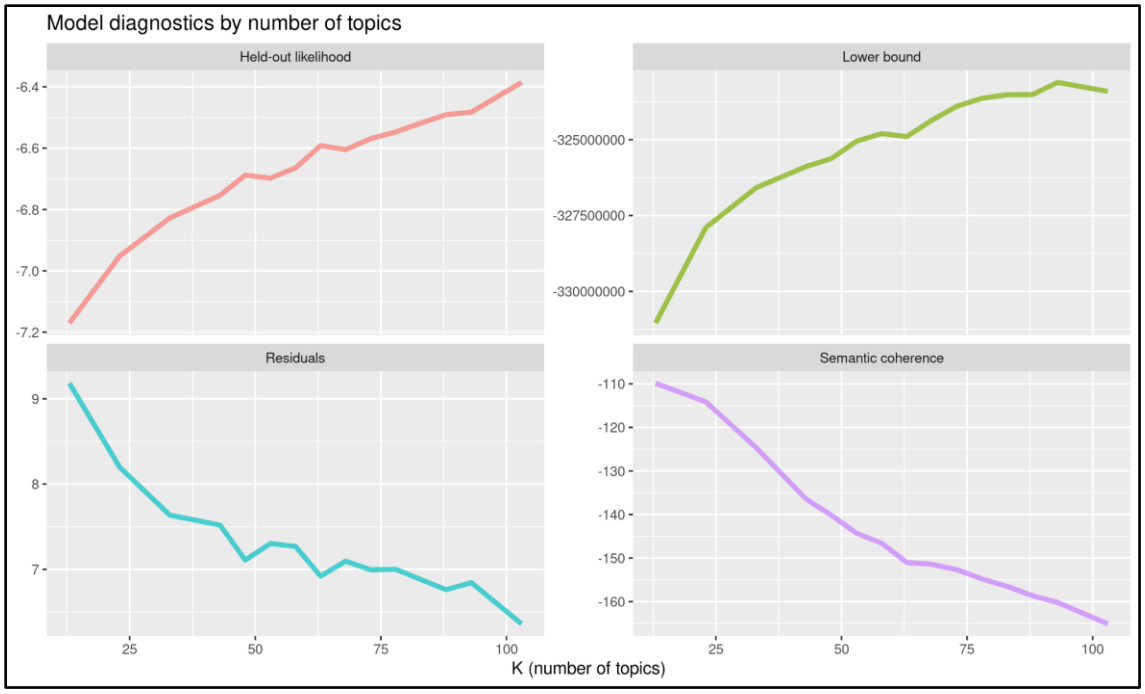


Figure A3.2. "Defense" Topic Frequency in Senate Speeches Over Time (Quinn et al., 2010)