Essays on Individual and Collective Decision Making

by

Nathan J. Mather

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2023

Doctoral Committee:

Professor James R. Hines, Chair
Assistant Professor Ash Craig
Professor Tanya Rosenblat
Professor Joel Slemrod

Nathan Mather

njmather@umich.edu

ORCID iD 0000-0001-8946-6197

“To my partner, Faith and my son, Booker”

ACKNOWLEDGEMENTS

I would like to first thank my academic advisors and mentors. Thank you to Professor Jim Hines for his extensive support and mentorship throughout the job market. In addition to supporting my non-conventional research topic, he helped me navigate the norms of the academic job market and successfully guided me through the difficult and stressful process of finding a job. I am also grateful to Assistant Professor Ash Craig for his support. Especially for taking the time to organize weekly research discussions with all of his advisees. These were an excellent learning opportunity. Thank you to Professor Joel Slemrod for his many insightful comments on my work. I always left your office with exciting new ideas to pursue. Thank you to Professor Tanya Rosenblat for her comments on my work and for agreeing to serve on my committee at the last minute and without knowing me very well. Graduate school is full of surprises, and I am grateful she was willing to step in when I needed the support of a professor outside my department.

In addition to my formal committee, there are several others who offered me significant academic mentorship on my path here. Thank you to my coauthor Professor Julian Betts at the University of California San Diego. He has been an unwavering supporter of our project since Mike, Tanner, and I came to him with the idea, and I have learned so much from our frequent meetings together. Thank you to Professor Jim Adams who mentored me in my role as an instructor during my graduate studies. While formally the dissertation does not require teaching, informally much of what I learned working with Professor Adams has made me a better writer, communicator, and researcher. I am also deeply grateful to Professor Robert Meyer, Dr. Andrew Rice, and Dr. Michael Christian at Education Analytics, who provided excellent support as I considered and applied for graduate programs. Andrew also provided invaluable advice on the job market that helped me to think about the career that is best suited for me and my family. I would also like to thank the entire research team at Education Analytics. I would not have been able to complete this dissertation without the skills I gained during my time as a research analyst. I would also like to thank Professor Nils Ringe at the University of Wisconsin for supporting my academic growth as an undergraduate student interested in research.

I would also like to thank my fellow graduate students who have supported me. Thank you to my coauthors Assistant Professor Michael Ricks and Tanner Eastmond. Working with the both

of you has truly been a pleasure. Thank you to Assistant Professor Dylan Moore who read an early and, to be honest, very sloppy version of my job market paper. He provided me with early and incredibly valuable feedback (including catching a mistake in my model) that allowed me to get my project to a place I was happy with. Thank you to Stephanie Karol, Caitlin Hagarty, Leticia Juarez, and Aaron Kaye for working together to review our papers and presentations as well as their continued support throughout the job market. I would also like to thank my office mates, Anirudh Yadav, Russell Morton, and Max Huppertz for the many conversations we had about economics (and not economics).

While I will not list them all out here, I am also grateful to the many friends I have made during my time as a PhD Student. Graduate school can be a trying time. During those times, it was thinking of all the wonderful people I have met here that helped me regain confidence in my decision to pursue a PhD.

Thank you to the Graduate Employees' Organization (GEO) for ensuring all graduate workers, including myself, have healthcare. While I considered this to be a small bonus when choosing a program, I could have been financially ruined by the medical issues I was surprised with during my time as a student. I am also thankful for all the GEO members who continue to fight for a fair wage for graduate workers.

Finally, thank you to my family. Words cannot describe my gratitude to my parents who supported me and encouraged me to learn and excel in my education throughout my entire life. I am especially grateful that they were able to encourage me to learn and commit myself to my goals without ever pressuring me to do more than I needed or wanted. Thank you to my partner, Faith, who supported me in so many ways throughout this entire program. From literal financial support, to listening to me drone about research problems on our walks. Thank you to my son Booker. Even before he was born, my drive to give him a stable and secure life helped give me the final push I needed to be ready for the job market and complete my dissertation.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

This dissertation contains three essays on individual and collective decision making. Understanding and modeling how individuals make choices is at the heart of economics. In chapter I, I empirically estimate the degree to which the marginal utility of income changes across income groups. The estimation is based on survey responses indicating willingness to pay to avoid unpleasant experiences and relies on the assumption that the associated disutility is equal on average across income groups. This assumption implies that any differences in average willingness to pay for relief are entirely driven by differences in marginal utilities of income. The results suggest that marginal utility of income is constant across income groups, implying that (cardinal) utility is roughly linear in dollars. While this paper is primarily about how individuals make choices, a better understanding of individual preferences and motivation can inform collective decisions, where the preferences or needs of different people must be considered simultaneously.

How to provide effective information to policymaker's who are making collective decisions is the focus of chapter II where we show how to get unbiased estimates of welfare from observational outcomes with heterogeneous populations. 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 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.

While chapter II focuses on how to aid in collective decision making, Chapter III focuses on understanding how, in real life, collective decisions are made. I examine how preference formation may directly lead to state fiscal policy interdependence. I start by laying out a formal model for state fiscal interdependence. The model is built on two core ideas. First, voters look at “similar" states via news coverage to determine what a normal level of public spending is. Second, governments respond to these shifting preferences by maximizing the probability of reelection. The model informs an empirical analysis which uses state newspaper articles to form a new metric of state inter-connectivity. This metric is compared against established metrics used in the literature.

CHAPTER I

Is Utility Concave?

1.0 Abstract

This paper empirically estimates the degree to which the marginal utility of income changes across income groups. The estimation is based on survey responses indicating willingness to pay to avoid unpleasant experiences and relies on the assumption that the associated disutility is equal on average across income groups. This assumption implies that any differences in average willingness to pay for relief are entirely driven by differences in marginal utilities of income. The results suggest that marginal utility of income is constant across income groups, implying that (cardinal) utility is roughly linear in dollars.

1.1. Introduction

Diminishing marginal utility is a foundational idea in economics stretching back to 1854 (*Gossen*, 1854; *Marshall*, 1890). According to this idea, as a person's income grows, the value they gain from an additional dollar should fall. This matches with the normative views many have of what is fair or just. According to most people, a dollar seems to mean more to a family with less income (*Kimball, Ohtake, Reck, Tsutsui, and Zhang*, 2015). When combined with a model of utility maximization, however, diminishing marginal utility is about more than normative perceptions of fairness. Utility must now describe behavior as well, and diminishing marginal utility implies that higher income folks will pay more for the same utility gain. Whether or not pacemaker's believe higher income folks deserve a dollar more or less on the margin does not need to coincide with individual consumer behavior. People's perceptions of their own money and the relative value individuals place on goods they purchase dictate their actions. While people with higher income might buy things lower income folks see as extravagant or wasteful, it is the value the higher income folks place on those goods that dictates their own behavior, and there is little to no evidence on the concavity of utility over income that is dictated by that behavior.

We cannot observe individual utility. This is the key problem that makes clear evidence on how marginal utility changes over income elusive. I overcome this problem by finding a particular class of goods that enable estimating the average concavity of utility over characteristics like income. For now, let's just call this special good a widget. The key characteristic of this widget is that the welfare benefit of receiving a widget is the same, on average, across income. That is, utility from widgets can vary between individuals but is independent from income. Now if we see that low-income Americans are willing to pay $X, and high-income Americans are willing to pay $Y for a widget, we know that they value X and Y dollars the same as well. This means the ratio of the average utility of money between high-income and low-income Americans is Y/X. That ratio identifies the average concavity of total utility. More generally, any differences in the average willingness to pay for relief over income is entirely driven by differences in marginal utilities of income since the average marginal utility for the good is the same on average.

While this may seem surprisingly simple, the difficulty lies in finding real goods like the widget described. The marginal utility of any given good almost always depends on the quantity already consumed and the consumption of compliments and substitutes[1]. To illustrate how difficult this really is, consider the following example. Suppose income does not change one's taste for chocolate. Of course it could, but suppose it does not. In this special case, it may seem that chocolate would work well as the widget from the example above, but it is unlikely to satisfy the necessary assumption. The marginal utility of a chocolate bar still depends on how much chocolate I already have as well as how many other desserts, food, or any compliments or substitutes to chocolate that I consume. A change in income relaxes a consumer's budget constraint. With a relaxed budget constraint, consumers will change the quantity and quality of compliment and substitute goods they consume, and, by extension, this changes the marginal utility of chocolate. Even in the exceptional case where income does not impact the taste of chocolate, the marginal utility of chocolate is unlikely to be independent from income.

Typical consumption goods generally face the same problem as the above example; so, I field a survey that elicits the willingness to pay for relief from common minor pains. These questions are carefully chosen to plausibly satisfy the assumption that the marginal utility of pain relief is actually independent from income. Consider the following example from the survey.

> Imagine you bump your shin badly on a hard edge, for example, on the edge of a glass coffee table. What is the most you would pay in U.S. dollars to completely and immediately eliminate any pain caused by the situation described, as if the event never happened?

[1] This refers to compliments and substitutes in utility when utility is cardinal and is equivalent to a cross partial in cardinal utility. This is how the terms were originally conceptualized prior to the ordinal revolution (*Auspitz and Lieben*, 1889; *Moscati*, 2018)

Is the marginal utility of this hypothetical good, relief from the pain of hitting one's shin, independent from income? First, consider if there are any compliment or substitute goods that would make bumping one's shin more or less desirable. There is nothing available to buy to immediately relieve the pain. Any medications will not be accessible or take affect before the pain has subsided naturally. Given the lack of alternative treatments, this means that no one, regardless of income, is already consuming some quantity of pain relief because it does not actually exist. Everyone is considering a change from zero relief after bumping their shin, to total relief. The marginal utility is not dependent on a person's consumption bundle, and so the only difference in marginal utility comes down to personal pain tolerance. While pain tolerance likely varies from person to person, it is arguably uncorrelated with income. I fielded a survey asking four open ended questions like the one above as well as three binary yes or no questions with a price randomly selected. The full survey can be seen in appendix A.1.

Surprisingly, the results indicate that the willingness to pay for relief from these situations does not change significantly with income. This is an unexpected and interesting result on its face, but the exact interpretation depends on what assumptions we make about utility and people's behavior. First, if we only assume that the utility from relief in these situations is equal on average over income, then the results imply that all sampled income levels will pay roughly the same amount per util. Richer people do not simply pay more for the same or less utility value. If we additionally assume that people are maximizing utility, then the results indicate the marginal utility of income does not change significantly across income, or that the average cardinal utility is roughly linear over income.

While either conclusion is at odds with typical economic assumptions about utility, it seems supported by some observed market behavior. For example, product marketing towards the wealthy. Luxury goods do not distinguish themselves as small improvements for a high marginal cost to isolate wealthy buyers. Popular writing and research on luxury marketing suggests that consumers need to be convinced of real significant value through better quality, craftsmanship, or taste (*Atwal and Williams*, 2017; *Kapferer*, 2014; *Vela*, 2019). What people are willing to pay for a fixed amount of utility does not obviously increase with income. While just one example, this demonstrates a pattern of behavior consistent with my findings.

The main motivation for this paper and main takeaway from the surprising results are a better understanding of human decision making and preferences. A clear understanding of how people behave and how they value and trade-off goods is at the heart of economic analysis. While the results may create as many questions as they answer, this is a significant step toward a better understanding and model of human behavior. Beyond this fundamental economic motivation, the results inform our interpretation of models of normative policy assessment. A utilitarian policy-maker, who values only the sum total of utility, would have linear welfare weights. By extension, any policymaker that favors redistribution is not a pure utilitarian and must care about the distribution of utility, equity, directly. It is also possible that policymakers value redistribution because the collective well-being that they value is simply different than the utility individuals maximize with their decisions. While entirely possible, this would be a big departure from convention in normative economics, which typically assumes individual decision utility reflects the collective goals of policymakers.

This paper's model and my discussion of the results build on the assumption that the marginal utility of pain relief in the specific survey questions[2] I pose is, on average, equal across income. Interpersonal utility comparisons are normative and so, of course, this assumed equality need not hold. However, regardless of the reader's personal views on interpersonal utility comparisons, these results should be interesting. First, because this assumption reflects a common or conventional way to think about utility. Many, including myself, would argue the marginal utility from reliving these minor pains should be considered equal, and so understanding the implications of that common assumption is important. Second, as I will show, the results imply a marginal utility of income that is orders of magnitude off from common assumptions like log utility. In order to maintain these common assumptions, and write off the results of my survey, the value of pain relief would not only need to be different across income, it would need to be orders of magnitude different (around 12 times worse for those with an income below $25k per year than those above $100k per year). This could be the case, again, interpersonal utility comparisons are normative, but I believe it would involve a considerable and important change in the way many economists think about utility. So, regardless of an individual's views on the necessary normative assumptions in the paper, the results improve our understanding of utility and human decision making.

The rest of the paper precedes as follows. Before diving into the analysis, section 1.2, "Measuring Utility", discusses how economists use the term utility, why cardinal preferences and interpersonal comparisons are useful, and some of the background literature on measuring utility. While I hope this section proves helpful and interesting, it can be skipped without missing any details about my model or empirical strategy. Section 1.3, "Theoretical Foundation", outlines the main assumptions of the model and lays the theoretical groundwork for the empirical model. Section 1.4, "Empirical Model", outlines the empirical model used to recover estimates for the concavity of utility. Section 1.5, "Survey Data", outlines the data from the survey, the survey method, and summary statistics about the population. Section 1.6, "Survey Analysis", details the analysis and results from my survey while section 1.7, "Discussion of Results", discusses the implications of the findings. Finally, section 1.8, "Conclusion", provides some final thoughts.

[2] As I explain later, it is not the case that pain relief in general must be equal. Certain ailments likely have costs tightly correlated with income.

1.2. Measuring Utility

Before understanding why we would want to measure utility, it's important to understand what I mean by utility in this paper. Despite playing a ubiquitous role in economic theory, the definition of utility and its use within that theory varies. Most economics theory models individual behavior as the result of utility maximization. When utility is defined as ordinal numbers corresponding to the preference ranking of a rational individual (with complete and transitive preferences) than utility maximization is true by definition. This is a great strength of ordinal utility as a flexible and adaptable modeling tool. It doesn't matter why people choose something, if they do in fact choose it consistently, we can assign it more ordinal utility. Despite ordinal utility's flexibility, normative economics frequently requires the additional assumption that utility correspond to an individual's welfare or well-being in some consistent and meaningful way. We can also see this pop up with the way many economists talk about people "seeking" or "obtaining" utility. For this to make any sense, that utility needs to be something meaningful, and, in my case, comparable across people. The following hypothetical scenario will help to show what options economists have with only ordinal utility, and why they often have room for improvement.

A city has enough money set aside in their budget to either expand public transit or widen the roads to allow more cars through at a time. Of course, which policy they should implement is partially a normative question with a wide range of potential considerations. Despite being a question of ethics, economics can still objectively aid in decision making by laying out the outcomes from each policy in a way that allows policymakers to discern the option that best coincides with their ethical normative goals. While economic work can take a general instrumental rationality approach and consider ethical concerns like racial disparities, inequality, justice, rights, or fairness, it often focuses on consequentialist and welfarist approaches to policy assessment that rely on aggregating preferences.

Economists have tried to find a helpful way to aggregate individual ordinal preferences for social choices. Sticking to ordinal preferences has practical appeal since preference rankings are directly observable through actions and require assuming only that people have complete and transitive preferences[3]. *Arrow* (1950), however, showed that aggregating ordinal preferences is not possible without the social choice mechanism having properties generally considered undesirable. This motivated even him to look for a way to think about cardinal utility and interpersonal comparisons (*Arrow*, 1978). Beyond Arrow showing the necessity of cardinality for social choice, ordinal preferences simply do not capture all of the relevant information for many normative decisions.

Returning to the above example about expanding public transit or widening roads, suppose

[3] Although these are not necessarily weak assumptions

the only information we have is that more people prefer widening roads. With only two options Arrow's impossibility theorem does not bind, and I expect many policymakers would propose widening the roads to reflect the majority of preferences (*Arrow*, 1950). However, suppose the folks favoring widening the roads favor it because it shaves a few minutes off their morning commute, and they do not use public transit at all. Now on the other hand, if the people in favor of expanding public transit would now be able to commute to more lucrative job opportunities or save considerable time on their commutes, reasonable policymakers may change their decision in favor of public transit expansion. The level of benefits going to each person in addition to their ranking often provides use-full, relevant information[4].

The need for more information than ordinal preferences provide is consistent with a welfarist approach to policy assessment. A welfarist policymaker has some weight they place on different people's utility and summing those weighted values up will lead to the outcome that coincides with their values. Doing this is, of course, easier said than done. There are many potential outcomes to consider and weigh against each other. Implicitly, the information about commuting times and job opportunities is conveying something about the intensity of the benefits, but what about other concerns like carbon emissions? How can policymakers compare or aggregate the impact of all the various relevant outcomes?

One way to collapse these concerns into a single dimension is by using willingness to pay (WTP) [5]. Willingness to pay has a clear theoretical connection to an individual's decision utility since trading money for a good implies one prefers the money less than or equal to the good traded for it. This is also consistent with the economic practice of using people's own preferences to infer what increases their well-being. While the connection to an individual's decision utility is clear, how to handle interpersonal comparisons is not.

Rather than addressing the tricky issue of interpersonal comparisons, economic work often resorts to efficiency arguments that use willingness to pay but rely only on an ordinal conception of utility for aggregation. This practice stems from the idea put forth by *Hicks* (1940) and *Kaldor* (1939) who proposed measuring economic efficiency as the sum of aggregate real income. The core idea of their argument is that, of course policy creates winners and losers, but if the winners can compensate the losers, the size of the economic pie has increased, and everyone can be made better off. This idea has become widely accepted in economic analysis and modern practitioners often use such arguments to justify using the sum of consumer and producer surplus as a measure of economic efficiency. The practice is ubiquitous, but some examples include industrial organization's analysis of consumer surplus in merger litigation (*Glick*, 2018; *Wilson*, 2019), cost benefit analysis, or theoretical policy arguments for things like price gouging (*Zwolinski*, 2008).

[4] *Pearce* (2021) explained the insufficiency of ordinal preferences in a similar way that was very helpful

[5] compensated and equivalent variation or consumer surplus

While an ethics free[6] efficiency is appealing, the idea does not hold up in a setting with heterogeneous preferences. *Samuelson* (1950) shows how even in the case where everyone is truly made better off, an outcome cannot be said to be efficient unless it completely expands the utility possibility frontier for every heterogeneous group. The logic here is that although everyone may be better off than the status quo after implementing a policy, that policy might also make redistributing utility to a particular group more difficult. If you are a policymaker that wants to redistribute to that group, the policy is a bad idea despite the Pareto gain relative to the status quo. This can be illustrated by a policy that shifts the slope of the utility possibility frontier between two groups and creates an intersection rather than efficiently expanding the utility possibility frontier. In his own words, *Samuelson* (1950, pg 10) says that without comparing an “Infinite number of points, no acceptable definition of an increase in potential real income can be devised at the non-ethical level of the new welfare economics."

Samuelson's critique applies even when allowing for theoretically cost-less lump sum transfers, but as he continues to explain, movement along a utility possibility frontier would require “an ideally perfect and unattainable system of absolutely lump-sum taxes or subsidies (*Samuelson*, 1950, pg. 18)". Lump sum transfers to increase equity simply do not exist. Rather than mapping out a utility possibility frontier, then, an efficiency approach would require mapping out utility feasibility frontiers. On this front, there has been some interesting work. *Coate* (2000) lays out a theoretical foundation that *Hendren* (2020) follows up on. The core idea being that, rather than use lump sum transfers as the benchmark, a given policy needs to be compared to redistribution through other government levers. *Hendren* (2020) looks at the cost of redistribution through the income tax code as a benchmark for other policies. While this is a promising improvement on the lump sum transfer arguments, it does not absolve the need for interpersonal comparisons. This approach is effective for eliminating re-distributive policies when redistribution could be more efficiently done though the income tax code. It eliminates some policies from consideration, but even a perfect efficiency measure can only partially order outcomes by eliminating Pareto dominated policies. A policymaker may still face a difficult and complex decision among efficient outcomes and objective measures can facilitate a decision that is in line with the policymaker's goals.

A slightly different type of efficiency argument is also frequently employed in an attempt to separate economic analysis from interpersonal comparisons. Introductory economics textbooks frequently present policy analysis as a balance between equity and efficiency. Betsey Stevenson and Justin Wolfers' book lays out the idea clearly.

> One argument for focusing on efficiency is that whenever economic surplus rises it's

[6] It's not really ethics free since Pareto efficiency is itself a fairly strong consequentialist ethical assumption, but it avoids normative interpersonal comparisons

> possible for those who benefit to compensate those who were harmed, and to do so in a way that ensures everyone's better off...In reality, it's rare for new policies to compensate the people they harm. Thus, the argument that it's possible to make everyone better off is just that, a possibility...consequently, real-world policy debates are rarely just about efficiency. They also focus on equity (*Stevenson and Wolfers*, 2019, Section 7.1).

This logic is helpful, but it also requires making interpersonal comparisons. If we are not actually achieving a Pareto gain, then we are not avoiding interpersonal comparisons or sticking to ordinal preferences. Weighing a gain in consumer surplus against a change in economic inequality, for example, requires assuming that surplus corresponds to welfare or well-being in a way that can be compared across people and can be compared against the distribution of that well-being, equity. This is a fine and common assumption to make, but it is not an ordinal efficiency argument. It is an interpersonal comparison using cardinal utility.

Hendren and Sprung-Keyser (2020) provide a helpful framework for thinking about this type of trade-off. They propose measuring the marginal value of public funds for individual policies and using the following logic for policy comparisons: "Given two policies, A and B, suppose $MVPF_A$ = 2 and $MVPF_B = 1$. Then one prefers more spending on policy A financed by less spending on policy B if and only if one prefers giving $2 to policy A beneficiaries over giving $1 to policy B beneficiaries." As we discuss in *Eastmond, Mather, Ricks, and Betts* (2022), this logic works well if policies impact homogeneous groups but places a high informational burden on policymakers for heterogeneous populations. Most importantly, it again assumes utility is cardinal and corresponds to individual well-being. Understanding the concavity of utility under this assumption will allow us to better conceptualize these trade-offs.

The above discussion is meant to make it clear that interpersonal comparisons are an unavoidable aspect of collective decision making in most practical applications. I build on the typical economic framework that assumes individual behavior is the result of maximizing cardinal utility where utility corresponds to individual preferences and individual well-being. What does this assumption imply about how utility changes with income or across individuals? How does this assumption fit into the re-distributive preferences of policymakers? Are policymakers true utilitarians under this definition of utility, or do they consider the total distribution of utility as well? Our understanding of these questions and by extension the core assumption being made would be greatly improved if we knew the concavity of utility over income. Before proceeding to the next section where I state and explain the assumptions needed to measure that concavity, it is important to draw a clear distinction between this paper and related literatures that propose alternative ideas for the concavity of utility and it's measurability.

One such alternative is measuring the concavity of utility using risk aversion (*Becker, DeGroot,* *and Marschak*, 1963; *Davidson, Suppes, and Siegel*, 1957; *Dolbear*, 1963; *Mosteller and Nogee*, 1951). This approach is built on the idea of expected utility theory laid out by *Von Neumann and Morgenstern* (1947) who originally saw "themselves to be contributing to an area of research where not much progress had been made, namely, the measurement of utility" (*Risse*, 2002, p. 563). However, in the immediate aftermath of the paper "[t]here was widespread, but not universal, agreement that von Neumann and Morgenstern's expected utility theory has little or no welfare significance" (*Weymark*, 2005, p. 533). Despite this widespread agreement, two highly influential papers by *Harsanyi* (1953,5) use von Neumann-Morgenstern utility for his axiomization of utility. A critique of this use was presented by *Sen* (1976) who pointed out that, although it is often thought of as a cardinal theory, Von Neumann-Morgenstern expected utility is actually an ordinal theory. This critique was later clarified by *Weymark* (1991, 2005). The key intuition is that the linear utility scale used by Von Neumann-Morgenstern is chosen out of convenience, but other forms could just as effectively describe risk taking behavior. This intuition is explained clearly by *Arrow* (2012, p. 10)

> This theorem does not, as far as I can see, give any special ethical significance to the particular utility scale found. For instead of using the utility scale found by von Neumann and Morgenstern, we could use the square of that scale; then behavior is described by saying that the individual seeks to maximize the expected value of the square root of his utility. This is not to deny the usefulness of the von Neumann-Morgenstern theorem; what it does say is that among the many different ways of assigning a utility indicator to the preferences among alternative probability distributions, there is one method (more precisely, a whole set of methods which are linear transforms of each other) which has the property of stating the laws of rational behavior in a particularly convenient way. This is a very useful matter from the point of view of developing the descriptive economic theory of behavior in the presence of random events, but it has nothing to do with welfare considerations, particularly if we are interested primarily in making a social choice among alternative policies in which no random elements enter. To say otherwise would be to assert that the distribution of the social income is to be governed by the tastes of individuals for gambling.

In addition to this fundamental point, later work calls into question how much behavior aligns with von Neumann-Morgenstern utility. This further differentiates risk aversion from the concavity of a cardinal and interpersonally comparable utility function (*Becker, DeGroot, and Marschak*, 1964; *Grether and Plott*, 1979; *Kahneman and Tversky*, 2013; *Karmarkar*, 1974; *MacCrimmon*, 1965; *MacCrimmon and Larsson*, 1979; *Machina*, 1982; *Moscati*, 2018; *Slovic and Lichtenstein*, 1968; *Tversky*, 1969). Both points show that, while there is a literature measuring risk aversion (which can be considered the concavity of Von Neumann-Morgenstern utility), it is not necessarily describing the concavity of a cardinal interpersonal representation of utility as I am doing in this paper.

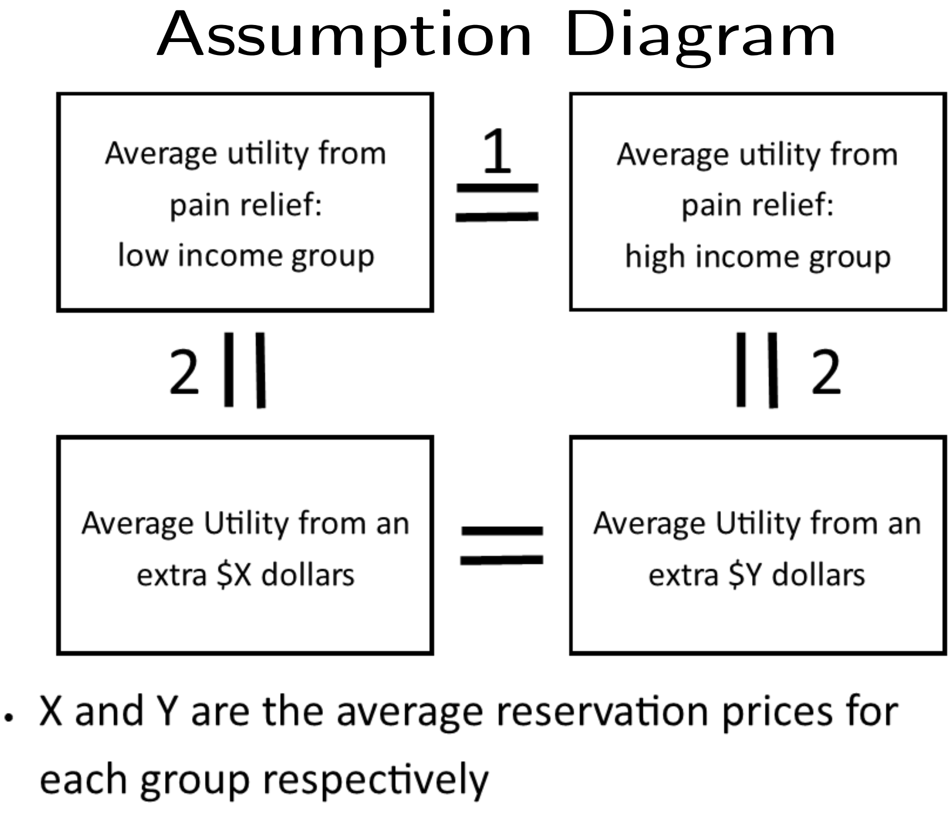
Another literature that is related, but distinct from my approach is the subjective well-being literature. This is a vast and interesting literature (*Baker and Ricciardi*, 2014; *Diener and Biswas-Diener*, 2002; *Diener, Sandvik, Seidlitz, and Diener*, 1993; *Diener, Wirtz, Tov, Kim-Prieto, Choi, Oishi, and Biswas-Diener*, 2010; *Diener, Tay, and Oishi*, 2013; *Diener, Oishi, and Tay*, 2018b; *Dunn and Norton*, 2014; *Dunn, Aknin, and Norton*, 2014; *Furnham and Argyle*, 1998; *Gilovich, Kumar, and Jampol*, 2015; *Haushofer, Reisinger, and Shapiro*, 2015; *Jebb, Tay, Diener, and Oishi*, 2018; *Kahneman, Diener, and Schwarz*, 1999; *Kimball et al.*, 2015; *Layard, Mayraz, and Nickell*, 2008; *Lucas and Schimmack*, 2009; *Luttmer*, 2005; *Ng*, 2013; *Oishi and Diener*, 2001; *Stevenson and Wolfers*, 2013; *Weiman, Knabe, and Schob*, 2015). For a thorough review, readers should see *Diener, Lucas, and Oishi* (2018a). As I will discuss later on, subjective well-being measures may coincide with utility for certain substantive definitions of utility (happiness, pleasure, or joy), but how happy or satisfied people say they are (or actually are) and how they treat money and make purchasing decisions need not be the same. It might be that they are the same (in many ways if these ideas coincide it would be very convenient for welfare economics), but there is no reason that they must be the same. The subjective well-being literature, then, provides an interesting comparison but it is asking and answering a fundamentally different question.

The differences between risk aversion and subjective well-being should only become more clear as I outline the theoretical foundations for this paper in the following section.

1.3. Theoretical Foundation

Key considerations like, “what is utility", are not clearly communicable with just mathematical equations, but are a vital aspect of measuring the concavity of utility. This section covers the two assumptions needed to identify the concavity of utility and maps out the theoretical and philosophical groundwork for the empirical model to come. This conceptual map can be seen in the diagram in figure 1.1 which walks through a simple two income case. As I explained in the introduction, if the average marginal utility of pain relief is the same for the two income groups, and each is willing to pay \$X and \$Y dollars respectively, then the ratio of the marginal utility of money between the two income groups is Y/X. Each of these equalities linking the boxes, however, requires an assumption.

Figure 1.1: Conceptual Map



Assumption Diagram

Average utility from pain relief: low income group

Average utility from pain relief: high income group

Average Utility from an extra $X dollars

Average Utility from an extra $Y dollars

- X and Y are the average reservation prices for each group respectively

The first assumption needed for identification, and corresponding to equality 1 in the diagram, is the following.

Assumption 1. *The Marginal utility of the pain relief described in the survey questions is independent of income.*

In other words, the average utility from pain relief in the survey scenarios is equal across income groups. Recall that the scenarios are things like bumping one's shin badly on a hard edge (full survey available here). There are no clear compliments or substitutes that make this experience better or worse and so different consumption bundles across income groups do not need to be considered to infer the marginal benefit of relief from this pain. We need only consider the experience itself.

Irving Fisher recognized that eliminating cross partials in cardinal utility, having no compliments or substitutes, facilitates identifying utility from revealed preferences[7] (*Fisher*, 1927). Instead of finding a particular good with this trait, however, he proposed using aggregates. The

[7] For an interesting and more thorough overview of research on measuring utility see (*Dimand*, 2019) and (*Moscati*, 2018). I focus on Fisher's approach as it most closely mirrors my own.

assumption being that “the utility derived from the consumption of each commodity group depends only upon the quantity of that commodity group that is consumed" (*Morgan*, 1945, pg 135). For example, the marginal utility of food must not depend on the type of housing or entertainment a person has. This simplifies the interpersonal comparison since we need only consider the marginal value of additional food, but higher-income people consume more and higher quality food; so, the marginal benefit of a bit more is conceivably different even without substitutes or compliments. To overcome this problem, Fisher proposes using different locations with a different price vector where people consume the same amount of food as the low-income group, but the same amount of housing as the high-income group. *Morgan* (1945) attempts to implement this method[8].

Using this method, however, requires solving a difficult if not impossible index number problem to determine equivalent food rations across regions with different populations and price vectors. What is the equivalent consumption of quality adjusted food for an American in dollars and an English person in pounds? If their preferences are not identical, the ability to purchase a given consumption basket in either country does not make the utility value of a given dollar expenditure on food equivalent to the average person in the two countries. This is a general problem with index numbers or measures of inflation under heterogeneous preferences (*Samuelson*, 1950; *Samuelson and Swamy*, 1974).

For the goods in my survey, like relief from bumping one's shin on a hard edge, there is no need to aggregate to avoid compliments and substitutes. Moreover, the marginal utility of relief is arguably equal across income levels without an intermediary comparison group. This avoids the index number problem required in Fisher's approach. I have stated that marginal utility is arguably equal across income levels, but what does that actually mean? What is utility? An advantage of my survey questions is that they satisfy assumption 1 for a wide range of philosophical definitions of utility.

Utility is a concept that is frequently used in economics and philosophy, but its definition is not consistent, and it is often used in differing and conflicting ways. Assumption 1 requires a cardinal definition where the experiences of two different people, or the average of two groups to be more exact, can be considered equal; but even among cardinal definitions of utility, there are a wide variety of views about what exactly utility is.

Hausman and McPherson (2006) provide a clarifying framework for grouping theories of well-being. Substantive theories say what things are inherently good. For example, happiness or pleasure could be considered the actual meaning of utility. The substantive approach fits with the work of utilitarian philosophers Jeremy Bentham, John Stuart Mill, and Henry Sidgwick (*Driver*,

[8] Or vise versa. The consumer with alternative prices needs to act as a link between any two broad class of goods

2014). Assumption 1 is likely satisfied for these substantive theories of utility. It is reasonable to think, for example, that two people hitting their shins are experiencing the same pain or lack of pleasure.

Formal theories instead specify how to find out what is good, but not what is inherently good (*Hausman and McPherson*, 2006). In economics well-being is often considered to be the satisfaction of preferences and utility is the extent to which those preferences are satisfied. This is a formal theory that says what people want must give them utility. Irving Fisher actually preferred terms like "wantability" to avoid conflation of what economics is considering utility with the substantive theories of "Benthom and his school" (*Fisher*, 1927). However, some moral philosophers have accepted this as their preferred understanding of utility (*Hare*, 1981). If utility is the satisfaction of preferences how can we compare that across people?

Harsanyi (1955,8) proposed a way to think about how to compare the intensity of preferences across people using "extended preferences". This idea was even considered by *Arrow* (1978) as a way past the independence of irrelevant alternatives assumption in his impossibility theorem (*Arrow*, 1950). The idea proposes a thought experiment for interpersonal comparisons of utility. To compare the utility cost of two people hitting their shins on a coffee table we must imagine being each person and hitting our shin and compare those experiences. *Harsanyi* (1986) explains further that you must imagine yourself with that person's preferences. *MacKay* (1986) calls this the mental shoehorn trick (*Hausman and McPherson*, 2006).

While this seems to be the dominant theory for rationalizing interpersonal comparisons of preferences, it is controversial. *Arrow* (2012, pg 115) says of it, "The principle of extended sympathy as a basis for interpersonal comparisons seems basic to many of the welfare judgments made in ordinary practice. But it is not easy to see how to construct a theory of social choice from this principle." Moreover, theoretical work has considered it unsatisfactory (*Hausman*, 1995; *MacKay*, 1986). The advantage of the particular assumption in this paper, however, is that we need not be capable of comparing every possible situation, only the ones in my survey which have been specifically chosen because the comparisons are relatively simple to make. Imagining oneself as a high- or low-income person hitting their shin seems to fall into the category of basic welfare judgments arrow referred to. Since there are no compliments and substitutes, to compare the utility from relief across income we need only consider how pain sensitivity or preferences for pain relief differ across income.

The medical literature provides some helpful references for considering how painful experiences would change with income. First, what is pain? The International Association for the Study of Pain (IASP) defines pain as "An unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage" (*Aydede*, 2019, Section 1.2). Under this definition, people have "epistemic authority with respect to their pain: they seem to be incorrigible, or even infallible, about their pains and pain reports" (*Aydede*, 2019, Section 1.2). This means pain cannot be measured with a medical instrument, it must be elicited from the person experiencing it. The pain sensitivity questionnaire (PSQ) is a validated measure of pain sensitivity. First, respondents were asked to fill out the survey questions. Then, their responses were compared with questions asked during actual painful experiences (heat, cold, pressure, and pinprick) (*Ruscheweyh, Marziniak, Stumpenhorst, Reinholz, and Knecht*, 2009). These questions are used both for the willingness to pay questions, and to construct a control for general pain sensitivity in my survey.

While the very nature of pain makes it difficult to measure, the medical literature seems to suggest that, if anything, lower income people may be less tolerant to pain. *Miljković, Stipčić, Braš, Đorđević, Brajković, Hayward, Pavić, Kolčić, and Polašek* (2014) shows that participants with a lower household possession index, suggesting lower income, are more sensitive to pain. Research has also shown that chronic pain increases pain sensitivity (*Ruscheweyh et al.*, 2012). These results align with the idea that higher-income folks may have access to goods that decrease pain sensitivity. While nothing can immediately relieve the pain from bumping my shin, perhaps the knowledge that I can reward myself for enduring that pain with an expensive treat makes the experience less difficult. Moreover, the research on chronic pain suggests that perhaps high- and low-income folks are not coming from the same starting quantity of daily pain. If there are diminishing returns to pain relief, this would impact that benefit of relief from a marginal painful event.

If these relationships are true, estimates without controls will be biased since low-income folks will be getting more of a benefit from pain relief. Despite the findings described, I do not find the medical literature to be conclusive. The chronic pain research, for example, suffers from selection bias. If 10 people actually have a chronic pain condition, the people that are least tolerant to pain are the ones who will go and get a diagnosis. This is not really addressed in the medical studies I have read. This may be because it is not important for their purposes. Medical studies are interested in describing attributes of patients with a chronic pain diagnosis, and this is true even if it is because of selection. Additionally, studies assessing pain tolerance, like how long you can endure holding your hand in ice water, do not incentivize performance. Because of this, variance in performance may be attributable to variance in commitment levels to the study or a desire to look tough as much as differences in the epistemic experience (*Stephens and Robertson*, 2020). While the points above cast some doubt on how applicable the medical literature results are to this particular setting, if these studies are measuring pain sensitivity differences, then we can (and I do) control for pain sensitivity using the validated survey questions in the pain sensitivity questionnaire.

Assumption 1 is the only assumption we need to show that higher income people do not pay significantly more per util in all cases (because in this case they do not). If there is no unique behavioral deviation taking place for only my survey questions, that is people have a consistent willingness to pay per util across goods, then we can identify how willingness to pay per util changes over income with just those assumptions. In order to go one step farther and identify the marginal utility of income, we need a stronger neoclassical assumption.

The second assumption needed to identify the marginal utility of income is about connecting our concept of welfare from assumption 1 to people's preferences and action. Specifically,

Assumption 2. *Individual behavior is the result of maximizing utility where utility holds the same meaning as in assumption 1. This implies if a person is willing to trade two goods, they provide the same cardinal utility.*

This assumption gives us the equalities labeled with a 2 in figure 1.1. Since people are willing to trade $X or $Y dollars for pain relief on average, they must provide the same utility.

For this to hold, people need to do what gives them the most utility under the same theory of utility used in assumption 1. The assumption that people maximize their own well-being goes hand in hand with the formal, preference utilitarian definition of utility commonly used in economics. If utility is defined to be higher for the things people prefer, then people's choices maximize their utility by definition. That being said, this assumption does not preclude a substantive theory, such as utility being pleasure, from also holding. If, for example "happiness is the ultimate object of preference, then it could be true both that well-being is the satisfaction of preference and that well-being is happiness" (*Hausman and McPherson*, 2006, Pg 119)

This assumption is a utilitarian theory of action, and it is important to clearly distinguish this from a utilitarian theory of ethics. Utilitarian ethics says the ethical action is the one that maximizes the most good for the most people. A utilitarian theory of ethics would apply to a policymaker who wants to maximize collective utility. A utilitarian theory of action requires people to act in their daily lives as if they ascribe to ethical egoism,[9] since an individual's utility can often come into conflict with what maximizes collective utility. The utilitarian theory of action is what is required by assumption 2.

This is a common assumption. The MVPF framework in *Finkelstein and Hendren* (2020); *Hendren* (2020); *Hendren and Sprung-Keyser* (2020) is a prominent example in public finance, but the consumer welfare standard in industrial organization or cost benefit analyses using a willingness to pay measure are all implicitly making this assumption.

Despite being commonly used in economics, this is also not an uncontroversial or unchallenged assumption. The potential problems come in two flavors. The first is broadly considered by the

[9] They must act like an egoist, but philosophical egoism requires a stronger statement on causality that is not important for our purposes. They may not choose an action because it maximizes self interest, but the assumption is that the actions happen to coincide with self interest regardless (*Sen*, 1977).

field of behavioral economics. People may not maximize their own well-being because they are bad at making choices. They would like to maximize their own well-being, but bounded rationality or behavioral mistakes get in their way. The second possibility is that, even if well informed, people may simply not act to maximize their own well-being. *Sen* (1977) makes this point clearer with a distinction between sympathy and commitment.

> The former corresponds to the case in which the concern for others directly affects one's own welfare. If the knowledge of torture of others makes you sick, it is a case of sympathy; if it does not make you feel personally worse off, but you think it is wrong and you are ready to do something to stop it, it is a case of commitment" (*Sen*, 1977, pg 326).

In either case, an ordinal utility defining choice would put a higher utility index on intervening to stop torture, but only in the former does the choice indicate a higher personal welfare from intervening. Assumption 2 assumes all actions are coming from sympathy, and are not being made in error. This discussion also clarifies what we can say without assumption 2. Even if people make behavioral mistakes or act as (*Sen*, 1977) describes, as long as this is not a unique feature to my survey questions, we have still identified how willingness to pay per util changes over income.

The final equality in figure 1.1 follows by the transitive property and gives us the ratio of the average marginal utility of income between the two groups. While this diagram is a simple example with two groups, the same general logic applies to more groups or a continuous function across income. The next section builds on the theoretical foundation discussed above with an empirical estimation model.

1.4. Empirical Model

With these assumptions in mind, we can now consider a more explicit empirical model. The model starts with a specific functional form for cardinal utility, and thus, takes assumption 2 as given.

$$U(m_i, q_i, X_i, \epsilon_i) = \phi(m_i) + r(q_i, X_i, \epsilon_i) \tag{1.4.1}$$

Where ϕ is a function for the utility of income m and r is a function for the utility for relief from one of the painful experiences in my survey. q_i is the quantity of pain relief ranging from partial to total relief, X_i is a vector of observable characteristics that influence pain tolerance, and ϵ_i is an error term that represents differences in pain tolerance that are not captured by observable characteristics.

As a starting point, I have assumed that the marginal utility of income is constant across people, but I will relax that assumption below. It is crucial that $m_i \notin X_i$. That is, pain relief cannot also be a function of income. It is also crucial that the utility of pain relief and income are additively separable. This ensures there are not cross partials between income and pain relief. With these assumptions, we can see that the reservation price for person i for pain relief is the following

$$P_i^r = \frac{r'(q_i, X_i, \epsilon_i)}{\phi'(m_i)} \tag{1.4.2}$$

The price of money is normalized to one. $\phi'(m_i)$ is the marginal utility of income, and $r'(q_i, X_i, \epsilon_i)$ is the marginal utility of pain relief from one of the events in the survey questions. The equality is the indifference condition that comes from setting the marginal rate of substitution equal to the price ratio.

The math is simplified here by assuming that the change in pain relief is marginal. The questions in my survey are actually discrete, zero or total pain relief, but the responses are generally small relative to the concavity of utility and so should not pose a significant bias for estimation. Appendix A.3.2 shows this in more detail.

Building on equation 1.4.2, the following theorem shows when marginal utility of income can be identified with a conditional average.

Theorem 1. *If $r'(q_i, X_i, \epsilon_i) \perp\!\!\!\perp m_i$,*

then the following holds up to a normalization α

$$\phi'(m_i) = \frac{\alpha}{\mathbb{E}[P_i^r(m_i)|m]} \tag{1.4.3}$$

Note that the if statement in the theorem is equivalent to assumption 1. The proof can be seen in Appendix A.3.3.1

What is the interpretation of the above result given the normalization? The ratio of $\phi'(m_1)$ to $\phi'(m_2)$ is the same for any normalization regardless of m_1 and m_2. This means we can say things like the marginal utility at m_1 is twice the marginal utility at m_2. More generally, we have identified the concavity of utility.

An important point that is implicit in this model is that the price vector is fixed. In this simplistic model where money is a single good, this is not apparent, but in appendix A.3.1, I present the same model using indirect utility, preference heterogeneity, and infinitely many goods. With that approach, it is clearer that the model is conditional on a price vector, but the conclusions are otherwise the same. The implication of being conditional on a price vector is that a change in the relative price of goods, in particular, the ratio of the average price of goods consumed by the high- and the low-income folks will shift the marginal utility of income function. I discuss the implications of this in more detail in section 1.7, "Discussion of Results".

The above equations say that the marginal utility of income is identified by the inverse of the conditional expectation of the reservation price for the special goods in my survey. This conditional average can be estimated in several ways. If income is categorical, we can simply take the average reservation price across income bins. Alternatively, the relationship between income and the reservation price can be estimating with any parametric or non-parametric estimation technique for a conditional average.

One simplification made in the work above is that everyone in the same income group has the same marginal utility of income. This might be a reasonable normalization for some government policy like income tax that redistributes based mainly on income, but how does it hold up when we think about the mental model of utility we used for comparisons in assumption 1? Does extended sympathy, for example, tell us it is identical? Probably not. For example, suppose two people have the same income but one has a lot of wealth. Likely the experience of losing some amount of money will be worse the less wealth one has. We can enrich the model by making the marginal utility of income individual specific and estimating the conditional average. The following theorem shows that, with an assumption on the error distribution somewhat analogous to homoscedasticity, the estimation strategy is unchanged.

Theorem 2. *Let*

$$U(m_i, q_i, X_i, \epsilon_i) = \phi_i(m_i) + r(q_i, X_i, \epsilon_i)$$

be individual utility where

$$\phi_i(m_i) = \mu_i \phi(m_i)$$

is individual specific utility from income and $\phi(m_i) = \mathbb{E}[\phi_i(m_i)|m]$ *is the average utility at income* m *and* μ_i *dictates the individual specific deviation from the average.*

If $\frac{r'(q_i,X_i,\epsilon_i)}{\mu_i} \perp\!\!\!\perp m_i$ *then the following holds up to a normalization* α

$$\mathbb{E}[\phi_i(m_i)|m] = \frac{\alpha}{\mathbb{E}[P_i^r(m_i)|m]} \tag{1.4.4}$$

The proof can be found in Appendix A.3.3.2.

In words, the required independence assumption is saying that the utility from pain relief is independent of income, and that the deviation in individual marginal utility from the mean, as a multiple of the mean, is independent of income. An implication of this is that if the marginal utility of income doubles, the variance of the marginal utility would double as well. While not a directly testable assumption, if this holds the variance of the reservation price will change in proportion to the marginal utility of income. Since we also expect the mean to change in proportion to the reservation price, the variance of the reservation price should change in proportion with the mean.

If we relax the assumption about how the error in the marginal utility of income is distributed, then we get the following:

Theorem 3. *Let $\phi_i(m_i)$ be defined as in Theorem 2*

If $r'(q_i, X_i, \epsilon_i) \perp\!\!\!\perp m_i$ and $r'(q_i, X_i, \epsilon_i) \perp\!\!\!\perp \phi_i(m_i)|m$ then the following holds up to a normalization α

$$\mathbb{E}[\phi(m_i)|m] = \alpha \mathbb{E}\Big[\frac{1}{P_i^r(m_i)}|m\Big] \tag{1.4.5}$$

The proof for this is in appendix A.3.3.3. For each empirical estimation in the paper, I include a robustness check with the estimation strategy from theorem 3. Before showing the results or building on this model, the next section reviews the data collection, population, and summary statistics.

1.5. Survey Data

The data for my main analysis comes from a new survey fielded to 1747 respondents through the survey panel company Centiment[10]. Respondents are recruited through Facebook, LinkedIn, and partner networks to fill out surveys for money. While not a truly random sample, my respondents are matched to the census on age, race, gender, and region and these demographics are provided to me by Centiment.

The full survey can be seen in appendix A.1. The first two sets of questions are variations of validated questions on the pain sensitivity questionnaire (PSQ) *Ruscheweyh et al.* (2009). These are open ended questions asking respondents to report their willingness to pay. While I recommend looking at the exact wording in appendix A.1, the painful scenarios are getting lemon juice in a minor wound, picking up a hot pot, burning your tongue on a very hot drink, and bumping your shin on a hard edge.

The second set of questions asks a simple “yes or no" question of the form, “would you pay $X to relieve that pain", where X is randomly selected. The painful scenarios in this section are hitting your funny bone, biting your tongue or cheek, and slamming your finger in a drawer.

The next set of questions is a subset of the PSQ asking respondents to say how painful scenarios would be on a 0-10 scale. Many of these overlap with the WTP questions above to get a sense of individual aversion to particular events.

[10] website: https://www.centiment.co/

I chose the PSQ questions that were reported as the most severe, but also that resolve in a relatively short amount of time. The pain resolving in a short amount of time helps satisfy the assumption that the marginal utility of relief from these experiences is uncorrelated with income because other goods do not impact the experience. To better see why these are a good fit, consider an excluded question that asks about pain from a sunburn. Sunburn is much less likely to satisfy the independence assumption because the respondent's job or access to soothing medication might make a difference on the actual pain sunburn causes.

One question in the first section and one in the third ask about a scenario that typically would not be painful at all. I refer to these questions as "catch questions". These "catch questions" are used to determine who is engaging in good faith in the questions and paying attention. The scenario presented is shaking hands with someone who has a normal grip. How exactly these are used to exclude responses is outlined in section 1.5.1. Additionally, in the third section of the survey there is a question asking respondents to enter 9 to ensure their full attention. Respondents who did not enter 9 for this are, at that point, removed from the survey.

The next section asks a few basic demographics. Marital status, number of children, education, employment. The final section of the survey asks about income and financial health. The financial health questions come from the Fin-health survey[11].

1.5.1 Protest Answers and Outliers

It is common in open ended willingness to pay surveys to receive protest answers to questions, yet there is not a unique strategy in the literature to handle them (*Boyle*, 2017, pg.110-111). The general ad hoc solution is to include additional questions that allow analysis to identify respondents who are not engaging with the questions in the desired way. I remove respondents from the survey based on their response to the "catch" questions I mentioned above. Specifically, shaking hands with someone who has a normal grip should not be painful. It may make sense in theory to exclude anyone who did not answer 0 to both of these catch questions. The situations are not painful and so rational utility maximizing respondents should not pay anything to relieve the pain and should enter 0 when asked. I do not use this strict of a cutoff for my base specification. It is possible some people perceive some pain from these situations or simply did not consider 0 to be a viable option.

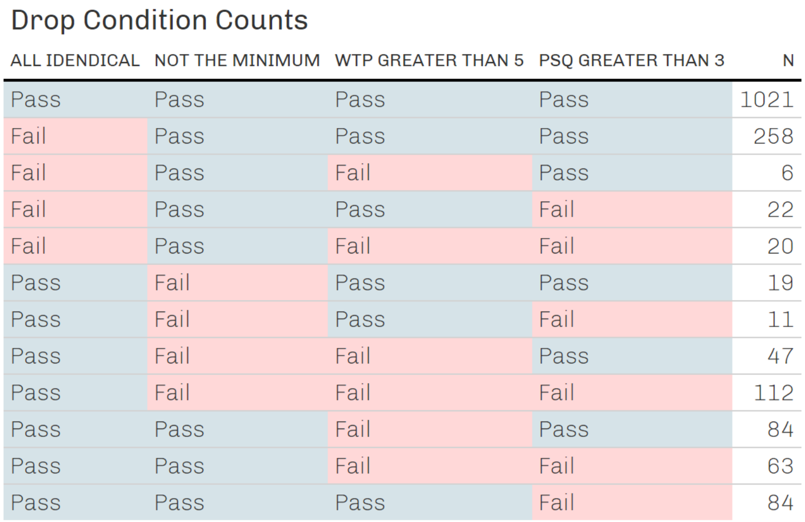
Instead of removing respondents with non-zero answers to the catch questions, I use the following conditions. First, I look at the response to the open-ended catch question. I exclude anyone who answered higher on this than any of the other painful open-ended questions. The thought here is that, while respondents may not think to enter zero, they should not say a

[11] Website: https://finhealthnetwork.org/

greater amount than for questions that are clearly painful. Second, to check for attention, I remove anyone who answered the same thing for every question. This includes 263 respondents who entered zero for every question. Finally, I drop anyone who said more than $5. This is, admittedly, a bit of an arbitrary cutoff, but the goal is to allow people who defaulted to low, non-zero answers without included people who were not thinking about the question.

The last condition is based on the response to the pain sensitivity questionnaire catch question. Here, I ask again about shaking hands with someone who has a normal grip but ask them to rank the pain from 0-10. I drop anyone who's answer to this question is greater than 3. I tried a few other more complicated conditions, like being less than the mean or less than all other PSQ questions, but most of the cases overlapped with the three or less condition anyway. Table 1.1 shows how many responses were dropped for meeting each condition. 1021 responses remain after all of the drop conditions.

Table 1.1: Each Row indicates what conditions are passed or failed. One failure leads to the response being dropped. The total for each combination of conditions are in the N column.

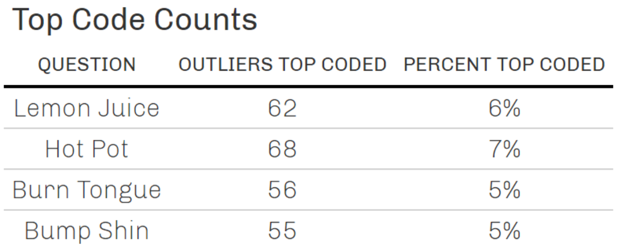


Drop Condition Counts

ALL IDENDICAL	NOT THE MINIMUM	WTP GREATER THAN 5	PSQ GREATER THAN 3	N
Pass	Pass	Pass	Pass	1021
Fail	Pass	Pass	Pass	258
Fail	Pass	Fail	Pass	6
Fail	Pass	Pass	Fail	22
Fail	Pass	Fail	Fail	20
Pass	Fail	Pass	Pass	19
Pass	Fail	Pass	Fail	11
Pass	Fail	Fail	Pass	47
Pass	Fail	Fail	Fail	112
Pass	Pass	Fail	Pass	84
Pass	Pass	Fail	Fail	63
Pass	Pass	Pass	Fail	84

For the open response questions, I also top coded responses. The maximum answer, for example, was one trillion dollars. I top-code them rather than dropping them since top-coding uses some of the information. Whoever entered such a high amount likely has a true amount that is high as well. A standard rule of thumb is to add 1.5 times the inter quartile range to the 75th percentile and treat anything above that as an outlier. As my data is skewed, I use 4.5 times the IQR for each income group. The number of top codes can be seen in table 1.2. The distributions for each open response question after removing protest answers and top coding outliers can be seen in figure 1.2.

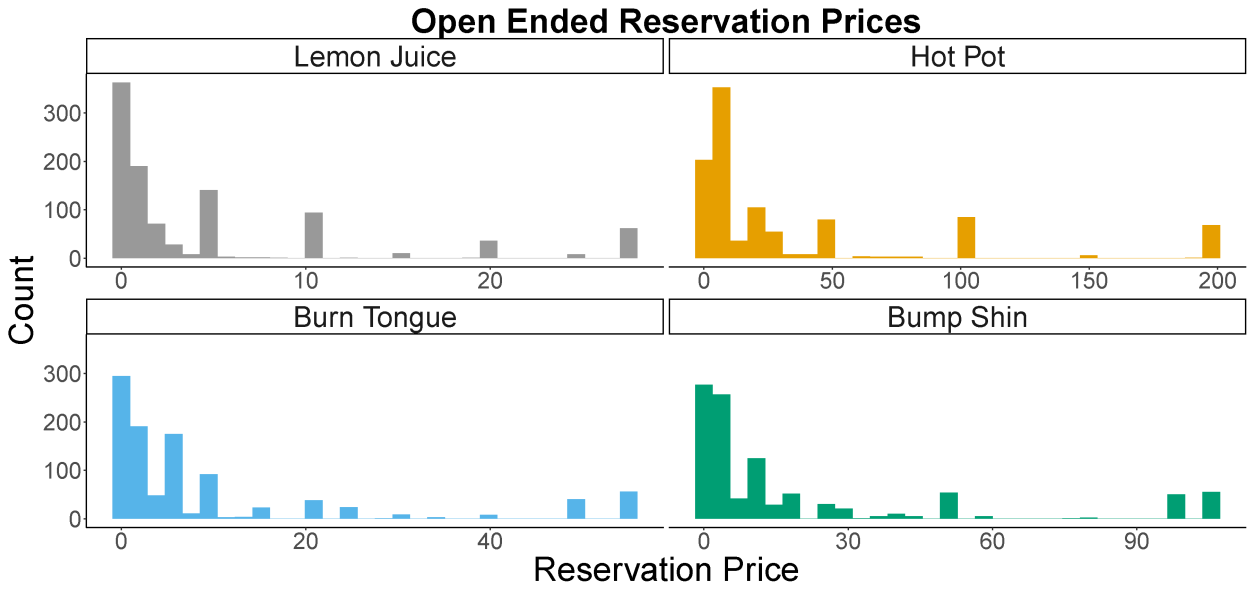
Table 1.2



Top Code Counts

QUESTION	OUTLIERS TOP CODED	PERCENT TOP CODED
Lemon Juice	62	6%
Hot Pot	68	7%
Burn Tongue	56	5%
Bump Shin	55	5%

Figure 1.2



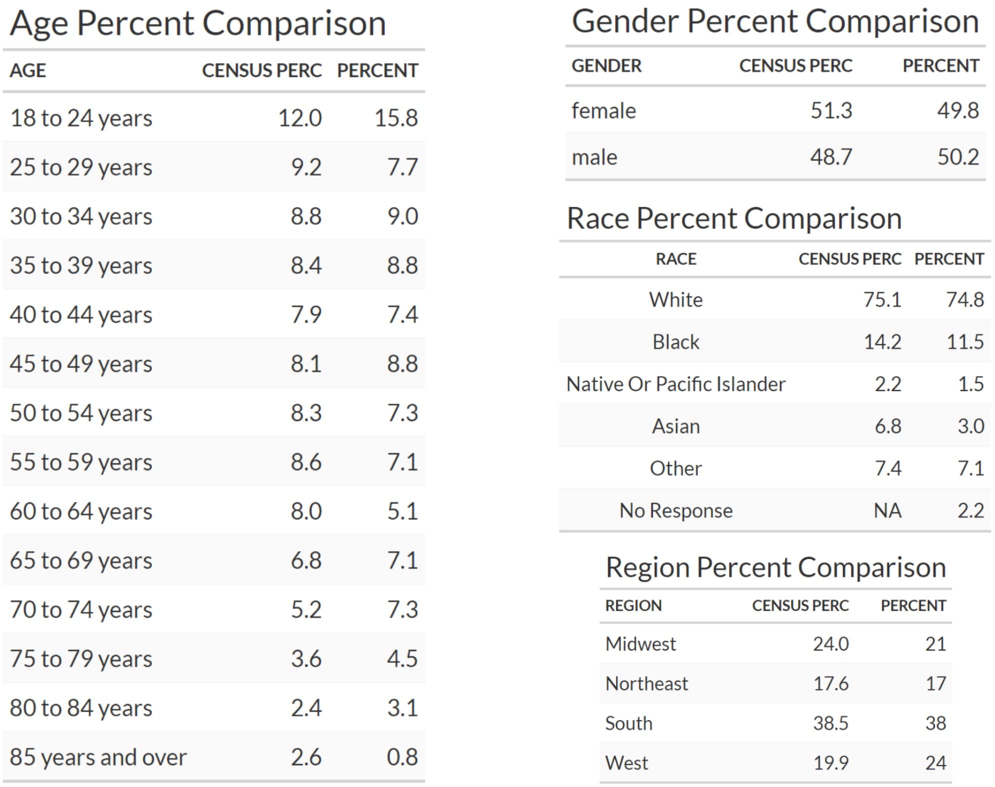
*Values are after protest removal and top coding

1.5.2 Summary Statistics

Before jumping into the analysis of willingness to pay and marginal utility, it is important to understand the population. While Centiment matched my survey sample to the census on age, race, gender, and census region, some respondents were dropped from the sample as described above. Table 1.3 shows rates for the matched demographic variables in my final sample compared to the census. Income was not matched explicitly, but figure 1.3 compares the results of my survey to the CPS family income. As is expected, I am slightly oversampled at lower income levels and undersampled at higher income levels. To gauge internal validity, we also want to consider how pain sensitivity differs across income.

Figure 1.4 shows the relationship between pain sensitivity, measured by the mean pain sensitivity questionnaire score (PSQ) from 0-10, and income. While the point estimate for the slope is negative, it is not statistically significant. The table within the graph shows the results of the

Table 1.3: Balance of Observable Traits



Age Percent Comparison

AGE	CENSUS PERC	PERCENT
18 to 24 years	12.0	15.8
25 to 29 years	9.2	7.7
30 to 34 years	8.8	9.0
35 to 39 years	8.4	8.8
40 to 44 years	7.9	7.4
45 to 49 years	8.1	8.8
50 to 54 years	8.3	7.3
55 to 59 years	8.6	7.1
60 to 64 years	8.0	5.1
65 to 69 years	6.8	7.1
70 to 74 years	5.2	7.3
75 to 79 years	3.6	4.5
80 to 84 years	2.4	3.1
85 years and over	2.6	0.8

Gender Percent Comparison

GENDER	CENSUS PERC	PERCENT
female	51.3	49.8
male	48.7	50.2

Race Percent Comparison

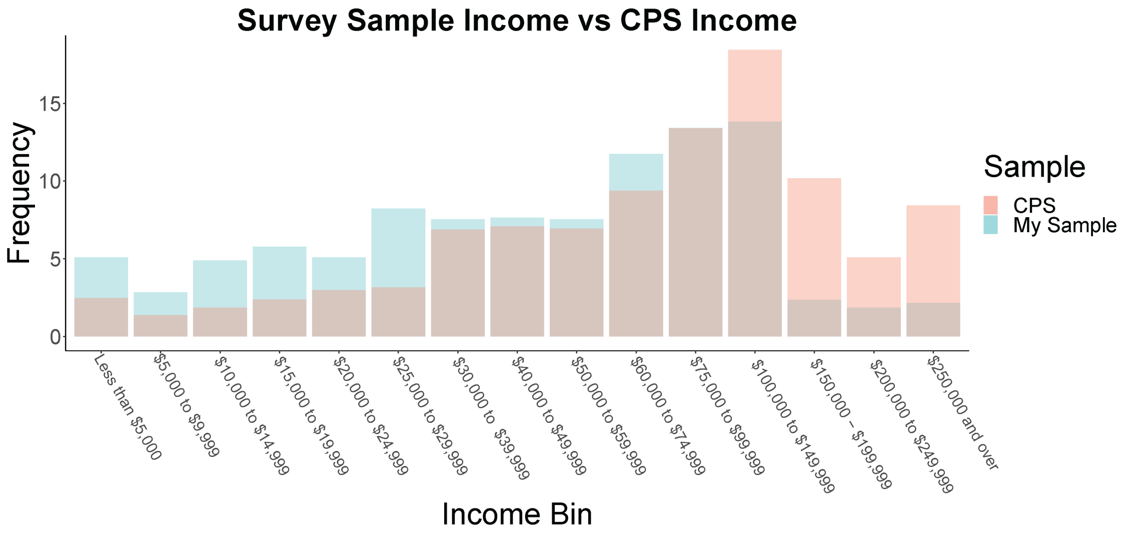
RACE	CENSUS PERC	PERCENT
White	75.1	74.8
Black	14.2	11.5
Native Or Pacific Islander	2.2	1.5
Asian	6.8	3.0
Other	7.4	7.1
No Response	NA	2.2

Region Percent Comparison

REGION	CENSUS PERC	PERCENT
Midwest	24.0	21
Northeast	17.6	17
South	38.5	38
West	19.9	24

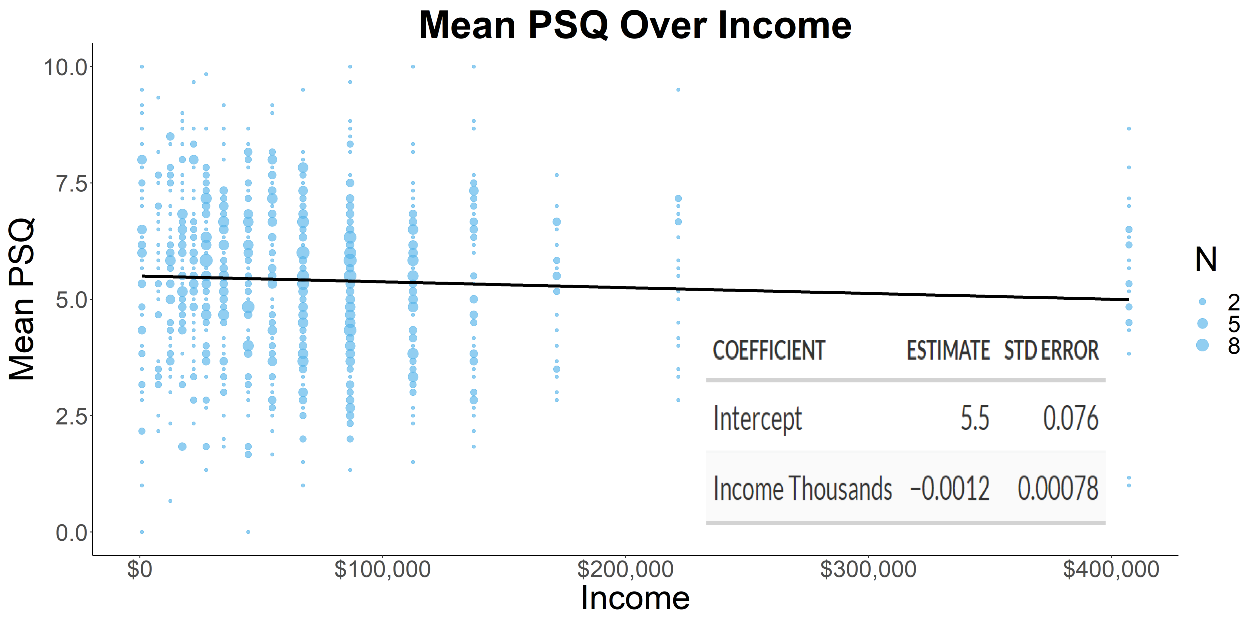
Note: Values are after protest answer removal

Figure 1.3



Grey is the overlap of both histograms simple regression from the plot. A one hundred thousand dollar increase in income is associated with a .12 decrease in average PSQ response (on a scale from 0-10). Not only is this statistically insignificant, but it is also practically small relative to the within income variation. Nevertheless, I do control for the average responses to these questions in section 1.6.2.2.

Figure 1.4



COEFFICIENT	ESTIMATE	STD ERROR
Intercept	5.5	0.076
Income Thousands	-0.0012	0.00078

Note: Mean PSQ is each respondents average over the 0-10 style pain sensitivity questions

1.6. Survey Analysis

1.6.1 Open Ended Mean Results

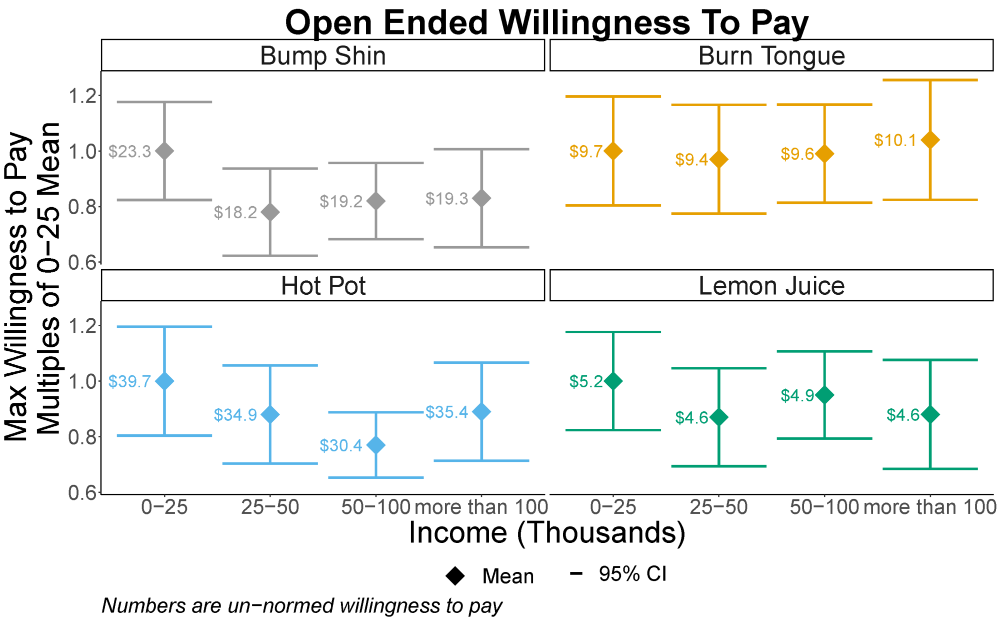
As I showed in theorem 2, the average marginal utility of income as a function of income is equal to the reciprocal of the expected reservation price conditional on income. The conditional expectation of the reservation price can be calculated in a variety of ways. The first method is to estimate the mean for four income groups. The size of each group can be seen in table 1.4.

Table 1.4

Aggregated Income Counts	
Income	Count
0-25	242
25-50	239
50-100	334
more than 100	206

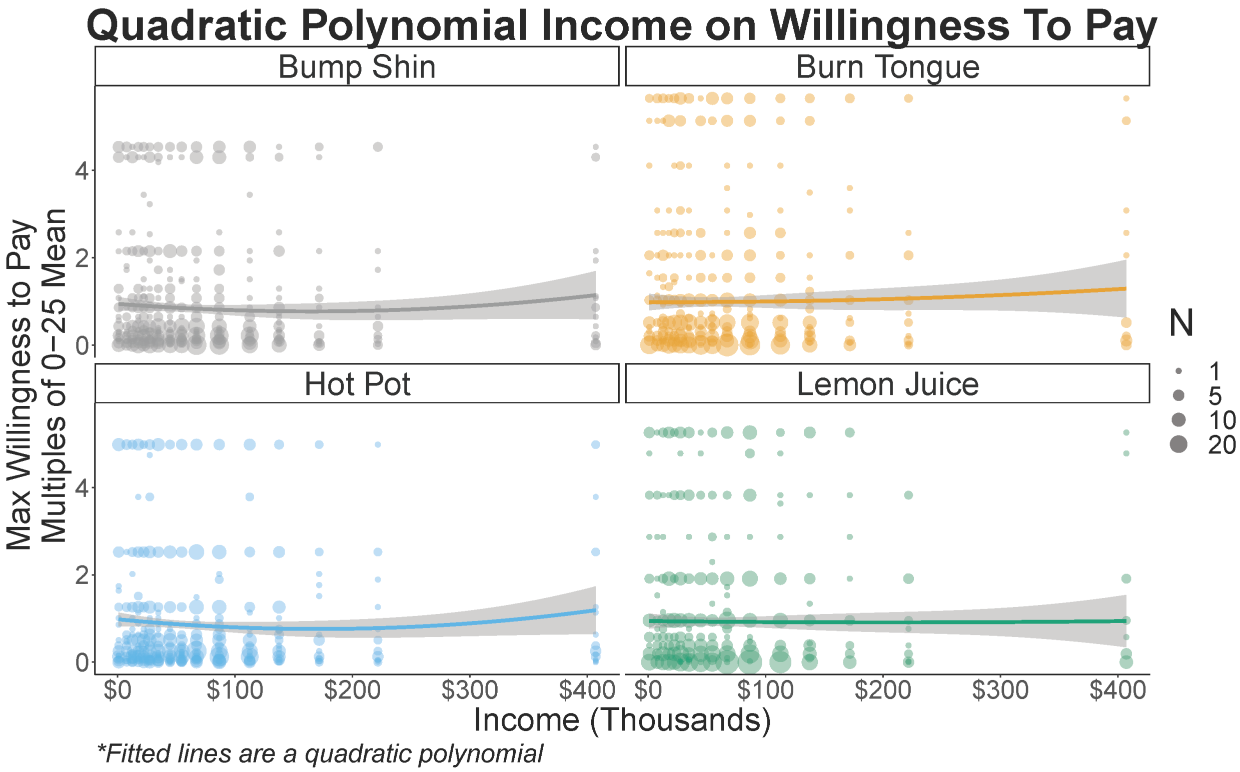
Figure 1.5 plots the mean willingness to pay for relief for each income bin in each question. The vertical axis normalizes the lowest income group to 1. This is done because the willingness to pay only identifies the marginal utility of income, up to a normalization. So, what is actually of interest is the ratio of responses across incomes, not the absolute values. Normalizing to one makes it easier to compare the implied marginal utility of income across questions. The normalized mean response is marked with the large diamond, and bootstrap standard errors are included for each mean. The un-normalized values are displayed in text to the left of each point as well. The theory of diminishing marginal utility of income would predict a rise in WTP, but, surprisingly, the point estimates are relatively flat or even decreasing a bit.

Figure 1.5



The second method for estimating the conditional expectation of the reservation price is to make a parametric assumption about the conditional average. Figure 1.6 shows every response to the open-ended questions across the full range of income categories. Each point is sized for the number of people in that income group with that response and a quadratic polynomial is fit for each question. Again, the results indicate little to no increase in WTP.

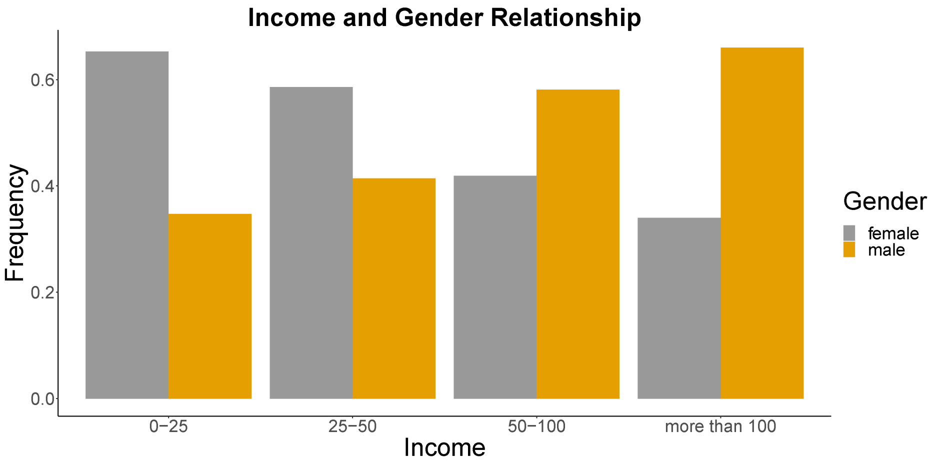
Figure 1.6



One concern with this approach discussed in section 1.4, is that the assumption from theorem 2, that the error in marginal utility is independent from income, does not hold. As I also explained above, if the error is independent from income, we would expect the standard deviation and mean to change in lockstep with one another over income. Table A.1 in appendix A.2 includes the mean and standard deviation for each income group and each question as a ratio of the $0-25k group. They both appear relatively constant across income, which supports the assumption. Given this support, I will continue to rely on this assumption for the main analysis, but I include robustness checks for each estimation strategy. The first of these alternative strategies can be seen in appendix A.2 table A.2 where the mean of the inverse of willingness to pay is shown. This is in line with the result from theorem 3.

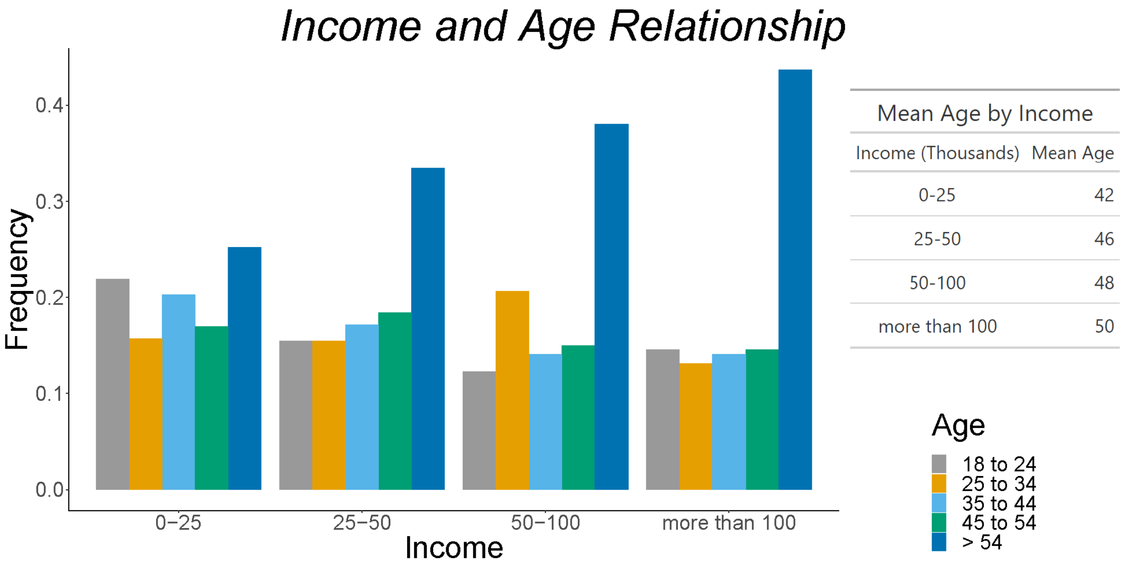
Another concern with this simple approach is that perhaps pain tolerance is not independent from income. This would be an issue for every theorem in section 1.4. While income may not itself change the marginal utility of pain relief in these situations, there may be characteristics that both impact pain tolerance and correlate with income. Gender and age have been suggested to affect pain tolerance and by extension the utility from pain relief (*Bartley and Fillingim*, 2013; *Lautenbacher, Peters, Heesen, Scheel, and Kunz*, 2017). These are both related to income in my sample. Figure 1.7 shows that men are more concentrated in the high-income categories. Figure 1.8 shows that age increases some with income. The average goes from 42 in the lowest income group to 50 in the highest. Surprisingly, neither age nor gender has a significant relationship with the pain sensitivity questionnaire (PSQ). One way to further investigate pain tolerance is to look at the unconditional relationship with WTP for relief to see if there are differences across groups. There is not a significant relationship for gender and WTP on the open-ended questions. There is, however, a significant relationship with age. Older respondents are willing to pay less on average, as we can see in figures 1.9, despite having higher average incomes. These observations at the very least, motivate adding gender and age controls as a robustness check. While there is no apparent relationship between average PSQ responses and income, I include the PSQ as a control as well.

Figure 1.7



Unfortunately, we cannot simply add controls as a linear parameter into our model of conditional averages using ordinary least squares (OLS). This is because our outcome is in terms of WTP, but the difference between groups, like age, is in utility. Suppose men are more willing to endure pain on average. Pain relief would give men some fixed α fewer utils than women. Figure 1.10 shows a hypothetical relationship that assumes marginal utility is diminishing with income. As income increases, that α util gap translates into a larger and larger gap in WTP. In fact, that gap grows in proportion to the marginal utility of income. This means a simple binary control would not be sufficient. An OLS model with interactions would allow for the differing slopes in figure 1.10, but this sacrifices an opportunity to use the difference in the slopes to help identify the marginal utility of income. To fully take advantage of this relationship, I use a maximum likelihood estimation (MLE) technique.

Figure 1.8



Mean Age by Income

Income (Thousands)	Mean Age
0-25	42
25-50	46
50-100	48
more than 100	50

Figure 1.9

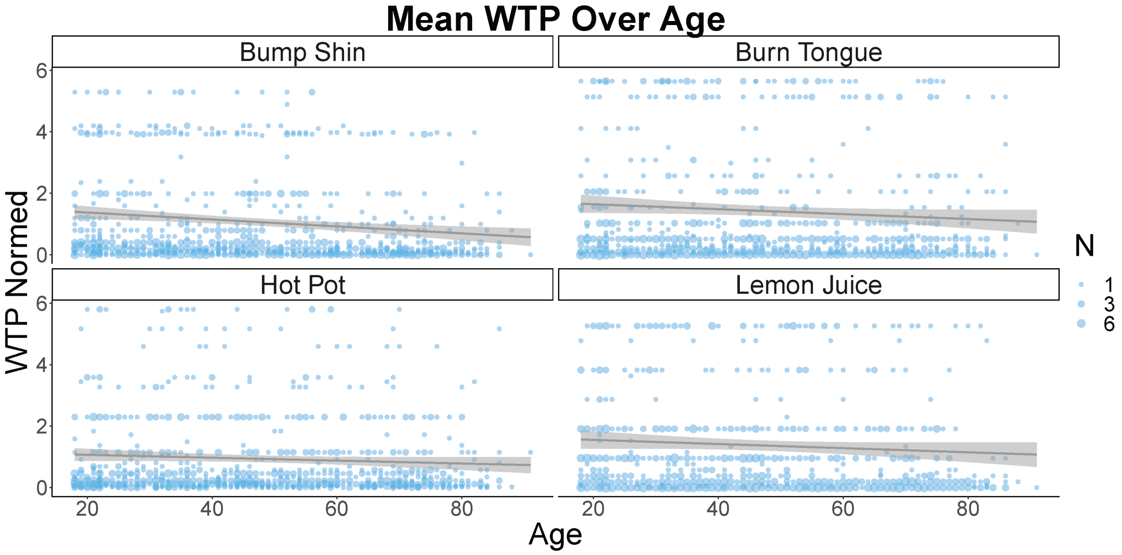
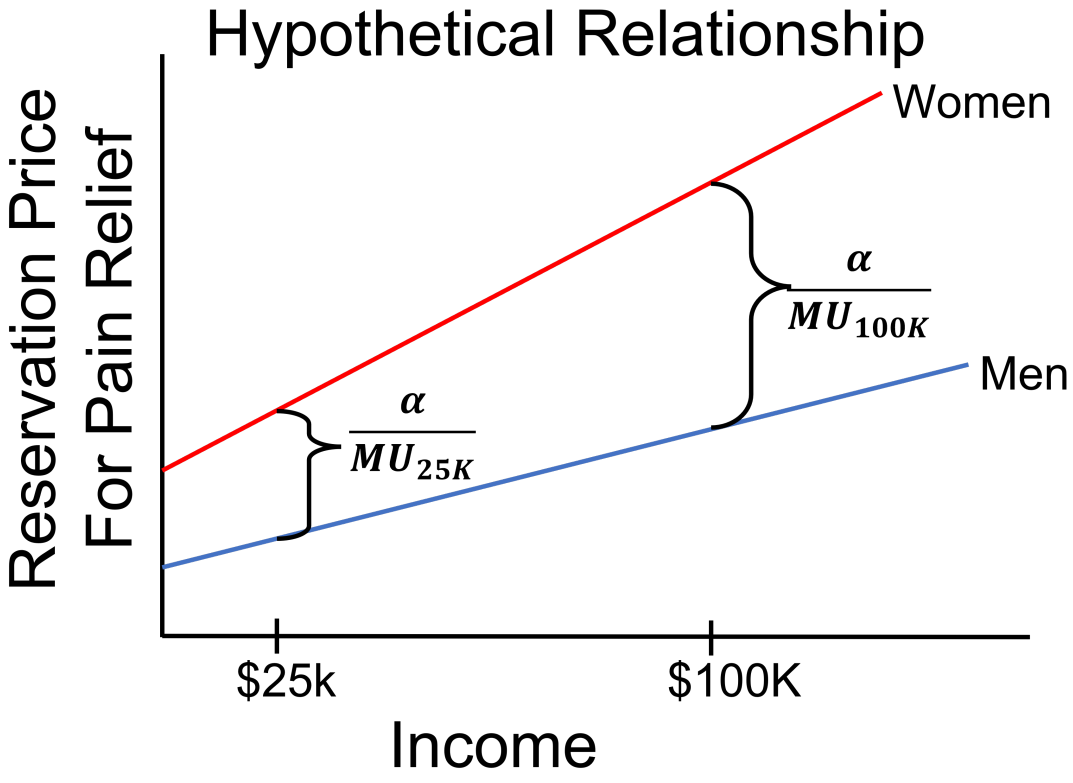


Figure 1.10


Hypothetical Relationship
Reservation Price For Pain Relief
Women
Men
$\frac{\alpha}{MU_{100K}}$
$\frac{\alpha}{MU_{25K}}$
$25k
$100K
Income

1.6.2 Open Ended Response With Controls

To move beyond simple averages, we need to further parameterize our model. As with the means, I aggregate income groups into b income bins. I assume the average marginal utility of income function is

Assumption 3.

$$\phi_i'(m_i) = \sum_{k=1}^{b} \mathbb{1}_{ik}(m_i \in k)\phi_k'$$

This gives b average marginal utility of income parameters, ϕ_k'. In this case b is four for the four income groups.

Now for the utility impact of r_i. I assume X_i and enters the model with linear parameters in the following form

Assumption 4.

$$r(X_i, \epsilon_i) = \beta_1 + \boldsymbol{\beta} X_i + \epsilon_i \tag{1.6.1}$$

Where

$$\epsilon_i \sim \mathcal{N}(0,\, \sigma^2) \tag{1.6.2}$$

and

$$\epsilon_i \perp\!\!\!\perp m_i \tag{1.6.3}$$

Together, these assumptions give the following equation for the reservation price of pain relief

Definition 1. *Given assumption 3 and 4 the reservation price for an individual is*

$$P_i^r = \frac{\beta_1 + \boldsymbol{\beta} X_i + \epsilon_i}{\sum_{k=1}^{B} \mathbb{1}_{ik}\phi_k'} \tag{1.6.4}$$

In matrix form, the expected price vector $\mathbb{P}^r$ for the full population can be written as

$$\mathbb{P}^r = (\mathbb{X}_i\boldsymbol{\beta} + \epsilon_i) \oslash \mathbb{M}\phi' \tag{1.6.5}$$

Where $\mathbb{X}$ is the matrix of X_i traits influencing pain tolerance and a constant for the intercept of the numerator. In matrix form, let $\boldsymbol{\beta}$ be $[\beta_1 \quad \boldsymbol{\beta}]$ from equation 1.6.4. Let $\mathbb{M}$ be an n by b matrix indicating what income bin each person i is in. Let $\boldsymbol{\phi'}$ be a b by 1 matrix of marginal utility parameters. $\oslash$ is element wise matrix division or Hadamard division.

Appendix A.3.4 proves that the model is identified, but the basic intuition is that it is identified so long as demographic characteristics do not show up in both the numerator and the denominator.

For example, differences in pain relief across age cannot be separately identified from differences in marginal utility of income across age. This is why the independence of income and the utility from pain relief is crucial.

1.6.2.1 Likelihood Functions

We can estimate the model with MLE. Definition 1 gives the following probability density function.

Definition 2.

$$P(P_i^r|\mathbb{X}_i, m_i \in k) = \pi \frac{1}{\sqrt{2\pi \frac{\sigma^2}{\phi'_{k_i}}}} e^{-\frac{(P_i^r - \hat{P_i^r})^2}{2(\frac{\sigma}{\phi'_{k_i}})^2}} \tag{1.6.6}$$

and the following log likelihood for a population of size n

Definition 3.

$$L(\mathbb{X}_i, m_i) = -\frac{n}{2}log(2\pi) + \sum_i^n -log(\frac{\sigma}{\phi'_{k_i}}) - \frac{(P_i^r - \frac{1+\boldsymbol{\beta}\mathbb{X}_i}{\phi'_{k_i}})}{2(\frac{\sigma}{\phi'_{k_i}})^2} \tag{1.6.7}$$

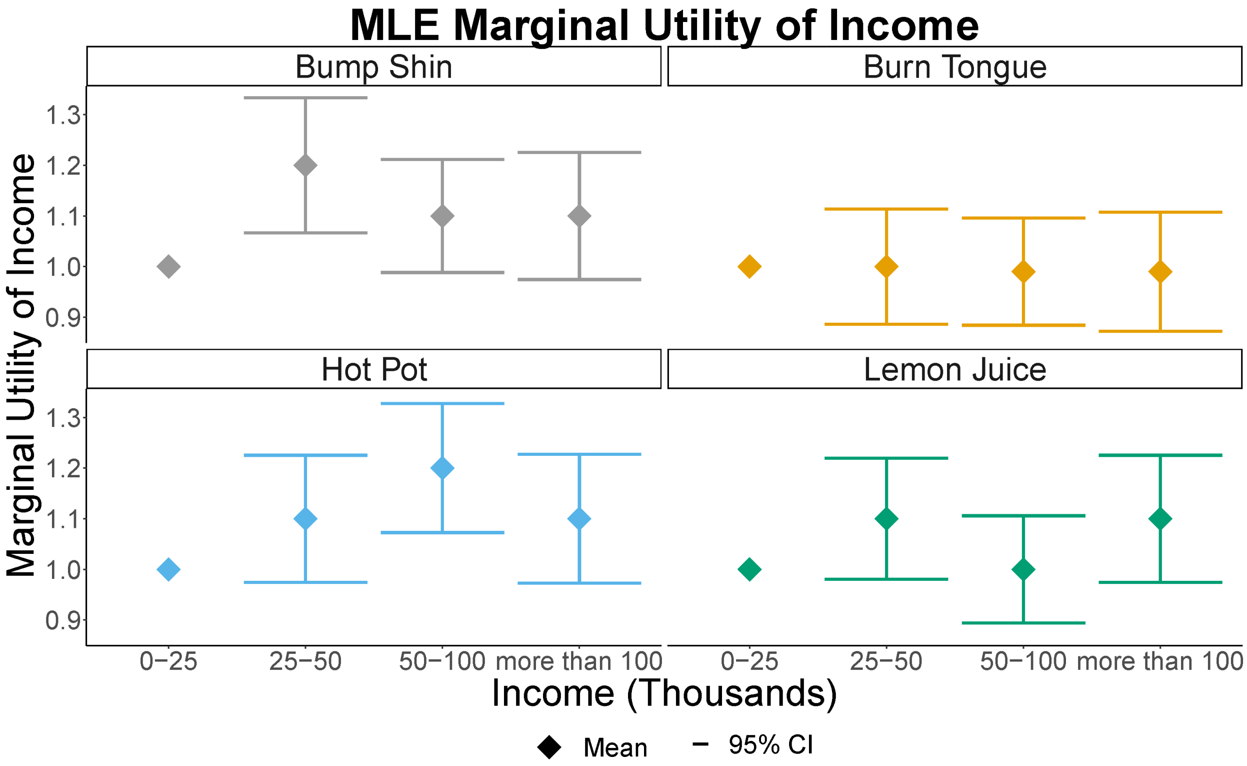
Where ϕ'_{k_i} indicates the marginal utility parameter for the corresponding bin for m_i. Recall that in assumption 4 the variance of the error term, and by extensions the variance in the utility from pain relief, is constant across income. This assumption allows the ϕ'_k terms to be identified off of mean differences in reservation prices, differences in the variance, and differences in reservation price gaps between control groups. We can see this since the marginal utility of income enters the same terms as σ and the term with $1 + \boldsymbol{\beta}\mathbb{X}_i$. The intuition here is that the underlying mean, variance, and coefficients in utility from pain relief are constant across income. So, observed differences in the mean, variance, or coefficients for the reservation prices across incomes must be because of differences in the marginal utility of income.

1.6.2.2 MLE Results

I run this model with controls for age, gender, and mean PSQ. The results of the model for each open-ended question are shown in figure 1.11. The full coefficient table is also included in appendix A.2 table A.3. The vertical axis in figure 1.11 are direct estimates of marginal utility ϕ_k and not WTP like in the unconditional means above. Here, a higher value indicates a higher marginal utility. All the questions are generally flat. For example, none of even the largest income groups are statistically different than 1. A nice sanity check is that the coefficients on "Mean

PSQ", which is the respondent's average response to the 0-10 pain sensitivity questions, are positive and significant for every question. For example, a one-point higher average is associated with a .25 increase in utility from pain relief for bumping one's shin. With utility normalize to one this means a one-point increase in PSQ is correlated with a $.25 increase in the reservation price for relief from bumping one's shin. This indicates that people who are more sensitive to pain will pay more to avoid it and indicates people are actually considering the questions and implications and answering in a logically consistent way throughout. Table A.4 in appendix A.2 shows the coefficients for a model where the variance of the error in utility is not assumed to be uniform across income groups, reflecting the conclusion from theorem 3. The results are similar.

Figure 1.11



The contingent valuation literature suggests that open ended questions are more difficult to respond accurately to and have high rates of protest answers (*Boyle*, 2017, pg 110-111). With this in mind, I also ask a series of binary choice, yes or no, questions. The price proposed is randomized across respondents, but each respondent only answers each question one time, for one price. This is the opposite extreme to open ended questions in that it carries less information, but places much less burden on the respondents.

1.6.3 Binary Choice Model

While a binary choice question may be more familiar for respondents, leading to more accurate answers, assessing the results is more difficult and requires more assumptions. The underlying economic model is the same as in theorem 1 and 2, but now we must model and estimate the reservation price rather than just observing it in the data. I follow the general strategy from *Hanemann* (1984), but the technique outlined in that paper is to calculate an overall average or median WTP. I update the technique to analyze the change in WTP with respect to income rather than a single collective estimate. The first step is to run a random utility logit regression of the following form.

$$V_i = \sum_{j=1}^{4} \delta_j \mathbb{1}(M_i = j) + \gamma X_i + \sum_{j=1}^{4} \beta_j \mathbb{1}(M_i = j) * P_i + \epsilon_i \tag{1.6.8}$$

Where δ_j is an intercept coefficient for income level j, γ is a vector of coefficients for controls X_i and β_j is the price coefficient for income group j. It might seem incorrect, at first pass, to include unique intercepts for income in the utility model. This seems to imply different utility levels across incomes (the opposite of the identifying assumption). However, it is important to remember that V_i is ordinal utility that should not be compared across individuals and is, in my opinion, better thought of as just modeling choice probability. The differing utility intercepts allow for different choice probability levels across incomes while the differing price coefficient allows the choice probability to decrease deferentially with price across incomes. The intercepts allow the model to more flexibly estimate WTP with less strict functional form assumptions. A nice clarifying example is to imagine a good where people of different incomes buy it with the same probability at a price near zero (indicating a similar choice probability intercept). This observation would not in any way indicate that the good provides the same marginal benefit to both people. Similarly, the converse, having different purchase probability intercepts, does not imply different marginal benefits.

Putting our above equation in terms of *Hanemann* (1984), let α_1 and α_0 be the utility from pain relief and no pain relief respectively. We can write the difference in utility from paying for pain relief as

$$\Delta V = (\alpha_1 - \alpha_0) - \beta_j P \tag{1.6.9}$$

The CDF for the change in utility $F_n(\Delta V) = (1+e^{-\Delta V})^{-1}$ gives the probability of purchasing relief. By extension the CDF of the reservation price is $F_n(\Delta V) = G_{P_r}(P)$. Now the mean reservation price, $\bar{P_r^j}$, for an income group j is represented by

$$\bar{P_r^j} = \int_0^\infty [1 - G_{P_r}(P)] dP \tag{1.6.10}$$

Connecting this back to the actual logit model, we can write this as

$$\bar{P_r^j} = \int_0^\infty [1 - \frac{1}{1 + e^{\delta_j + \bar{X}\gamma + \beta_j P}}] dP \tag{1.6.11}$$

Where $\bar{X}$ is the average value of the controls, age and gender and mean PSQ in this case, for the entire population. This assigns each group the ordinal utility for the average age and gender, but varies the ordinal utility associated with the income group and then scales the ordinal utility by the income specific ordinal marginal utility of income. The difference across groups shows the difference in WTP attributable to changes in income holding the controls fixed at the global means.

A slight variation of this I also use, which is used in *Bishop and Heberlein* (1979) and mentioned in *Hanemann* (1984), is to cap the integral at the maximum price.

1.6.3.1 Binary choice Results

A nice first check for the binary choice questions is to confirm that people are price sensitive. Do fewer people agree to pay for relief as the price increases? Figure 1.12 has log price on the horizontal axis and the percent of respondents who said they would pay that price on the vertical axis. This percent falls in all questions for all income categories. This indicates people did engage with the question and consider the price; however, a relationship between income groups is not clear.

The willingness to pay estimates with controls for age, gender, and the mean response to the 0-10 PSQ questions and using the truncated mean (where the integral is capped at the highest bid) are displayed in figure 1.13. These estimates are normalized in the same way as the open-ended means are. The first two questions seem to support the same story as the open-ended responses while the last suggests a slight increase in WTP, and so a decrease in marginal utility. Between the lowest and highest income groups the WTP roughly doubles for the question regarding slamming your finger in a drawer, indicating marginal utility of income is halved. Table A.5 in appendix A.2 shows the results for the standard mean, truncated mean, and median willingness to pay estimates. The mean and truncated mean are what are described above, but the median estimate is described in *Hanemann* (1984).

For the binary choice questions, the error assumption in theorem 2 might not hold as well as in the free response questions. Appendix A.3.5 presents a model to estimate $\mathbb{E}\left[\frac{1}{P_i^r(m_i)}|m\right]$ in line with theorem 3, and the results can be seen in appendix A.2 figure A.1. Using this method, all three questions indicate little change over income.

Figure 1.12

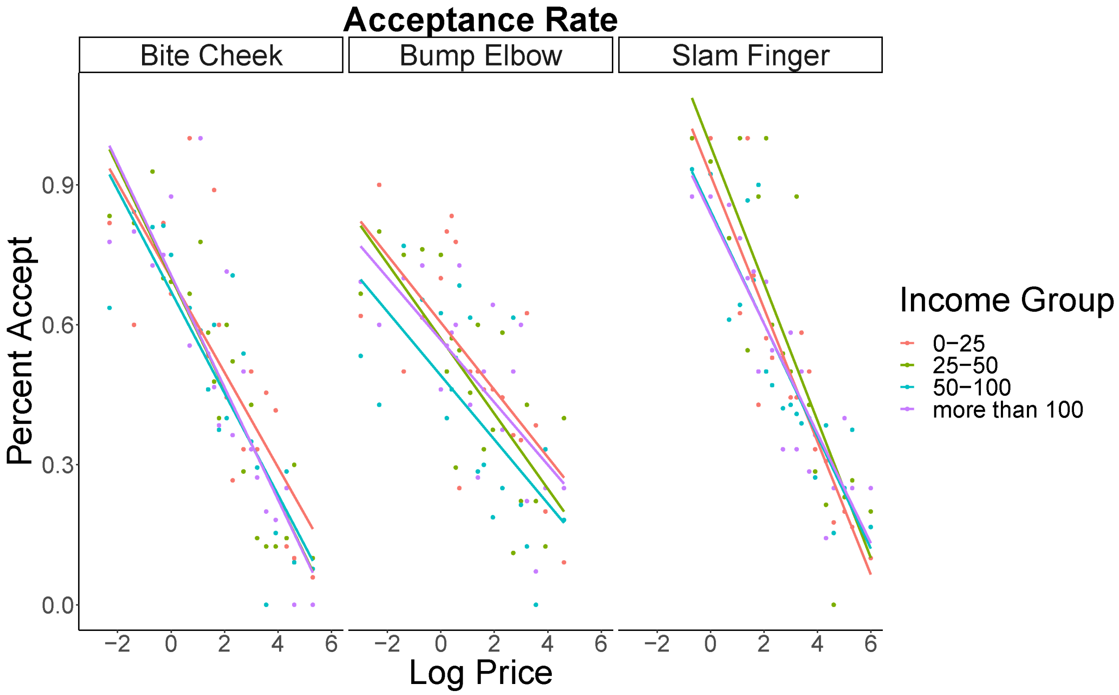
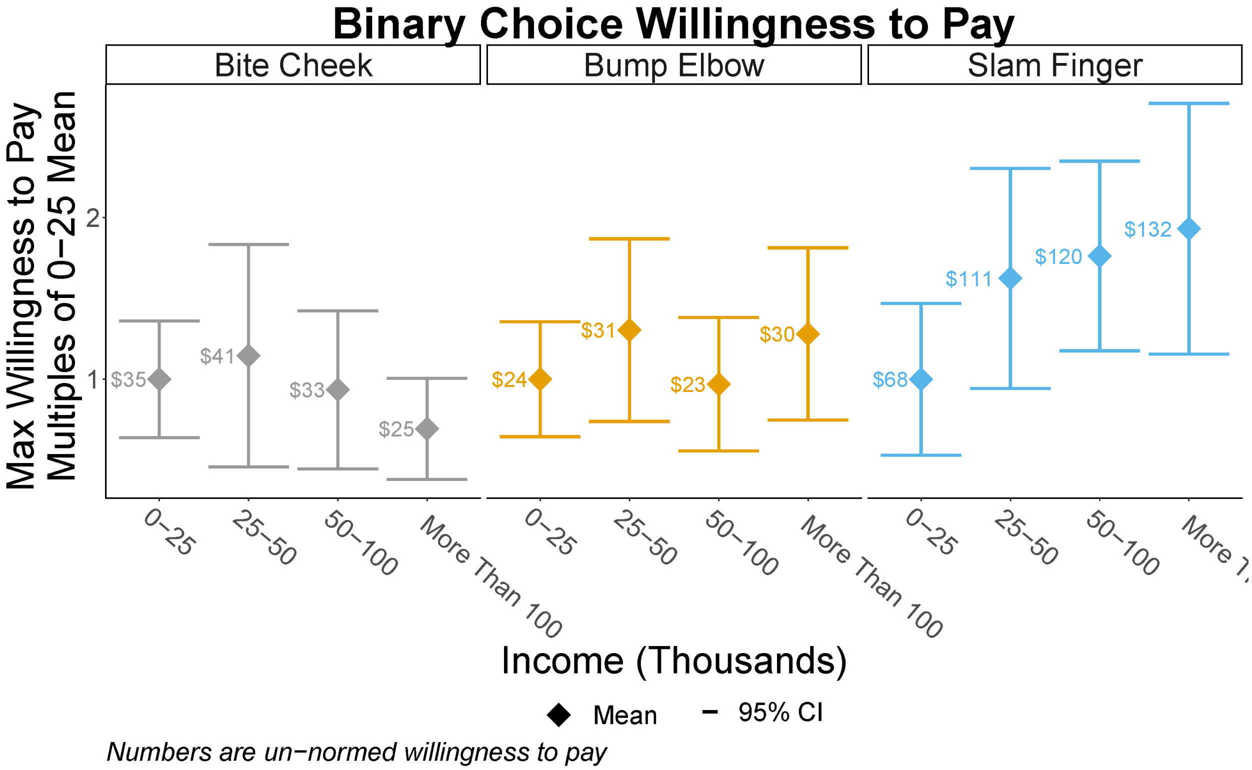


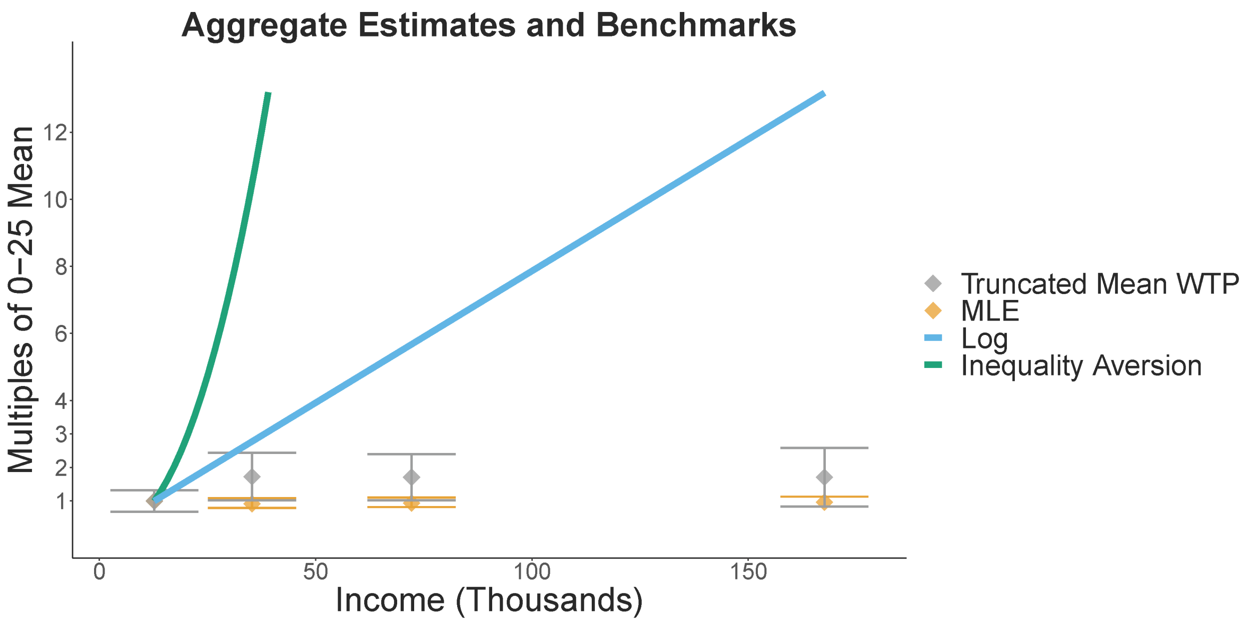
Figure 1.13



1.6.4 Aggregate Estimates and Benchmarks

The results indicate that there is little to no change in the willingness to pay for relief over income. While it is not possible to prove the null hypothesis that there is no effect, figure 1.14 shows that the confidence intervals for the point estimates are far outside of the standard assumptions for utility. Figure 1.14 shows both the willingness to pay implied by the marginal utility in the MLE estimates and the binary choice estimates using the truncated mean willingness to pay. I've aggregated the estimates across questions to provide a benchmark mean of all the questions and updated the standard errors to be clustered at the individual level. As a reference I first include log utility. The implication from log utility is that willingness to pay should increase linearly with income. I also included the implied results from the survey conducted in *Kimball et al.* (2015), labeled "Inequality Aversion" which asked people how much money given to a family with half their income is roughly equal to the impact of $1000 to family like theirs. This shows people's normative perceptions of what a dollar should be worth at various income levels. The graph shows that my estimates are clearly well outside the range of either benchmark. Note that I have normalized each estimate to intersect at the lowest income level. Any point could be used, but this would not lessen the difference between the estimates.

Figure 1.14



1.7. Discussion of Results

While there is some indication of slight diminishing marginal utility in one of the binary choice questions, the broad implication of the results is that higher income people will not pay significantly more for the same utility benefit. Using both assumption 1 and 2, this implies the marginal utility of income is constant across income groups, or as section 1.6.4 shows, at least diminishes slower than expected. How to interpret this result and what it means for economics depends on the exact assumptions we make about individual behavior as well as what we assume about the normative views of policymakers. While I could focus my discussion on the case where both assumptions are met, such a singular focus would short sell the findings. The purpose of this paper is not to convince behavioral economists that individuals maximize utility or to convince neoclassical economists that they do not. Either way, the findings have significant implications for our understanding of decision making and presenting the implications under stronger and weaker sets of assumptions demonstrates why this study and findings should not be ignored.

To start, I will consider what we can take away from these results with only the first assumption and not the second. That is, if utility for the survey questions is equal across income, but people do not generally maximize utility. Then, I will discuss the results once we add in the second neoclassical assumption of utility maximization.

1.7.1 Without Utility Maximization

This is a surprising outcome, and it has strong implications even with minimal assumptions. With only assumption 1, that the utility from pain relief in the survey scenarios is unrelated to income, we can say that, for the goods in my survey, higher income people do not pay more per util. With just this one assumption I have, at a minimum, identified a case where behavior deviates from economic intuition in a way not previously documented[12].

Next, if we also assume that this pattern is not due to some behavioral or psychological effect that applies only to my survey question, then we can generalize the finding to say higher income people do not pay more per util for any goods. Why would higher income people not pay more per util? There are many potential explanations, but the possibilities, outside of the neoclassical framework that I will cover in section 1.7.2, fit into two broad categories.

First, people might be making some kind of mistake. Suppose that higher income people would pay more (or lower income people less) if they were fully informed, there may be some psychological barrier preventing them from maximizing their own well-being. The exact reason or mechanics for this mistake could very widely, but one example would be an aversion to paying prices that seem unfair.

Alternatively, people may be willfully deviating from their own well-being. Recall from section 1.3 that utility maximization for assumption 2 requires more than that there exists some utility ranking that describes behavior. People need to actually be maximizing utility where utility

[12] I am of course also assuming that the responses on the hypothetical survey reflect real behavior.

has the same meaning we used to establish the interpersonal comparison in assumption 1. If that is done with preference utilitarianism and extended sympathy, assumption 2 holds almost by definition. However, if the policymaker is not a preference utilitarian, and instead has some substantive definition of utility, this may just not be the case. Individuals may have a different goal in mind.

To make this idea clearer, consider two simple but conflicting definitions of utility, pleasure, and happiness. Suppose individuals make choices to maximize their pleasure. So individual decision utility, and the utility corresponding to our utilitarian theory of action, is pleasure. At the same time, suppose a policymaker values the sum total of happiness. For this policymaker's utilitarian theory of ethics, which dictates what policy is the best and morally good policy, and equates utility across people for assumption, utility is happiness. This could cause a conflict for economists trying to use one utility function to represent both actions and policy assessment. How would we know if we were in this situation? How would we know if policymakers value something different than individuals? If this policymaker agrees with assumption 1, but also knows that \$100 creates more happiness, under their definition, when given to a high-income person than a low-income person. In this case, their definition conflicts with my estimates which shows that their definition of happiness does not correspond to individual decision utility. While thinking in terms of somewhat concrete ideas like pleasure and happiness makes the distinction more apparent, any differences in the definitions of utility could lead to the same result. In these cases, the policymaker's definition of utility is "wrong" in the sense that it does not correspond to the way economists think about utility, through preferences and individual action. That being said, from a moral or philosophical perspective, they cannot really be "wrong". I have no desire to dictate their values, but the results show their values do not align to individual decision utility.

Another way to think about this distinction is that people may, at times, consciously choose to do things that are not in their own best interest. Recall the above discussion from *Sen* (1977) where he draws this distinction with the ideas of sympathy and commitment. Sympathy is when an individual does a seemingly selfless act because it would make them feel badly not to or they feel good for doing it (like a warm glow). Commitment is when someone chooses an action despite it actually making them worse off than the alternatives. In any case, be it intentional or accidental, we have people consistently deviating from their own well-being. While the distinction matters for normative and philosophical considerations, in many cases they will be observationally equivalent and lead to a situation where there is a function describing individual behavior, and a second function describing actions that maximize their well-being.

What would it mean for there to be two different utilities. One for individual decisions and one for ethics and policymaking? Amartya Sen finds this idea appealing, saying

> A person is given one preference ordering, and as and when the need arises this is

> supposed to reflect his interests, represent his welfare, summarize his idea of what should be done, and describe his actual choices and behavior (*Sen*, 1977).

This is similar to the points made in *Kahneman* (1994) separating decision utility from experience utility. A single utility function to describe what is, ultimately, multiple different things may just be too simple. These two categories, mistakes or purposeful deviation, have different philosophical and normative implications, but are observationally equivalent in many cases.

Consider the following example. Suppose people will only buy a good if it both provides more well-being than its price and if the price is seen as fair. *Kahneman, Knetsch, and Thaler* (1991) document what people consider fair pricing to be, and the results do not clearly relate to individual well-being. If the fair price limit binds prior to the welfare limit, the practice of deriving well-being from the maximum willingness to pay, or decision utility generally, will be hindered. Even though higher income people value money less, they won't trade more of it for equal utility if the price seems unfair. This example could be viewed as either a commitment to fair prices, or a behavioral mistake that people are making. Either way, people are not doing what is in their best interest. The higher-income people in my survey might be better off if they paid more for relief, but a commitment not to pay what they view as an unfair price might hold them back. Or, a behavioral mistake of only purchasing fair prices, might keep them from a welfare enhancing purchase at a higher price.

This interpretation, while comfortable for some, would be a stark shift from conventional economic thinking. With that in mind, the following section returns to the neoclassical world and assumes individuals maximize their own utility and well-being.

1.7.2 With Neoclassical Utility Maximization

Now suppose assumption 2 holds. That is, individuals maximize their own utility. The results show that cardinal decision utility, which dictates individual choices, is roughly linear. The following welfare function will help to provide some context for the possible implications of a constant marginal utility of income.

$$\sum_i W(m_i) = \sum_i \gamma(U(m_i)) \tag{1.7.1}$$

In this equation, $W(m_i)$ is the total weight the social planner places on a person with income m_i. $U(m_i)$ is the utility at income level m_i and γ is a function that expresses the policymaker's preferences over the distribution of utility. Using these equations, a pure utilitarian would have $\gamma(U(m_i)) = U(m_i)$ and value only the sum total of utility[13]. A welfarist, but one who is not

[13] In economics the term utilitarian is sometimes used to describe any welfarist preferences since a specific

strictly utilitarian, might have a preference over not just the sum total utility, but how that utility is distributed. The policymaker's preference for the distribution of utility is represented with the γ function. Suppose a policymaker agrees bumping one's shin on a table costs the same amount of utility regardless of income, but this policymaker would rather a high-income person endure this pain than a low-income person since the higher-income person is at a higher total utility level. In this case $\gamma(U)$ will be concave in utility. I will refer to this type of policymaker as an egalitarian.

The results in this paper provide information about the shape of the $U(m_i)$ function, implying $U(m_i) = \alpha m_i + C$ where α and C are constant normalizations. This means a true utilitarian policymaker would have linear welfare weights and not desire redistribution. Conversely, this means that any policymaker that values redistribution is not a utilitarian. Rather, they must be an egalitarian. They have welfare weights that are concave in income as they value giving one util to someone with less income more than to someone with more income. As *Kimball et al.* (2015) showed, most people would fall into this category as most people think giving \$1 to a low-income person is better than giving it to a high-income person. In this survey specifically, they use the phrase "it would make a bigger difference" to the lower income person. While this has utilitarian connotations, it's not actually obvious what is meant by that phrase (*Kimball et al.*, 2015). Do they mean the lower-income person desires it more in the sense of a preference utilitarian, or do they mean it in some other way? Perhaps people see the things lower-income folks spend the money on as worthy or more impactful beyond preference or desire. In this case, it makes sense to think of these responses and the stated preference for redistribution as concavity in $\gamma(U)$. My results inform us about how people's motivations fit into the welfarist framework built on economic assumptions. We can see that $\gamma(U)$ is more concave than typically assumed. In other words, the welfarist, but non-utilitarian, portion of people's re-distributive preferences, i.e. γ, is more concave than expected. This also implies that very few people subscribe to a pure utilitarian view (under the economic definition of utility) because very few people favor no redistribution. This means the typical policymaker will need to be modeled as an egalitarian.

In the above discussion I am holding $W(m_i)$ fixed. That is, I am taking the re-distributive preferences of the policymaker as given and using the results of the paper to update how we model and think about those preferences. Alternatively, we can consider a policymaker with a fixed welfarist and preference utilitarian philosophy. In this case, how would these results cause them to update their re-distributive preferences, $W(m_i)$ relative to the prior with concave utility? A simple case is a pure utilitarian who only values the sum-total of utility, i.e. $\gamma(U(m_i)) = U(m_i)$, This policymaker goes from a concave to a linear social welfare function and would no longer

cardinal utility function is not typically know. However, this is actually an imprecise use. A utilitarian would value the sum total of utility and not have any welfare weights. If the policymaker has welfare weights, they are still welfarist, but not strictly speaking a utilitarian since they do not wish to simply maximize total or average utility.

value redistribution. For an egalitarian, however, the change in $W(m_i)$ is ambiguous. On the one hand, a linear utility means that a marginal dollar to a high-income person creates as much utility as a marginal dollar to a low-income one. Compared to a concave utility, this would push our egalitarian towards a less concave W. However, the impact on the distribution of total utility is unclear. If the slope of the linear utility is large relative to the policymaker's priors under concavity, total utility may be much more unequal than under our concave prior assumption. This would push W towards being more concave. How exactly one should think about the level changes in utility across income is not addressed in this paper.

Is a constant marginal utility of income plausible? While the idea does conflict with common economic intuition, it is important to take note of a few key points. First, a constant marginal utility of income still allows for diminishing marginal utility for particular goods. For example, consuming a tenth apple can be less enjoyable than the first because of convex preferences. People want to spread their money out rather than spend it on more apples, but this tells us nothing about what will happen if their income doubles. Cobb Douglass utility provides a simple example of preferences that are convex but, depending on the monotonic transformation used, can either have increasing, decreasing, or constant marginal utility of income.

Common intuition might suggest that even so, doubling consumption of everything, not just one good at a time, would still less than double cardinal utility. In the real world, however, if someone's income were to double, they almost certainly would not just double their consumption of the same goods they already purchase. People do not have homothetic preferences and so the quality and type of goods will change significantly as their income changes. This suggests that while the prior intuition may even be correct, it does not really apply to income changes in real life.

Non-homothetic preferences play another key role in why linear utility is plausible. The average marginal utility of income depends on the price vector in the economy. Thinking through how a price change could impact the marginal utility of income illustrates the role prices play. Suppose tomorrow all the goods lower income folks purchase on the margin, like necessities, become more expensive and lower quality. Suppose at the same time there are big efficiency gains to expensive luxury items purchased on the margin by higher-income folks (things low-income folks could never afford). After these changes, a dollar might just buy a lot more value in the hands of a wealthy person than it did in the previous period[14]. This idea is also exacerbated by the related idea that it is expensive to be poor. Berkouwer and Dean, for example, show that households in Nairobi are only willing to pay \$12 for a stove that would save \$237 over two years, and that a low interest loan increases willingness to pay to the actual savings over the life of the loan (*Berkouwer and*

[14] Similarly to the welfare function above, what this means for a policymaker's desire to redistribute between the two groups depends on if they value how utility is distributed in addition to its societal sum

Dean, 2021). Lower income folks may end up buying lower quality goods that do not last as long or provide as much value and so receive a lower quantity of quality adjusted goods per dollar on the margin.

Another point to consider is that our intuition about what utility is under assumption 2, that preferences reveal well-being, may simply not be very good. Under this assumption, the marginal utility of income is really about people's attitudes towards spending. How willing is a given consumer to part with money for a given gain? How flippant are they with cash? How much do they desire money relative to particular goods? It appears that higher income people do still desire money. This desire just might not stack up well with policymaker's ethical beliefs about who is deserving of this money or where it will do the most "good". As explained above, we can adjust the $\gamma(U)$ in our welfare functions to realign our policymaker's perceptions of fairness even with a linear utility function 2. Crucially, however, this is only possible so long as individuals are still maximizing their own well-being. If desire and individual actions are not even intrapersonally consistent with what a policymaker considers "good" or welfare maximizing, then this cannot be reconciled by manipulating $\gamma(U)$. This case brings us back to the discussion in section 1.7.1 where people are not maximizing well-being.

1.7.3 The Potential For Bias

A third possibility that I have not yet touched on is that the results are biased. I do not see this as the most likely outcome, but I do want to point out weaknesses in the analysis that I hope can be improved on in future work.

The biggest area of concern is selection bias into the survey. What kind of high-income person does a survey for a few dollars? One who is willing to do things for little money. This is a characteristic that will lead them to systematically give lower answers than their average peer. This motivated me to originally go with the company Pollfish. They collect responses from phone app users and pay the users with "in app" benefits related to the apps purpose like a free yoga lesson or news article. My expectation is that a wealthy person is more likely to do a survey to get free lives while playing a game like candy crush, because it is viewed more as an ad experience, than to sign up to do surveys for money, which may be viewed more as a job. This is speculation on my part, and the sample in the pilot I did with Pollfish was not representative of the United States on observables. The non-representative sample is what convinced me to switch to Centiment. However, the results from the Pollfish survey show more of an increase marginal utility of income with income (albeit still outside the range of log utility). The online appendix shows the mean figures and MLE results for the questions in my pilot. For the pain questions, the marginal utility for the 0-25 group was high, which I expected was due to non-working spouses or temporary unemployment making up a disproportionate share of that group. However, after this initial income group, we see an increase in WTP, implying a decrease in marginal utility of income. The sample is smaller and does not reflect national demographics, and so I do not place much weight on these. However, the difference across these two panels suggests another sampling approach would be a worthy endeavor. Other ways to address sampling bias in the future would be to increase the reward for participation in the survey so that most people would be willing to do it, or to decrease the monetary reward to zero so the selection into the survey is based on altruism only and not a willingness to do tasks for little monetary reward.

Another general concern is that it is a hypothetical survey and not properly incentivized. This is partially alleviated because the analysis does not require the absolute value of reservation prices to be at all accurate. I just need it to be the case that the ratio across incomes is consistent with real purchasing activity. This issue could be solved with an incentivized experiment, which I hope to follow up with in the future.

It is also possible pain tolerance changes significantly across income. Beyond showing that the PSQ does not change across income, there is not much that is possible to say on the matter. The best way to address this concern would be with follow up studies using other goods that are conceivably uncorrelated with income, but not based on pain. In the Pollfish pilot I also asked questions about disgusting scenarios. The variance in these questions was much higher. My intuition is that there is much more variance in what people find disgusting and how much of it they can handle. This convinced me to focus on the pain questions in my final survey in order to make the most of my limited budget, but perhaps another look at things like disgust is warranted in future work.

1.8. Conclusion

Measuring utility is a difficult problem with a long history (*Moscati*, 2018). Without any structure, measuring the concavity of utility with respect to income is not even a well-defined problem, but with the right structure and data it is empirically possible. I present a model for a feasible identification strategy for the average marginal utility of income and implement that strategy with new survey data. The results suggest that the marginal utility of income is constant across income groups, implying that utility is roughly linear in dollars. this result has significant implications for our understanding of individual preferences, utility, well-being, and distributional ethics.

CHAPTER II

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

Written with Tanner S. Eastmond, Michael David Ricks, and Julian Betts

2.0 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 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, and Walters*, 2022; *Bacher-Hicks and Koedel*, 2022); doctors, hospitals, and nursing homes (*Chan, Gentzkow, and Yu*, 2022; *Chandra, Finkelstein, Sacarny, and Syverson*, 2016; *Doyle, Graves, and Gruber*, 2019; *Einav, Finkelstein, and Mahoney*, 2022; *Hull*, 2020); and even judges, prosecutors, and defense attorneys (*Abrams and Yoon*, 2007; *Harrington and 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, and Sims*, 2014), attendance and suspensions (*Jackson*, 2018), and work ethic and learning skills (*Petek and Pope*, forthcoming). 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, and 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, and Novan*, 2021; *Hollingsworth and Rudik*, 2019; *Sexton, Kirkpatrick, Harris, and 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

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

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 teachers 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,*

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

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 education literature have often focused on math scores for a variety of reasons (*Bates et al.*, 2022; *Chetty, Friedman, and 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 and 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 et al.*, 2009; *Staiger and 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 preferences. 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 and Wager*, 2021; *Kitagawa and Tetenov*, 2018) and is reflected in a recent explosion of empirical inquiry about targeting

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

treatments as varied as social safety programs (*Alatas, Purnamasari, Wai-Poi, Banerjee, Olken, and Hanna*, 2016; *Finkelstein and Notowidigdo*, 2019), costly energy efficiency interventions (*Ida, Ishihara, Ito, Kido, Kitagawa, Sakaguchi, and Sasaki*, 2022; *Ito, Ida, and Tanaka*, 2021), promoting entrepreneurship in developing countries (*Hussam, Rigol, and Roth*, 2022), and even resources to reduce gun violence (*Bhatt, Heller, Kapustin, Bertrand, and 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 measuring 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, and 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

$$\Delta\widetilde{\mathcal{W}}^p \equiv \sum_{i=1}^{n} \gamma_i(S_i^p, S_i^0)\Delta S_i^p \tag{2.2.1}$$

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 B.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 and 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, introducing bias. Second, as shown in Appendix B.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 4. *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 2.2.1:*

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

Proof in Appendix B.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 B.2.4.

[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.

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 5. *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:*

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

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 4.

First, consider the covariance terms. The covariance term in Theorem 4 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

$$\left|\sum_X P_x \mathrm{Cov}(\gamma^p, \Delta S^p|x)\right| < \left|\sum_X P_x \mathrm{Cov}(\gamma^p, \Delta S^p|x) + \mathrm{Cov}(\mathbb{E}[\gamma^p|x], \mathbb{E}[\Delta S^p|x])\right| = \left|\mathrm{Cov}(\gamma^p, \Delta S^p)\right| \tag{2.2.2}$$

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

$$2 * \left|\sum_X P_x \mathrm{Cov}(\gamma^p, \Delta S^p|x)\right| < \left|\mathrm{Cov}(\mathbb{E}[\gamma^p|x], \mathbb{E}[\Delta S^p|x])\right| \tag{2.2.3}$$

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 unconditioned 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.

$$\sum_x P_x E[\gamma^p|x]\left(\mathbb{E}[\Delta S^p|x] - \widehat{CATE(x)}\right) < \mathbb{E}[\gamma^p]\left(\mathbb{E}[\Delta S^p] - \widehat{ATE}\right) \tag{2.2.4}$$

A more formal treatment can be seen in Appendix B.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.

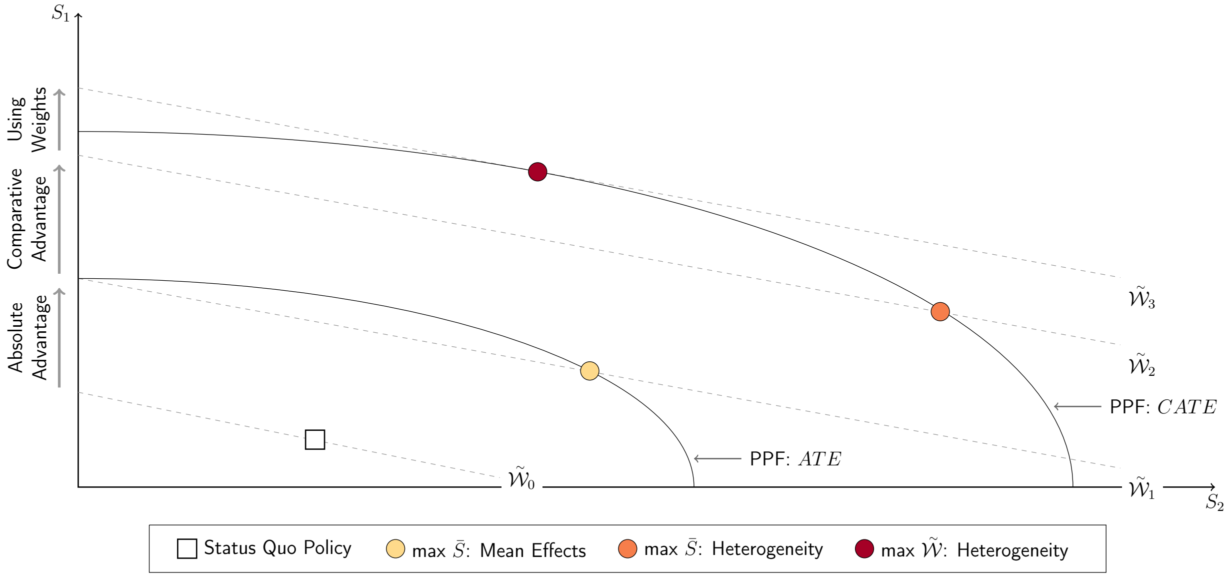
Given these differences, it is worth noting that there is no reason one could not condition the welfare weights and the estimates on different subsets of $\boldsymbol{X}$. for example, $E[\gamma^p|x_1]\ \widehat{CATE(X_2)}$. It might be the case that a variable is not meaningful in the welfare weight, 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.

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

[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.

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 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

[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.

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

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

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, and 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 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.

2.3.2.1 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 B.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:

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

[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.

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.

$$\hat{\tau}_s^{j,t} = \hat{\boldsymbol{\psi}}_s \bar{\boldsymbol{A}}_s^{j,-t} \tag{2.3.1}$$

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 B.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]

2.3.2.2 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 B.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:

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

[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.

$$\hat{\tau}_{g,s}^{j,t} = \hat{\boldsymbol{\psi}_{g,s}} \bar{\boldsymbol{A}}_{s}^{j,-t} \tag{2.3.2}$$

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 B.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.

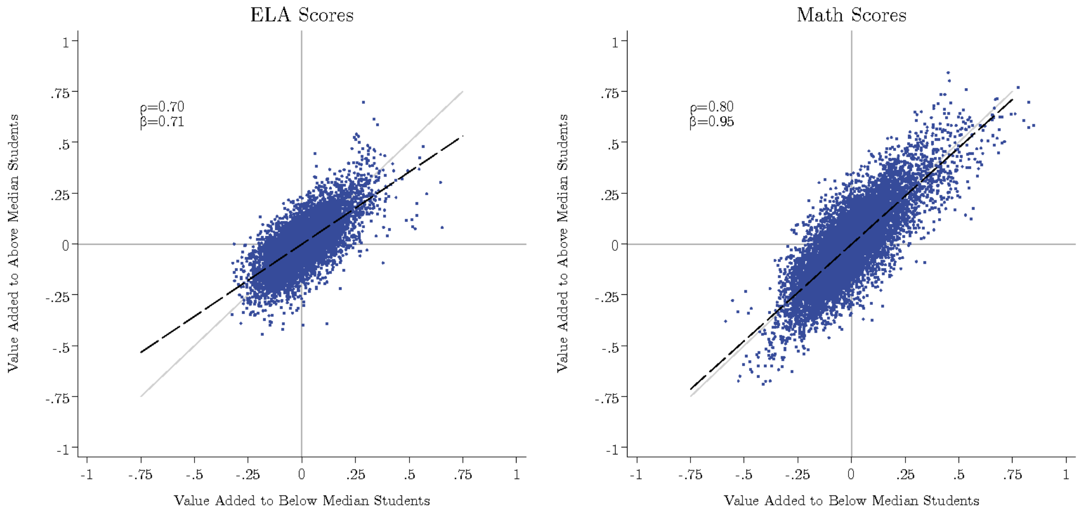
2.3.3 Heterogeneity Highlights the Importance of Comparative Advantage

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, suggesting 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.

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.

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 B.1 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.

2.3.3.1 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.

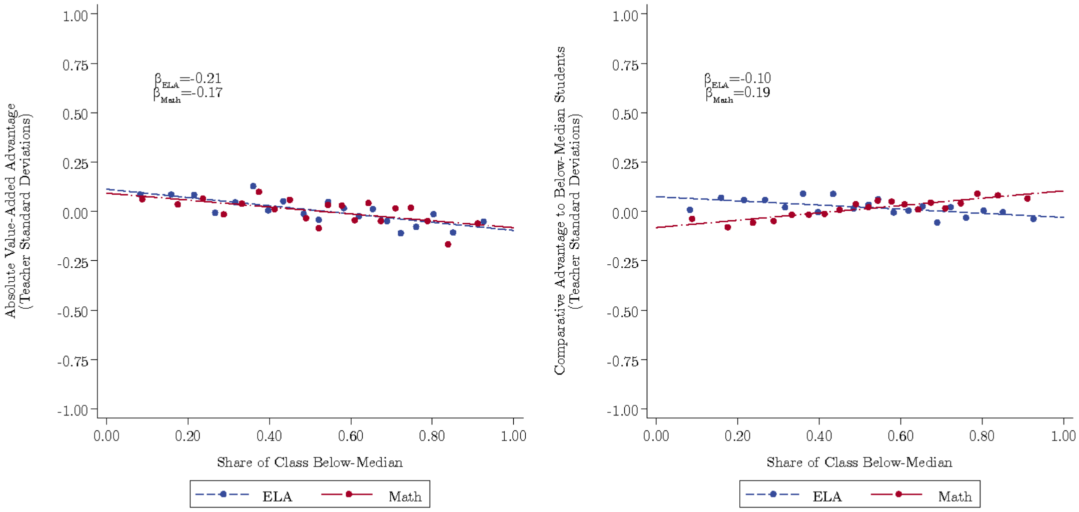
First, Appendix Figure B.6 reports patterns of persistence over time. For example, over 40% of teachers have a comparative advantage for teaching one group of students in *all* years, and the year-to-year correlation is between 0.78-0.90 for all estimates. Additionally, Appendix Figure B.7 leverages 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.

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 average 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 B.2 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.

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).

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 B.2). 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 B.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 $\widetilde{\mathcal{W}}$ 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):

$$\max_{\mathcal{J} \in \mathscr{J}} \widetilde{\mathcal{W}}(\mathcal{J}; \omega) = \max_{\mathcal{J} \in \mathscr{J}} \frac{1}{N_{i,t}} \sum_{(i,t)} \omega_L \, L_{i,t} \, \hat{\tau}_L^{\mathcal{J}(i,t)} + (1 - \omega_L)\,(1 - L_{i,t})\, \hat{\tau}_H^{\mathcal{J}(i,t)} \tag{2.4.1}$$

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 2.4.1 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 and Ralphs*, 2009).

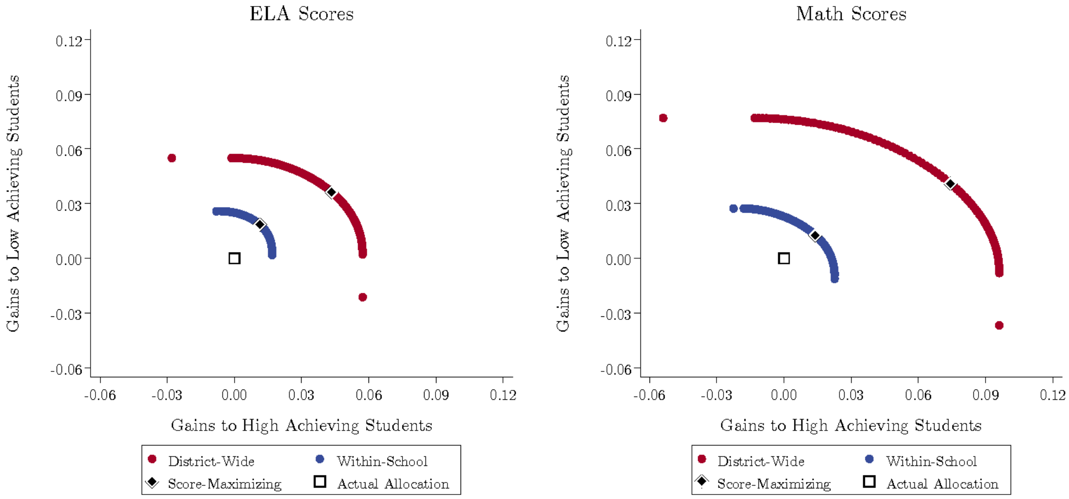
2.4.1 Allocations Incorporating Heterogeneous Impacts Increase Test Scores

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 2.4.1 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 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.

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.

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*, 2011; *Hanushek et al.*, 2009).

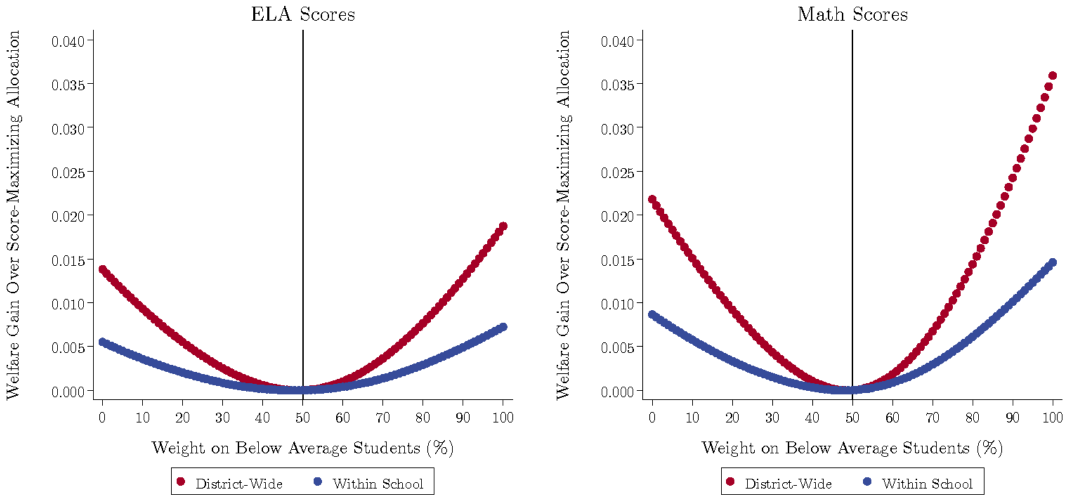
The second pattern visible in Figure 2.4 is that the curvature of the PPFs demonstrates the value in explicitly considering the distributional goals of a policymaker. These gains are dependent on the extent to which distributional goals deviate from the mean scores objective but are large for more extreme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

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

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.

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.

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 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, and 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:

$$\max_{\mathcal{J}\in\mathscr{J}} \tilde{\mathcal{W}}_{VA}(\mathcal{J};\omega) = \max_{\mathcal{J}\in\mathscr{J}} \frac{1}{N_{i,t}} \sum_{(i,t)} \omega_L \, L_{i,t} \, \hat{\tau}_{VA}^{\mathcal{J}(i,t)} + (1-\omega_L)\,(1-L_{i,t})\,\hat{\tau}_{VA}^{\mathcal{J}(i,t)} \qquad (2.4.2)$$

where $\hat{\tau}_{VA}^{j}$ is the standard value estimate described in section 2.3.2.1 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.

2.4.2.1 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.

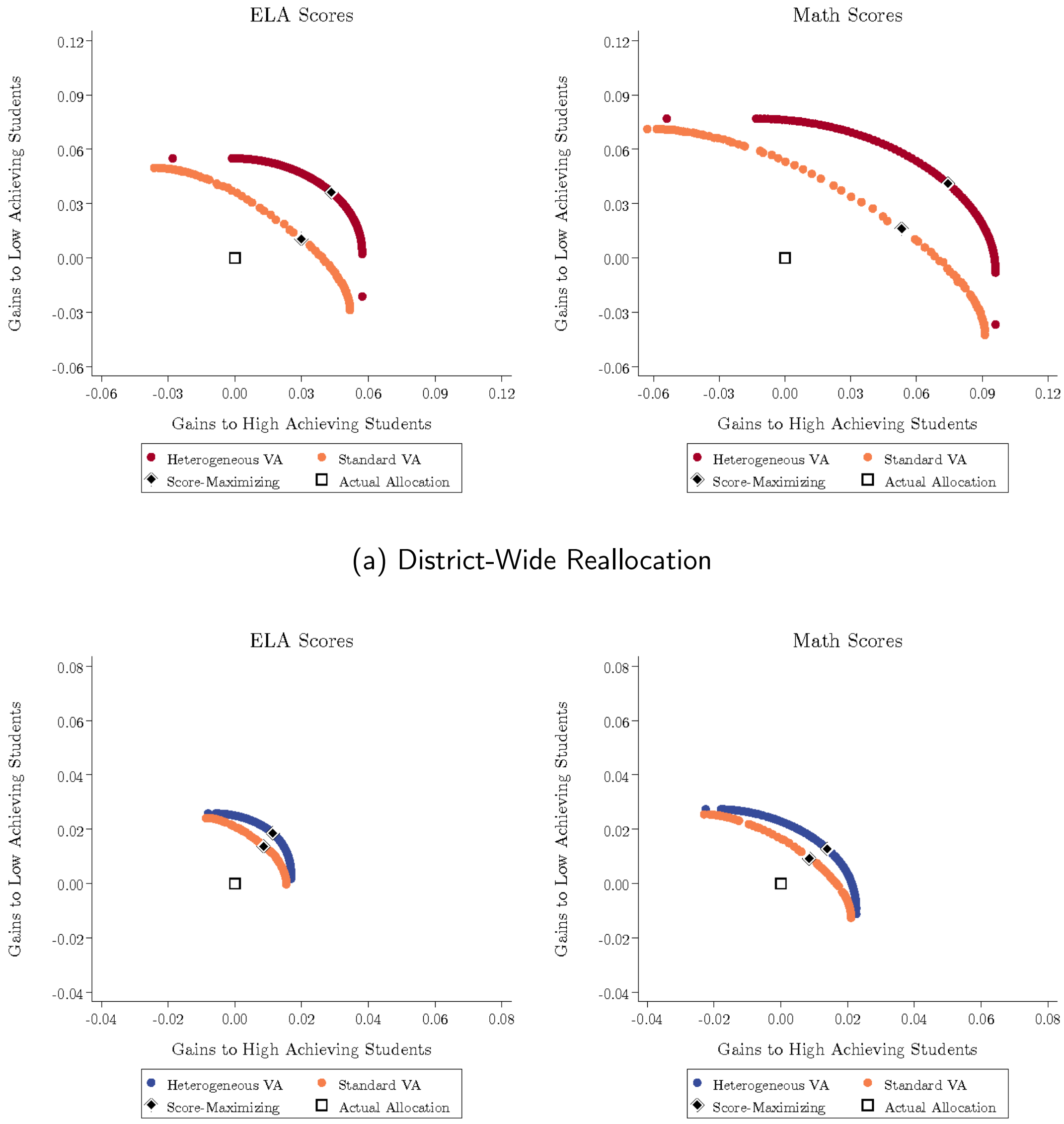
To approach this question, we depict and compare the production-possibility frontiers for average achievement gains to each group using heterogeneous and standard value-added in Figure 2.6. Here again each point presents the average change in lower-scoring students' test scores in the optimal allocations (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 advantage 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 be-

Figure 2.6: Using Heterogeneous Estimates Produces Larger Gains from Reallocation



(a) District-Wide Reallocation

(b) Within-School 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.

tween 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.

2.4.2.2 The Interaction of Distributional Goals and Comparative Advantage

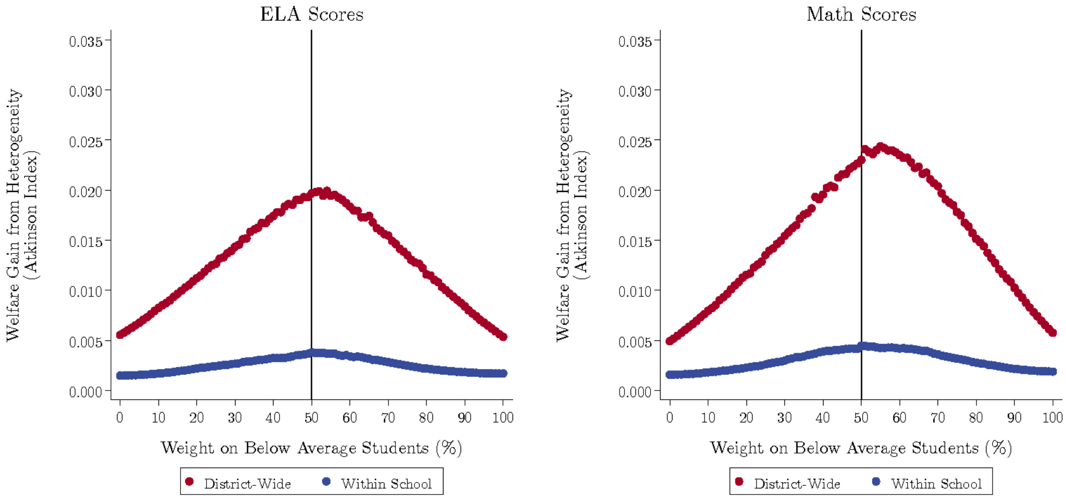
The above section shows that when the goal is to maximize average scores, using heterogeneous value-added leads to significant gains. We also know from section 2.4.1 that when policymakers favor one group over another, considering their distributional goals leads to significant welfare gains. Putting 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 B.3, we present a simpler measure: the true (unweighted) difference in average scores for each pair of allocations.

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-

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.

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 B.3. 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 comparative 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

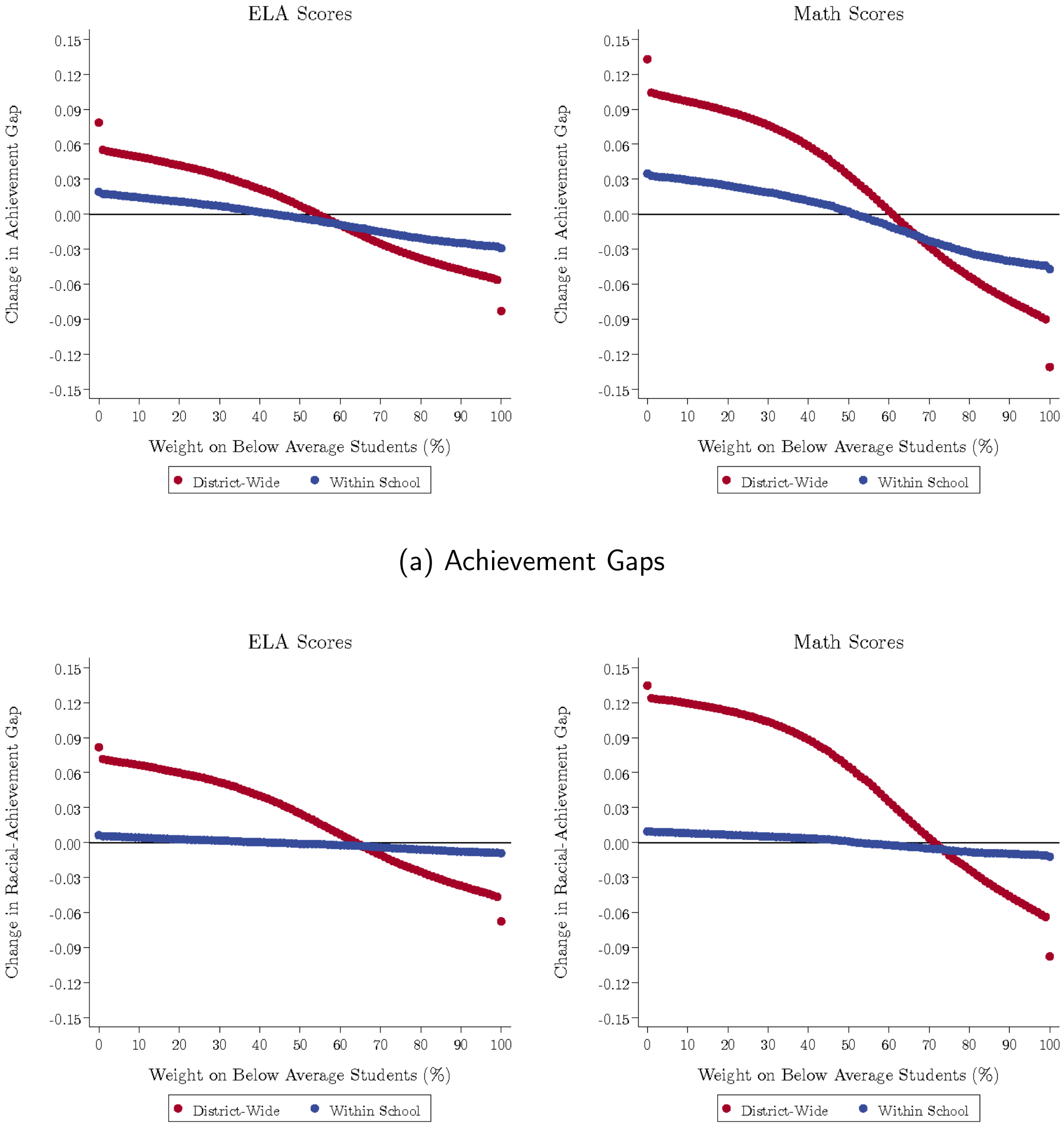
Having described the optimal reallocations and decomposed the welfare gains from them, our final task is to explore other equity implications that the proposed reallocations would have. Specifically, we study how our reallocations affect overall achievement gaps and racial achievement gaps, and we describe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

2.4.3.1 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 white and Asian students are smaller: at 0.72σ in ELA and 0.63σ in Math, and these gains

Figure 2.8: Reallocations Can Shrink Persistent Gaps in Student Performance



(a) Achievement Gaps

(b) Racial-Achievement Gap

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.

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 utilitarian, 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.

2.4.3.2 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

[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.

"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 B.4 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 B.4 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

[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.

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 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, and Pardo*, 2023; *Jacob and Lefgren*, 2007; *Petek and Pope*, forthcoming). 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, and Walters*, 2020) or housing markets (*Imberman and 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 5. 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 2.2.2 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 and 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 subsection demonstrates our approach for finding the optimal achievement gains for students' lifetime earnings and performing allocations that maximize those income gains.

[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.

2.5.1.1 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]

$$\widetilde{\mathcal{W}}(\mathcal{J};\omega) = \frac{1}{N_{i,t}} \sum_{(i,t)} \sum_{s} \omega_s \left[L_{i,s,t}\, \hat{\tau}_{L,s}^{\mathcal{J}(i,t)} + (1 - L_{i,s,t})\, \hat{\tau}_{H,s}^{\mathcal{J}(i,t)} \right] \tag{2.5.1}$$

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 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

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

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.

2.5.1.2 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 define a new social welfare function to optimize

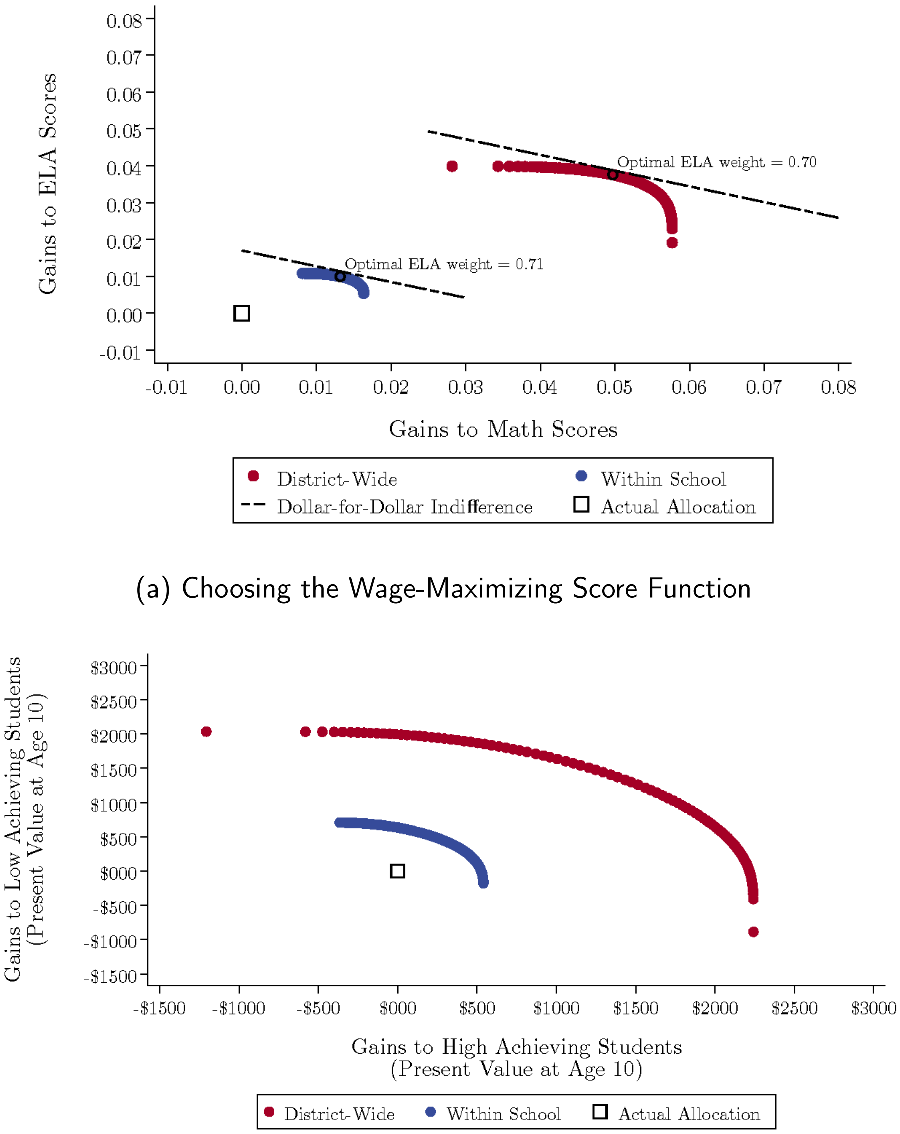
$$\begin{aligned}\widetilde{\mathcal{W}}(\mathcal{J};\omega) = \frac{1}{N_{i,t}} \sum_{(i,t)} \omega_L & \left[\omega_{\mathsf{ELA}}\, L_{i,\mathsf{ELA},t}\, \hat{\tau}^{\mathcal{J}(i,t)}_{L,\mathsf{ELA}} + (1-\omega_{\mathsf{ELA}}) L_{i,\mathsf{Math},t}\, \hat{\tau}^{\mathcal{J}(i,t)}_{L,\mathsf{Math}}\right]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

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 $\omega_{\mathsf{ELA}} = 0.75$ and $\omega_{\mathsf{Math}} = 0.25$ 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

[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.

Figure 2.9: Reallocations Can Shrink Persistent Gaps in Student Performance



(a) Choosing the Wage-Maximizing Score Function

(b) Present Value Income Gains

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).

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 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 B.5, 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 welfare under the resulting allocation would be the marginal value of public funds (MVPF, *Hendren and 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

$$MVPF^j(b) = \frac{\sum_i (1-t)\Delta S_i^p)}{N_j b - t\Delta S^p_)} \tag{2.5.2}$$

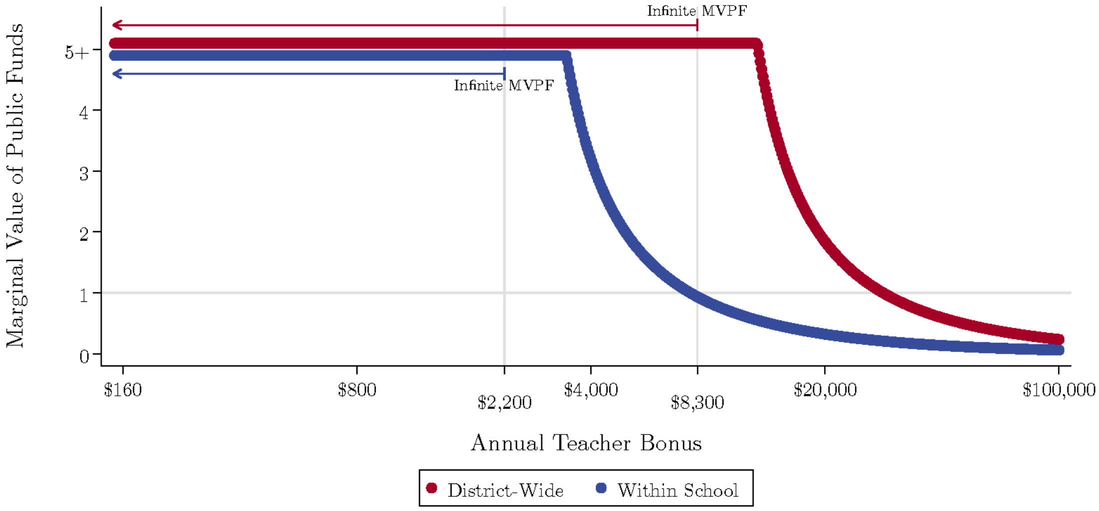
where $(1-t)\Delta S_i^p$ are the after-tax present-value monetary gains to each student from allocation j (given tax rate t), N_j is the number of teachers and $t\Delta S_i^p$ is the present-value of gains recouped as tax revenue. The key assumption required for this statistic to be meaningful in this 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, and Ly*, 2023), we focus on this simplified case as in other work (*Hendren and Sprung-Keyser*, 2020).(*Hendren and Sprung-Keyser*, 2020).[18]

We combine our estimates of present-value monetary gains with data from the Opportunity Atlas (*Chetty, Friedman, Hendren, Jones, and 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 and 0.06, implying $t = 0.28$ for in equation 2.5.2.

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.

[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.

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, and 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*, 2011; *Hanushek et al.*, 2009). Although deselecting 5% of teachers with the lowest value-added could produce large gains, there are concerns about the ethics of mistakes (*Staiger and 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 III

I'll Have What They're Having: State Fiscal Policy Interdependence

3.0 Abstract

This paper considers how preference formation may directly lead to state fiscal policy interdependence. I start by laying out a formal model for state fiscal interdependence. The model is built on two core ideas. First, voters look at "similar" states via news coverage to determine what a normal level of public spending is. Second, governments respond to these shifting preferences by maximizing the probability of reelection. The model informs an empirical analysis which uses state newspaper articles to form a new metric of state inter-connectivity. This metric is compared against established metrics used in the literature.

3.1. Introduction

If a coworker or neighbor buys a new car, house, or television I might decide it is a good time for me to do the same. Part of the reason I might feel that way could be driven by a psychological drive to "keep up with the Jones's". This colloquial term draws on the idea of conspicuous consumption first laid out by *Veblen* (1899). The core idea is that I may buy things specifically to signify social status, but conspicuous consumption also implies that my consumption and perhaps utility is relative to my neighbor's consumption. This type of behavioral reaction is likely familiar to many people, but the observed behavior of "keeping up with the Jones's" could also be driven by a more rational economic motivation. Information and our ability to process it is limited. We may look to others as an example of what is possible and what a specific decision would look like if we were to follow through on it. I may know my coworker is about as well off as me, if they can buy a house, why can't I? It's not necessarily that I need to keep up with the Jones's, but that I should be able to.

While preferences are often treated as fixed, immutable characteristics in economics, these ideas demonstrate the importance of considering how preferences are formed and evolve. Preference formation can clearly play a role in individual decision making, but preference formation could also lead to changes in collective decision making. Suppose that individuals in one state look to individuals in other states to both get a sense of what a standard level of taxes or government spending are or to gain information about what a feasible set of policies for the government to offer are. If governments respond to these changes in individual preferences, we will see states react as if they too directly respond to their neighbors. While this could be expressed in a number of policy channels, this paper specifically examines the following question. To what extent is state fiscal policy influenced by other states' policies, and could this relationship be driven by individuals looking to other states to form their own preferences over policy.

This question may be interesting on its face to researchers of government decision making and public policy, but there are implications beyond this as well. If this effect is significant and predictable, any study analyzing the impact of an event on state spending should control for it, and it could already be biasing a number of research designs. While previous research has shown this to be the case, more recent evidence is more mixed (*Agrawal, Hoyt, and Wilson*, 2022; *Baskaran*, 2014; *Lyytikäinen*, 2012). The changing and inconsistent dynamics are a good reminder that estimates of state interdependence are not structural parameters, and will likely very with the timing, policy, and a myriad of other factors. This is why continuing to probe and understand the mechanisms behind these relationships is so important.

To answer this question, I start with a formal model. The purpose of this model is to demonstrate minimal assumptions for this preference formation mechanism to be plausible. The model is structured around the idea that an individual's preferences for taxes and government spending are influenced by what they see happening in neighboring states. These changes in preferences translate into changes in the probability an individual will vote for the incumbent government. The government proposes policy in order to maximize the probability of being elected. Together, these imply the government responds to neighboring states as well. using monotone comparative statics, I show that with some general assumptions on the shape of an individual's vote plurality function, when one state increases government spending, neighboring states will as well.

Building on this model, I develop an empirical strategy to test my theory that preference formation drives interstate fiscal policy transmission. I run regressions, including instrumental variable regressions, to determine the impact of state fiscal policy on neighboring states' fiscal policy. Throughout the discussion I use the term "neighbor" to refer to states that influence spending, and not necessarily geographic neighbors. This is an important distinction as a crucial first step in determining if states are impacted by their neighbors is determining who state's see as a neighbor. My hypothesis is that the interstate connection is driven by individuals updating their beliefs and preferences about government policy based on what they see other people doing. If this is the case, then the network of state connections will be determined by who (meaning which states) people observe as an example. I capture this by analyzing how often state newspapers mention other states in policy relevant news. Once I have determined which states are neighbors with which, I use state spending data and an IV strategy to test for state fiscal interdependence. My new metric for connectedness is then compared to previous metrics used in the literature, which found significant state interdependence *Baicker* (2005); *Case, Rosen, and Hines* (1993).

This empirical exercise does not find a significant relationship between states for either my new network or the networks used in older research applied to more recent data. These results lay the groundwork for additional research investigating state fiscal policy transmission. These results are a valuable reminder that state fiscal policy diffusion, like policy diffusion more generally, is not an immutable characteristic that can be measured and marked for all of history. Rather, it is likely the result of a complex process depending on many factors that change and adapt over time. As such, continuing to investigate and probe these relationships as politics, society, and policy making change is vital to an up to date understanding of state policy interdependence.

3.2. Background

The question of how policy interacts and diffuses across states is complex and, as such, has produced an extensive literature. Discussing this full body of work is beyond the scope of this article and is large enough to fill multiple survey papers[1]. I will, however, point to specific papers, models, and results that contextualize and motivate my approach and contribution to this large literature.

There are multiple theories for why and how policies may diffuse across localities, but to my knowledge, the idea of preference formation driving these patterns has not been seriously considered in the economics literature. Perhaps one of the closest, and most general theories, is the spillover model. This theory was developed in *Case et al.* (1993). They model state interaction when the optimal expenditure of one state is a function of expenditures in another state. For this model, the author's give examples of direct spillovers like the idea that "one state's expenditures on roads may provide benefits to the residents of neighboring states who can use the roads." While these are the types of examples attributed to this model, for example in *Agarwal, Vyacheslav, and Scholnick* (2016), it is a much more general model. The optimal expenditures could depend on neighboring states for a variety of reasons, including my theory

[1] For a detailed review, readers should see *Agrawal et al.* (2022); *Brueckner* (2000,0); *Devereux and Loretz* (2013); *Gresik* (2001); *Keen and Konrad* (2013); *Revelli* (2005); *Revelli et al.* (2006); *Wilson* (1999,9); *Wilson and Wildasin* (2004)

that individual preferences shift based on what neighboring states are doing, and, because of this shift in preferences, the optimal shifts. Due to this connection, I will refer to my hypothesis as individual preference spillovers and refer to more direct examples, like roads, as benefit spillovers. The advantage of such a broad model is that it can incorporate a broad range of potential explanations. However, this advantage is simultaneously a challenge. Such a broad model cannot speak to the idiosyncratic assumptions needed or observable changes to expect from preference formation spillovers compared to benefits spillovers. By narrowing the model, I can identify the assumptions necessary for preference spillovers to play a key role and motivate an empirical approach to try and differentiate the two effects.

While I consider my approach to be a subset of the classical spillover model, it shares characteristics of other existing models. The Yardstick Model more closely matches my theory in spirit. In the yardstick model, voters attempt to discern between the two types of governments (*Besley and Case*, 1995; *Besley and Smart*, 2002). The main idea behind this model is that governments are either good or bad. Good governments deliver public goods at cost and charge necessary taxes while bad governments are rent seeking and try to charge extra and, somehow, keep the difference (*Besley and Case*, 1995). How exactly they extract the rents is unclear, perhaps by fixing contracts or just bags of cash in kickbacks. Given government can be good or bad, voters want to determine what type their government is and either reelect them or vote them out. The trouble is states receives economic shocks unobservable to voters. So, in isolation, voters cannot tell if taxes are high because the economy is bad, or because they have a corrupt government. Their neighbors, however, have correlated economic shocks. So, to make this distinction, voters can observe the government in those neighbor states to gain information and update their beliefs about their own government. The yardstick model incorporates the idea of individuals looking to their neighbors and changing their behavior based on what they see, but it also assumes people have fixed preferences and are updating their beliefs about their own government. It also makes the assumption that governments are either competent or incompetent and voter's primary goal is to discern the government's type. While there is evidence that introducing performance assessments lowers regional spillovers in the UK (*Revelli*, 2006), I have not seen any evidence presented to support this as a primary driver of voter behavior in the united states, and it does not seem to match up to the advertisements and focus for most political campaigns.

My model moves away from the dichotomy of a benevolent or selfish government altogether. Politicians are not choosing to be benevolent or extract rents. What governments are actually choosing is where to allocate resources, and voters respond based on their preferences for public goods and taxes. My model also incorporates the government budget constraint to explicitly include fiscal policy into the model in addition to tax levels. I describe this idea in more detail along with the mathematical model in the next section.

While these two models are the closest starting point for individual preferences spillovers, the other main class of model that may play a role in state interdependence are Tiebout Models. In Tiebout Models, households costlessly move between states in order to find a match that fits their desired level of public spending (*Tiebout*, 1956). While this explanation is common in tax competition, I am not convinced state fiscal policy is the appropriate setting for this model. Interstate mobility is not a large enough factor in political decision making to account for the empirical relationships historically observed in fiscal policy (*Baicker*, 2005; *Case et al.*, 1993). Interstate mobility is low. It was between 1.4 and 2 percent from 2005 to 2017 (*Frey*, 2017). Moreover, I've seen no evidence it plays a significant role in political campaigns at the state level. For smaller jurisdictions, where movement is more prevalent, or for taxes focusing on more mobile businesses rather than fiscal policies impacting individuals, these models make more sense.

My model shows that the preference formation hypothesis requires plausible assumptions and could reasonably play a role in interstate fiscal dependence, but going beyond plausibility and discerning its actual impact requires an empirical exercise. I follow the thread of empirical work starting with *Case et al.* (1993). They found that the "effect of a dollar of increased spending by a state's neighbors increases its own spending by about 70 cents" (*Case et al.*, 1993). This approach shows a pattern of correlated state spending, but differentiating between types of spillovers, preference formation or benefit spillovers, is difficult as they would both lead to correlated spending. A key step in their empirical approach is to choose a weighting matrix that determines who is neighbors with whom and how to weight these relationships. This step provides an opportunity to build evidence for a particular causal mechanism (even if it cannot provide a formal statistical test). *Case et al.* (1993) settle on using racial composition in the form of percent Black as a metric for determining which states are connected. I would characterize their argument for this measure as insisting it is a proxy for some underlying connection between the states rather than the primary mechanism linking states (*Case et al.*, 1993). I, instead, build my weighting matrix using the theory of preference formation. I analyze how often state newspapers mention other states in policy relevant news to determine the extent to which residents in one state are likely to look into the happenings of another. If this weight matrix, which connects directly to preference formation spillovers, leads to a stronger relationship, this will provide evidence for it as the main mechanism for those spillovers.

While *Case et al.* (1993) started the empirical work I build on, in an important follow up *Baicker* (2005) again addresses this question. She replicates the weighting matrices used in *Case et al.* (1993) but also tries to connect states using interstate mobility. While this interstate mobility fits in with the general spillover effects hypothesized in Case Hines and Rosen 1993, this method shifts the theoretical focus to more of a Tiebout type competition model (*Tiebout*, 1956). If states are in fact competing for residents, then it makes sense that a given state's fiscal policies would most closely reflect states where their residents are considering moving. I closely follow the advancements she made with her empirical strategy. She addresses the issue of correlated errors by adding instrumental variable regressions using instruments based on changes to Medicaid spending commitments and neighboring state's covariates.

An important point to keep in mind while looking through past and more recent literature on policy spillovers is that even in the specific context of states and fiscal policy, the amount of spillovers and even the mechanism is not a fixed immutable characteristic. Over time, how policies spread across states very well may change. Not only could the mechanism change, but all of the hypothesized mechanisms depend on a myriad of factors that may be changing over time as well. Given this reality, frequent research probing policy spillovers for a variety of policies is essential for maintaining a working knowledge of how and why policy spreads across the country. Recent evidence presented by *DellaVigna and Kim* (2022) shows that the ever-changing nature of policy spillovers is more than just a possibility. They show that for a large set of binary adoption policies, proximity best predicted policy diffusion from 1950-2000. Since 2000, however, political alignment has better predicted diffusion for the policies in question. By looking at the impact of party control, they argue political polarization is likely the driving factor in the change in trends.

3.3. Model

3.3.1 Model Intuition

My model for how individual preference formation translates into state fiscal policy spillovers stems from the following ideas. The first is that politicians are trying to win elections. They do this by splitting up the state's pool of resources between the private sector and the provision of various public goods to different groups. Ultimately, their decisions are all about resource allocation, and, more specifically, matching that resource allocation to the preferences of their electorate.

Second, all else equal, a given voter is more likely to vote for a politician the lower their own taxes are, and the more money is spent on public goods. While individuals are willing to trade taxes for public services at different rates [2], if you offer someone lower taxes and no other changes or higher services with no other changes, their opinion of you should not worsen[3]. Together these imply a politician wants to allocate resources between programs and tax breaks to maximize their chance of reelection.

Third, voters' preferences and beliefs are influenced by their neighboring state's policies. How

[2] see the two political parties in the United States

[3] at least in most cases

might a typical voter decide if they have enough spending on education? I expect at least part of a typical voter's strategy is to look to states that they see as similar. Anecdotally, I saw this type of behavior growing up in Wisconsin. Our policies were constantly compared to Minnesota and sometimes to Illinois or Iowa or other Midwestern states. This happened casually in news stories or with more rigorous detail. For example, comparing jobs numbers or other indicators across state to determine if Wisconsin was doing enough.

Part of the reason voters might do this is psychological. People have a preference to be as well off or better than others they see as their peers or to "keep up with the Joneses". This idea is modeled with reference dependent utility functions similar to prospect theory where a voter gains utility for a good relative to what they see as the norm (*Kahneman and Tversky*, 1979). I use the idea of a reference point from prospect theory but not the idea of loss aversion.

Another reason voters may look to their neighbors is that people may not have a good sense of what is reasonable to expect the state to provide. More education spending is better all else equal, but all else is never equal. Voters need to estimate how much money can be allocated to policies they care about without drawing from their personal taxes or losing the election. Looking to existing examples in neighboring states could assist in that estimation. This would alter a person's probability of voting for a given candidate at the same level of spending. This seems like a reasonable strategy to me given limited information and attention of voters. Both of these mechanisms could also be accentuated by interest groups and politicians using neighboring state's policies as rhetorical ammunition to stir up support for their interests.

Putting these ideas together guides a reasonable path to the hypothesis that state spending is connected between states. Preferences of voters change relative to their neighboring states. For example, when Minnesota spends more on education, Wisconsinites, on average, want to spend more as well. Politicians see this and realize they will lose votes if they do not shift resources to increase education and respond to these shifting preferences.

Empirically this should lead to a correlation in state policies, both fiscal policies and taxes, between states that see each other as neighbors. This will be more than the correlations we would expect from just correlated economic shocks or general similarities between states. The states that are neighbors in this model are states that voters are paying attention to and making political comparisons to. This could be states that are fundamentally similar, but it could also just be states where information is easily accessible. For example, states that are mentioned frequently in the news. It could even be states with a particularly strong cultural influence on voter's reference point for their utility functions (preference taste makers).

While the story here is hopefully clear, the assumptions and mechanics of the theory are not. Laying out this process in a clear mathematical framework will make the conditions and conclusions more apparent and can guide our path forward in the empirical strategy.

3.3.2 Formal Model

I will start by laying out the equations needed to describe a single government and its citizens actions, and then describe what happens if that state has a neighboring state governed by the same principles.

3.3.2.1 Budget Constraint

First, the government has to balance its budget between public goods provision G and taxes τ_i. So, if there are n citizens with incomes M_i then

$$G = \sum_{i=1}^{n} \tau_i M_i \tag{3.3.1}$$

3.3.2.2 Citizens

Voters have relativistic utility[4] where their level and marginal utilities are dependent on reference points $\bar{G}_i$ and $\bar{\tau}_i$.

$$U_i(G - \bar{G}_i, \tau_i - \bar{\tau}_i) \tag{3.3.2}$$

The individual utility function can be characterized with the following:

$$\frac{\partial U_i}{\partial (G - \bar{G}_i)} > 0 \qquad \frac{\partial U_i}{\partial (\tau_i - \bar{\tau}_i)} < 0$$

$$\frac{\partial U_i}{\partial (G - \bar{G}_i)^2} < 0 \qquad \frac{\partial U_i}{\partial (\tau_i - \bar{\tau}_i)^2} > 0$$

This implies voters compare their taxes and public goods to some reference point and get utility based on how their own level compares to this reference point. All else equal voters want more public goods and lower taxes, and they care less about marginal changes the farther they are from those reference points. Taxes are a "bad". We could instead characterize this with consumption C rather than taxes τ. There is a one to one mapping between tax rates and consumption since $C_i = (1 - \tau_i)M_i$ so utility can be expressed either way. However, working with the tax rate directly will make more sense when deriving the behavior of the government, which is our primary interest. From this utility function voters derive a probability of voting for a given government.

[4] This is similar to prospect theory but without the loss aversion

3.3.2.3 Voting

The utility that voters receive from a given policy set (G, τ_i) has a one to one correspondence to an expected vote plurality function that indicates the expected plurality for the incumbent party of this citizens probabilistic voting behavior. Citizens can vote for the incumbent, an opposition party, or stay home and not vote[5]. So, a given voter's expected vote plurality function takes values between negative one and one. That is

$$P_i(G - \bar{G}_i, \tau_i - \bar{\tau}_i) \in (-1, 1) \tag{3.3.3}$$

It is reasonable to think that the citizen's vote is impacted not just by the incumbent's policy set, but other parties offers as well. For example, an opposition party's policy set (G^O, τ^O). this would complicate our plurality function as it now depends on both party platforms:

$$P_i(G - \bar{G}_i, \tau_i - \bar{\tau}_i, G^O - \bar{G}_i, \tau_i^O - \bar{\tau}_i) \tag{3.3.4}$$

This idea, however, is implicit in equation (3.3.3) if we think of the opposition as best responding and let equation (3.3.3) be the incumbent plurality given the opposition's best response. We can take the oppositions best response functions $G^{O*}(G, \tau_i)$ and $\tau_i^{O*}(G, \tau_i)$ and plug them in to (3.3.4). This gives us a function independent of opposition policy sets equivalent to equation (3.3.3). We want to abstract away from the inter party conflict and focus on how a given party responds to voters and governments in neighbor states. Assuming we know the opposition's best response allows us to avoid the unhelpful and complicated question of intrastate party competition.

How voters come about this function is not of particular importance to the question at hand and so we will take this plurality function as given. However, we can think of this as the result of an individual optimization problem. Voters can either vote for the incumbent, the opposition party, or stay home. Casting a vote has a low cost which leads some voters to abstain and voting perhaps bequeaths some warm glow on the voter. Solving a problem like this would give us (3.3.4) from (3.3.2). There are of course some concerns about why people would vote under this model since their vote does so little to alter the odds, and certain behaviors may not correspond perfectly to warm glow preferences. The fact is that people do vote, and these concerns are not of primary importance to the conclusion.

While not explicitly solving this problem, some natural assumptions from this line of reasoning regarding the relationship between vote plurality and utility are:

[5] It is the expected plurality because we are thinking of voters as probabilistic voters. That is they have some probability of each action that leads to an expected plurality

$$\frac{\partial P_i}{\partial U} > 0 \qquad \frac{\partial^2 P_i}{\partial U^2} < 0 \tag{3.3.5}$$

P_i may not actually be differentiable, but the general idea of these conditions can still hold (as I will formally show below). They imply that the plurality of voting for a candidate goes up with the utility a voter gets from their proposals. Additionally as utility increases, the gain in plurality is decreasing. This implies that the dynamics from the utility function will translate to the vote plurality function.

3.3.2.4 Government optimization

The government's goal is to maximize their probability of winning. In large elections, maximizing the plurality of votes is roughly equal to maximizing the probability of voting (*Ledyard*, 1984, p. 20-21). Plurality, however, is much easier to work with. This is why we considered plurality and not probability in the previous section. So the government's objective is to

$$\max_{G,\tau} \quad \sum_{i=1}^{n} P_i(G - \bar{G}_i, \tau_i - \bar{\tau}_i) \tag{3.3.6}$$

A government maximizing the probability of winning over static voter preferences will lead to static policy based on voters' preferences. However, if we assume people's preferences receive random shocks or even trend in certain directions, we can get policy and political changing stochastically over time. Either way, the model so far does not address our question. For that we need to introduce a second neighbor government.

3.3.2.5 Two Governments

Now that we have set up how a single government selects policy, we want to consider how neighboring states interact with one another. Consider state A and state B with policy sets (G^A, τ^A) and (G^B, τ^B) respectively. Continuing with the framework we have built and with a few additional assumptions we will show that when G^B increases G^{A*} increases as well. The first assumption dictates the direct impact these states have on one another. The only direct impact we assume is that a change in policy in a neighboring state impacts the reference point for voters. Specifically,

Assumption 5. *Citizens reference points for spending and taxes are increasing in a neighbor states spending and tax levels. That is,*

$$\frac{\partial \bar{G}^A}{\partial G^B} > 0 \qquad \frac{\partial \bar{\tau}_i^A}{\partial \tau_i^B} > 0$$

$$\frac{\partial \bar{G}^B}{\partial G^A} > 0 \qquad \frac{\partial \bar{\tau}_i^B}{\partial \tau_i^A} > 0$$

Assumption 5 fits the story laid out so far. Voters see higher spending in the neighbor state and raise their reference point for what is acceptable to them. It is possible that in some cases this impact is reversed. Perhaps voters see a neighboring state raise spending and the effort ends badly. This may cause voters to actually lower their reference points. The conclusions we reach would simply be reversed in this case.

Next, we need some assumptions about the structure of citizens vote plurality functions.

Assumption 6. *Vote plurality functions are supermodular in G and τ. That is: $\forall\ G, G' \in \mathbb{R}^+$ and $\tau, \tau' \in [0,1]$*

$$P_i(\min\{G, G'\} - \bar{G}_i, \min\{\tau_i, \tau_i'\} - \bar{\tau}_i) + P_i(\max\{G, G'\} - \bar{G}_i, \max\{\tau_i, \tau_i'\} - \bar{\tau}_i) \geq$$

$$P_i(G - \bar{G}_i, \tau_i - \bar{\tau}_i) + P_i(G' - \bar{G}_i, \tau_i' - \bar{\tau}_i)$$

In words what this is doing is comparing the total plurality for two sets of two policy proposals. The first set has one option with low spending and low taxes and another option with high spending and high taxes. Total plurality for this set of proposals is higher than the second set which has one proposal with high taxes and low spending and a second proposal with low taxes and high spending. Another way of putting it is that people are less upset by low government spending if they also have low taxes and are less upset by high taxes if they also have good government services. This is conceptually similar to a convex preferences assumption without requiring differentiability. If P_i is differentiable this is equivalent to

$$\frac{\partial^2 P_i}{\partial G \tau_i} > 0$$

We also need another assumption about vote plurality.

Assumption 7. *P_i has increasing differences in $(G, \bar{G}_i)$. That is:*

$$\begin{aligned}&\forall \quad G' > G \quad \textit{and} \quad \bar{G}_i' > \bar{G}_i \\ &P_i(G' - \bar{G}', \tau_i - \bar{\tau}_i) - P_i(G - \bar{G}', \tau_i - \bar{\tau}_i) \geq P_i(G' - \bar{G}, \tau_i - \bar{\tau}_i) - P_i(G - \bar{G}, \tau_i - \bar{\tau}_i)\end{aligned} \tag{3.3.7}$$

In words this means that the gain in plurality from increasing G is larger if the reference point $\bar{G}_i$ is higher. If P_i is differentiable this is equivalent to

$$\frac{\partial^2 P_i}{\partial G \bar{G}_i} > 0$$

Combining these assumptions, we get the following theorem.

Theorem 6. *If assumptions 1,2, and 3 hold, then when state B receive a preference shock such that G^B increases, G^A will increase as well.*

Proof. The only direct impact of an increase in G^B on state A is an increase in $\bar{G}_i^A$. We can see this directly from assumption 1. Now by assumption 6 and 7 and the Topkis's theorem we can say that G^{A*} is increasing in $\bar{G}_i^A$ (*Topkis*, 1978, p. 317). Putting this together we get that an increase in G^B leads to an increase in G^{A*}. □

3.4. Model Discussion

The core conclusion of my model could be summarized simply as "state fiscal policy will impact the fiscal policy of neighboring states". This is similar to the conclusions in *Baicker* (2005) and *Case et al.* (1993) regarding fiscal policy and *Besley and Case* (1995) Yardstick model regarding tax policy. While the core is the same, my model improves on the underlying theory in a few ways.

First, my model can be useful for informing empirical tests of the conclusion that were not as clear from previous models. The model motivates my data collection processes as it is directly tied to individuals comparing themselves to individuals in other states through the newspaper. While the intuition may have been clear from a simple story, the model formally shows that the assumptions needed to go from individual preference formation to collective state interdependence are plausible and, in fact, in line with the assumptions economists would expect.

The second reason why my model is an improvement is that in order for theory to be convincing we need the story to reflect reality in a believable way and to deviate from reality in ways that are secondary to the conclusion. My model moves the underlying story and justification closer to a plausible story of voter behavior. Consider the idea from the yardstick model of simply good or bad governments. This is a useful dichotomy for the purposes of the model, but it is unrealistic[6]. I know that this type of rent seeking corruption does exist. However, especially at the state level, I do not think pure rent seeking, essentially corruption, is at the top of most peoples' list of concerns. For example, at the national level, voters consider many policy issues to be "Very Important" (pew, 2022). While, corruption investigations are considered important to some, it is primarily the corruption investigations into the opposition party that interests them. Suggesting little scope for a general interest in government corruption that is unrelated to policy considerations. The idea that people are choosing their party based on a set of policy proposals is much closer to describing the common perception of political behavior. While some people of course may make some decisions based on things like competence or corruption, it is reasonable to think policy is a larger part of the decision for most people in state politics.

The set up of the yardstick model also does not directly transfer to the fiscal policy setting in a clear way. I also agree (and it is assumed in my model) that all else equal people want lower taxes, but once we start considering fiscal policy all else is not equal. State governments, especially those with balanced budget requirements, have to weigh the benefit of lower taxes with the cost of fewer services. How would a benevolent or Leviathan government feel about this trade off? In order to extend the yardstick model directly we would need, I believe, normative statements about good and bad policy. My model works both taxes and spending into the government's decision and avoids subjective determinations about what a "good" or "bad" government would do.

While not necessarily an improvement, my model also leads to differences in secondary implications. In the yardstick model voters must be looking to states with similar economic shocks (*Revelli*, 2005). The idea that, for example, California influences a large number of states in the country, does not really make sense. However, in my model this would be perfectly feasible. California may have a large cultural influence and so people update their reference point for public goods quality in order to keep up with California. Whether or not this is an improvement is an empirical question.

Both my model and the yardstick model differ from the Tiebout model where government are trying to optimize community size and voters move to areas with their preferred policies (*Tiebout*, 1956). The biggest difference I see is that in a pure Tiebout model leads to the conclusion that intra-state political competition is not necessary (*Besley and Case*, 1995) . While the other two say the exact opposite. While intra state competition is not explicit in my model, it is the underlying mechanism that justifies the assumption that government tries to maximize the probability of election. "Government" acts as if it is maximizing the probability of winning

[6] I think it is perhaps more realistic in the context of tax policy alone. Which is how the model is formulated

the election because if an individual party stops doing this, it loses to the party that does.

As mentioned above, the spillover model pioneered in *Case et al.* (1993) is sufficiently broad that my model actually within the general framework. Their model characterizes the relationship between states in terms of the complementarity between neighbor state spending and home state spending or consumption. My model describes the necessary assumptions for this complementarity to hold in the context of individual vote preferences. It also gives a more specific theoretical explanation for why those complementarities exist.

The implications of the different models are not different enough to provide a conclusive empirical test. All models would lead to state interdependence. However, there are a few things that may be suggestive. If the weighting matrix I derive has states that are not economically linked, but who's spending does seem to be linked, this would suggest the yardstick model is not sufficient. I also think in the absence of a clear empirical test, we should prefer the model that has more plausible assumptions about how people behave. People voting based on government's policy decisions seems, on its face, more plausible than people selecting parties solely on the level of perceived corruption.

3.5. Data

3.5.1 Data Sources

Before discussing the data collection process, I will give a brief outline of the empirical methods. This is only meant to give context to the data section. A more thorough discussion of the empirical methods can be found in the following section. In order to test how much the spending in one state influences its neighbors, we first need to decide which states are neighbors with which. This decision leads to a matrix with a weight for each state's degree of "neighborliness" with one another. Determining the degree of "neighborliness" could be done in many ways, and testing how the result varies across different networks will allow some insight into the primary mechanism behind any relationship.

To facilitate a clean comparison to existing work, I replicate the methods in Case, Hines, and Rosen by using geographic neighbors and the percentage of the population that is Black (*Case et al.*, 1993). To test the individual preference formation theory, I collect new data on newspaper mentions of other state to create a weight matrix that is meant to capture the level of attention one state pays to another. Once I have this matrix, I regress the weighted mean of a states neighbor's spending on its own. I include the following covariates: the fraction of the population over 65 and under 18, per capita income, and the fraction of the population that is Black.

While a basic OLS regression will give us a sense of co-movement, an instrumental variables approach will get us closer to a causal assertion. I follow the methods laid out in Baicker 2005 to derive my IV specifications (*Baicker*, 2005). That is, I use medicaid spending in the year prior to my spending data and then inflate that by total healthcare spending in the following years. I also run a separate IV regression using the weighted neighbor's covariates as the instrument.

The state expenditures data comes from the Unites States Census Bureau Annual Survey of State and Local Government Finances. I am using the summary tables which include state and local spending for various categories including total expenditure, Education, public welfare, hospitals, health, natural resources and more (b), (2004-2017). State Medicaid spending data, used in creating the instrument, comes from the National Association of State Budget Officers (), (2004-2017), and national spending data, also used for an instrument, is from the U.S. Centers for Medicare and Medicaid (), (1960-2018)

Demographic data comes from the America Community survey one-year estimates. This includes things like the fraction of the population over 65 and under 18, the fraction of the population that identifies as Black, and per capita income (a), (2004-2017).

3.5.2 New Data Collection

In addition to these data sources, I collect new data from newspapers to construct a state neighbor weight matrix. While this process involves many steps, the core concept is straightforward. For multiple newspapers in each state I search Bing.com for the number of policy relevant stories on their site. I then perform an additional search for each potential neighbor state that searches that same site for policy relevant news that also mentions the potential neighbor state. The results of these searches are used to create a weight matrix that is driven from the theoretical motivation laid out in the previous section. If voters are updating their beliefs based on neighboring state policies, they need to be informed about what those policies are. So, we should see those states mentioned in local newspapers.

The first step in this process is to scrape a list of newspapers and their website and twitter accounts [7] for each state (), (2017). This gave me 5428 newspapers. This is too many to search even for my computer. In order to narrow down my list, I found the newspapers in each state with the top ten most twitter followers (which requires having a twitter page). This left me with 482 papers since not every state had ten.

The next step is to determine which of these papers are local and which are national. If a paper's readership is dispersed across multiple states or the entire country it will not make sense to make inferences about the population of that state based on the contents of the paper. For example, what the New York Times publishes is influenced by a national audience and not just residents of New York. In order to do this systematically I take the top ten most popular papers

[7] I also have their Facebook, and YouTube links but have not used them in any way but could be used in future work

from each state[8] and put them into google trends. This tells me how much each newspaper name is searched in each state relative to other states with the maximum state being 100. For example, a paper only searched for in Wisconsin would have a 100 for Wisconsin and a zero for everywhere else. A paper searched for in every state equally except twice as much in New York would have a 100 in New York and a 50 everywhere else.

Based on these results, I create an HHI type index of state searches to determine the level of national readership where

$$CI = \sum_{I=1}^{50} Trend\left(S_i\right)^2$$

There are some issues to consider with this approach. First, google trends does not use the URL, but rather the name of the newspaper. For certain papers like "Mercury", the majority of searches are probably not looking for the San Jose newspaper. This is something that could certainly be corrected with more time by searching on a case by case bases and adding terms like "newspaper" until the paper's website is the first result. While some paper names may simply be too generic, other papers may not be true news. I noticed that the popular Madison based satirical newspaper "The Onion" was included in the list. I may be able to resolve this by finding a better primary source for newspaper websites or again with a manual case by case approach[9]. In the specific case of "The Onion" it was removed for being a national, rather than local, paper.

This brings me to the third issue with the google trends locations: determining where exactly to draw the line of a local vs national paper. The cutoff is inherently arbitrary, and I do not have a systematic way to make this determination; so I instead rely on the frequently used "eyeball" method. That is, I looked at examples and decided if I thought it looked regional or national. With my eyeball method I decided on CI < 11649 as a "state" newspaper. Appendix C.2 Table C.1 shows a paper, "The Times-Union" that is just below that cutoff next to "The New York Times", which is far above the cutoff. The full tables for all newspapers can be explored in the interactive data appendix [10]. Enforcing this cutoff left me with 390 state newspapers.

Now that we have a list of truly local newspapers, we can proceed with the Bing searches and counts that will lead to our weight matrix. Recall that the purpose of the newspaper list is so we can determine how often the news in Alabama, for example, mentions California, for example.

[8] Based on twitter as I describe above

[9] For example I considered searching specifically for state news in every state and recording the top google answers. "Wisconsin state news" for example.

[10] Click on the words "interactive data appendix" to go to the link. You will need to download the file and open it in a web browser

For that, I use Bing searches [11]. I start with a search of key terms meant to narrow the results to policy relevant stories from a given paper. Below is an example:

Search 1. *site:www.montgomeryadvertiser.com (government|"government spending"|"public policy"|*
legislature|"fiscal policy"|spending)

After each search, the number of results is recorded for each paper. This gives us a baseline for the number of policy relevant stories that can be found in that newspaper. This search was done on the ten most popular papers from each state. The three most popular papers that also returned more than 100 results are included in the analysis. Searches were performed for the year 2019. This limits each paper to a fixed timeline despite differing histories of online publishing. It is worth noting that since this is after the spending data has been collected, it is possible some stories could be referring to events that happened during the time period we are analyzing. That being said, the most recent search results give us the largest number of results. The full list of included papers and counts are shown in the interactive data appendix .

Now for each paper, an additional search is made for each possible neighbor. It contains all the previous terms plus the potential neighbor state's name. In the case of Washington State, it included some additional terms in an attempt to distinguish it from Washington D.C. [12] An example search for the Montgomery Advertiser in Alabama and California is shown below.

Search 2. *site:www.montgomeryadvertiser.com (government |"government spending" |public policy" |*
legislature | "fiscal policy" | spending) & "California"

The end result is that for each newspaper in each state I have the number of results for my policy relevant search as well as 47 additional searches including each of the contiguous united states. For now, these terms were chosen by me to be fairly general and include a large number of stories. There are some conceivable data driven strategies for selecting these terms, but it is not totally obvious they would be an improvement. One way would be to get a training data set of policy relevant and irrelevant stories. This could be done by skimming a large number of stories or by taking advantage of existing categorical classifications in papers, like a "Politics" vs "sports" section. While this would be an ideal approach to find words distinguishing the articles, it is not necessarily true that those same words will translate to the best search terms in the black box of Bing.

[11] Google is much quicker to restrict automated searches

[12] a full search for Washington state would look something like this: site:www.montgomeryadvertiser.com & (government | "government spending" | "public policy" | legislature | "fiscal policy" | spending) & Washington -DC -D.C. -"district -of -columbia

The black box of Bing searches can cause additional problems as well. For example, some searches that add terms with an and condition, in theory making the search more restrictive, actually lead to more results, not fewer. In the above example, the second search has 25,500 results while the first has only 14,300[13]. I attempt to create weights that recognize the irregularity of theses search counts.

To create a weight of one state on another, for example California on Alabama, I start by dividing the number of results from Search 2 by the number of results in Search 1. The hope is that the number of results in Search 1 will normalize differences in online publishing and the number of political articles across papers. I then weight this newspaper level statistic by the number of twitter followers to get a state level statistic. Finally, I normalize the total weight for each state to one.

In mathematical terms: let S_{1ip} indicate the number of results for search 1 in state i and paper p. let S_{2ipj} indicate the number of results for search 2 in state i paper p and the additional term added for state j. Let T_{pi} be the number of twitter followers for paper p in state i. Then the weight is

$$w_{ij}^{(1)} = \frac{\sum_{p=1}^{3} \frac{S_{1ip}}{S_{2ipj}} * T_{pi}}{\sum_{j=1}^{47} \sum_{p=1}^{3} \frac{S_{1ip}}{S_{2ipj}} * T_{pi}}$$

Normalizing to one does come at a cost. The collective influence of every state's set of neighbors is now equalized. That means California, for example, is just as influenced by other states as Wisconsin. While I expect that is not exactly the case, the total weights (before scaling to 1) are highly variable and do not match my priors in any way. They are shown in Appendix C.2 Table C.2. While it may have been ideal to allow for this additional flexibility, this is no more restrictive than any previously used weighting scheme. In fact, it is still more flexible as the matrices are still asymmetric. That is California may impact Oregon without Oregon impacting California.

In addition to this weight, I make more transformations to get a second set of weights. It is not obvious that twice as many stories should lead to twice the attention from readers. This second set of weights seeks to lessen the difference in weights between a given states neighbors and to lessen the number of states with non-zero weights. In order to do this, I start by finding the median weight for a given state. I take any weights below that and set them equal to zero and subtract off the median for the rest. This is meant to account for some states simply showing up more in Bing searches because they are more popular states. Next, I take the cubed root of all the weights in order to lessen the difference between the remaining neighboring states. Finally, I

[13] Something similar to this could throw off the machine learning approach to search term selection

take this new set of weights and re scale them to sum to one. The end result is a set of weights that are less extreme across states and less variable within a selected group of "neighbors".

In mathematical terms: in addition to the variables above, let M_j be the median of $w_{ij}^{(1)}$ $\forall$ i

$$w_{ij}^{(2)} = \max\left(w_{ij}^{(1)} - M_j, 0\right)^{\frac{1}{3}}$$

In addition to the weights using the internet searches I replicate two weights from previous literature. The first is geographic neighbor. For each state i, every state that is geographically connected to state i[14] receives a weight of $\frac{1}{n_i}$ where n_i is the total number of states touching state i.

The second replicated weight uses the fraction of the state population that is Black. The weight for state i is

$$w_{ij}^{(B)} = \frac{1}{|\%Black_i - \%Black_j|S_i} \quad \textbf{Where} \quad S_i = \sum_j \frac{1}{|\%Black_i - \%Black_j|} \tag{3.5.1}$$

(*Baicker*, 2005). Since the new weight measures are not symmetric, they cannot be easily represented in a single graph as in Conley (*Conley*, 1999). I find these graphs a bit hard to digest. I instead include maps of the weights for every state in the interactive data appendix . A hand full of maps are also included in figures C.1 to C.7. As these figures show, there is significant variation between the three methods[15]. In general, I would say the Bing weights tend to be more diffuse. This is especially true for California. California and Kansas have a shockingly close percentage of their population that is Black, and these leads the above metric to weight them almost exclusively with each other. This is not true in every case, for example New York.

Overall, I think the Bing measures weight specific states more heavily. Washington and California for example. In the case of Washington, I expect much of this is due to the general use of the word Washington not referring to the state. In the case of California, I am less sure. It seems plausible that people pay particular attention to California. I tried a number of variations on the searches above. For example, including governor's names or state capital names instead or in addition to state names. None of these lead to results that were viable or seemed like a significant improvement. While there are other conceivable ways to create this matrix, they would indubitably have problems of their own.[16]. Given that I find clear null results, I do not expect

[14] including water boarders

[15] The two Bing search methods are clearly closer

[16] In the future work section I talk about the possibility of using google trends only rather than explicit search numbers

minor changes in this approach to lead to significant differences, but it is a potential area for future exploration.

3.6. Empirical Methods

To estimate state interdependence with the above data I start by following the strategy of *Baicker* (2005). First I run a baseline OLS model of state expenditures on weighted neighbor expenditures and a series of controls.

$$E_{it} = \phi W_i E\bar{E}_t + \beta X_{it} + \phi_i + \delta_t + \epsilon_{it} \tag{3.6.1}$$

Where E_{it} is real state per capita expenditures [17], W_i is a vector of "neighborliness" weights, $\bar{E}_t$ is a vector of neighbor state spending, X_{it} is a vector of state time controls, ϕ_i are state fixed effects, δ_t are time fixed effects, and ϵ_{it} is random noise. The controls included in X_{it} are the fraction of the population over 65 and under 18, per capita income, and the fraction of the population that is Black. This is fewer controls than in previous studies[18], but with additional controls I was having issues with model over specification. I also cluster all standard errors at the state level.

As pointed out in Baicker 2005, the OLS specification is suspect. Positively or negatively correlated standard errors are a reasonable concern that would bias these estimates. If the states that we identify as "neighbors" receive correlated economic shocks, this would bias the estimates up. If, on the other hand, neighbor states are competing for businesses and businesses move between states with shifts in political climate, we may bias our estimates downward (*Baicker*, 2005). While these regressions may be suspect, I include them to see a baseline measure of interstate correlations and to facilitate comparisons to existing work.

To address the issues biasing OLS I run two different IV specifications in line with Baicker 2005. The first relies on unanticipated spending on Medicaid. The idea behind this is that states make prior commitments to certain levels of care when it comes to Medicaid rather than a specific dollar amount. If healthcare costs rise more than expected, states find they have less money than anticipated. This is, ideally, like a negative exogenous shock to the state budget. While Baicker supplements this concept with some law changes that take place in the 1980s, those changes are clearly not available to me in 2005-2016 and I have not found a suitable substitute (*Baicker*, 2005). The second instrument is to use neighbors' covariates as an exogenous indicator of their budget changes. I find the exclusion restriction less convincing here, but I include it for the sake

[17] In 2013 dollars

[18] income squared and population density were excluded and have not yet found a comparable measure for federal grants. I need to dig into the state and local government finance data sets a little deeper

of comparison. These IV regressions include the same controls as the OLS regressions.

For an easy comparison to existing literature, I consider what I outlined above my baseline specification. However, I run a number of variations on these models. While states may react rather quickly to one another, these interactions may also take time. To test this, I also run all the regressions with one-, two-, and three-year lagslags. Finally, in addition to overall state spending I also run regressions on education spending and health and human services[19]. In total that is an OLS and two IV regressions for each of the 4 weights on spending levels, each done on the three types of spending and with 0-3 years of lag. In total that makes $3 * 4 * 3 * 4 = 144$ regressions. I include some of these in tables 3-5, and the full table of regressions is including in the interactive data appendix[20] .

The models outlined above are meant to test the relationship between states' spending, but they do not provide a clear test of which of the weighting schemes are correct. To get a better understanding of this question I show for each model the fraction of randomly generated weight matrices that are significant. This shows how sensitive the results are to the weight matrix. If they are not very sensitive, then we will have less confidence that the weight matrix used is a crucial part of the model results. If the results are very sensitive to the weight used, we can be more confident that the weight matrix used is playing a key role.

I generate random matrices in two ways. For the first method I simply select a random number from a uniform distribution between zero and one for each connection and then scale them so the total weight per state sums to one. For the second method I take the weights used in a given model and permute those numbers within the matrix. I start by shuffling the entire set of state weights. For example, the entire set of weights from Wisconsin could be assigned to any state. Then, I shuffle the weights within each state. This ensures the distribution is similar throughout the new randomized matrix. For example, all state weights will still sum to 1 without any rescaling, but the relationships will be random. The fraction of these random weights that are significant at a $.05$ level are included in the regression tables. These placebo tests demonstrates the probability that I would see a significant result had I drawn the weighting matrix at random. If they are small, it means it would be unlikely for the weighting matrix I generated from newspapers to have a significant result by chance rather than because it is related to the mechanism driving the relationship.

[19] This is defined as the sum of spending on Health, Hospitals, and Public welfare in the state and local finance aggregate tables

[20] Simply download the file and open it in a web browser

3.7. Results

The results show no significant relationship between states using my new metric for state inter-connectivity. More surprisingly, the networks used in older research also find no relationship in the more recent 2005-2015 data. Very few of the regressions ran are statistically significant and there does not seem to be a clear pattern among the significant results or even the point estimates in general. Let's start by examining the results that mirror the positive results in *Baicker* (2005) and *Case et al.* (1993). Table 3 Shows OlS and the two IV methods on total spending levels. The OLS estimates are expected to be biased but provide a first look at spending correlations. The estimates in *Baicker* (2005) were positive or close to zero while mine are all negative. While the standard errors on my estimates are for the most part large, 24% of randomly generated matrices were significant at a 5% level and around 10% of randomly permuted matrices were significant as well.

Table 3.1: Regression Results

OLS On Total Spending

Weight	Lag	Estimate	SE	P Value	Percent Randomly Significant	Percent Permuted Significant
Percent Black	0	-0.007	0.03	0.82	0.24	0.09
Geographic	0	-0.115	0.19	0.54	0.24	0.09
Scaled bing	0	-0.607*	0.24	0.02	0.24	0.12
Scaled Bing 2	0	-0.498	0.34	0.14	0.24	0.14

Medicaid IV On Total Spending

Weight	Lag	Estimate	SE	P Value	Percent Randomly Significant	Percent Permuted Significant
Percent Black	0	-0.029	0.11	0.79	0.06	0.04
Geographic	0	1.335	0.77	0.09	0.06	0.00
Scaled bing	0	-0.987	0.79	0.22	0.06	0.00
Scaled Bing 2	0	1.378	3.09	0.66	0.06	0.00

Neighbor's covariates IV On Total Spending

Weight	Lag	Estimate	SE	P Value	Percent Randomly Significant	Percent Permuted Significant
Percent Black	0	0.072	0.06	0.25	0.04	0.06
Geographic	0	0.086	0.34	0.80	0.04	0.07
Scaled bing	0	-0.571	0.34	0.10	0.04	0.09
Scaled Bing 2	0	-0.535	0.44	0.23	0.04	0.09

We should expect a less biased results from the IV specification. Unlike in Baicker, however, these are largely inconsistent (*Baicker*, 2005). None are significant at a 5% level. Additionally, very few of the 1000 permutations of the weight matrix were significant. This suggests first that state's may not be as clearly linked as we thought, and that most weight matrices with similar distributions we could generate would find similar results. The lack of positive relationships in the randomly permuted matrices suggest that some minor data hiccup in the exact form of the Bing weights is not hiding a significant result[21]. The direction of the effects does not seem to tell a consistent story either. The geographic weights produce a large positive result, consistent with previous work, but the percent Black weights are actually slightly negative. The Bing weights I constructed are not even consistent with one another. The first is large and positive, the second large and negative. Of course, with large standard errors these results should not be interpreted too closely, but it does suggest a that changes in the weight matrix can lead to large swings in the estimates.

What if the effects are delayed? It may take a while for information and preferences to transmit between states, in which case the effect would be slow to manifest. To check this, I also run every model with one, two, and three years of lags. This is shown in Table 5. I do not see a clear picture emerging from these results either. Results on other spending measures, education and health and human services are comparable and are all shown in the interactive data appendix[22].

3.8. Discussion

While these are not the results I was looking for, these null findings are interesting given the extensive literature referencing the positive results in previous work *Baicker* (2005); *Case et al.* (1993). There are a number of possible reasons I am getting null results. One potentially interesting interpretation is that since the 1980's the relationship between states has not disappeared but has become more complicated. Specifically, negative relationships between states could be becoming more common in a more polarized political environment. This would be consistent with the model if, as I noted earlier, people actually lowered their reference points in response to neighbors increasing spending. If the impact of neighboring states were all positive or zero in the past it would have been easier to pick up the relationship even with an imperfect weighting matrix. With negative relationships the potential of misallocation a negative neighbor as a positive one could obscure the connection. In this case, a weighting matrix based on political affiliation or control of state government may lead to more consistent results (*DellaVigna and Kim*, 2022).

It's also possible that there has been some significant change since the data used in *Baicker* (2005) and *Case et al.* (1993). It could be that states are less connected. Or, it could be that the network of states has changed in a way that is not captured by any of the four weights I used. Perhaps the increased connectivity through the internet means that traditional fixed networks are less the case today. It may be that people in a given state look to wherever is in the news that day regardless of where it is. This would mean the networks may change frequently as different stories

[21] Although the true set of possible matrices are massive

[22] Download and open in browser to view

enter and exit the news cycle. In this case the entire research design I laid out would struggle to capture this more fluid network. Rather than thinking of fixed networks between states, we would need to consider the diffusion of information more generally and how that pattern may change from policy to policy.

It is also possible that slight differences in methodology are accounting for the differences. I use fewer controls as my models were becoming over-fitted[23]. I excluded income squared and population density and have not yet found a comparable measure for federal grants. I am also not totally sold on the idea of controlling for federal grants as I imagine states have some control over that through lobbying efforts. As I explained above, I also had to use fewer Medicaid instruments since the law changes in Baicker happening in the 1980s (*Baicker*, 2005). While changing these could theoretically bring about a significant result, it would indicate that the findings are not particularly robust as I believe the choices I made are justifiable.

3.9. Future Work

This surprising result opens up a number of avenues for future work. The first is to test the sentiment of the relationship between states. That is, are states likely to see "connected" states actions and adjust in the same, or the opposite direction. This could be done by adding a sentiment index to the weight matrices. It would not be possible to scrape every Bing result, but it might be reasonable to scrape a handful from each search and run a sentiment index on that random sample. This could be used to scale the weight between negative one and one and incorporate the idea that some states may move in opposite directions rather than together.

While I currently use google trends only to find local newspapers, it might also make sense to try using google trends directly to get the linkages. The google trends actually show me that people in, for example Minnesota, search for the Milwaukee Journal Sentinel. This is a strong indication that Minnesotans are paying attention to Wisconsin policy. If I were to do this, I would need some other way to verify papers are local. One way would be to manually search "Wisconsin News" and find some of the first results rather than using a cite like USNLP to get an entire list. This would also simulate more directly how any person would go about getting state news without a specific paper in mind. It would also include popular results like Mlive.com which is a site that aggregates different newspapers and people are likely to use for online news.

I also think that my instruments could be improved. While future work may find many interesting new instruments, one possibility is to use the paper by Moretti and Wilson on taxing billionaires *Moretti and Wilson* (2019). While not the primary purpose of the paper, they identify that the death of a billionaire can be a somewhat significant windfall for state budgets in states

[23] I had perfect collinearity in some cases

with estate taxes. These deaths are also, conceivably, exogenous to state policies. While a shock to revenues is not the same as a shock to spending, my bet is that because of the flypaper effect a significant amount of this money will indeed be spent (*Courant, Gramlich, and Rubinfeld*, 1978). The data was not accessible at the time of writing, but it may be available for future work.

Since writing this paper, more work has been done to look at policy spillovers on a broad range of binary adoption policies. Specifically, (*DellaVigna and Kim*, 2022) shows how, since 2000, political alignment has outperformed other interstate networks in predicting the adoption of the set of policies they consider.

3.10. Conclusion

The surprising null result does not leave this work with a clear and well supported explanation, however, that does not make the results any less impactful. State interdependence was once a robust observation on state behavior. My work is an important reminder that treatment effects are not static objects to be discovered and held on to indefinitely but are instead ever changing and evolving. The political landscape has evolved over the last 15 or 20 years, and our observations on political behavior must evolve as well.

APPENDICES

APPENDIX A

Appendix to Chapter 1

A.1. Full Survey

Below is the full survey. The light grey text indicates the start and end of question "blocks". A block is a page and navigating from block to block required clicking a “Next" button. The red text was not shown in the survey, but explains some of the mechanics of the survey. The grey text box after Q9 explains the logic for the attention check question. If respndents did not select 9 as requested, they were sent to the end of the survey (and removed from the sample).

Centiment Survey

The instructions where first written alone on an initial page and required respondents to wait five seconds before continuing.

Instructions copy ***Instructions Repeated (if needed for reference):***

Imagine yourself in each of the following situations.

Consider if each situation would be painful for you and if yes, how painful it would be.

Then, enter the MOST you would pay in U.S. dollars to completely and immediately eliminate any pain caused by the situation, as if the event described never happened.

Q1 Imagine you have a minor cut on your finger and you accidentally get lemon juice in the wound.

○ $ ____

Q2 Imagine you pick up a hot pot by accidentally grabbing its equally hot handles.

○ $ ____

Q3 Imagine you shake hands with someone who has a normal grip.

○ $ ____

Q4 Imagine you burn your tongue on a very hot drink.

○ $ ____

Q5 Imagine you bump your shin badly on a hard edge, for example, on the edge of a glass coffee table.

○ $ ____

End of Block: WTP Pain

Start of Block: bump your elbow

Q6 Imagine you bump your elbow on the edge of a table ("funny bone"). Would you pay $X to completely and immediately eliminate any pain caused by this situation, as if it never happened?

○ Yes, I would pay $X

○ No, I would not pay $X

NOTE: X is chosen Randomly from (0.05 0.10 0.25 0.50 1 1.25 1.50 1.75 2 3 4 5 7 10 15 20 25 35 50 100) and a single value is displayed

End of Block: bump your elbow

Start of Block: Bite cheek

Q7 Imagine you accidentally bite your tongue or cheek badly while eating. Would you pay $X to completely and immediately eliminate any pain caused by this situation, as if it never happened?

○ Yes, I would pay $X

○ No, I would not pay $X

NOTE: X is chosen Randomly from (0.10 0.25 0.50 0.75 1 2 3 4 5 6 8 10 15 20 25 35 50 75 100 200) and a single value is displayed

End of Block: Bite cheek

Start of Block: Trap finger drawer

Q8 Imagine you slam your finger in a drawer. Would you pay $X to completely and immediately eliminate any pain caused by this situation, as if it never happened?

○ Yes, I would pay $X

○ No, I would not pay $X

NOTE: X is chosen Randomly from 0.5 1 2 3 4 5 6 8 10 15 20 25 30 40 50 75 100 150 200 400) and a single value is displayed

End of Block: Trap finger drawer

Start of Block: PSQ Match

Q9 Imagine yourself in each of the following situations. Decide if each situation would be painful for you and if yes, how painful it would be. Let 0 stand for no pain; 1 is just noticeable pain and 10 is the most severe pain that you can imagine.

	0 (no pain) (1)	1 (2)	2 (3)	3 (4)	4 (5)	5 (6)	6 (7)	7 (8)	8 (9)	9 (10)	10 (most severe pain imaginable) (11)
Imagine you slam your finger in a drawer. (1)	○	○	○	○	○	○	○	○	○	○	○
Imagine you accidentally bite your tongue or cheek badly while eating. (2)	○	○	○	○	○	○	○	○	○	○	○
Imagine you shake hands with someone who has a normal grip. (4)	○	○	○	○	○	○	○	○	○	○	○
Imagine you pick up a hot pot by accidentally grabbing its equally hot handles. (3)	○	○	○	○	○	○	○	○	○	○	○
Imagine you burn your tongue on a very hot drink. (5)	○	○	○	○	○	○	○	○	○	○	○
To ensure your full attention, please select 9 (8)	○	○	○	○	○	○	○	○	○	○	○

Imagine you bump your elbow on the edge of a table (“funny bone”). (6)	○	○	○	○	○	○	○	○	○	○	○
Imagine you bump your shin badly on a hard edge, for example, on the edge of a glass coffee table. (7)	○	○	○	○	○	○	○	○	○	○	○

Skip To: End of Block If Imagine yourself in each of the following situations. Decide if each situation would be painful f... != To ensure your full attention, please select 9 [9]

End of Block: PSQ Match

Start of Block: Demographics

Q10 Which of the following best describes your status?

○ Married

○ Divorced

○ Widowed

○ Separated

○ Single, never married

○ Living with a partner in a long term relationship

*

Q11 How many children do you have?

____

Q13 What is the highest level of school you have completed or the highest degree you have received?

○ Less than high school

○ High School Graduate- high school DIPLOMA or the equivalent (For example: GED)

○ Some college but no degree

○ Occupational/vocational program

○ Associate degree in college

○ Bachelor's degree (For example: BA, AB, BS)

○ Master's degree (For example: MA, MS, MEng, MEd, MSW, MBA)

○ Professional School Degree (For example: MD,DDS,DVM,LLB,JD)

○ Doctorate degree (For example: PhD, EdD)

Q14 What is your current employment status? Mark all that apply.

- □ Employed for wages full time
- □ Employed for wages part time
- □ Self-employed
- □ Out of work and looking for work
- □ Out of work but not currently looking for work
- □ A home-maker
- □ A student
- □ Military
- □ Retired
- □ Unable to work

End of Block: Demographics

Start of Block: Financial Health FHN

Title For the following questions, think of a HOUSEHOLD as all people related by birth, marriage, or adoption and residing together. If you live alone, or do not consider anyone else to be a member of your HOUSEHOLD, please answer these questions as an individual.

Q15 Which category represents the total combined income of all members of your HOUSEHOLD during the past 12 months? This includes money from jobs, net income from business, farm or rent, pensions, dividends, interest, social security payments and any other money income received by members of your HOUSEHOLD who are 15 years of age or older.

○ Less than $5,000

○ $5,000 to $9,999

○ $10,000 to $14,999

○ $15,000 to $19,999

○ $20,000 to $24,999

○ $25,000 to $29,999

○ $30,000 to $39,999

○ $40,000 to $49,999

○ $50,000 to $59,999

○ $60,000 to $74,999

○ $75,000 to $99,999

○ $100,000 to $124,999

○ $125,000- $149,999

○ $150,000 - $199,999

○ $200,000 to $249,999

○ $250,000 and over

Q16 Which of the following statements best describes how your household’s total spending compared to total income over the last 12 months?

○ Spending was much less than income

○ Spending was a little less than income

○ Spending was about equal to income

○ Spending was a little more than income

○ Spending was much more than income

Q17 Which of the following statements best describes how your household has paid its bills over the last 12 months? My household has been financially able to:

○ Pay all our bills on time

○ Pay nearly all our bills on time

○ Pay most of our bills on time

○ Pay some of our bills on time

○ Pay very few of our bills on time

Q18 At your current level of spending, how long could you and your household afford to cover expenses, if you had to live on only the money you have readily available, without withdrawing money from retirement accounts or borrowing?

○ 6 months or more

○ 3-5 months

○ 1-2 months

○ 1-3 weeks

○ Less than 1 week

Q19 Thinking about all of your household's current debts, including mortgages, bank loans, student loans, money owed to people, medical debt, past-due bills, and credit card balances that are carried over from prior months...

As of today, which of the following statements describes how manageable your household debt is?

○ Do not have any debt

○ Have a manageable amount of debt

○ Have a bit more debt than is manageable

○ Have far more debt than is manageable

Q20 Thinking about all of the types of insurance you and others in your household currently might have, including health insurance, vehicle insurance, home or rental insurance, life insurance, and disability insurance...

How confident are you that those insurance policies will provide enough support in case of an emergency?

○ Very confident

○ Moderately confident

○ Somewhat confident

○ Slightly confident

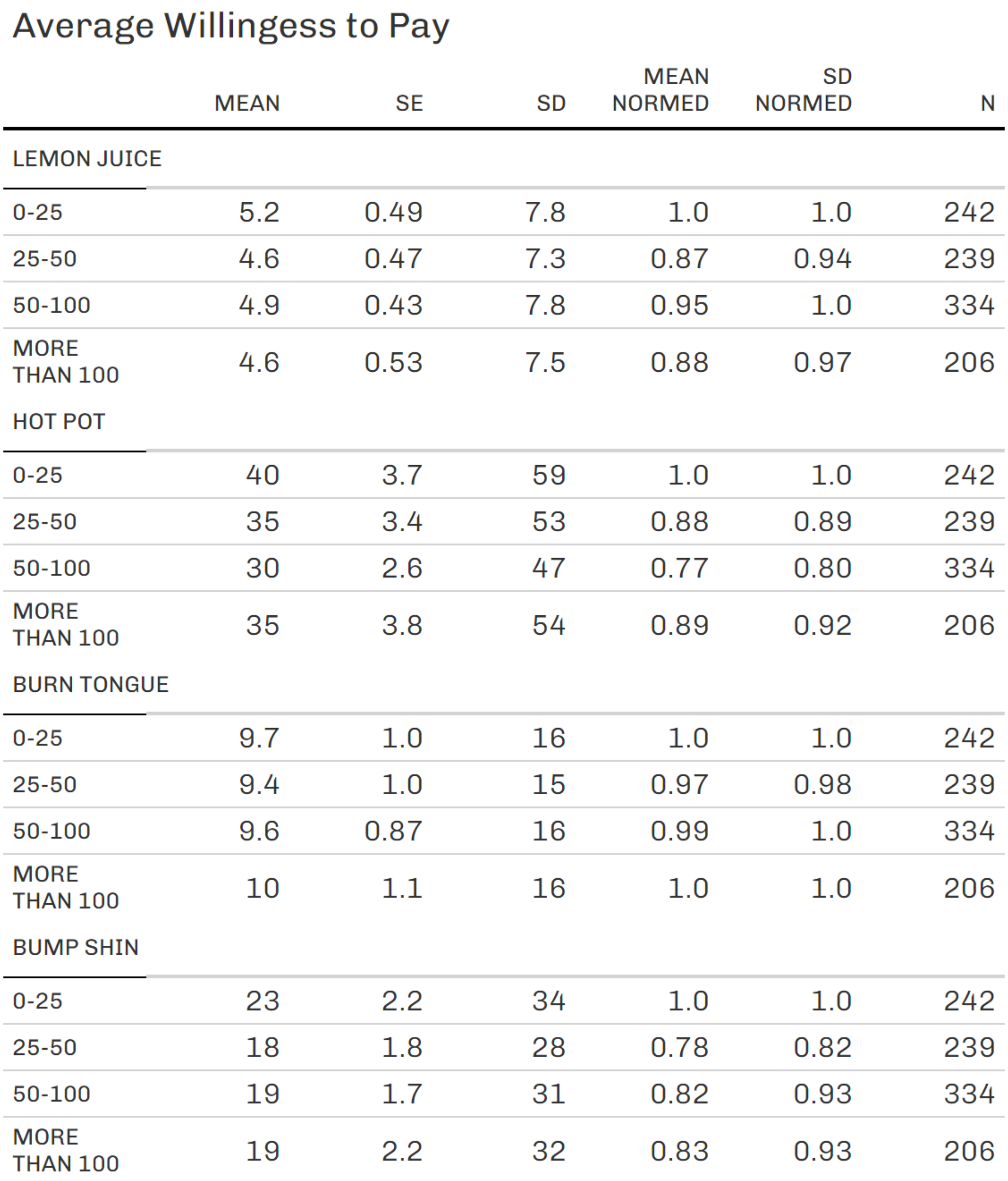
○ Not at all confident

○ No one in my household has any insurance

End of Block: Financial Health FHN

A.2. Auxiliary Tables and Figures

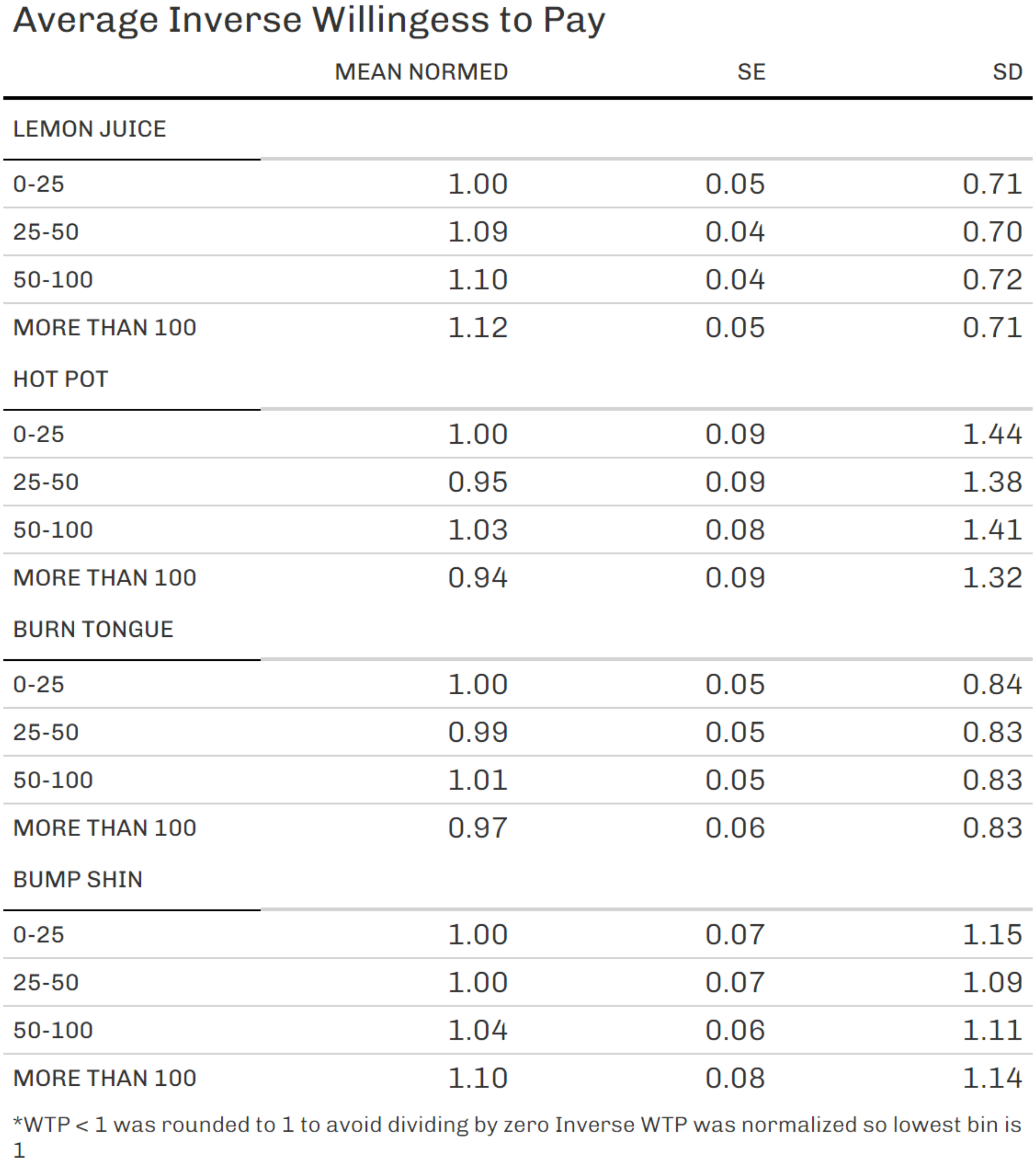
Table A.1: Mean Open Response Table



Average Willingess to Pay

	MEAN	SE	SD	MEAN NORMED	SD NORMED	N
LEMON JUICE						
0-25	5.2	0.49	7.8	1.0	1.0	242
25-50	4.6	0.47	7.3	0.87	0.94	239
50-100	4.9	0.43	7.8	0.95	1.0	334
MORE THAN 100	4.6	0.53	7.5	0.88	0.97	206
HOT POT						
0-25	40	3.7	59	1.0	1.0	242
25-50	35	3.4	53	0.88	0.89	239
50-100	30	2.6	47	0.77	0.80	334
MORE THAN 100	35	3.8	54	0.89	0.92	206
BURN TONGUE						
0-25	9.7	1.0	16	1.0	1.0	242
25-50	9.4	1.0	15	0.97	0.98	239
50-100	9.6	0.87	16	0.99	1.0	334
MORE THAN 100	10	1.1	16	1.0	1.0	206
BUMP SHIN						
0-25	23	2.2	34	1.0	1.0	242
25-50	18	1.8	28	0.78	0.82	239
50-100	19	1.7	31	0.82	0.93	334
MORE THAN 100	19	2.2	32	0.83	0.93	206

Table A.2: Mean Open Response Table

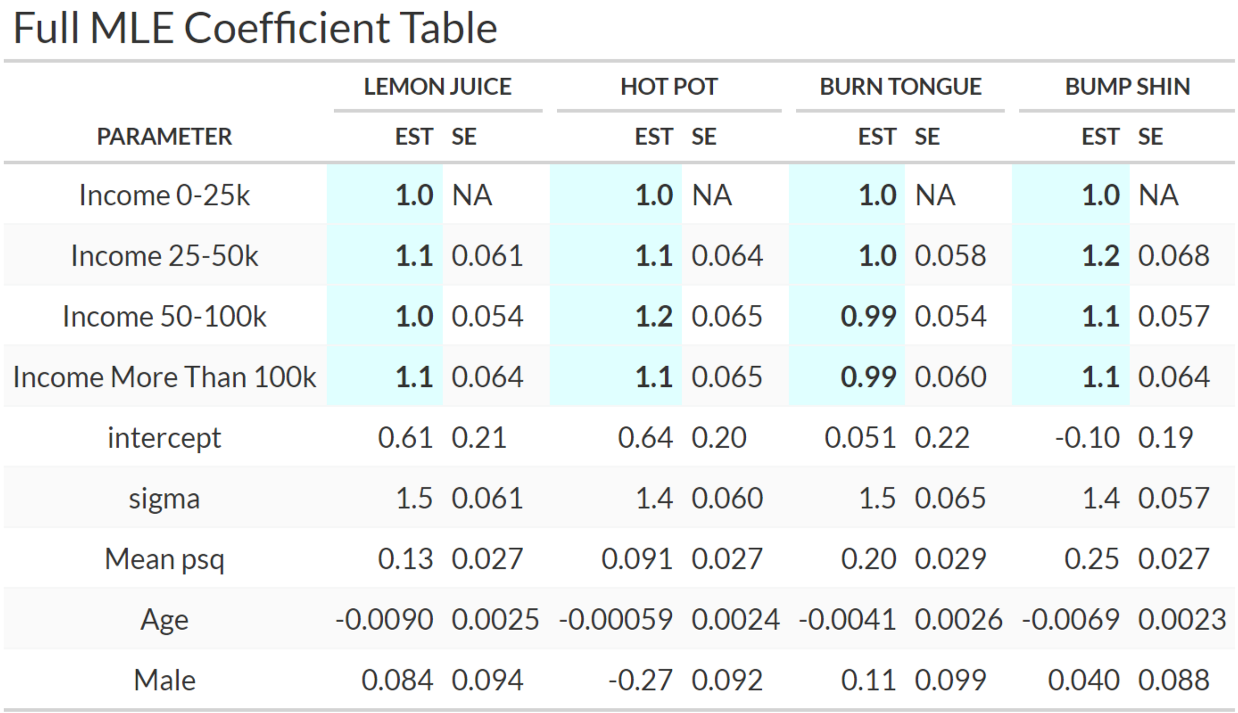


Average Inverse Willingess to Pay

	MEAN NORMED	SE	SD
LEMON JUICE			
0-25	1.00	0.05	0.71
25-50	1.09	0.04	0.70
50-100	1.10	0.04	0.72
MORE THAN 100	1.12	0.05	0.71
HOT POT			
0-25	1.00	0.09	1.44
25-50	0.95	0.09	1.38
50-100	1.03	0.08	1.41
MORE THAN 100	0.94	0.09	1.32
BURN TONGUE			
0-25	1.00	0.05	0.84
25-50	0.99	0.05	0.83
50-100	1.01	0.05	0.83
MORE THAN 100	0.97	0.06	0.83
BUMP SHIN			
0-25	1.00	0.07	1.15
25-50	1.00	0.07	1.09
50-100	1.04	0.06	1.11
MORE THAN 100	1.10	0.08	1.14

*WTP < 1 was rounded to 1 to avoid dividing by zero Inverse WTP was normalized so lowest bin is 1

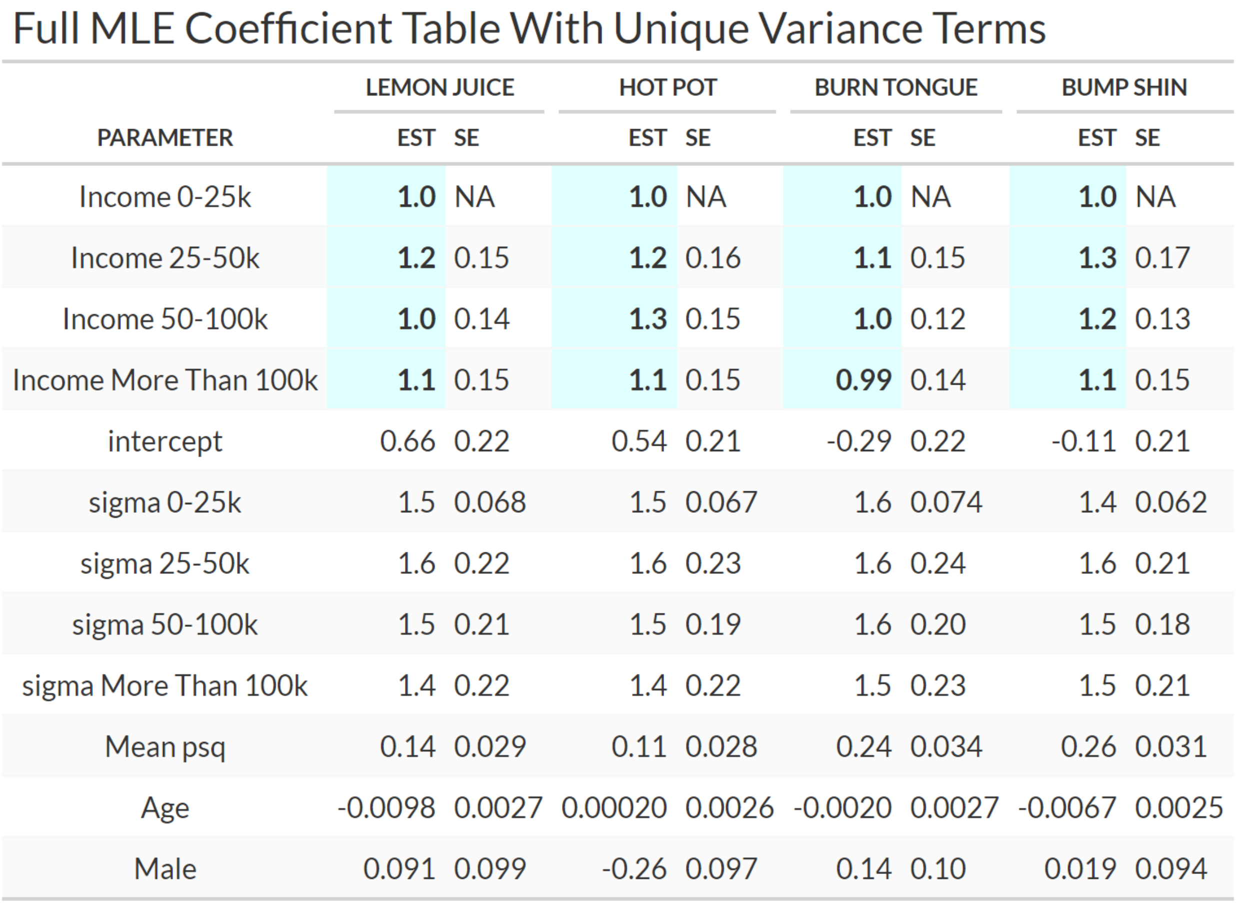
Table A.3: Full MLE Coefficient Table



Full MLE Coefficient Table

	LEMON JUICE		HOT POT		BURN TONGUE		BUMP SHIN	
PARAMETER	EST	SE	EST	SE	EST	SE	EST	SE
Income 0-25k	**1.0**	NA	**1.0**	NA	**1.0**	NA	**1.0**	NA
Income 25-50k	**1.1**	0.061	**1.1**	0.064	**1.0**	0.058	**1.2**	0.068
Income 50-100k	**1.0**	0.054	**1.2**	0.065	**0.99**	0.054	**1.1**	0.057
Income More Than 100k	**1.1**	0.064	**1.1**	0.065	**0.99**	0.060	**1.1**	0.064
intercept	0.61	0.21	0.64	0.20	0.051	0.22	-0.10	0.19
sigma	1.5	0.061	1.4	0.060	1.5	0.065	1.4	0.057
Mean psq	0.13	0.027	0.091	0.027	0.20	0.029	0.25	0.027
Age	-0.0090	0.0025	-0.00059	0.0024	-0.0041	0.0026	-0.0069	0.0023
Male	0.084	0.094	-0.27	0.092	0.11	0.099	0.040	0.088

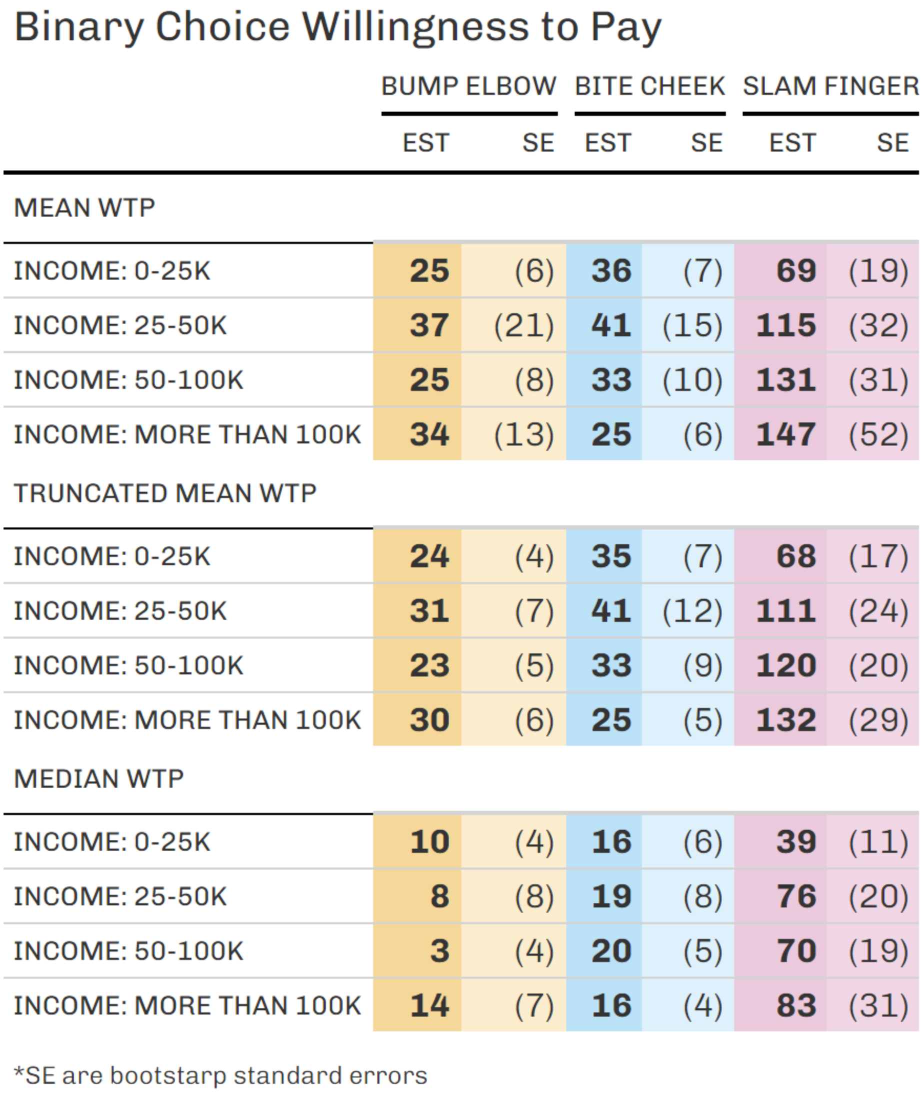
Table A.4: Full MLE Coefficient Table: Income specific Standard Deviation



Full MLE Coefficient Table With Unique Variance Terms

	LEMON JUICE		HOT POT		BURN TONGUE		BUMP SHIN	
PARAMETER	EST	SE	EST	SE	EST	SE	EST	SE
Income 0-25k	**1.0**	NA	**1.0**	NA	**1.0**	NA	**1.0**	NA
Income 25-50k	**1.2**	0.15	**1.2**	0.16	**1.1**	0.15	**1.3**	0.17
Income 50-100k	**1.0**	0.14	**1.3**	0.15	**1.0**	0.12	**1.2**	0.13
Income More Than 100k	**1.1**	0.15	**1.1**	0.15	**0.99**	0.14	**1.1**	0.15
intercept	0.66	0.22	0.54	0.21	-0.29	0.22	-0.11	0.21
sigma 0-25k	1.5	0.068	1.5	0.067	1.6	0.074	1.4	0.062
sigma 25-50k	1.6	0.22	1.6	0.23	1.6	0.24	1.6	0.21
sigma 50-100k	1.5	0.21	1.5	0.19	1.6	0.20	1.5	0.18
sigma More Than 100k	1.4	0.22	1.4	0.22	1.5	0.23	1.5	0.21
Mean psq	0.14	0.029	0.11	0.028	0.24	0.034	0.26	0.031
Age	-0.0098	0.0027	0.00020	0.0026	-0.0020	0.0027	-0.0067	0.0025
Male	0.091	0.099	-0.26	0.097	0.14	0.10	0.019	0.094

Table A.5

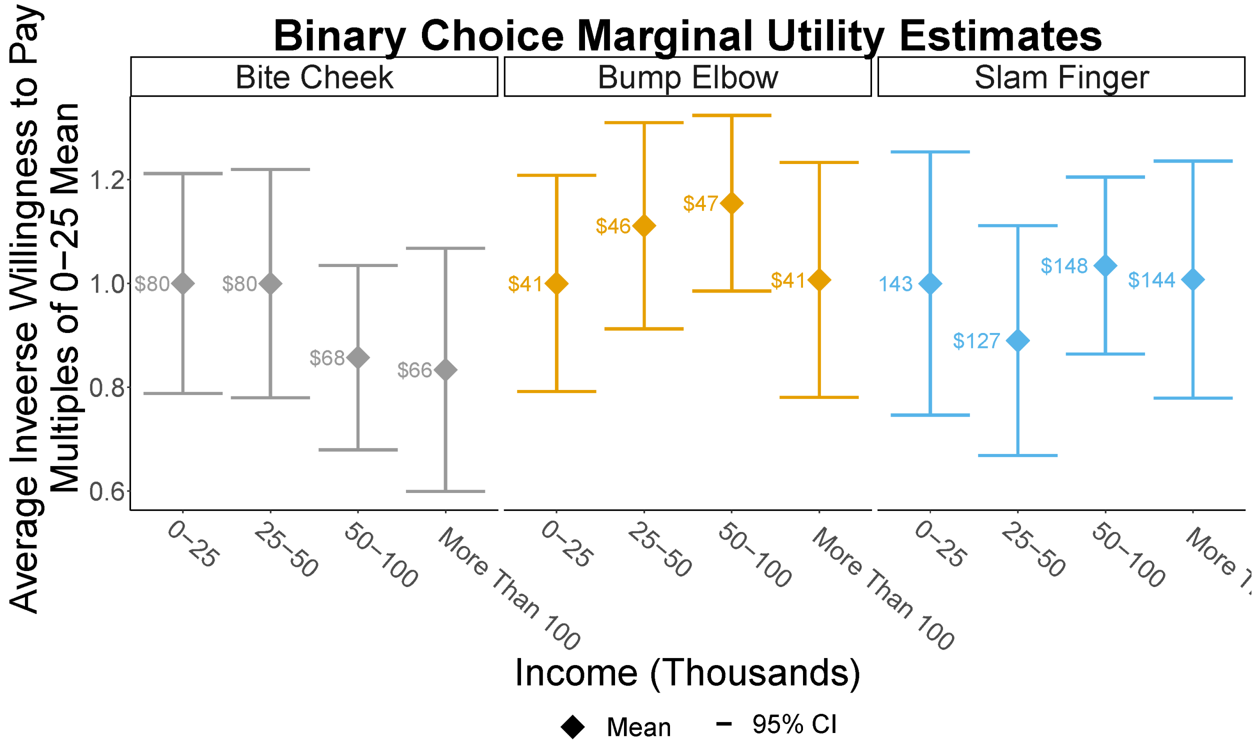


Binary Choice Willingness to Pay

	BUMP ELBOW		BITE CHEEK		SLAM FINGER	
	EST	SE	EST	SE	EST	SE
MEAN WTP						
INCOME: 0-25K	**25**	(6)	**36**	(7)	**69**	(19)
INCOME: 25-50K	**37**	(21)	**41**	(15)	**115**	(32)
INCOME: 50-100K	**25**	(8)	**33**	(10)	**131**	(31)
INCOME: MORE THAN 100K	**34**	(13)	**25**	(6)	**147**	(52)
TRUNCATED MEAN WTP						
INCOME: 0-25K	**24**	(4)	**35**	(7)	**68**	(17)
INCOME: 25-50K	**31**	(7)	**41**	(12)	**111**	(24)
INCOME: 50-100K	**23**	(5)	**33**	(9)	**120**	(20)
INCOME: MORE THAN 100K	**30**	(6)	**25**	(5)	**132**	(29)
MEDIAN WTP						
INCOME: 0-25K	**10**	(4)	**16**	(6)	**39**	(11)
INCOME: 25-50K	**8**	(8)	**19**	(8)	**76**	(20)
INCOME: 50-100K	**3**	(4)	**20**	(5)	**70**	(19)
INCOME: MORE THAN 100K	**14**	(7)	**16**	(4)	**83**	(31)

*SE are bootstarp standard errors

Figure A.1



Numbers are un−normed mean inverse willingness to pay

A.3. Empirical Extensions and Clarifications

A.3.1 Empirical Model With Indirect Utility

The simple empirical model laid out in the body of the paper includes money as a numeraire good. However, the same estimation strategy can be supported using heterogeneous preferences and indirect utility. I think this model is more accurate to the real world where there are many goods and a whole set of prices, but the estimation is ultimately identical, and it is more confusing. Thus, it is here in the appendix.

First consider an indirect utility function V that is a function of prices P and income y, but also other characteristics θ like gender, age, pain sensitivity, or anything else that might influence preferences. This allows V to be heterogeneous across different people. This gives

$$V(P, Y, \theta) \tag{A.1}$$

Now for the pain relief described in our questions needs to be re-characterized as a price change. Let P be the price vector in our current world and let P' be a price vector where immediate pain relief is free. Now we can define the change in utility from pain relief as follows.

$$V(P', Y, \theta) = V(P, Y, \theta) = \Delta V^j \tag{A.2}$$

Since these are relatively small changes we can, as in the empirical model in the body, treat this as a marginal change. That means the following equality holds

$$EV(P', P, Y, \theta)U'_y(P, Y) = \Delta V^j \tag{A.3}$$

Where U'_y is the marginal utility of income and EV is equivalent variation. To shorten the notation, let $EV(P', P, Y, \theta) = EV^j$.

An important note regarding the previous step is that the marginal utility of income is not a function θ. So, while the preference for pain relief may vary by characteristics like age, the marginal utility of income cannot.

Now to identify the marginal utility of income we need the same assumption as in the simpler model. That is, we need the utility from pain relief to be independent of income. With that we get the following

Theorem 7. *if* $\Delta V^j \perp\!\!\!\perp Y$ *then*

$$\mathbb{E}[U'_y(P, Y)|Y, P] = \frac{\alpha}{\mathbb{E}[EV(P', P, Y, \theta)|Y, P]} \tag{A.4}$$

Proof. Given $\Delta V^j \perp\!\!\!\perp Y$ we get that

$$\mathbb{E}[\Delta V^j|P, Y] = \mathbb{E}[\Delta V^j|P] = \alpha \tag{A.5}$$

□

The first equation follows from the independence of the change in utility, ΔV^j and income Y. The second equality is a bit of an abuse of notation, but here the price vector is actually set to our current price vector P, and so this is a constant we can normalize to any level. Now, given this equality we get

$$\alpha \tag{A.6}$$
$$= \mathbb{E}[\Delta V^j|Y, P] \tag{A.7}$$
$$= \mathbb{E}[EV(P', Y, P, \theta)U'_y(Y, P)|Y, P] \tag{A.8}$$
$$= \mathbb{E}[EV(P', Y, P, \theta)|Y, P]E[U'_y(Y, P)|Y, P] \tag{A.9}$$
$$\implies \mathbb{E}[U'_y(P, Y)|Y, P] = \frac{\alpha}{\mathbb{E}[EV(P', P, Y, \theta)|Y, P]} \tag{A.10}$$

The second line is the normalization explained above, the third line is the identify also explained above, the last line comes from the fact that $U'_y(Y, P)$ is only a function of Y and P and so conditioning on those makes it a constant. Rearranging the equation gives us the theorem.

One point this more complex model makes clearer is that the model estimates the marginal utility of nominal dollars, conditional on a price vector. A nominal dollar might not provide the same purchasing power to everyone in a world of non-linear pricing, quality variation, credit constraints, non-convex preferences, and other complications outside of basic economics models. In particular, a marginal dollar may have more, or less, purchasing power the more dollars someone has. A simple example would be to consider geographic sorting. Suppose richer people live in more expensive areas. In this case richer people value dollars less because the marginal value of consumption is lower, but also because a dollar literally buys less at stores in their area. This example might theoretically be controlled for with geographic price indices, but other examples are more complicated.

The following story popularized by novelist Terry Pratchett illustrating why the rich are able to spend less. Suppose a quality pair of boots that will last ten years is $50, but a cheap pair that will last only a year is $10. A poor person, with a marginal dollar, may only be able to purchase the cheaper option despite it being more costly in the long run (*Flood*). The poor person may appreciate the boots more, in line with diminishing marginal utility of consumption, but the richer person is able to purchase boot years at half the price, making their marginal consumption per dollar higher. This example supports the colloquial saying[1], "it's expensive to be poor". Higher credit rates, an inability to buy in bulk or take advantage of off-peak sales might mean the poor just can't buy as much with an additional dollar. Berkouwer and Dean, for example, show that households in Nairobi are only willing to pay $12 for a stove that would save $237 over two years, and that a low interest loan increases willingness to pay to the actual savings over the life of the loan (*Berkouwer and Dean*, 2021).

Perhaps a more fundamental consideration is that consumption is not homothetic and so the types of goods people consume with their first $10,000 a year look very different than after making $100,000. The first priorities for consumption are essentials like food and shelter. If food staples and housing become more expensive relative to luxury or entertainment goods, than we will see the marginal utility of a dollar for the poor fall relative to the rich. So, in general, the marginal utility of a nominal dollar also captures the relative cost of the differing consumption baskets of each income group. Interestingly, looking at these utility estimates over time could capture to what extent inflation has been concentrated on essentials or low-quality items compared to luxury goods.

[1] I've at least heard this a lot among family and friends. Not sure how common it actually is

A.3.2 Marginal Relief Assumption

While the empirical model in the body of the paper treats pain relief as a marginal change, in truth, the questions are a binary choice. Pay and receive total pain relief, or don't. How does this change the model? In words, I am making a local linearity assumption. I average the reservation prices within an income group and so a reservation price that is twice as high is treated as twice as much utility lost. If utility is concave, this is not correct since losing twice as much money should be more than twice as bad. The extent to which this biases the result is a function of the concavity of utility and the size of the difference. Over small changes, approximately marginal, the linearity assumption will be not so wrong. While our priors may be that someone with twice the income has very a different marginal utility of income, it is typically not assumed that say, a $100 difference in income will drastically alter the marginal utility of income. To formalize this, we can re-characterize equation 1.4.1 as

$$U(m_i, q_i, X_i, \epsilon_i) = \phi(m_i) + r(X_i, \epsilon_i) \tag{A.11}$$

Where now r is a binary choice good for full relief or no relief. Now, the indifference condition is characterized by

$$\phi(m_i) = \phi(m_i - P_i^r(m_i)) + r(X_i, \epsilon_i) \tag{A.12}$$

Taking a first order Taylor approximation gives

$$\phi(m_i - P_i^r(m_i)) \approxeq \phi(m_i) - \phi'(m_i)P_i^r(m_i) \tag{A.13}$$

Inserting this into the indifference condition gives

$$\phi(m_i) \approxeq \phi(m_i) - \phi'(m_i)P_i^r(m_i) + r(X_i, \epsilon_i) \tag{A.14}$$

and finally rearranging gives us

$$P_i^r(m_i) \approxeq \frac{r(X_i, \epsilon_i)}{P_i^r(m_i)} \tag{A.15}$$

The Taylor approximation will be closer to correct the smaller the change in utility and the closer to linear utility is over the range from m_i to $m_i - P_i^r(m_i)$

An alternative way to see this is to consider the following exact equation

$$\frac{1}{P_i^r(m_i)} \int_{m_i - P_i^r(m_i)}^{m_i} \phi'(m_i) = \frac{r(X_i, \epsilon_i)}{P_i^r(m_i)} \tag{A.16}$$

Technically, what we are identifying with the inverse of the reservation price is the average marginal utility of income over the range from m_i to $m_i - P_i^r(m_i)$. Since the prices are small, this average is probably close to the marginal utility.

Suppose this really were a big concern. For example, suppose the reservation prices were larger and/or I was finding more concave utility. One potential solution would be to iteratively estimate the marginal utility of income function and, for each iteration, use the previous estimate to compute the integral in equation A.16 until the estimates converge. Given my results indicate a linear utility, I have not done this or formalized the econometrics, but I expect it would provide more accurate estimates in this hypothetical case.

A.3.3 Proof of Theorems 1 through 3

A.3.3.1 Theorem 1 Proof

Proof. Start by taking the conditional expectation of both sides of equation 1.4.2. The expectations here are expectations across people for a given income level m. This gives

$$\begin{aligned}
\mathbb{E}[P^r(m_i)|m] &= \mathbb{E}[\frac{r'(q_i, X_i, \epsilon_i)}{\phi'(m_i)}|m] \\
&= \frac{\mathbb{E}[r'(q_i, X_i, \epsilon_i)|m]}{\phi'(m_i)} \\
&= \frac{\mathbb{E}[r'(q_i, X_i, \epsilon_i)]}{\phi'(m_i)} \\
&= \frac{\alpha}{\phi'(m_i)} \\
\implies \phi'(m_i) &= \frac{\alpha}{\mathbb{E}[P_i^r(m_i)|m]}
\end{aligned}$$

After conditioning on m, $\phi'(m_i)$ is a constant and so it can be removed from the expectation. Next, $r_i' \perp\!\!\!\perp m_i$ by assumption, and so we can remove the condition from the numerator. Finally, $\mathbb{E}[r_i']$ is a constant that we can normalize to α

□

A.3.3.2 Theorem 2 Proof

Proof.

$$\begin{aligned}
&\mathbb{E}[P^r(m_i)|m] = \mathbb{E}[\frac{r'(q_i, X_i, \epsilon_i)}{\phi_i(m_i)}|m] \\
&= \mathbb{E}[\frac{r'(q_i, X_i, \epsilon_i)}{\mu_i \phi(m_i)}|m] \\
&= \mathbb{E}[\frac{r'(q_i, X_i, \epsilon_i)\frac{1}{\mu_i}}{\phi(m_i)}|m] \\
&= \frac{\mathbb{E}[\frac{r'(q_i, X_i, \epsilon_i)}{\mu_i}|m]}{\phi(m_i)} \\
&= \frac{\mathbb{E}[\frac{r'(q_i, X_i, \epsilon_i)}{\mu_i}]}{\phi(m_i)} \\
&= \frac{\alpha}{\phi(m_i)} \\
&= \frac{\alpha}{\mathbb{E}[\phi_i(m_i)|m]} \\
&\implies \mathbb{E}[\phi_i(m_i)|m] = \frac{\alpha}{\mathbb{E}[P_i^r(m_i)|m]}
\end{aligned}$$

The second line comes from the definition of $\phi_i(m_i)$, the third is algebra, the fourth comes from $\phi(m_i)$ being a constant conditional on m, the fifth lines follows from the independence assumption, the sixth is just normalizing a constant utility level to α, and the seventh is just rewriting $\phi(m_i)$ using it's definition. □

A.3.3.3 Theorem 3 Proof

Proof. Rearranging the indifference condition from equation 1.4.2 gives

$$\frac{\phi_i(m_i)}{r'(q_i, X_i, \epsilon_i)} = \frac{1}{P_i^r(m_i)}$$

Now taking the conditional expectation of both sides, we get

$$\begin{aligned}
&\mathbb{E}\Big[\frac{1}{P_i^r(m_i)}|m\Big] \\
&= \mathbb{E}\Big[\frac{\phi_i(m_i)}{r'(q_i, X_i, \epsilon_i)}|m\Big] \\
&= \mathbb{E}[\phi_i(m_i)|m]\mathbb{E}\Big[\frac{1}{r'(q_i, X_i, \epsilon_i)}|m\Big] \\
&= \mathbb{E}[\phi_i(m_i)|m]\mathbb{E}\Big[\frac{1}{r'(q_i, X_i, \epsilon_i)}\Big] \\
&= \mathbb{E}[\phi_i(m_i)|m]\frac{1}{\alpha} \\
&\implies \mathbb{E}[\phi(m_i)|m] = \alpha\mathbb{E}\Big[\frac{1}{P_i^r(m_i)}|m\Big]
\end{aligned}$$

where the third line comes from $r'(q_i, X_i, \epsilon_i) \perp\!\!\!\perp \phi_i(m_i)|m$, the fourth line comes from $r'(q_i, X_i, \epsilon_i) \perp\!\!\!\perp m_i$, and the fifth line comes from normalizing utility.

□

A.3.4 Maximum Likelihood Identification

Equations 1.6.4 and 1.6.5 lead to the following theorem specifying the identification of the parameters. Let $\theta = (\boldsymbol{\beta}, \boldsymbol{\phi'})$ be the full set of parameters in the model.

Theorem 8. *If the conditions in assumption 3 and 4 and definition 1 hold and we also have that the matrix $[\mathbb{X} \quad \mathbb{M}]$ is full rank, then the ratio of any two parameters in θ is identified. If we normalize the marginal utility of income for the lowest income group to one, that is $\phi'_1 = 1$, than the remaining parameters in θ are identified.*

To prove this Suppose there exists a $\theta^* \neq \theta$ s.t. $\mathbb{E}[\mathbb{P}^r|\theta^*] = \mathbb{E}[\mathbb{P}^r|\theta]$. This implies

$$\mathbb{X}_i\boldsymbol{\beta}^* \oslash \mathbb{M}\phi'^* = \mathbb{X}_i\boldsymbol{\beta} \oslash \mathbb{M}\phi' \tag{A.17}$$

or that

$$\frac{\beta_1^* + X_i\boldsymbol{\beta}^*}{\sum_{k=1}^b \mathbb{1}_{ik}\phi_k'^*} = \frac{\beta_1 + X_i\boldsymbol{\beta}}{\sum_{k=1}^b \mathbb{1}_{ik}\phi_k'} \quad \forall \quad i \tag{A.18}$$

Now it is true that $\theta = \alpha\theta^*$, where α is any constant, satisfies the condition since α cancels out in the numerator and denominator. Once we have normalized $\phi_1 = 1$, however, α no longer appears in the denominator for i's in income group 1 and so does not cancel. If $\phi'_2 = .5$, for example, it implies the marginal utility of a dollar for income group 2 is half that of group 1.

With the $\phi_1 = 1$ normalization, any change to a parameter in the numerator would alter the expected reservation price for those in group one and violate the equality in equation A.18, assuming, as in a regression, that $\mathbb{X}$ is full rank. Any change to the other marginal utility parameters ϕ'_k could be cancelled out for that group by appropriately scaling the numerator, but, since the marginal utility of group one is fixed and they share the numerator parameters, this would again change the expected reservation price for group 1 and violate A.18. Thus, there does not exist a $\theta^* \neq \theta$ satisfying the condition.

Can this same identification strategy be used for any good? No, the marginal utility of income cannot be identified from just any reservation price. Recall the assumption that $\epsilon_i \perp\!\!\!\perp m_i$ and notice that income does not enter the utility function for pain relief and the factors impacting pain relief to not impact the marginal utility of income. If either of these appeared in both the numerator and denominator, we would not not be able to uniquely identify the parameters.

A.3.5 Binary Choice Expectation of Inverse Price

The average reservation price is found by using the integral of one minus the CDF for the reservation price $G_{P_r}(P)$ like so:

$$\bar{P_r^j} = \int_0^\infty [1 - G_{P_r}(P)]dP \tag{A.19}$$

To get an estimate of the average inverse of the reservation price in line with theorem 3 we need the CDF of $Y = \frac{1}{x}$. Using the following, we get

$$\begin{aligned}
&F_n(Y) = P(Y < p) \\
&= P(\frac{1}{x} < p) \\
&= P(\frac{1}{p} < x) \\
&= 1 - G_{P_r}(\frac{1}{p}) \\
&\implies \mathbb{E}[\frac{1}{p}] = \int_0^\infty [G_{P_r}(\frac{1}{p})]dp \\
&= \int_0^\infty [\frac{1}{1 + e^{\delta_j + \bar{X}\gamma + \beta_j \frac{1}{p}}}]
\end{aligned}$$

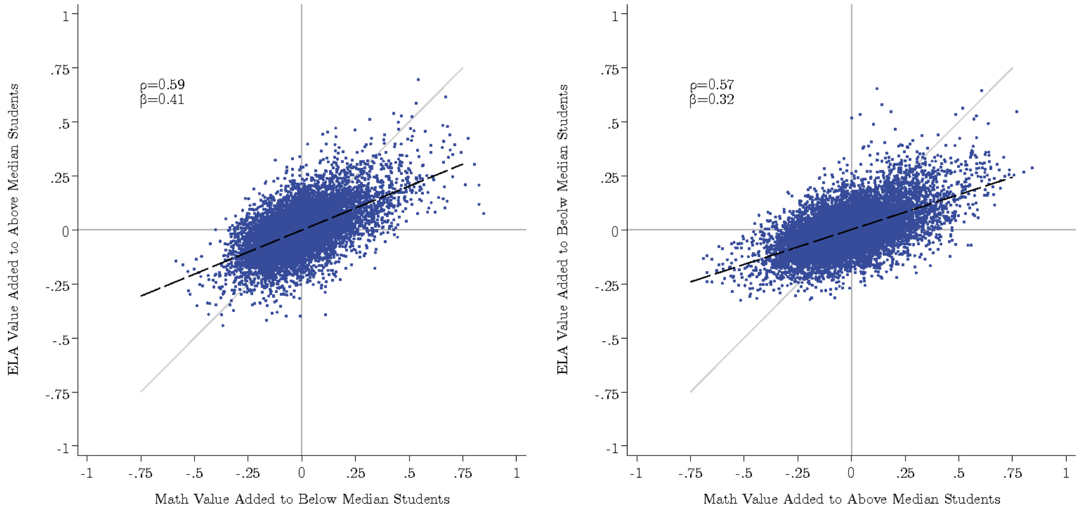
The results of this estimation are in figure ??. In my Monte Carlo Simulations for this estimate, the truncated estimate, which limits the integral to the highest bid, was the only stable and accurate estimator for the average inverse reservation price. This is what is presented in the figure.

APPENDIX B

Appendix to Chapter 2

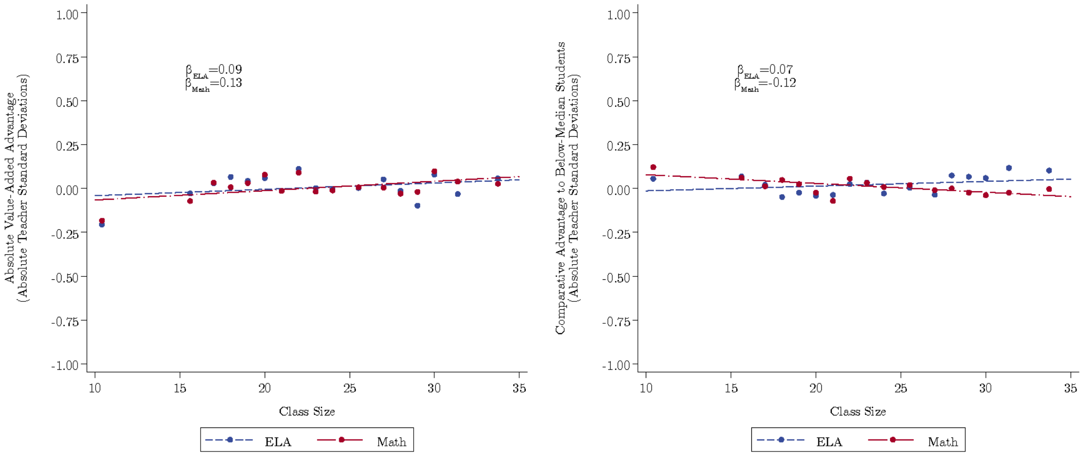
B.1. Additional Tables and Figures

Figure B.1: Cross-Subject and Cross-Type Value Added Is Much Less Correlated



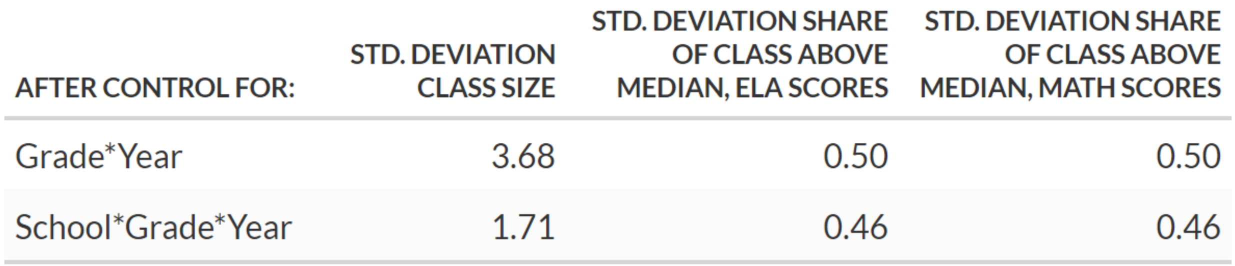
Note: This figure shows our heterogeneous estimates of teacher value added on both English Language Arts (ELA) and Math test scores. Note that in this Figure Math and ELA scores are plotted against each other. Each dot represents one teacher-year estimate of value added on higher- and lower-scoring students. 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.

Figure B.2: 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.

Table B.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.

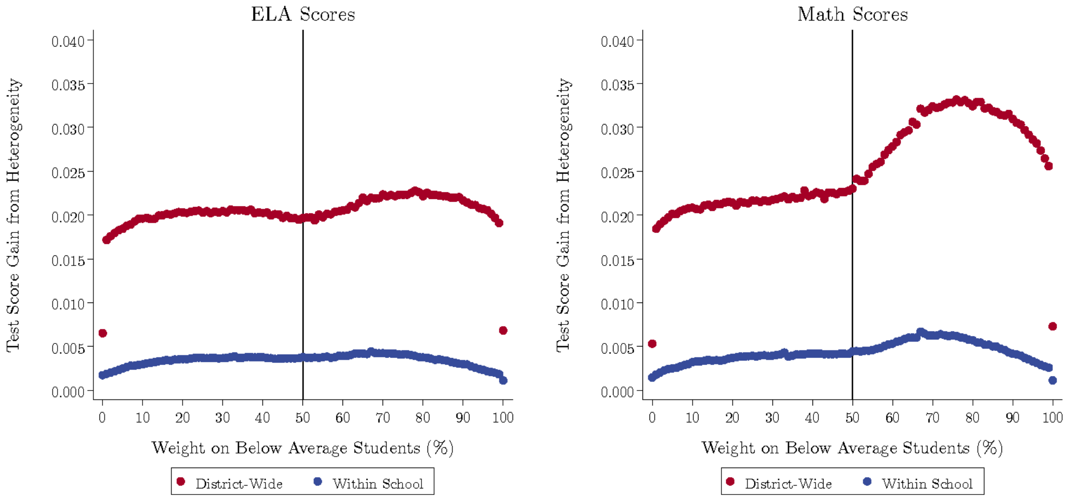
B.2. Theory Appendix

B.2.1 From Test Scores to Welfare Details

Below is a more detailed version of definition 2.2.1

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 $\Delta\widetilde{\mathcal{W}}^j$ from

Figure B.3: 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.

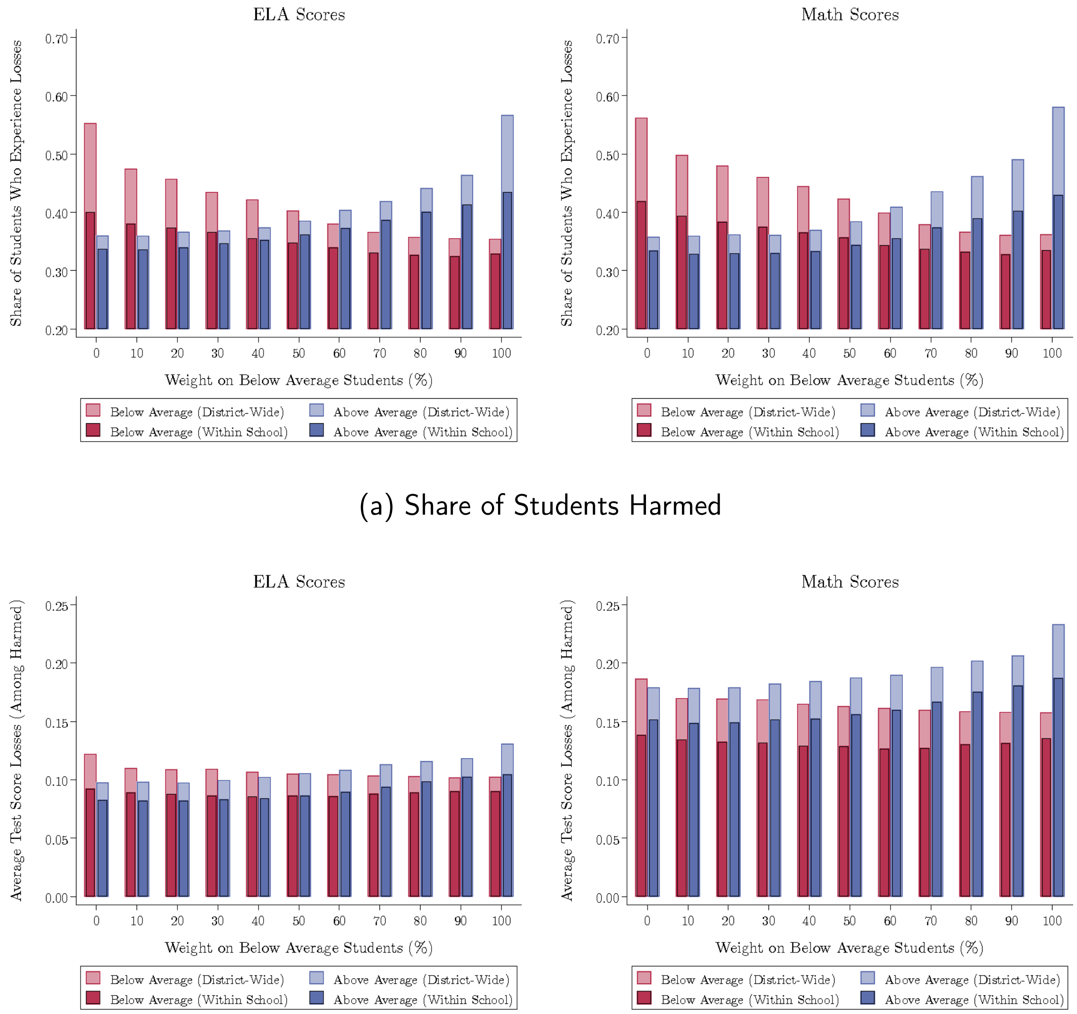
the status quo policy ($j = 0$) to policy j is

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

□

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

Figure B.4: While Reallocations Help Many Students, They Will Harm Others

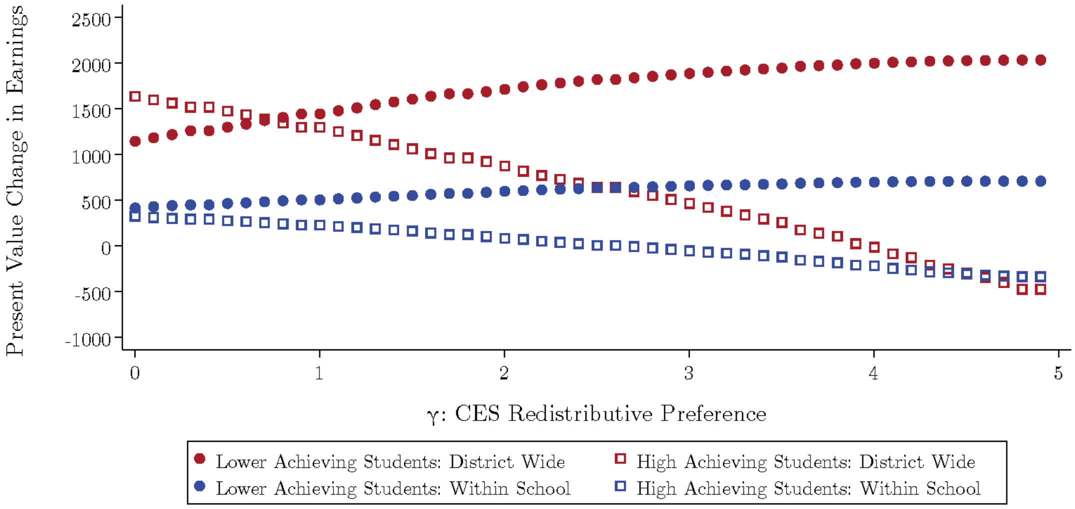


(a) Share of Students Harmed

(b) Mean Score Change among Harmed Students

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 B.5: 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.

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 2.2.1 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 2.2.1 provides a ground truth reference that incorporates a large amount of of potential heterogeneity, individual differences.

B.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

$$
\begin{aligned}
&\Delta\mathcal{W}^j && \text{(B.1)}\\
&= \int_0^1 \gamma_i(S_i^j, S_i^0)\Delta S_i^p \mathsf{d}i && \text{(B.2)}\\
&= \frac{\int_0^1 \gamma_i(S_i^j, S_i^0)\Delta S_i^p \mathsf{d}i}{\int_0^1 \Delta S_i^p \mathsf{d}i}\int_0^1 \Delta S_i^p \mathsf{d}i && \text{(B.3)}\\
&= \tilde{\gamma}^j ATE^j && \text{(B.4)}
\end{aligned}
$$

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.

B.2.3 Theorem 4 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 2.2.1. 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. □

B.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.

B.2.5 Conditional Average Treatment Effect Bias Explained

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

$$\mathbb{E}[\Delta S^p] - \widehat{ATE} > \sum_x P_x \Big(\mathbb{E}[\Delta S^p | x] - C\widehat{ATE(X)} \Big) \tag{B.5}$$

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

$$
\begin{aligned}
& ATE_d^j - \widehat{ATE}_d^j && \text{(B.6)}\\
& = \frac{1}{K_d^j}\sum_{i=1}^{K_d^j}\Delta S_i^d - \frac{1}{K_d^0}\sum_{l=1}^{K_d^0}\Delta S_l^d && \text{(B.7)}\\
& > \sum_X P_{dx}^j \left(\frac{1}{K_{dx}^j}\sum_{i=1}^{K_{dx}^j}\Delta S_i^d - \frac{1}{K_{dx}^0}\sum_{l=1}^{K_{dx}^0}\Delta S_l^d \right) && \text{(B.8)}\\
& \sum_X P_{dx}^j \left(\widehat{ATE}_{dx}^j - \widehat{ATE}_{dx}^0 \right) && \text{(B.9)}
\end{aligned}
$$

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).

B.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.

B.3.1 Estimators

B.3.1.1 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

$$S_{i,s,t} = \alpha_{j(i,s,t)} + \beta_s X_{i,t} + \epsilon_{i,s,t} \tag{B.1}$$

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[1]. $j(i,t)$ 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

$$\bar{A}_{jt} = \frac{1}{n} \sum_{i \in i:j(i,t)=j} A_{it} \tag{B.2}$$

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)"

$$\psi = \arg\min_{\psi} \sum_j \left(\bar{A}_{jt} - \sum_{s=1}^{t-1} \psi_s \bar{A}_{js}\right)^2 \tag{B.3}$$

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

$$\hat{\mu}_{jt} = \psi' \mathbf{A}_j^{-t} \tag{B.4}$$

B.3.1.2 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

$$S_{i,t} = \alpha_{j(i,b,t)} + \beta X_{i,t} \tag{B.5}$$

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

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,

$$\bar{A}_{jBt} = \frac{1}{n} \sum_{i \in i : j(i,B,t)=j} A_{it} \tag{B.6}$$

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

$$\psi = \arg\min_{\psi} \sum_{j} \left(\bar{A}_{jBt} - \sum_{s=1}^{t-1} \psi_s \bar{A}_{jBs} \right)^2 \tag{B.7}$$

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

$$\hat{\mu}_{jBt} = \psi' \mathbf{A}_{jB}^{-t} \tag{B.8}$$

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.

B.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 5.

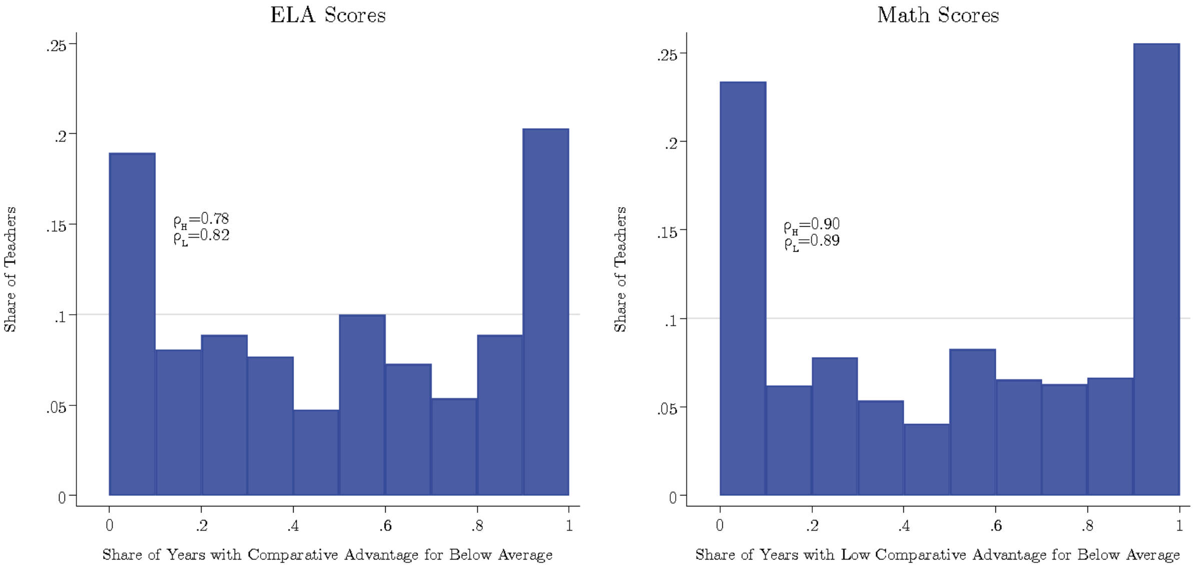
However, in some cases, value added is also used for general teacher ranking and assessment. If teacher heterogeneity is significant, is there still a way to objectively rank teachers according to a particular set of heterogeneous welfare weights? There is not a perfect single solution since their impact 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

$$VA_j = \sum_{k} \gamma_k \bar{\omega_k} \alpha_{j,k} \tag{B.9}$$

This gives the welfare benefit a teacher would have on an average class. This is the same as

Figure B.6: Measures of Comparative Advantage Persistent



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 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 B.9 does help to rank teachers based on what is under their control.

B.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:

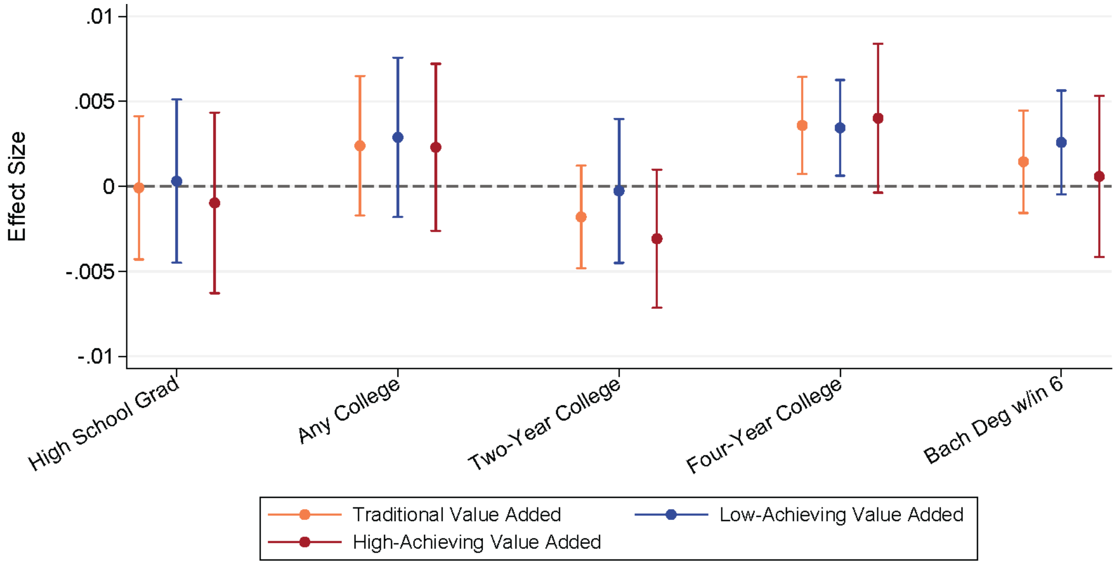
$$\begin{aligned} y_{i,j,s,t} &= \tau_{VA}\hat{\gamma}^{VA}_{j,s,t}\mathbb{1}(k_i = g) + \beta_2 X_i + \nu_{i,j,s,t} \\ y_{i,j,s,t} &= \sum_{g=H,L} \tau_g \hat{\gamma}^{g}_{j,s,t}\mathbb{1}(k_i = g) + \beta_3 X_i + \nu_{i,j,st} \end{aligned} \tag{B.1}$$

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 B.7 reports the results from the regression in equation B.1 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 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.

Figure B.7: Our Estimates Predict Long Term Effects as Well as Standard VA



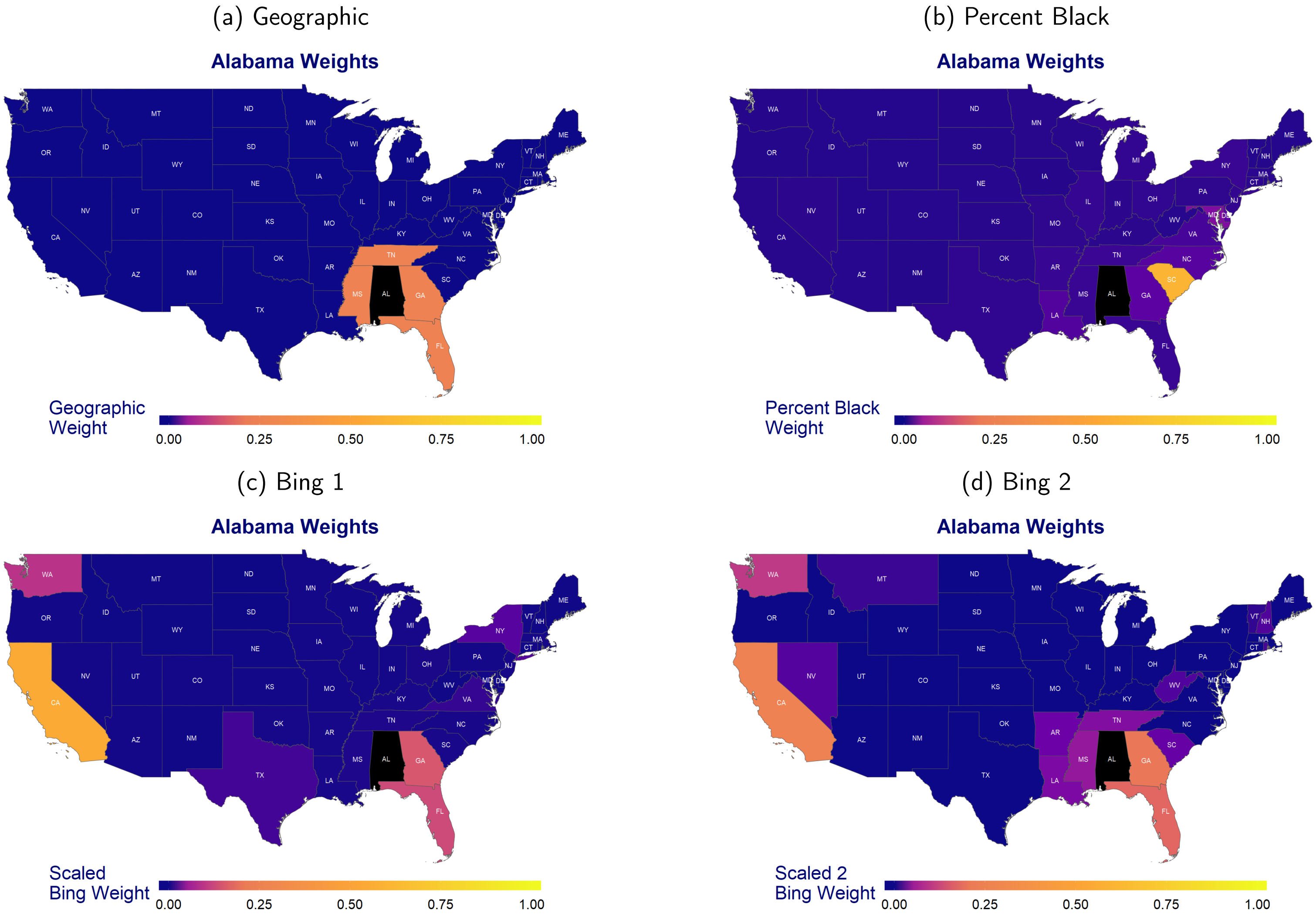
Note: This figure compares the effect of different measures of teacher value added on long-term outcomes. All regressions follow equation B.1 and include all controls from the value added estimation. For the outcomes, High School Grad is an indicator for whether the student graduated from high school, 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.

APPENDIX C

Appendix to Chapter 3

C.1. Figures

Figure C.1: Alabama Maps



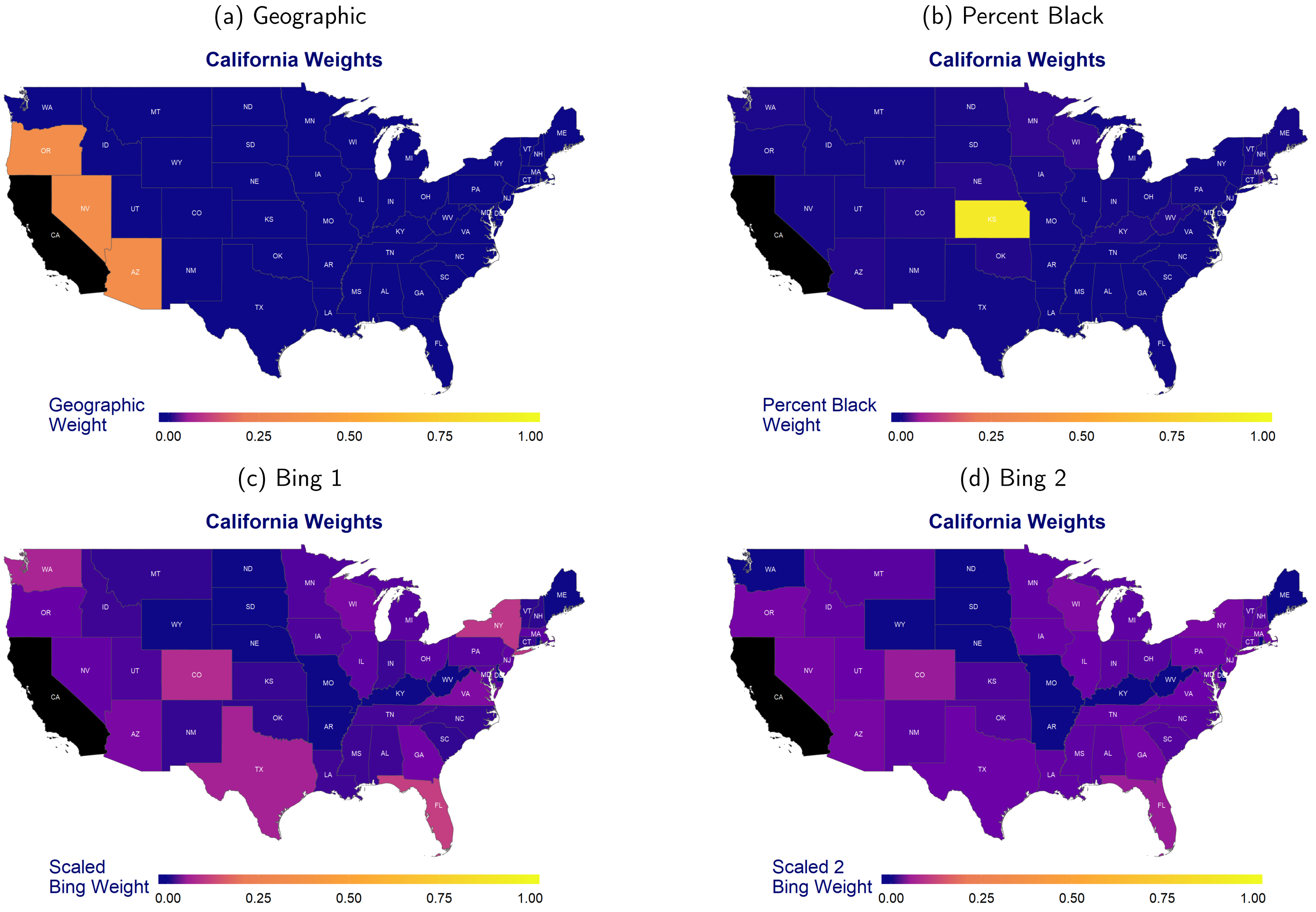
(a) Geographic

(b) Percent Black

(c) Bing 1

(d) Bing 2

Figure C.2: California Maps



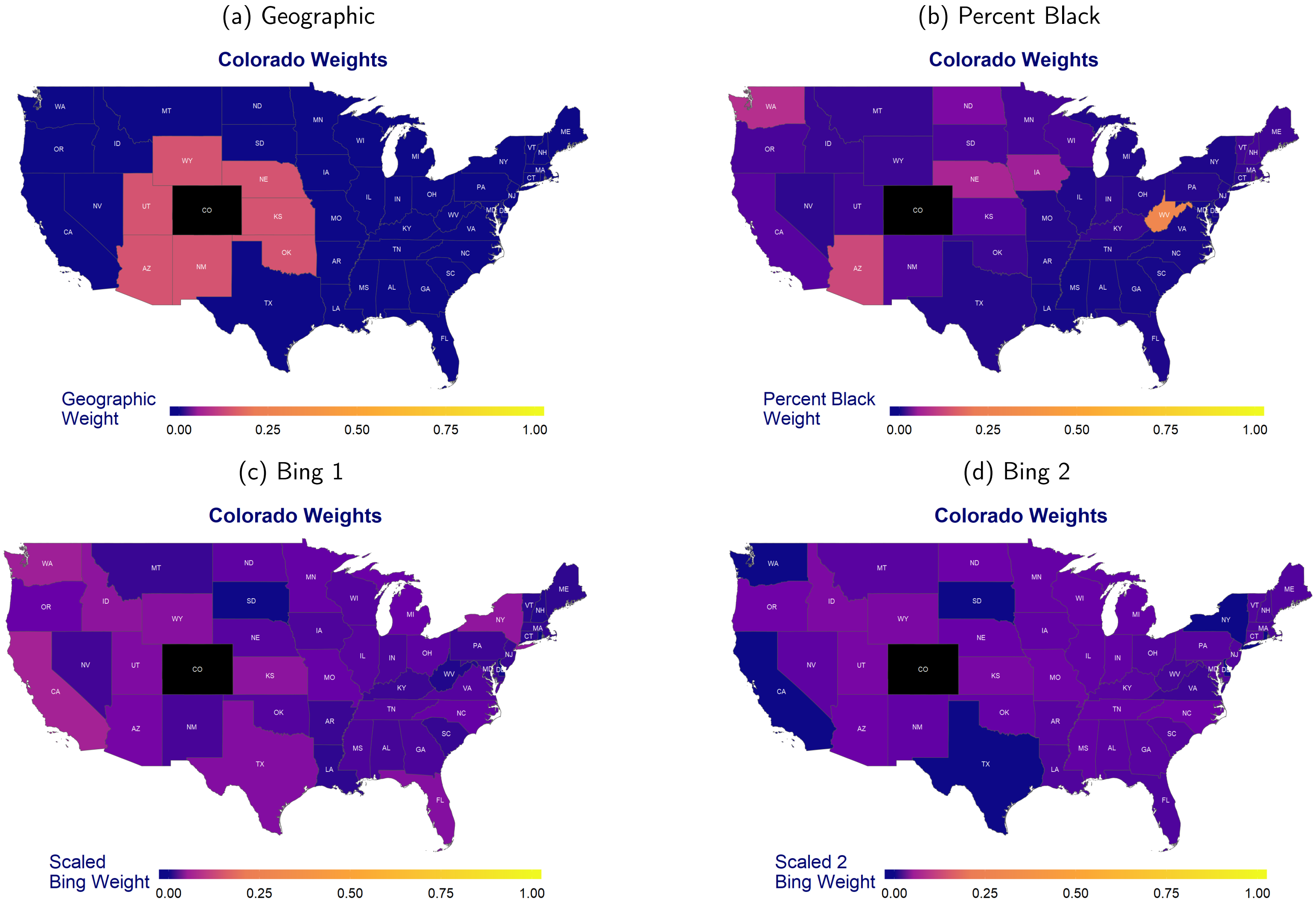
(a) Geographic

(b) Percent Black

(c) Bing 1

(d) Bing 2

Figure C.3: Colorado Maps



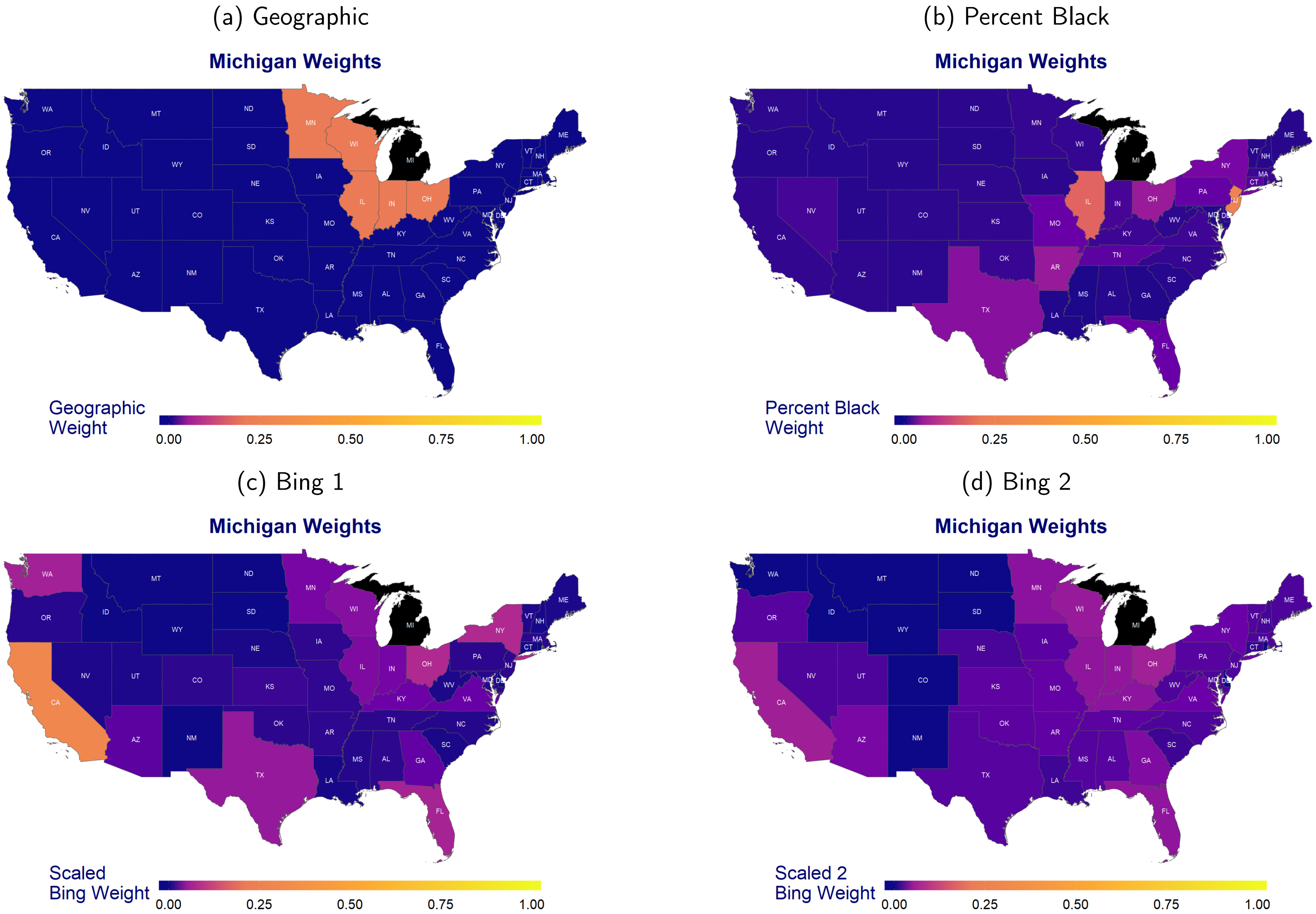
(a) Geographic

(b) Percent Black

(c) Bing 1

(d) Bing 2

Figure C.4: Michigan Maps



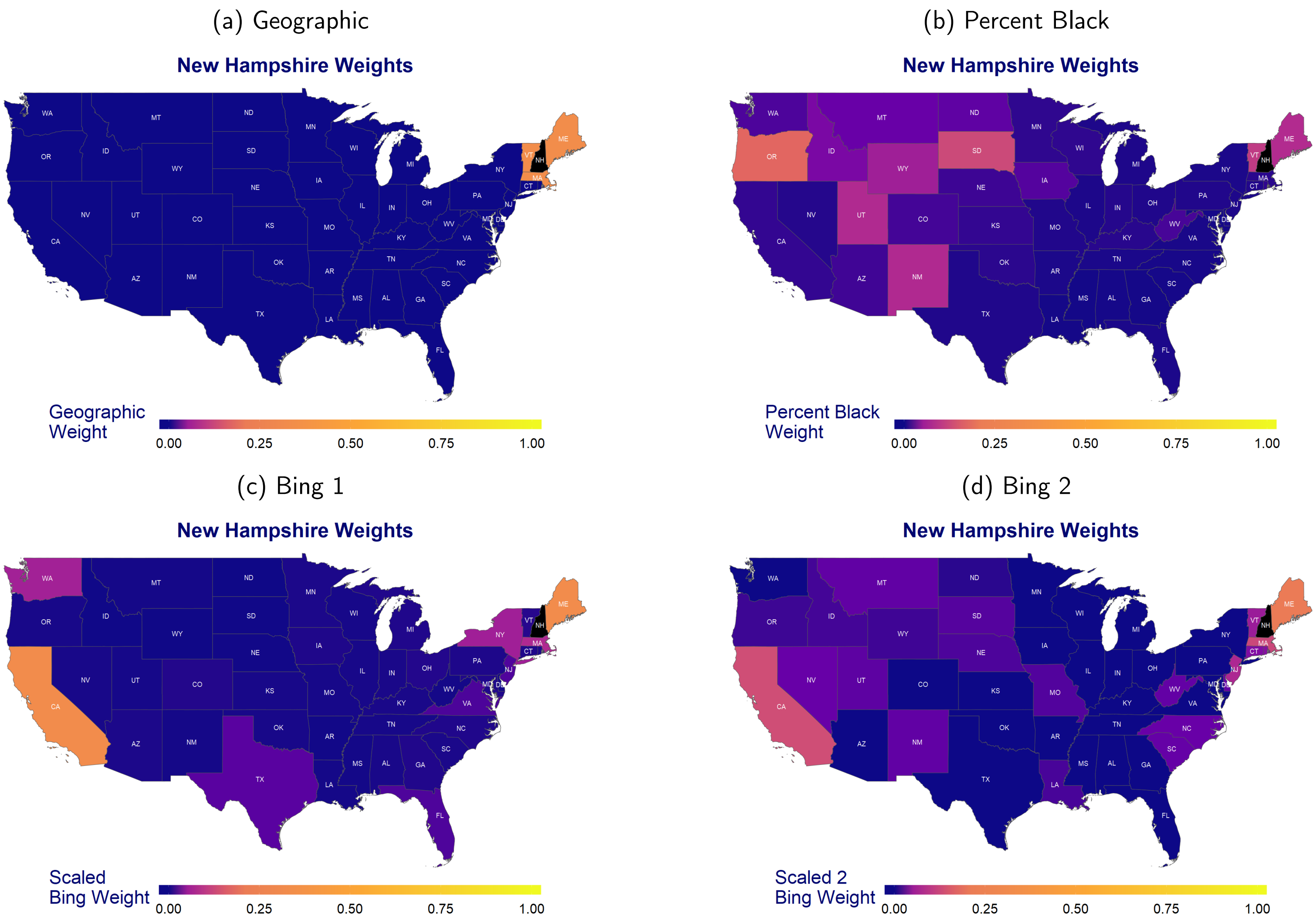
(a) Geographic

(b) Percent Black

(c) Bing 1

(d) Bing 2

Figure C.5: New Hampshire Maps



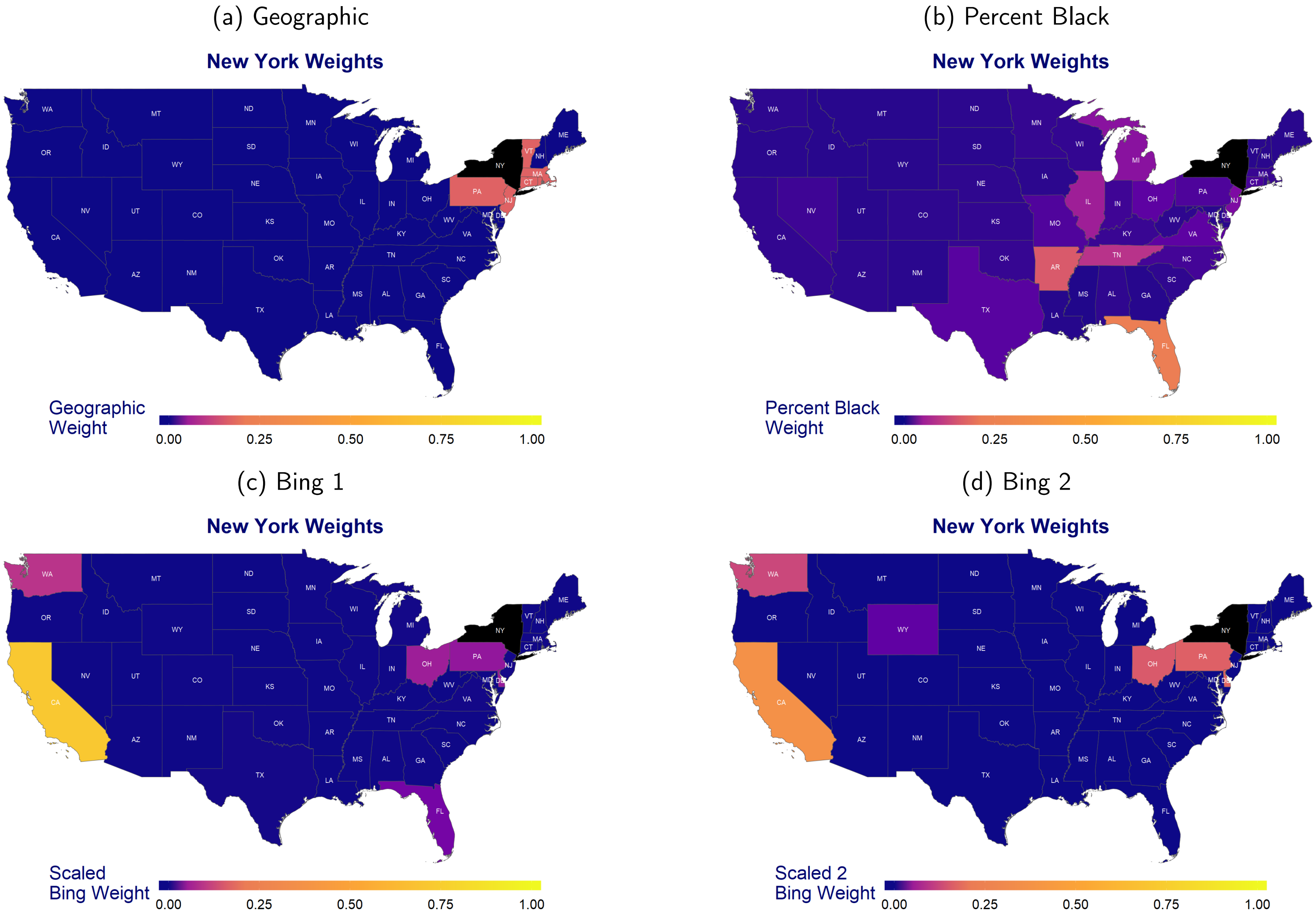
(a) Geographic

(b) Percent Black

(c) Bing 1

(d) Bing 2

Figure C.6: New York Maps



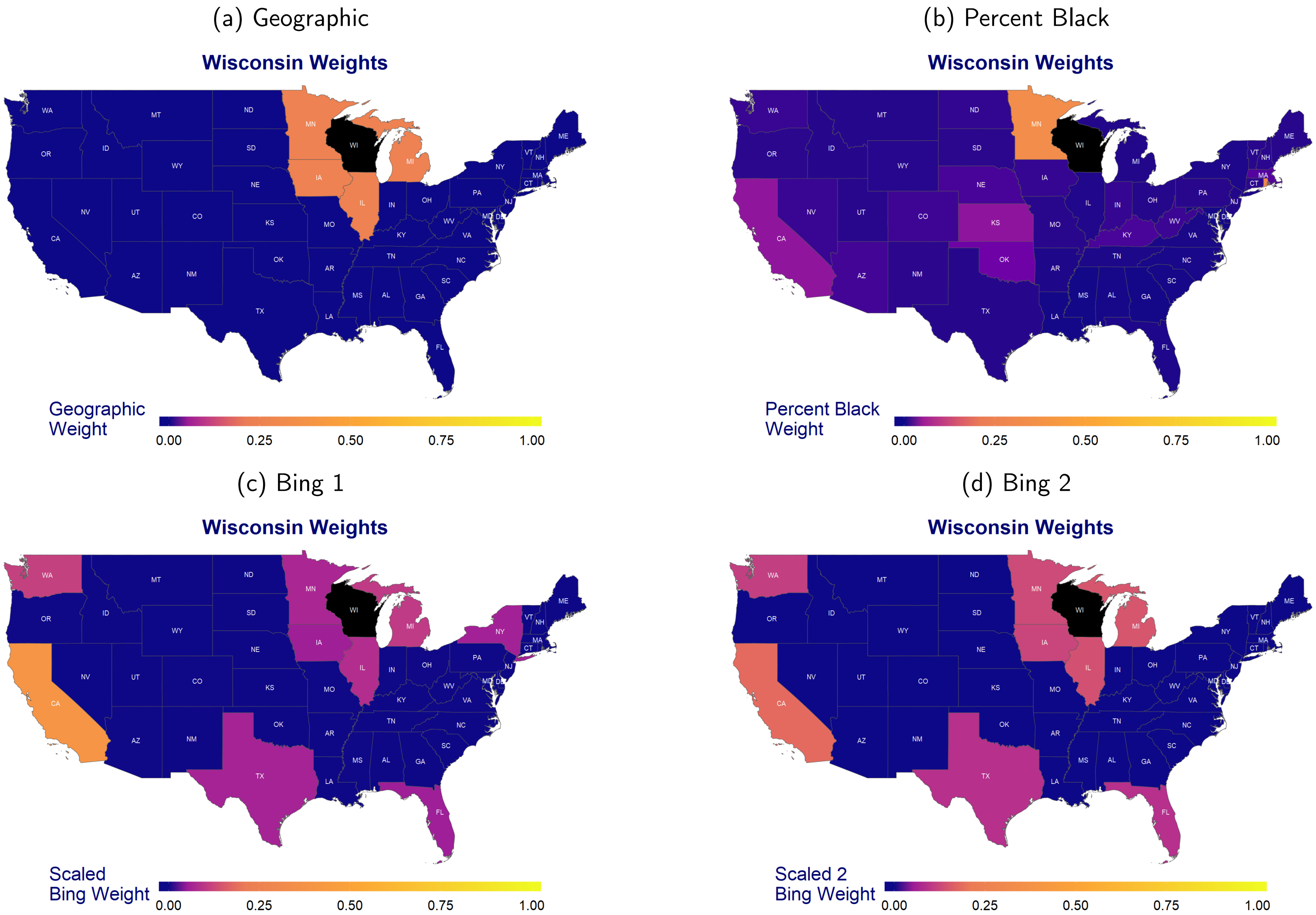
(a) Geographic

(b) Percent Black

(c) Bing 1

(d) Bing 2

Figure C.7: Wisconsin Maps



(a) Geographic

(b) Percent Black

(c) Bing 1

(d) Bing 2

C.2. Tables

Table C.1: Regional vs Local Paper State Trends

NYT Regional Google Trends Concentration of 99509

Region	Trend
Vermont	100.00
New York	95.00
District of Columbia	91.00
Connecticut	76.00
Maine	76.00
Massachusetts	71.00
Rhode Island	59.00
New Jersey	55.00
Maryland	54.00
Washington	50.00
New Hampshire	48.00
Montana	46.00
Pennsylvania	44.00
Oregon	44.00
Minnesota	42.00
Colorado	42.00
Delaware	42.00
New Mexico	41.00
Alaska	41.00
Hawaii	39.00
California	39.00
Wisconsin	37.00
Virginia	37.00
Illinois	37.00
Wyoming	35.00
Iowa	33.00
Michigan	32.00
Ohio	32.00
Utah	31.00
Missouri	30.00
North Carolina	30.00
Florida	29.00
Idaho	29.00
Arizona	29.00
Indiana	29.00
Nebraska	28.00
Kansas	26.00
Kentucky	26.00
Georgia	25.00
South Carolina	25.00
Louisiana	25.00
Tennessee	25.00
North Dakota	24.00
Texas	24.00
South Dakota	24.00
West Virginia	23.00
Alabama	23.00
Nevada	22.00
Oklahoma	20.00
Arkansas	20.00
Mississippi	20.00

Times-Union Google Trend Concentration of 11648

Region	Trend
Florida	100.00
Georgia	26.00
New York	25.00
Massachusetts	12.00
Virginia	11.00
Illinois	8.00
California	3.00
Texas	3.00
Vermont	0.00
Indiana	0.00
District of Columbia	0.00
North Carolina	0.00
Delaware	0.00
South Carolina	0.00
Wyoming	0.00
New Jersey	0.00
South Dakota	0.00
Ohio	0.00
Hawaii	0.00
New Hampshire	0.00
Arkansas	0.00
West Virginia	0.00
Tennessee	0.00
Connecticut	0.00
Rhode Island	0.00
North Dakota	0.00
Alaska	0.00
Mississippi	0.00
Maryland	0.00
Wisconsin	0.00
Pennsylvania	0.00
Missouri	0.00
Minnesota	0.00
Montana	0.00
Kentucky	0.00
Maine	0.00
Alabama	0.00
Arizona	0.00
Nebraska	0.00
Louisiana	0.00
Colorado	0.00
Nevada	0.00
Utah	0.00
Washington	0.00
Oregon	0.00
Oklahoma	0.00
Idaho	0.00
Michigan	0.00
Kansas	0.00
Iowa	0.00
New Mexico	0.00

Table C.2

Total Unscaled Weights

state	Total Weight
Illinois	40.07
Oregon	29.77
Colorado	26.97
California	24.83
Minnesota	23.24
Kentucky	22.42
Maryland	21.82
Connecticut	17.91
New Jersey	17.16
Ohio	12.40
Michigan	11.50
Rhode Island	9.26
Utah	8.66
Iowa	7.77
Vermont	6.93
Maine	6.74
Florida	6.63
Pennsylvania	6.47
Missouri	6.10
Texas	4.84
Indiana	4.77
Massachusetts	4.75
Wisconsin	4.38
Nevada	3.68

state	Total Weight
North Carolina	3.13
Alabama	3.12
New Hampshire	3.04
Washington	2.83
Arkansas	2.78
Louisiana	2.76
Mississippi	2.74
Tennessee	2.42
South Dakota	2.39
New York	2.14
New Mexico	1.98
Virginia	1.71
Arizona	1.59
West Virginia	1.51
South Carolina	1.36
Delaware	1.35
Oklahoma	1.33
Idaho	0.87
Nebraska	0.85
North Dakota	0.68
Kansas	0.66
Wyoming	0.18
Montana	0.10
Georgia	0.08

Lagged Regression Tables

Table C.3

Regression Type	IV	Spending Type	Weight	Lag	Estimate	SE	P Value	Percent Randomly Significant	Percent Permuted Significant
OLS	NA	Total	Percent Black	1	0.043	0.04	0.24	0.18	0.08
OLS	NA	Total	Geographic	1	-0.184	0.20	0.37	0.18	0.09
OLS	NA	Total	Scaled bing	1	-0.694**	0.22	0.00	0.18	0.12
OLS	NA	Total	Scaled Bing 2	1	-0.631*	0.31	0.04	0.18	0.12
IV	Medcaid	Total	Percent Black	1	-0.031	0.11	0.79	0.05	0.04
IV	Medcaid	Total	Geographic	1	1.496	0.93	0.12	0.05	0.00
IV	Medcaid	Total	Scaled bing	1	-1.209	1.05	0.26	0.05	0.00
IV	Medcaid	Total	Scaled Bing 2	1	3.862	10.89	0.72	0.05	0.01
IV	Neighbor's Controls	Total	Percent Black	1	0.071	0.05	0.19	0.06	0.04
IV	Neighbor's Controls	Total	Geographic	1	0.025	0.41	0.95	0.06	0.08
IV	Neighbor's Controls	Total	Scaled bing	1	-0.459	0.38	0.23	0.06	0.10
IV	Neighbor's Controls	Total	Scaled Bing 2	1	-0.447	0.48	0.36	0.06	0.09
OLS	NA	Total	Percent Black	2	0.002	0.04	0.95	0.11	0.06
OLS	NA	Total	Geographic	2	-0.186	0.21	0.37	0.11	0.08
OLS	NA	Total	Scaled bing	2	-0.74***	0.20	0.00	0.11	0.12
OLS	NA	Total	Scaled Bing 2	2	-0.708*	0.28	0.02	0.11	0.10
IV	Medcaid	Total	Percent Black	2	0.011	0.11	0.92	0.04	0.04
IV	Medcaid	Total	Geographic	2	1.898	1.09	0.09	0.04	0.00
IV	Medcaid	Total	Scaled bing	2	-2.538	3.25	0.44	0.04	0.00
IV	Medcaid	Total	Scaled Bing 2	2	-8.488	26.97	0.75	0.04	0.01
IV	Neighbor's Controls	Total	Percent Black	2	0.015	0.05	0.75	0.07	0.06
IV	Neighbor's Controls	Total	Geographic	2	0.005	0.39	0.99	0.07	0.08
IV	Neighbor's Controls	Total	Scaled bing	2	0.054	0.67	0.94	0.07	0.11
IV	Neighbor's Controls	Total	Scaled Bing 2	2	-0.349	0.53	0.52	0.07	0.10

Table C.4

Regression Type	IV	Spending Type	Weight	Lag	Estimate	SE	P Value	Percent Randomly Significant	Percent Permuted Significant
OLS	NA	Total	Percent Black	3	0.002	0.03	0.95	0.07	0.06
OLS	NA	Total	Geographic	3	-0.166	0.21	0.44	0.07	0.08
OLS	NA	Total	Scaled bing	3	-0.637**	0.20	0.00	0.07	0.10
OLS	NA	Total	Scaled Bing 2	3	-0.641*	0.29	0.03	0.07	0.07
IV	Medcaid	Total	Percent Black	3	-0.145	0.11	0.21	0.04	0.03
IV	Medcaid	Total	Geographic	3	2.066	1.19	0.09	0.04	0.00
IV	Medcaid	Total	Scaled bing	3	29.698	410.94	0.94	0.04	0.00
IV	Medcaid	Total	Scaled Bing 2	3	-3.095	5.16	0.55	0.04	0.00
IV	Neighbor's Controls	Total	Percent Black	3	-0.047	0.05	0.34	0.07	0.06
IV	Neighbor's Controls	Total	Geographic	3	-0.038	0.37	0.92	0.07	0.07
IV	Neighbor's Controls	Total	Scaled bing	3	0.591	0.87	0.50	0.07	0.12
IV	Neighbor's Controls	Total	Scaled Bing 2	3	-0.391	0.51	0.44	0.07	0.09

C.3. Interactive Data Appendix

The Interactive Data appendix can be downloaded here:
https://drive.google.com/file/d/1K_1LjlnfFTPsQmBuxPU_0eUU1WhkMUuK/view?usp=sharing
Google drive does not preview HTML documents so you will need to download it and then open it in a web browser.

BIBLIOGRAPHY

BIBLIOGRAPHY

(2022), The economy remains the top issue for voters in the midterms, https://www.pewresearch.org/fact-tank/2022/11/03/key-facts-about-u-s-voter-priorities-ahead-of-the-2022-midterm-elections/ft_2022-11-03_election-roundup_01/, accessed: 2023-04-04.

Abdulkadiroğlu, A., P. A. Pathak, J. Schellenberg, and C. R. Walters (2020), Do parents value school effectiveness?, *American Economic Review*, *110*(5), 1502–39.

Abrams, D. S., and A. H. Yoon (2007), The luck of the draw: Using random case assignment to investigate attorney ability, *University of Chicago Law Review*, *74*, 1145.

Agarwal, S., M. Vyacheslav, and B. Scholnick (2016), Does Inequality Cause Financial Distress? Evidence from Lottery Winners and Neighboring Bankruptcies, *FRB of Philadelphia Working Paper*, *No. 16-4*.

Agrawal, D., W. Hoyt, and T. Ly (2023), A new approach to evaluating the welfare effects of decentralized policies, *Working paper*.

Agrawal, D. R., W. H. Hoyt, and J. D. Wilson (2022), Local policy choice: theory and empirics, *Journal of Economic Literature*, *60*(4), 1378–1455.

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

Angrist, J., P. Hull, and C. R. Walters (2022), Methods for measuring school effectiveness.

Arrow, K. J. (1950), A difficulty in the concept of social welfare, *Journal of political economy*, *58*(4), 328–346.

Arrow, K. J. (1978), Extended sympathy and the possibility of social choice, *Philosophia*, *7*(2), 223–237.

Arrow, K. J. (2012), *Social choice and individual values*, vol. 12, Yale university press.

Athey, S., and S. Wager (2021), Policy learning with observational data, *Econometrica*, *89*(1), 133–161.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (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.

Atwal, G., and A. Williams (2017), Luxury brand marketing–the experience is everything!, *Advances in luxury brand management*, pp. 43–57.

Auspitz, R., and R. Lieben (1889), *Untersuchungen über die Theorie des Preises*, Duncker & Humblot.

Aydede, M. (2019), Pain, in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta, Spring 2019 ed., Metaphysics Research Lab, Stanford University.

Bacher-Hicks, A., and C. Koedel (2022), Estimation and interpretation of teacher value added in research applications, *Working paper*.

Baicker, K. (2005), The spillover effects of state spending, *Journal of Public Economics*, 10.1016/j.jpubeco.2003.11.003.

Baker, H. K., and V. Ricciardi (2014), *Investor behavior: The psychology of financial planning and investing*, John Wiley & Sons.

Bartley, E. J., and R. B. Fillingim (2013), Sex differences in pain: a brief review of clinical and experimental findings, *British journal of anaesthesia*, *111*(1), 52–58.

Baskaran, T. (2014), Identifying local tax mimicking with administrative borders and a policy reform, *Journal of Public Economics*, *118*, 41–51.

Bates, M. D., M. Dinerstein, A. C. Johnston, and I. Sorkin (2022), Teacher labor market equilibrium and student achievement, *Tech. rep.*, National Bureau of Economic Research.

Becker, G. M., M. H. DeGroot, and J. Marschak (1963), Stochastic models of choice behavior, *Behavioral science*, *8*(1), 41–55.

Becker, G. M., M. H. DeGroot, and J. Marschak (1964), Measuring utility by a single-response sequential method, *Behavioral science*, *9*(3), 226–232.

Berkouwer, S. B., and J. T. Dean (2021), Credit, attention, and externalities in the adoption of energy efficient technologies by low-income households, *Unpublished manuscript, University of Pennsylvania, Philadelphia, PA*.

Besley, T., and A. Case (1995), Incumbent Behavior: Vote-Seeking, Tax-Setting, and Yardstick Competition, *Tech. Rep. 1*.

Besley, T., and M. Smart (2002), Does tax competition raise voter welfare?, *Discussion Paper Series-Centre for Economic Policy Research London*, (3131), 1–21.

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., C. K. Jackson, L. Navarro-Sola, and F. Pardo (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., S. B. Heller, M. Kapustin, M. Bertrand, and C. Blattman (2023), Predicting and preventing gun violence: An experimental evaluation of readi chicago, *Tech. rep.*, National Bureau of Economic Research.

Bishop, R. C., and T. A. Heberlein (1979), Measuring values of extramarket goods: Are indirect measures biased?, *American journal of agricultural economics*, *61*(5), 926–930.

Boyd, D., H. Lankford, S. Loeb, and J. Wyckoff (2005a), The draw of home: How teachers' preferences for proximity disadvantage urban schools, *Journal of Policy Analysis And Management*, *24*(1), 113–132.

Boyle, K. J. (2017), Contingent valuation in practice, in *A primer on nonmarket valuation*, pp. 83–131, Springer.

Brueckner, J. K. (2000), Welfare reform and the race to the bottom: Theory and evidence, *Southern Economic Journal*, *66*(3), 505–525.

Brueckner, J. K. (2003), Strategic interaction among governments: An overview of empirical studies, *International regional science review*, *26*(2), 175–188.

Bureau of the Census ((2004-2017)a), American community survey 1-year estimates [data file], generated by Nathan Mather; using American FactFinder.

Bureau of the Census ((2004-2017)b), State and local government finances datasets and tables [data file].

Case, A. C., H. S. Rosen, and J. R. Hines (1993), Budget spillovers and fiscal policy interdependence. Evidence from the states, *Journal of Public Economics*, 10.1016/0047-2727(93) 90036-S.

Chan, D. C., M. Gentzkow, and C. Yu (2022), Selection with variation in diagnostic skill: Evidence from radiologists, *The Quarterly Journal of Economics*, *137*(2), 729–783.

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2016), Health care exceptionalism? performance and allocation in the us health care sector, *American Economic Review*, *106*(8), 2110–2144.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a), Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates, *American Economic Review*, *104*(9), 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b), Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood, *American Economic Review*, *104*(9), 2633–79.

Chetty, R., J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter (2018), The opportunity atlas, *Opportunity Insights*.

Coate, S. (2000), An efficiency approach to the evaluation of policy changes, *The Economic Journal*, *110*(463), 437–455.

Condie, S., L. Lefgren, and D. Sims (2014), Teacher heterogeneity, value-added and education policy, *Economics of Education Review*, *40*, 76–92.

Conley, T. G. (1999), GMM estimation with cross sectional dependence, *Tech. rep.*

Courant, P., E. Gramlich, and D. Rubinfeld (1978), The stimulative effects of intergovernmental grants: Or why money sticks where it hits, *Fiscal Federalism and Grants-in-Aid*.

Dahlstrand, A. (2022), Defying distance? the provision of services in the digital age, *Tech. rep.*

Davidson, D., P. Suppes, and S. Siegel (1957), Decision making; an experimental approach.

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.*

DellaVigna, S., and W. Kim (2022), Policy diffusion and polarization across us states, *Tech. rep.*, National Bureau of Economic Research.

DeNegre, S. T., and T. K. Ralphs (2009), A branch-and-cut algorithm for integer bilevel linear programs, in *Operations research and cyber-infrastructure*, pp. 65–78, Springer.

Devereux, M. P., and S. Loretz (2013), What do we know about corporate tax competition?, *National Tax Journal*, *66*(3), 745–773.

Diamond, P. A. (1973), Consumption externalities and imperfect corrective pricing, *The Bell Journal of Economics and Management Science*, pp. 526–538.

Diener, E., and R. Biswas-Diener (2002), Will money increase subjective well-being?, *Social indicators research*, *57*, 119–169.

Diener, E., E. Sandvik, L. Seidlitz, and M. Diener (1993), The relationship between income and subjective well-being: Relative or absolute?, *Social indicators research*, *28*, 195–223.

Diener, E., D. Wirtz, W. Tov, C. Kim-Prieto, D.-w. Choi, S. Oishi, and R. Biswas-Diener (2010), New well-being measures: Short scales to assess flourishing and positive and negative feelings, *Social indicators research*, *97*, 143–156.

Diener, E., L. Tay, and S. Oishi (2013), Rising income and the subjective well-being of nations., *Journal of personality and social psychology*, *104*(2), 267.

Diener, E., R. E. Lucas, and S. Oishi (2018a), Advances and open questions in the science of subjective well-being, *Collabra: Psychology*, *4*(1).

Diener, E., S. Oishi, and L. Tay (2018b), *Handbook of well-being*, Noba Scholar.

Dimand, R. (2019), Irving fisher, ragnar frisch and the elusive quest for measurable utility.

Dolbear, F. T. (1963), Individual choice under uncertainty-an experimental-study, *Yale economic essays*, *3*(2), 418–469.

Doyle, J., J. Graves, and J. Gruber (2019), Evaluating measures of hospital quality: Evidence from ambulance referral patterns, *Review of Economics and Statistics*, *101*(5), 841–852.

Driver, J. (2014), The history of utilitarianism.

Duflo, E., P. Dupas, and M. Kremer (2011), Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya, *American economic review*, *101*(5), 1739–1774.

Dunn, E., and M. Norton (2014), *Happy money: The science of happier spending*, Simon and Schuster.

Dunn, E., L. Aknin, and M. Norton (2014), " prosocial spending and happiness: Using money to benefit others pays off": Corrigendum.

Eastmond, T., N. Mather, M. D. Ricks, and J. Betts (2022), Effect heterogeneity and optimal policy: Getting welfare added from teacher value added.

Einav, L., A. Finkelstein, and N. Mahoney (2022), Producing health: Measuring value added of nursing homes, *Tech. rep.*, National Bureau of Economic Research.

Fell, H., D. T. Kaffine, and K. Novan (2021), Emissions, transmission, and the environmental value of renewable energy, *American Economic Journal: Economic Policy*, *13*(2), 241–72.

Finkelstein, A., and N. Hendren (2020), Welfare analysis meets causal inference, *Journal of Economic Perspectives*, *34*(4), 146–67, 10.1257/jep.34.4.146.

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

Fisher, I. (1927), *A statistical method for measuring" marginal utility" and testing the justice of a progressive income tax*.

Flood, A. (), Terry pratchett estate backs jack monroe's idea for 'vimes boots' poverty index, *The Guardian*.

Frey, W. H. (2017), U.s. migration still at historically low levels, census shows.

Furnham, A., and M. Argyle (1998), *The psychology of money*, Psychology Press.

Gilovich, T., A. Kumar, and L. Jampol (2015), A wonderful life: Experiential consumption and the pursuit of happiness, *Journal of Consumer Psychology*, *25*(1), 152–165.

Glazerman, S., A. Protik, B.-r. Teh, J. Bruch, and J. Max (2013), Transfer incentives for high-performing teachers: Final results from a multi-site randomized experiment, *Tech. rep.*, U.S. Department of Education.

Glick, M. (2018), The unsound theory behind the consumer (and total) welfare goal in antitrust, *The Antitrust Bulletin*, *63*(4), 455–493.

Gossen, H. H. (1854), The laws of human relations: and the rules of human action derived therefrom, translated by rc blitz.

Gresik, T. A. (2001), The taxing task of taxing transnationals, *Journal of Economic Literature*, *39*(3), 800–838.

Grether, D. M., and C. R. Plott (1979), Economic theory of choice and the preference reversal phenomenon, *The American Economic Review*, *69*(4), 623–638.

Griffith, R., M. O'Connell, and K. Smith (2019), Tax design in the alcohol market, *Journal of Public Economics*, *172*, 20–35.

Hanemann, W. M. (1984), Welfare evaluations in contingent valuation experiments with discrete responses, *American journal of agricultural economics*, *66*(3), 332–341.

Hanushek, E. A. (2011), The economic value of higher teacher quality, *Economics of Education review*, *30*(3), 466–479.

Hanushek, E. A., et al. (2009), Teacher deselection, *Creating a new teaching profession*, *168*, 172–173.

Hare, R. M. (1981), *Moral thinking: Its levels, method, and point*, Oxford: Clarendon Press; New York: Oxford University Press.

Harrington, E., and H. Shaffer (2023), Estimating prosecutor effects on incarceration and reoffense, *Tech. rep.*, Working Paper.

Harsanyi, J. C. (1953), Cardinal utility in welfare economics and in the theory of risk-taking, *Journal of Political Economy*, *61*(5), 434–435.

Harsanyi, J. C. (1955), Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility, *Journal of political economy*, *63*(4), 309–321.

Harsanyi, J. C. (1986), *Rational behaviour and bargaining equilibrium in games and social situations*, CUP Archive.

Haushofer, J., J. Reisinger, and J. Shapiro (2015), Your gain is my pain: Negative psychological externalities of cash transfers, *Online paper, retrieved on May*, *13*, 2016.

Hausman, D. M. (1995), The impossibility of interpersonal utility comparisons, *Mind*, *104*(415), 473–490.

Hausman, D. M., and M. S. McPherson (2006), *Economic Analysis, Moral Philosophy and Public Policy.*, vol. 2nd ed, Cambridge University Press.

Hendren, N. (2020), Measuring economic efficiency using inverse-optimum weights, *Journal of Public Economics*, *187*, 104,198, https://doi.org/10.1016/j.jpubeco.2020.104198.

Hendren, N., and B. Sprung-Keyser (2020), A unified welfare analysis of government policies, *135*(3), 1209–1318, 10.1257/jep.34.4.146.

Hicks, J. R. (1940), The valuation of the social income, *Economica*, *7*(26), 105–124.

Hollingsworth, A., and I. Rudik (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., N. Rigol, and B. N. Roth (2022), Targeting high ability entrepreneurs using community information: Mechanism design in the field, *American Economic Review*, *112*(3), 861–98.

Ida, T., T. Ishihara, K. Ito, D. Kido, T. Kitagawa, S. Sakaguchi, and S. Sasaki (2022), Choosing who chooses: Selection-driven targeting in energy rebate programs, *Tech. rep.*, National Bureau of Economic Research.

Imberman, S. A., and M. F. Lovenheim (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., T. Ida, and M. Tanaka (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., and L. Lefgren (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.

Jebb, A. T., L. Tay, E. Diener, and S. Oishi (2018), Happiness, income satiation and turning points around the world, *Nature Human Behaviour*, *2*(1), 33–38.

Johnson, A. C. (2021), Preferences, selection, and the structure of teacher pay, *Working paper*.

Kahneman, D. (1994), New challenges to the rationality assumption, *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, pp. 18–36.

Kahneman, D., and A. Tversky (1979), Prospect Theory: An Analysis of Decision under Risk, *Tech. Rep. 2*.

Kahneman, D., and A. Tversky (2013), Prospect theory: An analysis of decision under risk, in *Handbook of the fundamentals of financial decision making: Part I*, pp. 99–127, World Scientific.

Kahneman, D., J. L. Knetsch, and R. H. Thaler (1991), Anomalies: The endowment effect, loss aversion, and status quo bias, *Journal of Economic perspectives*, *5*(1), 193–206.

Kahneman, D., E. Diener, and N. Schwarz (1999), *Well-being: Foundations of hedonic psychology*, Russell Sage Foundation.

Kaldor, N. (1939), Welfare propositions of economics and interpersonal comparisons of utility, *The Economic Journal*, *49*(195), 549–552.

Kapferer, J.-N. (2014), The future of luxury: Challenges and opportunities, *Journal of Brand Management*, *21*(9), 716–726.

Karmarkar, U. S. (1974), The effect of probabilities on the subjective evaluation of lotteries.

Keen, M., and K. A. Konrad (2013), The theory of international tax competition and coordination, *Handbook of public economics*, *5*, 257–328.

Kimball, M. S., F. Ohtake, D. Reck, Y. Tsutsui, and F. Zhang (2015), Diminishing marginal utility revisited, *Available at SSRN 2592935*.

Kitagawa, T., and A. Tetenov (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.

Lautenbacher, S., J. H. Peters, M. Heesen, J. Scheel, and M. Kunz (2017), Age changes in pain perception: a systematic-review and meta-analysis of age effects on pain and tolerance thresholds, *Neuroscience & Biobehavioral Reviews*, *75*, 104–113.

Layard, R., G. Mayraz, and S. Nickell (2008), The marginal utility of income, *Journal of Public Economics*, *92*(8-9), 1846–1857.

Ledyard, J. O. (1984), The pure theory of large two-candidate elections, *Public Choice*, *44*(1), 7–41.

Lucas, R. E., and U. Schimmack (2009), Income and well-being: How big is the gap between the rich and the poor?, *Journal of Research in Personality*, *43*(1), 75–78.

Luttmer, E. F. (2005), Neighbors as negatives: Relative earnings and well-being, *The Quarterly journal of economics*, *120*(3), 963–1002.

Lyytikäinen, T. (2012), Tax competition among local governments: Evidence from a property tax reform in finland, *Journal of Public Economics*, *96*(7-8), 584–595.

MacCrimmon, K. R. (1965), *An experimental study of the decision making behavior of business executives*, University of California at Los Angeles.

MacCrimmon, K. R., and S. Larsson (1979), Utility theory: Axioms versus 'paradoxes', *Expected Utility Hypotheses and the Allais Paradox: Contemporary Discussions of the Decisions under Uncertainty with Allais' Rejoinder*, pp. 333–409.

Machina, M. J. (1982), " expected utility" analysis without the independence axiom, *Econometrica: Journal of the Econometric Society*, pp. 277–323.

MacKay, A. F. (1986), Extended sympathy and interpersonal utility comparisons, *The Journal of philosophy*, *83*(6), 305–322.

Marshall, A. (1890), *The Principles of Economics*, McMaster University Archive for the History of Economic Thought.

Miljković, A., A. Stipčić, M. Braš, V. Đorđević, L. Brajković, C. Hayward, A. Pavić, I. Kolčić, and O. Polašek (2014), Is experimentally induced pain associated with socioeconomic status? do poor people hurt more?, *Medical science monitor: international medical journal of experimental and clinical research*, *20*, 1232.

Moretti, E., and D. J. Wilson (2019), Taxing Billionaires: Estate Taxes and the Geographical Location of the Ultra-Wealthy, 10.24148/wp2019-25.

Morgan, J. N. (1945), Can we measure the marginal utility of money?, *Econometrica: Journal of the Econometric Society*, pp. 129–152.

Moscati, I. (2018), *Measuring utility: From the marginal revolution to behavioral economics*, Oxford Studies in History of E.

Mosteller, F., and P. Nogee (1951), An experimental measurement of utility, *Journal of Political Economy*, *59*(5), 371–404.

National Association of State Budget Officers ((2004-2017)), 1991-2019 state expenditure report data.

Ng, W. (2013), The duality of wealth: Is material wealth good or bad for well-being?, *Journal of Social Research & Policy*, *4*(2), 7.

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

Oishi, S., and E. Diener (2001), Goals, culture, and subjective well-being, *Personality and Social Psychology Bulletin*, *27*(12), 1674–1682.

Pearce, D. (2021), Individual and social welfare: A bayesian perspective.

Petek, N., and N. G. Pope (forthcoming), The multidimensional impact of teachers on students, *Journal of Political Economy*.

Revelli, F. (2005), On spatial public finance empirics, *International Tax and Public Finance*, *12*, 475–492.

Revelli, F. (2006), Performance rating and yardstick competition in social service provision, *Journal of Public Economics*, *90*(3), 459–475, https://doi.org/10.1016/j.jpubeco.2005.07.006, special issue published in cooperation with the National Bureau of Economic Research: Proceedings of the Trans-Atlantic Public Economics Seminar on Fiscal Federalism 20–22 May 2004.

Revelli, F., et al. (2006), Spatial interactions among governments, *Handbook of fiscal federalism*, pp. 106–130.

Ricks, M. D. (2022), Strategic selection around kindergarten recommendations, *Tech. rep.*

Risse, M. (2002), Harsanyi's' utilitarian theorem'and utilitarianism, *Noûs*, *36*(4), 550–577.

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.

Ruscheweyh, R., M. Marziniak, F. Stumpenhorst, J. Reinholz, and S. Knecht (2009), Pain sensitivity can be assessed by self-rating: Development and validation of the pain sensitivity questionnaire, *Pain*, *146*(1-2), 65–74.

Ruscheweyh, R., et al. (2012), Validation of the pain sensitivity questionnaire in chronic pain patients, *Pain*, *153*(6), 1210–1218.

Samuelson, P. A. (1950), Evaluation of real national income, *Oxford Economic Papers*, *2*(1), 1–29.

Samuelson, P. A., and S. Swamy (1974), Invariant economic index numbers and canonical duality: survey and synthesis, *The American Economic Review*, *64*(4), 566–593.

Sen, A. (1976), Welfare inequalities and rawlsian axiomatics, *Theory and decision*, *7*(4), 243–262.

Sen, A. K. (1977), Rational fools: A critique of the behavioral foundations of economic theory, *Philosophy & Public Affairs*, pp. 317–344.

Sexton, S., A. J. Kirkpatrick, R. I. Harris, and N. Z. Muller (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.

Slovic, P., and S. Lichtenstein (1968), Relative importance of probabilities and payoffs in risk taking., *Journal of experimental psychology*, *78*(3p2), 1.

Staiger, D. O., and J. E. Rockoff (2010), Searching for effective teachers with imperfect information, *Journal of Economic perspectives*, *24*(3), 97–118.

Stephens, R., and O. Robertson (2020), Swearing as a response to pain: Assessing hypoalgesic effects of novel "swear" words, *Frontiers in Psychology*, *11*, 10.3389/fpsyg.2020.00723.

Stepner, M. (2013), VAM: Stata module to compute teacher value-added measures, Statistical Software Components, Boston College Department of Economics.

Stevenson, B., and J. Wolfers (2013), Subjective well-being and income: Is there any evidence of satiation?, *American Economic Review*, *103*(3), 598–604.

Stevenson, B., and J. Wolfers (2019), *Principles of macroeconomics*, Macmillan Higher Education.

Tiebout, C. M. (1956), A Pure Theory of Local Expenditures, *Tech. Rep. 5*.

Tomlinson, C. A. (2017), *How to differentiate instruction in academically diverse classrooms*, third ed., ASCD.

Topkis, D. M. (1978), Minimizing a Submodular Function on a Lattice, *Operations Research*, *26*(2), 305–321, 10.1287/opre.26.2.305.

Tversky, A. (1969), Intransitivity of preferences., *Psychological review*, *76*(1), 31.

Unspl ((2017)), United states newspaper listing.

U.S. Centers for Medicare and Medicaid ((1960-2018)), The national health expenditure accounts: Nhe summary, including share of gdp, cy 1960-2018 [data file].

Veblen, T. (1899), The theory of the leisure class: an economic study in the evolution of institutions.

Vela, G. (2019), Council post: Six tips for marketing to the super-rich.

Von Neumann, J., and O. Morgenstern (1947), Theory of games and economic behavior, 2nd rev.

Weiman, J., A. Knabe, and R. Schob (2015), Money does buy happiness: Where happiness economics gets it wrong.

Weymark, J. A. (1991), A reconsideration of the harsanyi–sen debate on utilitarianism, *Interpersonal comparisons of well-being*, *255*.

Weymark, J. A. (2005), Measurement theory and the foundations of utilitarianism, *Social Choice and Welfare*, *25*, 527–555.

Wilson, C. S. (2019), Welfare standards underlying antitrust enforcement: What you measure is what you get, in *In: Luncheon Keynote Address at George Mason Law Review 22nd Antitrust Symposium. Arlington: Antitrust at the Crossroads*.

Wilson, J. D. (1999), Theories of tax competition, *National tax journal*, *52*(2), 269–304.

Wilson, J. D., and D. E. Wildasin (2004), Capital tax competition: bane or boon, *Journal of public economics*, *88*(6), 1065–1091.

Zwolinski, M. (2008), The ethics of price gouging, *Business Ethics Quarterly*, *18*(3), 347–378.