UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays in Applied Economics

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Economics

by

Tanner Scott Eastmond

Committee in charge:

Professor Gordon Dahl, Co-Chair
Professor Yizhak Fadlon, Co-Chair
Professor Julian Betts
Professor Julie Cullen
Professor Sally Sadoff

2024

The Dissertation of Tanner Scott Eastmond is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

This dissertation is dedicated to my wonderful wife, Allie, who is my partner in everything and made it possible for me to complete this degree, and to my two boys, Tate and Asher.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Professors Gordon Dahl and Itzik Fadlon for their support as my committee co-chairs. They have both been pivotal in helping and guiding me through this process.

I also want to thank the rest of my committee, Professors Julian Betts, Julie Cullen, and Sally Sadoff. They have been and are great mentors for me.

Chapter 1, in part, is currently being prepared for submission for publication of the material and is coauthored with Ricks, Michael. The dissertation author was the primary researcher and author of this material.

Chapter 2, in part, is currently being prepared for submission for publication of the material and is coauthored with Ricks, Michael; Mather, Nathan; and Betts, Julian. The dissertation author was the primary researcher and author of this material.

Chapter 3, in part, is currently being prepared for submission for publication of the material and is coauthored with Bonheur, Amanda. The dissertation author was the primary researcher and author of this material.

VITA

2018 Bachelor of Arts, Brigham Young University

2024 Doctor of Philosophy, University of California San Diego

FIELDS OF STUDY

Major Fields: Labor and Public Economics

ABSTRACT OF THE DISSERTATION

Essays in Applied Economics

by

Tanner Scott Eastmond

Doctor of Philosophy in Economics

University of California San Diego, 2024

Professor Gordon Dahl, Co-Chair
Professor Yizhak Fadlon, Co-Chair

This dissertation explores the causes and consequences of inequality and bias in the labor market as well as light-touch solutions to ameliorate those consequences.

In Chapter 1 my coauthors and I seek to understand the extent to which beliefs about various groups of people are malleable in the long term through exposure to different types of places. We study this using variation from the location assignments of volunteer missionaries for The Church of Jesus Christ of Latter-day Saints. Administering an original survey to former volunteers, we find noticeable changes in attitudes about underrepresented minorities, political out-partisans, and women in the workforce.

In Chapter 2 my coauthors and I study the consequences for evaluating public policy when

there are heterogeneous impacts of that policy and when the social planner has distributional preferences over the subjects of the policy. In particular, we study teacher allocations and compare them to the baseline case of using mean-based "value-added" measures. Using data from the San Diego Unified School District we estimate heterogeneity in teacher value-added over between students with above- and below-median test scores. Because a majority of teachers have significant comparative advantage across student types, allocations that use a heterogeneous estimate of value-added can significantly raise student test scores.

Chapter 3 describes a field experiment where my coauthor and I explore a light-touch intervention that modifies the language in job postings surrounding required qualifications in order to reduce application gaps for women and other underrepresented individuals. We do so using a large-scale, "reverse audit study" field experiment where we randomize the content of job ads and observe job seeker behavior. Specifically, we established a non-profit firm to act as an intermediary in the job search process. This firm reposts real job ads and collects information from job seekers interested in applying. We randomize whether we encourage people to apply even if they don't meet all of the listed qualifications and whether we inform them that companies routinely hire individuals who do not have all qualifications.

Chapter 1

Broader Horizons: The Long-Run Impacts of Exposure to New Places

Tanner Eastmond and Michael Ricks[0]

Abstract

We study how volunteering in different cultural environments shapes individuals' social attitudes and actions using variation from the location assignments of volunteer missionaries for The Church of Jesus Christ of Latter-day Saints. Administering an original survey to former volunteers, we find noticeable changes in attitudes about underrepresented minorities, political out-partisans, and women in the workforce. We find that volunteering in places with high Black or Latino populations increases positive sentiment towards these groups and that volunteering in places with higher government spending change real life behaviors like donating to or volunteering for political causes. These effects persist for decades after missionary service. Although we don't find large effects on attitudes about working women, women who volunteer in more gender-equitable places

[0]Department of Economics, University of California San Diego and the Department of Economics, University of Nebraska-Lincoln. Authors can be reached at teastmond@ucsd.edu and mricks4@unl.edu. This research is conducted under UCSD IRB #808966.

may have more children.

1.1 Introduction

Young adulthood is a critical time for preference and identity formation, as many youth explore their role as increasingly independent behavioral agents. With this in mind, many organizations provide formative opportunities for young adults to travel, work, study, and volunteer in new places. These opportunities drive interesting development for young individuals. For example, social attitudes, labor market outcomes, and migration are strongly affected by horizon-expanding experiences like college enrollment (Malamud & Wozniak, 2012); study abroad programs (Di Pietro, 2012; Oosterbeek & Webbink, 2011; Parey & Waldinger, 2011); national, religious, and humanitarian service (Berinsky, Karpowitz, Peng, Rodden, & Wong, 2022; Mo & Conn, 2018); and military service (Card & Cardoso, 2012; Ertola Navajas, López Villalba, Rossi, & Vazquez, 2022). Interestingly, there is growing evidence that in addition to the extensive-margin effects of these programs, exposure to different types of places has notable effects on a variety of outcomes like social attitudes about national identity (e.g. Bagues & Roth, 2023; Bazzi, Gaduh, Rothenberg, & Wong, 2019; Okunogbe, forthcoming). Despite this growing understanding that these experiences in new locations matter, much less is understood about what types of experiences contribute to these changing economic preference and for whom. This paper aims to study the formation of attitudes towards under-represented minorities, mothers who work outside of the home, and political out-partisans and changes in behaviors related to these attitudes.

To understand how place shapes social attitudes, we explore the extent to which prolonged exposure to different places has long-run effects on young adults' attitudes and behaviors. Estimating these geographically specific treatment effects is challenging, however, since individuals intentionally choose to live and work in different locations. We address this difficulty by exploiting a large scale natural experiment: volunteer missionaries for The Church of Jesus Christ of Latter-day Saints (the Church). These 18-25 year old volunteers are assigned to locations around the world without regard to their preferences and serve in their assigned location for up to two years. Assignments are made by church leaders who do not know volunteers to staff volunteer

locations around the world. We collect the information available to those leaders at the time of assignment and condition on this information in our analysis. We also show conditional balance across location assignments on characteristics not available to the leaders making the assignment.

To measure the effects of volunteering in different places, we administered an original survey to over 15,000 former volunteers who volunteered in the last 50 years. Although this setting provides an excellent opportunity to study how prolonged exposure to different places impacts young adult's views and behaviors in the long run, identifying a sample of past volunteer missionaries is challenging. To solve this issue, we sample from a population that is highly likely to have participated in this volunteering in the past: alumni of Brigham Young University (BYU), a university affiliated with the Church of Jesus Christ of Latter-day Saints. We deploy an original survey to BYU alumni and identify past volunteers of all ages. Our survey elicits information on where each person served, what information they provided to the Church when they applied to volunteer, and their present views and behaviors pertaining to race, politics, and gender roles. We combine this survey information with public data about place characteristics by digitizing maps of mission boundaries over time.

We find long-term impacts of exposure to different locations on a person's attitudes and behaviors. In particular, we find that assignment to places with a high shares of Black or Latin American people increases positive sentiment towards those groups. Moving from the 25th to the 75th percentile of either the local Black or Latin American population share increases attitudes toward the respective group by 3.5 percentiles. Changes are not limited to stated attitudes. Volunteering in these places also increases the probability of living in a diverse zipcode as an adult, volunteering or donating to social justice causes, and voting for minority candidates for national office. As we begin to explore mechanisms we do not seem to find any impacts of volunteering in places with more favorable racial attitudes, but instead find effects stemming from volunteering in places with more people who would be under-represented minorities in the United States. It seems that these effects are driven by having a more positive, personal experiences with local residents in these places. Volunteers assigned to these locations spent more time visiting people in their homes

and report that people are more kind and more receptive.

Although the results are preliminary, we also consider attitudes about political out-partisans and about working mothers. For example, volunteering in a place with more extensive government spending shapes feelings about political out-partisans. Because volunteer missionaries tend to have fairly conservative backgrounds, moving from the 25th to the 75th percentile of per-capital expenditure reduces affective polarization (measured as the gap between feelings towards republicans and democrats) by 28%. This is substantiated in voting and donation behaviors as well. In contrast, while we do find that mission assignment also affects stated attitudes about mothers who pursue education or careers (and not those who stay at home), the effects are not related to the social or economic standing of women in the assigned locations. Despite the null results on attitudes, the results do suggest that for female missionaries volunteering in these missions increases their optimism about marriage—increasing marriage rates and fertility.

Our paper relates most closely to a set of papers that study the impact of exposure to new places on immigration views (Berinsky et al., 2022) and national integration (Bagues & Roth, 2023; Bazzi et al., 2019; Okunogbe, forthcoming). Methodologically, the most closely related paper to ours is Berinsky et al. (2022)[1], which explores the same context as ours—that of Latter-day Saint missionaries—to examine the impact of contact with immigrants on immigration attitudes. They survey BYU students planning to serve a mission before they are assigned to any location then follow up directly after. They find that missionaries assigned to places with a higher likelihood of interacting with immigrants express more pro-immigrant attitudes. They also provide strong evidence that location assignments in this missionary program are conditionally independent of prior social views. In addition to studying outcomes beyond immigration, our contribution is measuring long-term effects on attitudes and actions—we are able to study the long run effects of exposure to different places since we survey former volunteers 10-50 years after service rather than directly upon their return. Furthermore, rather than relying on stated attitudes, which are prone to social desirability bias, we measure impacts on related behaviors.

[1]Crawfurd (2021) studies a similar question using a convenience sample.

Other closely related studies use random variation in youth service in Nigeria (Okunogbe, forthcoming), military service in Spain (Bagues & Roth, 2023), and population resettlement in Indonesia (Bazzi et al., 2019) to explore the short and long run impacts of intergroup exposure on national integration. Because our variation is worldwide, our respondents can be assigned to over 400 different "treatments" (locations). With this broad variation, our survey responses, and our detailed ancillary data about place characteristics, we are able to understand how exposure to different places impacts individuals broadly and the mechanisms through which these effects may propagate, including how a person's background characteristics play an important role in shaping the changes they experience.

We also speak to the broader literature on "contact theory" (Allport, 1954), where interaction with individuals from different groups can reduce prejudice towards those groups. This has been studied in a variety of contexts, including random roommate assignment (for example Baker, Mayer, & Puller, 2011; Boisjoly, Duncan, Kremer, Levy, & Eccles, 2006; Carrell, Hoekstra, & West, 2019; Marmaros & Sacerdote, 2006; Van Laar, Levin, Sinclair, & Sidanius, 2005), assignment across schools (Billings, Chyn, & Haggag, 2021; Kaplan, Spenkuch, & Tuttle, 2019; Rao, 2019), sports teams (Lowe, 2020; Mousa, 2020), or military assignments (Dahl, Kotsadam, & Rooth, 2021; Schindler & Westcott, 2021). We build on these studies by finding that while contact with people from a different background is one important driver of changes in attitudes, it is only one piece of a bigger picture when considering the impact of exposure to different places. We find that attitudes of residents with similar demographic characteristics and prevailing local institutions may also play an important role in changing a person's views and behavior.

Finally, we relate to the literature on place effects more broadly. Because of the endogeneity of location choice, many place effect papers often use mover designs. Prominent movers designs include estimating the effects of place on children's later-life earnings (e.g., Chetty & Hendren, 2018a; 2018b), workers earnings (e.g., Card, Rothstein, & Yi, 2023), consumer behavior (Bronnenberg, Dubé, & Gentzkow, 2012), and medical spending and mortality (e.g., Finkelstein, Gentzkow, & Williams, 2016). We have two main contributions to this literature. First, because

location assignment is independent in our setting, we are able to estimate place effects without relying on a movers design. Second, rather than focusing on the effects of place on *economic outcomes* (like income, consumption, and health), our study is trying to measure the effects of place on *economic primitives* measured in attitudes towards different groups.

The rest of the paper is organized as follows. Section 2 describes the institutional context, followed by a discussion of the data in Section 3. Thereafter we describe the empirical strategy in Section 4 and the results in Section 5. We finish by describing mechanisms for the results in Section 6 and conclude in Section 7.

1.2 Missionary Service in The Church of Jesus Christ of Latter-day Saints

In this section, we describe the Church's missionary program and the process for assigning volunteers to locations around the world.

1.2.1 The Mission Program

The Church's missionary program is a global program for unpaid volunteers ages 18-25.[2] The primary purposes of the program include proselytizing, strengthening church members globally, providing community service, and the personal development of volunteers. The missionary program has been running for more than a century and in 2024, over 65,000 full-time missionaries volunteer in 450 different locations (called "missions") across world. Young missionaries are drawn from the membership of the Church. Among actively participating church members we estimate that approximately 75-85% of young men and 45-50% of young women volunteer as missionaries[3] (participation is considered a responsibility for young men—who serve for 24

[2]Although volunteers pay some their own expenses, the Church averages and subsidizes costs. This means that the volunteers pay a standard amount based on where they come from rather than where they are assigned to (this way someone assigned to live in Tokyo pays the same amount as someone assigned to live in the Dominican Republic despite cost of living differences). Those who cannot pay the flat fee are provided for by local donations or general church funds.

[3]Based on statistics published by the Church, there are roughly 30,000 missionaries out of an estimated 130,000-150,000 nominal members of any given age. Combining that with the fact that in recent years 35% of young missionaries are women and that roughly one third of nominal members are actively participating in the Church, and assuming that members are roughly evenly split between men and women yields 76-86% for men and 44-50% for women.

months—and is optional for young women—who serve for 18). The young volunteers in each mission are organized under a few adult volunteers who also oversee mission logistics.[4]

After prospective volunteers fill out an application, they are assigned specific service dates and a mission location from church headquarters. The size of this location varies from part of a metro-area to multiple countries. Over the course of their service, volunteers rotate to different locations or congregations within the mission every few months. Despite this regular rotation *within* mission, volunteers generally do not leave the mission location during their service. At the conclusion of their service, volunteers are required to return to their hometown.

Missionary service begins with 2-9 weeks of standardized training, religious study, and language learning. Thereafter, volunteers go to their assigned missions, where they are paired up with another volunteer and assigned to serve a local congregation. Typical daily activities include talking to people in the community about the Church and Christianity, teaching individuals in their homes, visiting with members of the local congregation, attending local church meetings and meetings with other missionaries, participating in religious and language study and preparing lessons, and taking part in formal and informal community service.[5] An important aspect of this program is that volunteers are given the explicit commission to interact with as many people as possible, to get to know them, and seek to develop love and understanding for them. This is particularly interesting because these interactions will often be particularly salient and authentic (relative to a typical person living and working in the area). This commonly leads to lasting relationships and communication between the volunteers and those whom they met volunteering.

1.2.2 Mission Location Assignment

Our identification hinges crucially on the quasi-experimental assignment of volunteer missionaries to service areas. Specifically, to explore the causal impact of place on a person's social attitudes and behavioral outcomes, location assignment must be conditionally independent of the

[4] Including housing, travel, stipends, visas, etc.

[5] Although these activities constitute most of what missionaries will do anywhere in the world, the specific mix of activities varies greatly by location.

unobserved determinants of these outcomes (such as baseline characteristics). There is precedent for using this variation, as several studies have done to explore different questions (Berinsky et al., 2022; Crawfurd, 2021; Pope, 2008).

Prospective volunteers initiate the mission assignment process by filling out an application. The specific questions changes over time, but this application has generally included availability dates for service and information on basic demographics, church participation, education, language learning, family living situation, and information from general medical check-ups. Applicants interview with local ecclesiastical leaders who know the volunteer personally. These interviews are standardized and are intended to accomplish at least three purposes: (1) determine whether the person is living church standards, (2) understand whether applicants' physical and mental health are sufficient for the rigors of volunteering, and (3) elicit a commitment from the volunteer to go wherever they are ultimately assigned. Local leaders may make comments on the application and send them in to church headquarters. Importantly, during the application process neither the volunteer nor the ecclesiastical leaders make a requests or recommendations for specific (types of) service area.

Mission assignments are made at church headquarters by one of twelve senior church leaders who have many responsibilities. The leader making the assignment begins with a list of staffing needs in mission locations across the world, the volunteer's picture, and application information from that volunteer then makes a location assignment based on thoughtful consideration of that information—including constraints like when the volunteer is available and whether any moderately serious mental- or physical-health challenges limit possible locations. We estimate that the Church made roughly 750 assignments each week over the last 10 years, suggesting that leaders spend less than 3 minutes per application.[6] Volunteers are then told their location assignments,

[6]Over this time period the Church had roughly 70,000 volunteer missionaries serving at a time. The average # assignments to make weekly = 70,000/(average # weeks volunteered), where average # weeks volunteered is just under 104 for men and 78 for women (the Church recommendation is 2 years volunteering for men and 1.5 years for women). Women comprise roughly 25-35% of the volunteer workforce, so average # weeks volunteered is between 93-98. Stated information from the Church also suggests that only 2-4 of these leaders handle this task each week, so unless these senior executives are all spending more than 8 work hours each week assigning volunteers, it must be less than 2.5 minutes per volunteer.

start date, and basic information about preparing to serve. There is extremely low attrition of volunteers after reception of the assignment, so there is little concern of differential attrition based on social views.

In Section **??** we demonstrate the conditional independence of assignment. We document which application characteristics are associated with assignment decisions, but present evidence that church leaders seem to be making the matches based on mission and volunteer characteristics based on openings. Furthermore, we show that conditional on the information in the application, mission assignments are not correlated with volunteers' other baseline characteristics. As such we condition on the information provided at the time of assignment in our analyses and consider location assignments independent of the unobserved determinants of social attitudes and related actions.

1.3 Gathering Data on Volunteers and Locations

Our primary data for this project are collected using an original survey. The survey, sample frame, and mode of administration were all designed after extensive piloting from August 2021-July 2023 (see Appendix 1.A.2 for details). This section discusses our data collection, outlines the survey instrument, and gives a description of the data used for analysis. All surveys were administered via email using the survey platform Qualtrics.

1.3.1 Survey Administration

Identifying and contacting former volunteer missionaries presents a significant challenge. Because the Church does not share confidential data with researchers, we sample from a population that is highly likely to have participated in this program: former students of Brigham Young University (BYU). BYU is owned and operated by the Church, and we estimate that about 50% of former students at BYU served as missionaries, providing a sample frame with a high hit rate for former volunteers.

We identified these likely former volunteers by collecting public directory information and

verifying contact information online. In 2021 the directory information included about 400,000 living BYU alumni, or about 80-95%.[7] We were able to successfully scrape or verify contact information for about 150,000 of these former students from social media, online employee listings, and online white pages.

We sent our survey to 111,950 former students between November 2023 and March 2024. Each individual was sent an invitation to participate with the screening question embedded in the email. Those who had not completed the survey one and two weeks after the original email received short follow-up emails. Qualtrics reported that 74,626 of the individuals opened an email we sent, and 32,586 former students answered the screening question. Of these, 18,321 served a mission as a young adult and consented to take the survey. Given our estimate that 50% of our sample frame was eligible, this constitutes a roughly 33-49% response rate.[8] We consider this rate is very high given that respondents were *not* paid for participation in the survey.

1.3.2 Survey Instrument

After consenting to participate respondents answered six main blocks of questions. The median time to completion was 25 minutes and 13% of individuals who began the survey did not complete it.

Mission Information. The first block of the survey was about mission assignment. First we asked individuals what location they were assigned to, what language they volunteered in, and what years they volunteered. Then they reported what they put in their mission applications including gender; education and work experience (including high school GPA and college applications/enrollment decisions); experience and interest in traveling out of the country; interest, experience, and aptitude for learning a language; leadership experience; activity in the Church; previous family mission assignments; and whether there were any medical conditions flagged in the application process. This block concluded by asking about additional baseline characteristics

[7] This number comes from comparing our data to IPEDS data. Since our coverage of alumni drops off substantially after 2016, we limit our sample to those who started their volunteering during or before 2010 (two years for volunteering, four years to graduate BYU).

[8] 18,321 out of 55,975 total possible email addresses and out of 37,313 opened emails.

that were not reported on the mission application such as childhood zipcode, parental education and employment, and subjective assessments of the underlying motivations for choosing to serve.

Life Outcomes. The second block of questions was the shortest. Because we were worried that some individuals would choose to atritt once they saw questions related to race, we asked our key behavioral questions first: current zipcode of residence, marital status, number of children, and current attachment to the Church. We use the zipcode to measure what type of location individuals choose to live in. We are also in the process of linking our survey responses to voting registration and donation records to obtain additional behavioral outcomes.

Stated Attitudes. The third block of questions focused on stated attitudes about underrepresented minorities, political partisans, and mothers working outside of the home. We measured all stated attitudes using the standard "feelings thermometer" from 0-100 as used in the American National Election Studies and the General Social Survey (). Figure 1.1 shows the graphic respondents interacted with.[9] We asked for feelings toward Blacks, Latinos, and Whites; Republicans, Democrats, and Political Independents; and Mothers Pursuing Careers, Mothers Pursuing Education, and Stay-at-Home Mothers.[10]

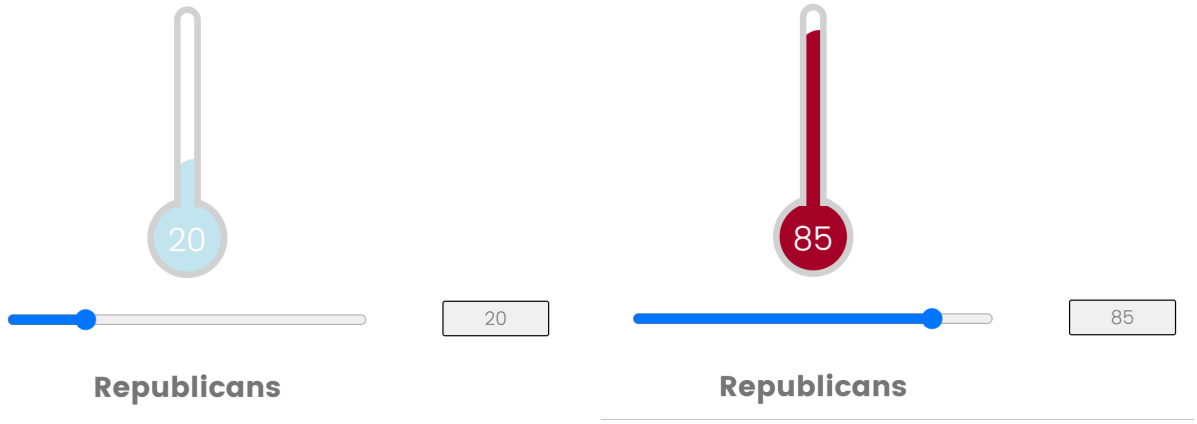


Figure 1.1: **Feelings Thermometer Example**

[9]Respondents were given the following instructions: "Thank you for sharing information about yourself and your mission. Now, we would like to gauge your feelings toward different groups of people using a 'feeling thermometer' rating system. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm towards the group. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable towards them. You would rate the group at the 50 degree mark if you feel neutral towards them."

[10]To avoid anchoring affects we randomized the order of the domains and groups within domains at the individual level.

Related Behaviors. After measuring stated attitudes about these groups, the fourth block of questions asked individuals about behaviors related to their social attitudes. We asked individuals whether they had participated in various activities in the last five years. The activities included (1) reading a book or listening to a podcast (about race, gender norms, or Republican/Democratic policies), (2) donating or volunteered (for a social justice cause, Republican/Democratic political cause, or the Church), (3) voting in a national election (for a minority, female, or Republican/Democratic candidate), and (4) protesting police violence or masking/vaccine mandates. For behaviors relating to working women we also ask about respondents' and spouses' time allocations between working for pay, housework, dependent care, and leisure.[11]

The block of related behaviors also included a real-stakes donation activity. We adapt this activity from Exley (2020) and allow participants to choose between donations of different sizes to different charitable organizations (see Figure 1.2).[12] Each respondent was asked whether they would prefer any donation, a $0.25 donation, a $0.50 donation, a $1.00 donation, or no donation to a given nonprofit relative to a $0.50 donation to the American Red Cross. The organizations were the National Association for the Advancement of Colored People (NAACP), the National Partnership for Women and Families, the Republican party, the Democratic party, and the Church.

Mechanisms. The fifth block of survey questions aimed to clarify the channels through which effects operate. As such we asked respondents about their experiences while volunteering. For example, we asked how they spent their time; what their interactions with others were like; whether they had conversations about race, politics, or gender; and to what extent they felt like related institutions functioned better or worse in their mission locations relative to at home.

Other Demographics. The survey concludes with some additional demographics, like race, education, household income, political orientation, and perceived parental orientation.

[11] About 89 percent of our sample are currently married.

[12] Respondents were given the following instructions: "The following questions will present you with a series of choices. You can choose to have us donate 50 cents to the American Red Cross (on the left) or donate 50 cents to a different organization (on the right). When our survey is complete, we will donate $950 to these organizations based on the answers you and other participants give."

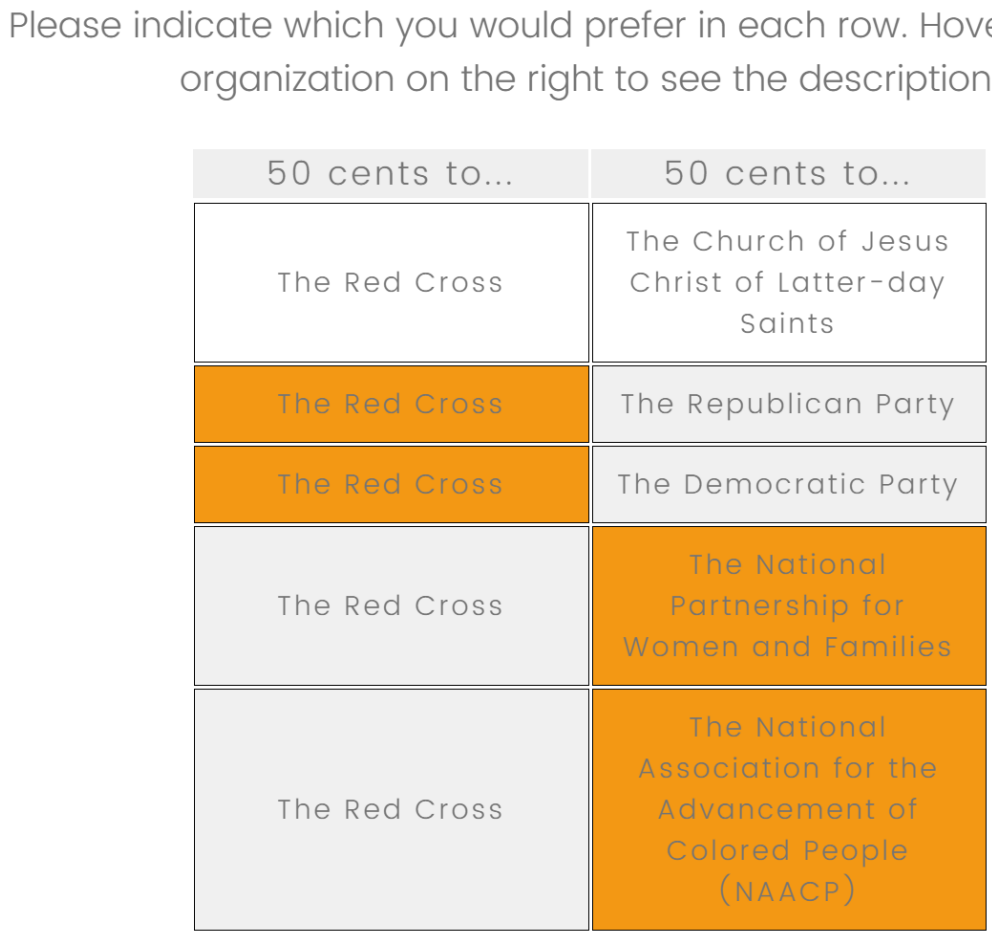


Figure 1.2: **Donation Activity Example**

1.3.3 Information about Mission Locations

In addition to the survey data, we collect information about mission locations to characterize different places. Conceptually speaking, a volunteer who was assigned to a location today would have 450 different possible "treatments" since, in principle, they could be assigned to any one of the possible mission locations around the world. Each location comes with a bundle of characteristics, that may characterize the effects. We collect measures of these places across the world and over time.

To collect information on the characteristics of each location, we digitized and geotagged maps for each mission over time. This allowed us to aggregate information for each mission to the mission level. We include demographic information, social attitudes, and measures of institutions

related to race, partisans, and working mothers. Demographic information came from census data for countries around the world. These items include the fraction of people who were Black or Latin American, average family size and average age at first marriage, and average age, fraction rural, and fraction with less than college education. Social attitudes come from the World Values Survey, the General Social Survey (GSS), and Project Implicit. These include things like 'Black-White', 'Light Skin-Dark Skin', 'Gender Career', and 'President Popularity' implicit association tests[13]; feelings thermometers towards Black people, Republicans, and feminists; and direct questions about attitudes such as asking whether children suffer when the mother works outside of the home or if the person would be uncomfortable with a neighbor of a different race. For institutions we use the Freedom House Civil Liberties Index (a measure of how well an area's institutions support civil liberties for underrepresented groups), the Gender Equality Index from the Human Development Reports, and the amount of government spending per capita.[14]

1.3.4 Sample Description

Our main analysis sample come from 15,647 former volunteers whose began their missionary service between 1970-2010. Appendix Tables 1.6 and 1.7 describe the characteristics of these volunteers before volunteering and now, separated by which decade they started their volunteering. Our sample includes about 20% women over time and is predominantly white. An decreasing fraction attended some college before leaving to complete their volunteering, and three quarters report having grown up with Republican parents. About a third spoke another language at the time of application. Currently, most former volunteers are married with children, over 85% completed a Bachelor's Degree and over 50% of the total completed a graduate degree. A majority of the former volunteers are Republicans, but the share of Democrats increases across cohorts to just under one fourth in the 2000-2010 cohort. Nearly 85% of respondents report being actively participating members of the Church of Jesus Christ of Latter-day Saints.

[13]These are meant to be measures of implicit preference over groups. In our case we use them to measure implicit preference against people of color, women in careers, and Democrat Presidents.

[14]Generally we have these data aggregated to the county-year level within the US and to the country-year level elsewhere. The data from project implicit are currently aggregated over time to increase precision.

1.4 Empirical Strategy

1.4.1 The Conditional Independence of Mission Assignment

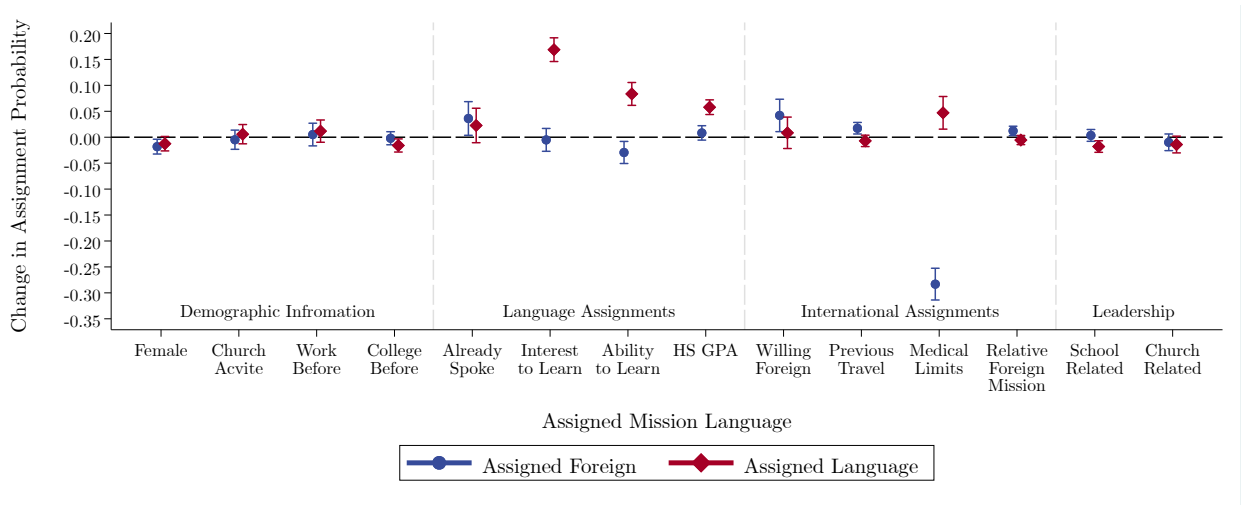
Identification depends on whether the unobserved determinants of social attitudes are truly independent of location assignment conditional on the information in the missionary application. Speaking to this question first requires understanding how the available information is used in deciding where to assign volunteers. We estimate two regressions estimating

$$\begin{aligned} \text{leave\_US}_i &= \gamma_1 X_i + \text{non\_English}_i + u_{1,i} \\ \text{non\_English}_i &= \gamma_2 X_i + \text{leave\_US}_i + u_{2,i} \end{aligned} \tag{1.1}$$

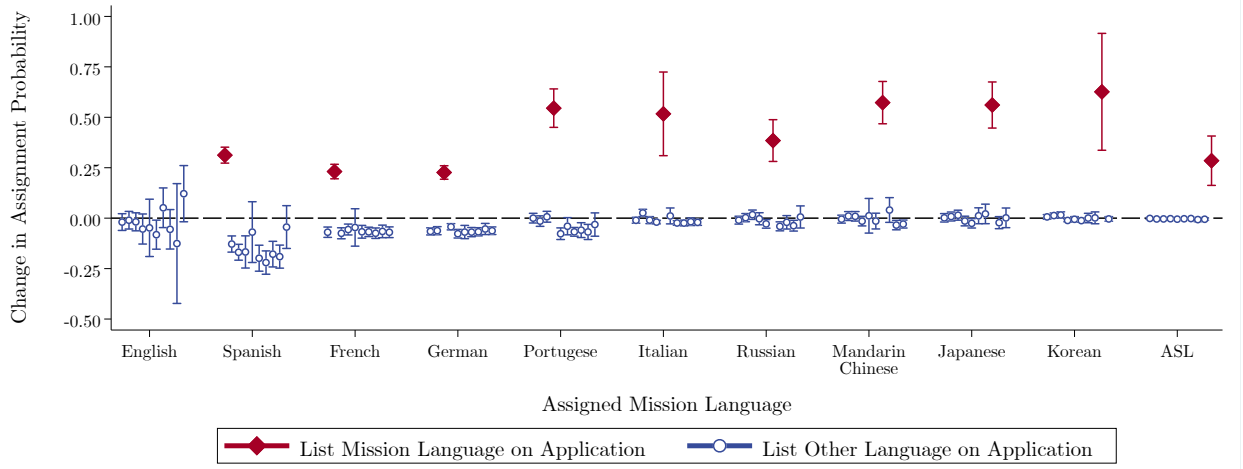
where X_i are mission application characteristics including demographics (gender, education, work experience), language preparedness (intermediate proficiency, language indicators, interest in learning a language, ability to learn a language, high-school GPA), international preparedness (willingness to leave country, foreign travel experience, family members with international missions, and location limiting medical conditions), other experiences (leadership in high school, leadership in the church, activity in the church, completion of church curricula), and year and month of application fixed effects. Because assignment to learn a new language and to leave the country are correlated with one another, we control for the other in each outcome's regression.

Panel (a) of Figure 1.3 shows that the information from the volunteer application is very important in making location assignments. Demographics have relatively small but statistically significant effects.[15] Language-related information seems to be strongly (but not universally) taken into account for language assignment and seems to some what affect domestic versus foreign assignments. Experience abroad affects foreign assignments but not as strongly. Leadership experience is essentially independent of mission assignment, likely because church leaders intentionally want a roughly uniform distribution of potential leaders across missions.

[15] We find for example that women are slightly more likely to be assigned to the US. This finding is consistent with Berinsky et al. (2022), who report the same when considering mission assignments.



(a) Assignment to Foreign and Non-English Speaking Missions



(b) Assignment to Specific Mission Languages

Figure 1.3: **Mission Application Information Shapes Mission Assignments**

Examining specific language assignments, Panel (b) of Figure 1.3 demonstrates the importance of prior language experience. The figure is based off of analogous regressions to Equation 1.1 where the dependent variable is assignment to volunteer speaking a certain language, and we report coefficients for intermediate proficiency various languages before hand. We focus on the ten most common languages known before: Spanish, French, German, Portuguese, Italian, Russian, Mandarin Chinese, Japanese, Korean, and American Sign Language.[16] We find that intermediate proficiency with a given language before beginning missionary service is associated with 20-60

[16]These are similar to the 10 most commonly assigned languages, with the exception of ASL.

percentage point increases in assignment probability to volunteer in that language.

With an understanding of how application information influences location decisions, we also consider whether baseline characteristics that were not reported on the mission application are conditionally balanced across locations. To do this, we extend the regressions in Equation 1.1 by including additional unobserved characteristics: maternal employment (part-time or full-time versus stay at home), parental political leaning (strong Republican or independent/Democrat/strong Democrat versus Republican), parental education (no college degree or graduate/professional degree versus bachelors degree), and childhood zipcode characteristics (share Black or Latino, Republican vote share, and Gender Equality Index above sample mean).

Figure 1.4 shows that these baseline characteristics are generally balanced across US versus foreign assignments and English speaking versus foreign language speaking assignments. Note that these items are strongly correlated with social attitudes, but conditional on the other information in the application they do not covary with mission assignments. Importantly, the scale of Figure 1.4 is dramatically smaller than those in Figure 1.3, so we can rule out even relatively small changes in assignment probabilities related to these characteristics.

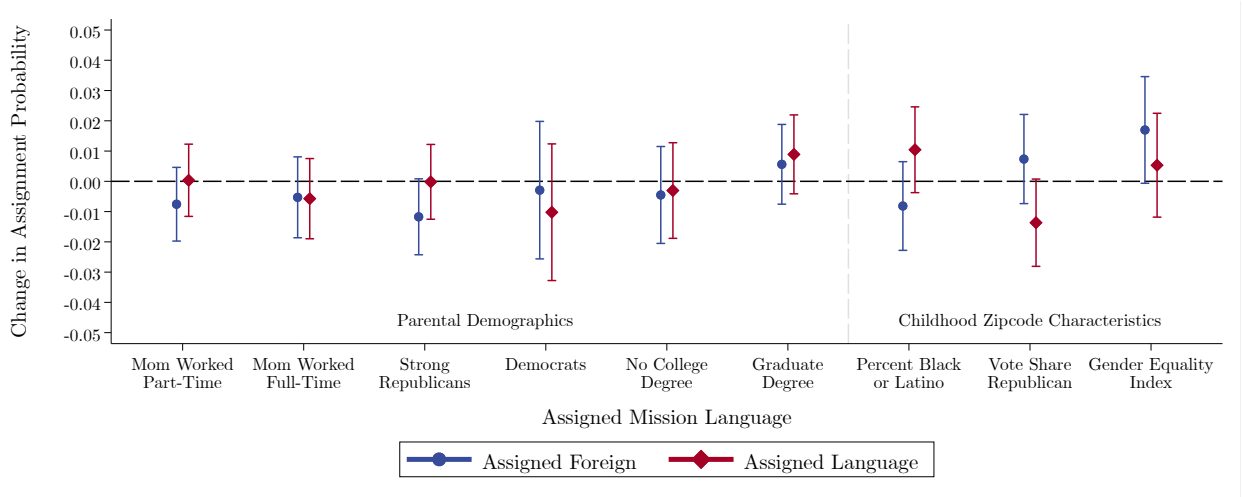


Figure 1.4: **Mission Characteristics Are Unrelated with "Unobserved" Baseline Characteristics**

1.4.2 Estimation

Using the plausibly exogenous location assignments, we can contrast volunteer missionaries assigned to different locations to compare their outcomes. Since we have many different individual outcomes for each domain, we combine each relevant outcome measure into an index measure. For stated attitudes we make a standardized index for each, e.g. stated racial attitudes. Our measure of behavioral outcomes is the sum of the number of relevant behaviors. For donations, we take each decision in our survey that the participant faced/could have faced (i.e. donate to Red Cross at any amount, donate when Red Cross is double, donate when equal to Red Cross, donate when organization is double, donate at any amount to organization) and we took the sum of each decision they made towards the organization. This means a person with a score of 5 would always donate to the organization and a person with a score of 0 would never donate, etc.

With these outcomes in mind, our pre-specified estimating equation is the following:

$$y^d_{i,t} = \beta_0 + \beta_1 Demographics_{j(i,t)} + \beta_2 Attitudes_{j(i,t)} + \beta_3 Institutions_{j(i,t)} + X'_{i,t}\delta + \gamma_{g(i} + u_i \quad (1.2)$$

where $y^d_{i,t}$ is the outcome index for domain $d \in \{race, gender, politics\}$ for individual i who started their mission in year t. Each individual is assigned to mission location $j(i,t)$. Location characteristics for each domain d are included by combining individual measures from each category of demographics, attitudes, and institutions into a standardized, covariance weighted index to provide one measure of each. We then include the standardized indices $Demographics^d_{j(i,t)}$, $Attitudes^d_{j(i,t)}$, and $Institutions^d_{j(i,t)}$ for assigned location $j(i,t)$ in year t. Each of these is a vector including the indices for each outcome domain, including race, political out-partisans, and working women.

As shown previously, conditioning on characteristics included in the mission application is crucial for identification. To this end we include X_i, a vector of mission application controls including willingness to be assigned outside of home country, whether any health conditions were

flagged in the application, whether they graduated from seminary for the Church,[17] whether they participated in extracurricular activities in high school, whether they performed well academically in high school, if they had leadership opportunities in their local church congregation and/or high school activities, whether they were willing to learn a language on their mission, their gender, pre-mission educational attainment, whether they spoke another language prior to assignment, frequency of church participation during high school, indicators for languages they spoke proficiently prior to their mission service, and indicators for where their family had previously served missions. These are the relevant items available to the senior church leader at the time they made the location assignment for the volunteer. This vector also includes fixed effects for year and month of service.

We also include fixed effects $\gamma_{g(i}$ for the area where the volunteer primarily grew up as a child. All standard errors are clustered at the mission level.

1.5 Results

1.5.1 Regional Patterns in Stated Outcomes

We now show the variation in outcomes for assignments across the world by showing the average outcomes for assignments (aggregated to broad geographic regions with roughly similar numbers of assigned missionaries). These patterns are shown in Figure 1.5. These maps show the average of the stated attitudes index for volunteers assigned to each indicated region conditional on the characteristics available at the time of location assignment. In our sample we have no missionaries assigned to Northern Africa, the Middle East, Central Asian Countries, China, and North Korea. Many of these countries do not allow proselytizing Christian missionaries, so the Church of Jesus Christ of Latter-day Saints does not send volunteer missionaries to these areas. In each figure blue colors indicate a positive effect on affect towards the indicated group, while red colors indicate a negative effect.

In panel (a) we see the variation in stated racial attitudes around the world, and the patterns

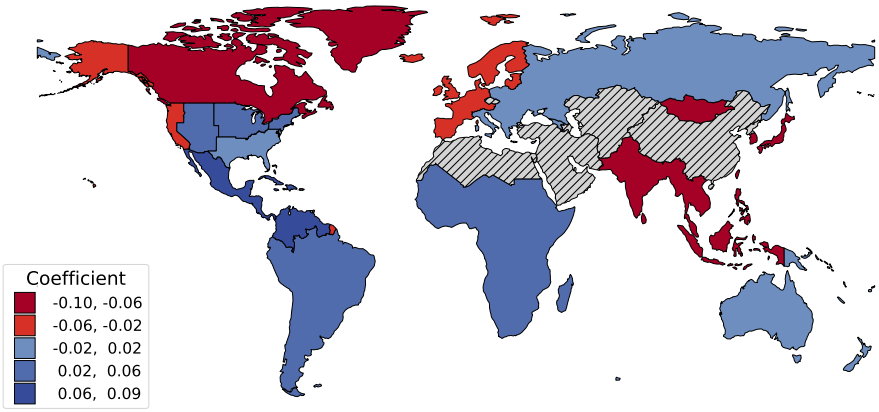
[17] This is a four year bible study class for high school students sponsored by the Church, nearly all teenagers in the Church participate, but not all graduate.

are quite striking. Volunteers assigned to Central America, South America, and Africa see the largest increases in affect towards Black and Latin American people. This is broadly consistent with the thinking that exposure to these groups should lead to more positive affect towards them.

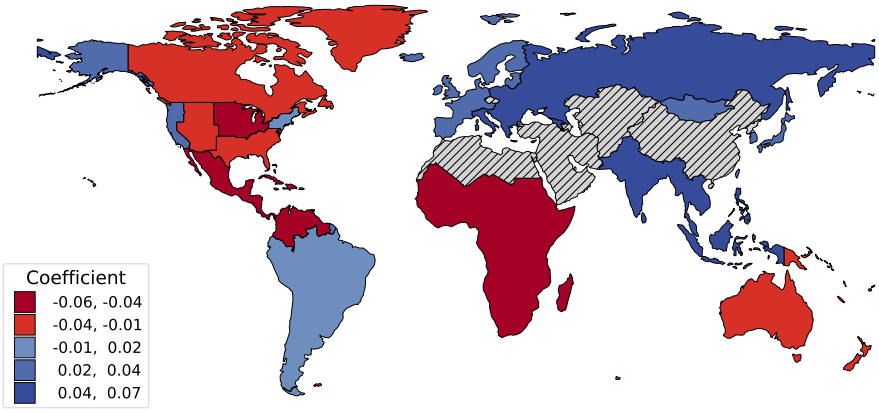
For feelings towards Democrats we see in panel (b) that assignments to Europe and Southeast Asia show the strongest increases in positive sentiment towards Democrats, followed by assignments to New England and the West coast in the US as well as South America. Other regions in the world show decreases in affinity towards Democrats.

Panel (c) shows positive sentiment towards mothers pursuing career or education outside of the home, with assignments to Europe, Africa, Australia, the US outside of the Midwest increasing these positive feelings.

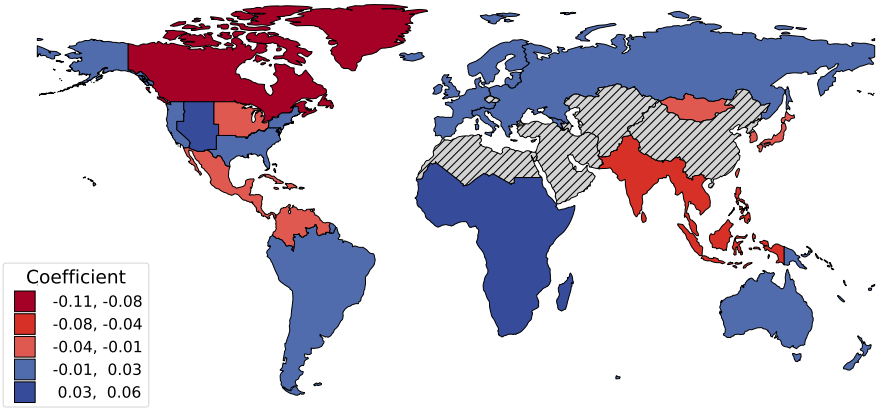
These results highlight one of the unique strengths of our setting. Most other studies exploring related questions relied on variation from people moving or assignment to places with limited differences (e.g. different majority ethnic groups). Conceptually, in our study a person today who decided to participate in the volunteer missionary program we study would have 450 different independent treatments where they could be assigned. This provides an opportunity to understand the different characteristics and experiences from the varied locations that drive the differences we observe in the long-term outcomes of volunteers assigned around the world.



(a) Under-Represented Minorities



(b) Democrats



(c) Mother's Working Outside of the Home

Figure 1.5: **Geographic Distribution of Stated Attitudes**
Notes: Coefficients are the average of the standardized stated attitudes index for each outcome domain conditional on our assignment variables.

1.5.2 Does Mission Assignment Affect Long-Term Outcomes?

Using the plausibly exogenous location assignments, we compare foreign volunteer's outcomes based on mission assignment. First, we examine whether location assignment actually affects attitudes and behaviors to do this we estimate regressions of the form:

$$y_{i,c} = \phi_{m(i,c)} + \beta X_i + e_i$$

where the outcomes of individual i who volunteered in cohor year c are a function of mission fixed effects, $\phi_{m(i,c)}$, and application controls X_i (which include cohort fixed effects as before). We consider regressions of this form with and without the mission fixed effects for nine stated attitudes and for nine behaviors. The stated attitudes are feelings thermometers towards Blacks, Latinos, Whites, Republicans, Democrats, Independents, moms pursuing careers, moms pursuing education, and stay at home moms. The behaviors are voting for a minority candidate, the Latino population share in the respondent's current zipcode, reading a book about race, voting for a Republican, voting for a Democrat, the Republican vote share in the respondent's current zipcode, voing for a women, being married, and the number of children.

Table 1.1 demonstrates the important role mission locations have in forming social attitudes. The for each of the nine stated attitudes and behaviors we report a Joint F-Test that the mission fixed effects are all equal to zero. The table also reports means, standard deviations, and the (unadjusted) R^2 from the regressions with and without mission fixed effects. This is one of our primary pre-specified analyses.

Stated Attitudes. Panel A shows that mission assignment is a statistically significant predictor of attitudes towards Blacks, Latinos, Democrats (only at the 0.1 level), mothers with careers, and mothers pursuing education. Because social attitudes are very idiosyncratic there is still significant residual variation, but the inclusion of mission fixed effects typically has about as much explanatory power as the baseline controls. Interestingly, across the board we find less evidence that missions move feelings towards groups these former volunteers were typically more exposed

Table 1.1: **Mission Assignment Shapes Attitudes and Behaviors**

Panel A: Stated Attitudes	Blacks	Latinos	Whites	Republicans	Democrats	Independents	Mothers with Careers	Mothers in School	Stay-at-Home Mothers
Mean	86.3	87.6	85.9	58.1	54.0	71.1	80.7	86.9	87.6
SD	(17.6)	(16.3)	(17.6)	(28.3)	(27.6)	(22.7)	(20.9)	(16.3)	(16.8)
Respondents	13,090	13,209	13,044	13,026	12,720	11,243	13,470	13,707	13,623
R^2 (Assignment Controls)	0.035	0.033	0.043	0.059	0.032	0.014	0.023	0.021	0.028
R^2 (+ Mission FE)	0.088	0.084	0.092	0.107	0.083	0.070	0.074	0.073	0.073
Joint F-test (FE=0)	$[p = 0.005]$	$[p = 0.023]$	$[p = 0.111]$	$[p = 0.123]$	$[p = 0.098]$	$[p = 0.308]$	$[p = 0.028]$	$[p = 0.003]$	$[p = 0.520]$

Panel B: Behaviors	Vote Minority	Zipcode % Latino	Read a Book on Race	Vote Republican	Vote Democrat	Zipcode % Republican	Vote Woman	Married	Number of Children
Mean	0.485	17.2	0.391	0.718	0.462	52.9	0.580	0.863	4.0
SD		(11.9)				(16.0)			(1.7)
N	14,235	13,913	14,268	14,253	14,239	13,859	14,236	14,802	13,819
R^2 (Assignment Controls)	0.018	0.017	0.031	0.061	0.037	0.041	0.017	0.032	0.097
R^2 (+ Mission FE)	0.062	0.066	0.080	0.107	0.081	0.089	0.059	0.079	0.143
Joint F-test (FE=0)	$[p = 0.515]$	$[p = 0.050]$	$[p = 0.008]$	$[p = 0.048]$	$[p = 0.349]$	$[p = 0.056]$	$[p = 0.792]$	$[p = 0.019]$	$[p = 0.029]$

to (Whites, Republicans, and Stay-at-home mothers). This suggests that part of what makes missions so formative for underlying preferences is that they expose individuals to new people and experiences. Also note that some participants were uncomfortable with the feelings thermometer, especially for race, so the changing sample sizes are due to differences in missingness.

Behaviors. If the changes in stated attitudes really reflect differences in underlying preferences and feelings towards other groups, we should expect mission assignment to also affect behaviors, which it does. Panel B shows that mission assignment changes the zipcodes former volunteers live in, whether they read books about social issues, who they vote for, and even whether they marry and how many children they have.

1.5.3 Attitudes towards Under-Represented Minorities and Related Behaviors

The previous sections demonstrated the importance of the assignments around the world on volunteer's long term outcomes and showed evidence that assignments to different mission locations do, in fact, cause long term changes in attitudes and behavior. We now estimate equation (1) for each individual outcome for the survey to more clearly understand and quantify the effects of exposure to different types of places. In each of the tables including individual survey items in this and the following two sections we present the Benjamini, Krieger, and Yekutieli (2006) sharpened

two-stage q-values in brackets to account for the fact that we are testing multiple hypotheses (see also Anderson, 2008). We use this procedure to adjust our p-values within each outcome domain (race, gender, and politics) for each treatment arm (demographics, attitudes, and institutions).

We show the stated outcomes for Race in Table 1.2 and the behavioral outcomes for Race in Table 1.3.

Table 1.2: **Individual Stated Outcomes for Race**

	FT Black	FT Black> Av FT Black	FT Black> FT White	FT Latino	FT Latino> Av FT Latino	FT Latino> FT White
Resident Frac Black/Latino	0.478	0.008	0.007	0.449	0.010	0.015*
	[0.207]	[0.252]	[0.250]	[0.207]	[0.209]	[0.080]
Resident IAT Black	-0.303	-0.014	0.003	-0.402	-0.009	-0.001
	[0.404]	[0.134]	[0.675]	[0.242]	[0.242]	[0.742]
Civil Liberties Index	-0.332	-0.011	0.007	-0.259	0.004	0.007
	[0.555]	[0.555]	[0.555]	[0.664]	[0.713]	[0.555]
Control Means:						
	86.2	0.481	0.104	87.2	0.312	0.132
Observations	11173	11173	11014	11276	11276	11055

Notes: * p<0.1, ** p<0.05, *** p<0.01. Benjamini et al. (2006) sharpened two-stage q-values in brackets. Estimates in each column reflect estimates from equation (1), which includes controls for reported medical conditions, race, graduation from seminary, participation in extracurriculars in high school, leadership opportunities in high school and church, whether they spoke a language pre-mission, whether they attended some college pre-mission, their participation level in the Church pre-mission, their sex, and fixed effects for decade of service. Each outcome with 'FT ...' is the number reported by the respondent on the corresponding feelings thermometer (FT). Each outcome with 'FT ... >Av FT' is an indicator for if the respondent put responded higher than the average respondent. 'Resident Frac Black/Latino' is a standardized index the fraction of Black and Latin American residents, 'Resident IAT Black' is a standardized index for the average Implicit Association Test value for feelings towards Black individuals demonstrated by residents in the area, and 'Civil Liberties Index' is the Freedom House Civil Liberties index for the assigned area, standardized.

We see little movement in stated racial attitudes for assignment to places with more equitable racial attitudes or institutions in rows 2 and 3, but we see some impacts of being assigned to places with more Black or Latino people. In our sample, about 80 percent of respondents answered precisely the same number across racial/ethnic groups. We anticipated that this could be an issue, so randomized the order in which the respondents would encounter the groups for the thermometers. This allows us to still understand the impact on the level of the thermometers, and we see a large and statistically significant increase in the fraction of volunteers feeling more warmly to-

wards Latinos than White people for those assigned to locations with more Black or Latino people. Though the rest of the measures are noisily measured, when we aggregate these into a standardized index we see a statistically significant increase of 0.4 standard deviations in stated warmth towards these groups for every standard deviation increase in the fraction of Black or Latino residents (see Table 1.8)

Table 1.3: **Individual Behavioral Outcomes Race**

	Read Book on Race	Podcast on Race	Donate to Social Just	Volunteer for Social Just	Vote for Minority	Protest Police	Zipcode Diversity	Donate to NAACP
Resident Frac Black/Latino	-0.020*	-0.011	-0.006	-0.007	-0.012	-0.001	-0.002	0.010
	[0.080]	[0.250]	[0.252]	[0.207]	[0.209]	[0.382]	[0.273]	[0.382]
Resident IAT Black	0.009	0.006	0.007	0.009	0.001	0.005	0.001	0.011
	[0.300]	[0.478]	[0.404]	[0.134]	[0.742]	[0.438]	[0.675]	[0.675]
Civil Liberties Index	0.017	0.007	0.013	0.000	-0.006	0.002	0.003	-0.005
	[0.199]	[0.697]	[0.199]	[0.872]	[0.697]	[0.713]	[0.555]	[0.855]
Control Means:								
	0.399	0.618	0.166	0.081	0.480	0.082	0.208	1.91
Observations	12182	12182	12086	12076	12152	12162	11865	10666

Notes: * p<0.1, ** p<0.05, *** p<0.01. Benjamini et al. (2006) sharpened two-stage q-values in brackets. Estimates in each column reflect estimates from equation (1), which includes controls for reported medical conditions, race, graduation from seminary, participation in extracurriculars in high school, leadership opportunities in high school and church, whether they spoke a language pre-mission, whether they attended some college pre-mission, their participation level in the Church pre-mission, their sex, and fixed effects for decade of service. 'Resident Frac Black/Latino' is a standardized index the fraction of Black and Latin American residents, 'Resident IAT Black' is a standardized index for the average Implicit Association Test value for feelings towards Black individuals demonstrated by residents in the area, and 'Civil Liberties Index' is the Freedom House Civil Liberties index for the assigned area, standardized.

A major concern with the results on stated attitudes is whether these reflect the respondent's actual feelings or something else, for example just learning to say more equitable things but not actually changing any core beliefs. The results in Table 1.3 provide some suggestive evidence that these concerns may be well founded. In fact, though we again see little impact for the attitudes and institutions exposure, we see, if anything, negative impacts on behaviors for those assigned to locations with more Black or Latino people. We can see that when we aggregate these into an index, Table 1.8 shows a statistically significant decrease in race related behaviors.

1.5.4 Attitudes towards Political Partisans and Related Behaviors

We turn our attention to stated political outcomes in Table 1.4 and find insignificant but substantial impacts on stated affect towards Democrats for those volunteers assigned to locations with higher per capita government spending. These effects, when aggregated, show a large and statistically significant impact, shown in Table 1.9.

Table 1.4: **Individual Stated Outcomes for Politics**

	FT Republican	FT Republican> Av FT Republican	FT Republican> FT Indep	FT Democrat	FT Democrat> Av FT Democrat	FT Democrat> FT Indep
Urban Index	-0.010	0.003	-0.008	0.270	0.003	-0.001
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Resident IAT Democrats	-0.738	0.000	0.002	-0.178	-0.003	0.004
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Gov. Spending	-0.684	-0.015	0.001	0.635	0.015	0.013
	[0.178]	[0.133]	[0.545]	[0.192]	[0.133]	[0.123]
Control Means:						
	58.9	0.477	0.235	53.7	0.458	0.120
Observations	11123	11123	9324	10865	10865	9242

Notes: * p<0.1, ** p<0.05, *** p<0.01. Benjamini et al. (2006) sharpened two-stage q-values in brackets. Estimates in each column reflect estimates from equation (1), which includes controls for reported medical conditions, race, graduation from seminary, participation in extracurriculars in high school, leadership opportunities in high school and church, whether they spoke a language pre-mission, whether they attended some college pre-mission, their participation level in the Church pre-mission, their sex, and fixed effects for decade of service. Each outcome with 'FT ...' is the number reported by the respondent on the corresponding feelings thermometer (FT). Each outcome with 'FT ... >Av FT' is an indicator for if the respondent put responded higher than the average respondent. 'Urban Index' is the standardized index measuring fraction urban, average age, and fraction with college education, 'Resident IAT Democrat' is the Implicit Association Test measure for residents towards Democrat presidents, and 'Gov spending' is a standardized measure of Government spending per capita.

The behavioral outcomes for politics are even more interesting in Table 1.5. While we do not observe any impacts for being assigned to younger, more urban, more educated areas, or for those assigned to places with more liberal attitudes, we find large, consistent impacts for those assigned to places with higher government spending. Assignments to those types of places lead to a 21% increase in the likelihood of donating to Democrat campaigns or candidates, a 4% higher likelihood to have voted for a Democrat in a national election, and a 15% higher chance to self identify as a Democrat.

These results are especially interesting when comparing to the race results, because while

Table 1.5: **Individual Behavioral Outcomes Politics**

	Book about Democrats	Podcast about Democrats	Donate to Democrats	Volunteer for Democrats	Vote Democrat	Donate Dem Score	Zipcode frac Democrat	Identifies as Democrat
Urban Index	0.004	0.008	-0.002	0.000	0.003	0.014	-0.001	0.003
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Resident IAT Democrats	-0.002	-0.002	0.003	0.001	0.004	0.005	0.002	0.002
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Gov. Spending	-0.003	-0.003	0.012**	0.001	0.016*	0.047**	-0.002	0.019**
	[0.545]	[0.545]	[0.028]	[0.545]	[0.095]	[0.029]	[0.542]	[0.020]
Control Means:								
	0.216	0.506	0.056	0.015	0.443	0.298	0.382	0.131
Observations	12180	12180	12079	12082	12155	10483	11819	12645

Notes: * p<0.1, ** p<0.05, *** p<0.01. Benjamini et al. (2006) sharpened two-stage q-values in brackets. Estimates in each column reflect estimates from equation (1), which includes controls for reported medical conditions, race, graduation from seminary, participation in extracurriculars in high school, leadership opportunities in high school and church, whether they spoke a language pre-mission, whether they attended some college pre-mission, their participation level in the Church pre-mission, their sex, and fixed effects for decade of service. 'Urban Index' is the standardized index measuring fraction urban, average age, and fraction with college education, 'Resident IAT Democrat' is the Implicit Association Test measure for residents towards Democrat presidents, and 'Gov spending' is a standardized measure of Government spending per capita.

the race results were about the demographic makeup of the assigned location, the results relating to politics are driven by the institutions in the area.

1.5.5 Alternate Measures of the Effects of Mission Characteristics on Long-Term Outcomes

Having documented the effects of mission locations on attitudes and actions decades into the future, we now explore the causal effects of place characteristics using an alternate specification. To do this we estimate treatment effects of the form:

$$y_i = \tau \text{PlaceCharacteristics}_{m(i,c)} + \beta X_i + e_i$$

where τ is the causal effect of being assigned to a place with a one percentile higher value of the place characteristic. As in other papers that follow this approach (e.g., Chetty & Hendren, 2018b), this exercise does not describe the causal effect of increasing the characteristic by one percentile, but rather the effect of being assigned to a place with a higher value of the characteristic *and everything that comes with it.* Since we have only just completed the data collection the results

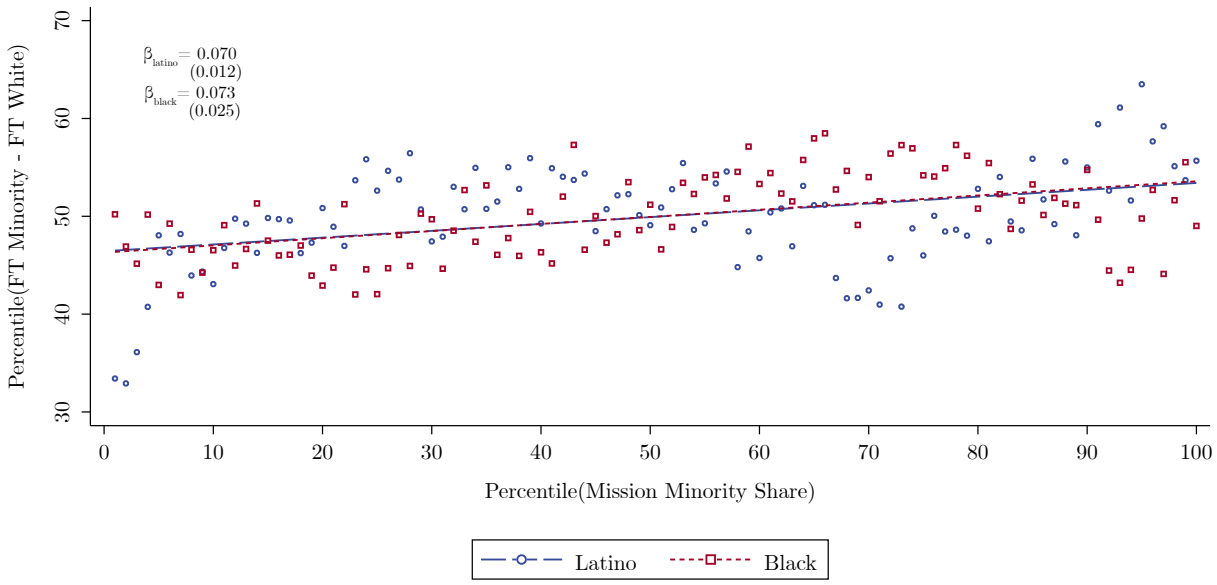
that follow are preliminary, but we have spent the most time on the results about race.

Racial Attitudes. For our analysis of race, we generate new outcome variables that are conditional percentiles. Furthermore because the anchoring effects are particularly pronounced for the racial/ethnic groups (many respondents choose to give the exact same thermometer ranking to all three groups), we specify the outcome as the thermometer towards the underrepresented minority group relative to whites. For behaviors we focus on each groups population share in the respondent's current zipcode. As the independent variable we use the conditional percentiles we use the percent of the mission population from each underrepresented minority group.

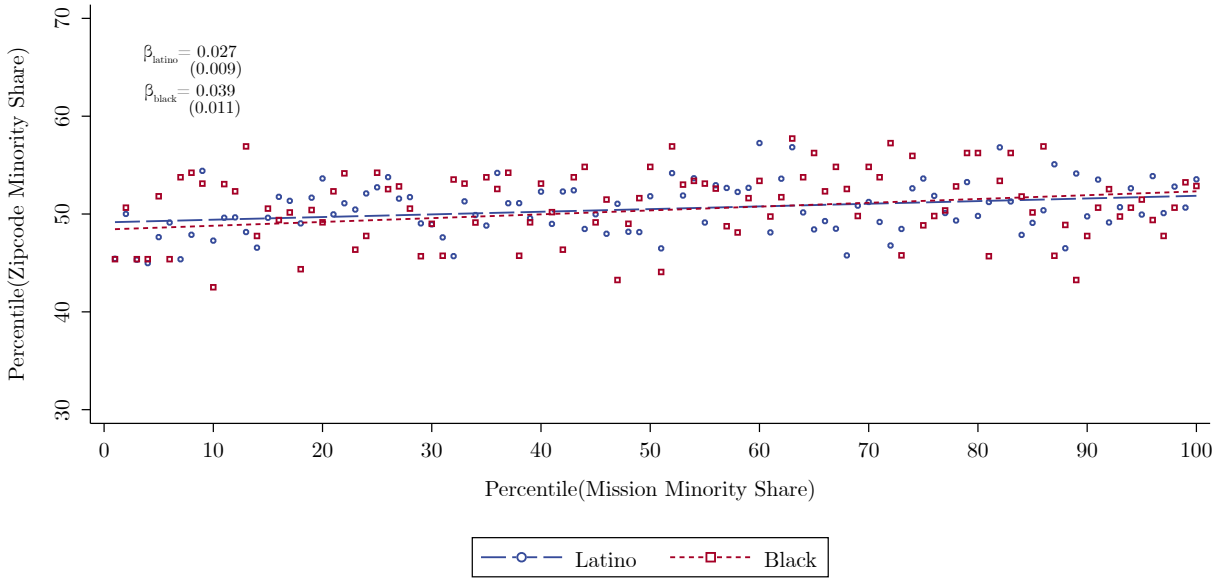
Figure 1.6 shows the results along with accompanying point estimates and block bootstrapped standard errors. Panel (a) shows that minority share affects stated attitudes about race. Volunteering in a mission with one percentile larger share of Latinos improves former volunteers' relative stated attitudes toward Latinos by 0.07 (0.01) percentiles. The effect is almost identical for serving in missions with more Black people: volunteering in a mission with one percentile larger Black population share increase relative attitudes by 0.07 (0.03). These magnitudes are hard to interpret, but the implication is that moving from the 25th to the 75th percentile of either the local Black or Latin American population share increases attitudes toward the respective group by 3.5 percentiles, and going from a mission in East Asia to one in Africa or Latin America increases attitudes by 7 percentiles.

Panel (b) reveals that volunteering in missions with larger shares of Black or Latino residents also changes life-long behaviors. Volunteering in a mission with one percentile larger share of Latinos increases the share of Latinos in the respondent's current zipcode by 0.03 (0.01) percentiles, and Volunteering in a mission with one percentile larger share of Blacks increases the share of Black in the respondent's current zipcode by 0.04 (0.01) percentiles. Note that, these effects are not limited to recently returned cohorts whose mission experiences are most salient. Instead, the effects are nearly identical for individuals who returned more recently as for those whose missionary service was 30-55 years ago.

Other Preliminary Results. We wanted to report a few interesting results regarding at-



(a) Stated Racial Attitudes (Feelings Thermometer)



(b) Current Zipcode Minority Share

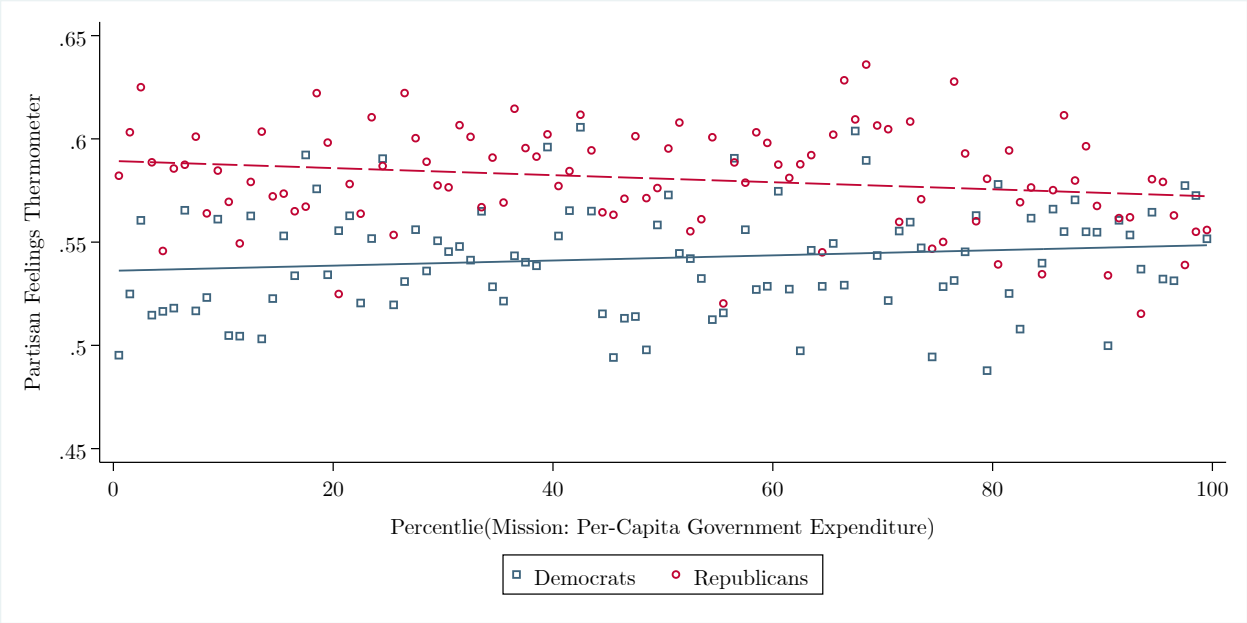
Figure 1.6: **Volunteering in Minority Rich Missions Changes Racial Attitudes**

titudes towards political partisans and working women using this alternate specification. First, when we compare feelings thermometers towards political partisans over the distribution of government size (measured in per-capita expenditure) we find that serving in places with larger government reduces affective polarization, or the gap between feelings thermometers to Republicans and Democrats. Panel (a) of Figure 1.7 presents binned scatter plots of the relationship. Because volunteer missionaries tend to have fairly conservative backgrounds, moving from the 25th to the 75th percentile of per-capital expenditure reduces affective polarization (measured as the gap between feelings towards republicans and democrats) by 28%. We have early results that this is substantiated in voting and donation behaviors as well.

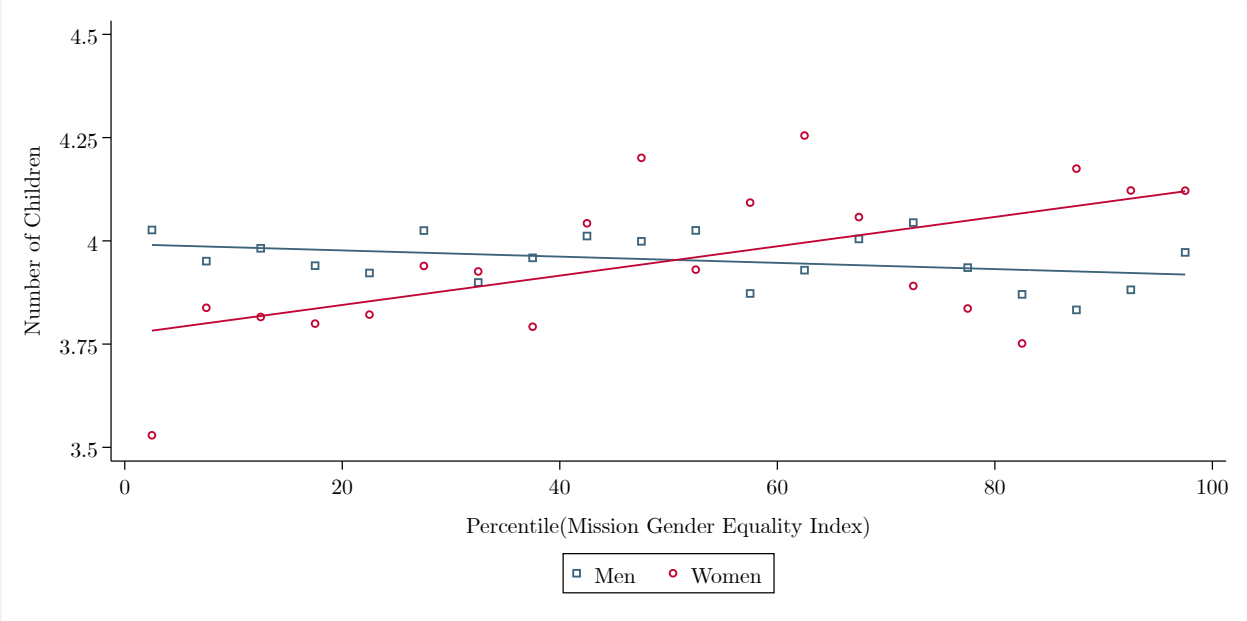
The other striking result we are seeing is that while we do find that mission assignment does affects stated attitudes about mothers who pursue education or careers, the effects are not related to the social or economic standing of women in the assigned locations. The Gender Equality Index we used was not related to the feelings thermometer answers. In stark contrast to this, we do find that despite the null results on attitudes, volunteering in these missions increases marriage rates and fertility—an effect entirely driven by women. Panel (b) of Figure 1.7 reports the results (recall that there are fewer women than men in the sample, so the ventiles are noisier for women). This result surprised us but given the cultural importance of marriage for Latter-day Saints, we wonder if seeing highly gender equitable marriages increased these female volunteers' optimism about marriage and family.

1.6 Behavioral Mechanisms and Implications

Our results provide insight into how different types of attitudes are influenced via exposure to unique characteristics of places and through different experiences. In particular, we find large impacts of exposure to places with different demographic characteristics on racial attitudes and related behaviors. In this section we seek to understand what it is about certain locations that catalyze these effects and for whom.



(a) Stated Political Attitudes (Feelings Thermometer)



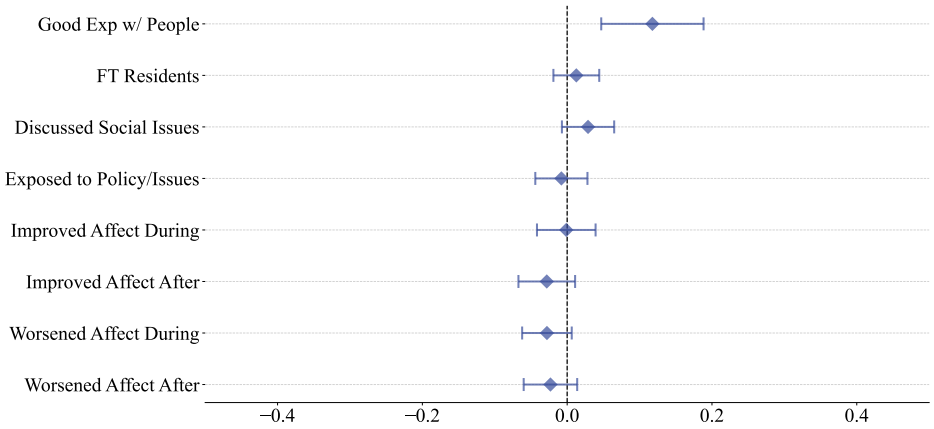
(b) Number of Children

Figure 1.7: **Mission Characteristics Affect Other Attitudes and Behaviors**

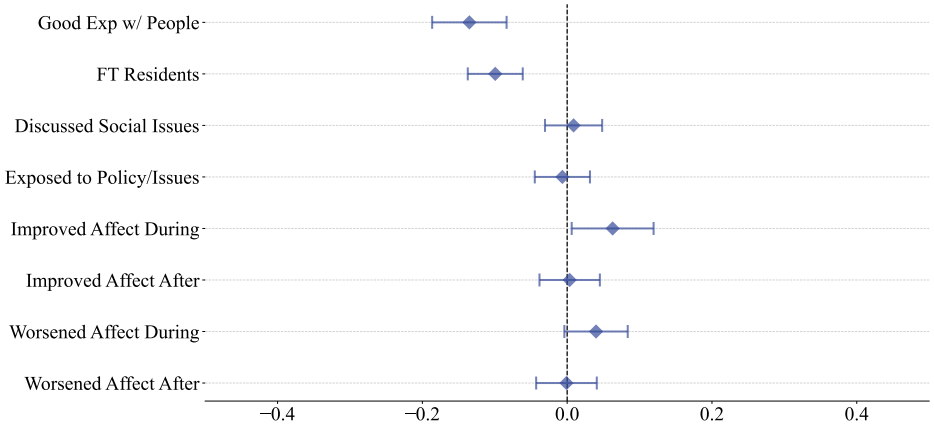
1.6.1 What Mechanisms Underpin the Observed Effects?

Understanding simply that people change as a result of exposure to the place where they live is important, but to then make normative statements about potential interventions we must understand something about the mechanisms by which these effects arise. To understand these mechanisms, we explore the impact of place through the demographics or the social attitudes of the people in the mission location as well as through the institutions in that place. We then estimate the impact of our various treatment indices on several potential mechanisms. These mechanisms include whether that was a positive experience; how they felt towards the residents in their area; discussing various social issues during their time volunteering; whether they stayed in contact with people from the location after the volunteering; how open they were to change; whether they changed during or after their mission; whether they observed policies, institutions, and issues while volunteering; and whether they are more familiar with church policy on the specific social issues.

We dig into these mechanisms for each of the main results that we find. In each of the following figures we show the reduced form impacts of assignments to the indicated types of places on each of the above items. In Figure 1.8 (a) we see that those volunteers who are assigned to places with the highest fraction of Black and Latin American residents are much more likely to report the they had a good experience with the residents in their area.



(a) SD Higher Frac. Black/Latino



(b) SD Higher Gov. Spending

Figure 1.8: **Mechanisms Suggest Positive Interactions and Political Discussions Matter**

Notes: This figure reports regressions of these standardized mechanism indices based on Equation 1. 'Good Exp w/ People' kind, receptive, time in people's homes; 'FT Residents' difference in FT at beginning and end of missions, 'Discussed Social Issues' talked about race, politics, or gender roles; 'Exposed to Policy/Issues' observed issues; 'Kept in Contact' still in contact with residents; 'Openness' openness to change; 'Change During' and 'Changed After' changed their mind on issues during or after their mission; and 'Correct Policy' know the Church policy on issues.

For our results relating to politics, we find distinct patterns in Panel (b) of Figure 1.8. Volunteers assigned to locations with higher government spending per capita have a worse experience overall with the people, but also report being more likely to have changed their mind on social issues after their service.

These patterns suggest that the impacts we observe for assignments to different places not only impact different types of people in different ways, but they also occur through different

channels. Our impacts on assignments to places with the most Black or Latino individuals seem to be driven by strong, positive experiences with people, whereas the impacts we saw on politics do not exhibit the same pattern.

1.7 Conclusion

In this paper we have explored the impact of where a person lives on their social views, including both what they say about important issues and their actions. We use the quasi-random assignment of volunteer missionaries for the Church of Jesus Christ of Latter-day Saints to fixed geographic locations to explore the impacts on those volunteer's views on race, gender roles, and politics. We find strong impacts of where a person lives on their views and actions related to race and politics, but find little impact on their views on gender roles. This is the our main contribution. We also add to our knowledge about place effects by showing some evidence that the novelty of the information matters when considering heterogeneity in the results (i.e. if a person is experiencing new things in the new location) and show that these impacts come through contact with others as well as through learning new information.

Though we find strong causal effects of where a person lives, the results about which characteristics really matter are still preliminary. We are also looking forward to exploring important questions like what mechanisms mediate the effects and which individuals are particularly susceptible to having their attitudes and behaviors influenced.

Although we are still considering policy implications, our work suggests that programs immersing individuals in communities with new demographics, attitudes, and institutions can be powerful tools in mitigating bias towards different groups. On a national level these programs could include national or military service as required in many countries. Additionally, educational institutions could use or expand existing study abroad programs to include service and community integration components. For example, scholarship programs exist for doctors who are willing to relocate to under-served areas. Since racial bias has been found to be prevalent in the medical profession (e.g. Williams & Wyatt, 2015), programs such as these could encourage doctors to

become involved in the community and specifically serve, help, and integrate with marginalized individuals in the community in addition to the important service of providing medical care to under-served areas.

Chapter 1, in part, is currently being prepared for submission for publication of the material and is coauthored with Ricks, Michael. The dissertation author was the primary researcher and author of this material.

1.8 References

Allport, G. W. (1954). *The Nature of Prejudice*. Addison-Wesley Reading, MA.

Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495.

Bagues, M., & Roth, C. (2023). Interregional Contact and the Formation of a Shared Identity. *American Economic Journal: Economic Policy*, *15*(3), 322–350.

Baker, S., Mayer, A., & Puller, S. L. (2011). Do more Diverse Environments Increase the Diversity of Subsequent Interaction? Evidence from Random Dorm Assignment. *Economics Letters*, *110*(2), 110–112.

Bazzi, S., Gaduh, A., Rothenberg, A. D., & Wong, M. (2019). Unity in Diversity? How Intergroup Contact Can Foster Nation Building. *American Economic Review*, *109*(11), 3978–4025.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive Linear Step-Up Procedures That Control the False Discovery Rate. *Biometrika*, *93*(3), 491–507.

Berinsky, A., Karpowitz, C., Peng, Z. C., Rodden, J., & Wong, C. (2022). How Social Context Affects Immigration Attitudes. *The Journal of Politics*.

Billings, S. B., Chyn, E., & Haggag, K. (2021). The Long-Run Effects of School Racial Diversity on Political Identity. *American Economic Review: Insights*, *3*(3), 267–84.

Boisjoly, J., Duncan, G. J., Kremer, M., Levy, D. M., & Eccles, J. (2006). Empathy or Antipathy? The Impact of Diversity. *American Economic Review*, *96*(5), 1890–1905.

Bronnenberg, B. J., Dubé, J.-P. H., & Gentzkow, M. (2012). The Evolution of Brand Preferences: Evidence from Consumer Migration. *American Economic Review*, *102*(6), 2472–2508.

Card, D., & Cardoso, A. R. (2012). Can Compulsory Military Service Raise Civilian Wages? Evidence from the Peacetime Draft in Portugal. *American Economic Journal: Applied Economics*, *4*(4), 57–93.

Card, D., Rothstein, J., & Yi, M. (2023). *Location, Location, Location* (Tech. Rep.). National Bureau of Economic Research.

Carrell, S. E., Hoekstra, M., & West, J. E. (2019). The Impact of College Diversity on Behavior Toward Minorities. *American Economic Journal: Economic Policy*, *11*(4), 159–82.

Chetty, R., & Hendren, N. (2018a). The Impacts of Neighborhoods on Intergenerational Mobility

I: Childhood Exposure Effects. *The Quarterly Journal of Economics*, *133*(3), 1107–1162.

Chetty, R., & Hendren, N. (2018b). The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates. *The Quarterly Journal of Economics*, *133*(3), 1163–1228.

Crawfurd, L. (2021). Contact and Commitment to Development: Evidence from Quasi-Random Missionary Assignments. *Kyklos*, *74*(1), 3–18.

Dahl, G. B., Kotsadam, A., & Rooth, D.-O. (2021). Does Integration Change Gender Attitudes? The Effect of Randomly Assigning Women to Traditionally Male Teams. *The Quarterly Journal of Economics*, *136*(2), 987–1030.

Di Pietro, G. (2012). Does Studying Abroad Cause International Labor Mobility? Evidence from Italy. *Economics Letters*, *117*(3), 632–635.

Ertola Navajas, G., López Villalba, P. A., Rossi, M. A., & Vazquez, A. (2022, 01). The Long-Term Effect of Military Conscription on Personality and Beliefs. *The Review of Economics and Statistics*, *104*(1), 133-141. Retrieved from https://doi.org/10.1162/rest_a_00930 doi: 10.1162/rest_a_00930

Exley, C. L. (2020). Using Charity Performance Metrics as an Excuse not to Give. *Management Science*, *66*(2), 553–563.

Finkelstein, A., Gentzkow, M., & Williams, H. (2016). Sources of Geographic Variation in Health Care: Evidence from Patient Migration. *The Quarterly Journal of Economics*, *131*(4), 1681–1726.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring Individual Differences in Implicit Cognition: the Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464.

Kaplan, E., Spenkuch, J. L., & Tuttle, C. (2019). *School Desegregation and Political Preferences: Long-run Evidence from Kentucky* (Tech. Rep.). Working Paper.

Kinder, D. R., Sanders, L. M., & Sanders, L. M. (1996). *Divided by Color: Racial Politics and Democratic Ideals*. University of Chicago Press.

Lowe, M. (2020). *Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration* (Tech. Rep.). Munich: CESifo Working Paper.

Malamud, O., & Wozniak, A. (2012). The Impact of College on Migration: Evidence from the Vietnam Generation. *Journal of Human Resources*, *47*(4), 913–950.

Marmaros, D., & Sacerdote, B. (2006). How do Friendships Form? *The Quarterly Journal of Economics*, *121*(1), 79–119.

Mo, C. H., & Conn, K. M. (2018). When do the Advantaged see the Disadvantages of Others? A

Quasi-Experimental Study of National Service. *American Political Science Review*, *112*(4), 721–741.

Mousa, S. (2020). Building Social Cohesion between Christians and Muslims through Soccer in post-ISIS Iraq. *Science*, *369*(6505), 866–870.

Okunogbe, O. M. (forthcoming). Does Exposure to Other Ethnic Regions Promote National Integration? Evidence from Nigeria. *American Economic Journal: Applied Economics*.

Oosterbeek, H., & Webbink, D. (2011). Does Studying Abroad Induce a Brain Drain? *Economica*, *78*(310), 347–366.

Parey, M., & Waldinger, F. (2011). Studying Abroad and the Effect on International Labor Market Mobility: Evidence from the Introduction of ERASMUS. *The Economic Journal*, *121*(551), 194–222.

Pope, D. G. (2008). Benefits of Bilingualism: Evidence from Mormon Missionaries. *Economics of Education Review*, *27*(2), 234–242.

Rao, G. (2019). Familiarity does not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools. *American Economic Review*, *109*(3), 774–809.

Schindler, D., & Westcott, M. (2021). Shocking Racial Attitudes: Black GIs in Europe. *Review of Economic Studies*.

Van Laar, C., Levin, S., Sinclair, S., & Sidanius, J. (2005). The Effect of University Roommate Contact on Ethnic Attitudes and Behavior. *Journal of Experimental Social Psychology*, *41*(4), 329–345.

Williams, D. R., & Wyatt, R. (2015). Racial Bias in Health Care and Health: Challenges and Opportunities. *Journal of the American Medical Association*, *314*(6), 555–556.

Chapter 1 Appendix

1.A.1 Appendix Tables

Table 1.6: **Sample Baseline Characteristics**

	1970-1979	1980-1989	1990-1999	2000-2010
Female	0.11	0.2	0.27	0.23
White	0.94	0.93	0.93	0.92
Black	0.00	0.00	0.00	0.00
Asian	0.01	0.01	0.02	0.03
Hispanic	0.02	0.02	0.03	0.05
Reported Medical Conditions	0.03	0.03	0.05	0.09
Spoke Language	0.18	0.21	0.26	0.28
Graduated Seminary	0.73	0.82	0.89	0.92
HS GPA	3.39	3.50	3.63	3.71
	(0.50)	(0.46)	(0.42)	(0.39)
Leadership Opportunities in HS	0.48	0.53	0.59	0.62
Leadership Opportunities in Church	0.77	0.84	0.86	0.86
Some College Before	0.40	0.32	0.29	0.24
Parents Republican	0.72	0.79	0.83	0.84
Parents Democrat	0.10	0.08	0.06	0.04
Observations	2662	3202	5077	4706

Notes: Each column gives the means for volunteers who started their volunteering in the given year range. Standard deviations are in parentheses.

Table 1.7: **Sample Current Characteristics**

	1970-1979	1980-1989	1990-1999	2000-2010
Married	0.89	0.89	0.87	0.82
Never Married	0.02	0.03	0.04	0.08
Number Children	4.52	4.18	4.04	3.33
	(1.86)	(1.68)	(1.64)	(1.57)
Active in Church	0.91	0.88	0.84	0.76
Bachelor's Degree	0.27	0.3	0.28	0.29
Graduate Degree	0.61	0.6	0.6	0.56
Earn less 50k	0.05	0.02	0.01	0.03
Earn 50-75k	0.08	0.04	0.03	0.05
Earn 75-100k	0.11	0.06	0.06	0.09
Earn 100-150k	0.21	0.18	0.18	0.22
Earn 150-200k	0.13	0.16	0.16	0.16
Earn 200-250k	0.30	0.28	0.27	0.29
Earn more than 250k	0.17	0.28	0.30	0.19
Republican	0.58	0.51	0.39	0.29
Democrat	0.08	0.11	0.14	0.22
Observations	2662	3202	5077	4706

Notes: Each column gives the means for volunteers who started their volunteering in the given year range. Standard deviations are in parentheses.

Table 1.8: **Aggregate Results for Stated Racial Attitudes and Related Behaviors**

	Stated Attitudes	Behaviors
Resident Frac Black/Latino	0.043***	-0.032**
	(0.016)	(0.014)
Resident IAT Black	-0.023	0.022
	(0.012)	(0.012)
Civil Liberties Index	-0.004	0.021
	(0.014)	(0.014)
Observations	11325	12449

Notes: * p<0.1, ** p<0.05, *** p<0.01. Standard errors are in parenthesis. Estimates in each column reflect estimates from equation (1), which includes controls for reported medical conditions, race, graduation from seminary, participation in extracurriculars in high school, leadership opportunities in high school and church, whether they spoke a language pre-mission, whether they attended some college pre-mission, their participation level in the Church pre-mission, their sex, and fixed effects for decade of service. 'Stated Attitudes' is a standardized index of the respondent's reported feelings thermometer values for Black people, White people, and Latino people. 'Behaviors' is the number of the following actions they reported: read a book on race, listen to a podcast on race, volunteer for social justice, donate to social justice, vote for a minority candidate, protest police violence, fraction Black/Hispanic in current zip code. 'Donations' is a measure of willingness to pay towards the NAACP. 'Resident Frac Black/Latino' is a standardized index the fraction of Black and Latin American residents, 'Resident IAT Black' is a standardized index for the average Implicit Association Test value for feelings towards Black individuals demonstrated by residents in the area, and 'Civil Liberties Index' is the Freedom House Civil Liberties index for the assigned area, standardized.

Table 1.9: **Aggregate Results for Stated Political Attitudes and Related Behaviors**

	Stated Attitudes	Behaviors
Rural Index	0.004	0.000
	(0.013)	(0.013)
Resident IAT Democrats	0.000	-0.008
	(0.014)	(0.015)
Gov. Spending	0.044**	0.004
	(0.017)	(0.020)
Observations	11292	12645

Notes: * p<0.1, ** p<0.05, *** p<0.01. Standard errors are in parenthesis. Estimates in each column reflect estimates from equation (1), which includes controls for reported medical conditions, race, graduation from seminary, participation in extracurriculars in high school, leadership opportunities in high school and church, whether they spoke a language pre-mission, whether they attended some college pre-mission, their participation level in the Church pre-mission, their sex, and fixed effects for decade of service. 'Stated Attitudes' is a standardized index of the respondent's reported feelings thermometer values for Republicans and Democrats (reverse coded). 'Behaviors' is the sum of the number of the following actions they reported: read a book on politics, listen to a podcast on politics, donate to political causes, volunteer for political causes, protest mask mandates, self report that they are a conservative. 'Donations' is a measure of willingness to pay towards the Republican party and the Democratic party (reverse coded). 'Urban Index' is the standardized index measuring fraction urban, average age, and fraction with college education, 'Resident IAT Democrat' is the Implicit Association Test measure for residents towards Democrat presidents, and 'Gov spending' is a standardized measure of Government spending per capita.

Table 1.10: **Results for Gender Attitudes, Behaviors, and Donations to the National Partnership for Women and Families**

	Stated Attitudes	Behaviors
Family Index	-0.014	0.026
	(0.022)	(0.019)
Resident IAT Working Women	-0.006	-0.017
	(0.013)	(0.014)
Gender Ineq Index	0.009	-0.018
	(0.017)	(0.014)
Observations	11797	12456

Notes: * p<0.1, ** p<0.05, *** p<0.01. Standard errors are in parenthesis. Estimates in each column reflect estimates from equation (1), which includes controls for reported medical conditions, race, graduation from seminary, participation in extracurriculars in high school, leadership opportunities in high school and church, whether they spoke a language pre-mission, whether they attended some college pre-mission, their participation level in the Church pre-mission, their sex, and fixed effects for decade of service. 'Stated Attitudes' is a standardized index of the respondent's reported feelings thermometer values for stay-at-home moms, mothers who work because they choose to, and feminists. 'Behaviors' is the sum of the number of the following actions they reported: read a book on gender, listen to a podcast on gender, husband in household responsible for childcare, husband in household responsible for cooking/cleaning, wife in household works full-time. 'Donations' is a measure of willingness to pay towards the National Partnership for Women and Families. 'Family Index' is a standardized index combining age at first marriage and average family size, 'Resident IAT Working Women' is a standardized measure of the Implicit Association Test for residents towards working women, and 'Gender Ineq Index' is the standardized gender inequality index from the Human Development Reports.

1.A.2 Pilot Surveys

Before running this project at-scale, we ran several waves of a pilot survey to show the viability of the project; explore important outcomes, heterogeneity, and mechanisms; understand possible sample frames for the project; refine and perfect the survey instrument; and work out details for survey administration. Across all waves of the pilot we collected information on mission service, including crucially when and where the person served, and basic demographics.

We ran the first wave of these pilots in August 2021. In this wave we started with a focus on racial attitudes for individuals assigned to volunteer within the United States. We focused on this group primarily to allow for cleaner comparisons across mission location assignments. This was intended as a simplification to show viability in a first pass and to explore important heterogeneity and mechanisms. Before the start of wave 1, we ran a set of field interviews with former full-time volunteers to design the survey to identify the most plausible dimensions of heterogeneity and mechanisms. Through these interviews we zeroed in on parent political leaning, parent education, and the diversity of the childhood hometown as key dimensions of heterogeneity. Furthermore, we identified the following as important possible mechanisms: information about the size of racial disparities (e.g. 'I didn't realize how large income gaps were between Black and White Americans, but now I know differently'), contact with different types of people (e.g. 'I feel more warmly towards people because I have now met them'), contact with people who have different beliefs and attitudes about the world, opening to future learning about race/racism after the duration of the volunteering, and confirmation of prior stereotypes.

With these in hand, we designed a first survey instrument to elicit attitudes about race and racism. The survey was 20-25 minutes long. The sample frame for this survey was gathered using a website called 'Lifey.org' and can be thought of as a convenience sample. When a volunteer participates in this mission service it is common practice for them to keep a record of their experience for friends and family, and many volunteers do this via an online blog. This website has lists of volunteers who have published blogs for every possible mission assignment location since roughly

2010. We took these lists of volunteers, starting with the most and least racially diverse assignment locations within the US and focusing on college age people, then found these volunteers on social media (primarily Facebook and LinkedIn). We subsequently messaged these individuals to elicit their participation in our survey. In the end we collected 497 responses. Unfortunately we were unable to track response rate (because it was unclear who actually received our messages and whether we correctly identified the person we were looking for who participated in this volunteer service) and we had very high attrition (about 30 percent of participants).

Table 1.11 shows the characteristics of the sample for pilot wave 1 on average and shows that those characteristics are very balanced across different types of mission assignments. Importantly, this balance holds not only for characteristics available to the mission leaders making location assignments (i.e. demographics, pre-mission experiences, and mission application information), but this also holds for characteristics unobserved by those leaders. These characteristics can be found in the last panel of Table 1.11. This is strong evidence that the assignment of volunteers to location is actually independent of the outcomes we care about.

To overcome some of the major pitfalls in wave 1, continue refining the survey instrument, understand response rates, and explore a different sample frame, we ran a second pilot wave. The revised survey was about 10 minutes long and included questions on attitudes towards race/racism, education, immigration, and government spending. We did not collect all of the same detailed information for balance in this sample since the main goals of this pilot were different than the first wave. The sample frame for the first half (waves 2a and 2b) of this wave was a list of Brigham Young University Idaho (BYU-I) Alumni who posted their information publicly to allow networking for students. BYU-I (and Brigham Young University (BYU) in Provo, Utah) are schools owned and operated by the Church of Jesus Christ of Latter-day Saints. As such, a large fraction of students and alumni are members of the Church, and about two thirds did this volunteer mission service. For the second half of this wave (waves 2c and 2d), the sample frame was the same used for the at-scale survey, namely BYU alumni. We collected email addresses for these individuals to administer the survey.

The BYU-I alumni in this sample ranged in age from young professionals to retirees, but all graduated from BYU-I. For wave 2a we still limited participation to volunteers assigned to locations within the US, which ultimately garnered us 145 responses. The response rate was very high (78 percent) and attrition was much lower, around 8 percent. We paid these participants $12 for participation in the 10 minute survey.

For wave 2b, we took the other half of the BYU-I sample (those who were assigned outside of the United States), and sent them the survey comparing response rates if we asked them to participate out of goodwill rather than paying them. This has several advantages over paying participants, including the ability to survey a larger sample of individuals. In this sample we received 114 responses, which reflects a 48 percent response rate[18]. Attrition was even lower in this sample, with about 4 percent of survey takers attriting.

Since the BYU-I sample is a group of people who are very likely to be more responsive to survey requests than the typical person, we turned to the sample of BYU alumni that we collected for the at-scale survey and sent two more waves testing compensation schemes; one offering a lottery incentive (wave 2c) and one asking individuals to participate out of goodwill (wave 2d). Response rates in each were not substantively different, i.e. 29 percent in wave 2c and 26 percent in wave 2d. This assumes that all of our emails were properly received, since we did not track who ultimately read the email as in pilot wave 3.

We also ran several smaller surveys on the online survey marketplace Prolific to test the viability of specific questions. In particular, we went through several iterations of stated attitudes where we compared trust questions across various groups of people with feelings thermometers. We also went through many iterations of the donation activity in Exley (2020). We found that our setting was quite different than that in Exley (2020) since we had a much shorter time to administer the activity, so ultimately settled on a simplified version that participants in our Prolific sample seemed to understand well.

[18]There was some ambiguity about how many non-respondents would have been eligible, but we bound response rates between 68-98% for wave 2a and 42-61% for wave 2b based on best or worst case scenarios of eligibility.

1.A.2.1 Pilot Results

Wave 1 Results

Wave 1 of our pilot had three primary goals: (1) establish the viability of our empirical strategy and explore suggestive results, (2) understand possible dimensions for heterogeneity and mechanisms, and (3) start to refine our survey instrument. Initially in this pilot we focused on racial attitudes and limited our analysis to former volunteers who were assigned within the United States. In particular, the primary outcome in this pilot is the commonly used Racial Resentment Index introduced by Kinder, Sanders, and Sanders (1996). This is an index of the respondent giving the following answers to four Likert questions:

- (Agree, Strongly Agree) Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. People of color should do the same without any special favors.
- (Disagree, Strongly Disagree) Over the past few years, people of color have gotten less than they deserve.
- (Disagree, Strongly Disagree) Generations of slavery and discrimination have created conditions that make it difficult for people of color to work their way out of the lower class.
- (Agree, Strongly Agree) It's really a matter of some people just not trying hard enough: if people of color would only try harder they could be just as well off as whites.

High values on this index indicate high levels of stated racial resentment: in particular, a value of four would be the highest level of stated resentment and a value of zero would be the lowest. Our primary results from using this outcome can be seen in Tables 1.12 and 1.13. Our dependent variable in each of these regressions is an indicator for whether the geographic area covered by the mission location has a higher fraction of Black or Hispanic individuals than the national average (13.4 and 18.5 percent respectively). In column 1 in each table we see that

relative to an individual assigned to a mission more white than the national average, those assigned to more racially diverse mission locations have a reduction in stated prejudice of -0.187. The mean for volunteers in our sample assigned to more white missions is 0.92, so this is a large reduction (about 20 percent). For comparison, individuals in our sample exhibit slightly less stated racial resentment than the national average measured in the ANES, both overall and for individuals under 30. Our sample is comparable to other Latter-day Saints surveyed for the ANES. Though this measure is somewhat noisy, it is a meaningful impact on stated prejudice and strongly suggestive of a treatment effect.

Perhaps more interestingly, though, this impact exhibits strong heterogeneity along important dimensions. We find no meaningful heterogeneity for men versus women, but the impact is 50 percent larger in magnitude for those individuals who reported that their parents were strongly conservative or for those who grew up in more white zip codes. We also see a very large effect for those with less educated parents, though the sample size is quite small, so should be interpreted cautiously. One important note about these results is that they are virtually identical when including or excluding the controls for the mission application items, providing strong evidence, in addition to the balance, that our identification is working properly.

In addition to possible sample size concerns, there is one very important caveat to this primary analysis. We had quite high attrition for this survey wave (about 30 percent), but for just over half of those who did not complete the survey we still received information on the Racial Resentment Index. In the above reported results, we limited to those who completed the survey. If we run the same analysis on the sample who never finished the survey, the coefficient on being assigned to a more racially diverse mission is large and positive. It is difficult to interpret this number because the mean for the control observations in this group is very low and there is a small sample size, but this suggests that reducing attrition is a key concern for our analysis.

Keeping this caveat in mind, we also explored the impact of being assigned to a more racially diverse mission on a broader set of outcomes, displayed in Table 1.14.

The additional outcomes in Table 1.14 include the Explicit Racial Resentment Index[19], the Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998), whether the participant has read a book on race/racism, whether they have voted for a minority candidate, and whether they voted for Biden in 2020.

The results on the Explicit Racial Resentment scale are even stronger than those on the standard scale presented above. Additionally, those assigned to more racially diverse missions exhibit a lower level of implicit bias towards Black people (a positive coefficient on the Implicit Association Test means a lower automatic preference for White people over Black people), are about 12.1 percentage points more likely to read a book on race, are 13.7 percentage points more likely to vote for a minority candidate at any point in time, and are 7.6 percentage points more likely to have voted for Biden in 2020 (though this measure is not statistically different than 0). These are large and meaningful impacts, not only on stated measures of racial attitudes measured in our survey, but also on a few behavioral outcomes.

We also explored how treatment moved each of the proposed mechanisms during this wave. These include 'Belief' - information about the size of racial disparities (e.g. 'I didn't realize how large income gaps were between Black and White Americans, but now I know differently'), 'Contact' - contact with different types of people (e.g. 'I feel more warmly towards people because I have now met them'), 'Softening' - opening to future learning about race/racism after the duration of the volunteering, and 'Confirmation' - confirmation of prior stereotypes.

Table 1.15 displays these results. The contact mechanisms, as measured by the fraction of people that the volunteer visited in their homes on a daily basis who were Black or Hispanic, is the only mechanism strongly moved by treatment. Beliefs (measured by their stated beliefs about the magnitude of racial disparities), Softening (measured by future engagement in learning about race/racism), and Confirmation (measured by agreement with a question asking how much they agreed that they realized on their mission that there is a reason for racial stereotypes) are all much

[19] An index of the following: (Agree, Strongly Agree) "I resent all of the special attention and favors that people of color receive. Other Americans have problems too.", (Agree, Strongly Agree) "I am concerned that the special privileges for people of color place me at an unfair disadvantage, even when I have done nothing to harm them", (Agree, Strongly Agree) "For people of color to succeed, they need to stop using race as an excuse".

smaller and statistically indistinguishable from zero.

Taken together our results from this wave suggest strongly that assignment to different types of locations moves beliefs in a meaningful way. Additionally, it provides important possible dimensions of heterogeneity (parent political leaning, parent education, and diversity of childhood zip code) as well as possible mechanisms (in particular, contact with a variety of different types of people). These results should be interpreted cautiously, given the nature of the convenience sample and the high attrition to the survey.

Wave 2 Results

We first present the analogous results to wave 1 for race and racism in this pilot in Table 1.16. These estimates again limit to volunteers assigned in the US, but divide the sample by individuals who finished their volunteering after 2006 (younger cohorts) and those who finished in 2006 or earlier (older cohorts). Though the estimates are all quite noisy, these results suggest interesting cross-cohort heterogeneity. In particular, the impact of being assigned to a more racially diverse location flips signs across the cohorts. We also see signs in the opposite direction for reading books on race and voting for minority candidates than we did in wave 1, but these are all statistically indistinguishable from zero.

We examine not only racial attitudes, but also extend our analysis to views on immigration, shown in Table 1.17, and views on politics, 1.18. For attitudes relating to immigration we limit to volunteers assigned outside of the United States and compare those assigned to developing countries as opposed to developed nations. Broadly speaking, effects are large but quite noisy, suggesting that further analysis is needed to draw any definitive conclusions. Our outcomes in table 1.17 are whether the volunteer says their views on immigration were strongly changed during their volunteering, an index of positive views towards immigrants[20], and whether the individual has

[20]This is an index of the following: (Disagree, Strongly Disagree) "Immigrants and refugees today are a burden on our country because they take our jobs and social benefits", (Agree, Strongly Agree) "The United States should accept more refugees and immigrants than in recent years", (Disagree, Strongly Disagree) "I would prefer to have fewer immigrants and refugees in my community", (Agree, Strongly Agree) "It is unfair to blame immigrants and refugees for crime more than other groups", and (Disagree, Strongly Disagree) "In general I would be happier to see a relative marry a US native than an immigrant or refugee".

volunteered to help refugees. Again, there may be a hint of cross-cohort heterogeneity in these attitudes, but more analysis is warranted.

The last set of attitudes, those concerning politics, compare volunteers assigned to the US across areas that voted Democrat on average in 2016, as opposed to those who voted Republican in 2016. Though younger cohorts are more strongly impacted, both younger and older cohorts are pushed in the same direction; exposure to people with more liberal political leanings moves the volunteers to vote and affiliate more liberally.

Overall the strength in this wave of the pilot was refining and pinpointing many of the survey administration pieces, but also providing suggestive evidence that many social views are moved by living in different types of places, and that they are moved in different ways.

Table 1.11: **Balance in Characteristics Across Different US Missions**

	Sample Average	More White	More Minority	Difference: Minority-White
Demographics:				
Female	0.400	0.412	0.391	-0.021 [p=0.672]
Non-White	0.121	0.124	0.119	-0.004 [p=0.898]
Pre-mission Experiences:				
Any College	0.595	0.563	0.618	0.055 [p=0.222]
Language Exposure	0.891	0.903	0.882	-0.021 [p=0.456]
Foreign Travel	0.586	0.579	0.592	0.013 [p=0.795]
Weekly Church Participation	0.955	0.953	0.956	0.003 [p=0.886]
Mission Application:				
Willing to go Foreign	0.960	0.965	0.956	-0.009 [p=0.669]
Willing to Learn Language	0.952	0.942	0.961	0.020 [p=0.383]
Family Mission in Region	0.196	0.181	0.209	0.027 [p=0.503]
Medical Issue Flagged	0.271	0.257	0.282	0.024 [p=0.598]
Parent Characteristics:				
Less than Bachelors	0.230	0.218	0.239	0.021 [p=0.616]
Graduate School	0.453	0.512	0.412	-0.100 [p=0.045]
Republicans	0.844	0.847	0.842	-0.005 [p=0.896]
Strong Republicans	0.448	0.488	0.419	-0.069 [p=0.167]
Pro Redistribution	0.114	0.094	0.131	0.038 [p=0.249]
Zipcode % Black or Hispanic	0.172	0.161	0.180	0.019 [p=0.241]

Notes: 'More White' refers to volunteers who were assigned to mission locations that have less Black/Hispanic individuals than the national average in the US, whereas 'More Minority' is the converse.

Table 1.12: **Pilot Wave 1 Results using the Racial Resentment Index**

	Full	Sex		Parent Political		Parent Education	
	Sample	Female	Male	Strong Rep	Other	Less than BA	BA +
Minority Rich Mission	-0.187	-0.167	-0.180	-0.279	-0.087	-0.975	-0.001
	(0.132)	(0.168)	(0.203)	(0.219)	(0.159)	(0.340)	(0.143)
Control Means:							
White Mission (Pilot)	0.92	0.75	1.09	1.14	0.72	1.43	0.81
National (ANES)	1.55	1.61	1.50				
Under 30 (ANES)	1.27	1.20	1.35				
Latter-day Saint (ANES)	0.91	0.79	1.04				
Observations	299	155	144	129	168	60	239

Notes: Standard errors in parentheses. 'Minority Rich Mission' indicates assignment to a mission with more Black/Hispanic individuals than the national average in the US. The outcome is the Racial Resentment Index from Kinder et al. (1996).

Table 1.13: **Pilot Wave 1 Results using the Racial Resentment Index (cont.)**

	Full	Childhood Zipcode		Interracial Friendship		One
	Sample	Diverse	White	Yes	No	Caveat
Minority Rich Mission	-0.187	-0.039	-0.309	-0.329	-0.145	0.627
	(0.132)	(0.188)	(0.193)	(0.161)	(0.269)	(0.291)
Control Means:						
White Mission (Pilot)	0.92	0.94	0.86	0.96	0.93	0.64
National (ANES)	1.55	1.78	1.43			
Under 30 (ANES)	1.27	1.28	1.25			
Latter-day Saint (ANES)	0.92	0.79	0.90			
Observations	299	146	136	192	79	65

Notes: Standard errors in parentheses. 'Minority Rich Mission' indicates assignment to a mission with more Black/Hispanic individuals than the national average in the US. The outcome is the Racial Resentment Index from Kinder et al. (1996).

Table 1.14: **Additional Outcomes from Pilot Wave 1**

	Explicit Resentment	Implicit Association	Book on Race	Vote for Minority	Vote for Biden
Minority Rich Mission	-0.203	0.309	0.121	0.137	0.076
	(0.116)	(0.176)	(0.057)	(0.056)	(0.058)
Control Means:					
White Mission (Pilot)	0.94	-1.39	0.35	0.57	0.33
Observations	299	266	299	296	282

Notes: Standard errors in parentheses. 'Minority Rich Mission' indicates assignment to a mission with more Black/Hispanic individuals than the national average in the US.

Table 1.15: **Impact of Assignment to a Racially Diverse Location on Possible Mechanisms**

	Belief	Contact	Softening	Confirmation
Minority Rich Mission	0.067	0.514	0.160	-0.064
	(0.065)	(0.114)	(0.116)	(0.117)

Notes: Standard errors in parentheses. 'Minority Rich Mission' indicates assignment to a mission with more Black/Hispanic individuals than the national average in the US.

Table 1.16: **Pilot 2 Results on Racial Attitudes**

	Younger Cohorts			Older Cohorts		
	Racial Resentment	Book on Race	Vote for Minority	Racial Resentment	Book on Race	Vote for Minority
Minority Rich Mission	-0.242	-0.056	-0.071	0.664	-0.056	-0.112
	(0.316)	(0.139)	(0.134)	(0.378)	(0.145)	(0.121)
Control Means:						
White Mission (Pilot)	0.94	0.54	0.63	1.09	0.65	0.80
Observations	63	63	63	54	54	54

Notes: Standard errors in parentheses. 'Minority Rich Mission' indicates assignment to a mission with more Black/Hispanic individuals than the national average in the US. The outcome is the Racial Resentment Index from Kinder et al. (1996).

Table 1.17: **Pilot 2 Results on Immigration Attitudes**

	Younger Cohorts			Older Cohorts		
	Attitudes Changed	Positive Immigrant	Volunteer for Refugees	Attitudes Changed	Positive Immigrant	Volunteer for Refugees
Developing Country	0.132	0.158	0.073	0.056	0.192	-0.172
	(0.127)	(0.414)	(0.125)	(0.133)	(0.404)	(0.134)
Control Means:						
White Mission (Pilot)	0.36	2.05	0.32	0.39	2.59	0.55
Observations	59	59	59	56	56	56

Notes: Standard errors in parentheses. 'Developing Country' indicates assignment to a mission with in a developing country. The outcome is an index of attitudes towards immigration mirroring the Racial Resentment Index from Kinder et al. (1996).

Table 1.18: **Pilot 2 Results on Political Attitudes**

	Younger Cohorts		Older Cohorts	
	Report Republican	Vote for Biden	Report Republican	Vote for Biden
Area Voted Democrat	-0.035	0.178	-0.011	0.111
	(0.128)	(0.117)	(0.129)	(0.128)
Control Means:				
White Mission (Pilot)	0.59	0.32	0.70	0.31
Observations	61	63	53	54

Notes: Standard errors in parentheses. 'Area Voted Democrat' indicates assignment to a mission in the US that voted Democrat on average in the 2016 presidential election.

Chapter 2

Effect Heterogeneity and Optimal Policy: Getting Welfare Added from Teacher Value Added

Tanner Eastmond, Michael Ricks, Nathan Mather, and Julian Betts[0]

Abstract

Though ubiquitous in research and practice, mean-based "value-added" measures may not fully inform policy or welfare considerations when policies have heterogeneous effects, impact multiple outcomes, or seek to advance distributional objectives. In this paper we formalize the importance of heterogeneity for calculating social welfare and quantify it in an enormous public service provision problem: the allocation of teachers to elementary school classes. Using data from the San Diego Unified School District we estimate heterogeneity in teacher value-added over the lagged student test score distribution. Because a majority of teachers have significant comparative advantage across student types, allocations that use a heterogeneous estimate of value-added can raise scores by 34-97% relative to those using only standard value-added estimates. These gains are even larger if the social planner has heterogeneous preferences over groups. Because reallocations benefit students on average at the expense of teachers' revealed preferences, we also consider

[0]Department of Economics, University of California San Diego, the Department of Economics, University of Nebraska-Lincoln, and Secretariat. Authors can be reached at teastmond@ucsd.edu, mricks4@unl.edu, nmather@secretariat-intl.com, and jbetts@ucsd.edu.

a simple teacher compensation policy, finding that the marginal value of public funds would be infinite for bonuses of up to 14% of baseline pay. These results, while specific to the teacher assignment problem, suggest more broadly that using information about effect heterogeneity might improve a broad range of public programs—both on grounds of average impacts and distributional goals.

2.1. Introduction

When evaluating policies, programs, and institutions researchers often rely on mean impacts. While means are powerful summary measures, they can also mask economically important information. This paper seeks to understand how measuring heterogeneity can more fully inform welfare measures and better optimize policy choices. We ask two main questions. (1) Theoretically, when does heterogeneity (in effects, outcomes, and social preferences) matter for maximizing a social objective? (2) Empirically, how large are the welfare gains from using heterogeneous rather than average estimates of impacts to evaluate and refine public policy?

Although these questions have many applications, we explore them in the context of value-added scores for elementary school teachers. Many have used value-added scores (regression adjusted means) to measure the effects of teachers and schools (see reviews in Angrist, Hull, & Walters, 2022; Bacher-Hicks & Koedel, 2022); doctors, hospitals, and nursing homes (Chan, Gentzkow, & Yu, 2022; Chandra, Finkelstein, Sacarny, & Syverson, 2016; Doyle, Graves, & Gruber, 2019; Einav, Finkelstein, & Mahoney, 2022; Hull, 2020); and even judges, prosecutors, and defense attorneys (Abrams & Yoon, 2007; Harrington & Shaffer, 2023; Norris, 2019). We choose the elementary school setting because of mounting empirical evidence that value-added scores are both *multidimensional* and *heterogeneous* in the education context. For example, teachers affect student outcomes in multiple dimensions such as math and reading scores (Condie, Lefgren, & Sims, 2014), attendance and suspensions (Jackson, 2018), and work ethic and learning skills (Petek & Pope, in press). Furthermore, teachers also have heterogeneous effects on different types of students defined by factors such as race and gender (e.g., Dee, 2005; Delgado, 2022; Delhommer, 2019) and socioeconomic status (Bates, Dinerstein, Johnston, & Sorkin, 2022). Similar patterns have been found in health-related value-added (e.g. Hull, 2020).

This paper applies and extends insights from theoretical welfare economics to overcome the limitations that arise from multidimensionality and heterogeneity, allowing us to empirically evaluate the optimal allocation of teachers to classes based on this information. The critical issue

from a social welfare perspective is that in the presence of multidimensionality and heterogeneity, value-added measures only partially order the welfare of an allocation of teachers to students. Intuitively, this is because of ambiguity about whether the definition of a "better" teacher should prioritize gains in math versus reading scores or gains for high-achieving versus low-achieving students (See the impossibility-like results in Condie et al., 2014). Fortunately, whereas research in value-added has identified these problems, research in public finance has a long history of using welfare functions to aggregate over the heterogeneous effects of policies. We extend such insights from welfare economics for two purposes. First, we characterize the shortcomings of relying on mean-oriented measures of policy effects such as standard value-added to make welfare considerations in general. Then the bulk of the paper evaluates the optimal allocation of teachers to classes using measures of heterogeneous value-added that produce scalar, welfare-relevant statistics.

Our theoretical results show two ways that ignoring effect heterogeneity can lead to inaccurate inference about both policy counterfactuals and how policy can be improved. First, bias arises when mean effects are not externally valid to match effects from the policy. For example, imagine a medical treatment that did not have serious side effects in the population in general. If we are considering a policy that would target this treatment to new high-risk patients, it is not clear whether the impact will be the same. Second, bias also arises from the covariance across the target population of the heterogeneous effects of a policy and an individual's welfare weights. For example, consider a tax reform that raises post-tax incomes by \$3000 to the richest 50% of households but reduces incomes by \$1000 for the poorest 50% of households. Policymakers may consider this reform undesirable for equity reasons even though it increases average incomes. These biases can both be reduced or eliminated by estimating conditional average treatment effects along appropriate observable dimensions and allowing for heterogeneous welfare weights. When optimizing policy, correcting this bias can lead to significant gains through comparative advantage and allow policymakers to direct interventions towards people with the highest marginal welfare benefit.

These theoretical results highlight an interesting contribution of our paper. As empirical policy evaluations become increasingly common, our theoretical results characterize the trade-offs

implicit in relying on mean impacts. For example, using mean effects to predict the welfare of an allocation is biased in general because welfare depends not just on program impacts and welfare weights but the covariance of the two. Interestingly, this insight is reminiscent of similar results in optimal corrective taxation of heterogeneous consumption externalities (like alcohol). Griffith, O'Connell, and Smith (2019) show that the optimal corrective tax is the average consumption externality *plus* the covariance between individual contributions to the externality (the effect) and demand elasticities (the weight). Furthermore, in the externality context, conditioning (in this case tax differentiation by product) also reduces the bias, as it can in our setting.[1] The importance of heterogeneity and conditioning in these theoretical settings raises questions about whether using average "sufficient statistics" is appropriate when heterogeneous estimates could inform differentiated policies like corrective taxation of heterogeneous *production* externalities (Fell, Kaffine, & Novan, 2021; Hollingsworth & Rudik, 2019; Sexton, Kirkpatrick, Harris, & Muller, 2021). Crucially, we speak to these trade-offs by showing how both biases can be reduced by estimating conditional average treatment effects along observable dimensions to allow for heterogeneity in impacts.

Motivated by the importance of heterogeneity in general, we estimate heterogeneity in teacher value-added along the achievement distribution in the San Diego Unified School District, the second largest district in California. We find large gains from using heterogeneity to more optimally allocate teachers to students. In particular, we use the methods pioneered by Delgado (2022) to estimate the value-added of all third- through fifth-grade teachers on student math and English language arts (ELA) scores allowing for heterogeneous effects on students who had above- and below-median scores the previous year. Although these measures of value-added are correlated with standard (i.e. homogeneous value-added) measures, we find substantial heterogeneity. For example, the average within-teacher difference in value-added across groups (i.e. comparative advantage) is as large as 53% (48%) of a standard deviation in mean value-added for ELA (math). We use these estimates to consider welfare gains from two sets of possible policies: reallocating teach-

[1]The second insight is technically a generalization of the first, which was originally suggested in Diamond (1973).

ers to classes without changing school assignment or allowing for school reassignment.[2] There are enormous gains from reallocation. Over the course of third to fifth grade, using heterogeneous measures of value-added to improve district-wide teacher assignments could raise student math scores by 0.17 student standard deviations on average and ELA scores by 0.12. For context, both changes are roughly equivalent to an intervention improving all teachers' value-added by 30% of the (teacher) standard deviation in the relevant subject.

In this process, our paper makes three innovative contributions to the literatures on value-added and teacher value-added. First, we demonstrate how important achievement is as a dimension of effect heterogeneity in our education context. Whereas many papers have found evidence of "match effects" between students and teachers sharing observable characteristics like gender or race (Dee, 2005; Delhommer, 2019), other results reveal that these match effects only explain part of the heterogeneity in teacher effects on the same dimensions (Delgado, 2022). Our results suggest that focusing on demographic match is incomplete because it overlooks how instructional differentiation along the achievement distribution (well documented in the education literature) interacts with these characteristics. This insight reflects other evidence from health economics that in general lagged outcomes are one of the most important dimensions for match effect heterogeneity (as in Dahlstrand, 2022).

Second, our results highlight how combining information from multiple outcomes substantially improves the welfare gains from reallocations. Although it is not obvious *ex ante* how to address this multidimensionality, our theory suggests combining outcomes based on how they affect long-term outcomes of interest. To this end, we aggregate teacher effects using estimates of the differential impact of elementary school gains in math and ELA on lifetime earnings from Chetty, Friedman, and Rockoff (2014b). Back of the envelope calculations suggest that over three years the allocation of teachers that maximizes present-valued lifetime earnings would generate over $4000 in present valued earnings per student or over $83.7 million in total.[3] Whereas interventions in the

[2]In all reallocations the assignment of students to classes is held constant, as is the grade in which the teacher teaches.

[3]Here present valuation is discounted at 3% following back to age 10 following Krueger (1999) and Chetty et al. (2014b).

education literature have often focused on math scores for a variety of reasons (Bates et al., 2022; Chetty, Friedman, & Rockoff, 2014a; Delgado, 2022; Ricks, 2022), our contribution is accounting for the separate marginal effects of math and reading outcomes, which generates 34% larger wage impacts (value-added of $21 million) relative to focusing only on math.

Third, these results have implications for the discussion of using value-added in teacher (and doctor and hospital) compensation and extend our understanding of the welfare implications of such policies. Motivated by the large earnings gains from reallocations, we explore the welfare implications of using lump-sum transfers to compensate teachers for the possibility of being reallocated. We consider varying sizes of bonus payments to all teachers and find enormous gains measured in the marginal value of public funds (or MVPF (Hendren & Sprung-Keyser, 2020)). The MVPF of bonuses in the district-wide reallocation is infinite for up to $8300 per teacher (roughly 14% of salary for SDUSD teacher with 10 years of experience). For within-school-grade reallocations—which have smaller gains but which should be all but costless to teachers—we find that the MVPF is infinite for bonuses of up to $2200. These ideas combine insights from two literatures on teacher labor markets: one focusing on dismissal (Chetty et al., 2014a; Hanushek, 2009; Staiger & Rockoff, 2010), but sometimes ignoring teacher supply decisions (as pointed out in Rothstein, 2010) and the other characterizing teacher demand (Johnson, 2021) but sometimes ignoring teacher impacts on students (as addressed in Bates et al., 2022, where both are combined). Our contribution is characterizing the welfare effects of policies that use teacher value-added but compensate teachers for the possible disutility of the resulting allocation.

Taken together, our results highlight the first-order importance of considering heterogeneity in empirical welfare analysis. In our theory we show how the gains possible from allocations based on heterogeneous effects may be much larger than those based on means only. We document this empirically in our setting where considering just one dimension of heterogeneity increases test score gains by 34-97% relative to only using the standard value-added measure. While the critical role of comparative advantage has been acknowledged for centuries, our contribution to welfare theory is in connecting treatment effect heterogeneity, comparative advantage, and social prefer-

ences. These connections capture and formalize the growing understanding that heterogeneity is a key consideration for allocating scarce resources according to a social objective by means of targeting. This has been explored theoretically (Athey & Wager, 2021; Kitagawa & Tetenov, 2018) and is reflected in a recent explosion of empirical inquiry about targeting treatments as varied as social safety programs (Alatas et al., 2016; Finkelstein & Notowidigdo, 2019), costly energy efficiency interventions (Ida et al., 2022; Ito, Ida, & Tanaka, 2021), promoting entrepreneurship in developing countries (Hussam, Rigol, & Roth, 2022), and even resources to reduce gun violence (Bhatt, Heller, Kapustin, Bertrand, & Blattman, 2023). Our results suggest that in these settings and others ignoring heterogeneity may have serious welfare ramifications and that considering heterogeneity in effects and social preferences presents a clear path forward for future welfare analyses.

This paper is organized into 6 sections. Section 2 introduces our framework for welfare and value-added with the implications of heterogeneity. Section 3 contains our estimation procedure and a description of value-added in the San Diego Unified School District. Section 4 leverages our welfare theory to explore the reallocation of teachers to classes and measures the welfare gains from using information about heterogeneity. Finally, Section 5 draws the pieces together to explore the implications for welfare and Section 6 concludes.

2.2. A Welfare Theory of value-added

This section formalizes the implications of estimating mean-oriented statistics for use in welfare analyses and the benefits of estimating heterogeneous impacts. We begin by showing how a welfare-theoretical framework can allow a social planner to aggregate over multidimensional policy impacts on a heterogeneous population. Second, we show how relying on average effects and average welfare weights can lead to biased welfare estimates. This bias has two sources: average treatment effects have imperfect external validity in different allocations (for example assigning teachers to classes with different compositions), and average welfare weights ignore heterogeneous gains to groups with different welfare weights (for example, differential valuation of an identical test-score increase for struggling versus advanced students). Third, we show how mea-

suring heterogeneity along key dimensions can minimize the bias. Finally, we show graphically how correcting this bias leads to better policy optimization through comparative advantage and targeting interventions towards the recipients with the highest marginal benefit.

2.2.1 Welfare with Heterogeneity and Multidimensionality

Consider a social planner selecting a policy $p \in \mathcal{P}$. This policy could be assigning teachers to classes (our application), defining an eligibility threshold for a means-tested program like health insurance, or choosing between various public works projects. The welfare under policy p is a function of the lifetime utilities U_i^p and welfare weights ϕ_i^p of each person i under each policy p. With a population of size n welfare is

$$\mathcal{W}^p = \sum_{i=1}^{n} \phi_i^p U_i^p$$

If the policy p has heterogeneous effects on utility for different people, using welfare weights ϕ_i^p is a long-standing method to allow the social planner to aggregate over individuals and recover a scalar measure of welfare.

In practice neither policymakers nor economists observe lifetime utility directly. Instead, they usually rely on observable outcomes Y like earnings, health outcomes, or test scores as proxies. We let the social planner evaluate policies using a "score function" $S_i^p = s(\boldsymbol{Y}_i^p, \boldsymbol{X}_i)$ which produces an individual-level score for the policy based on observable outcomes and characteristics. Note that while this score could represent any social objective, identifying the expected lifetime utility or earnings would be particularly useful in many cases (see the related work on surrogate indices by Athey, Chetty, Imbens, & Kang, 2019). Just as the welfare weights allow the social planner to aggregate over the heterogeneous effects of the policy, the score function allows the social planner to aggregate over the multidimensional effects of the policy.

Under this setup, a policymaker can evaluate each policy p based on observable outcomes. Assuming an individuals' outcomes $\boldsymbol{Y}_i^p$ only impact their own utility and weights, the expected

change in welfare from the status quo ($p = 0$) to policy p is

$$\Delta\tilde{\mathcal{W}}^p \equiv \sum_{i=1}^{n} \gamma_i(S_i^p, S_i^0)\Delta S_i^p \tag{2.1}$$

where $\gamma_i(S_i^p, S_i^0)$ is a new welfare weight and ΔS_i^p is the effect of policy p on individual i's score. The weight γ_i^p reflects the average welfare gain from marginal score changes over $[S_i^0, S_i^p]$, incorporating the change in expected utility and the relevant welfare weights, ϕ_i^p. A detailed explanation of this derivation can be found in Appendix 2.A.2.1.

Unfortunately, estimating this welfare metric has a major complication: The effects of the policy ΔS_i^p and the proper weights γ_i^p are both individual specific. The impact of the policy on the score, ΔS_i^p, and the impact of the score on lifetime utility, γ_i^p, may both vary from student to student. Even though these individual-level measures provide a more accurate theoretical framework, using individual welfare weights and individual outcomes to assess policy is typically not feasible. Because of this limitation, policies are often evaluated with aggregate measures. We now characterize the bias that this aggregation produces and how estimating heterogeneous effects can reduce that bias.

2.2.2 Bias from Ignoring Match Effects or Individual Welfare Weights

Empirical analyses often simplify the weights and treatment effects to means in order to measure welfare. This approach multiplies an estimate of the average treatment effect of a policy $\widehat{ATE}^p$ with the average welfare weight for the impacted population (see intuition in Hendren & Sprung-Keyser, 2020). Assuming the average welfare weight is known $\mathbb{E}[\gamma^p] = \frac{1}{n}\sum_{i=1}^{n} \gamma_i(S_i^p, S_i^0)$, this approach allows for two sources of bias.[4] First, because the true ATE^p is rarely known (and never known *ex ante*), other estimates such as rules-of-thumb and estimates from different times or populations are used. For example, in the value-added setting a teacher's average impact on a different class in the past is often used to infer their impact on another class in the future,

[4]In practice the average welfare weight needs to be estimated as well, which could introduce a third source of bias, so we assume that policymakers have prior knowledge about the average welfare weight.

introducing bias. Second, as shown in Appendix 2.A.2.2, the welfare weights that convert a true ATE^p into welfare are a function of the joint distribution of the individual-level treatment effects and individual welfare weights. By instead using the simple population mean $\mathbb{E}[\gamma^p]$, more bias is introduced. In general, these simplifications lead to a biased measure of welfare:

Theorem 1. If welfare is estimated using the product of an average outcome from a different population $\widehat{ATE}$ and an average welfare weight $\mathbb{E}[\gamma^p]$, then the estimate will contain the following bias relative to the more general benchmark in Equation 2.1:

$$\begin{aligned}\textbf{Average Bias}_{ATE} &= \frac{\Delta\widetilde{\mathcal{W}}^p}{n} - \mathbb{E}[\gamma^p]\widehat{ATE} \\ &= \mathbb{E}[\gamma^p]\left(\mathbb{E}[\Delta S^p] - \widehat{ATE}\right) + \mathrm{Cov}(\gamma^p, \Delta S^p)\end{aligned}$$

Proof in Appendix 2.A.2.3

With the equation for the bias in hand, we see that these common simplifications lead to two sources of bias. First, one source of bias comes from the difference in the expected change in our outcome of interest, and the $\widehat{ATE}$ estimate used. While these statistics could differ for any reason relating to the external or internal validity of our estimate, our paper is most interested in a specific concern with external validity: Whether averages of heterogeneous effects apply in different populations. For example, if teachers have heterogeneous impacts on students, then estimating the average treatment effect on their current class will not give an unbiased estimate of their average impact on a class of very different students. If, for example, we change the class composition to better match the teacher's comparative advantage, their average impact will increase. A more formal explanation of this impact can be seen in Appendix 2.A.2.4.

Second, using the population average welfare weight ignores any covariance between welfare weights and treatment. While not the case in general, there are some situations where the covariance would be zero. For example, when the effects of a policy are uniform (or random) there can be no covariance. Perhaps more relevant to policy the covariance will also be zero when there

is no variation in welfare weights among the impacted population. This may approximately hold, for example, for targeted programs like SNAP, Medicaid, and TANF. The covariance is likely to matter in many other settings. For example, in our setting teacher reassignment has the potential to disproportionately help low-performing students. If low-performing students have higher welfare weights, the covariance term in the bias would be positive and means would understate the value of the reallocation.

2.2.3 The Case for Estimating Heterogeneity

Measuring heterogeneous impacts along key dimensions can lower the bias outlined above. By choosing features that explain the most variation in welfare weights and policy impacts, we may be able to lower the bias significantly. In practice, this method requires estimates of the conditional average treatment effect and welfare weights by subgroup ($\widehat{CATE(x)}$ and $E[\gamma^p|x]$) rather than using average treatment effects and weights. Incorporating this, the bias can be characterized in the following way:

Theorem 2. If mean welfare is estimated using the weighted mean of a conditional average treatment effect $\widehat{CATE(x)}$ and a conditional average welfare weight $E[\gamma^p|x]$ weighted by the fraction of the population with characteristic x, P_x, the mean welfare estimate will contain the following bias:

$$\begin{aligned}\textbf{Average Bias}_{CATE} &= \frac{\Delta\tilde{\mathcal{W}}^p}{n} - \sum_X P_x E[\gamma^p|x]\widehat{CATE(x)} \\ &= \sum_x P_x \left(\text{Cov}(\gamma^p, \Delta S^p|x) + E[\gamma^p|x] \left(\mathbb{E}[\Delta S^p|x] - \widehat{CATE(x)} \right) \right)\end{aligned}$$

If the features in x are chosen carefully, both portions of the bias can be lowered while still being identifiable. To be more precise, we will again consider the two bias terms separately and compare them to the unconditional counterpart in Theorem 1.

First, consider the covariance terms. The covariance term in Theorem 1 has been replaced

by the weighted sum of conditional covariance terms. Using the law of total covariance, we can see that this portion of the bias will be smaller after conditioning, when

$$\left|\sum_X P_x \text{Cov}(\gamma^p, \Delta S^p|x)\right| < \left|\sum_X P_x \text{Cov}(\gamma^p, \Delta S^p|x) + \text{Cov}(\mathbb{E}[\gamma^p|x], \mathbb{E}[\Delta S^p|x])\right| = \left|\text{Cov}(\gamma^p, \Delta S^p)\right| \tag{2.2}$$

This means that when the average within group covariance between γ^p and ΔS^p is smaller than the total covariance, the bias will be reduced. The middle term breaks up the total covariance into two parts. The first term is the within group covariance, and the second is the covariance of the group means. To better connect these terms to applications, it is helpful to think through cases. First, if both of these terms are the same sign, the condition will be met. Consider a case where we condition on pre-test scores, like our paper, but race also impacts γ and is not conditioned on. If the gains from a teacher allocation are positively (or negatively) correlated with both the welfare weights on both pre-test scores and race, the condition is met. Now suppose they are opposite signs. That is, the gains are positively associated with test score and negatively associated with the welfare weights on race or visa-versa. In this case, the inequality may or may not be satisfied. It will still be satisfied when

$$2 * \left|\sum_X P_x \text{Cov}(\gamma^p, \Delta S^p|x)\right| < \left|\text{Cov}(\mathbb{E}[\gamma^p|x], \mathbb{E}[\Delta S^p|x])\right| \tag{2.3}$$

Put simply, this holds when the within group covariance is small relative to the group mean covariance. In keeping with our example, the within group covariance would be small if the un-conditioned feature, race, either does not impact γ^p very much after conditioning on pretest scores, has little association with ΔS^p after conditioning on pretest scores, or their relationship happens to be randomly distributed after conditioning on pre-test scores. The group mean covariance will be large if the conditioned factor, pre-test-scores, plays a large role in the relationship between γ^p and ΔS^p. For example, suppose pre-test groups with large welfare weights also see large test

score gains because teachers are sorted according to their comparative advantage along the pre-test dimension.

Now to consider the second term. As before, this could come from any external or internal validity issue with $\widehat{CATE}(x)$, but we focus on the bias from population changes interacted with heterogeneous treatment effects. If a teacher has different impacts on different types of students, for example, and the class composition changes, their average impact will change. By conditioning on the observable, x, we can adjust for compositional and treatment effect differences over X. The new estimator takes a teacher's average impact on group x and weights that impact by the composition of their new class. The remaining bias, then, would need to come from differences in treatment effects along other dimensions and variation in composition within a group x across classes. Pulling out the terms, this will be smaller when the following holds.

$$\sum_x P_x E[\gamma^p|x] \left(\mathbb{E}[\Delta S^p|x] - \widehat{CATE}(x)\right) < \mathbb{E}[\gamma^p] \left(\mathbb{E}[\Delta S^p] - \widehat{ATE}\right) \tag{2.4}$$

A more formal treatment can be seen in Appendix 2.A.2.5.

Putting these ideas together, there are two special cases that are helpful to think through. first, the case where welfare weights really only depend on x. For example, if x is pretest scores and the policymakers want to treat every student with the same pre-test score equally. In this case, the first term goes to zero since there is no covariance within test score groups. There could still, however, be differences in treatment effects and class composition within a test score group x. For example, if teachers have differential impact by race (Delgado, 2022). This would lead to a non-zero value for the second term. If there is no heterogeneity within x, either because the treatment effects are the same or the class compositions are the same within x, the second term would also be zero and we would have a completely unbiased estimator. These special cases help to highlight how the first term is driven by the policymaker's re-distributive preferences while the second is driven by the heterogeneous treatment effects and compositional differences between sup-populations.

Given these differences, it is worth noting that there is no reason one could not condition the

welfare weights and the estimates on different subsets of $\boldsymbol{X}$. for example, $E[\gamma^p|x_1]\ \widehat{CATE(X_2)}$. It might be the case that a variable is not meaningful in the welfare weight, but is a factor in estimating an accurate treatment effect. While this could be done, we focus on the case where the same variable, pre-test scores, is being considered for both.

2.2.4 Graphical Intuition of the Welfare-Relevant Components

Having illustrated how to reduce bias for welfare estimates of a given policy intervention, this section considers the welfare gains from decreased bias when comparing different policies. We present a simple example with two groups to show how heterogeneous estimates allow welfare improvements relative to evaluations based on means. For simplicity of exposition, we assume that all effect heterogeneity and heterogeneity in social preference relates to these two groups. This highlights three channels for gains from reallocations—some of which are only possible by estimating heterogeneity.

We illustrate these three channels for improving welfare in Figure 2.1. The two axes of Figure 2.1 depict the average change in the score function for two groups. In our example it would depict the average change in math scores for lower- and higher-scoring students. Connecting these two axes are two production possibility frontiers (PPFs—depicted as curves). Allocations between the origin and "PPF: ATE" are possible by using information about mean effects that capture absolute advantage—such as a teacher's average test-score value-added on students.[5] In our setting this would mean assigning teachers with higher overall value-added to larger classes, and teachers with lower value-added to smaller classes. Allocations within the "PPF: $CATE$" are possible by using information about heterogeneous effects that capture both absolute and comparative advantage. In our setting this would mean also assigning teachers to classes with larger shares of the group they have a comparative advantage in teaching. This PPF is at least weakly dominant because it allows for additional gains from matching teachers to classes in ways that leverage their heterogeneous value-added across student groups.

[5]Technically, a valid value-added estimator is only a consistent estimate for this parameter as the set of students a teacher teaches approaches a representative sample.

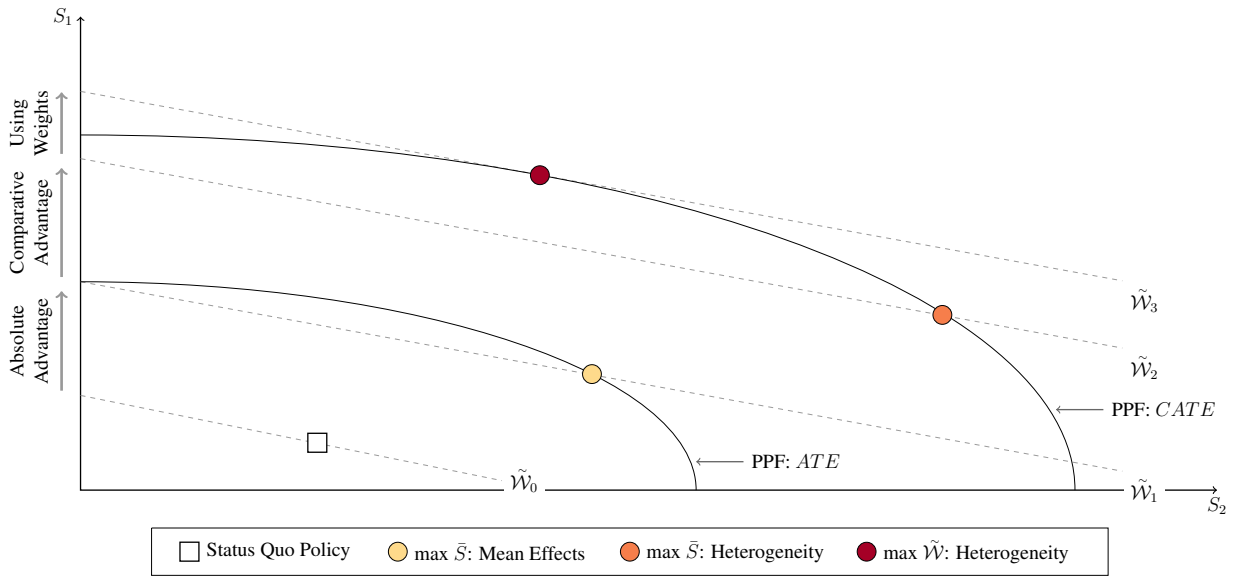


Figure 2.1: **Absolute Advantage, Comparative Advantage, and Social Preferences Contribute to Welfare**

Note: This figure illustrates the welfare gains allocations using heterogeneous effects and welfare weights. The two axes present the outcome score of interest, S, for individuals of two types. The graph contains two production possibility frontiers and some indifference curves. The interior production possibility frontier is attained by allocations made with the constant-effects model, like traditional value-added measures. These mean estimates could enable welfare gains from allocations based on the absolute advantage (possibly weighted by social preferences). The second, dominant frontier is attained by allocations using information about effect heterogeneity and, thus, comparative advantage. The indifference curves show the welfare value of four allocations: (1) the status quo, (2) the average-score maximizing allocation using mean effects, (3) the average-score maximizing allocation using heterogeneous effects, and (4) the welfare maximizing allocation using heterogeneous effects.

Now consider a policymaker with indifference curves corresponding to the dotted lines. The slope of these indifference curves indicates the relative preferences given to one group versus the other. In this example, the slope is higher than -1, indicating that the policymaker places greater weight on group 1. Figure 2.1 presents the status quo and three possible reallocations (a white box and colored circles) and their corresponding welfare (indicated with dashed indifference curves).

First, a policymaker trying to maximize test scores (despite having re-distributive goals) using standard value-added measures can experience welfare gains from the absolute advantage of teachers. Figure 2.1 represents this reallocation as a movement from the white box to the yellow circle on PPF: ATE with welfare gains corresponding to a move from $\tilde{\mathcal{W}}_0$ to $\tilde{\mathcal{W}}_1$[6]. This movement reflects the gains from making allocations based on absolute advantage.

Second, a policymaker maximizing test scores with heterogeneous estimates of teacher value-added (but still ignoring their re-distributive preferences) can experience further gains from the comparative advantage of teachers. With heterogeneous estimates, the policy makers can assess how a teacher would impact students in each group in addition to students on average. This knowledge would allow them to reallocate teachers based on absolute and comparative advantage, indicated as a movement from the white box to the orange circle on PPF: CATE with welfare gains corresponding to a move from $\tilde{\mathcal{W}}_0$ to $\tilde{\mathcal{W}}_2$.[7] Compared to the allocation on PPF: ATE, the gains from $\tilde{\mathcal{W}}_1$ to $\tilde{\mathcal{W}}_2$ reflect the additional gains from making allocations based on comparative advantage.

Finally, a policymaker can produce further welfare gains by directly considering their distributional goals. In our example, the policymaker wants to focus on lower-scoring students for educational remediation (although a focus on higher-scoring students, perhaps for prestige, is also

[6]Note that, in our case, for these gains to be non-zero, two things must be true: it must be the case that (1) some classes have different sizes, and that (2) some teachers have different value-added scores. If these conditions are met a policymaker would expect to increase the scores for students in both groups by assigning higher-value-added teachers to the larger classes. Such reallocations can lead to meaningful impacts in the real world setting we use, where class size averages about 27 with a standard deviation of about 6.

[7]Note that, in our case, for these gains to be larger than the gains from absolute advantage, two more things must be true: it must be the case that (1) some classes have different compositions of student types, and (2) that some teachers have different value-added on each type of student. If these conditions are met a policymaker would expect to further increase the scores for students in both groups by assigning better matched teachers to classes.

possible). If this is the case, both score-maximizing allocations are sub-optimal. This loss is visualized in Figure 2.1 where the indifference curves at $\tilde{\mathcal{W}}_1$ and $\tilde{\mathcal{W}}_2$ are not tangent to either PPF. As such, the policymaker can increase welfare by trading off the possible test-score gains for one group against gains to the other groups. The optimal consideration moves them to the red point, with the largest welfare of $\tilde{\mathcal{W}}_3$.

Although each of these pieces could generate large welfare gains in theory, whether there are meaningful gains from estimating heterogeneity in practice remains an empirical question. For example, if teacher effects are homogeneous or highly correlated there would be no gains from making allocations based on comparative advantage. Furthermore, even if there are differences or distributional objectives, if the status-quo allocation already takes them into account, there would be no gains from reallocations since the welfare gains have already been captured. The remaining sections of the paper measure the amount of heterogeneity in teacher impacts and describe the welfare effects of possible reallocations.

2.3. Estimating Heterogeneous value-added for Teachers in San Diego Unified

Having established how measuring effect heterogeneity could be useful for informing welfare and policy, this section sets the groundwork for determining to what extent heterogeneity in teacher value-added matters in practice for the allocations of teachers to classes in elementary school. To that end, we describe the data from the San Diego Unified School District, present our estimation strategy for value-added, and summarize patterns in value-added—including the extent of comparative advantage and how it is at play in the status quo allocation of teachers to classes.

2.3.1 Background and Administrative Data

To consider socially optimal allocations of teachers to classes, we use administrative data on the universe of students attending schools in the San Diego Unified School District (SDUSD). For our main analyses we focus on 1,816 teachers who are the main instructors in third, fourth, or

fifth grade classes in the 2002-03 through 2012-13 school years.[8] We link all teachers to their students each year and we restrict our attention to students with test scores in both English Language Arts (ELA) and math for two consecutive years. This leaves us with 196,452 student-year observations in 10,447 class-year groups. The administrative data also contain relevant information about student demographics and academics as well as long-term outcomes. We provide more descriptive statistics and information about the current allocation of teachers to classes in Section 2.3.4.

2.3.2 Estimation Overview

We use the data from San Diego Unified to evaluate the importance of estimating heterogeneity in optimally assigning teachers to classes. While there are many dimensions over which we could estimate heterogeneous effects, we focus on lagged student scores. Specifically, we estimate the value-added of each teacher on the Math and ELA scores of students with below-median scores (lower-scoring students) and students with above-median scores (higher-scoring students). Our theory suggests that to be welfare improving the dimension we choose should capture a lot of the variance in impacts and be relevant to the social planner. We estimate heterogeneity along the achievement distribution because it meets these criteria.

First, measuring heterogeneity in teachers' effects on lower- and higher-scoring students captures the most salient dimension of instructional heterogeneity. This intuition is not just based on anecdotes; indeed, the large education literature about instructional differentiation suggests that teaching lower- and higher-scoring students requires very distinct skills. See for instance the large literature on differentiated instruction (see Betts, 2011; Duflo, Dupas, & Kremer, 2011; Tomlinson, 2017, for review and examples). Furthermore, while many papers have found evidence of "match effects" between students and teachers sharing observable characteristics like gender or race (Dee, 2005; Delhommer, 2019), results from Delgado (2022) shows that these match effects only explain part of the heterogeneity in teacher effects on students of different genders and races. This suggests that focusing on demographic match may be overlooking something key. We suggest that the most

[8] We limit to these years because the state-mandated tests were stable and comparable over these years.

relevant dimension is related to differentiation along the test-score distribution.

Second, policymakers often expressly identify achievement as a dimension over which they have heterogeneous valuations of gains. For example, quintessential US policies like the federal No Child Left Behind Act of 2001 directly focused on accountability for and proficiency among lower-scoring students. The stated goal was to focus on raising the lower bound of student test scores, calling for corrective action based on whether the lowest performing groups met state standards.[9] At the same time, many national, state, and local policies promote gains to lower-scoring students while expressing nondiscriminatory, identical preferences for students of different genders, races, and socioeconomic statuses conditional on their achievement.

Standard value-added

For our traditional value-added estimates we follow the approach in Chetty et al. (2014a) and implement it with associated Stata package (Stepner, 2013). The details are presented in Appendix 2.A.3, but the general approach has three steps. First, we estimate the effects of student i's characteristics in year t, $X_{i,t}$, on test scores in subject s, $S_{i,s,t}$, in a regression of the form:

$$S_{i,s,t} = \beta_s X_{i,t} + u_{i,s,t}$$

Second, we obtain the average of the residuals implied by β_s by class and year:

$$\bar{A}_s^{j,t} = \frac{1}{n_{j,t}} \sum_{i:\mathcal{J}(i,t)=j} \left[S_{i,s,t} - \hat{\beta}_s X_{i,t} \right]$$

Finally, we estimate leave-year-out (jackknife) measures of teacher impact by predicting $\bar{A}^{j,t}$ with the residuals in all other years.

$$\hat{\tau}_s^{j,t} = \hat{\boldsymbol{\psi}}_s \bar{\boldsymbol{A}}_s^{j,-t} \tag{2.5}$$

[9]The fact that these policy objectives often find broad cross-partisan support could lead one to conclude that all policymakers have somewhat egalitarian preferences and that disagreements are not questions of direction but only magnitude.

The main assumption necessary to interpret these estimates as causal effects is that class-level shocks and idiosyncratic student-level variation are conditionally independent and a stationary process (given the controls, $X_{i,t}$). It must also be the case that the variance in teacher value-added is stationary (as outlined in Chetty et al., 2014a, —again formal details are in Appendix 2.A.3).

To the end of establishing this conditional independence, we follow the controls of Chetty et al. (2014a), documented to have unbiased estimates of teacher effects. In our setting $X_{i,t}$ includes cubic polynomials in prior year test scores in math and ELA, those polynomials interacted with student grade level, as well as controls for ethnicity, gender, age, the lagged percentage of days absent, indicators for past special education and English language learner status, cubic polynomials in class and school-grade means of prior test scores in both subjects (also interacted with student grade level), class and school means of all the other covariates, class size, and grade and year indicators.[10]

Heterogeneous value-added

For our estimates of heterogeneous value-added, we follow the approach pioneered in Delgado (2022) and applied in Bates et al. (2022), implemented with extensions we made to the Stepner (2013) Stata package. The details are also presented in Appendix 2.A.3, but the general approach also has three steps. The first step is identical, with the addition of indicators for group g to $X_{i,t}$ We then obtain the average of the residuals implied by β_s by class, type, and year:

$$\bar{A}^{j,t}_{g,s} = \frac{1}{n_{j,t,g}} \sum_{i:\mathcal{J}(i,t)=j,g_i=g} \left[S_{i,s,t} - \hat{\beta}_s X_{i,t} \right]$$

[10]The only notable difference from the controls in Chetty et al. (2014a) is their inclusion of information about free and reduced price lunch, which we omit in our research because of restrictions that SDUSD imposes on researchers' use of this information due to their perception of federal regulations on use of student level subsidy information.

Finally, we estimate leave-year-out (jackknife) measures of teacher impact by predicting $\bar{A}^{j,t}$ with the residuals in all other years using the observed auto-covariance.

$$\hat{\tau}_{g,s}^{j,t} = \hat{\boldsymbol{\psi}}_{g,s} \bar{\boldsymbol{A}}_{s}^{j,-t} \tag{2.6}$$

Here the main assumption necessary to interpret these estimates as causal effects is that, class-*type*-level and student-level variation are conditionally independent and stationary processes (as derrived in Delgado, 2022, —again formal details are in Appendix 2.A.3). Note that we differ from Delgado (2022) in one way: We impose a zero-covariance assumption about the idiosyncratic teacher value-added components across groups, similar to the assumptions implicit in the measurement of value-added across subjects in both Chetty et al. (2014a) and Delgado (2022) for internal consistency.

2.3.3 Heterogeneity Highlights the Importance of Comparative Advantage

We use these techniques to estimate the heterogeneous effects of 1,816 teachers on 109,125 lower-and higher- scoring students from 127 elementary schools in SDUSD. These teachers taught grades 3-5 in the 2002-03 to the 2012-13 school years. In this section, the mean value-added is normed to zero for each group, reflecting both the economic intuition that for the average student the "outside option" for the teacher she or he has is the average teacher and the econometric identification argument in Chetty et al. (2014a) implicit in our identifying assumptions.

We depict the main value-added results in Figure 2.2. This Figure reports two scatter plots—one for ELA and one for math—where each point represents one teacher. The teachers value-added on higher-scoring students is plotted on the y-axis over their value-added on lower-scoring students on the x-axis. Each plot also presents the correlation coefficient between the value-added on the two student groups as well as a slope coefficient for the line of best fit between the two.

Visual inspection of Figure 2.2 illustrates the differences within *and* across teachers, sug-

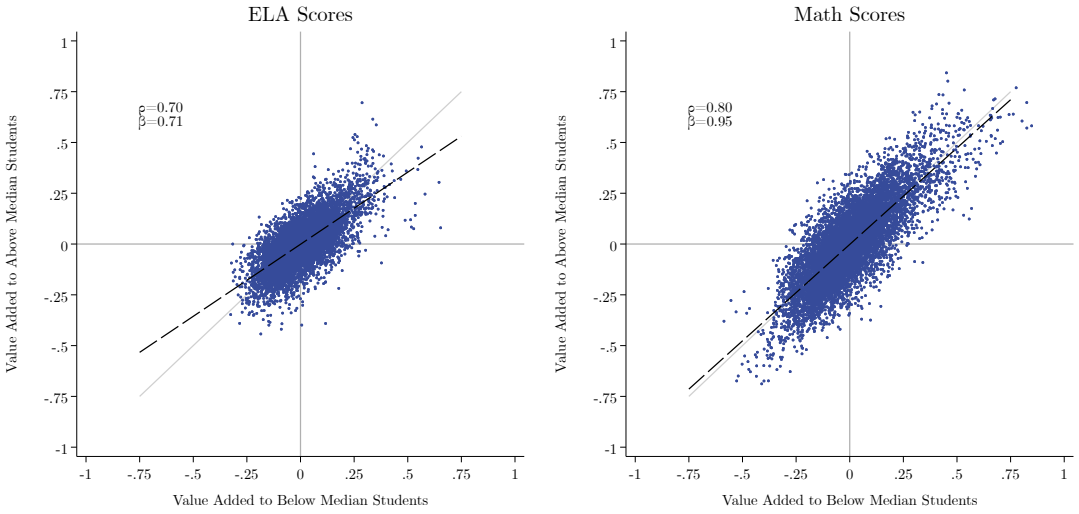


Figure 2.2: **Value-added Varies Significantly within and across Teachers**
Note: This figure shows our heterogeneous estimates of teacher value-added on both English Language Arts (ELA) and Math test scores. Each dot represents one teacher-year estimate of value-added on high- and low-scoring students. The correlation coefficients is for the entire population stacked by year. The dashed line shows the line of best fit with the slope reported. For reference a line with slope one is plotted in the background.

gesting we should reject the standard "constant effects" model of value in favor of one with appreciable comparative advantage. Differences across teachers, or absolute advantage, can be seen by comparing teachers along the gray 45-degree line. Teachers above and to the right generate larger testing gains compared to teachers below and to the left. Comparative advantage can also be seen visually. Teachers with dots above the gray 45-degree line have a comparative advantage in teaching higher-scoring students, and teachers with dots below that line have a comparative advantage in teaching lower-scoring students. The size of the average comparative advantage is large: 53% the size of the cross-teacher standard deviation in standard teacher value-added for ELA and 48% for math.

The differences within and between teachers are what will generate gains for the reallocation exercises. We estimate that teacher value-added to higher- and lower-scoring students is correlated at 0.7 for ELA and 0.8 for Math. The fact that this correlation is less than one allows for gains from allocating teachers by comparative advantage. Even though the correlations are high,

there are still significant margins for gains. For comparison, our cross-group correlations are lower than those by socioeconomic status (0.9 for math in Bates et al., 2022) but larger than those by race (0.7 for math and 0.4 for ELA in Delgado, 2022). Furthermore, our theoretical framework suggests there is value in combining information from multiple outcomes. In that light, it is also worth noting that the cross-subject correlations are lower. For example, Figure 2.11 shows that the cross-subject, cross-group correlations are both around 0.6, suggesting even larger gains from cross-subject comparative advantage.

It is also interesting to note that Figure 2.2 reveals that value-added to math is much more dispersed than value-added to ELA. This is consistent with evidence from similar value-added papers (e.g., Chetty et al., 2014a). Our results further show that teachers' value-added is more highly correlated across achievement groups for Math than for ELA. This is also consistent with absolute advantage being more important and variable with Math teaching than with ELA teaching.

Validation and Robustness

Although these results suggest striking patterns of comparative advantage, our reallocation exercises and welfare estimates would be meaningless if these estimates reflected idiosyncratic noise rather than persistent heterogeneity within and across teachers. Although the use of shrinkage assuages these concerns, we also perform three additional exercises demonstrating the stability and credibility of our heterogeneous estimates. Each result reinforces our confidence that the value-added scores are fitting systematic patterns in causal differences and not just idiosyncratic noise.

First, Appendix Figure 2.16 reports patterns of persistence over time. For example, over 40% of teachers have a comparative advantage for teaching one group of students in *all* years, and the year-to-year correlation is between 0.78-0.90 for all estimates. Additionally, Appendix Figure 2.17 leverages the longitudinal nature of our data to show that heterogeneous value-added estimates carry the same information about long term outcomes as traditional value-added estimates (Chetty et al., 2014b). These results show striking similarities between the effects of our estimates and traditional value-added. Furthermore, estimates for each student group are no less precise

suggesting that the variance is loading on the dimension of heterogeneity we specified.

2.3.4 The Status-Quo Allocation of Teachers and Students

This section shows how teachers are allocated to classes in the status quo, whether this allocation is efficient or equitable, and presents descriptive evidence that there may be gains from reallocation. Figure 2.3 presents a binned scatter plot of value-added for each subject over the share of lower-scoring students for that subject. Absolute advantage is reported as the average of teacher value-added on lower- and higher-scoring students, and comparative advantage is reported as the difference.

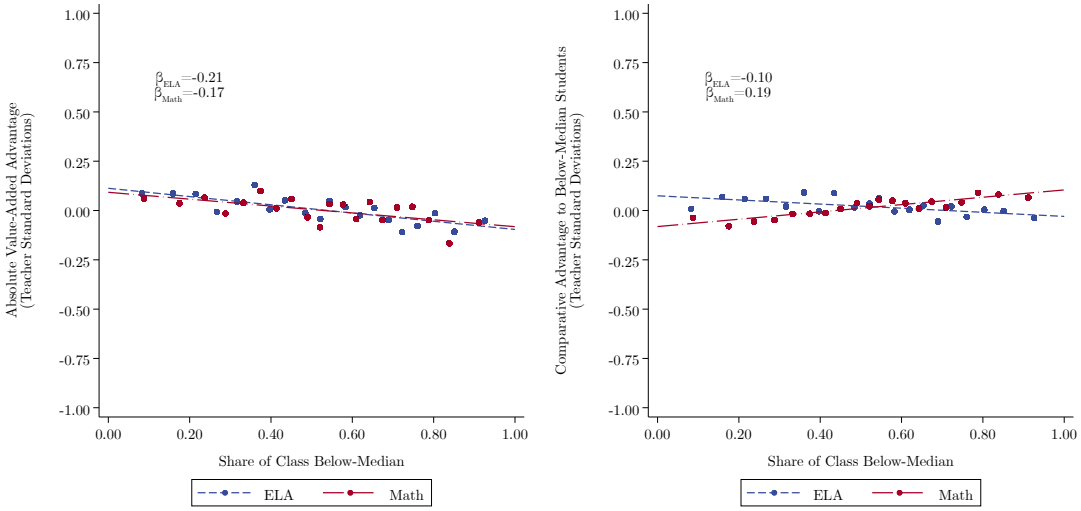


Figure 2.3: **Teacher value-added Only Varies Somewhat with Class Composition**
Note: This figure shows how our heterogeneous estimates of teacher value-added on both English Language Arts (ELA) and Math test scores relate to class composition. The panel on the left shows teacher absolute advantage (average of value-added on lower- and higher-scoring students) and the panel on the right shows the comparative advantage (difference of value-added on lower-scoring students minus value-added on higher-scoring students). both panels plot the ventiles of value-added (measured in teacher standard deviations in absolute advantage) over the share of students who are lower-scoring (i.e. have below-median lagged test scores).

These patterns suggest that classes with larger shares of lower-scoring students do not tend to have teachers with substantially different absolute or comparative advantage. Overall teachers with a higher average value-added are somewhat more likely to sort into classes with higher aver-

age test scores at baseline. This suggests the current allocation is inequitable, but the effects are small: the slope only predicts that students in a class with an additional lower-scoring student in one subject will experience 0.001σ smaller gains in that subject on average. Interestingly, there is some evidence that this slightly inequitable sorting may be according to absolute advantage. Appendix Figure 2.12 shows analogous results by class size revealing that better teachers teach in slightly larger classes, suggesting some allocative efficiency from sorting better teachers in bigger classes, but again the differences are small. These two patterns are likely connected as larger classes tend to be in more affluent schools with higher average test scores.

There is also no clear evidence of sorting on comparative advantage. Figure 2.3 also depicts the difference in value-added to lower- and higher-scoring students along the class test score distribution. In math, teachers who are comparatively better at teaching lower-scoring students are sorting into classes with slightly larger shares of lower-scoring students, but the opposite is true in ELA. Neither of these patterns is economically large. The differences by class size are similarly signed but even smaller (see Appendix Figure 2.12). The combination of heterogeneity in teacher effects and the absence of significant sorting in the status quo suggest large gains from reallocation.

The current allocation of students to classes also suggests that there will be gains from reallocations. Variance in class size and class composition will both increase the gains from reallocation. Appendix Table 2.1 reports the standard deviations of class size and the share of higher-scoring students in math and ELA at a district-wide level and within schools (controlling for variation by grade and year), revealing ample variation even within school. This suggests that although reallocating teachers across schools necessarily allows for bigger test-score gains, much of the potential gains may be achievable by reallocating teachers within their current school and grade.

2.4. Efficiently Allocating Teachers to Classes

Although our general theoretical framework could be applied in many settings, with estimates of the heterogeneous teacher effects we now use our theory to consider the public service

provision problem of allocating teachers to classes. This section defines the allocation problem, presents the gains possible under the optimal allocations, and compares the gains obtained from using our estimates relative to using standard value-added measures.

We parameterize the social objective $\tilde{\mathcal{W}}$ using higher- and lower- scoring students to compare different allocations and find the relevant optima. Let $\mathcal{J} : (i,t) \to j$ be an allocation function, telling us which teachers teach each student in each year. We define the following optimization problem for weighted test score gains in a given subject (s subject subscripts suppressed):

$$\max_{\mathcal{J} \in \mathscr{J}} \tilde{\mathcal{W}}(\mathcal{J}; \omega) = \max_{\mathcal{J} \in \mathscr{J}} \frac{1}{N_{i,t}} \sum_{(i,t)} \omega_L \, L_{i,t} \, \hat{\tau}_L^{\mathcal{J}(i,t)} + (1 - \omega_L) \, (1 - L_{i,t}) \, \hat{\tau}_H^{\mathcal{J}(i,t)} \tag{2.7}$$

where $\omega_L \in [0.0, 1.0]$ represents the weight on lower-scoring students in the social objective, $L_{i,t}$ is an indicator for whether student i is lower-scoring, and $\hat{\tau}_H^j$ and $\hat{\tau}_L^j$ are our estimates of heterogeneous value-added. The set $\mathscr{J}$ is the social planner's choice set made up of feasible allocations. In our setting, we focus only on reallocating teachers to existing classes in the grade they actually taught without changing the composition of those classes in any way. We do this to avoid introducing peer-effect biases into our welfare estimates. The single-ω parameterization of welfare imposes linear indifference curves that trade off performance for lower- and higher-scoring students where the weight on each group reflects the degree to which the social planner wishes to target gains to one group of students relative to the other. It also assumes that the social planner only values gains to students in the given subject—something we will relax in Section 5.

This allocation problem captures three distinct trade-offs that have been mentioned in the value-added literature but never fully addressed together. First, the optimal allocation must account for the *comparative advantage* of teachers because of differences in *class composition* (as pointed out in Delgado, 2022). Second, the optimal allocation must also account for the *absolute advantage* of teachers because of differences in *class size*. This crucial detail has been accounted for at the school level (see Bates et al., 2022), but class size and class composition vary both across *and* within schools. Because of these differences, we are interested in both within-school and district-

wide reallocation exercises. Finally, the optimal allocation must account for possible heterogeneity in the social value of gains to different types of students—something unique to our paper.

We solve this allocation problem for two sets of possible reallocations: within-school and district-wide. For both, we restrict $\mathscr{J}$ so that every year the students in each class and the grade assignments of each teacher do not change. We leave class composition fixed so that changes in within-class peer effects do not contaminate the outcomes in predicted counterfactual allocations. For the within-school reallocation we further require that teachers do not change schools. Whereas this within-school problem can be solved easily by iterating over school-grade(-year) cells, the district-wide reallocation problem has over 3×10^{1830} allocations to search over. Because the optimal policy depends on both absolute and comparative advantage when both class sizes and class compositions vary, this problem cannot be solved by simply assigning teachers to classes with large shares of students they have a comparative advantage in teaching or simply assigning the best teachers to the largest classes. The social planner problem in equation 2.7 can be re-characterized as a mixed-integer linear programming problem and solved using the COIN-OR Branch and Cut solver implemented by the Python package Pulp (see, for example, DeNegre & Ralphs, 2009).

2.4.1 Allocations Incorporating Heterogeneous Impacts Increase Test Scores

We create a production-possibility frontier (PPF) for the gains to each group from the within-school and district-wide reallocations. To do this, we solve the optimization problem in Equation 2.7 for 101 different values of the social weights ω_L ranging from 0.0 to 1.0. We then recover the average value-added received by lower- and higher-scoring students and calculate the gain beyond the status quo. By comparing the optimal gains attained under different weights, this analysis characterizes how reallocation gains to lower-scoring students trade off with those to higher-scoring students, creating the PPFs.

We depict these production-possibility frontiers in Figure 2.4. We plot the PPF for change in ELA scores on the left and Math scores on the right. Each point presents the average one-year change in lower-scoring students' test scores in the optimal allocations (on the y-axis) over

average change for higher-scoring students (on the x-axis), all relative to the status quo (noted with the square marker). Allocations that would reduce a group's scores relative to the status quo are denoted with negative numbers. Allocations above and/or to the right of the status quo are preferred by the social planner. The lighter (blue) PPF denotes the within-school reallocations and the darker (red) PPF the district-wide reallocations. Unsurprisingly, the district-wide reallocations produce gains that are further out in both dimensions.

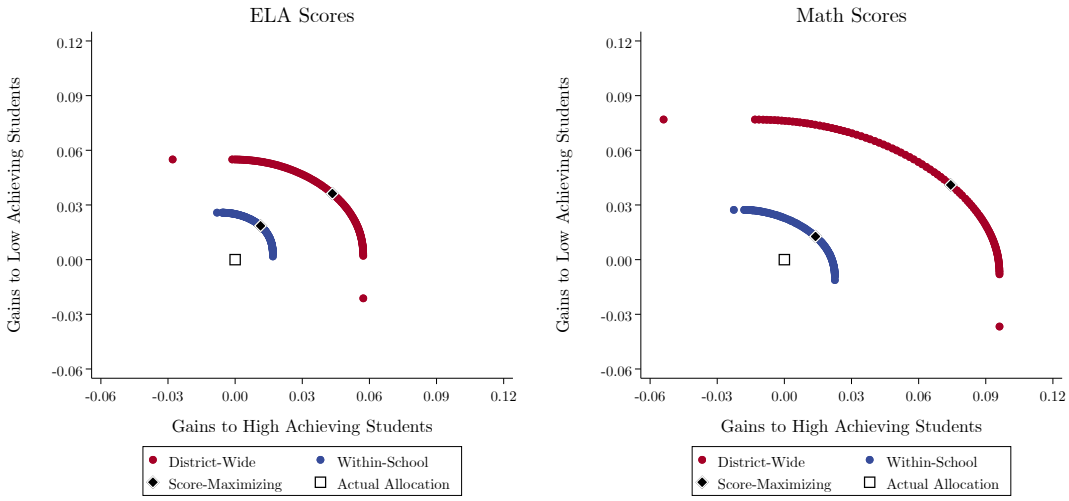


Figure 2.4: **Optimal Allocations Can Create Large Gains to High- and Low-scoring Students** Note: This figure shows the test score gains from optimal allocations relative to the status quo. Two production possibility frontiers are presented, one for reallocating teachers within school-grade cells and one reallocating teachers across schools (still within grade). Each PPF is constructed by finding the optimal allocation given relative weights on lower- and higher-scoring students [0.0,1.0] by solving the optimal mixed-integer linear programming problem. Gains are reported as average changes in scores measured in student standard deviations per school year that the reallocation is performed.

Figure 2.4 reveals three striking patterns. First, there are large gains possible from both reallocations. For example, in the district-wide reallocation a social planner seeking to raise average scores (i.e., a utilitarian planner with $\omega_L = \omega_H = 0.5$) could increase both lower- and higher-scoring students' scores by 0.04 student standard deviations. Gains from math are even larger: 0.04 for lower-scoring students and 0.07 for higher. Similarly, the simpler within-school reallocation could raise ELA and Math scores for both groups by more than 0.01 standard deviations.

Recalling that these represent one-year gains, a policy that optimally allocated teachers could increase average math scores by 0.12σ in ELA and 0.17σ in math.[11] These are large gains—almost identical to the gains that would result from improving the value-added of *every teacher* in the district by one teacher standard deviation (but retaining status quo assignments) for one year, and triple the gains from proposed teacher screening programs that "deselect" (i.e., fire) teachers with the lowest 5% standard value-added (as considered in Chetty et al., 2014b; Hanushek, 2009; 2011).

The second pattern visible in Figure 2.4 is that the curvature of the PPFs demonstrates the value in explicitly considering the distributional goals of a policymaker. These gains are dependent on the extent to which distributional goals deviate from the mean scores objective but are large for more extreme distributional goals.

We compare the total welfare achieved under an optimal allocation for a given set of welfare weights (the optimal point on a PPF in Figure 2.4 for a given indifference curve) to the test-score maximizing allocation (the black diamond mark on the relevant PPF). To normalize these welfare gains, we construct an "Atkinson index" type measure such that the social planner would be indifferent between the optimal allocation and an allocation where every student experienced a given test score gain. Figure 2.5 shows the difference in this Atkinson index for each allocation on the comparative advantage frontier compared to the test-score maximizing allocation. As expected, the gains are small for similar weights and grow as the social planner favors one group more or less. At the tail ends, where the policymaker favors one group almost exclusively, the gains for the district-wide (within-school) reallocations are 85% (20%) larger in math and 50% (35%) larger in ELA. Of course, the true weights for policymakers may not be near these tails, but Figure 2.5 demonstrates significant potential for gains in the right setting. These potential welfare gains highlight the fact that choosing the allocation that maximizes average scores isn't necessarily a neutral choice. For example, in math it benefits higher-scoring students more.

Estimating these gains highlights three interesting implications for our understanding of teacher allocations. First, the gains to math scores are larger than the gains to ELA scores. This

[11] Where the annual means and standard deviations scores are normalized by those in the entire state of California.

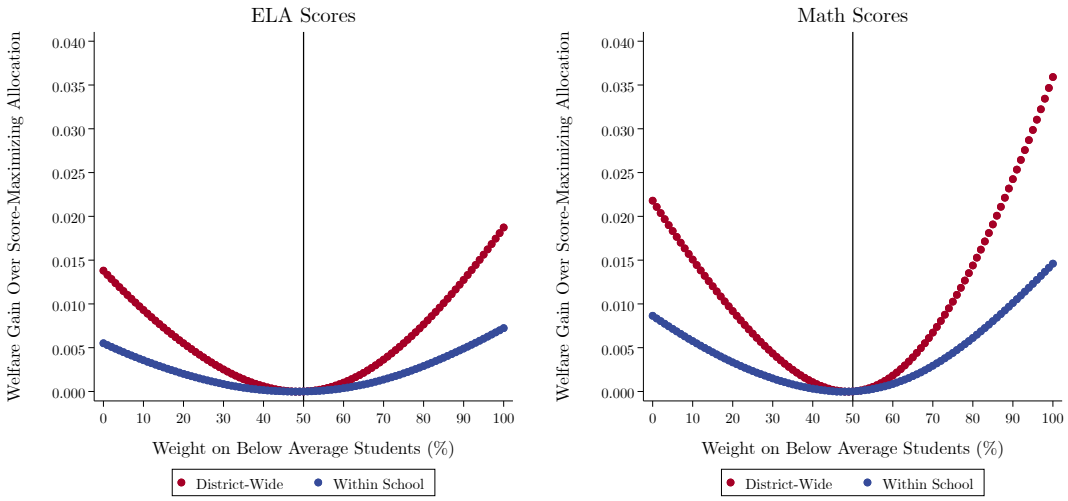


Figure 2.5: **Welfare Gains from Considering Distributional Objectives**
Note: This figure shows the differences in welfare attained under the score maximizing allocation and the optimal allocation using heterogeneous value-added. The unit is an Atkinson Index indifference, i.e., how much would test scores have to increase for all students to generate equivalent welfare gains. We report differences for both within-school and district-wide reallocations.

is because the variance in teacher value-added on math is larger as shown in Figure 2.2 and in prior work (e.g., Chetty et al., 2014a). This suggests that for one-subject reallocations like Bates et al. (2022), it is indeed better to focus on math in order to raise average scores. Second, the allocations that optimize math scores and ELA scores are distinct. This is because the teachers that are the best at teaching each group of students math are not always the best at teaching those students in ELA. As such, the gains highlighted in papers that do reallocations using one subject at a time like Delgado (2022) and Bates et al. (2022) only give a lower bound to the gains from using information on both outcomes simultaneously. This will motivate our analyses in Section 2.5 where we aggregate gains over multidimensional outcomes. Finally, note that the largest possible gains to each group are different. This asymmetry highlights the welfare implications of structural features of the education system such as the fact that higher-scoring students tend to be in larger classes compared to lower-scoring students. This class-size dimension becomes particularly important when comparing these allocations to those made using only information about absolute advantage from traditional value-added estimates.

Before proceeding, we want to note three caveats in considering these reallocations. First, note that because we do not change class composition, these gains could be significantly larger in a district that employs class-level tracking because of greater variance in class composition. Second, the district-wide reallocations might be infeasible. For example, in SDUSD the union contract gives teachers with seniority higher priority in hiring. Furthermore, teachers have strong preferences over locations (Boyd, Lankford, Loeb, & Wyckoff, 2005a) and schools (Bates et al., 2022) that could impede some allocations from being incentive compatible. Finally, the new allocations must be interpreted in the light of partial equilibrium, barring families re-sorting to classes (via requests), schools (via school choice), or districts (via in- or out-mobility).

2.4.2 What Value Does Estimating Heterogeneity Add?

The previous subsection quantified large gains from teacher reallocations, but how much of these gains would be possible without knowing the heterogeneous effects? If all of these gains simply come from moving better teachers to larger classes, there is no need to estimate heterogeneous effects. To evaluate the importance of estimating heterogeneity, we compare the best allocations using heterogeneous estimates with those possible using only standard estimates of value-added. This allows us to decompose the welfare gains from the best allocations into the absolute advantage, comparative advantage, and redistribution components.

To find the optimal allocations with the standard value-added we use the same set of social objective functions and same solution concept, but we replace the estimates of each teacher's value-added on both higher- and lower-scoring students with the standard estimates:

$$\max_{\mathcal{J}\in\mathscr{J}} \tilde{\mathcal{W}}_{VA}(\mathcal{J};\omega) = \max_{\mathcal{J}\in\mathscr{J}} \frac{1}{N_{i,t}} \sum_{(i,t)} \omega_L \, L_{i,t} \, \hat{\tau}_{VA}^{\mathcal{J}(i,t)} + (1-\omega_L)\,(1-L_{i,t})\,\hat{\tau}_{VA}^{\mathcal{J}(i,t)} \tag{2.8}$$

where $\hat{\tau}_{VA}^{j}$ is the standard value estimate described in section 2.3.2 and where we again solve the problem for 101 different values of the social weights ω_L ranging from 0.0 to 1.0. Intuitively, the gains from using absolute advantage as captured in the standard measures come from putting the

higher value-added teachers in larger classes to maximize average scores—or using ω_L-weighted class size when the social planner has heterogeneous preferences over groups' gains. The gains attained and reported at each point are calculated using our heterogeneous estimates to avoid compromising the external validity of our score predictions that would occur if using standard estimates to predict the effect of sending teachers to very different classes.

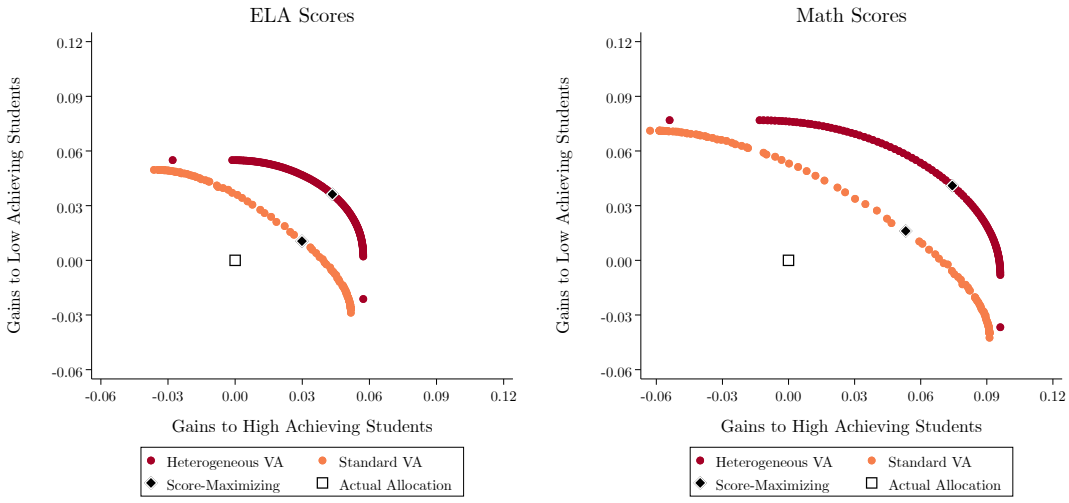
Estimating Heterogeneity Increases Average Test Scores

As illustrated in Figure 2.1, using heterogeneous value-added could increase average scores beyond what is possible using standard value-added via comparative advantage. This subsection explores the extent to which information about comparative advantages can raise average scores in practice. We document large gains beyond what can be accomplished using the information about absolute advantage that standard value-added measures provide.

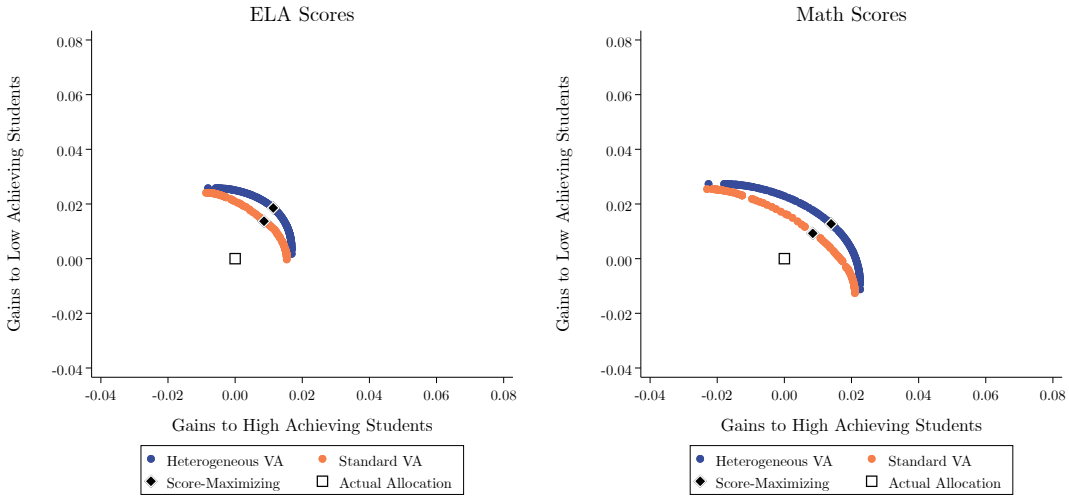
To approach this question, we depict and compare the production-possibility frontiers for average achievement gains to each group using heterogeneous and standard value-added in Figure 2.6. Here again each point presents the average change in lower-scoring students' test scores in the optimal allocations (on the y-axis) over average change for higher-scoring students (on the x-axis). relative to the status quo (noted with the square marker). Panel (a) presents the results from the district-wide reallocation, Panel (b) presents those from the within-school reallocation. These figures also mark the allocations that maximize test scores with a black diamond for reference—which is obtained by placing the highest value-added teachers in the largest classes.

Note that the empirical results in Figure 2.6 are analogous to the theoretical depiction in Figure 2.1. For each panel the outer PPF presents the changes in test scores possible by using information about both absolute and comparative advantage based on the heterogeneous teacher effects whereas the interior PPF presents the changes in test scores possible by using only the information about absolute advantage contained in standard value-added estimates. Again, the current allocation is denoted with a square.

Comparing the optimal allocations reveals that using information about comparative ad-



(a) District-Wide Reallocation



(b) Within-School Reallocation

Figure 2.6: **Using Heterogeneous Estimates Produces Larger Gains from Reallocation**
Note: This figure shows the test score gains from optimal allocations relative to the status quo. In each panel two production possibility frontiers are presented, one for reallocating teachers based on our estimates of value-added (absolute and comparative advantage) and one reallocating teachers only based on traditional value-added (absolute advantage). Panel (a) displays the result for reallocating teachers across schools and panel (b) the results for reallocating teachers within schools (both always keep teacher in the same grade). Each PPF is constructed by finding the optimal allocation given relative weights on low- and high-scoring students [0.0,1.0] by solving the optimal mixed-integer linear programming problem. Gains are reported as average changes in scores measured in student standard deviations per school year that the reallocation is performed.

vantage can as much as double the achievement gains from reallocations. In the district-wide reallocation, allocations using comparative advantage generate 97.3% higher ELA scores and 66.4% higher Math scores than allocations using only absolute advantage. These are large gains: an average gain of 0.020σ in ELA or 0.023σ in Math for students in the district would be an impressive policy victory, especially considering this policy could be implemented year-over-year for compounding gains. Gains to the within-school reallocations are smaller in absolute terms, but comparative advantage is still critical. Using heterogeneous effects boosts average ELA scores by 34.1% and math scores by 50.3% (both about 0.0045σ).

Interestingly, even for a social planner trying to maximize average scores the choice between standard and heterogeneous value-added measures has striking distributional implications in the district-wide allocations. On one hand, the average-score gains from reallocations using only information about absolute advantage (from standard value-added) are concentrated among higher-scoring students. For example, the higher-scoring students' gains of 0.03σ in ELA and 0.05σ in Math are almost exactly three times larger than the corresponding gains to lower-scoring students. On the other hand, the large gains from using comparative advantage in the district-wide reallocations accrue disproportionately to lower-scoring students. For example, the 0.02σ ELA gain is split almost 0.03σ to lower-scoring students and just over 0.01σ to higher-scoring students. Figure 2.6 depicts these observations visibly: Whereas the expansion path from the status quo through the two PPFs is almost linear for the within-school reallocations in Panel (b), it is extremely non-linear for the district-wide reallocations Panel (a). These asymmetries motivate a direct focus on the equity implications of using heterogeneity.

The Interaction of Distributional Goals and Comparative Advantage

The above section shows that when the goal is to maximize average scores, using heterogeneous value-added leads to significant gains. We also know from section 2.4.1 that when policymakers favor one group over another, considering their distributional goals leads to significant welfare gains. Putting these together, we now address how different distributional objectives

impact the gains from comparative advantage, and using heterogeneous value-added.

Using Figure 2.6 as a reference, we now compare the welfare from the optimal points on the inner PPF relying on mean effects and the outer PPF using heterogeneity for a given distributional goal. Reporting the difference in the Atkinson index between the optimal allocations reveals the welfare gains from using heterogeneous value-added estimates for each distributional goal. Figure 2.7 reports the results. In Appendix Figure 2.13, we present a simpler measure: the true (unweighted) difference in average scores for each pair of allocations.

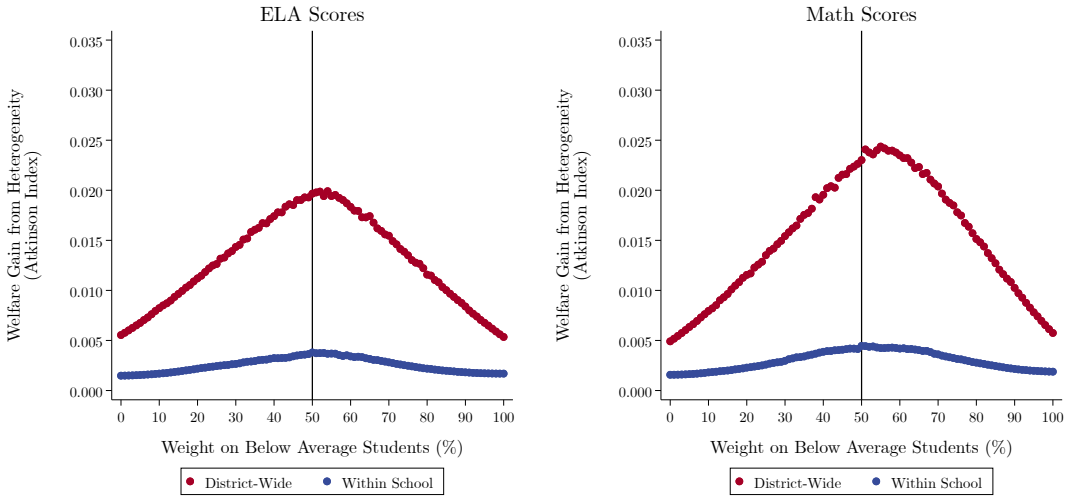


Figure 2.7: **Welfare Gains from Comparative Advantage Along Distributional Objectives** Note: This figure compares the welfare attained at the optimal allocations based on our measures of value-added with those attained at allocations based on standard value-added measures. The unit is an Atkinson Index indifference, i.e., how much would test scores have to increase for all students to generate equivalent welfare gains. We report differences for both within-school and district-wide reallocations.

These analyses reveal that using heterogeneous value-added matters most when the social planner has slightly egalitarian preferences. This is visible in Figure 2.7 where for the district-wide reallocation the highest points on each upside-down U shape are slightly to the right of utilitarian preferences denoted with the gray line (at $\omega_L = \omega_H = 0.5$). Although the maxima, where using heterogeneous value-added is most useful, are at ω_L =0.54 for ELA and 0.55 for math, the entire region between $\omega_L \in [0.30, 0.70]$ show gains equivalent to over 0.015σ of gains to all students.

The comparative advantage gains from estimating heterogeneous value-added are only

large if the social planner cares about both groups. For example, if the social planner only cares about lower- or higher-scoring students ($\omega_L \in \{0.0, 1.0\}$), there are essentially no gains from comparative advantage using heterogeneous value-added. This is because lower- and higher-scoring value-added are positively correlated, so a policy that puts the highest absolute advantage teachers in the class with the most lower-scoring students will have a very similar effect on lower-scoring students to a policy that puts the teachers with the highest lower-scoring value-added in the same classes. This is visible in how close the frontiers are in Figure 2.6 and in the upside-down U-shape in the gains reported in Figure 2.7.

The key driver of these differences are the relative shapes of the PPFs and how they affect scores. As seen in Figure 2.6, the best attainable allocations using standard value-added create a much flatter frontier than those using information about heterogeneity. As a result, the "price" of an additional score increase to one group is much more expensive if the social planner relies only on information from standard value-added measures. This has direct implications for average test scores, as seen in Appendix Figure 2.13. Here we depict the change in average scores generated from moving from the optimal allocation attained using standard value-added to the optimal allocation attained using our heterogeneous estimates. Rather than being U-shaped like the welfare gains, these suggest an M-shape where the score gains are biggest when on these flat regions of the interior PPF, but away from the center where average scores (and thus class sizes) are all that matter.

In summary, comparative advantage and distributional goals are both potentially important to consider, but how each effect interacts with a policymaker's welfare weights means one effect may play a much bigger role for a given policymaker. Redistribution is important when the social planner has very strong preferences for gains to one group relative to another; however, the standard measures of value-added are able to capture most of these gains because value-added heterogeneity is positively correlated within teachers. There is little scope for welfare gains from comparative advantage. Conversely, when a policymaker values gains to each group roughly equally, there is little scope for distributional gains to matter, but significant scope for welfare gains from com-

parative advantage. Since policy suggests some social objectives may be more nuanced, we also turn our attention to the implications of our reallocations for achievement gaps and the creation of winners and losers.

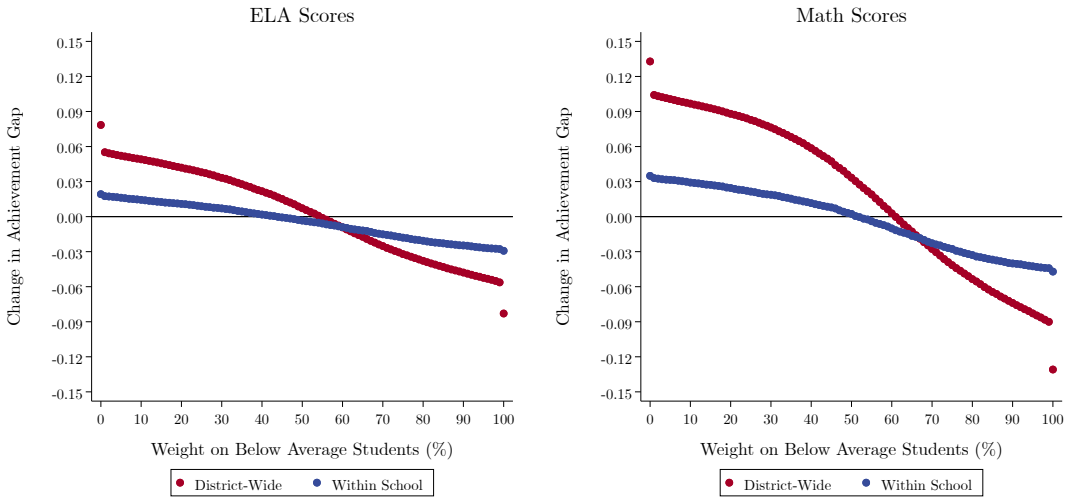
2.4.3 Other Equity Implications from Reallocations

Having described the optimal reallocations and decomposed the welfare gains from them, our final task is to explore other equity implications that the proposed reallocations would have. Specifically, we study how our reallocations affect overall achievement gaps and racial achievement gaps, and we describe how certain allocations that generate gains on average still create significant heterogeneity for winners and losers masked by that average.

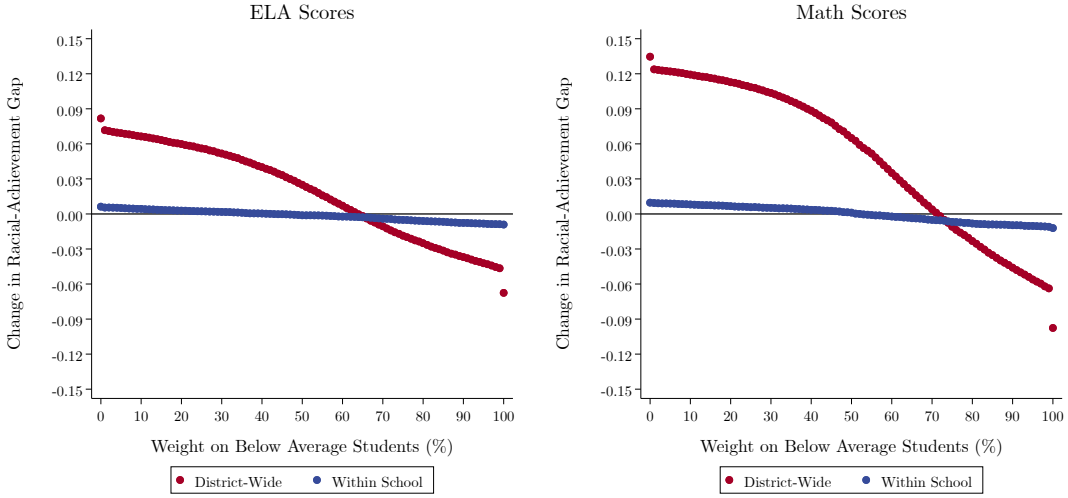
Shrinking Achievement Gaps

Many education policies—including those that motivated our welfare theory—propose interventions that will lower the achievement gaps between lower- and higher-scoring students. To consider this we plot out the change in two policy-relevant achievement gaps in Figure 2.8. First, in Panel (a) we show how the optimal within-school and district-wide reallocations for each ω_L would change the achievement gap between students who performed above and below median in the previous year. We also report similar changes in the racial achievement gap in Panel (b). We define this gap as the difference in average scores between Black and Hispanic students versus White and Asian students. Interestingly, we show that our completely race-blind policies can reduce average racial test score gaps just as much as the race focused reallocations in Delgado (2022).

The main takeaway from these analyses is that a social planner who cares about gaps can partially control the size of the gaps by making allocations that are on the efficiency frontier based on comparative advantage. For example, the baseline gap between students who scored above and below the median last year is 1.27σ in ELA and 1.19σ in Math. A social planner focused on raising lower-scoring students' scores without, on average, hurting higher-scoring students could shrink those gaps by 4.4 and 7.6% *every year*. The gap between Black and Hispanic students versus



(a) Achievement Gaps



(b) Racial-Achievement Gap

Figure 2.8: **Reallocations Can Shrink Persistent Gaps in Student Performance**
Note: This figure shows how optimal reallocations would change achievement gaps between students. Each panel plots the change in the gaps of interest over the relative weights on higher- and lower-scoring students. Panel (a) displays the change in the average difference in test scores between students who scored below versus above the median in the previous year (relative to about 1.2σ), and Panel (b) displays the change in the average difference in test scores between Black and Hispanic students versus white and Asian students (relative to about 0.7σ). Both gaps are measured in student standard deviations.

white and Asian students are smaller: at 0.72σ in ELA and 0.63σ in Math, and these gains could be reduced by 6.5% and 9.7% per year. These changes are strikingly similar to those in Delgado (2022) where allocations are made to explicitly shrink racial gaps in math scores subject to not lowering average scores. Delgado (2022) finds a 0.068σ reduction in the racial gap with no change in average scores, but using a race blind policy our district-wide reallocations would shrink the gap by 0.064 and *raise* average test scores by 0.032σ.[12]

There are three additional points we want to highlight from this figure with implications for which gaps are effected. First, whereas both the within-school and district-wide reallocations could change the achievement gap, only the district-wide reallocations could meaningfully affect the racial achievement gap. This makes sense because there is more variance in racial composition across schools than within.

Second, it is interesting to note that the welfare weights that hold gaps constant vary a lot across allocations. For the within-school reallocations attaining similar gaps requires a weight on lower-scoring students between 40-43% for ELA and 52-53% for Math. On the other hand, the district-wide reallocations require much larger weights on lower-scoring students. For example, it takes 55% and 61% to shrink the achievement gaps in ELA and math, and even more to shrink the racial gaps: 64% and 72%. For context, this means that to control the racial-achievement gap in math, a social planner would have to forego 0.007σ in average gains.

Finally, although average-test-score maximizing reallocations ($\omega_L = \omega_H = 0.5$) within school tend to not affect either gap significantly,[13] district-wide reallocations to maximize test scores will actually expand both the achievement and racial achievement gaps. Intuitively this is because of cross-school co-variation in achievement (or race) and class size as discussed above.

[12]Note that in our context larger reductions in gains are obviously possible if the social planner is willing to choose allocations that actually reduce the average scores of certain groups while staying on the frontier. While it is likely that there are interior allocations in which gaps could be further reduced, we restrict our focus to allocations that are on the frontier of gains to higher- and lower-scoring students.

[13]In fact, if anything they would slightly shrink the achievement gap.

Reallocation winners and losers

As noted above, because there are so many students, no reallocation—even one creating large average gains—is a Pareto gain in the sense that it helps, or leaves unaffected, all students. Despite the net gains from matching teachers to their comparative advantages and putting stronger teachers in larger classes, reallocations will assign some students to less effective teachers or to teachers who are a worse match for them (despite the teacher being a better match for their class).

Before communicating these results, we want to highlight the fact that *any* allocation of teachers to students will assign some students better teachers than others. In that sense the "harms" presented here should be benchmarked by the fact that in the status quo roughly one third of students are assigned to a teacher with below-median value each year (among teachers teaching the relevant grade in the student's school), and for these students, the average "loss" (relative to the expectation) is about 0.10 student standard deviations in their scores on tests of each subject.

With that context in mind, Appendix Figure 2.14 shows that just as some students experience lower test score growth because of the year-to-year allocations of teachers in the status quo, some also receive lower value-added teachers in our reallocations. For example, the optimal within-school reallocations assign between 35-38% of students to lower value-added teachers, with 39-47% for the district-wide reallocations. Unsurprisingly, more egalitarian allocations reduce the achievement gains of higher-scoring students relative to the status quo whereas more elitist allocations reduce the gains to lower-scoring students. Appendix Figure 2.14 also reports the average achievement loss among students who are harmed. In the optimal district-wide (within-school) allocations, students who receive lower value-added teachers than they would in the status quo experience 0.104-0.120σ (0.085-0.099σ) smaller ELA testing gains on average and 0.173-0.204σ (0.140-0.165σ) smaller math gains on average, per year. While these figures sound large in terms of educational interventions, it's important to remember that they are relatively similar to the "losses" that are occurring in the status quo. Our reallocations change which students receive teachers with lower absolute advantage or poorly matched comparative advantage, but on average these changes are more than offset by even larger average gains to other observably similar students.

One implication of this depiction of winners and losers is that our reallocative policies have a strong redistributive component. For a social planner who only cares about higher- versus lower-scoring students this consideration is irrelevant, but in practice districts may want to preserve some horizontal equity.[14] For example, because our reallocations tend to put teachers with higher absolute advantage in larger classes and because larger classes tend to be in schools with more higher-scoring students, our optimal reallocations will tend to benefit lower-scoring students in these schools slightly more than lower-scoring students in schools with lower average achievement. As discussed in Section 2.2, this may be troubling if the policymaker has preferences over multiple dimensions of student characteristics. For example, this could be problematic if the policymaker is most concerned about lower-scoring students in schools with lower achievement.

The fact that there are indeed winners and losers among students, in addition to the observation that teachers, administrators, and teachers' unions—by revealed preference—weakly prefer the status quo to any reallocation raises the question of welfare implications from these reallocation policies. Can schools reallocate teachers in ways that matter for welfare? How could they make such reallocations incentive compatible for families and teachers? What would be the cost of smoothing such incentive compatibility constraints? And would the reallocation still be worth doing? These are questions we consider in the following section.

2.5. From value-added to Welfare Added

We have provided a welfare theory, estimated the relevant parameters, and demonstrated the test score gains from reallocations along a single subject. Our empirical findings so far can be interpreted as statements about a popular outcome of interest, test scores. With some assumptions, however, our findings on test score gains can be interpreted as an unbiased, or less biased than the mean, welfare estimate using our welfare theory.

First, we need to make an assumption about family preferences and their behavior in light of our policy change. We assume that families—the main decision-makers for students—value the

[14] At least relative to the status quo. In an obvious sense, the opportunity cost of the current allocation is that it harming (or at least not benefiting) many students that a different allocation could be making better off.

average achievement of the school they enroll in. This means that students will not re-sort to new schools after we have rearranged teachers within a school. This is obviously restrictive as parents may value many aspects of education, some idiosyncratic, like having a teacher an older sibling took classes, and others more systematic, like sociability and non-cognitive value-added (e.g., Beuermann, Jackson, Navarro-Sola, & Pardo, 2023; Jacob & Lefgren, 2007; Petek & Pope, in press). Nevertheless, the vast majority of families do not request specific teachers, and even when they do, not all requests are honored. This assumption is analogous to the "no spillovers" condition assumed in Section 2.2. Given extensive evidence that families do not respond to information about value-added in school choice (Abdulkadiroğlu, Pathak, Schellenberg, & Walters, 2020) or housing markets (Imberman & Lovenheim, 2016), we think this assumption is not too restrictive. Readers critical of this assumption should consider all welfare gains in partial equilibrium terms.

Second, we need to consider the bias terms from Theorem 2. First, consider the covariance term. It is important to remember that this term is dependent on the policymaker's welfare weights. As mentioned above, the covariance terms would be zero if our policymaker truly cared about only average lower- and higher-scoring students. If this is not the case, for a completely unbiased estimate, we need the conditional covariance of the true welfare weights (that consider all factors important to the policymaker) and student gains to be uncorrelated. We know that different allocations impact racial test score gaps and that gains from some reallocations accrue to lower-scoring students primarily in higher-scoring schools. While the estimates may not be unbiased in this case, satisfying Equation 2.2 would still ensure they are better than simple means. Conditioning on additional factors like race and school average scores could further assuage these concerns, but for tractability, we stick to conditioning on test scores.

Next, we consider the estimation bias between our estimated conditional average treatment effect and the truth. While we know teacher impacts differ along different dimensions (Delgado, 2022), we believe conditioning on test scores captures much of the variation without over-fitting. While race also plays a role, finding common support for all teachers can be practically challenging. Gender may play a role in teacher impacts as well; however, gender composition does not

change significantly between most classes, limiting the bias introduced by teacher heterogeneity.

There are still two significant shortcomings that we address in the following section. First, these teachers teach both ELA and Math, and so an optimal reallocation policy would consider the impact on both simultaneously. To combine both of these subjects into a single score function, we map achievement gains to lifetime earnings, which we do using the subject-specific estimates from Chetty et al. (2014a) of how value-added affects lifetime earnings.

The second shortcoming to address is the impact of reallocations on teachers. We need to consider the welfare component attributable to teachers' disutility from the reallocations. We treat teacher's preferences as an incentive compatibility constraint and assume they will need to be compensated enough to willingly switch classes. Using a revealed preference argument, if teachers willingly move, they will have been made better off. Assuming all teachers must be compensated for changing assignments will likely overstate the cost to teachers because at least some may prefer their new assignments,[15] the main challenge is how to price this disutility. Some papers have attempted to price the disutility to teachers from various policies (e.g., Bates et al., 2022; Rothstein, 2015), but highly structured wages in teacher labor markets often make this difficult in practice. We will focus on the marginal value of public funds (MVPF, Hendren & Sprung-Keyser, 2020) for a hypothetical universal bonus program.

Note that by restricting our focus on families and teachers in this way, we implicitly assume that other considerations like union concerns or the administrative costs of performing the reallocations are negligible. While these considerations are likely important, we argue that welfare gains of a large enough magnitude could allow transfers or interventions to alleviate these concerns or pay these costs.

2.5.1 Students: Earnings Implications of Reallocations

We begin with the welfare implications for students under the assumptions outlined above. These results are most closely tied to our previous analyses focused on student gains. This sub-

[15]For example, some teachers will be sent to schools they would like to teach at but cannot because of opening and union tenure requirement.

section demonstrates our approach for finding the optimal achievement gains for students' lifetime earnings and performing allocations that maximize those income gains.

Choosing an Income-Optimal Score Function

Because there are numerous allocations, all of which would generate different earnings outcomes, our first objective is choosing a welfare "score" function to maximize income. To do so we use the subject-specific estimates of the effects of value-added in Math or ELA on student earnings from Chetty et al. (2014a). They estimate that a one standard deviation increase in ELA scores in elementary school generates an additional $1,524 in earnings in early adulthood and that the corresponding gains in Math are $650.

Because of the fundamental trade-off between the facts that our reallocations generate larger gains in math, but gains to ELA matter more for earnings, we take a principled approach to defining the income-optimal allocation. We consider the following set of utilitarian score functions that take into account value-added in two subjects, s, ELA and Math.[16]

$$\tilde{\mathcal{W}}(\mathcal{J};\omega) = \frac{1}{N_{i,t}} \sum_{(i,t)} \sum_{s} \omega_s \left[L_{i,s,t}\, \hat{\tau}_{L,s}^{\mathcal{J}(i,t)} + (1 - L_{i,s,t})\, \hat{\tau}_{H,s}^{\mathcal{J}(i,t)} \right] \tag{2.9}$$

where ω_s represent the weight on each subject and $\sum_s \omega_s = 1$. And now $L_{i,s,t}$ indicates whether the student is low scoring in that particular subject.

Solving the optimization problem for a range of $\omega_{ELA} \in [0.0, 1.0]$ generates a production possibility frontier similar to those in the reallocation exercises in Section 2.4. Whereas the previous PPF plotted the trade-offs of possible gains between higher- and lower-scoring students, the PPF in Panel (a) of Figure 2.9 presents the trade-offs between gains to average Math and average ELA scores. For example, an allocation focused entirely on Math scores could raise average math scores by 0.058σ (0.016σ within schools). Because Math and ELA value-added are somewhat correlated, this allocation would also raise ELA scores by 0.019σ (0.005σ within schools). The focus

[16] We will soon relax the assumption about a utilitarian social planner.

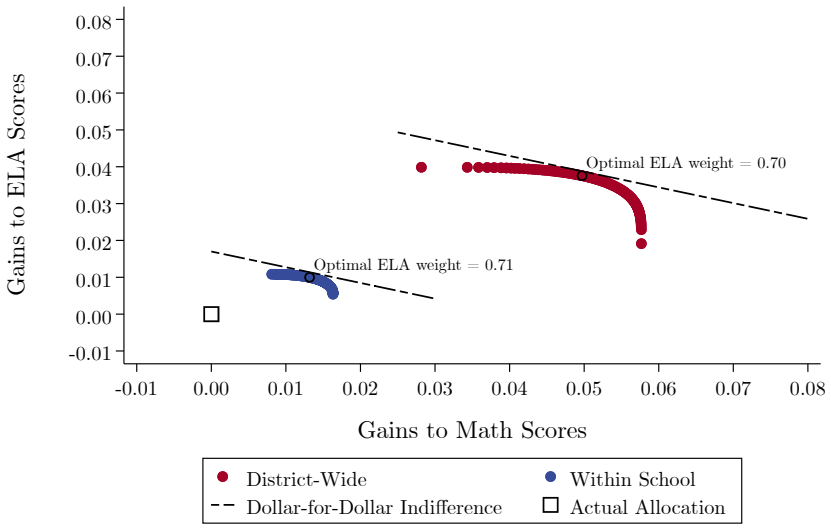
on math scores only, however, forgoes large ELA gains. This could be particularly problematic as ELA gains are nearly 2.5 times more important for earnings.

We combine the information on possible gains with the estimates of the subject-specific income effects of those gains to calculate the weight each subject that maximizes income gains. The estimates from Chetty et al. (2014a) create relative "prices" of gains to scores in each subject measured in earnings. As such, the income-maximizing weight sets the marginal rate of substitution between ELA and math scores equal to the relative price. We illustrate this graphically in Panel (a) of Figure 2.9 using a dashed line with a slope of the relative price. This line is tangent to the within-school PPF at $\omega_{ELA} = 0.71$ and to the district-wide PPF at $\omega_{ELA} = 0.70$. These values favor ELA gains, but do not focus exclusively on ELA value-added because the value of marginal gains to ELA scores from increasing ω_{ELA} beyond 0.71 are smaller than the value of the larger gains to increasing math scores.

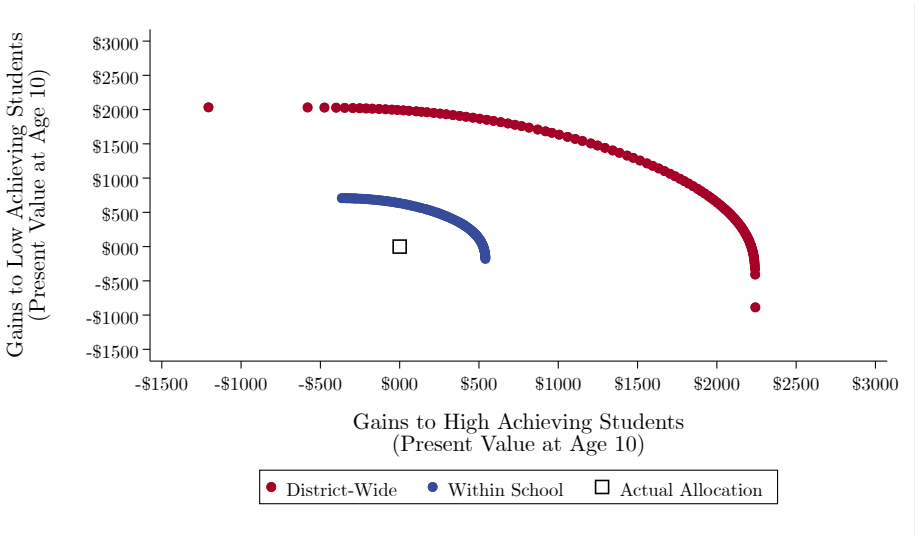
The combination of gains from both subjects significantly increases the income gains from students. The facts that math value-added scores have higher variance and result in larger achievement gains from reallocations might motivate a social planner to focus only on math scores in their objective function. In fact, this intuition plays out in the policy experiments considered in Delgado (2022) and Bates et al. (2022) which both focus only on math. Surprisingly, our results overturn this intuition. We will discuss the details of how we obtain these numbers below, but we find that a district-wide allocation that focuses only on math scores increases average present-valued earnings by \$1030. The insight that we can incorporating information about both math and ELA optimally generates gains of \$1390 per student. This \$360 (34%) gain is large and is costless once one allows the social planner to optimally weight value-added to both test scores.

Characterizing Possible Income Gains

With information about the income-optimal score function in hand, we return to the question of optimal policy with heterogeneous social preferences. Combining all of the pieces we



(a) Choosing the Wage-Maximizing Score Function



(b) Present Value Income Gains

Figure 2.9: **Multiple Outcomes Increase Achievement more in Reallocations**
Note: This figure shows how we combine math and ELA scores to estimate the frontier of possible earnings gains. Panel (a) displays the PPF of math versus ELA gains (assuming equal weights). The tangent lines are those implied by the subject-specific estimates of Chetty et al. (2014a). Panel (b) shows the implied effect on lifetime earnings from reallocations with a score of S =0.75 ELA + 0.25 Math (present valued at age 10).

define a new social welfare function to optimize

$$\tilde{\mathcal{W}}(\mathcal{J};\omega) = \frac{1}{N_{i,t}} \sum_{(i,t)} \omega_L \left[\omega_{\text{ELA}}\, L_{i,\text{ELA},t}\, \hat{\tau}^{\mathcal{J}(i,t)}_{L,\text{ELA}} + (1-\omega_{\text{ELA}}) L_{i,\text{Math},t}\, \hat{\tau}^{\mathcal{J}(i,t)}_{L,\text{Math}} \right]$$
$$+(1-\omega_L) \left[\omega_{\text{ELA}}\, (1-L_{i,\text{ELA},t})\, \hat{\tau}^{\mathcal{J}(i,t)}_{H,\text{ELA}} + (1-\omega_{\text{ELA}})\, (1-L_{i,\text{Math},t})\, \hat{\tau}^{\mathcal{J}(i,t)}_{H,\text{Math}} \right]$$

where now we explicitly sum test score gains over both subjects and both student types with their respective weights. Because this formulation exponentially increases the dimensionality of ω, we use our evidence about income-optimal weights to choose $\omega_{\text{ELA}} = 0.75$ and $\omega_{\text{Math}} = 0.25$ in this section. To the extent to which the optimal ω^*_{ELA} varies over ω_L, our results provide a lower bound on the true earnings gains.[17]

After calculating the efficient allocations for each ω, we use the process in Chetty et al. (2014a) to map the test score improvements into the present value of lifetime earnings. We outline our approach as follows. First, we assume that individuals may choose to work between the ages 20 and 65. We also assume that the average income gains implied from test scores apply to all of these earning. Finally, we assume that families discount these earnings gains at a 3% (i.e., with a 5 percent discount rate partially offset by 2 percent wage growth) back to age 10, the average age of students in our sample. Empirically this implies a multiplier of 15.5 on the baseline gains implied from test scores.

The results, depicted in Panel (b) of Figure 2.9 show that optimally reallocating teachers could create millions of dollars of gains per year. Based on our calculation, the income-maximizing district-wide allocation would generate over $1140 in present valued earnings for low scoring students and over $1630 for high-scoring students. Since there are 10,150 students of each type each year (on average), this implies the value of the reallocation across all students is $27.9 million. While smaller, the gains from the within school reallocations are not insignificant: over $400 for

[17]Note that because not all students are low scoring in Math and ELA the achievement weight ω_L may not apply uniformly to each student. In practice this means that there are four implicit weights generated by this welfare function. One conceptually simple way to think of this function is treating each student's score as a different student and then weighting the welfare from gains to that "student" by both their achievement and which test it is.

lower-scoring students and over $300 for high-scoring students, implying $7.4 million across the district.

Policy makers concerned about inequality can also create large redistributive gains. For example in the district-wide reallocation, a social planner could increase the present value of lower-scoring students' earnings by $1990 without hurting high scoring students on average. A similar comparison reveals gains of $600 from within school reallocations. Compounded year-over year gains like these could be powerful tools at reducing not only achievement, but also earnings inequality among students coming out of the district. In Appendix Figure 2.15, we compare these results to those of a social planner with continuous CES preferences across students rather than discrete preferences across groups and show similar patterns.

Taken together the gains from this policy are enormous. Even if the 27.9 million dollar gain is infeasible because of teacher or union preferences, the within-school reallocation is an essentially costless program generating nearly quarter of those gains. This underscores the power of using information about comparative advantage to improve policy. Furthermore, if there are ways to make the 27.9 million dollar gains attainable, a discussion of how to do so is of first-order importance. The following subsection provides that discussion.

2.5.2 Teachers: Welfare Value of a Teacher Bonus Program

We now turn to the welfare implications for teachers. Rather than trying to price teacher disutility, we focus on a teacher bonus thought experiment. One advantage of considering this experiment is that it allows us to separately consider welfare and incentive compatibility. Our estimates reflect the welfare attainable for each policy and would allow policymakers to choose the optimal one based on their understanding of the incentive constraints (e.g., teacher supply, wages, amenities, seniority, unions, etc.).

Imagine a policy that paid all teachers a certain bonus for participating in a reallocation. Teachers would be paid this bonus whether or not their school or class assignment changed. If the bonus was sufficient to ensure incentive compatibility, then one way to characterize the wel-

fare under the resulting allocation would be the marginal value of public funds (MVPF, Hendren & Sprung-Keyser, 2020). This characterizes a lower bound on an envelope of possible incentive programs that could be improved by targeting bonuses the teachers with the highest impacts from reallocation or by relaxing the requirement to participate in the reallocation (for example, for teachers with very strong preferences to their current assignment.

The MVPF is a "bang-for-the-buck" measure of the bonus program, calculated as the present value of the total program benefits divided by the net cost of implementing it. Specifically, for a bonus of size b the MVPF of allocation j is

$$MVPF^j(b) = \frac{\sum_i (1-t)\Delta S_i^p)}{N_j b - t\Delta S_)^p} \tag{2.10}$$

where $(1-t)\Delta S_i^p$ are the after-tax present-value monetary gains to each student from allocation j (given tax rate t), N_j is the number of teachers and $t\Delta S_i^p$ is the present-value of gains recouped as tax revenue. The key assumption required for this statistic to be meaningful in this policy thought experiment is internalizing the fiscal externality of the district's policy. For example, this could be interpreted as the national value of the district administering the reallocation policy. Although it is possible to compare national and local MVPFs (e.g., see Agrawal, Hoyt, & Ly, 2023), we focus on this simplified case as in other work (Hendren & Sprung-Keyser, 2020).(Hendren & Sprung-Keyser, 2020).[18]

We combine our estimates of present-value monetary gains with data from the Opportunity Atlas (Chetty, Friedman, Hendren, Jones, & Porter, 2018) to calculate these MVPF empirically. For the changes in earnings, we focus on the utilitarian, earnings-maximizing, within-school and district-wide reallocations as described in the previous subsection. To compute the tax rate, we note that for children growing up in San Diego county, the median income at age 35 is \$43,000. Because the majority of these individuals are unmarried (56%) and still living in the same commuting zone (68%), we apply the marginal tax rates from the United States and California for single filers, 0.22

[18]Note that the two could be equivalent if the state and federal governments were to transfer the marginal tax revenue generated by the policy back to the SDUSD.

and 0.06, implying $t = 0.28$ for in equation 2.10.

We present the results in Figure 2.10. Figure 2.10 plots the Marginal Value of Public Funds over a broad support of possible bonus sizes (using a log scale on the x-axis). The two series represent the MVPF of a bonus program of a given size for the district-wide or within-school reallocations. The curve showing the value of bonuses for the within-school reallocations is lower because those reallocations produce smaller gains. For each point, the MVPF can be interpreted as dollars of social benefit produced for each dollar spend on the teacher bonus program. Values of the MVPF above 5 are reported at the same height on the y-axis.

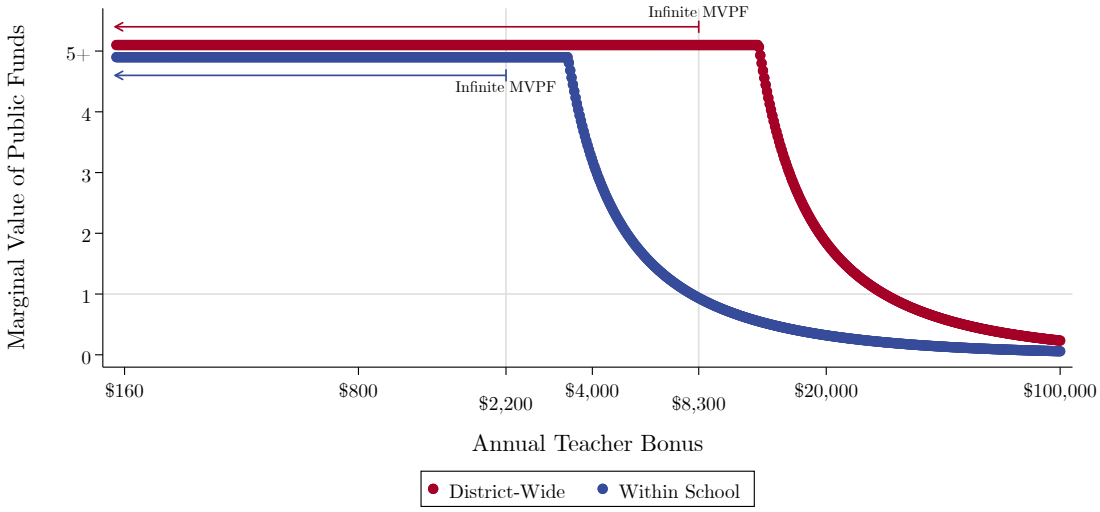


Figure 2.10: **Compensating Teachers for Reallocations Could Have Enormous Welfare Impacts**

Note: This figure shows the marginal value of public funds for teacher bonus programs of different sizes (for either the within-school or district-wide reallocation). Values are capped at 5 on the figure, the range for which the MVPF is infinite is indicated with arrows, and the x-axis shown on a log scale.

The main takeaway from Figure 2.10 is that for a broad range of bonus sizes the policy of reallocations and bonuses has an infinite MVPF. An infinite MVPF occurs when the net cost of the program is negative and the benefits are positive. in other words, the district would be *making* money by paying to reassign teachers—and would be increasing student earnings in the process. For the district-wide reallocation, the MVPF is infinite for a bonus of up to $8,300, and

it is infinite for bonuses up to $2,200 for the within-school reallocation. This second number is particularly striking because despite being noninvasive the within-school reallocation is still generating substantial gains.

A second important insight from Figure 2.10 is that even when the MVPF is not infinite it is still large even for very costly bonus programs. For example, for the district-wide reallocation, a bonus program of paying *every teacher* in the district $20,000 to participate in the reallocation would still have an MVPF of roughly 2. In other words, it would generate $2 of present valued earnings gains for every dollar spent on bonuses. This is a marked pay increase – equivalent to a one-third salary increase for a teacher in the 2010-11 school year with 10 years of teaching experience and the middle tier of education in the district's collective bargaining agreement.

Note that some of these bonus policies may not be incentive compatible, but other research suggests that reallocations with large and even infinite gains could be attainable. For example, while $20,000 may sounds enormous, it amount was shown to be more than enough inducing teachers to move to very low performing schools in a large randomized controlled trial (Glazerman, Protik, Teh, Bruch, & Max, 2013). On the other hand, it's likely that almost all of the within-school reallocations are incentive compatible for most bonuses. First this is because teachers seem to care much more about which school they teach at than which class they teach—in large part because of commuting (Bates et al., 2022)—and this is not affected in the within-school reallocation. Furthermore, in the within-school reallocation most teachers do not even switch classes, suggesting that the utility impact of the reallocation would be particularly small.

Taken together the teacher bonus thought experiment suggests that the large gains from reallocations are more than an impossibility. Although some teachers would be worse off because of certain reallocations, generating structures that appropriately compensate them for teaching to their comparative advantage could generate tremendous gains. In fact, many of the policies we explore generate large enough earnings gains to students to justify lavish teacher bonuses on the grounds of added tax revenue alone.

2.6. Conclusion and Implications for Policy

This paper set out to answer two questions: When does heterogeneity matter for maximizing a social objective in general? And how large are the welfare gains from using heterogeneous estimates for refining education policy in particular? We employed and extended tools from public finance to think about aggregating teacher effects on multidimensional outcomes and heterogeneous student types into welfare relevant statistics and implemented them in the context of a large urban school district. In reallocation exercises, using information about both multidimensionality and heterogeneity produce up to double the gains for test scores or for later-life outcomes relative to using standard measures that assume teachers have homogeneous impacts on students, and which focuses on one student outcome rather than two. This highlights the importance of incorporating such information into welfare considerations and policy.

We conclude by exploring three policy trade-offs that our results highlight and discussing possible directions for continued inquiry.

In the specific context of education value-added, our results highlight the power of comparative advantage relative to other policy proposals. Historically researchers have benchmarked the importance of teacher value-added with the a policy "deselecting" (i.e., firing) low-performing teachers (see Chetty et al., 2014b; Delgado, 2022; Hanushek, 2009; 2011). Although deselecting 5% of teachers with the lowest value-added could produce large gains, there are concerns about the ethics of mistakes (Staiger & Rockoff, 2010) and the implications for teacher labor markets (Rothstein, 2015), in the sense that it is not obvious who the replacement teachers will be, and their own teaching effectiveness. An interesting implication of our results, however, is that by relaxing the traditional assumptions of constant effects and equal class sizes we can reallocate rather than release teachers. In our setting a district-wide reallocation would produce gains more than three times larger than the gains from deselecting 5% of teachers. Furthermore, because deselection using standard value-added penalizes teachers who happen to be allocated to worse-matched classes, reallocations prevent incorrect dismissals—16-19% of those targeted. A reallocation-based policy

would be less costly to teachers and more beneficial. A within-school reallocation would be even less costly and would still generate 50% of the gains from deselection. In other words, our results suggest that in some, and perhaps many, cases, teachers in the bottom 5-10% need not be deselected, but rather provided an assignment that better matches their comparative advantage. In other cases, where absolute advantage is extremely low, deselection could still be an option.

A second, more general, policy-insight is that our theory can show policymakers how mean evaluations of existing policies may (or may not) apply to new policy considerations. For example, we show that mean-based welfare estimates can be biased when based on estimates that are not externally valid, or when there is a covariance between welfare weights and treatment effects. While our results clearly indicate the value of considering heterogeneity, even without information beyond the means, policymakers can use these conditions to assess the severity of the bias. For example, using estimates from an expansion of Medicaid to beneficiaries similar to those who are eligible in another state may be very reasonable, whereas assuming that both welfare weights and the elasticity of taxable income are homogeneous along the income distribution may not be. Furthermore, policy can be further improved by conditioning on the relevant dimensions of heterogeneity. Admittedly, using characteristics to condition the estimates often reduces precision—although this type of tradeoff between bias and variability is hardly unique to our setting.

A final policy consideration can be taken from our results at large. Since value-added and other mean evaluations are useful in so many contexts, we hope many practitioners will extend the use of heterogeneous estimates. As they do our research can provide a framework for the gains from adding heterogeneity and which dimensions of heterogeneity and multidimensionality to add and which to ignore. While our results highlight striking patterns in how value-added heterogeneity specifically may affect the long-term outcomes of students, we note that assessing the optimality of reallocation policies in the long run will depend on heterogeneity in the long-term effects. We think an important next step in this literature is directly assessing the effect of multi-dimensional measures of teacher quality on various life-long outcomes and particular the heterogeneity in these relationships across groups.

Taking a step back, our results also highlight the value of testing for and estimating heterogeneous estimates of teacher impacts, and of causal effects more broadly. Whether it is allocating teachers to classes, assessing racial health disparities in care, comparing possible social services, or measuring the effects of firms on earnings growth, the mean is rarely enough to characterize the full question of interest. Although estimating and implementing these evaluations can be costly, researchers have their own comparative advantage in such analyses, and our results suggest enormous gains from finding ways to leverage that knowledge to improve allocation in public programs of many types.

Chapter 2, in part, is currently being prepared for submission for publication of the material and is coauthored with Ricks, Michael; Mather, Nathan; and Betts, Julian. The dissertation author was the primary researcher and author of this material.

2.7 References

Abdulkadiroğlu, A., Pathak, P. A., Schellenberg, J., & Walters, C. R. (2020). Do Parents Value School Effectiveness? *American Economic Review*, *110*(5), 1502–39.

Abrams, D. S., & Yoon, A. H. (2007). The Luck of the Draw: Using Random Case Assignment to Investigate Attorney Ability. *University of Chicago Law Review*, *74*, 1145.

Agrawal, D., Hoyt, W., & Ly, T. (2023). *A New Approach to Evaluating the Welfare Effects of Decentralized Policies* (Working Paper).

Alatas, V., Purnamasari, R., Wai-Poi, M., Banerjee, A., Olken, B. A., & Hanna, R. (2016). Self-targeting: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy*, *124*(2), 371–427.

Angrist, J., Hull, P., & Walters, C. R. (2022). Methods for Measuring School Effectiveness.

Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2019). *The Surrogate Index: Combining Short-term Proxies to Estimate Long-term Treatment Effects more Rapidly and Precisely* (Tech. Rep.). National Bureau of Economic Research.

Athey, S., & Wager, S. (2021). Policy Learning with Observational Data. *Econometrica*, *89*(1), 133–161.

Bacher-Hicks, A., & Koedel, C. (2022). *Estimation and Interpretation of Teacher Value Added in Research Applications* (Working Paper).

Bates, M. D., Dinerstein, M., Johnston, A. C., & Sorkin, I. (2022). *Teacher Labor Market Equilibrium and Student Achievement* (Tech. Rep.). National Bureau of Economic Research.

Betts, J. R. (2011). The Economics of Tracking in Education. In *Handbook of the Economics of Education* (Vol. 3, pp. 341–381). Elsevier.

Beuermann, D. W., Jackson, C. K., Navarro-Sola, L., & Pardo, F. (2023). What is a Good School, and Can Parents Tell? Evidence on the Multidimensionality of School Output. *The Review of Economic Studies*, *90*(1), 65–101.

Bhatt, M. P., Heller, S. B., Kapustin, M., Bertrand, M., & Blattman, C. (2023). *Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago* (Tech. Rep.). National Bureau of Economic Research.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2005a). The Draw of Home: How Teachers' Preferences for Proximity Disadvantage Urban Schools. *Journal of Policy Analysis And Management*, *24*(1), 113–132.

Chan, D. C., Gentzkow, M., & Yu, C. (2022). Selection with Variation in Diagnostic Skill: Evidence from Radiologists. *The Quarterly Journal of Economics*, *137*(2), 729–783.

Chandra, A., Finkelstein, A., Sacarny, A., & Syverson, C. (2016). Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector. *American Economic Review*, *106*(8), 2110–2144.

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R., & Porter, S. R. (2018). The Opportunity Atlas. *Opportunity Insights*.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, *104*(9), 2593–2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–79.

Condie, S., Lefgren, L., & Sims, D. (2014). Teacher Heterogeneity, Value-Added and Education Policy. *Economics of Education Review*, *40*, 76–92.

Dahlstrand, A. (2022). *Defying Distance? The Provision of Services in the Digital Age* (Tech. Rep.).

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

Delgado, W. (2022). Heterogeneous Teacher Effects, Comparative Advantage, and Match Quality: Evidence from Chicago Public Schools.

Delhommer, S. (2019). *High School Role Models and Minority College Achievement* (Tech. Rep.).

DeNegre, S. T., & Ralphs, T. K. (2009). A Branch-and-Cut Algorithm for Integer Bilevel Linear Programs. In *Operations Research and Cyber-Infrastructure* (pp. 65–78).

Diamond, P. A. (1973). Consumption Externalities and Imperfect Corrective Pricing. *The Bell Journal of Economics and Management Science*, 526–538.

Doyle, J., Graves, J., & Gruber, J. (2019). Evaluating Measures of Hospital Quality: Evidence from Ambulance Referral Patterns. *Review of Economics and Statistics*, *101*(5), 841–852.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American economic review*, *101*(5), 1739–1774.

Einav, L., Finkelstein, A., & Mahoney, N. (2022). *Producing Health: Measuring Value Added of Nursing Homes* (Tech. Rep.). National Bureau of Economic Research.

Fell, H., Kaffine, D. T., & Novan, K. (2021). Emissions, Transmission, and the Environmental Value of Renewable Energy. *American Economic Journal: Economic Policy*, *13*(2), 241–72.

Finkelstein, A., & Notowidigdo, M. J. (2019). Take-up and Targeting: Experimental Evidence from SNAP. *The Quarterly Journal of Economics*, *134*(3), 1505–1556.

Glazerman, S., Protik, A., Teh, B.-r., Bruch, J., & Max, J. (2013). *Transfer Incentives for High-Performing Teachers: Final Results from a Multi-site Randomized Experiment* (Tech. Rep.). U.S. Department of Education.

Griffith, R., O'Connell, M., & Smith, K. (2019). Tax Design in the Alcohol Market. *Journal of Public Economics*, *172*, 20–35.

Hanushek, E. A. (2009). Teacher Deselection. *Creating a New Teaching Profession*, *168*, 172–173.

Hanushek, E. A. (2011). The Economic Value of Higher Teacher Quality. *Economics of Education Review*, *30*(3), 466–479.

Harrington, E., & Shaffer, H. (2023). *Estimating Prosecutor Effects on Incarceration and Reoffense* (Tech. Rep.). Working Paper.

Hendren, N., & Sprung-Keyser, B. (2020). A Unified Welfare Analysis of Government Policies. , *135*(3), 1209-1318. Retrieved from `https://academic.oup.com/qje/artjournal={QuarterlyJournalofEconomics}` doi: 10.1257/jep.34.4.146

Hollingsworth, A., & Rudik, I. (2019). External Impacts of Local Energy Policy: The Case of Renewable Portfolio Standards. *Journal of the Association of Environmental and Resource Economists*, *6*(1), 187–213.

Hull, P. (2020). *Estimating Hospital Quality with Quasi-Experimental Data* (Tech. Rep.). Working Paper.

Hussam, R., Rigol, N., & Roth, B. N. (2022). Targeting High Ability Entrepreneurs Using Community Information: Mechanism Design in the Field. *American Economic Review*, *112*(3), 861–98.

Ida, T., Ishihara, T., Ito, K., Kido, D., Kitagawa, T., Sakaguchi, S., & Sasaki, S. (2022). *Choosing Who Chooses: Selection-Driven Targeting in Energy Rebate Programs* (Tech. Rep.). National Bureau of Economic Research.

Imberman, S. A., & Lovenheim, M. F. (2016). Does the Market Value Value-Added? Evidence from Housing Prices after a Public Release of School and Teacher Value-Added. *Journal of Urban Economics*, *91*, 104–121.

Ito, K., Ida, T., & Tanaka, M. (2021). *Selection on Welfare Gains: Experimental Evidence from Electricity Plan Choice* (Tech. Rep.). National Bureau of Economic Research.

Jackson, C. K. (2018). What do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. *Journal of Political Economy*, *126*(5), 2072–2107.

Jacob, B. A., & Lefgren, L. (2007). What do Parents Value in Education? An Empirical Investigation of Parents' Revealed Preferences for Teachers. *The Quarterly Journal of Economics*, *122*(4), 1603–1637.

Johnson, A. C. (2021). *Preferences, Selection, and the Structure of Teacher Pay* (Working Paper).

Kitagawa, T., & Tetenov, A. (2018). Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, *86*(2), 591–616.

Krueger, A. B. (1999). Experimental Estimates of Education Production Functions. *The Quarterly Journal of Economics*, *114*(2), 497–532.

Norris, S. (2019). Examiner Inconsistency: Evidence from Refugee Appeals. *University of Chicago, Becker Friedman Institute for Economics Working Paper*(2018-75).

Petek, N., & Pope, N. G. (in press). The Multidimensional Impact of Teachers on Students. *Journal of Political Economy*.

Ricks, M. D. (2022). *Strategic Selection Around Kindergarten Recommendations* (Tech. Rep.).

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214.

Rothstein, J. (2015). Teacher Quality Policy when Supply Matters. *American Economic Review*, *105*(1), 100–130.

Sexton, S., Kirkpatrick, A. J., Harris, R. I., & Muller, N. Z. (2021). Heterogeneous Solar Capacity Benefits, Appropriability, and the Costs of Suboptimal Siting. *Journal of the Association of Environmental and Resource Economists*, *8*(6), 1209–1244.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, *24*(3), 97–118.

Stepner, M. (2013, October). *VAM: Stata module to compute teacher value-added measures.* Statistical Software Components, Boston College Department of Economics. Retrieved from `https://ideas.repec.org/c/boc/bocode/s457711.html`

Tomlinson, C. A. (2017). *How to Differentiate Instruction in Academically Diverse Classrooms* (Third ed.). ASCD.

Chapter 2 Appendix

2.A.1. Additional Tables and Figures

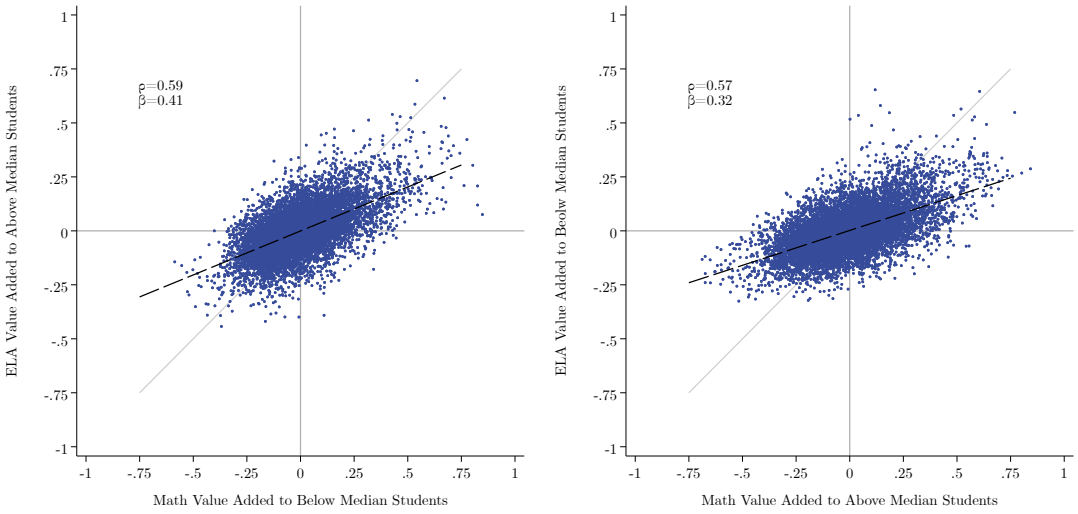


Figure 2.11: **Cross-Subject and Cross-Type value-added Is Much Less Correlated**
Note: This figure shows our heterogeneous estimates of teacher value-added on both English Language Arts (ELA) and Math test scores. Note that in this Figure Math and ELA scores are plotted against each other. Each dot represents one teacher-year estimate of value-added on higher- and lower-scoring students. The correlation coefficients is for the entire population stacked by year. The dotted line shows the line of best fit with the slope reported. For reference a line with slope one is plotted in the background.

Table 2.1: **The Standard Deviation of Class Size and the Share of Students in the Class Who Are High-Scoring in ELA and Math**

AFTER CONTROL FOR:	STD. DEVIATION CLASS SIZE	STD. DEVIATION SHARE OF CLASS ABOVE MEDIAN, ELA SCORES	STD. DEVIATION SHARE OF CLASS ABOVE MEDIAN, MATH SCORES
Grade*Year	3.68	0.50	0.50
School*Grade*Year	1.71	0.46	0.46

Note: This figure shows the within year-grade standard deviations in class size and composition at a district-wide level and a within-school level.

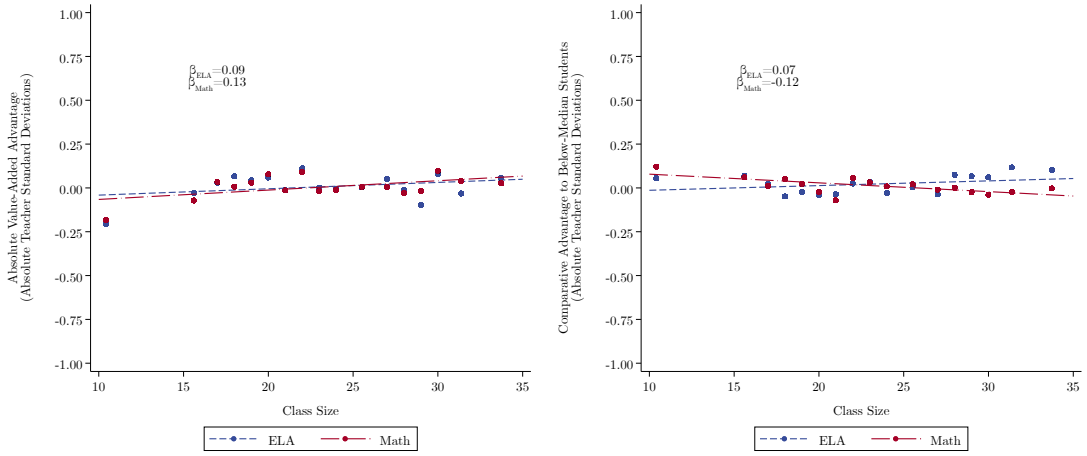


Figure 2.12: **Value-added Only Varies Somewhat Across Class Sizes**
Note: This figure shows how our heterogeneous estimates of teacher value-added on both English Language Arts (ELA) and Math test scores relate to class composition. The panel on the left shows teacher absolute advantage (average of value-added on higher- and lower-scoring students) and the panel on the right shows the comparative advantage (difference of value-added on below-median students minus value-added on higher-scoring students). both panels plot the ventiles of value-added (measured in teacher standard deviations in absolute advantage) over the share of number of students in each class. Both β report the change from a 25-student change in class size.

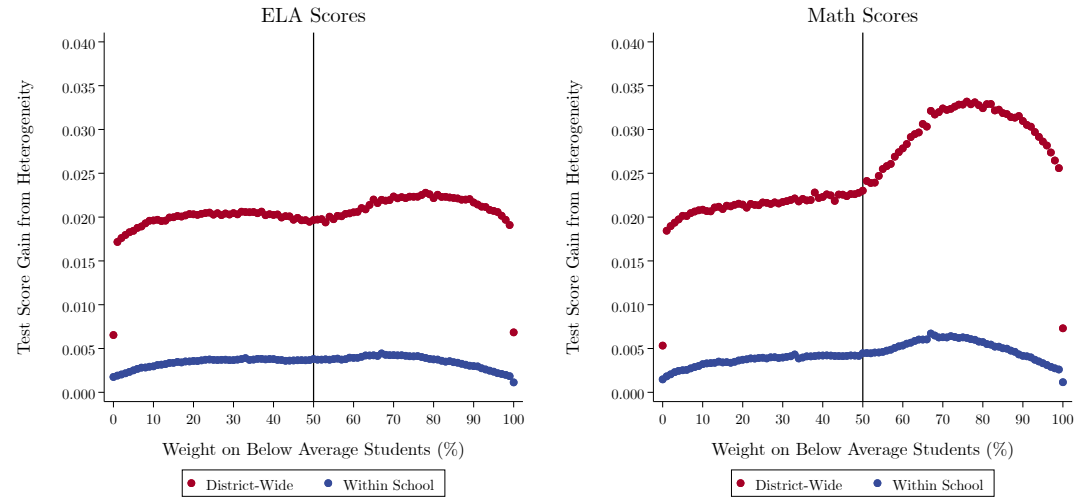
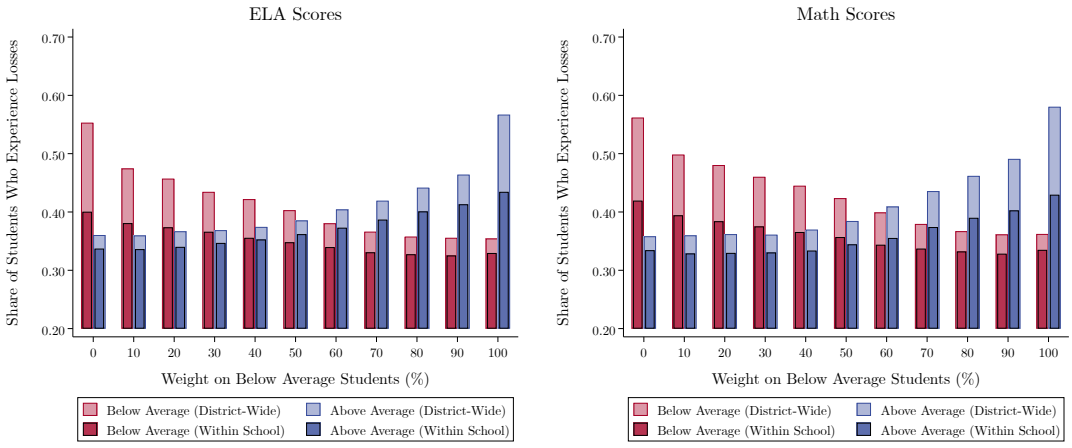
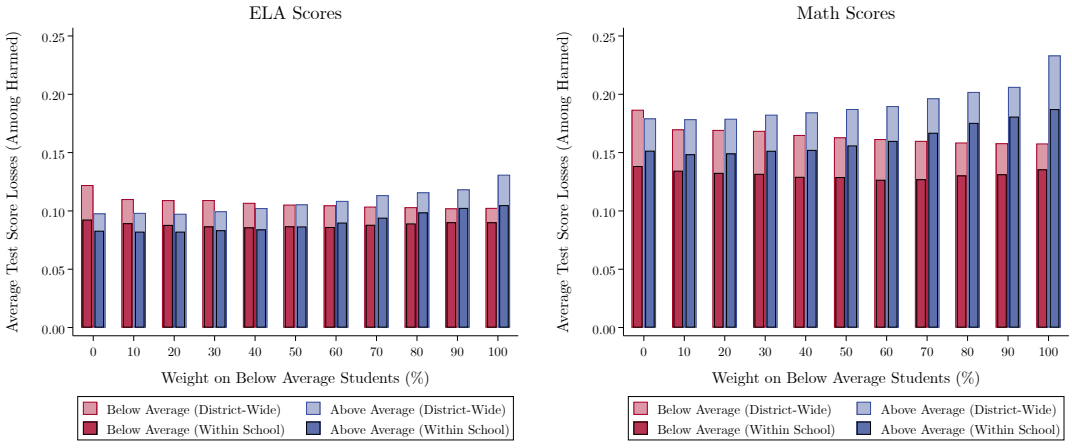


Figure 2.13: **Test-Score Gains from Using Heterogeneity**
Note: This figure shows the test scores gains from using our measures of heterogeneous value-added to make allocations relative to standard measures over various social preferences.



(a) Share of Students Harmed



(b) Mean Score Change among Harmed Students

Figure 2.14: **While Reallocations Help Many Students, They Will Harm Others**
Note: This figure shows information about which students are made worse off by the reallocations. Panel (a) reports the share of students whose scores would be lowered by each reallocation and Panel (b) reports the average change in scores among those harmed.

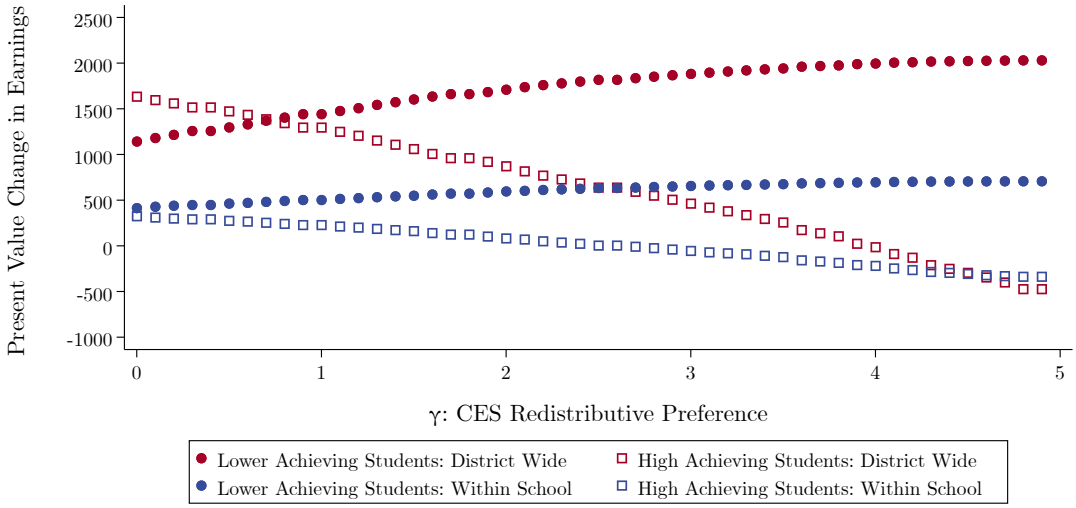


Figure 2.15: **Comparing to a CES Benchmark**
Note: This figure shows the present-value earnings gains from optimal reallocations based off of continuous CES preferences over student types rather than discrete preferences between higher- and lower-scoring students.

2.A.2. Theory Appendix

2.A.2.1 From Test Scores to Welfare Details

Below is a more detailed version of definition 2.1

Proof. If a change in an individual's outcomes $\boldsymbol{Y}_i$ only impacts the utility and welfare weights of that individual i, then for a given score function S, the expected change in welfare $\Delta\tilde{\mathcal{W}}^j$ from the status quo policy ($j = 0$) to policy j is

$$\begin{aligned}
\Delta\tilde{\mathcal{W}}^j &\equiv \mathbb{E}[\mathcal{W}^j|\boldsymbol{S}^j] - \mathbb{E}[\mathcal{W}^0|\boldsymbol{S}^0] \\
&= \sum_{i=1}^{n} \mathbb{E}[\psi_i^j U_i^j|S_i^j] - \mathbb{E}[\psi_i^0 U_i^0|S_i^0] \\
&= \sum_{i=1}^{n} \frac{\mathbb{E}[\psi_i^j U_i^j|S_i^j] - \mathbb{E}[\psi_i^0 U_i^0|S_i^0]}{\Delta S_i^p} \Delta S_i^p \\
&\equiv \sum_{i=1}^{n} \gamma_i(S_i^j, S_i^0)\Delta S_i^p
\end{aligned}$$

The last line is simply redefining the first term as a test score welfare weight $\gamma_i(S_i^j, S_i^0)$. $\boldsymbol{S}^j$ is the vector of test scores for every student under policy j. This means the expectations on the first line are conditional on the entire vector of test scores. This means the relationship between test scores and utility is fully flexible, and each student's utility can be uniquely impacted by a given test score change. Note that γ_i is an average over test score points for a given student, not an average across students. To understand this term, it is helpful to think through a simple example. Suppose $\mathbb{E}[\psi_i^j U_i^j | S_i^j] = S_{it}$ for all students. That is, expected welfare is linear in test scores. In this case, $\gamma_i(S_i^j, S_i^0) = 1$ because all students gain 1 util per score over the entire range of scores, and test scores are equivalent to welfare. Although welfare weights are often based off of earnings or earnings ability, the implication of definition 2.1 is that we can theoretically apply weights to a short term outcomes like test scores, rather than utility, and still have an unbiased estimate of welfare. Of course, in practice, getting individual weights is likely impossible. The later theory sections address the best way to overcome this problem with conditional aggregation, but definition 2.1 provides a ground truth reference that incorporates a large amount of of potential heterogeneity, individual differences.

2.A.2.2 Welfare Weighting the ATE

Using a similar approach to Hendren and Sprung-Keyser (2020), the following equation shows how it is possible to estimate welfare from an average treatment effect if the proper weight is applied

$$\Delta \mathcal{W}^j \tag{11}$$

$$= \int_0^1 \gamma_i(S_i^j, S_i^0) \Delta S_i^p \mathrm{d}i \tag{12}$$

$$= \frac{\int_0^1 \gamma_i(S_i^j, S_i^0) \Delta S_i^p \mathrm{d}i}{\int_0^1 \Delta S_i^p \mathrm{d}i} \int_0^1 \Delta S_i^p \mathrm{d}i \tag{13}$$

$$= \tilde{\gamma}^j ATE^j \tag{14}$$

The trouble is that the first term, $\tilde{\gamma}^j$ depends, not just on the test score welfare weights γ_i, but also on the joint distribution of those weights with the changes in test scores for policy j. It is a complex object that involves a deep understanding of the distribution of heterogeneous impacts resulting from policy j. If a policymaker already has this deep knowledge, it is not clear how much giving them the average treatment effect will help.

2.A.2.3 Theorem 1 proof

Proof.

$$\begin{aligned}
\textbf{Average Bias}_{ATE} &= \frac{\Delta\tilde{\mathcal{W}}^j}{n} - \mathbb{E}[\gamma^p]\widehat{ATE} \\
&= \frac{1}{n}\sum_{i=1}^{n}\gamma_i(S_i^j, S_i^0)\Delta S_i^p - \mathbb{E}[\gamma^p]\widehat{ATE} \\
&= \mathbb{E}[\gamma^p\Delta S^p] - E[\gamma^p]\widehat{ATE} \\
&= \mathbb{E}[\gamma^p]\mathbb{E}[\Delta S^p] + \mathrm{Cov}(\gamma^p, \Delta S^p) - E[\gamma^p]\widehat{ATE} \\
&= \mathrm{Cov}(\gamma^p, \Delta S^p) + \mathbb{E}[\gamma^p]\left(\mathbb{E}[\Delta S^p] - \widehat{ATE}\right)
\end{aligned}$$

The first line is how we are defining bias. It is the benchmark with individual heterogeneity minus our common estimator of the mean welfare weight and the average treatment effect. The second line comes from definition 2.1. The third line comes from recognizing that the first term in line two is the population average, or expectation, of $\gamma^p\Delta S^p$. The fourth line uses the general definition of covariance, that is $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[y]$. The last line just rearanges the terms.

2.A.2.4 Averate Treatment Effect Bias Explained

The specific source of average treatment effect bias we are consider can be a concern for any policy j that involves assigning specific sub-treatments d (teachers) to subsets of the population of size K_d^j (classes). First note that the average treatment effect is the following weighted average of sub-treatment effects ATE_d^j

$$ATE^j = \frac{1}{n}\sum_d K_d^j ATE_d^j$$

The bias comes in from incorrect estimates of the average sub-treatment effect (teacher impact) ATE_d^j characterized by the following

$$ATE_d^j - \widehat{ATE}_d^j = \frac{1}{K_d^j}\sum_{i=1}^{K_d^j} \Delta S_i^d - \frac{1}{K_d^0}\sum_{l=1}^{K_d^0} \Delta S_l^d$$

Here we can see the bias comes from different individual impacts between the existing class and the class in the policy counterfactual. It is helpful to think through the two cases where this difference goes to zero. First, if there is no treatment effect heterogeneity. For example, a teacher impacts all students equally on average and so $\Delta S_i^d = \Delta S_l^d \quad \forall \quad i, l$. Second, even if there is treatment effect heterogeneity, if the classes have similar characteristics the means may still be the same. For example, a teacher may be very bad at teaching English language learners (ELA). However, if both classes have the same fraction of ELA students, the teacher's mean impact will be the same.

2.A.2.5 Conditional Average Treatment Effect Bias Explained

The bias in the second term will be lower after conditioning when

$$\mathbb{E}[\Delta S^p] - \widehat{ATE} > \sum_x P_x \left(\mathbb{E}[\Delta S^p|x] - C\widehat{ATE(X)}\right) \tag{15}$$

As in the previous section, we can zero in on a specific teacher or sub-treatment and see that, for a given teacher, conditioning reduces bias when

$$ATE_d^j - \widehat{ATE}_d^j \tag{16}$$

$$= \frac{1}{K_d^j}\sum_{i=1}^{K_d^j}\Delta S_i^d - \frac{1}{K_d^0}\sum_{l=1}^{K_d^0}\Delta S_l^d \tag{17}$$

$$> \sum_X P_{dx}^j \left(\frac{1}{K_{dx}^j}\sum_{i=1}^{K_{dx}^j}\Delta S_i^d - \frac{1}{K_{dx}^0}\sum_{l=1}^{K_{dx}^0}\Delta S_l^d \right) \tag{18}$$

$$\sum_X P_{dx}^j \left(\widehat{ATE}_{dx}^j - \widehat{ATE}_{dx}^0 \right) \tag{19}$$

The left side is the difference in mean treatment effects between the baseline class and the counterfactual class, as described above. The right hand side is the difference in the mean treatment effects for a given x, weighted by the portion of students in the counterfactual class in group x. Bias in this case comes from differences within a group x between the baseline and counterfactual treatment effects. There is no longer any bias from differences in the fraction of students with characteristics x. If a teacher is worse at teaching struggling students, for example, and their new class has many more struggling students, the left hand side will overestimate their impact on the new class. The right hand side will only be biased if there is variation within performance groups in both the teachers impact and the student compositions. For example, teachers may have different impacts on students based on race, even within a pretest group, and racial composition could differ across class (Delgado, 2022).

2.A.3. Value-added Estimation Details

The above discussion shows the theoretical importance of measuring test score heterogeneity, but of course, measuring heterogeneity increases the variance of estimates. Weather or not it can be effectively measured to improve policy analysis is a practical empirical question. Below we cover two different methods for measuring test score heterogeneity, but first, a quick review of our benchmark traditional value-added estimation.

2.A.3.1 Estimators

Standard value-added

In order to reference our estimates against an up to date and rigorously tested value-added approach, we follow the baseline practices used in Chetty et al. (2014a) and implement it using the associated Stata package (Stepner, 2013). The general approach of these authors is as follows. First regress test scores $S_{i,t}$ on controls $X_{i,t}$ which gives test score residuals A_{it}. This is obtained from a regression on test scores of the form

$$S_{i,s,t} = \alpha_{j(i,s,t)} + \beta_s X_{i,t} + \epsilon_{i,s,t} \tag{20}$$

Where $X_{i,t}$ includes cubic polynomials in prior year test scores in math and ELA, those polynomials interacted with student grade level, ethnicity, gender, age, lagged suspensions and absences, indicators for special education and English language learner status, cubic polynomials in class and school-grade means of prior test scores in both subjects each interacted with grade, class and school means of all the other covariates, class size and type indicators, and grade and year dummies[19]. $j(i,t)$ is the index for the teacher who has student i in her class at time t, so $\alpha_{j(i,t)}$ are year-specific teacher fixed effects.

Next, we average the residuals within each class year to get

$$\bar{A}_{jt} = \frac{1}{n} \sum_{i \in i: j(i,t)=j} A_{it} \tag{21}$$

The last step is to use the average residuals in every year but year t, denoted $\mathbf{A}_j^{-t}$, to predict $\bar{A}_{jt}$. Specifically, we choose coefficients $\psi = (\psi_i, ..., \psi_{t-1})$ to "minimize the mean squared error of the forecast test scores (Chetty et al., 2014a)"

[19]The covariates match those used in (Chetty et al., 2014a) closely. Means and standard deviations of the underlying variables appear in Appendix Table **??**.

$$\psi = \arg\min_{\psi} \sum_{j} \left(\bar{A}_{jt} - \sum_{s=1}^{t-1} \psi_s \bar{A}_{js} \right)^2 \tag{22}$$

This then gives the estimate for teacher j's value-added in year t of

$$\hat{\mu}_{jt} = \psi' \mathbf{A}_j^{-t} \tag{23}$$

Binned Estimator

A simple way to add heterogeneity into this model is to include an indicator for each student's type and estimate teacher affects separately for each type. This gives each teacher an estimate for each student type. We separate students into above and below median prior year test score bins. All of the above math works out essentially the same except we now have twice as many parameters to estimate. We now estimate residuals from the equation

$$S_{i,t} = \alpha_{j(i,b,t)} + \beta X_{i,t} \tag{24}$$

where $j(i,b,t)$ indicates if student i is assigned to teacher j in bin b at time t. Next we group residuals for teacher, year, bin,

$$\bar{A}_{jBt} = \frac{1}{n} \sum_{i \in i: j(i,B,t)=j} A_{it} \tag{25}$$

and we do the leave-one-out estimator with teacher bin estimates across years

$$\psi = \arg\min_{\psi} \sum_{j} \left(\bar{A}_{jBt} - \sum_{s=1}^{t-1} \psi_s \bar{A}_{jBs} \right)^2 \tag{26}$$

This then gives the estimate for teacher j's bin B value-added in year t of

$$\hat{\mu}_{jBt} = \psi' \mathbf{A}_{jB}^{-t} \tag{27}$$

We also apply statistical shrinkage, using the variance within each bin so that if the variance of one bin is higher it does not get shrunk more relative to the other bins.

2.A.3.2 Aggregating Estimates

The above method gives multiple estimates for each teacher's impact on the different types of students. For specific policy interventions, like teacher reassignment, these can be combined by summing up the conditional expected treatment with the conditional average welfare weight such as the weights described in theorem 2.

However, in some cases, value-added is also used for general teacher ranking and assessment. If teacher heterogeneity is significant, is there still a way to objectively rank teachers according to a particular set of heterogeneous welfare weights? There is not a perfect single solution since their impact depends on the class or policy environment. However, one solution that puts teachers on an even playing field is to rank teachers on the expected welfare impact they would have on an average representative class, rather than on the average impact on test scores for the class they have, which may depend on class composition, which is outside of the teacher's control and does not reflect their welfare impact.

In the discrete setting, let $\bar{\omega_k}$ and γ_k be the average proportion of students in group k and the welfare weight for group k respectively. Let $\alpha_{j,k}$ be teacher j's group specific value-added for group k. Than we can aggregate their group specific test scores as

$$VA_j = \sum_k \gamma_k \bar{\omega_k} \alpha_{j,k} \tag{28}$$

This gives the welfare benefit a teacher would have on an average class. This is the same as A_j from definition **??**. Now, choosing the average class composition for every teacher may or may not be the right normative choice. Suppose that a teacher has a big comparative advantage with high scoring students in a district with, on average, very high scoring students, but their class is primarily low scoring. What is the right way to assess their performance? They may not be

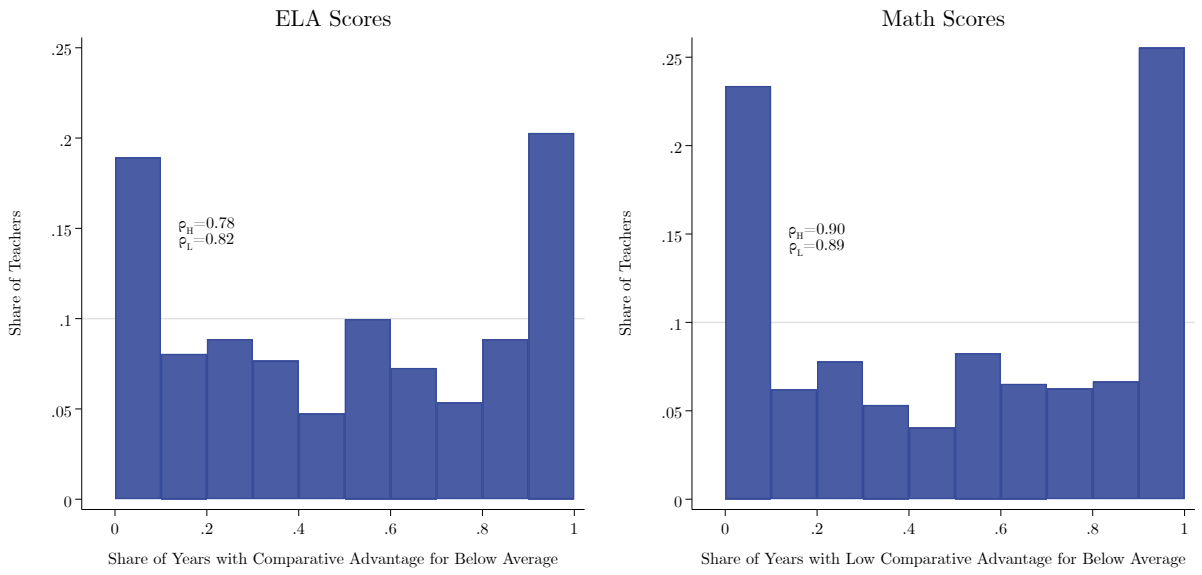


Figure 2.16: **Measures of Comparative Advantage Persistent**

bad relative to their well matched peers, which the above metric could tease out, but they may still in fact be doing a poor job helping the students they have, which the above metric ignores. This emphasizes that in a world of heterogeneity, no metric will be perfect. However, equation 28 does help to rank teachers based on what is under their control.

2.A.4. Validation and Robustness of Heterogeneous Estimates

In addition to these standard exercises we leverage the longitudinal nature of our data to show that our heterogeneous estimates capture the same correlations with long term outcomes as do standard value-added does—despite being identified off of only half of the students. In the spirit of Chetty et al. (2014b), we focus on five main outcomes: high school graduation, college enrollment in the year after twelfth grade (two-year, four-year, and any), and completion of a bachelors degree within six years of (anticipated) high school graduation. If our heterogeneous estimates corresponds to future outcomes in a similar way to standard value-added, then the predictive power has not been diminished and the estimated effects are fitting on true value-added rather than idiosyncratic noise.

To test the predictive power of value-added, we regress each outcomes teacher value-added and the controls from equation **??** in a student-subject-grade level regression. For the binned estimates, we include terms for the high- and low-bin value-added interacted with an indicator for whether the student is a high scoring:

$$y_{i,j,s,t} = \tau_{VA}\hat{\gamma}^{VA}_{j,s,t}\mathbb{1}(k_i = g) + \beta_2 X_i + \nu_{i,j,s,t} \tag{29}$$
$$y_{i,j,s,t} = \sum_{g=H,L} \tau_g \hat{\gamma}^{g}_{j,s,t}\mathbb{1}(k_i = g) + \beta_3 X_i + \nu_{i,j,st}$$

This is analogous to treating the each teacher-subject-bins as a separate class where the coefficients on value-added indicate the predictive power of high-bin value-added in each subject on high-scoring students' outcomes and low-bin value-added on low-scoring students' outcomes.

Figure 2.17 reports the results from the regression in equation 29 on each outcome variable. Our results show striking similarities between traditional value-added and our estimates, despite the fact that we split our sample to estimate above- and below median effects. Surprisingly, none of the measures are predictive of high school graduation. One explanation for this might be that SDUSD has an unusually high graduation rate, averaging 90 percent for our sample, creating ceiling effects. While not statically significant, standard value-added and both of our binned estimates track closely with an increase in any college, primarily from four year college with potentially a drop in two year college, and an increase in a bachelor's degree within 6 years. We can also see that the standard errors for each student group are not actually much bigger than for the mean as a whole suggesting that the variance is loading on this achievement dimension. On a whole these effects are similar with those in Chetty et al. (2014b) and **?** for traditional value-added.

Although imprecise, these effects point to patterns in college enrollment that are independently interesting beyond this validation exercise. For example, the effect on two-year college enrollment is higher for below-median students, which makes sense if they are more likely to be on the margin of not going to any college. On the other hand, for high-scoring students, well matched value-added may decrease the probability of two-year college enrollment and increase in

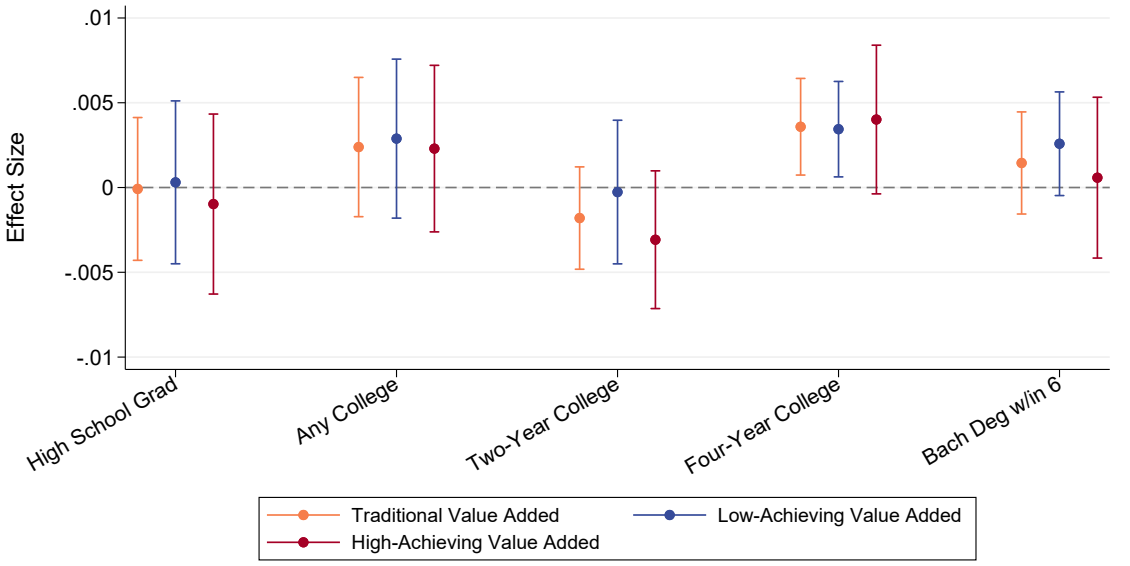


Figure 2.17: **Our Estimates Predict Long Term Effects as Well as Standard VA**
Note: This figure compares the effect of different measures of teacher value-added on long-term outcomes. All regressions follow equation 29 and include all controls from the value-added estimation. For the outcomes, High School Grad is an indicator for whether the student graduated from high school, Two Year College is an indicator for whether the student enrolled in a two-year college within a year following high school graduation, Four-Year College is an indicator for whether the student enrolled in a four-year college within a year following high school graduation, and Any College is an indicator for either Two Year College or Four-Year College. Finally, we model an indicator for whether the student obtained a Bachelor's degree within six years of high school graduation.

the probability of four-year college enrollment. These patterns are consistent with well-matched teachers increasing the quality of post-secondary education, moving students on one margin from no college to two-year colleges and on another margin from two-year colleges to four-year colleges.

Chapter 3

The Hidden Cost of Strict Job Qualification Requirements: Application Gaps, Diversity, and Perceptions about Hiring

Tanner Eastmond and Amanda Bonheur [0]

Abstract

Despite years of policy and revised corporate practice intended to correct inequality in the hiring process, application gaps persist for women and individuals from underrepresented racial minority groups. This study explores whether it is possible to narrow this application gap and promote diversity in the applicant pool by including encouraging and informative language around qualification requirements in job ads. We do so using a large-scale, "reverse audit study" field experiment where we randomize the content of job ads and observe job seeker behavior. Specifically, we established a non-profit firm to act as an intermediary in the job search process. This firm reposts real job ads and collects information from job seekers interested in applying. We randomize whether we encourage people to apply even if they don't meet all of the listed quali-

[0]Department of Economics, University of California San Diego. Authors can be reached at teastmond@ucsd.edu and abonheur@ucsd.edu. This research is conducted under UCSD IRB #806291. Pre-registration of the pilot can be found on AsPredicted #148892. This work has been supported (in part) by Grant #2301-41761 from the Russell Sage Foundation and by the W.E. Upjohn Institute for Employment Research. Any opinions expressed are those of the principal investigator alone and should not be construed as representing the opinions of either funder. We also thank the Yankelovich Center for Social Science Research and the UC San Diego Department of Economics Diversity and Research fund for generous financial support with this research.

fications and whether we inform them that companies routinely hire individuals who do not have all qualifications. Preliminary results show that this light touch intervention nudges more people into applying. Further analysis will study how the intervention changes perceptions of the hiring process and whether it has larger impacts on women, individuals from underrepresented racial minority groups, and people with non-traditional employment backgrounds.

3.1 Introduction

Despite years of policy and revised corporate practice intended to correct inequality in the hiring process, gaps persist for women and individuals from underrepresented racial minority groups at every level of the hiring process. Past work has shown that qualified workers in underrepresented groups are not only less likely to be hired after the interview stage (MORE CITES Goldin & Rouse, 2000), but their resumes are screened out at higher rates (MORE CITES Kline, Rose, & Walters, 2022) and they are even less likely to apply to the job in the first place (MORE CITES Burn, Firoozi, Ladd, & Neumark, 2022). This paper focuses on the first step of the hiring process, seeking to understand who applies to a job in the first place.

In particular, we explore the true extent of the application gap for underrepresented workers across a variety of industries and ask whether changing the language surrounding the listed qualifications in the job ad can induce more of these workers to apply for the job. This is motivated in part by a finding from a Hewlett Packard internal report that says "men apply for a job promotion when they meet 60% of the qualifications, but women apply only if they meet 100% of them" (Clark, 2014; Mohr, 2014; Sakowitz, 2018). Furthermore, women are 16% less likely to apply to a job after viewing it, apply to 20% fewer jobs overall, and are less likely to apply for 'stretch roles' (Tockey & Ignatova, 2019). Taken together, these findings would not be concerning if companies only ever hired workers that met every qualification from the job ad. Though companies are very thoughtful about the qualifications they put in their listings, there is no way to fully describe a perfect candidate with a short list of possible experiences and traits. This is not only true in theory, but in practice as well: Half (2019) finds that 84% of companies are willing to hire and train employees up where needed and 62% of employees have been offered a position even when they did not satisfy all listed required qualifications. By not applying in the first place, workers who would ultimately be an excellent match for the prospective employer remove themselves from the applicant pool, even though companies frequently extend offers to similar people.

In our study we evaluate whether high quality job seekers can be induced to apply for jobs

by varying the language in job ads around the list of qualifications. We do this using a non-profit corporation that we set up, the Job Connections Project (JCP)[1]. The Job Connections Project re-posts open job ads from other companies for full time positions and advertises on various online job boards, thus resembling a recruiting firm. Job seekers who click through our job ads from online job boards are randomly assigned to treatment or control versions of the job ad then are shown a brief survey before being routed to the actual application. If assigned to the control version, nothing is changed about the job ad. The various treatment arms add language encouraging the candidate to apply even if they do not meet all qualifications, informing them that companies routinely hire people without all listed qualifications, and informing them that women have been shown to be less likely to apply without all qualifications. This step is interposed in their standard job application process and is minimally disruptive to their job search, but to compensate them for their time we offer several free services to help job seekers. This design is similar in spirit to resume audit studies, where realism is preserved by randomizing the content of resumes sent in to real hiring managers. Our 'reverse' audit study instead randomizes content in real job listings, which allows us to observes job seeker behavior towards real jobs, measure real-time perceptions of the job, and understand job seeker attitudes towards the job market more broadly.

Crucially, our randomization does not change any of the listed qualifications in each listing. Keeping the listed qualifications fixed is important and relevant since companies thoughtfully choose the qualifications in job postings and altering qualifications is not feasible for many roles. Additionally, previous research has found that changing the qualifications can change the perceived rigor of the role and is not guaranteed to increase applicant diversity (Abraham, Hallermeier, & Stein, 2023). We focus on encouragement and information about the hiring process because, while some have suggested that this difference in applying behavior is a professional confidence problem (Rojas, 2021; Sandberg & Scovell, 2013), differing perceptions about the hiring process are a more likely culprit. Mohr (2014) surveys people about their application behavior and found that common reasons for not applying to a job include not wanting to waste time if they do not have

[1]Our company serves two major goals: help job seekers better match with jobs and study the labor market to improve the search and match process broadly.

a chance, fear of failure, and simply following the rules, with women being more likely to report these feelings. In other words, women "thought that the required qualifications were ... well, required qualifications. What held them back from applying was not a mistaken perception about themselves, but a mistaken perception about the hiring process." Not only does this impact women, but there is some evidence that other historically disadvantaged groups may be impacted as well (Avery & McKay, 2006; Wille & Derous, 2017).

Initial pilot results suggest that job seekers exposed to treatment are more likely to continue toward applying for the job, and furthermore that this is driven primarily by the simplest treatment arm that only encourages them to still apply if they do not meet all listed qualifications with no other information. We are still collecting data for the pilot, but will soon also be able to speak to the demographic composition of prospective applicants,[2] their quality, their perceptions of the job ad itself, and their broader perceptions of the job market. These data are collected from engagement with JCP's website, survey answers, and job seeker resumes.

We hypothesize that women, BIPOC individuals (Black, Indigenous, and People of Color),[3] and people who are Skilled Through Alternative Routes rather than a bachelor's degree (STARs)[4] will be more affected by encouragement and information about 'required' qualifications. We further hypothesize that the treatment will nudge qualified individuals (e.g. those who meet 7 to 9 out of 10 qualifications) into applying. Overall, this means that we expect a change in the demographic composition of the applicant pool with an equal or higher number of applicants who could perform the job well. This project aims to improve outcomes for those traditionally disadvantaged in the labor market.

Understanding who applies for jobs under different ways of presenting qualifications will both (i) help people who have been historically disadvantaged in the labor market and (ii) help

[2]Throughout this paper we call job seekers 'prospective applicants' if they click through our site indicating that they intend to apply, though we do not actually observe whether or not they do ultimately apply to the job.

[3]Pager and Pedulla (2015) find that Black people cast a wider net in their search relative to similarly situated White people, as an adaptation to deal with labor market discrimination. Due to this fact, that Black workers may already apply to a large breadth of roles, we may not see a large change for Black job seekers.

[4]I use this language since Opportunity @ Work, a non-profit organization that works to advance economic mobility, refers to these individuals as Skilled Through Alternative Routes or STARs.

companies with their hiring initiatives. In the two-sided job match process, employers choose how they present openings in job ads, and potential employees make application decisions. Just as employees navigate the hiring process, many employers struggle to recruit diverse applicants (Kessler, Low, & Sullivan, 2019). Job ad language is one of the common themes in articles about how-to-attract-a-diverse-workforce,[5] but no rigorous test has been performed yet. This project will provide valuable insight into how job seekers behave in the current job market. It has the potential to provide employers with a tool to attract a more diverse applicant pool and mitigate application gaps.

There are three main contributions of this work. The first is that we observe real job seeker behavior towards real, full-time positions. Our ability to preserve realism expands upon previous literature that has used fake job ads (Burn et al., 2022) or short-term positions (Castilla & Rho, 2023; Del Carpio & Guadalupe, 2021). Additionally, we are able to study effects across companies and occupations while abstracting away from company-specific reputation (Abraham et al., 2023).[6] The second contribution of this work is that the intervention is a partial solution that is easily implementable across a variety of contexts. We keep listed qualifications the same without removing or changing requirements, which is less effort from companies and is less likely to affect the perceived rigor of the role (Abraham et al., 2023). If treatment is effective in attracting different types of workers who could perform the job well, then this is a simple, readily adoptable policy tool for firms. No test scores or additional certifications are required to encourage women to apply (K. B. Coffman, Collis, & Kulkarni, 2019). Given that changing a sentence or two in a job ad is low-cost from a firm's perspective, the resulting policy recommendation may have a high likelihood of being adopted. Third, our survey design enables us to study mechanisms behind application decisions. Specifically, we measure perceptions of the hiring process, confidence, feelings of wasting time, interest in the role, self-assessments versus how they think hiring managers will view their ability, and more.

[5] See, for example https://bit.ly/3AjOhVn, https://bit.ly/3dxmtUG, and https://bit.ly/3JXLPHj.

[6] Abraham et al. (2023) randomizes information for corporate jobs for Uber. Since Uber is a well known company, people likely view their job ads through the lens of their perception of them.

In addition to the strong policy relevance, this project is at the frontier of methodologies to study inequality. We have created a new reverse audit study experimental design and gather original data.

Anticipated Contribution to the Literature

We are not the first to study application decisions of women and other traditionally disadvantaged groups in the labor market, but we innovate in realism, breadth, and depth.

Disparities arise throughout the hiring pipeline; there are gaps in who applies, who gets interviews, how interviews are rated, promotions, and more (Abraham et al., 2023; Avery & McKay, 2006; Benson, Li, & Shue, in press; Bertrand & Mullainathan, 2004; Burn et al., 2022; Chaturvedi, Mahajan, & Siddique, 2021; Clark, 2014; Mohr, 2014; Sakowitz, 2018; Tockey & Ignatova, 2019; Wille & Derous, 2017). This paper focuses on application gaps and its connection to perceptions about the hiring process.

We contribute to the social sciences literature about the existence, causes, and remedies for application gaps by gender and race (Barbulescu & Bidwell, 2013; Ekstrom, 1981; Llinares-Insa, González-Navarro, Córdoba-Iñesta, & Zacarés-González, 2018; Pager & Pedulla, 2015; Reskin & Bielby, 2005). Most notable for this project, women are less likely to apply for a job unless they have all of the qualifications listed, whereas men apply with a fraction of the qualifications (Mohr, 2014; Sakowitz, 2018). A Gender Insights Report from LinkedIn (Tockey & Ignatova, 2019) similarly found that women are 16% less likely to apply to a job after viewing it, apply to 20% fewer jobs overall, and are less likely to apply for 'stretch' roles. At first, people thought this was likely a professional confidence problem (Rojas, 2021; Sandberg & Scovell, 2013), but more recent research has shown it is more about their beliefs about the hiring "rules" (Mohr, 2014; Zucker, 2020). Further, there is little research into application behavior of non-binary and other gender minority individuals.[7] Aksoy, Exley, and Kessler (2024) shows that gender minority individuals exhibit less confidence and less favorable self-evaluations, and we hope to be the first to document application gaps, depending on sample size. Relatedly, African American job seekers

[7]For example, trans individuals and people who identify as genderqueer.

cast a wider net in their job search, as a response to labor market discrimination (Pager & Pedulla, 2015). This evidence demonstrates that perceptions about the hiring manager and hiring process are crucial components of application decisions.

These application gaps are likely inefficient because people are sometimes hired into stretch roles. Half (2019) found that 84% of companies are willing to hire and train up and 62% of employees have been offered a position when they were 'underqualified'.

Previous studies have shown that altering aspects of job advertisement language can inadvertently turn away or attract certain types of job seekers. For example, job applicants are affected by stereotyped language or gendered skills (Burn, Button, Menguia Corella, & Neumark, 2019; Burn et al., 2022; Chaturvedi et al., 2021; Kuhn, Shen, & Zhang, 2020), the inclusion of diversity and EEO statements (Dover, Major, & Kaiser, 2016; Flory, Leibbrandt, Rott, & Stoddard, 2021; Gaucher, Friesen, & Kay, 2011; Hurst, 2022; Leibbrandt & List, 2018), information about the success of marginalized groups (Choi, Pacelli, Rennekamp, & Tomar, 2022; Del Carpio & Guadalupe, 2021), highlighting personal benefits (Linos, 2017), and the framing of factual information about the job (L. C. Coffman, Featherstone, & Kessler, 2017; Dal Bó, Finan, & Rossi, 2013; Flory, Leibbrandt, & List, 2015; Gee, 2019; Marinescu & Wolthoff, 2020; Samek, 2015). These papers answer different questions than ours, but suggest that less strict language around required qualifications could be effective in modifying behavior. On the other hand, Castilla and Rho (2023) find negligible effects of the gendering of job postings or of the recruiter, implying that small language changes may not very important for online job postings.

The closest work is Abraham et al. (2023), which finds that making listed qualifications less demanding encourages people to apply and reduces the skill gap between male and female applicants, but also changed perceptions of the rigor of the role. We complement this work by keeping the listed qualifications fixed while changing the wording around the qualifications. This distinction is important since companies choose which qualifications are needed for job postings, so removing or altering the qualifications themselves is not feasible for many roles. Similarly, K. B. Coffman et al. (2019) find that adding strict test score cutoffs reduces ambiguity around

qualifications and reduces the gender gap in willingness to apply, but test scores are not readily available for most roles.[8]

This paper innovates on previous audit study designs by proposing an infrastructure for a 'reverse' audit study which focuses on job seeker response instead of firm behavior. Audit studies have been used to research discrimination from firms by sending fake resumes to companies (Gaddis, 2018). These have found gaps in callback rates by race (Bertrand & Mullainathan, 2004; Kline et al., 2022; Quillian, Pager, Hexel, & Midtbøen, 2017); age (Farber, Silverman, & von Wachter, 2016); gender (Bohren, Imas, & Rosenberg, 2019); religion, attractiveness, sexual orientation, and more (Bertrand & Duflo, 2017).

Most other research in this space have either used fake job ads, studied short-term work such as internships, or partnered with one company.[9] Burn et al. (2022) uses a similar concept and method to our reverse audit study design, and we innovate on their approach to ensure realism and reproducibility. They answer a different question and encountered a number of implementation difficulties due to posting fake job ads.[10] We learn from their work and create an experimental design that not only dodges many of their implementation hurdles, but also posts ads for full-time jobs across firms, meaning that we can establish realistic effects independent from a company's reputation.

Overall, our paper expands upon the current literature by studying job seekers within their job search process, across types of roles, and with an in-depth study of mechanisms.

3.2 Methodology

We design a large-scale 'reverse' audit study field experiment where we randomize the content in real job ads to explore how job seekers respond to information in jobs ads about the listed

[8]For instance, there is no test or score cutoff to define someone as a 'strong communicator'.

[9]For examples, see Abraham et al. (2023); Burn et al. (2022); Del Carpio and Guadalupe (2021); Flory et al. (2021). One paper that does use real job seekers, real jobs, and across firms is Kuhn et al. (2020). They study explicit requests for applicants of a particular gender in a Chinese context. Alternatively, we study language surrounding job qualifications that imply varying levels of strictness in the US context.

[10]These include technical difficulties such as having to have phone numbers and research assistants for every fake job ad they posted, which limited their sample size.

qualifications. This methodology is in the spirit of the large set of resume audit studies in the literature (e.g. Bertrand & Mullainathan, 2004; Kline et al., 2022). To accomplish this we established a non-profit company, the Job Connections Project (JCP), that acts as an intermediary within the regular job application process to preserve realism. The JCP serves two primary purposes simultaneously. First, it serves as a platform to post jobs to learn about real job seeker responses to the content of job ads and the labor market more broadly. Second, it provides services to those job seekers to compensate them for the small incursion into their time and to help them in their job search.

Using data from the Current Population Survey, we first identify 10 occupational categories with differing levels of baseline representation of women, Black workers, and Hispanic workers for inclusion in our study. These occupations are shown in Table 3.1. With these occupations in hand, we conduct our experiment using our platform as follows.

For each batch of job postings, we find 3-5 open job listings in the chosen category, each that are located in the same randomly chosen city in the US, are for full time jobs, and that have clear lists of qualifications.

For open job listings used in the study, we remove identifying information for the hiring company and repost the job ad on the JCP website.[11] We also post versions on multiple large job listing websites and, to prevent contamination of the treatment, do not include the qualifications in these initial listings. Job seekers then come across our ads on these sites and click through if they are interested in applying. Once they click through, they are redirected to our company's website, https://jobconnectionsproject.org/, and are randomized by IP address to see either the control or one of the treatment versions of the job ad and told that we are a non-profit trying to learn about the job market (Figure 3.1).

Our control and treatment settings describe the listed qualifications using whatever heading

[11] We remove identifying information for the hiring company to limit reputation-related confounding factors. This is common practice among recruiting firms and so will not be out of the ordinary.

Table 3.1: **Occupations used in the study and their fraction of representation**

Occupation	Fraction of employed persons who are: Women	Black/African American	Hispanic/ Latino
Engineers, various	Low	Low	Low
Credit counselors and loan officers**	Median	Median	Median
Human resources manager	High	High	High
Preschool and kindergarten teachers	High	High	Median
Elementary and middle school teachers	High	Median	Median
Secondary teachers	Median	Low	Low
Postsecondary teachers	Median	Median	Low
Public relations specialists	High	Median	Low
Wholesale and retail buyers	Median	Low	High
Training and Development specialists	Median	High	Low
Computer/data/software occupations**	Low	Median	Low
Mental health and guidance counselors	High	High	Median

Low means the fraction of that identity is less than the 36th percentile of that group's participation across all occupations. Median is between the 36th and 63rd percentile, while High is above the 64th percentile.

**Occupations used in pilot.

Figure 3.1: **Example of what job seekers see when read job ad on JCP's website**

the original job post had.[12] In the control arm, the job ad is otherwise completely unaltered.[13] The

[12]Examples include: Qualifications, Required Qualifications, Knowledge and Skills, Required Skills/Competencies, Experience, and similar.

[13]Except for removing the identifying information of the hiring company, as noted earlier.

first treatment (which we call hereafter the 'Encourage' treatment) includes a blurb saying: "Don't meet every single requirement? If you're excited about this role but your past experience doesn't align perfectly with the job description, we encourage you to apply anyways. You may be just the right candidate for this role." This blurb is embedded in the job ad at the end of the qualifications section. The second treatment ('Encourage + Hiring Info' treatment) is the same, except we add the statement "Most companies routinely hire individuals who lack some of the stated required skills" to the blurb. The third treatment arm ('Encourage + Hiring Info + Women Info' treatment) includes the "Most companies routinely hire…" statement, along with "Studies have shown that women are less likely to apply to jobs unless they meet every single qualification." The fourth treatment arm matches the third ('Encourage + Hiring Info + Women Info'), but the blurb comes from the JCP. In this fourth treatment ('Encourage + Hiring Info + Women Info from JCP'), instead of being embedded in the job advertisement itself, the blurb appears in a popup box under the text "Tip from the Job Connections Project:". Overall, treatment interventions are a variation of the following language:

> "Don't meet every single requirement? Studies have shown that women are less likely to apply to jobs unless they meet every single qualification, but most companies routinely hire individuals who lack some of the stated required skills. So if you're excited about this role but your past experience doesn't perfectly align with the job description, we encourage you to apply anyways. You may be just the right candidate for this role."

After reading the full job ad with randomized intervention language, interested job seekers click 'Continue' and are presented with the free services offered by the JCP to job seekers (free resume feedback and access to a Chrome extension that automatically fills out their information on other job applications).[14] Directly below these services on the same page, the job seeker is asked to participate in a survey about their views toward this role to help us learn about the job match process. This 2-minute survey asks for their likelihood of applying to the position, likelihood of accepting the position if offered, current employment status, basic demographics (race, gender

[14]The Chrome Extension is available in the Chrome Web Store at the following link https://chromewebstore.google.com/detail/job-connections-project-j/apjpojgndgefpkgpmkifhhgmepkgieki.

identity), and a set of questions about how they fit with the role. These last questions give us self-assessments of quality and fit for the specific role. Specifically, we elicit how well they believe they could perform the job, their own perceptions of the proportion of qualifications they meet, whether they have other relevant skills that weren't specifically asked for, their guess of whether the hiring manager will recognize their potential, and their guess as to their chances of being invited for an interview.

The last part of the survey elicits perceptions of the job ad itself. We ask agree-disagree Likert scale questions of how they view the following: the hiring company's leniency when reviewing candidates relative to other companies, whether they will be wasting their time if they apply to this job, whether they want to apply to more 'stretch' roles, how people should apply when they meet most but not all of listed qualifications, and whether companies in general stick to required qualifications.

Survey answers combined with how they interact with our website allow us to determine whether the presentation of required qualifications can affect perceptions and the likelihood that people apply for jobs they could perform well.

The field experiment design allows us to measure a few different outcomes of how application behavior is affected by treatment using interaction with our website, survey answers, and job seeker resumes. The outcomes are the number, composition, and quality of job seekers. The quantity of likely applicants is tracked using both self-reported likelihood of applying and click-through rate to the hiring company's webpage. Specifically, we measure the proportion of those who view a job ad on our website and click 'Continue', and the proportion that clicks 'Apply now on Company website' versus not continuing or selecting 'No longer interested, show me other jobs'. We can measure click-through rates for all individuals who encounter our website and self-reported likelihood of applying for survey respondents. While we do not observe actual application rates,[15] these two proxies will be proportional and able to detect treatment effects. Demographic and em-

[15]Those who select 'Apply on Company Website' will be redirected to the original job ad on the hiring company's website where they will be able to apply for the open job. At this point we will not interact with or receive information from the job seeker any longer. Those who select 'Not interested, show me other jobs' will be redirected to a page on our website that lists multiple related jobs.

ployment status characteristics are collected in the survey to identify composition effects. Self assessments from the Madlib style question of how they fit with the job being advertised, including how well they believe they could perform the job and whether they think the hiring manager will see their potential give us some measures of quality. We also measure quality using listed employment history, education, and skills from individuals who share their resume with us. We believe that our treatment will lead to more people applying across the skill distribution. While people who are less likely to perform the job well may also apply more, we predict that the number of excellent candidates that would perform the job well will also increase, giving employers a larger desired applicant pool.

Importantly, we care about who is most affected by job ad language, and whether the intervention can encourage more women to apply. We will explore treatment effects by race, gender, education, and employment status. We hypothesize that these modifications will have larger impacts on women, BIPOC individuals, and STARs (skilled individuals without a bachelor's degree). We know that women tend to take themselves out of the running, and we hope to find evidence that more women have intent to apply when presented with less strict language surrounding required qualifications.[16]

The strength and believability of the encouragement language may vary depending on the industry and/or current amount of representation of women and BIPOC individuals. For this reason, we select a handful of occupations with varying levels of current diversity, such that we can identify heterogeneous effects along this margin. This is important for understanding when effects translate to other contexts. The occupations are chosen using 2023 data from the Current Population Survey (CPS).[17] Table 3.1 shows the 10 chosen occupations and whether their gender and race/ethnicity fractions are average, low, or high. Due to the smaller size of our pilot, our pilot includes 2 of these occupations: Data Science/Computer Occupations and Loan Officer/Credit

[16]We anticipate that we may also find larger treatment effects for individuals who are currently employed, since they can be more discretionary with respect to job ad language than currently unemployed individuals.

[17]The CPS detailed occupation data by race/ethnicity and sex can be found at https://www.bls.gov/cps/cpsaat11.htm .

Counselor.[18]

Further, our setup allows us to unpack mechanisms behind behavior. We will use the survey to examine whether people's perceptions move in conjunction with their application behavior. As stated above, we gather data on perceptions of own fit, how others view them as a candidate, if this job is worth applying to, and more. If people who are nudged into applying are also less likely to think they are wasting their time when they view treated job ads, then fear of wasting time is a factor driving application gaps that can be influenced by job ad language.

We are at the frontier of studying real job seekers in relation to actual full-time jobs across multiple employers. The contributions of this project include: (1) We quantify the impact of language surrounding job qualifications, without changing the qualifications themselves, on applicant pool diversity in a real job setting. (2) We explore mechanisms behind the results using survey evidence, including feelings of wasting time, perceived rules, etc. (3) We provide concrete policy recommendations to attract a wider group of applicants and reduce gaps. (4) We innovate on previous methods by developing a reverse audit study methodology and founding the Job Connections Project non-profit. This setup creates a novel dataset, allows analysis at-scale across industries and job types, and will enable future work to deepen understanding of job seeker behavior.

3.3 Results

This section describes preliminary results from the pilot up to April 18, 2024, and are subject to change. At this time we have 203 observations, where each observation is a unique person and job ad pairing. This is from 196 unique job seekers who have come across our website.[19] Of these 203 people looking at a job ad on our website, 75 (37%) clicked continue at the bottom of the job description, 40 (20%) filled out some part of the survey or uploaded their resume, and 47 (23%) clicked 'Apply now on company website'.

[18]The data science/tech was chosen over engineers, who also have low representation among women and non-White individuals, since there have been layoffs around the time of our pilot, such that our ability to reach job seekers is higher.

[19]This means that 96.5% of job seekers interact with the JCP once, looking at only one JCP job ad, while a minority also look at other jobs posted by the JCP.

Roughly half of job seekers were randomized into seeing control ads (89, 42.4%), while the other half saw a treatment ad (117, 57.6%). Within the treated individuals, roughly 30 job seekers (15% of entire sample) saw each version of Treatment Encourage, Encourage + Hiring Info, Encourage + Hiring Info + Women Info, and Encourage + Hiring Info + Women Info from the JCP. Note that this randomization is occurring at the IP and job ad level, meaning that we can also make comparisons across type within each open position.

Preliminary results show that the treatment language may encourage applications. There are 3 ways we measure application likelihood; clicking continue, self-reporting high likelihood of applying; and clicking 'Apply now on company website'.

The first outcome is whether treatment language (encouragement and information around required qualifications) induces more people to click 'Continue'. This would be the first step showing more intention to apply to the position. Individuals who are shown job ads with treatment variation are 40% more likely to click continue after reading the job ad (Table 3.2). While this is not statistically significant, we expect more power as the sample size grows. Similarly, job seekers in treatment 1, 2, and 3 are more likely to click continue. Interestingly, job seekers who see treatment 4, the encouragement and information intervention listed as a tip from the JCP, are not more likely than control individuals to click continue.

Another measure of intent to apply is the number who click 'Apply now on company website' on the next page.[20] Conditional on clicking continue, treated individuals are equally likely to click 'Apply now' as control individuals. As our sample grows, we could also look at self-reported answers to "How likely are you to apply to this job?" and whether they vary by treatment.

Our main interest is whether we can disproportionaly encourage more women and other traditionally disadvantaged groups into applying more. We currently do not have enough sample size, particularly of women, to answer this pertinent question.

Forty job seekers filled out some portion of the survey or uploaded their resume, with most of them filling out every question of the 2-minute survey. Descriptive statistics of those who chose

[20]After clicking continue, job seekers are presented with JCP's survey and/or can click 'Apply now on company website'.

Table 3.2: **Treated job seekers are more likely to click continue**

Clicked Continue	(1) (Odds Ratio)	(2) (Odds Ratio)
Treated	1.391	
	(0.448)	
Treatment Encourage		2.071**
		(0.765)
Encourage+Hiring Info		1.308
		(0.390)
Encourage+Hiring+Women Info		1.381
		(0.741)
Encourage+Hiring+Women Info (JCP)		1.036
		(0.300)
N	203	203
N clusters	18	18
Pseudo R^2	0.005	0.011

Exponentiated coefficients; Standard errors in parentheses

Standard errors are clustered by job ad. Results are similar if use robust standard errors.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

to share their information is shown in Table 3.3.

Overall, we see substantial variation in our pool of job seekers and a few patterns (Table 3.3). We observe more men than women are intending to apply to these positions, which makes sense given the current gender make-up of these occupations. There is strong alignment between showing intent to apply by clicking continue and self-reporting a high likelihood of both applying to the job and accepting the job if offered. We have a variety of racial groups, a range of years of experience, and both people who are currently employed and unemployed. We have mostly people with college degrees or higher, which fits with the types of job ads used in the pilot.

Table 3.4 shows our potential mechanism questions about own confidence, self-assessment of own skills, perceptions of the hiring manager, perceptions of this company, and more general assessments of the hiring process. From this you can see that there is substantial variation in most categories, meaning these variables have the potential to provide insights into behavior.

Table 3.3: **Descriptive statistics of job seekers who shared additional information**

	Summary
N	40
Clicked Continue	1.000 (0.000)
Clicked Apply Now	0.900 (0.304)
Number of survey questions answered	14.275 (5.875)
Uploaded resume	0.800 (0.405)
Gender	
Man	29 (85.3%)
Woman	5 (14.7%)
Non-binary/genderqueer/other	0 (0%)
Race	
A race/ethnicity not listed here	1 (2.9%)
Asian or Pacific Islander	13 (38.2%)
Black or African American	6 (17.6%)
Hispanic or Latino	6 (17.6%)
Multiracial or Biracial	1 (2.9%)
White or Caucasian	7 (20.6%)
Likelihood Apply	
Equally Likely and Unlikely	1 (2.6%)
Likely	8 (21.1%)
Very Likely	29 (76.3%)
Likelihood Accept	
Likely	6 (16.7%)
Very Likely	30 (83.3%)
Highest education attained	
High School	1 (2.9%)
Some college	2 (5.9%)
Bachelor's degree	18 (52.9%)
Graduate degree	13 (38.2%)
Experience (years)	4.714 (3.886)
Employment Status	
Employed (full-time)	16 (47.1%)
Employed (part-time)	5 (14.7%)
Unemployed	13 (38.2%)

3.4 Discussion & Policy Implications

Findings from this study have clear and useful policy implications for all parties involved in the job search-and-match process. This paper will either provide an evidence-backed solution to mitigate application gaps or insights into why it doesn't work and proposals for other solutions.

If using accessible language around qualifications has the hypothesized effect, then this intervention provides more opportunities for job seekers traditionally disadvantaged in the labor market. The slight interventions can be used to improve equality in overall placement outcomes for women, BIPOC workers, and STARs. The more that companies adopt these tested job ad modifications, the more job seekers who are influenced by the changes can benefit by being encouraged

Table 3.4: **Variation in perceptions of job seekers**

	Summary
N	40
Think perform on job	
I could do the job well	22 (73.3%)
the job would be a stretch, but I could learn quickly	7 (23.3%)
this job would be a real challenge	1 (3.3%)
How many qualifications meet	
meet all	14 (45.2%)
meet most	14 (45.2%)
meet some	3 (9.7%)
Have other relevant skills	
many	25 (80.6%)
some	6 (19.4%)
Think hiring manager see my potential	
would	30 (93.8%)
may or may not	1 (3.1%)
would not	1 (3.1%)
Likelihood interview	
very likely	20 (62.5%)
likely	9 (28.1%)
equally likely and unlikely	1 (3.1%)
unlikely	2 (6.2%)
This company more lenient	
Strongly disagree	4 (13.8%)
Disagree	3 (10.3%)
Neither agree nor disagree	14 (48.3%)
Agree	6 (20.7%)
Strongly agree	2 (6.9%)
Feel wasting time if apply	
Strongly disagree	11 (37.9%)
Disagree	10 (34.5%)
Neither agree nor disagree	6 (20.7%)
Agree	2 (6.9%)
Want to apply more stretch roles	
Strongly disagree	2 (7.1%)
Disagree	4 (14.3%)
Neither agree nor disagree	6 (21.4%)
Agree	11 (39.3%)
Strongly agree	5 (17.9%)
Think everyone should apply when meet most qualifications	
Strongly disagree	3 (10.3%)
Neither agree nor disagree	4 (13.8%)
Agree	9 (31.0%)
Strongly agree	13 (44.8%)
Sticking to required qualifications less common	
Strongly disagree	5 (16.7%)
Disagree	1 (3.3%)
Neither agree nor disagree	9 (30.0%)
Agree	11 (36.7%)
Strongly agree	4 (13.3%)

to give themselves a chance at 'stretch roles'.[21] Job seekers could also infer which firms care about diversity through the use of language encouraging a broader range of people to apply.

Results will inform best practices for companies to attract a diverse set of applicants that

[21] Applying to 'stretch roles' is a good thing because sometimes people who do not meet all qualifications are hired. In fact, 62% of survey respondents in Half (2019) were offered jobs when they didn't have all qualifications.

can be implemented without financial, time, or capacity barriers. These minor changes to job ads are a low-cost intervention from the firm's perspective.[22] At the same time, making these changes will benefit companies through more diverse desired applicant pools. Further, the experimental language around job qualifications can be applied to all types of roles for any industry. We will analyze heterogeneous effects to tailor recommendations.

Policymakers at various levels can integrate insights into policies. Online job boards could create guides for companies that want to advertise open positions on their platform. Hiring consultants and Diversity & Inclusion (D&I) officers can advocate for use of this empirically-supported tool to mitigate application gaps. This benefits companies since they will see a larger range of applicants, which could lead to better matches since fewer people would be left out of the applicant pool.

3.5 Conclusion

Overall, this research has important ramifications for the labor market. We hope to identify one lever than can be used strategically to alleviate application gaps. Current job ads may be discouraging women and others from applying to some jobs, particularly stretch roles. This research tests a simple and hopefully effective way to reduce this problem, namely, using encouragement around job qualifications and information about the hiring process.

Results are subject to change as data collection is currently ongoing.

Chapter 3, in part, is currently being prepared for submission for publication of the material and is coauthored with Bonheur, Amanda. The dissertation author was the primary researcher and author of this material.

[22] One potential cost is the company needing to sort through a larger number of applications. We anticipate that the interventions will induce applications from capable candidates and also less qualified individuals. We will analyze the size of this trade-off to inform our policy recommendation.

3.6 References

Abraham, L., Hallermeier, J., & Stein, A. (2023). Words matter: Experimental Evidence from Job Applications. *Forthcoming (Revise & Resubmit), Journal of Economic Behavior & Organization*. Retrieved from https://drive.google.com/file/d/1mOl_P9ezjGY0Dfo1Bzus47yfCUY8LyYf/view

Aksoy, B., Exley, C. L., & Kessler, J. B. (2024, January). The gender minority gaps in confidence and self-evaluation. *National Bureau of Economic Research Working Paper Series*(32061). Retrieved from http://www.nber.org/papers/w32061 doi: 10.3386/w32061

Avery, D. R., & McKay, P. F. (2006). Target Practice: An Organizational Impression Management Approach to Attracting Minority and Female Job Applicants. *Personnel Psychology*, *59*(1), 157-187. doi: https://doi.org/10.1111/j.1744-6570.2006.00807.x

Barbulescu, R., & Bidwell, M. J. (2013). Do Women Choose Different Jobs From Men? Mechanisms of Application Segregation in the Market for Managerial Workers. *Organization Science*, *24*(3), 737-756.

Benson, A., Li, D., & Shue, K. (in press). 'Potential' and the Gender Promotion Gap. *R&R, American Economic Review*. Retrieved from https://danielle-li.github.io/assets/docs/PotentialAndTheGenderPromotionGap.pdf

Bertrand, M., & Duflo, E. (2017). Chapter 8 - Field Experiments on Discrimination. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of field experiments* (Vol. 1, p. 309-393). North-Holland. doi: https://doi.org/10.1016/bs.hefe.2016.08.004

Bertrand, M., & Mullainathan, S. (2004, September). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, *94*(4), 991-1013. Retrieved from https://www.aeaweb.org/articles?id=10.1257/0002828042002561

Bohren, J. A., Imas, A., & Rosenberg, M. (2019, October). The Dynamics of Discrimination: Theory and Evidence. *American Economic Review*, *109*(10), 3395-3436. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.20171829 doi: 10.1257/aer.20171829

Burn, I., Button, P., Menguia Corella, L. F., & Neumark, D. (2019, December). *Older Workers Need Not Apply? Ageist Language in Job Ads and Age Discrimination in Hiring* (Working Paper No. 26552). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w26552 doi: 10.3386/w26552

Burn, I., Firoozi, D., Ladd, D., & Neumark, D. (2022, July). *Help Really Wanted? The Impact of Age Stereotypes in Job Ads on Applications from Older Workers* (Working Paper No. 30287). National Bureau of Economic Research. doi: 10.3386/w30287

Castilla, E. J., & Rho, H. J. (2023). The Gendering of Job Postings in the Online Recruitment Process. *Management Science*, *0*(0). Retrieved from https://doi.org/10.1287/mnsc.2023.4674

Chaturvedi, S., Mahajan, K., & Siddique, Z. (2021). Words Matter: Gender, Jobs and Applicant Behavior. *IZA Institute of Labor Economics Discussion Paper*(14497).

Choi, J. H., Pacelli, J., Rennekamp, K. M., & Tomar, S. (2022). Do Jobseekers Value Diversity Information? Evidence from a Field Experiment. *Journal of Accounting Research*. Retrieved from https://www.utah-wac.org/2022/Papers/choi_UWAC.pdf

Clark, N. F. (2014, April 28). Act Now To Shrink The Confidence Gap. *Forbes*. Retrieved from https://www.forbes.com/sites/womensmedia/2014/04/28/act-now-to-shrink-the-confidence-gap/?sh=547c2ed05c41

Coffman, K. B., Collis, M., & Kulkarni, L. (2019, November). Whether to Apply. *Harvard Business School Working Paper*(20-062). (Revised June 2021)

Coffman, L. C., Featherstone, C. R., & Kessler, J. B. (2017). Can Social Information Affect What Job you Choose and Keep? *American Economic Journal: Applied Economics*, *9*(1), 96–117.

Dal Bó, E., Finan, F., & Rossi, M. A. (2013, 04). Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service. *The Quarterly Journal of Economics*, *128*(3), 1169-1218. Retrieved from https://doi.org/10.1093/qje/qjt008

Del Carpio, L., & Guadalupe, M. (2021). More Women in Tech? Evidence from a Field Experiment Addressing Social Identity. *Management Science*.

Dover, T. L., Major, B., & Kaiser, C. R. (2016). Members of High-Status Groups are Threatened by Pro-Diversity Organizational Messages. *Journal of Experimental Social Psychology*, *62*, 58–67.

Ekstrom, R. B. (1981). Psychological and Sociological Perspectives on Women's Paid and Unpaid Work Choices. *Advances in Consumer Research*, *08*, 580-584. Retrieved from https://www.acrwebsite.org/volumes/5863/volumes/v08/NA-08

Farber, H. S., Silverman, D., & von Wachter, T. (2016, May). Determinants of Callbacks to Job Applications: An Audit Study. *American Economic Review*, *106*(5), 314-18. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.p20161010 doi: 10.1257/aer.p20161010

Flory, J. A., Leibbrandt, A., & List, J. A. (2015). Do Competitive Workplaces Deter Female Workers? A large-scale Natural Field Experiment on Job Entry Decisions. *The Review of Economic Studies*, *82*(1), 122–155.

Flory, J. A., Leibbrandt, A., Rott, C., & Stoddard, O. (2021). Increasing Workplace Diversity Evidence from a Recruiting Experiment at a Fortune 500 Company. *Journal of Human Resources*, *56*(1), 73–92.

Gaddis, S. M. (2018). An Introduction to Audit Studies in the Social Sciences. In S. M. Gaddis (Ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance* (p. 3-44). Cham, Switzerland: Springer International Publishing.

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality. *Journal of Personality and Social Psychology*, *101*(1), 109–128. Retrieved from https://doi.org/10.1037/a0022530

Gee, L. K. (2019). The More you Know: Information Effects on Job Application Rates in a Large Field Experiment. *Management Science*, *65*(5), 2077–2094.

Goldin, C., & Rouse, C. (2000, September). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review*, *90*(4), 715-741. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.90.4.715 doi: 10.1257/aer.90.4.715

Half, R. (2019, March 19). Survey: 42 Percent Of Job Applicants Don't Meet Skills Requirements, But Companies Are Willing To Train Up. *Cision PR Newswire*. Retrieved from https://www.prnewswire.com/news-releases/survey-42-percent-of-job-applicants-dont-meet-skills-requirements-but-companies-are-willing-to-train-up-300813540.html

Hurst, R. (2022, January 31). *Workplace Backlash? Workforce Diversity, Status Threat, and the Contractionary Effects of Pro-Diversity Claims* (Working Paper). Retrieved from https://ssrn.com/abstract=3789682

Kessler, J. B., Low, C., & Sullivan, C. D. (2019, November). Incentivized Resume Rating: Eliciting Employer Preferences without Deception. *American Economic Review*, *109*(11), 3713-44. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.20181714 doi: 10.1257/aer.20181714

Kline, P., Rose, E. K., & Walters, C. R. (2022). Systemic Discrimination Among Large US Employers. *The Quarterly Journal of Economics*, *137*(4), 1963–2036.

Kuhn, P., Shen, K., & Zhang, S. (2020). Gender-Targeted Job Ads in the Recruitment Process: Facts from a Chinese Job Board. *Journal of Development Economics*, *147*, 102531.

Leibbrandt, A., & List, J. A. (2018, September). *Do Equal Employment Opportunity Statements Backfire? Evidence From A Natural Field Experiment On Job-Entry Decisions* (Working Paper No. 25035). National Bureau of Economic Research. doi: 10.3386/w25035

Linos, E. (2017). More Than Public Service: A Field Experiment on Job Advertisements and Diversity in the Police. *Journal of Public Administration Research and Theory*, *28*(1), 67-85. Retrieved from https://doi.org/10.1093/jopart/mux032 doi: 10.1093/jopart/mux032

Llinares-Insa, L. I., González-Navarro, P., Córdoba-Iñesta, A. I., & Zacarés-González, J. J. (2018). Women's Job Search Competence: A Question of Motivation, Behavior, or Gender, journal=Frontiers in Psychology. , *9*. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00137

Marinescu, I., & Wolthoff, R. (2020). Opening the Black Box of the Matching Function: The Power of Words. *Journal of Labor Economics*, *38*(2), 535-568. Retrieved from https://doi.org/10.1086/705903 doi: 10.1086/705903

Mohr, T. S. (2014). Why Women Don't Apply for Jobs Unless They're 100% Qualified. *Harvard Business Review*, *25*. Retrieved from https://hbr.org/2014/08/why-women-dont-apply-for-jobs-unless-theyre-100-qualified

Pager, D., & Pedulla, D. S. (2015). Race, Self-Selection, and the Job Search Process. *American Journal of Sociology*, *120*(4), 1005–1054. Retrieved from https://doi.org/10.1086/681072

Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (2017). Meta-analysis of Field Experiments Shows no Change in Racial Discrimination in Hiring over Time. *Proceedings of the National Academy of Sciences*, *114*(41), 10870-10875. doi: 10.1073/pnas.1706255114

Reskin, B. F., & Bielby, D. D. (2005). A Sociological Perspective on Gender and Career Outcomes. *The Journal of Economic Perspectives*, *19*(1), 71–86. Retrieved 2023-01-16, from http://www.jstor.org/stable/4134993

Rojas, M. (2021). Dear Female Job Seeker: Apply for the Job, Ignore the 'Qualifications'. *Fast Company*. Retrieved from https://www.fastcompany.com/90661349/dear-female-jobseeker-apply-for-the-job-ignore-the-qualifications

Sakowitz, J. (2018). Uncovering the Gendered Dimensions of Job Hunting. *Stanford News & The Clayman Insitute for Gender Research*. Retrieved from https://gender.stanford.edu/news/uncovering-gendered-dimensions-job-hunting

Samek, A. (2015). A University-Wide Field Experiment on Gender Differences in Job Entry Decisions. *Manuscript, UWM*.

Sandberg, S., & Scovell, N. (2013). *Lean In*. Alfred A. Knopf.

Tockey, D., & Ignatova, M. (2019). *Gender Insights Report: How Women Find Jobs Differently* (Tech. Rep.). LinkedIn Talent Solutions. Retrieved from https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions-lodestone/body/pdf/Gender-Insights-Report.pdf

Wille, L., & Derous, E. (2017). Getting the Words Right: When Wording of Job Ads Affects Ethnic Minorities' Application Decisions. *Management Communication Quarterly*, *31*(4), 533-558. doi: 10.1177/0893318917699885

Zucker, R. (2020, January 27). Is That Stretch Job Right for You? *Harvard Business Review*. Retrieved from https://hbr.org/2020/01/is-that-stretch-job-right-for-you

Chapter 3 Appendix

3.A.1 Survey Instrument

Please help us learn about the job market by completing this 2-minute survey.

We are the Job Connections Project, not the hiring company, and we will never share your data with them.

How likely are you to apply to this job?

- ○ Very Likely
- ○ Likely
- ○ Equally Likely and Unlikely
- ○ Unlikely
- ○ Very Unlikely

Figure 3.2: **Survey page details**

If this job were offered to you right now, how likely would you be to accept it?

- ○ Very Likely
- ○ Likely
- ○ Equally Likely and Unlikely
- ○ Unlikely
- ○ Very Unlikely

Finish the following statement to best describe you:

I think that --- . I meet --- of the qualifications listed and I have --- other relevant skills that were not explicitly asked for in the ad. I think that, were I to apply, the hiring manager --- recognize my full potential and that I would be --- to get an interview.

where the dropdown menus contain:

Finish the following statement to best describe you:

I think that --- . I meet --- of the qualifications listed and I have --- othe r in the ad. I think that, were I to apply, the hiring manager --- be --- to get an interview.

I could do the job well
the job would be a stretch, but I could learn quickly
I'm not sure how I would perform in this role, but I want to try
this job would be a real challenge

Finish the following statement to best describe you:

I think that --- . I meet --- of the qualifications listed and I have --- other relevant skills that were not explicitly asked for in the a at, were I to apply, the hiring manager --- recognize my full potential and that I would be --- to get an interview.

---
all
most
some
few or none

Finish the following statement to best describe you:

I think that --- . I meet --- of the qualifications listed and I have --- other relevant skills that were not explicitly asked for in the ad. I think that, were I to apply, the hiring manager --- recognize my full potential and that I would be --- to get an interview.

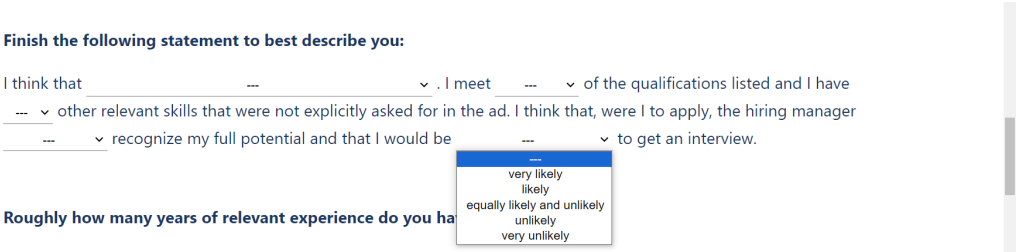
---
many
some
few
no

Finish the following statement to best describe you:

I think that --- . I meet --- of the qualifications listed and I have --- other relevant skills that were not explicitly asked for in the ad. I think that, were I to apply, the hiring manager --- recognize my full potential and that I would be --- to get an interview.

---
would
may or may not
would not

Figure 3.2: **Survey page details (Continued)**



Roughly how many years of relevant experience do you have for this role?

Which of the following best describes your gender identity?

- ○ Man
- ○ Woman
- ○ Non-binary
- ○ Genderqueer or gender fluid
- ○ A gender identity not listed here

Which of the following best describes your race?

- ○ Asian or Pacific Islander
- ○ Black or African American
- ○ Hispanic or Latino
- ○ White or Caucasian
- ○ Multiracial or Biracial
- ○ A race/ethnicity not listed here

What is your highest level of education completed?

- ○ Less than high school
- ○ High school diploma or equivalent
- ○ Associates degree or Some college
- ○ Bachelor's degree (BA, BS)
- ○ Graduate degree (MA, PhD, EdD, JD, etc.)

Figure 3.2: **Survey page details (Continued)**

What is your current employment status?

- ○ Employed (full-time)
- ○ Employed (part-time)
- ○ Unemployed
- ○ On temporary leave

Indicate the extent to which you agree or disagree with the following statements:

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
I think this company will be more lenient than others when reviewing candidates.	○	○	○	○	○
I feel like I will be wasting my time if I apply for this job.	○	○	○	○	○
I want to put myself out there and apply to more 'stretch' roles.	○	○	○	○	○
I think everyone should apply to jobs when they meet most (not necessarily all) of the listed qualifications.	○	○	○	○	○
I think companies are realizing that candidates have diverse backgrounds, so sticking to required qualifications is becoming less common.	○	○	○	○	○

Thank you! We are grateful for the information that you have shared with us. It will help us understand the labor market and continue to support people looking for work.

Apply now on company website

No longer interested, show me similar jobs

Figure 3.2: **Survey page details (Continued)**