Essays in Applied Econometrics and Labor Economics

Leonard Goff

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

© 2021

Leonard Goff

All Rights Reserved

Abstract

Essays in Applied Econometrics and Labor Economics

Leonard Goff

Recent decades have seen great advances in the methods we use to understand cause and effect in the world of work. Building on that tradition, this dissertation explores two broad topics in econometrics as tools to address specific questions in labor economics. The main econometric contributions are to extend identification results for research designs based on bunching (Chapter 1) and those that make use of instrumental variables (Chapters 2 and 3). The empirical questions that compel them are described below.

Chapter 1 examines the effect of overtime regulation on hours of work in the United States, extending a recently popularized technique that uses bunching observed at kinks in agents' choice sets for identification. In the U.S., most workers are required to be paid one-and-a-half times their typical rate of pay for any hours in excess of forty within a week. While prominent and long-standing, this policy has not been meaningfully reformed since it was first established at the federal level in 1938. As a result, few studies have been able to leverage causal research designs to assess its labor market impacts. I use bunching in the distribution of weekly hours at forty—where the policy introduces a convex "kink" in firms' costs—to estimate this effect. To do so, I develop a framework in which bunching at a choice-set kink is informative about causal effects under substantially weaker assumptions than those maintained in existing work. This allows the effect of the overtime policy to be partially identified without making parametric assumptions about firms' objective functions, or about the distribution of hours they would set in the absence of the policy. Using an administrative dataset of weekly hours derived from payroll records, I find that the bounds are informative and that covered hourly workers in the U.S. work an average of at least half an hour less as a result, in affected weeks.

Chapter 2 turns to a still-more popular strategy in applied microeconomics: the instrumental variables research design. I propose a new method for estimating causal effects

when a researcher has more than one such instrument, and apply it to reassess the labor market returns to college education. The method is motivated by the following issue. When treatment effects are heterogeneous, it is known that instruments can be used to identify local average treatment effects under an assumption known as "monotonicity" (Imbens and Angrist, 1994). However, when a researcher wishes to use multiple instruments together, this assumption can become quite restrictive, and empirical conclusions may be misleading if it is violated. I propose an alternative assumption that I call "vector monotonicity", which is quite natural in typical settings with multiple instruments. I show that vector monotonicity leads to identification of a useful class of treatment effect parameters, but the two-stage-least-squares estimator popular in applied work does not consistently estimate them. I propose an alternative estimator, and apply it to the classic question of the returns to schooling. I find that the approach based upon vector monotonicity reveals new patterns of heterogeneity in the earnings effect of college education.

Chapter 3, with coauthors Ashna Arora and Jonas Hjort, considers the effects of a worker's first job on outcomes later in their career. This is typically a difficult question to answer empirically, as workers entering the labor force are not randomly assigned to employers. We make use of a unique opportunity to study this question in the context of medical residencies in Norway. For decades, medical school graduates in Norway were matched to residencies based on a random serial dictatorship mechanism, in which doctors could choose—in an order determined by lottery—among available positions in the country. We develop an econometric framework in which the random choice set a doctor is presented with provides a collection of instruments for their choice of residency hospital, and hence first job as a doctor. Because we only observe choices and not a doctor's full preferences, this requires new methods—related to those of Chapter 2. We find persistent effects of a doctor's first job on earnings, specializations, and mid-career moves. We use the estimates to assess the replacement of the serial-dictatorship by a decentralized labor market in 2013, which we find led to a small increase in resident welfare.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I'm meta-grateful to have this space to express my appreciation for those who have made this dissertation possible. Firstly, I thank my co-sponsors Simon Lee and Suresh Naidu. With them guiding me along, my PhD experience has been full of the excitement that comes with exploring ideas and the encouragement and grit to see them through. I don't take for granted that I came through even more interested than when I started. Bernard Salanié also played a central role advising me, and I appreciate his generosity in carefully analyzing the many things I put down on paper. I thank Daniel Hamermesh, whom I had the great fortune of having next-door at Barnard and always helped to keep me rooted in reality. I am grateful to Jonas Hjort for being an invaluable guide and collaborator in the final years of my PhD.

Beyond my committee I have benefited from many informal advisors: Michael Best for the many useful and encouraging discussions on bunching, Joshua Angrist for igniting my research interests in IV, and to Glen Weyl and Jaron Lanier for challening me to think towards the future. Staying in touch with John Helliwell and Allen Blackman has kept me concerned with things that matter. I am grateful to the many other faculty at Columbia who have given me feedback along the way, especially the econometrics and applied micro groups—always making me feel that I had a home in both. In particular, I want to thank Doug Almond, Jushan Bai, Sandra Black, Gregory Cox, François Gerard, Wojciech Kopczuk, Bentley MacLeod, Serena Ng, José Luis Montiel Olea, Christoph Rothe and Miguel Urquiola. I have also learned so much from the rest of my cohort at Columbia and my deep friendships within it, and from those in years ahead and behind. Junlong Feng and Eric Adebayo always gave me a preview of what to expect next year. I am grateful to the incredible administrative and support staff in Columbia economics department. I thank the provider of the payroll data I use in Chapter 1, as well as the Bureau of Labor Statistics, Statistics Norway, and the Research Council of Norway.

I am grateful to the people of New York City: a place that I was wary to move to but grew so fond of, and has been a constant source of inspiration in studying economics. Thanks to my roommates Iain Bamford and Dilip Ravindran for being inspired with me. I am grateful to all those who I have become close with and loved here, and cherish the relationships that have deepened through phone calls and visits through these past six years. Above all I thank my parents and Tim. I was so fortunate to grow up in a household that embraced both curiosity and adventure, things that have since shaped what I value in life.

Chapter 1: Treatment Effects in Bunching Designs: The Impact of the Federal Overtime Rule on Hours

1.1 Introduction

Many countries require premium pay for long work hours. In the U.S., this takes the form of the "time-and-a-half" rule of the Fair Labor Standards Act (FLSA): workers must be paid one and a half times their normal hourly wage for any hours they work in excess of 40 within a week. While some workers are exempt from the overtime rule (and some of their employers are not covered by the FLSA), the time-and-a-half rule applies to a majority of the U.S. workforce, including nearly all of its 82 million hourly workers (U.S. Department of Labor, 2019). Given the prevalence of long workweeks in the U.S., the total number of hours paid at the overtime rate is substantial. Workers in many industries average several overtime hours per week, making overtime the largest form of supplemental pay in the U.S. (Bishow, 2009).[1]

Nevertheless, only a small literature has addressed the effects of the federal overtime rule on the U.S. labor market. This stands in marked contrast to the large literature on the minimum wage, which was also introduced at the federal level by the FLSA in 1938. A likely reason for this gap is that the overtime rule has varied little since then: the basic parameters have remained throughout as time-and-a-half after 40 hours within a week.[2] This lack of variation has afforded few opportunities to leverage research designs that

[1] Hart (2004) reports an average of 3 overtime hours per week among non-supervisory production workers. See Table A.2 for new estimates by industry from my sample. From a separate representative survey I estimate in Section 1.3 a grand average of about one overtime hour per week per worker, among all employed.

[2] While there are supplemental state overtime rules that vary somewhat by state (e.g. Minnesota has a 48 hour threshold), these rules bind for relatively few workers since the federal rules supersede the state rules.

exploit policy changes over time to identify causal effects, particularly on hours worked.[3] Unlike with the minimum wage, reforms to overtime policy have been rare and have left the central parameters of the rule unchanged.

In this paper, I take a new approach to assessing the effect of the FLSA overtime rule on hours by making use of variation within the rule itself: given a fixed hourly wage, hours in excess of 40 within a week for a single worker are more expensive to the firm than those below 40. Rather than attempt to explicitly control for confounding factors affecting hours or exploit reforms to whom is covered by the rule, I leverage the sharp discontinuity in the marginal cost of a worker-hour at 40 for identification. This methodology requires two new ingredients that this paper adds to the existing literature: first, high resolution data on the hours of individual workers within a single given week, allowing me to observe the distribution of hours close to 40. I obtain this from a novel dataset of paycheck records from a large payroll processing company that records the exact number of hours that a worker was paid for in a given week. Second, my approach requires a way to translate the observed hours distribution near 40 into credible causal estimates of the rule's effect, given reasonable assumptions about how weekly work hours are determined.

While wages change only occasionally in the data, hours are quite variable among my sample of hourly workers, suggesting that hours are set dynamically each week with the overtime rule generating a convex kink in pay as a function of hours. A well-known prediction is that if firms set hours (at least for a subset of workers), there will be a mass of observations located exactly at the kink at 40 hours. I show that the size of this mass of paychecks is informative about the joint distribution of two *counterfactual* choices: the number of hours the firm would choose for the worker if the worker's normal wage rate

[3] A notable exception is Hamermesh and Trejo (2000), who apply a difference-in-differences approach over the expansion of a daily overtime rule in California to include men in 1980, estimating a price elasticity of demand for overtime hours of roughly -0.5. Costa (2000) and Johnson (2003) also consider the impact of federal overtime regulation on hours worked, studying the phase-in of the FLSA and a supreme court decision clarifying the eligibility of public sector workers, respectively. Quach (2020) looks at recent reforms to eligibility criteria for exemption from the FLSA, estimating effects of the change on employment and the incomes of salaried workers.

applied to all hours, and the hours that the firm would choose if all were paid at the worker's overtime rate. This generalizes a popular research design that has used bunching at kink points to identify elasticities, which I refer to as the "bunching design".[4] The bunching design originated in public finance to assess the labor supply effects of taxation (Saez 2010; Chetty et al. 2011), but variations have since been applied in many other settings.[5] In my context, the bunching design uncovers the effect on hours of the wage variation induced by the FLSA overtime rule, providing an estimate of its reduced form causal effect.

One of the main contributions of this paper is thus to extend and reinterpret the kink bunching-design methodology, which has gained popularity with the increasing availability of administrative data—and the ubiquity of policy thresholds at which incentives change discontinuously. Here I make four main contributions on the econometrics of the bunching design. First, while bunching designs are typically motivated by a choice model featuring an explicit functional form for decision-makers' utility, I require only *convexity*, both of the kink itself and agents' possibly heterogeneous preferences. Secondly, I show that the bunching design can allow for multiple (possibly unknown) underlying margins of choice, yielding a single outcome variable observed to the researcher. Inference about counterfactual choices is thus robust to a large class of choice models, though this robustness can make it difficult to isolate a single structural interpretation of the estimates.[6] This is turn makes a potential outcomes framework a natural language for analyzing the bunching design. Third, I propose a way to confront a challenge to identification in the bunching design leveled by Blomquist and Newey (2017)—that it requires extrapolation of observed densities into a region where they are not. To perform this extrapolation I

[4]This paper considers only the bunching design for kinks, and not the related method for bunching at *notches* (e.g. Kleven and Waseem 2013b). Bunching can also be used to overcome endogeneity in settings where the variable exhibiting bunching is the treatment, as recently shown by Caetano et al. (2020).

[5]Examples include cell phone plan pricing (Huang, 2008), fuel economy standards (Ito and Sallee, 2017), prescription drug spending (Einav et al., 2017) and Social Security (Gelber et al., 2020).

[6]This provides a response to the point made by Einav et al. (2017) that alternative models calibrated from the bunching-design can yield very different predictions about counterfactuals. I define a particular type of counterfactual question that can be answered robustly across a class of such models.

impose a weak non-parametric shape constraint— *bi-log-concavity*—that can be verified within the support of observations and allows the researcher to place bounds on a local average treatment effect among individuals who locate at the kink. Finally, I show that these same restrictions are informative about policy counterfactuals, for example changing the location of the kink or how "sharp" it is.

The empirical context of overtime pay involves an additional challenge that is not typical to the bunching design: the kink occurs at a location that may have independent salience to firms and workers. Bunching in the hours distribution at 40 may arise in part from factors other than the FLSA rule, for example to coordinate the hours of workers across firms. I use two strategies to estimate the amount of bunching at 40 that would exist absent the FLSA, to deliver clean estimates of the rule's causal effect. First, I use the fact that when hours are paid out as holidays, sick pay, or paid-time off, they do not count towards a week's 40 hours. This "moves" the location of the kink in total hours paid during weeks when a worker is paid for non-work hours. I outline assumptions under which this yields the bunching that would occur absent the overtime rule. Second, I present a strategy that assumes alternative explanations for bunching are time-invariant to pin down the distinct contribution of the FLSA bunching at 40.

I find that the FLSA indeed has effects on hours worked, as predicted by labor demand theory. My preferred estimate suggests that just one quarter of the bunching observed in the sample (of hourly workers) at 40 is due to the FLSA, and employees working at least 40 hours work, on average, about 30 minutes less than they would absent the time-and-a-half rule. While a detailed analysis of the employment effects of the FLSA is beyond the scope of this paper, a back-of-the-envelope calculation using this estimate suggests that FLSA regulation creates about 700,000 jobs. The implied effects are larger when I use less conservative estimates of the contribution of the FLSA to the observed bunching, and overall I estimate that the local wage elasticity of hours demand close to 40 falls in the range -0.04 to -0.19. I also estimate that a reform from time-and-a-half to double pay would introduce further hours effects of a similar magnitude to those from the current FLSA, and that lowering the hours threshold from 40 to 35 would nearly eliminate bunching due to the FLSA, in the short run.

These effects speak directly to the substitutability of hours of labor between workers. The primary justifications for overtime regulation have been to reduce excessively long workweeks, while encouraging hours to be distributed over more workers (Ehrenberg and Schumann, 1982). This past year has seen a renewed interest in work sharing programs, which also pair per-worker hours reductions to keeping more workers on payroll. How well these types of policy play out in practice hinges on how easily an hour of work can be moved from one worker to another or across time, from the perspective of the firm. The effects of federal overtime policy provide a potentially large body of relevant evidence for this question. My results suggest that hours demand is relatively inelastic and that hours cannot be easily so reallocated. The estimates are also relevant to ongoing efforts to expand coverage of the FLSA overtime rule (by increasing the earnings threshold at which some salaried workers are exempt), which has resulted in one very recent major reform.[7]

The structure of the paper is as follows. Section 1.2 lays out a motivating conceptual framework that draws on the existing theory and empirical literature on overtime. Section 1.3 introduces the payroll data I use in the empirical analysis. In Section 1.4 I describe the empirical strategy, with Appendix A.1 developing some of the supporting formal results. Section 3.5 applies these results to obtain estimates of effect of the FLSA overtime rule on hours worked, as well as the effects of hypothetical reforms to the FLSA. Section 1.6 discusses the empirical findings from the standpoint of policy objectives, and 1.7 concludes.

[7]In particular, the salary threshold for employers to be free from overtime obligations for executive, administrative or professional workers was increased substantially at the beginning of 2020. Quach (2020) studies this change along with a previous attempt at an increase in 2016 that was never ultimately executed, finding evidence that salaries are moved up to the threshold and that some workers are reclassified as hourly, accompanied by a modest reduction in employment. He does not study effects on hours.

1.2 Conceptual framework

This section outlines a framework for thinking about the role of overtime policy in determining hours, which then motivates the bunching design identification strategy of Section 1.4. The framework is centered around two observations from the data in Section 1.3: weekly hours vary considerably between pay periods for an individual hourly worker, and wages tend to remain fixed with only infrequent adjustment.

I thus propose a conceptual model that views hour determination as a two stage-process. First, workers are hired with an hourly wage set along with an anticipated number of hours they will work per week. Then, with that hourly wage fixed in the short-run, final scheduling of hours is controlled by the firm and varies by week given fluctuation in their demand for each worker's labor. It is at this stage that the bunching design comes into play, given the kink in each each week's costs.

Wages and anticipated hours set at hiring

It is natural to expect both workers and firms to have preferences over the hours each employee works in a given week. Workers derive utility from non-work time, and firms may not be able to costlessly move hours between workers in production.

I bring both sides of the market together through an ex-ante "earnings-hours" posting model, which is spelled out more fully in Appendix A.4. For simplicity, workers are here taken to be homogeneous within the firm and all non-exempt from the FLSA. Each firm faces a labor supply function that takes as arguments both the total weekly compensation z it offers to each new worker, and the number of hours h they are expected to work per week at hiring.[8] The firm makes a choice of (z^*, h^*) and the corresponding employment level given the labor supply function and their production technology.

[8] This labor supply function can be viewed as an equilibrium object that reflects both worker preferences over income and leisure and the competitive environment for labor. In Supplemental Appendix A.6, I endogenize this function in a simple extension of the imperfectly competitive Burdett and Mortensen (1998) search model.

While labor supply is viewed as a function over *total* compensation z and hours, there is always a unique wage w associated with a particular (z, h) pair, such that h hours at that rate yields earnings of z, given the FLSA overtime rule

$$w_s(z, h) = \frac{z}{h + \mathbb{1}(h > 40)0.5(h - 40)} \tag{1.1}$$

I refer to this normal hourly rate of pay w, which applies to the first 40 hours, as the *straight-time wage* or simply *straight wage*. Assume that upon hiring, a worker's straight-time wage is set endogenously according to Eq. (1.1) given the firm's chosen z^* and h^*. The bunching design outlined in Section 1.4 will itself only require that *some* straight-time wage is agreed upon and is fixed in the short-run, a phenomenon that is indeed observed in the data. However, assuming that hourly wages are set based on a target total earnings z^* will play a role in my overall evaluation of the FLSA, and helps fix ideas.

In particular, the earnings-hours posting model allows us to distinguish different views that have been proposed on the effects of overtime policy. The first is what Trejo (1991) calls the *fixed-job* view of overtime: if straight time wages are set according to (1.1), then for a generic labor supply function the FLSA has no effect on employment, earnings, or hours *if* workers are in fact ultimately paid for exactly h^* hours each week (and the implied $w_s(z^*, h^*)$ is above any applicable minimum wage). The job package (z^*, h^*) posted by the firm is the same as the one that would exist absent the overtime rule, as the hourly wage rate simply adjusts to fully neutralize the additional cost of overtime pay.[9] Note that the fixed-job view abstracts away from any dynamics or uncertainty, such that the hours workers actually work is equivalent to the h^* used to determine the straight wage.

The fixed-jobs view can be contrasted with what Trejo (1991) calls the *fixed-wage* view, in which the firm faces an exogenous straight-time wage when determining hours.[10] Con-

[9]In Appendix A.4 I give a closed form expression for (z^*, h^*) when both labor supply and production are iso-elastic: hours and earnings are each increasing in the elasticity of labor supply with respect to earnings, and decreasing in the magnitude of the elasticity of labor supply with respect to pay.

[10]Versions of this idea are considered in Brechling (1965), Rosen (1968), Ehrenberg (1971), Hamermesh (1996), Hart (2004) and Cahuc and Zylberberg (2014).

tinuing with a static view of labor demand, this can be captured in our earnings-hours model by a labor supply function that reflects perfect competition on the quantity $w_s(z, h)$. In Appendix A.4 I show that in this case h^* and z^* are pinned down by the concavity of production with respect to hours and the scale of fixed costs (e.g. training) that do not depend on hours. The fixed-wage job makes the clear prediction that the FLSA will cause a reduction in hours, and bunching at 40; Figure 1.1 depicts the intuition. In a fixed-wage model the overall effect on employment is positive given plausible assumptions on the substitution between labor and capital (Cahuc and Zylberberg, 2014), though the total number of labor-hours will decrease (Hamermesh, 1996).

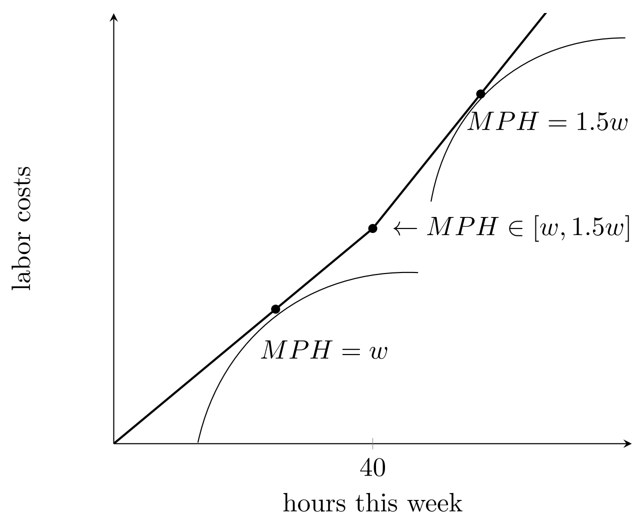

Figure 1.1: With a given worker's wages fixed at w labor costs as a function of hours have a convex kink at $h = 40$, given the overtime rule. A simple model of hours choice yields bunching when the marginal product of an hour at 40 is between w and $1.5w$ for a mass of workers—see Section 1.4.1.

A small existing literature has investigated whether the fixed-job or fixed-wage model better accords with the observed joint distribution of hourly wages and hours. Trejo (1991) and Barkume (2010) find evidence that wages do tend to be lower among jobs with overtime pay provisions and more overtime hours, however these estimates could be driven by selection of lower skilled workers into covered jobs with longer hours. In Appendix A.5, I conduct a novel empirical test of Equation (1.1) that is instead based on assuming that the conditional distribution of z is smooth across $h = 40$. Consistent with the previous findings, I find evidence of adjustment in wages, but this adjustment is far from complete. Since my data records hours at the individual paycheck level, this partial adjustment can be explained by straight wages tending to remain fixed in the short run while hours vary, as I now discuss.

Dynamic adjustment to hours by week

While the previous section considers anticipated hours and earnings at hiring, the data reveals that the hours workers are actually paid for vary considerably from week to week. Indeed the anticipated hours h^* that affect a worker's wage rate through Equation (1.1) might place little to no constraint on the hours actually scheduled in a given week.

There are many reasons why hours may vary from week to week throughout a worker's tenure at the firm. As time passes, shocks to product demand or productivity can change the number of weekly hours that would be optimal that week from the firm's perspective. For example, if demand for the firm's products is seasonal or volatile, it may not be worthwhile to hire additional workers only to reduce employment later. Similarly, cross-sectional variation in worker productivity may only become apparent to supervisors after straight wages have been set. In this case, it might be worthwhile for the firm to ask particularly productive workers to work overtime, despite the need to pay their higher overtime rate. Finally, workers may experience time-variation in their desire to work longer hours, and take advantage of overtime premium pay.

I make two main assumptions regarding the choice of a worker i's hours h_{it} in a given week t. The first is that h_{it} is a flexible choice variable on the part of the firm rather than the worker, and the second is that the firm does not contemplate alternative straight-time wages w_{it} depending on alternative choices for h_{it}. In line with the second assumption, in my sample straight-time wages do not change within worker with nearly the frequency that hourly wages do, for my sample of hourly workers (see Table 1.3). This accords with the long literature on nominal wage rigidity (see e.g. Grigsby et al. 2021 for recent evidence from payroll data). Mounting evidence that hourly wages are often standardized among workers within a firm despite cross-sectional heterogeneity (Hjort et al., 2020), and bunched at round numbers (Dube et al., 2020), also dovetails with this assumption.

The assumption that firms (rather than workers) set hours in a given week is compatible with the earnings-hours posting model above, in which the preferences of *both* sides of the market matter for the determination of an initial employment contract. It simply supposes that the firm then retains the right to set the final schedule week-to-week given each worker's agreed-upon hourly wage.[11] This view is supported by available survey evidence,[12] and can be rationalized on the basis of workers generally having less bargaining power: if the worker and firm fail to agree on a worker's hours, the worker's outside option may be unemployment while the firm's outside option is having one less worker or making a costly replacement (Stole and Zwiebel, 1996).

In the empirical strategy presented in Section 1.4, I maintain this assumption that in all cases a worker's hours are set unilaterally by their employer, which eases notation and emphasizes the intuition behind my identification strategy. However, Appendix A.2 presents a generalization in which some fraction of workers choose their hours, along with intermediate cases in which the firm and worker bargain over hours each week. If some workers have complete control over their hours, the empirical approach described in Section 1.4 will only be informative about effects of the FLSA among workers whose hours are chosen by the firm. However, the fraction of such workers appears to be small (see footnote 12), despite recent increases in flexible work arrangements.

[11] If workers are aware that the firm cannot commit to h^* every week this may affect the labor supply function; e.g. systematic departures from h^* could affect equilibrium wages. I do not attempt to model this.

[12] For example, the 2017-2018 Job Flexibilities and Work Schedules Supplement of the American Time Use Survey asks workers whether they have some input into their schedule, or whether their firm decides it. Only 17% report that they have some input. In a survey of firms, about 10% report that most of their employees have control over their shifts (Society for Human Resource Management, 2018).

1.3 Data and descriptive patterns

The main dataset I use comes from a large payroll processing company. They provided anonymized paychecks for the employees of 10,000 randomly sampled employers, for all pay periods in the years 2016 and 2017. At the paycheck level, I observe the check date, straight wage, and amount of pay and hours corresponding to itemized pay types, including normal ("straight-time") pay, overtime pay, sick leave, holiday pay, and paid time off. The data also include state and industry for each employer. Finally, for the employees, the data include age, tenure, gender, state of residence, pay frequency and their salary if one is stored in the system.

1.3.1 Sample description

I construct a final sample based on two desiderata: i) the ability to observe hours within a single week; and ii) a sample only of workers who are non-exempt from the FLSA overtime rule. For the purposes of i), I keep paychecks from workers who are paid on a weekly basis (roughly half of the workers in the sample), and condition on paychecks that contain a record of positive hours for work, vacation, holidays, or sick leave, totaling fewer than 80 hours in a week.[13]

To achieve ii) I focus on hourly workers, since nearly all workers who are paid hourly are subject to FLSA regulation. However, while the data include a field for the employer to input a salary, there is no guarantee that they use it. Therefore, I use a combination of sampling restrictions to ensure I remove all non-hourly workers from the sample. First, I drop workers that ever have a salary on file with the payroll system. Second, I only keep workers at firms for whom *some* workers have a salary on file, reflecting an assumption that employers either don't use the feature at all or use it for all of their salaried employees. I drop paychecks from workers for whom hours are recorded as 40 in every week in the

[13]This final restriction removes about 2% of the sample after the other restrictions. While a genuine 80 hour workweek is possible, I consider these observations to likely correspond to two weeks of work despite the worker's pay frequency being coded as weekly.

sample,[14] as it is possible that these workers are simply coded as working 40 hours despite being paid on a salary basis. I also drop workers who never receive overtime pay.

I drop observations from California, which has a daily overtime rule that is binding for a significant number of workers, and could confound the effects of the weekly FLSA rule. The final sample includes 630,217 paychecks for 12,488 workers across 566 firms.

Table 1.1 shows how the final sample compares to survey data that is constructed to be representative of the U.S. labor force. Column (1) reports variable means in the sample used in estimation. Column (3) reports means from the Current Population Survey (CPS) for the same years 2016–2017, among those reporting hourly employment. The "has overtime" variable for the CPS sample indicates that the worker usually receives overtime, tips, or commissions.[15] The fourth column reports means for 2016–2017 from the National Compensation Survey (NCS), a representative establishment-level dataset accessed on a restricted basis from the Bureau of Labor Statistics. The NCS uses administrative data when available, and reports typical overtime worked at the quarterly level for each job in an establishment. Columns (3) and (4) both lack some variables, as the CPS does not specifically ask about number of overtime hours, while the NCS lacks worker-level information such as tenure, age and sex.

The sample I use is somewhat more male, earns lower straight-time wages, and works more overtime than a typical U.S. worker. The NCS does not distinguish between hourly and salaried workers, reporting only an average hourly rate that does not include overtime pay. This effective straight-time wage thus includes many salaried workers, who are on average paid more, likely explaining the higher value than the CPS and payroll samples. Column (2) in Table 1.1 also reveals that my sampling restrictions can explain why the estimation sample tilts male and has higher overtime hours than the workforce as a whole. In particular, conditioning on workers that are paid on a weekly basis over-

[14]For the purposes of this drop, I count the "40 hours" event as occurring when either hours worked or hours paid is equal to 40.

[15]The hourly wage variable for the CPS may mix straight-time and overtime rates, and is only present in the outgoing rotation group sample. The tenure variable comes from the 2018 Job Tenure Supplement.

	(1) Estimation sample	(2) Initial sample	(3) CPS	(4) NCS
Tenure (years)	3.21	2.81	6.34	.
Age (years)	37.15	35.89	39.58	.
Female	0.23	0.46	0.50	.
Weekly hours	38.92	27.28	36.31	35.70
Has overtime (fraction of workers)	1.00	0.37	0.17	0.52
Straight-time wage	16.16	22.17	18.09	23.31
Weekly overtime hours	3.56	0.94	.	1.04
Number of workers in sample	12488	149459	63404	228773

Table 1.1: Comparison at the worker level of the sample with representative surveys. Column 1 reports means from the administrative payroll sample used in estimation, Column 2 from the Current Population Survey and Column 3 from the National Compensation Survey). Column 2 uses a larger sample from the payroll data, before sampling restrictions.

samples industries that tend to have more men, and tend to pay somewhat lower wages. Appendix A.5 compares the industry and regional distributions of the estimation sample to the CPS.

1.3.2 Hours and wages

I turn now to the empirical inputs that I use in estimation. Figure 1.2 reports the empirical distribution of weekly hours in the pooled sample of paychecks. The graphs indicate a large mass of individuals who were paid for exactly 40 hours, amounting to about 11.6% of the sample.[16] Appendix Figure A.10 makes clear that overtime pay is present in nearly all weekly paychecks that report more than 40 hours, in line with the assumption that the workers in my final sample are non-exempt from the FLSA.[17]

Recall from the conceptual framework of Section 1.2 that firms face a kink in labor costs within a given pay period when there is short run wage rigidity, and that this mediates the main causal effect of the FLSA on hours worked. Table 1.3 documents that while the hours paid in 70% of all pay checks in the final estimation sample differ from those of

[16]The second largest mass occurs at 32 hours, and is explained by paid-time-off, holiday, and sick pay hours as discussed in Section 3.5.

[17]However, I cannot rule out that some of the overtime pay is based on voluntary firm overtime policies.

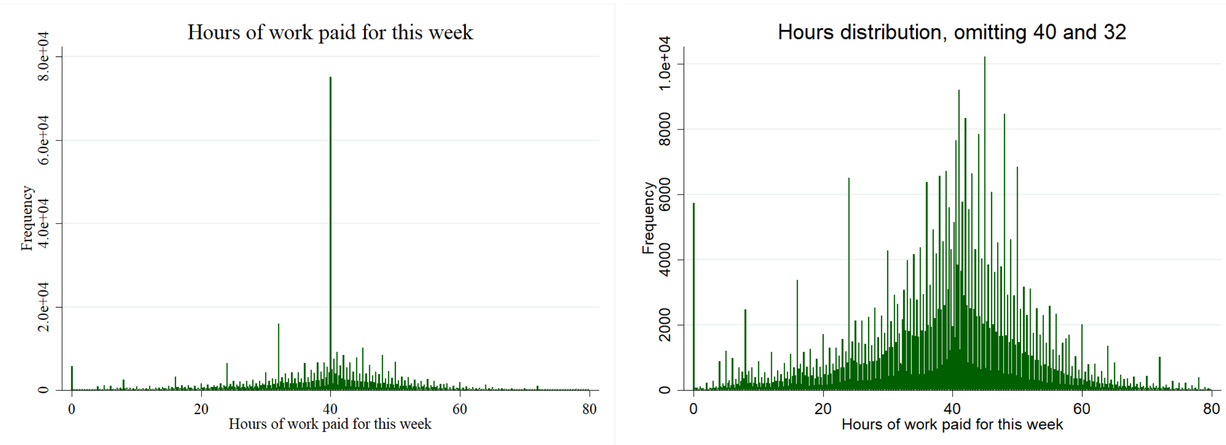

Figure 1.2: Empirical densities of hours worked pooling all paychecks in final estimation sample. Sample is restricted to hourly workers receiving overtime pay at some point (to ensure nearly all are non-exempt from FLSA, see text), and workers having hours variation. The right panel omits the points 40 and 32 to improve visibility elsewhere. Bins have a width of 1/8 of an hour, based on the observed granularity of hours (see Appendix Figure A.14 for details).

the last paycheck by at least one hour, just 4% of all paychecks record a different straight-time wage than the previous paycheck for the same worker. This figure is unchanged if I condition on the event of an hours change. Among the roughly 22,500 average wage change events, the average change is about a 45 cent increase. When hours change the magnitude is about 7 hours on average (see Supplemental Figure A.15 for the distribution of hours changes), with no average secular increase in hours over time.

Appendix Table A.5 reports a direct test of the Trejo (1991) model that straight-time wages are related to hours according to Equation (1.1). In particular, I show that under natural smoothness assumptions, the change in slopes of a regression function of straight wages on hours at 40 identifies the proportion of checks around 40 that reflect the wage-hours relationship described by Equation (1.1). This exercise suggests that about 25% of checks near 40 hours satisfy this relationship, consistent with straight wages being adjusted in response to overtime pay obligations but being updated only intermittently.

I report some further details on the variation present in the data in Appendix A.5. Appendix Table A.3 regresses hours, overtime hours, and an indicator for bunching on worker observables, and shows that after controlling for worker and date fixed effects bunching and overtime hours are both predicted by recent hiring at the firm. This lends credibility to the assumption that shocks to labor demand drive variation in hours. Appendix Table A.4 shows that overall, about 63% of variation in total hours can be explained by worker and employer by date fixed effects. Appendix Figure A.2 documents heterogeneity in the prevalence of overtime pay across industry classifications. Industries with the largest average overtime pay include Health Care and Social Assistance, Administrative and Support, and Transportation and Warehousing.

	Mean	Std. dev.	N
Indicator for hours changed from last period	0.84	0.37	630,217.00
Indicator for hours changed by at least 1 hour	0.70	0.46	630,217.00
Indicator for wage changed from last period	0.04	0.19	630,217.00
Indicator for wage changed, if hours changed	0.04	0.19	529,791.00
Difference in hours, if hours changed	-0.02	10.69	529,791.00
Absolute value of hours difference, if hours changed	6.83	8.23	529,791.00
Difference in wage, if wage changed	0.45	26.46	22,501.00

Figure 1.3: Changes in hours paid or straight time wages between consecutive paychecks, within worker.

1.4 Empirical strategy: a generalized kink bunching design

In this section I consider the firm making its week-to-week choice of hours for a given worker, with costs a kinked function of hours as depicted in Figure 1.1. I show that under weak assumptions, firms facing such a kink will make a choice that can be completely characterized by choices they *would* make under two counterfactual linear cost schedules that do not feature the kink, and differ with respect to a single worker's hourly wage. I then parlay the observable bunching at 40 hours into a statement about the joint distribution of these counterfactuals, which can be interpreted in the language of treatment effects. Finally, I use these treatment effects to estimate my main parameter of interest: the average effect of the FLSA on hours.

The identification results in this section hold in a much more general setting in which a generic decision-maker faces a kinked choice set and has convex preferences. I present this general model in Appendix A.1, and some of the formal assumptions are given there rather than in the main text. Throughout this section I refer to a worker i in week t as a *unit*: an observation of h_{it} for unit it is thus the hours recorded on a single paycheck. Probability statements are to be understood with respect to the pooled distribution of such paychecks across the sample period.

1.4.1 A benchmark model: hours chosen from marginal productivity

Let us begin with the conceptual framework introduced in Section 1.2. With the wage fixed, the firm in week t faces a kinked cost schedule in deciding hours h_{it} for a given worker. If the firm chooses less than 40 hours, it will pay $w = w_{it}$ for each hour, where w_{it} is the straight-time wage.[18] If the firm chooses $h > 40$, then it will pay $40w$ for the first forty hours and $1.5w(h-40)$ for the remaining hours, giving the convex shape to Figure 1.1. Let $B_{kit}(h) = w_{it}h + .5w_{it}\mathbb{1}(h > 40)(h-40)$ be the kinked pay schedule for unit it.

A natural view of weekly hours demand is that firms balance the cost $B_{kit}(h)$ against the value of h hours of the worker's labor, in order to maximize profits. Consider a single firm, and let $F_t(h, \mathbf{h}_{-i,t})$ denote production in dollars this week, where h are the hours for worker i and $\mathbf{h}_{-i,t}$ is the vector of hours for the other workers in the firm. Take F to be strictly concave in the total hours profile of its workers $\mathbf{h} = (h, \mathbf{h}_{-i,t})$, such that the marginal product of an hour $MPH_{it}(h) = \frac{\partial}{\partial h}F_t(h, \mathbf{h}_{-i,t})$ is declining as a function of h. If firms maximize weekly profits, they will choose $h < 40$ when MPH equals the straight time wage for some such value of h. This situation is depicted by the leftmost indifference

[18] A unit's straight-time wage w_{it} is fixed with respect to the choice of hours this week, but may depend on t due to e.g. occasional or automatic periodic raises.

curve in Figure 1.1. By concavity of production, MPH declines with h. If the MPH is still above $1.5w$ at $h = 40$, for a worker with wage w, then tangency with the budget constraint $B_{kit}(h)$ will occur for some $h > 40$ where $MPH(h) = 1.5w$. This is depicted by the rightmost indifference curve in Figure 1.1. If the MPH at $h = 40$ is between w and $1.5w$, then the firm will choose to locate that worker at the corner solution $h = 40$.

These predictions may be summarized as follows, separating the cases based on the marginal productivity of a worker's hours at 40:

$$h_{it} = \begin{cases} MPH_{it}^{-1}(w_{it}) & \text{if} \quad MPH_{it}(40) < w_{it} \\ 40 & \text{if} \quad MPH_{it}(40) \in [w_{it}, 1.5w_{it}] \\ MPH_{it}^{-1}(1.5w_{it}) & \text{if} \quad MPH_{it}(40) > 1.5w_{it} \end{cases} \tag{1.2}$$

Shocks to the function F_t, or to the hours $\mathbf{h}_{-i,t}$ worked by $i's$ colleagues within the firm, can be seen as determining which of the three types of outcome occurs in a given week.

While Equation 1.2 provides fairly general intuition, it is useful to consider a simpler context that ignores interdependencies between workers and assumes that heterogeneity in hours is driven by a scalar productivity parameter: $F_t(h, \mathbf{h}_{-i,t}) = a_{it} \cdot f(h)$ where $f' > 0$ and $f'' < 0$. Then $MPH_{it}(h) = a_{it} \cdot f'(h)$, where the function f is common across firms, workers, and time periods. If $f(h)$ is furthermore iso-elastic, we arrive at the canonical bunching-design approach from the literature (Saez, 2010; Chetty et al., 2011; Kleven, 2016).[19] The iso-elastic case is illustrative, and I will focus on it as a benchmark, before generalizing. In the iso-elastic model, firm profits take the form:

$$\pi_{it}(z, h) = a_{it} \cdot \frac{h^{1+\frac{1}{\epsilon}}}{1 + \frac{1}{\epsilon}} - z \tag{1.3}$$

where $\epsilon < 0$ is common across all units it, and c are labor costs for worker i in week

[19] Alternatively, they may allow heterogeneous elasticities by taking the kink to be suitably "small". My approach allows us to relax both assumptions at the same time.

t. Under any linear pay schedule $z = wh$, the profit maximizing number of hours is $\left(\frac{w}{a_{it}}\right)^{\epsilon}$, so ϵ can be interpreted as the elasticity of hours demand to the wage. Define $\eta_{it} = a_{it}/w_{it}$, the ratio of the current productivity factor to the straight-time wage. Then, by Equation (1.2) hours are ranked across units by their value of η_{it}. Namely, $h_{it} = \eta_{it}^{-\epsilon}$ if $\eta_{it} < 40^{-1/\epsilon}$, $h_{it} = 1.5^{\epsilon} \cdot \eta_{it}^{-\epsilon}$ if $\eta_{it} > 1.5 \cdot 40^{-1/\epsilon}$, and $h_{it} = 40$ if η_{it} falls in the intermediate region $[40^{-1/\epsilon}, 1.5 \cdot 40^{-1/\epsilon}]$. If η_{it} is continuously distributed over a region containing this interval, then the observed distribution of h_{it} will feature a point mass at 40: "bunching" – paired with a density elsewhere.

Now consider identifying the effect of the FLSA, in the context of this iso-elastic model. Let $h_{0it} = \eta_{it}^{-\epsilon}$ be the hours it would work if their employer faced the straight-time wage rate for all hours. I will refer to the difference $h_{it} - h_{0it}$ as the *effect of the kink*—the effect of the FLSA on unit it when ignoring changes to workers' straight-time wage, or complementaries between units (I account for effects on wages in Section 1.6). In the iso-elastic model, the effect of the kink is

$$h_{it} - h_{0it} = \begin{cases} 0 & \text{if} \quad h_{it} < 40 \\ 40 - h_{0it} & \text{if} \quad h_{it} = 40 \\ h_{it} \cdot (1 - 1.5^{-\epsilon}) & \text{if} \quad h_{it} > 40 \end{cases}$$

Given the value of ϵ, we could evaluate this effect for any paycheck recording overtime $h_{it} > 40$ using the worker's observed hours. We could then easily estimate, for example, the average treatment effect among paychecks having overtime hours.

Thus a natural starting place for evaluating the FLSA via the bunching design is to estimate ϵ. Assume that we have access to a random sample of paychecks h_{it}.[20]. If we were willing to suppose η_{it} belongs to a parametric family, then the entire model could be estimated by maximum likelihood (Bertanha et al., 2020). The method pioneered by Saez (2010) is more local: ϵ is related to the observable bunching probability $\mathcal{B} = P(h_{it} = 40)$.

[20] The empirical implementation relaxes this and only assumes independence between firms.

Figure 1.4 depicts the intuition, which is convenient to express in terms of the log-hours distribution.

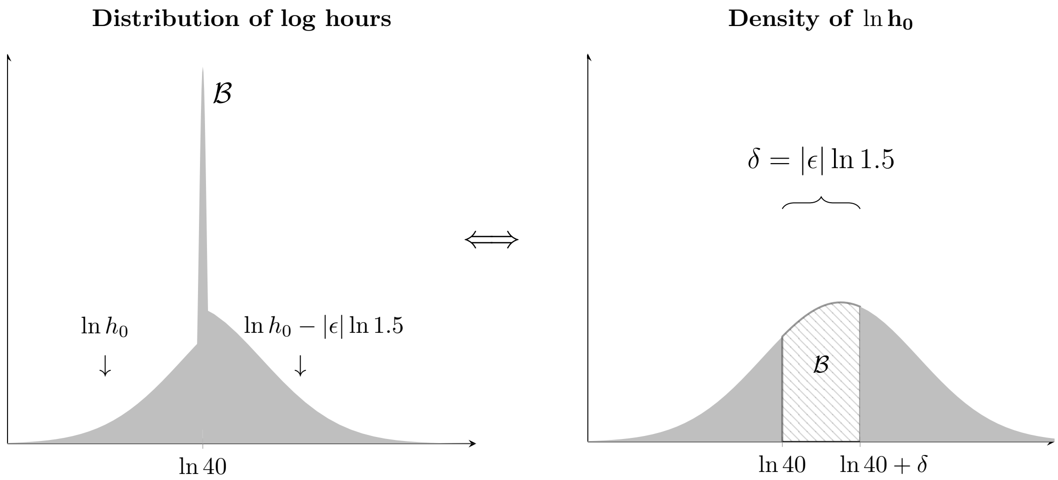

Figure 1.4: The left panel depicts the distribution of observed log hours $\ln h_{it}$ in the iso-elastic model, while the right panel depicts the underlying full density of $\ln h_{0it}$. The full density is related to the observed density by "sliding" the observed density for $h > 40$ out by the unknown distance $\delta = |\epsilon| \ln 1.5$. The density of h_{0it} is not observed in the missing region between $\ln 40$ and $\ln 40 + \delta$, but the area total therein must equal the observed bunching mass $\mathcal{B}$.

If the researcher unwilling to assume anything about the density of h_0 in the missing region of Figure 1.4, then the data are compatible with any finite $\epsilon < 0$, a point emphasized by Blomquist and Newey (2017) and Bertanha et al. (2020). In particular, given the integration constraint that $\mathcal{B} = P(\ln h_{0it} \in [\ln 40, \ln 40 + \delta])$, an arbitrarily small $|\epsilon|$ could be rationalized by a density that spikes sufficiently high just to the right of 40, while an arbitrarily large $|\epsilon|$ can be reconciled with the data by supposing that the density drops quickly to some very small level throughout the missing region.

Standard methods from the literature use parametric assumptions to point-identify ϵ in the iso-elastic model. The approach of Saez (2010) assumes that the density of h_{0it} (not in logs) is linear through the corresponding region $[40, 40 \cdot 1.5^{-\epsilon}]$. The popular method of Chetty et al. (2011) instead fits a global polynomial to the hours distribution. However, neither of these approaches is suitable for the overtime context. The linear method of Saez

(2010) implies monotonicity of the density in the missing region, which is unlikely to hold given that 40 appears to be near the mode of the latent hours distribution. The method of Chetty et al. (2011) ignores the "shift" by δ in the right panel of Figure 1.4, which would be problematic in this setting since the slope of the density is far from zero and the bunching at 40 is exact, rather than diffuse.

My approach instead imposes a non-parametric shape constraint: bi-log-concavity, on the distribution of h_{0it}. Bi-log-concavity (BLC) generalizes the familiar property of log-concavity, and importantly allows for a peak within the missing region (Dümbgen et al., 2017). I defer a detailed discussion of BLC to Section 1.4.3, after I generalize from the iso-elastic model, and indeed more generally from a model in which hours are chosen on the basis of productivity alone. The reason for this generalization is two-fold. First, its weakens the assumptions under which the effect of the FLSA on hours can be identified. Second, it enables a range of underlying models that might be used to rationalize the results.

The robustness over structural models is important in the overtime context. The iso-elastic model applied to the data described in Section 1.3 yields implausible values for ϵ, when interpreted in the context of the hours production function from Equation (1.3). Appendix A.5.4 reports estimates of the identified set of values for ϵ compatible with the data and BLC of h_0. The bounds are narrow and suggest a value of about $\epsilon = -0.2$, when all of the bunching observed at 40 is attributed to the FLSA.[21] This value would suggest that revenue as a function of hours is (up to an affine transformation): $f(h) = -\frac{1}{4}h^{-4}$, a production function with an unreasonable degree of concavity. Note that attributing just a portion of the observed bunching at 40 to the FLSA, as I do in Section 1.5.1 would further reduce the estimate of ϵ. The more general separable model in which $f(h)$ is arbitrary is also not much help here, since estimating the iso-elastic model then identifies an averaged

[21] This estimate is from the pooled sample across all industries. Also reported Appendix A.5.4, estimation by industry yields bounds on ϵ ranging from -0.26 to -0.06, which are similarly implausible as estimates of concavity of production. The estimates are similar when applying the linear density assumption from Saez (2010).

local inverse elasticity of $f(h)$. In particular: $h_{1it} - h_{0it} = h_{0it}\left(1.5^{\bar{\epsilon}_{it}} - 1\right)$ where $\bar{\epsilon}_{it}$ is a unit-specific weighted average of the inverse elasticity of production between $1.5\eta_{it}$ and η_{it}: $\bar{\epsilon}_{it} := \int_{\eta_{it}^{-1}}^{1.5\eta_{it}^{-1}} w(m) \cdot \epsilon(g(m)) \cdot dm$ where $\epsilon(h) := \frac{f'(h)}{f''(h)h}$ is the reciprocal of the local elasticity of production, $g(m) := (f')^{-1}(m)$ yields the hours h at which $f'(h) = m$, and $w(m) = \frac{1/m}{\ln 1.5}$ is a positive function integrating to one.

Put simply, the observed bunching is too small to be reconciled with an iso-elastic response in which ϵ parameterizes the concavity of production with respect to hours: it is better interpreted as a reduced form elasticity of demand for hours. The next section formalizes this idea, by showing how identification in the bunching design generalizes to a class of models that can include additional choice variables that may attenuate the observed labor demand response to overtime pay, as well as incorporate multi-dimensional heterogeneity.

1.4.2 Counterfactual choices in a larger class of choice models

The basic structure of what is observable in the bunching design is preserved when we not only relax the constant-elasticity assumption, but also when we allow the firm to have multiple choice-variables that may be responsive to the incentives created by the kink. Additional margins of response can have the effect of diminishing the hours response that would occur on the basis of production alone, which can explain the small elasticity reported in the last section.

Begin by observing that in the model of the last section, units who work overtime work the number of hours that they would work if their wage was 1.5 times their straight time wage: c.f. Equation (1.2). This property holds quite generally. Let h_{1it} be the hours that would be chosen for it if their straight-time wage were instead equal to $1.5w_{it}$.[22]. Appendix A.1 presents a generic model of choice for the bunching design in which Equation

[22]Specifically, the counterfactuals h_{0it} and h_{1it} replace the cost schedule for this week's hours with a linear wage w_{it} or $1.5w_{it}$, holding fixed both w_{it} and the hours of other units. See Section 1.4.4 for more details.

(1.2) can be seen as a special case of:

$$h_{it} = \begin{cases} h_{0it} & \text{if} \quad h_{0it} < 40 \\ 40 & \text{if} \quad h_{1it} \leq 40 \leq h_{0it} \\ h_{1it} & \text{if} \quad h_{1it} > 40 \end{cases} \tag{1.4}$$

This expression says that knowledge of the two counterfactual hours choices h_{0it} and h_{1it} are sufficient to pin down the actual hours chosen for any given unit. The worker will work h_{0it} when h_{0it} is less than 40, h_{1it} when it is greater than 40, and be located at 40 if and only if the two counterfactual outcomes "straddle" the kink, falling on either side.

Appendix Lemma A.1 shows that Equation 1.4 holds quite generally when an exogenous change to the hours-pay schedule would cause the firm to re-optimize on a vector **x** of choice variables that includes hours of work h as a component, and firm preferences are convex in the pair $(z, \mathbf{x})$, where z are this period's wage costs. To demonstrate the flexibility of this framework, I present some examples beyond the baseline model of the last section. These examples are illustrative, and each could apply to a different subset of units in the population.[23]

Example 1: Substitution from bonus pay

Let the firm's choice vector be $\mathbf{x} = (h, b)'$, where $b \geq 0$ indicates a bonus (or other fringe benefit) paid to the worker. Firms may find it optimal to offer bonuses to improve worker satisfaction and reduce turnover. Suppose firm preferences are: $\pi(z, h, b) = f(h) + g(z + b - \nu(h)) - z - b$, where z continues to denote wage compensation this week, $z + b - \nu(h)$ is the worker's utility with $\nu(h)$ a convex disutility from labor h, and $g(\cdot)$ increasing and concave. In this model firms will choose the surplus maximizing choice of

[23] Appendix A.2 discusses a further example in which the firm and worker bargain over this week's hours. This weekly bargaining can diminish the wage elasticity of hours since overtime pay gives the parties opposing incentives.

hours $h_m := \text{argmax}_h f(h) - \nu(h)$ regardless of the hourly wage, provided that the corresponding optimal bonus is feasible (e.g. non-negative). Bonuses may thus fully adjust to absorb the added costs of overtime pay, such that $h_0 = h_1 = h_m$.

Example 2: Off-the-clock hours and paid breaks

Suppose firms choose a pair $\mathbf{x} = (h, o)'$ with h hours worked and o hours worked "off-the-clock", such that $y(\mathbf{x}) = h - o$ are the hours for which the worker is paid. This model can include some firms voluntarily offering paid breaks by allowing o to be negative. Evasion is harder the larger o is, which we represent by firms facing a convex evasion cost $\phi(o)$, so that firm utility is $\pi(z, h, o) = f(h) - \phi(o) - z$. Note that the data observed in our sample are of hours of work $y(\mathbf{x})$ for which the worker is paid, when this differs from h. Appendix A.1 describes how Equation 1.4 still holds, but for counterfactual values of hours paid $y = h - o$ rather than hours worked h. The bunching design lets us investigate treatment effects on paid hours, without observing off-the-clock hours or break time o.

Example 3: Complementaries between workers or weeks

Suppose the firm simultaneously chooses the hours $\mathbf{x} = (h, g)$ of two workers according to production that is iso-elastic in a CES aggregate of the two worker's hours. I focus on the hours h for the first worker (g could also denote planned hours next week for the same worker): $\pi(z, h, g) = a \cdot \left((\gamma h^\rho + g^\rho)^{1/\rho}\right)^{1+\frac{1}{\epsilon}} - z$, where $\gamma > 0$ reflects a relative productivity shock for the first worker, and z are labor costs. Let g^* denote the firm's optimal choice of hours for the second worker. The firm's choice of h must maximize $\pi(z, h, g^*)$ subject to $z = B_k(h)$, as if the firm faced a single-worker production function of $f(h) = a \cdot \left((\gamma h^\rho + g^{*\rho})^{1/\rho}\right)^{1+\frac{1}{\epsilon}}$. This function is more elastic than the corresponding single-worker iso-elastic production function with the same $\epsilon < 0$ provided that $\rho < 1 + 1/\epsilon$, since $\frac{f''(h)h}{f'(h)} = \frac{1}{\epsilon} - \frac{1+1/\epsilon-\rho}{\gamma(h/g^*)^\rho+1}$, attenuating the response to an increase in w (with g^*

fixed) implied by a given ϵ, provided sufficient complementarity.[24]

1.4.3 Identifying treatment effects in the bunching design

Given the definitions in the last two sections, let $\Delta_{it} = h_{0it} - h_{1it}$. This is the difference between the firm's choice of hours for a given worker (this week) if they were paid at their straight-time rate for all hours, versus their overtime rate for all hours. I refer to Δ_{it} as it's *treatment effect*, interpreting h_0 and h_1 as potential outcomes. A unit's treatment effect can be contrasted with the "effect of the kink" quantity $h_{it} - h_{0it}$ introduced earlier: the effect of the kink is $-\Delta_{it}$ for those units working overtime.[25]

Beyond the iso-elastic model, Δ_{it} rather than ϵ is the quantity of interest in causal analysis. In the iso-elastic model $\Delta_{it} = h_{0it} \cdot (1 - 1.5^{\epsilon})$; this model thus delivers treatment effects in logs: $\ln h_{0it} - \ln h_{1it} = |\epsilon| \cdot \ln 1.5$ that are constant across all units (see Figure 1.4). In general we can expect Δ_{it} to vary across units, and a reasonable parameter of interest is some summary statistic of Δ_{it}. To ease notation, let $k = 40$ denote the location of the kink. We can see that bunching should be in some way informative about the distribution of Δ_{it} by using Equation (1.4) to write the bunching probability as:

$$\mathcal{B} = P(h_{1it} \leq k \leq h_{0it}) = P(h_{0it} \in [k, k + \Delta_{it}]) = P(h_{1it} \in [k - \Delta_{it}, k]) \tag{1.5}$$

Units bunch when either of their counterfactual outcomes lie within their individual treatment effect of the kink. Note that $\mathcal{B} = F_1(k) - F_0(k)$ provided that h_{0it} and h_{1it} are continuously distributed, where F_0 and F_1 are their cumulative distribution functions.

The existing literature on the bunching design contains few positive identification results that move beyond univariate heterogeneity and explicitly allow responsiveness to

[24] This expression overstates the degree of attenuation, since h_1 and h_0 maximize $f(h)$ above for different values g^*, which leads to a larger gap between h_0 and h_1 compared with a fixed g^* by the Le Chatelier principle (e.g. Milgrom and Roberts, 1996). However, given $\rho < 1 + 1/\epsilon$, maintaining productivity of the second worker gives the firm enough incentive against decreasing h that h_1/h_0 still increases on net.

[25] Both of these treatment effects are "partial equilibrium" in the sense that they hold the hours worked by units other than it fixed at their actual values. Section 1.4.4 discusses this further when evaluating the FLSA.

vary by individual unit. Saez (2010) and Kleven (2016) consider a "small-kink" approximation that allows one to estimate $\mathbb{E}[\Delta_{it}|h_{0it} = k]$, in the present notation.[26] In the overtime setting, a 50% increase in the hourly cost of labor is likely to produce large enough effects that this approximation would be quite poor. Blomquist et al. (2021) allow multi-dimensional heterogeneity in a labor supply model under taxation, by assuming the density of counterfactual choices at a kink is linear across tax rates. However this type of assumption can be hard to motivate.

One type of heterogeneity that it is important to allow in the context of overtime is some degree of non-responsiveness to the incentives introduced by the kink at 40 hours, since 40 is a particularly salient hours choice. Let $K^*_{it} = 1$ indicate a group of units such that $h_{0it} = h_{1it} = k$. I refer to these units as "counterfactual bunchers", since they would locate at the kink even in the counterfactual outcome distributions. These units are not of particular interest, but they complicate measurement of the bunching caused by kink when there is a positive mass $p := P(K^*_{it} = 1)$ of counterfactual bunchers. In this section, I treat p as known, and estimate it empirically in Section 1.5.1. Given p and the CDF $F(h)$ of the data, one can construct the conditional distribution for all other units (denoted by $K^*_{it} = 0$) by simply subtracting p from the observed bunching mass $\mathcal{B}$ and re-normalizing the distribution, i.e. $F_{h|K^*=0}(h) = \frac{F(h)-p\mathbb{1}(h\geq k)}{1-p}$.

I focus on partial identification of the average treatment effect among units who locate at the kink and are not counterfactual bunchers, what I call the "buncher LATE":

$$\Delta^*_k = \mathbb{E}[\Delta_{it}|h_{it} = k, K^*_{it} = 0]$$

To simplify the discussion, suppose for now that there are no counterfactual bunchers, so

[26] In particular, the density of h_0 is taken to be constant throughout the region $[40, 40 + \Delta_{it}]$ conditional on each value of Δ_{it}, leading to $\mathbb{E}[\Delta_{it}|h_{0i} = 40] = \mathcal{B}/\lim_{h\uparrow k} f(h)$, with $f(h)$ the density of observed hours (see Appendix A.1 for a derivation in my generalized framework). The uniform density assumption is hard to justify except in the limit that the distribution of Δ_{it} concentrates around zero. Lemma SMALL in Appendix C.4 makes this claim precise, while connecting the approach from Saez (2010) and Kleven (2016) to a non-parametric treatment without point identification by Blomquist et al. (2015).

that $\Delta_k^* = \mathbb{E}[\Delta_{it}|h_{it} = k]$. My approach to identifying bounds on Δ_k^* is based on assuming a weakened version of *rank invariance* between h_0 and h_1:

$$P\left(F_0(h_{0it}) = F_1(h_{1it})\right) = 1. \tag{1.6}$$

Equation (1.6) says that increasing each unit's wage by 50% does not change the rank of each unit's hours: for example, a worker at the median of the h_0 distribution also has a median value of h_1. This is satisfied by models in which there is perfect positive co-dependence between the potential outcomes, such-as the benchmark model from Section 1.4.1 with production $a_{it} \cdot f(h)$. The left panel of Figure 1.6 shows an example.

Rank invariance allows us to translate statements about Δ_{it} into statements about the marginal distributions of h_{0it} and h_{1it}. In particular, under rank invariance the buncher LATE is equal to the quantile treatment effect $Q_0(u) - Q_1(u)$ averaged across all u between $F_0(k)$ and $F_1(k) = F_0(k) + \mathcal{B}$, with Q_d the quantile function of h_{dit}:

$$\Delta_k^* = \frac{1}{\mathcal{B}} \int_{F_0(k)}^{F_1(k)} [Q_0(u) - Q_1(u)] du, \tag{1.7}$$

so long as $F_0(y)$ and $F_1(y)$ are continuous and strictly increasing. To place bounds on the buncher LATE, it is thus sufficient to place point-wise bounds on the quantile functions $Q_0(u)$ and $Q_1(u)$ throughout the range $u \in [F_0(k), F_1(k)]$, as depicted in Figure 1.5.

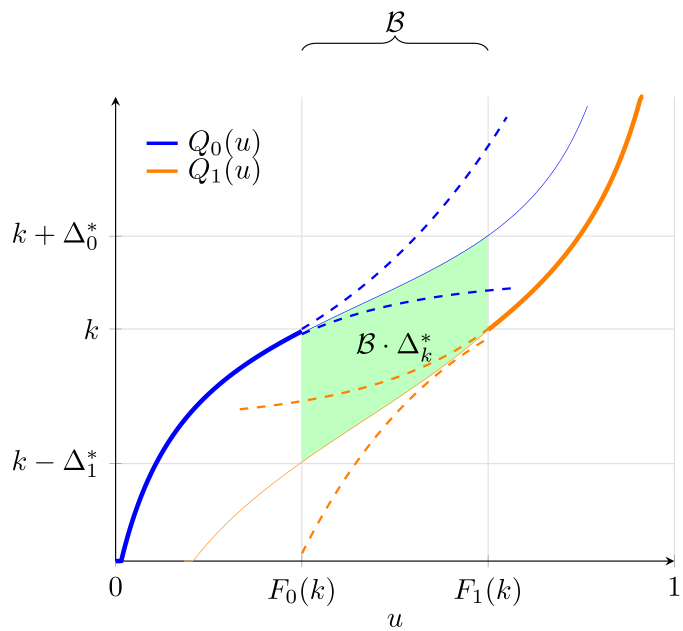

Figure 1.5: Extrapolating the quantile functions for h_0 and h_1 (blue and orange, respectively) to place bounds on the buncher LATE. The observed portions of each quantile function are depicted by thick curves, while the unobserved portions are indicated by thinner curves. The dashed curves represent upper and lower bounds for this unobserved portion implied by bi-log-concavity (see text below). The buncher LATE is equal to the area shaded in green, divided by the bunching probability $\mathcal{B}$. The quantities Δ_0^* and Δ_1^* are defined in Assumption RANK below.

I obtain such bounds by assuming that both h_0 and h_1 have *bi-log-concave* distributions. Bi-log-concavity is a non-parametric shape constraint that generalizes log-concavity, a property of many common parametric distributions:

Definition (BLC). *A distribution function F is is bi-log-concave (BLC) if both $\ln F$ and $\ln(1-F)$ are concave functions.*

If F is BLC then it admits a strictly positive density that is itself differentiable with the locally bounded derivative: $\frac{-f(h)^2}{1-F(h)} \leq f'(h) \leq \frac{f(h)^2}{F(h)}$ (Dümbgen et al., 2017). Intuitively, this rules out cases in which the density of either h_0 or h_1 ever spikes or falls *too* quickly on the interior of its support, leading to non-identification of the type discussed in Section 1.4.1.[27] The family of BLC distributions includes uniform and linear densities (as assumed

[27]Bertanha et al. (2020) propose partial identification in an iso-elastic model by specifying a Lipschitz

by Saez 2010), as well as all globally log-concave distributions such as the normal.[28] Importantly, the BLC property is partially testable in the bunching design, since $F_0(y)$ is identified for all $h < k$ and $F_1(h)$ is identified for all $h > k$. Appendix Figure A.9 shows that these observations in the data are indeed consistent with BLC. I will also refer to a random variable as "BLC" if its distribution is BLC. For each $d \in \{0, 1\}$, assuming h_{dit} is BLC yields point-wise upper and lower bounds on the quantile function $Q_d(u)$ appearing in Equation (1.7) that depend on $F_d(k)$ and $f_d(k)$, with f_d the density of h_{dit}.[29]

Assuming that each of h_0 and h_1 are separately BLC thus allows me to move beyond point-identification based on strong parametric assumptions while simultaneously accommodating heterogeneous treatment effects, requiring only rank invariance. But while rank invariance weakens the homogeneity assumptions typically made in the literature, it is nevertheless a restrictive assumption in the overtime setting. Fortunately, a still weaker assumption proves sufficient for the RHS of (1.7) to recover the buncher LATE:

Assumption RANK. *There exist values Δ_0^* and Δ_1^* such that $h_{0it} \in [k, k + \Delta_{it}]$ iff $h_{0it} \in [k, k + \Delta_0^*]$, and $h_{1it} \in [k - \Delta_{it}, k]$ iff $h_{1it} \in [k - \Delta_1^*, k]$.*

Note that Δ_0^* and Δ_1^* are fixed numbers that do not vary by unit it. If treatment effects were homogeneous with $\Delta_{it} = \Delta$, we would have $\Delta_0^* = \Delta_1^* = \Delta$, and Assumption RANK would simply echo Equation 1.5. With heterogeneous effects however, RANK allows ranks to be reshuffled by treatment among bunchers and on either side of the bunching region.[30] For example, suppose that a 50% increase in the wage of worker i would result in their hours

constant on the density of $\ln \eta_{it}$. This yields global rather than local bounds on f'.

[28] BLC distributions can have multiple modes however, relaxing the unimodality property of log-concave densities (Dümbgen et al., 2017). Note that any polynomial density with real roots is a log-concave function.

[29] It is worth noting that under rank invariance, assuming BLC of h_1 and h_0 is sufficient to calculate bounds on the treatment effect $Q_1(u) - Q_0(u)$ at any quantile $u \in [0, 1]$. However, these bounds quickly widen as one moves away from the kink in either direction. The narrowest bounds for a single rank are obtained for a "median" buncher roughly halfway between $F_0(k)$ and $F_1(k)$ when $f_0(k) \approx f_1(k)$. However, averaging over a larger group is more useful for meaningful ex-post evaluation of the FLSA, and reduces the sensitivity to departures from rank invariance (see Figure A.2). The buncher LATE balances these considerations.

[30] Given Equation (1.4), RANK is equivalent to the *rank-similarity* assumption of Chernozhukov and Hansen (2005), where the conditioning variable V_i indicates which of the three cases of Equation (1.4) hold for the unit.

being reduced from $h_{0it} = 50$ to $h_{1it} = 45$. If another worker j's hours are instead reduced from $h_{0jt} = 48$ to $h_{1jt} = 46$ under a 50% wage increase, workers i and j will switch ranks, without violating RANK. Note also that RANK is compatible with the existence of counterfactual bunchers $p > 0$.

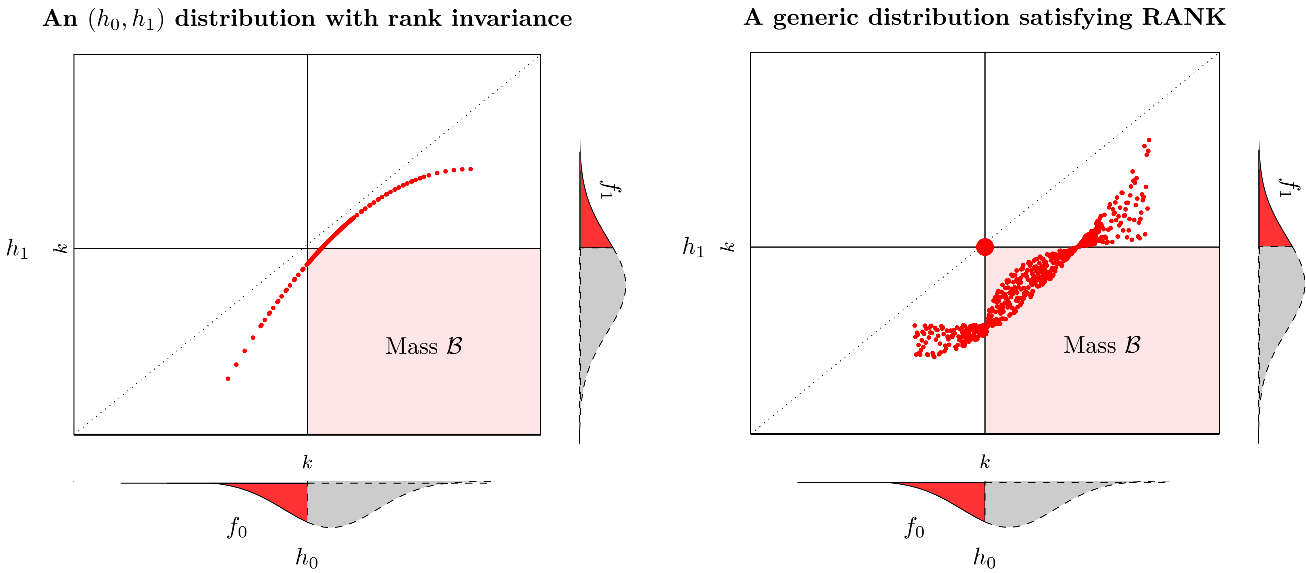

Figure 1.6: The joint distribution of (h_{0it}, h_{1it}), comparing an example satisfying rank invariance (left) to a case satisfying Assumption RANK (right). RANK allows the support of the joint distribution to "fan-out" from perfect co-dependence of h_0 and h_1, except when either outcome is equal to k. The large red dot in the right panel indicates a possible mass p of counterfactual bunchers. The observable data identifies the red portions of outcome's marginal distribution (depicted along the bottom and right edges), as well as the total mass $\mathcal{B}$ in the (shaded) south-east quadrant.

The right panel of Figure 1.6 shows an example of a distribution satisfying RANK. When RANK is not perfectly satisfied (e.g. when the support of (h_0, h_1) doesn't quite narrow to a point at each $h_d = k$), Δ^*_k can still be interpreted as an averaged quantile treatment effect across $[F_0(k), F_1(k)]$. Appendix Figure A.2 explains that this will then represent a lower bound on the true buncher LATE. Appendix Figure A.3 depicts a case in which some workers choose their hours, resulting in mass in the north-west quadrant.

Theorem 1.1 gives sharp bounds on the buncher LATE given RANK and bi-log-concavity. It requires two further assumptions that have so far been implicit: hours can be perfectly manipulated by firms, and firms' preferences are convex over available choice variables.

Appendix A.1 gives a formulation of these assumptions for more general kink settings, and shows that bunching still has identifying power without convexity of preferences.

Assumption CHOICE. *The outcomes h_{0it}, h_{1it} and h_{it} reflect choices the firm would make under counterfactual cost constraints $z \geq B(h)$, with $B(h)$ given by $B_{0it}(h) = w_{it}h$, $B_{1it}(h) = 1.5w_{it}h - 20w_{it}$, or $B_{kit} = \max\{B_{0it}(h), B_{1it}(h)\}$ respectively.*

Assumption CONVEX. *Firm choices maximize some $\pi_{it}(z, \mathbf{x})$, where π_{it} is strictly quasiconcave in $(z, \mathbf{x})$ and decreasing in z. Hours h are a continuous deterministic function of $\mathbf{x}$.*

Note that the importance of firms being the decision-maker for a unit enters in the assumption that utility π is decreasing, rather than increasing, in z. Appendix A.2 relaxes this to allow some workers to set their hours. The second term in the definition of h_{1it} keeps the firm indifferent between B_1 and B_0 at $h = 40$, and is only necessary for Equation 1.4 (and the subsequent analysis) to hold when preferences π are not quasi-linear in z.[31] Since quasi-linearity with respect to costs is implied by firms maximizing profits, h_{1it} can be thought of as hours under the simple pay schedule $1.5w_{it}h$.

Theorem 1.1 (bi-log-concavity bounds on the buncher LATE). *Assume CHOICE, CONVEX, RANK and that h_{0it} and h_{1it} are both bi-log-concave conditional on $K^*_{it} = 0$. Then:*

1. *Each of $F(h)$, $F_0(h)$ and $F_1(h)$ are continuously differentiable for $h \neq k$. When $p > 0$, define the density $f_d(y)$ of h_{dit} at $y = k$ to be $f_d(k) = \lim_{h \to k} f_d(h)$, for each $d \in \{0, 1\}$.*

2. *The buncher LATE $\Delta^*_k \in \left[\Delta^L_k, \Delta^U_k\right]$, where:*

$$\Delta^L_k := g(F_0(k) - p, f_0(k), \mathcal{B} - p) + g\left(1 - F_1(k), f_1(k), \mathcal{B} - p\right)$$

and

$$\Delta^U_k := -g(1 - F_0(k), f_0(k), p - \mathcal{B}) - g\left(F_1(k) - p, f_1(k), p - \mathcal{B}\right)$$

[31] This reflects the well-known observation that the bunching design yields a combination of compensated and uncompensated elasticities (Blomquist et al., 2015; Kleven, 2016).

with $g(a,b,x) = \frac{a}{bx}(a+x)\ln\left(1+\frac{x}{a}\right) - \frac{a}{b}$, *and the bounds are sharp.*

Proof. See Appendix C.4. □

Let $f(h)$ be the density of the data for $h \neq k$. Given p, the remaining quantities in Theorem 1.1 are identified: $F_0(k) = \lim_{h\uparrow k} F(h) + p$, $F_1(k) = F(k)$, $f_0(k) = \lim_{h\uparrow k} f(h)$ and $f_1(k) = \lim_{h\downarrow k} f(h)$.[32]

Inspection of the expressions appearing in Theorem 1.1 reveals that the bounds become wider the larger the net bunching probability $\mathcal{B} - p$. A second-order approximation to $\ln(1+\frac{x}{a})$ shows that when this probability is small, $\Delta_k^* \approx \frac{\mathcal{B}-p}{2f_0(k)} + \frac{\mathcal{B}-p}{2f_1(k)}$. This delivers a "small-bunching" approximation similar to one that has appeared in the literature (e.g. Kleven, 2016), and corresponds to the "excess mass" quantity in Chetty et al., 2011. When $f_0(k) \approx f_1(k)$ and $p = 0$, the bounds will tend to be narrower when $F_0(k)$ is closer to $(1-\mathcal{B})/2$, i.e. the kink is close to the median of the latent hours distribution.

1.4.4 Estimating policy relevant parameters

The buncher LATE yields an internally-valid answer to a particular causal question, among a well-defined subgroup of the population. Namely: how would hours among bunched units be affected by a counterfactual change from linear pay at the worker's straight-time wage to linear pay at their overtime rate? This section discusses how I use an estimate of this buncher LATE to both evaluate the overall ex-post effect of the FLSA on hours, as well as forecast the impacts of hypothetical changes to the FLSA. This requires some additional assumptions, which I continue to approach from a partial identification perspective.

[32] Since the bounds depend only on the CDFs at k and data local to k, point masses elsewhere in the distributions of h_0 and h_1 can be safely ignored provided that they are well-separated from the kink.

From the buncher LATE to the ex-post hours effect of the FLSA

To consider the overall ex-post hours effect of the FLSA among covered workers, I proceed in two steps. I first relate the buncher LATE to the average effect of introducing the overtime kink on all units, holding fixed the distributions of counterfactual hours h_{0it} and h_{1it}. Then, I allow straight-time wages to be affected by the FLSA, using the buncher LATE again to bound the additional effect of these wage changes on hours.

To motivate this strategy, let us first define the parameter of interest to be the difference in average weekly hours with and without the FLSA: $\theta := \mathbb{E}[h_{it}] - \mathbb{E}^*[h^*_{it}]$, where h^*_{it} indicates the hours unit it would work absent the FLSA, and the second expectation $\mathbb{E}^*$ is over the population of observational units of workers that would exist in the no-FLSA counterfactual–but would be eligible were it introduced.[33] I assume that the hours among workers who are hired because of the FLSA are not systematically different from those who would have existed anyways, so that we may rewrite θ as an average over individual-level effects in the actual population given the FLSA: $\theta = \mathbb{E}[h_{it} - h^*_{it}]$.

Next, I decompose this average effect as:

$$\theta = \mathbb{E}[h_{it}(w_{it}, \mathbf{h}_{-i,t}) - h_{0it}(w^*_{it}, \mathbf{h}^*_{-i,t})] = \mathbb{E}[\underbrace{h_{it}(w_{it}, \mathbf{h}_{-i,t}) - h_{0it}(w_{it}, \mathbf{h}_{-i,t})}_{\text{“effect of the kink”}}]$$
$$+ \mathbb{E}[\underbrace{h_{0it}(w_{it}, \mathbf{h}_{-i,t}) - h_{0it}(w^*_{it}, \mathbf{h}_{-i,t})}_{\text{“wage effects”}}] + \mathbb{E}[\underbrace{h_{0it}(w^*_{it}, \mathbf{h}_{-i,t}) - h_{0it}(w^*_{it}, \mathbf{h}^*_{-i,t})}_{\text{“interdependencies”}}], \quad (1.8)$$

where the notation makes explicit the dependence of h and h_0 on the worker's straight-time wage w_{it}, and possibly the hours $\mathbf{h}_{-i}$ of other workers in their firm. In the notation of the last section: $h_{it} = h_{it}(w_{it}, \mathbf{h}_{-i,t})$, $h_{0it} = h_{0it}(w_{it}, \mathbf{h}_{-i,t})$ and $h_{1it} = h_{1it}(w_{it}, \mathbf{h}_{-i,t})$; since pay is linear in hours in the no-FLSA counterfactual $h^*_{it} = h_{0it}(w^*_{it}, \mathbf{h}^*_{-i,t})$.

The first term in Equation (1.8) reflects the "effect of the kink" quantity $h_{it} - h_{0it}$ ex-

[33] Since the FLSA may itself change the population of workers who are covered by it (for instance, by encouraging the hiring of new covered workers), I first define θ as this difference in two population quantities. Note that h^*_{it} in this section differs from the "anticipated" hours quantity h^* in Section 1.2.

amined in Section 1.4.1, and is the primary object of interest. The second term reflects that straight-time wages w_{it} may differ from those that workers would face without the FLSA, denoted by w^*_{it}. The third term is zero when each worker's hours are chosen to solve a separate optimization problem, as in the benchmark model from Section 1.4.1 with linearly separable production. More generally however, it will capture interdependencies in hours across units, for instance due to non-separability in production. In Appendix A.3 I provide evidence that such effects do not play a large role in θ, and I do not attempt to account for them explicitly in estimation.

Turning first to the "effect of the kink" term, note that with straight-wages and the hours of other units fixed, the kink only has *direct* effects on those units working at least $k = 40$ hours:

$$h_{it} - h_{0it} = \begin{cases} 0 & \text{if} \quad h_{it} < k \\ k - h_{0it} & \text{if} \quad h_{it} = k \\ -\Delta_{it} & \text{if} \quad h_{it} > k \end{cases} \tag{1.9}$$

and thus $\mathbb{E}[h_{it} - h_{0it}] = \mathcal{B} \cdot \mathbb{E}[k - h_{0it}|h_{it} = k] - P(h_{it} > k)\mathbb{E}[\Delta_{it}|h_{it} > k]$. To identify this quantity we must extrapolate from the buncher LATE to obtain an estimate of $\mathbb{E}[\Delta_{it}|h_{it} > k]$, the average effect for units who work overtime. To do this, I assume that Δ_{it} of units working more than 40 hours are at least as large on average as those who work 40, but that the (reduced-form) *elasticity* of their response is no greater than that of the bunchers. The logic is that assuming a constant percentage change between h_0 and h_1 over units would imply responses that grow in proportion to h_1, eventually becoming implausibly large. On the other hand, it would be an underestimate to assume high-hours workers, say at 60 hours, have the same effect in levels $h_0 - h_1$ as those closer to 40.[34] To put bounds on the average effect of the kink among bunchers $\mathbb{E}[k - h_{0it}|h_{it} = k]$, I use the bi-log-concavity assumptions from Section 1.4.3. Details are provided in Supplemental Appendix A.8.

[34] In the benchmark model, constant treatment effects in levels corresponds to exponential production: $f(h) = \gamma(1 - e^{-h/\gamma})$ where $\gamma > 0$ and $h_{0it} - h_{1it} = \gamma \ln(1.5)$ for all units.

The "wage effects" term in Equation (1.8) arises because the straight-time wages observed in the data may reflect some adjustment to the FLSA, as we would expect on the basis of the conceptual framework in Section 1.2. While the "effect of the kink" term is expected to be negative, this second term will be positive if FLSA causes a reduction in the straight-time wages set at hiring on the basis of expected hours. However, both terms ultimately depend on the same thing: responsiveness of hours to the cost of an hour of work. I thus use the buncher LATE to compute an approximate upper bound on wage effects by assuming that all straight-time wages are adjusted according to Equation (1.1) with anticipated hours approximated by h_{it}, and an iso-elastic response. A lower bound on the "wage effects" term is zero. Supplemental Appendix A.8 gives the explicit formulas and provides a visual depiction of these definitions. Section 3.5 also reports results with and without this wage effect. The size of the wage effect $\mathbb{E}[h_{0it} - h^*_{0it}]$ is appreciable but still small in comparison with $\mathbb{E}[h_{it} - h_{0it}]$. This is because the average percentage wage change according to Equation (1.1) is fairly small near 40, where most of the mass is.

Forecasting the effects of policy changes

Apart from ex-post evaluation of the overtime rule, policymakers may also be interested in predicting what would happen if the parameters of overtime regulation were modified. Reforms that have been discussed in the U.S. include decreasing "standard hours" k at which overtime pay begins from 40 hours to 35 hours,[35] or increasing the overtime premium from time-and-a-half to "double-time" (Brown and Hamermesh, 2019).

I begin by considering changes to standard hours k. For now, I hold the distributions of h_0 and h_1 fixed across the policy change, and return to changes to the latent hours distributions at the end of this section. Inspection of Equation 1.4 reveals that as the kink is moved upwards, say from $k = 40$ hours to $k' = 44$ hours, some workers who were previously bunching at k now work h_{0it} hours: namely those for whom $h_{0it} \in [k, k']$. By

[35]Several countries have implemented changes to standard hours; Brown and Hamermesh (2019) provides a review of the evidence.

the same token, some individuals with values of $h_{1it} \in [k, k']$ now bunch at k'. Some individuals who were bunching at k may now bunch at k'—namely those workers for whom $h_{1it} \leq k$ and $h_{0it} \geq k'$. I assume that the mass of counterfactual bunchers p remains at $k = 40$ after the shift.[36] In the case of a reduction in overtime hours, say to $k' = 35$ this logic is reversed: some workers now work $h_{1it} \in [k', k]$, while workers with $h_{0it} \in [k', k]$ now bunch at k'. Figure 1.8 depicts both of these cases.

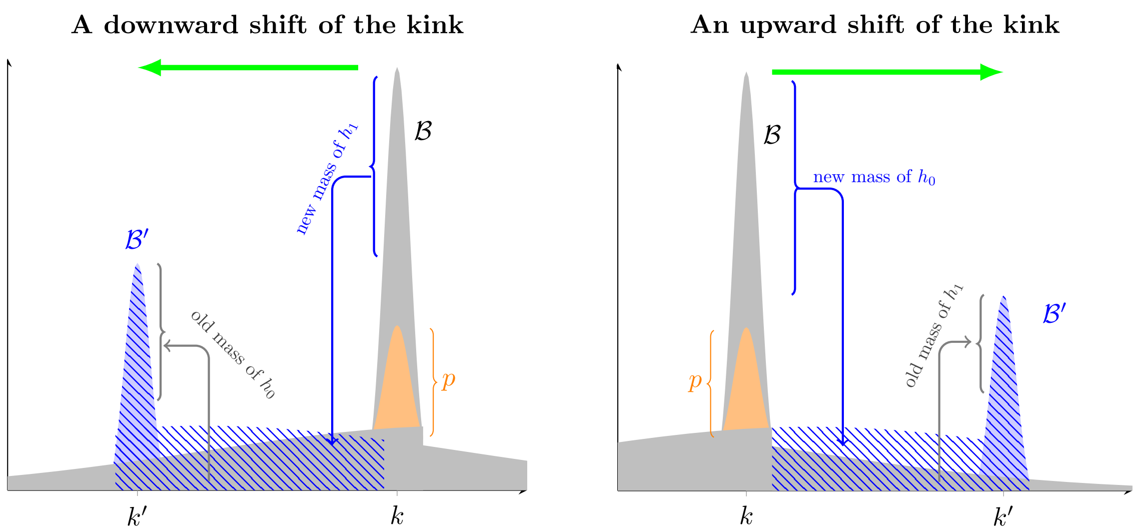

Figure 1.7: The left panel depicts a shift of the kink point downwards from k to k', while right panel depicts a shift of the kink point upwards. See text for details.

Quantitatively assessing a change to double-time pay requires us to move beyond the two counterfactual choices h_{0it} and h_{1it}: hours that would be worked under straight-wage and time-and-a-half. Let $h_{it}(\rho)$ be the hours that it would work if their employer faced a linear pay schedule at rate $\rho \cdot w_{it}$ (with both the straight-wage w_{it} and hours of other units fixed at their realized levels). In this notation, $h_{0it} = h_{it}(1)$ and $h_{0it} = h_{it}(1.5)$. Now consider a new overtime policy in which a premium pay factor of ρ_1 is required for hours in excess of k, e.g. $\rho_1 = 2$ for a "double-time" policy. Let $h_{it}^{[k,\rho_1]}$ denote realized hours

[36]It is conceivable that some or all counterfactual bunchers locate at 40 because it is the FLSA threshold, while still being non-responsive to the incentives introduced there by the kink. In this case, we might imagine that they would all coordinate on k' after the change. The effects here should thus be seen as short-run effects before that occurs.

under this overtime policy, and let $\mathcal{B}^{[k,\rho_1]} := P(h_{it}^{[k,\rho_1]} = k)$ the observable bunching that would occur.

Theorem 1.2 allows me to discuss the effects of small changes to k or ρ_1. Results for the effect of changing standard hours k make use of an explicit assumption that firm preferences are quasi-linear with respect to costs:

Assumption SEPARABLE. $\pi_{it}(z, \mathbf{x})$ *is additively separable and linear in* z*.*

I continue to assume that counterfactual bunchers $K_{it}^* = 1$ stay at $k^* := 40$, regardless of ρ and k. Let $p(k) = p \cdot \mathbb{1}(k = k^*)$ denote the possible mass of counterfactual bunchers as a function of k.

Theorem 1.2 (marginal comparative statics in the bunching design). *Under Assumptions CHOICE, CONVEX, SEPARABLE and SMOOTH:*

1. $\partial_k \left\{\mathcal{B}^{[k,\rho_1]} - p(k)\right\} = f_1(k) - f_0(k)$
2. $\partial_k \mathbb{E}[h_{it}^{[k,\rho_1]}] = \mathcal{B}^{[k,\rho_1]} - p(k)$
3. $\partial_{\rho_1} \mathbb{E}[h_{it}^{[k,\rho_1]}] = -\int_k^\infty f_{\rho_1}(h) \mathbb{E}\left[\frac{dh_{it}(\rho_1)}{d\rho} \middle| h_{it}(\rho_1) = h\right] dh$

Proof. See Appendix A.1. □

Assumption SMOOTH is a set of regularity conditions which imply that $h_{it}(\rho)$ admits a density $f_\rho(h)$ for all ρ – see Appendix A.1 for details. Theorem 1.2 also makes use of a stronger version of CHOICE that applies to all ρ, described therein.

Beginning from the actual FLSA policy of $k = 40, \rho_1 = 1.5$, the RHS of the first two objects above are point identified from the data, provided that p is known. Item 1 says that if the location of the kink is changed marginally, the bunching probability will change according to the difference between the densities of h_{1i} and h_{0i} at k^*, which are in turn equal to the left and right limits of the observed density $f(h)$ at the kink. This result is intuitive: given continuity of each potential outcome's density, a small increase in k will result in a mass proportional to $f_1(k)$ being "swept in" to the mass point at the kink, while a mass proportional to $f_0(k)$ is left behind. Item 2 aggregates this change in bunching with the changes to non-bunchers as k is increased. The $f_0(k)$ and $f_1(k)$ terms from the change in bunching end up being canceled, and the first- order effect of changing k is simply to transport the mass of inframarginal bunchers to the new value of k.[37] Making use of Theorem 1.2 for a discrete policy change like reducing standard hours to 35 requires integrating across the actual range of hypothesized policy variation. We lose point identification, but can use bi-log-concavity of the marginal distributions of h_0 and h_1 to retain bounds, as depicted by Figure 1.8.

Now consider the effect of moving from time-and-a-half to double time on average hours worked, in light of item 3. This scenario, similar to ex-post evaluation of the effect of the kink, requires making assumptions about the response of individuals who may locate far from the kink, and for whom the buncher LATE is less directly informative. Note that integrating item 3 over ρ we can write the average effect on hours from a move to double-time in terms of local average elasticities of response:

$$\mathbb{E}[h_{it}^{[k,\rho_1]} - h_{it}^{[k,\bar{\rho}_1]}] = \int_{\rho_1}^{\bar{\rho}_1} d\ln\rho \int_k^{\infty} f_\rho(h)h \cdot \mathbb{E}\left[\frac{d\ln h_{it}(\rho)}{d\ln\rho}\middle| h_{it}(\rho) = h\right] dh$$

Recall from the iso-elastic model that when the elasticity $\frac{d\ln h_{it}(\rho)}{d\ln\rho} = \frac{dh_{it}(\rho)}{d\rho}\frac{\rho}{h_{it}(\rho)}$ is constant across ρ and across units, it is partially identified. Just as an iso-elastic response is likely to overstate responsiveness at large $h_{it}(\rho)$, I argue it is likely to understate responsiveness to larger values of ρ, thus yielding a lower bound on the effect of moving to double-time. For an upper bound on the magnitude of the effect, I assume rather that in levels $\mathbb{E}[h_{it}(\rho_1) - h_{it}(\bar{\rho}_1)|h_{1it} > k]$ is at least as large as $\mathbb{E}[h_{0it} - h_{1it}|h_{1it} > k]$, and that the increase in bunching from a change of ρ_1 to $\bar{\rho}_1$ is as large as the increase from ρ_0 to ρ_1. I provide additional details in Supplemental Appendix A.8.

[37] Intuitively, in the limit of a small change in k bunchers who would choose exactly k under one of the two cost functions B_0 or B_1 cease to "bunch" as k moves to some $k' > k$, but they also do not change their realized value of h since the counterfactual hours choice that characterizes their new choice is equal to k.

In these calculations, I have held fixed the distributions of h_0 and h_1, which can be seen as describing the short-run before adjustment to straight-time wages or other factors that influence these latent hours distributions. In the empirical implementation I account for possible changes to straight wages when considering the average effects of policy changes on hours, as we saw with the ex-post effect of the FLSA. The effect of such corrections for the impact of changing k on the bunching probability is discussed in Section 1.6.

1.5 Implementation and Results

This section implements the empirical strategy described in the last section with the sample of administrative payroll data described in Section 1.3.

1.5.1 Identifying counterfactual bunching at 40 hours

Section 1.2 has argued that with wages fixed, the overtime kink should lead to bunching at 40 hours a week, while Section 1.4 has shown that this bunching is useful in identifying treatment effects and the impact of policy changes. However, there are other reasons to expect bunching at 40 hours. For one, 40 may be considered a status-quo choice by firms and/or workers, and it may be chosen even when it is not cost minimizing for the firm. It can also be important for firms to synchronize hours across workers, and thus have them coordinate on some number h^* of hours. Finally, for any salaried workers who were not successfully removed from the sample, firms may record the number of hours in a pay period as 40 even as actual hours worked vary.

In terms of the empirical strategy from Section A.1.2, all of these alternative explanations manifest in the same way: a point mass p at 40 in the distribution of hours that would occur even if workers were paid their straight-time wages for all hours. In the notation introduced in Section 1.4.3, these "counterfactual bunchers" are demarcated by $K^*_{it} = 1$; I refer to the $K^*_{it} = 0$ individuals who also locate at the kink as "active bunchers". The mass of active bunchers is $\mathcal{B} - p$. Theorem 1.1 shows that we can still partially identify the buncher LATE in the presence of counterfactual bunchers, so long as we know how many of the total bunchers are active and how many are counterfactual.

I leverage two strategies to provide plausible estimates for the mass of counterfactual bunchers p. My preferred estimate uses of the fact that when an employee is paid for hours that are not actually worked—including sick time, paid time off (PTO) and holidays—these hours do not contribute to the 40 hour overtime threshold of the FLSA. For example, if a worker applies PTO to miss a six hour shift, then they are not required to be paid overtime premium until they reach 46 total paid hours in that week, corresponding to 40 hours *worked*. These non-work hours thus shift the position of the kink in paid-hours.

The identifying assumption that I rely on is that individuals who still work 40 hours a week, even when they are paid for a positive number of non-work hours, are all active bunchers, and would not locate at forty hours in the counterfactuals h_{0it} and h_{1it}. This assumption reflects the idea that alternative reasons for bunching at 40 hours besides the overtime kink operate at the level of hours paid, rather than hours worked. Let n_{it} indicate non-worked hours for worker i in week t. Specifically, I make the following two assumptions:

1. $P(h_{it} = 40 | n_{it} > 0) = P(h_{it} = 40$ and $K^*_{it} = 0 | n_{it} > 0)$

2. $P(h_{it} = 40$ and $K^*_{it} = 0 | n_{it} > 0) = P(h_{it} = 40$ and $K^*_{it} = 0 | n_{it} = 0)$

The first item states that all of the individuals who locate at the kink, despite having a positive number of non-work hours are indeed active bunchers. I thus know the mass of active bunchers in the $n_{it} > 0$ conditional distribution of hours. The second item says that the $n_{it} > 0$ distribution is representative of the unconditional distribution, in the sense that the conditional mass of active bunchers does not vary based on whether non-work hours are positive or zero. Together, these two assumptions imply that $P(K^*_{it} = 0$ and $h_{it} = 40) = P(h_{it} = 40 | \eta_{it} > 0)$ and hence that $p = P(K^*_{it} = 1$ and $h_{it} = 40) = \mathcal{B} - P(h_{it} = 40 | \eta_{it} > 0)$.

I focus on paid time off as n_{it} because it is generally planned in advance, and has somewhat idiosyncratic timing. By contrast sick pay is often unanticipated, so the firm may not be able to re-optimize total hours within a week in which a worker calls in sick. Holiday pay is known in advance, holidays are unlikely to be representative in terms of product demand and other factors important for hours determination, threatening the second assumption.

Figure 1.8 shows the conditional distribution of hours paid for work when the paycheck contains a positive number of PTO hours ($n_{it} > 0$). The figure reveals that when moving from the unconditional (left panel) to positive-PTO conditional (right panel) distribution, most of the point mass at 40 hours moves away, largely concentrating now at 32 hours (corresponding to the PTO covering a single eight hour shift). Of the total bunching of $\mathcal{B} \approx 11.6\%$ in the unconditional distribution, I estimate that only about $P(h_{it} = 40|n_{it} > 0) \approx 2.7\%$ are active bunchers, leaving $p \approx 8.9\%$. Roughly three quarters of the individuals at 40 hours are counterfactual rather than active bunchers.

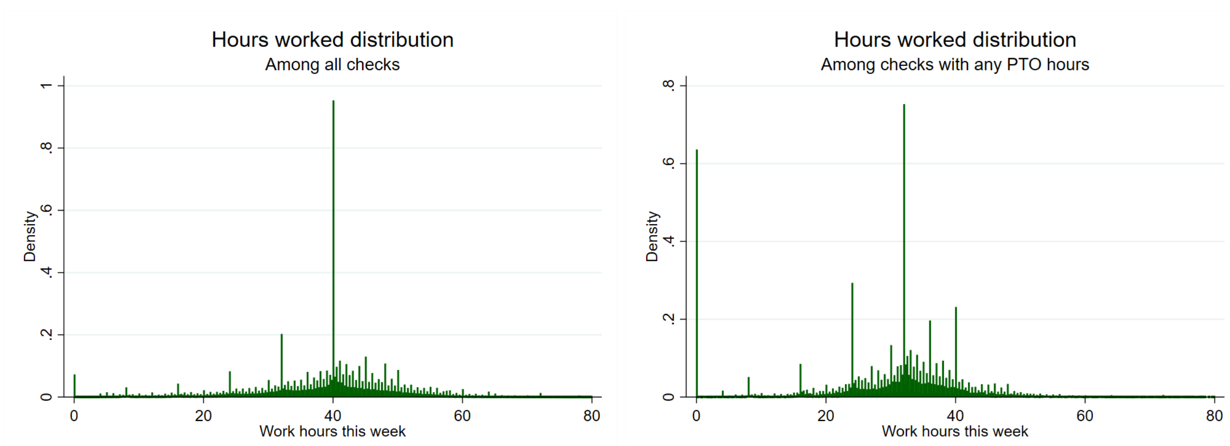

Figure 1.8: The right panel shows a histogram of hours worked when paid time off hours are positive. The left panel shows the unconditional distribution. Bin width is 1/8 hour.

As a secondary strategy, I estimate an upper bound for p by using the assumption that the potential outcomes of counterfactual bunchers are relatively immobile over time. The idea is that counterfactual bunchers have behavioral or administrative reasons for being at 40 hours, rather than 40 hours maximizing short run profits. I assume that these external considerations are fairly static over time, preventing latent hours h_{0it} from changing much between adjacent pay periods. In particular, assume that in a given period t nearly all of the counterfactual bunchers are also non-movers from $t-1$, i.e.

$$p = P(h_{0it} = 40) \approx P(h_{0it} = h_{0it-1} = 40) \leq P(h_{it} = h_{i,t-1} = 40)$$

where the inequality follows from $h_{0it} = 40 \implies h_{it} = 40$ by Lemma A.1. The probability $P(h_{it} = h_{i,t-1} = 40)$ can be directly estimated from the data, yielding $p \leq 6\%$.

1.5.2 Estimation and inference

Estimating bounds on the buncher LATE requires estimates of the CDF and density of hours worked, and in particular right and left limits of these objects at the kink. I use the local polynomial density estimator of Cattaneo, Jansson and Ma (2020) (CJM), which is well suited to estimating a CDF and its derivatives at boundary points. I work with the pooled distribution of paychecks over the full study period. The CJM estimator provides a smoothed estimate of the left limit of the CDF and density at k as:

$$(\hat{F}_{-}(k), \hat{f}_{-}(k)) = \underset{(b_1,b_2)}{\operatorname{argmin}} \sum_{i:h_{it}<k} (F_n(h_{it}) - b_1 - b_2 h_{it})^2 \cdot K\left(\frac{h_{it}-k}{h}\right) \tag{1.10}$$

where $F_n(y) = \frac{1}{n}\sum_{it} \mathbb{1}(h_{it} \leq y)$ is the empirical CDF function, $K(\cdot)$ is a kernel function, and h is a bandwidth. I use a triangular kernel, and choose h as follows: first, I use CJM's mean-squared error minimizing bandwidth selector to produce a bandwidth choice using the data on either side of $k = 40$ (for the left and right limits, respectively). I then average the two bandwidths, and use this as the bandwidth in the final calculation of both the right and left limits, to mitigate any dependence of the estimates on a differential bandwidth choice for each side. In the full sample, the bandwidth chosen by this procedure is about 1.7 hours, and is somewhat larger for subsamples that condition on a single industry.

To construct confidence intervals for parameters that are partially identified (e.g. the buncher LATE), I use the adaptive critical values proposed by Imbens and Manski (2004) and Stoye (2009) that are valid for the underlying parameter. In all cases, estimators of bounds or point identified quantities are functions of inputs that are $\sqrt{n}$-asymptotically normal.[38] To easily incorporate sampling uncertainty in both $(\hat{F}_-(k), \hat{f}_-(k), \hat{F}_+(k), \hat{f}_+(k))$ and in $\hat{p}$, I estimate the variances by a cluster non-parametric bootstrap that resamples at the firm level. This allows arbitrary autocorrelation in hours across pay periods for a single worker, and between workers within a firm. All standard errors use 500 bootstrap replications.

1.5.3 Results of the bunching estimator

Table 1.2 reports treatment effect estimates $h_{0it} - h_{1it}$, in a sample that pools across all industries, when p is either assumed zero or estimated by one of the two methods described in Section 1.5.1. The first row yields an estimate of the net bunching probability $\mathcal{B} - p$, while the second row reports the bounds on the buncher LATE $\mathbb{E}[h_{0it} - h_{1it} | h_{it} = k]$ based on bi-log-concavity. Within a fixed estimate of p, the bounds on the buncher LATE are quite informative: the upper and lower bounds are always close to each other and precisely estimated. Appendix A.5 reports estimates based on alternative shape constraints and assumptions about effect heterogeneity, which deliver similar results.[39]

The PTO-based estimate of p provides the most conservative treatment effect estimates, attributing roughly one quarter of the observed bunching to active rather than counterfactual bunchers. Nevertheless, this estimate still yields a highly statistically significant buncher LATE of about 2/3 of an hour, or 40 minutes. This estimate says that individuals who in fact work 40 hours given the overtime kink in a given pay period would work

[38] For the effect of changing the kink point, I censor CDF estimates at zero and one. In principle, this could undermine asymptotic normality, but these constraints are not typically binding so I ignore this issue.

[39] In particular, I present a point estimate based on Appendix Proposition A.1, which assumes that treatment effects are constant and that the density is linear in the missing region, as well as results under a weaker assumption that the density is monotonic in the missing region. Monotonicity is not likely to hold in the overtime context, since the kink appears to be located at the mode of both the h_0 and h_1 distributions. Nevertheless, the bounds based on monotonicity do not deliver vastly different results.

about 40 minutes more that week in a world in which they were paid their straight-time wage for all hours, compared with a world in which they were paid 1.5 times this wage for all hours. On the other side of the spectrum, if all of the observed bunching mass is attributed to active bunchers, corresponding to $p = 0$, then the estimated buncher LATE suggests a difference of at least 2.6 hours. The next section expresses these estimates as elasticities, by making the bi-log-concavity assumption on the distribution of log hours rather than hours.[40] In Appendix Table A.5 I report estimates of the buncher LATE for each of the largest industries in the sample, and also present estimates as a function of the assumed mass p of counterfactual bunchers at 40 hours.

	p=0	p from non-changers	p from PTO
Net bunching:	0.116	0.057	0.027
	[0.112, 0.120]	[0.055, 0.058]	[0.024, 0.030]
Buncher LATE	[2.614, 3.054]	[1.324, 1.435]	[0.640, 0.666]
	[2.493, 3.205]	[1.264, 1.501]	[0.574, 0.736]
Num observations	630217	630217	630217
Num clusters	566	566	566

Table 1.2: Estimates of net bunching $\mathcal{B} - p$ and the buncher LATE: $\Delta_k^* = \mathbb{E}[h_{0it} - h_{1it}|h_{it} = k, K_{it}^* = 0]$, across various strategies to estimate counterfactual bunching $p = P(K_{it}^* = 1)$. Unit of analysis is a paycheck, and 95% bootstrap confidence intervals (in gray) are clustered by firm.

[40] Appendix Table A.11 also shows estimates based on constant treatment effects in logs and monotonicity or linear interpolation.

1.5.4 Estimates of policy effects

I now use estimates of the buncher LATE to estimate the overall causal effect of the FLSA overtime rule, as well as simulate changes based on modifying standard hours or the premium pay factor. Table 1.3 reports an estimate of the buncher LATE expressed as a reduced form elasticity,[41] which I use as an input in these calculations. The next two rows report bounds on $\mathbb{E}[h_{it} - h^*_{it}]$ and $\mathbb{E}[h_{it} - h^*_{it}|h_{1it} \geq 40, K^*_{it} = 0]$, respectively. The first of these is the overall ex-post effect of the FLSA on hours, averaged over both workers and pay periods, while the second conditions on paychecks for which the FLSA premium has a direct effect (those reporting at least 40 hours aside from counterfactual bunchers). The final row reports an estimate of the effect of moving to double-time pay, also including a correction term to account for possible wage changes. I provide details of the calculations in Supplemental Appendix A.8.

Taking the PTO-based estimate of p as a lower bound on responsiveness, the estimates suggest that FLSA eligible workers work at least 1/5 of an hour less in any given week than they would absent overtime regulation: about one third the magnitude of the buncher LATE in levels. When I focus on those eligible workers that are directly affected in a given week, the figure is about twice as high: roughly 30 minutes. I estimate that a move to double-time pay would introduce a further reduction that may be comparable to the existing overall ex-post effect, but with substantially wider bounds. These estimates include the effects of possible adjustments to straight-time wages, which tend to attenuate the effects of the policy change. Appendix Table A.12 replicates Table 1.3 neglecting these wage adjustments, which might be viewed as a short-run response to the FLSA before wages have time to adjust.

Figure 1.9 breaks down estimates of the ex-post effect of the kink by major industry, revealing considerable heterogeneity between industries. The estimates suggest that the

[41] This is $\hat{\Delta}^*_k/(40\ln(1.5))$ where $\hat{\Delta}_k$ is the estimate of the buncher LATE presented in Table 1.2, which is numerically equivalent to the elasticity implied by the buncher LATE in logs $\mathbb{E}[\ln h_{0it} - \ln h_{1it}|h_{it} = k, K^*_{it} = 0]/(\ln 1.5)$ estimated under assumption that $\ln h_0$ and $\ln h_1$ are BLC.

	p=0	p from non-changers	p from PTO
Buncher LATE as elasticity	[-0.188,-0.161]	[-0.088,-0.082]	[-0.041,-0.039]
	[-0.198,-0.154]	[-0.093,-0.078]	[-0.045,-0.035]
Average effect of FLSA on hours	[-1.466, -1.026]	[-0.727, -0.486]	[-0.347, -0.227]
	[-1.535, -0.977]	[-0.762, -0.463]	[-0.384, -0.203]
Avg. effect among directly affected	[-2.620, -1.833]	[-1.453, -0.972]	[-0.738, -0.483]
	[-2.733, -1.750]	[-1.518, -0.929]	[-0.812, -0.434]
Double-time, average effect on hours	[-2.604, -0.569]	[-1.239, -0.314]	[-0.580, -0.159]
	[-2.707, -0.547]	[-1.285, -0.300]	[-0.638, -0.143]

Table 1.3: Estimates of the buncher LATE expressed as an elasticity, the average ex-post effect of the FLSA $\mathbb{E}[h_{it} - h^*_{it}]$,[41] the effect among directly affected units $\mathbb{E}[h_{it} - h^*_{it} | h_{it} \geq k]$ and predicted effects of a change to double-time. 95% bootstrap confidence intervals in gray, clustered by firm.

industries Real Estate & Rental and Leasing as well as Wholesale Trade see the highest average reduction in hours. The least-affected industries are Health Care and Social Assistance and Professional Scientific and Technical, with the average worker working just about 6 minutes less per week. Appendix Figure A.8 compares the hours distribution for Real Estate & Rental and Leasing with the distribution for of Professional Scientific and Technical, showing that the difference in their effects can be explained by $\mathcal{B} - p$ being larger for Real Estate & Rental and Leasing, while the density of hours close to the kink is smaller. Appendix Table A.6 reports numerical values as well as estimates based on assuming all of the bunching is due to the FLSA. Appendix A.5 reports estimates broken down by gender, finding that the FLSA has considerably higher effects on the hours of men.

Figure 1.10 looks at the effect of changing the threshold for overtime hours k from 40 to alternative values k'. The left panel reports estimates of the identified bounds on $\mathcal{B}^{[k',\rho_1]}$ as well as point-wise 95% confidence intervals (gray) across values of k' between 35 and 45, for each of the three approaches to estimating p. In all cases, the upper bound

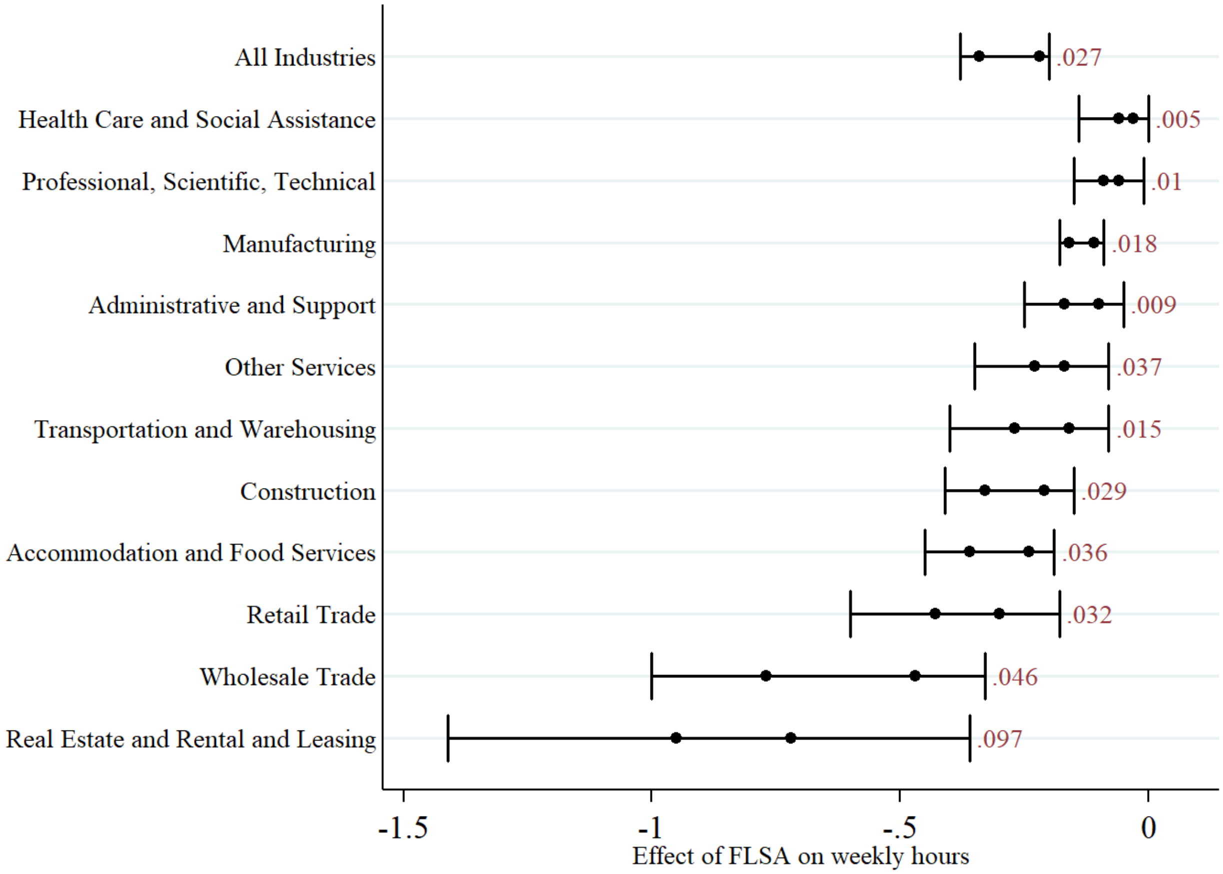

Figure 1.9: 95% confidence intervals for the effect of the FLSA overtime rule on hours by industry, using PTO-based estimates of p for each. Dots are point estimates of the upper and lower bounds. The number to the right of each range is the point estimate of the net bunching $\mathcal{B} - p$ for that industry.

on bunching approaches zero as k' is moved farther from 40. This is sensible if the h_0 and h_1 distributions are roughly unimodal with modes around 40: straddling of potential outcomes becomes less and less likely as one moves away from where most of the mass is. Appendix A.11 shows these bounds as k' ranges all the way from 0 to 80, for the $p = 0$ case. Since these estimates do not account for adjustment to straight-time wages, they should be viewed as short-run responses.

When p is estimated using PTO or non-changers between periods, we see that the upper bound of the identified set for $\mathcal{B}^{[k',\rho_1]}$ in fact reaches zero quite quickly. Moving standard errors to $k' = 35$ is predicted to completely eliminate bunching due to the overtime kink in the short run, before any adjustment to latent hours (e.g. through changes to straight-time wages). The right panel of Figure 1.10 shows estimates for the average effect on hours of changing k, inclusive of wage effects (see Appendix C.4 for details). Increases to k cause an increase in hours, as overtime policy becomes less stringent, and reductions to k reduce hours. The actual size of these effects ais not well-identified for changes larger than a couple of hours, however the range of statistically significant effects depends on p. Even for the preferred estimate of p from PTO, increasing the overtime threshold as high as 43 hours is estimated to increase average working hours by an amount distinguishable from zero.

1.6 Implications of the estimates for overtime policy

The estimates from the preceding section suggest that FLSA regulation indeed has real effects on hours worked, in line with labor demand theory when wages do not fully adjust to absorb the added cost of overtime hours. When averaged over affected workers and across pay periods, I find that hourly workers in my sample work at least 30 minutes less per week than they would without the overtime rule. A less conservative estimate of the bunching caused by the FLSA suggests the effect is between 1 and 1.5 hours. My preferred estimate of about half an hour is broadly comparable to the few causal estimates that exist in the literature, including Hamermesh and Trejo (2000) who assess the effects of expanding California's daily overtime rule to cover men in 1980, and Brown and Hamermesh (2019) who use the erosion of the real value of FLSA exemption thresholds over the last several decades.[42] By contrast, my estimates carry the strengths of an approach to identification that does not require a natural experiment, and use much more recent data.

From the perspective of a typical worker, a decrease in working hours of 30 minutes

[42]Hamermesh and Trejo (2000) and Brown and Hamermesh (2019) report estimates of -0.5 and -0.18 for the elasticity of overtime hours with respect to the overtime rate. My preferred estimate of -0.04 for the buncher LATE as an elasticity is the elasticity of *total* hours, including the first 40. An elasticity of overtime hours can be computed by multiplying this by the ratio of mean hours to mean overtime hours in the sample, resulting in an estimate of roughly -0.45.

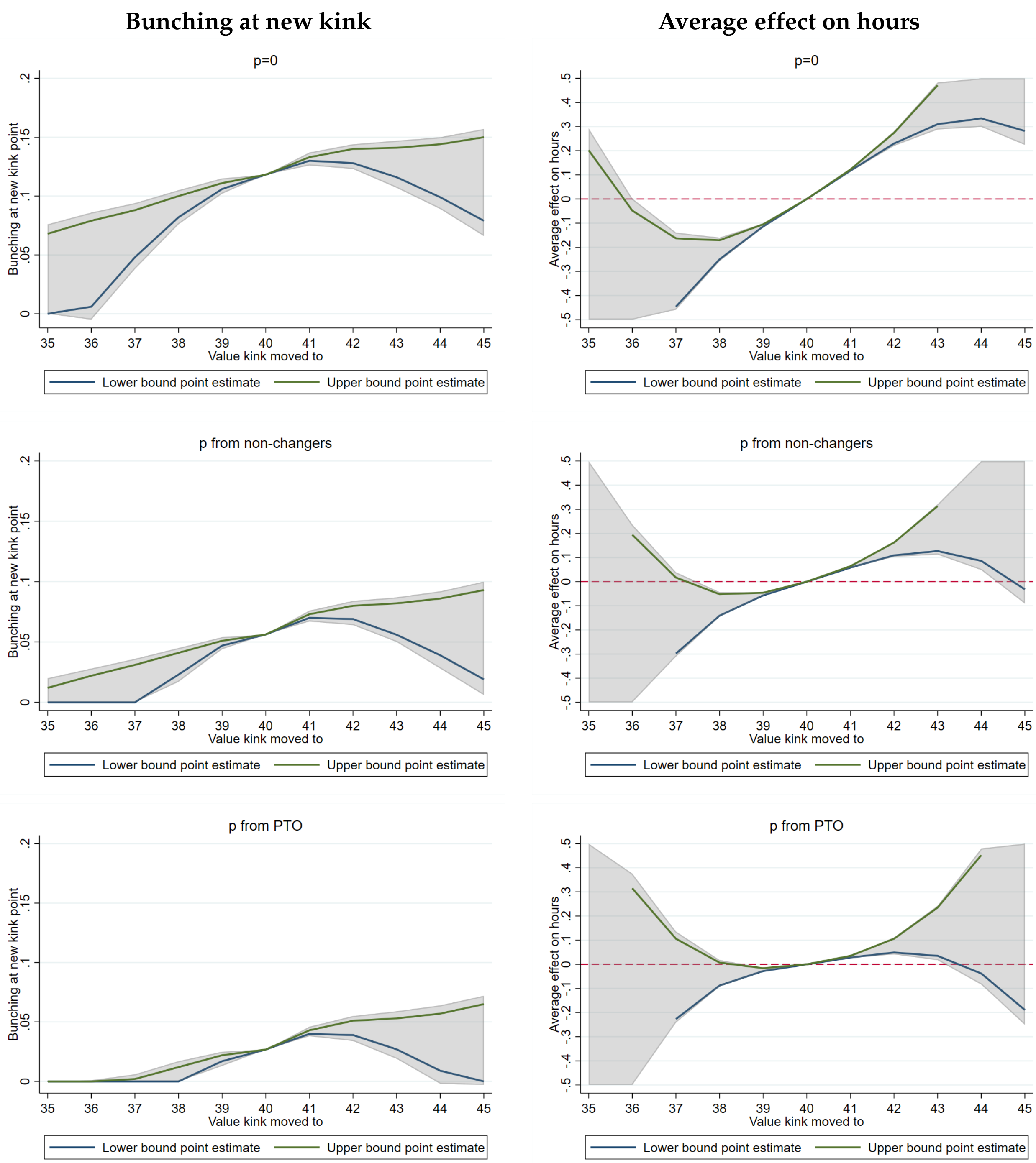

Figure 1.10: Bounds for the bunching that would exist at standard hours k if it were changed from 40 (left panel), as well as for the impact on average hours (right panel). Bounds of the effect on hours are clipped to the interval $[-0.5, 0.5]$ for visibility. Pointwise bootstrapped 95% confidence intervals, cluster bootstrapped by firm, are shaded gray.

per week may seem modest, but the overall effect of the policy could be quite large. The data suggest that at least about 3% and as many as about 11% of workers' hours are adjusted to the threshold introduced by the policy, indicating that the policy may have significant distortionary impacts. But the policy may also have quite substantial effects on unemployment. While a full assessment of the employment effects of the FLSA overtime rule is beyond the scope of this paper, the hours effects estimated here can be used to construct some back-of-the-envelope calculations.

If the average FLSA eligible worker works approximately 1/3 of an hour less per week because of the rule, hours per worker are reduced by just under 1% on average. If we ignore scale effects of the overtime rule on the total number of labor hours in FLSA-eligible jobs, this would suggest that employment among such jobs is 1% higher than it would be without the overtime premium. This serves as an upper bound, since overall hours worked may decrease due to overtime regulation. Hamermesh (1996) proposes a simple adjustment, based on assuming a value for the rate at which firms substitute labor for capital based on their relative prices, and the possibility of offsetting labor supply effects. In particular, the adjustment assumes the percentage change in employment is $\Delta \ln E|_{EH} - \eta \cdot \Delta \ln LC \cdot \frac{\eta}{\alpha - \eta}$ where η is a constant-output demand elasticity for labor (rather than capital), α is a labor supply elasticity, and $\Delta \ln LC$ is the percentage change in total labor costs from the introduction of the FLSA. Here $\Delta \ln E|_{EH}$ is the quantity implied by my estimates: the percentage change in employment that would occur were the total number of worker-hours EH unchanged.

Using plausible values from Hamermesh (1996) for the remaining parameters yields 0.17 percentage points for the substitution term $\eta \cdot \Delta \ln LC \cdot \frac{\eta}{\alpha - \eta}$, suggesting that the effect of the FLSA is attenuated from roughly 0.87 percentage points to about a 0.70 percentage point net increase in employment. This would represent about 700,000 jobs, assuming 100 million FLSA eligible workers. A reasonable range of parameter values rules out negative overall employment effects from the FLSA.[43] I can also put an overall upper bound on the size of employment effects, by attributing all of the bunching at 40 to the FLSA and assuming the total number of worker-hours is not reduced at all. By this estimate the FLSA increases employment by at most 3 million jobs, or 3% among covered workers.

This paper has also considered the likely effects of adjusting the two parameters that characterize the FLSA overtime rule: standard hours and the overtime premium factor. The effect of moving to double-pay for overtime is not as precisely identified as the ex-post effect of the FLSA, but estimates suggest an average additional effect on hours that is at least as large as the effect of the current FLSA regulation. I also find that moving time-and-a-half overtime pay to begin at 35 rather than 40 hours would nearly eliminate bunching due to the FLSA, given workers' current wages.[44] While my short run prediction under this policy counterfactual assumes away changes to straight-time wages, the reduction in bunching is likely to remain after allowing such adjustment over time. With 35 already to the left of the mode of the latent hours distributions h_{0i} and h_{1i}, it would become even further from the mode as these distributions move rightward due to lower wages. Moving the overtime premium away from the mode of the distribution of these latent hours choices may thus lead to efficiency benefits that are persistent over time.

1.7 Conclusion

This paper has analyzed the effects of U.S. overtime policy on hours worked by adapting the bunching-design method to address itself to questions of causal inference. While structural models of choice can help interpret estimates that use bunching at a kink, I have shown that the basic identifying power of the bunching design is robust to a variety of underlying choice models and functional form assumptions. Across such choices, the

[43] These "best-guess" values are $\eta = -0.3$, $\alpha = 0.1$, and $\Delta \ln LC$ calibrated assuming 80% of labor costs come from wages with overtime representing 2% of total hours. Generating a negative overall employment response by assuming higher substitution to capital requires $\eta = -1.25$, well outside of empirical estimates.

[44] Estimates of the average hours effect for changes to standard hours are consistent with estimates by Costa (2000), that hours fell by 0.2-0.4 on average during the phased introduction of the FLSA in which standard hours declined by 2 hours in 1939 and 1940.

identified parameter of interest is an appropriately-defined average treatment effect between two counterfactual choices, making the method useful for reduced-form program evaluation. This also opens the door to applying the bunching design in a broader variety of contexts, beyond those in which the researcher is prepared to posit a parametric model of decision-makers' preferences.

By leveraging these insights with a new payroll dataset recording exact weekly hours paid at the individual level, I estimate that U.S. workers subject to the FLSA indeed work shorter hours due to the overtime rule, which may lead to substantial employment effects. Given the large amount of within-worker variation in hours observed in the data, the modest size of the FLSA effects estimated in this paper suggest that firms do face significant incentives to maintain longer working hours, countervailing against the ones introduced by policies intended to reduce them.

Chapter 2: A Vector Monotonicity Assumption for Multiple Instruments

2.1 Introduction

The local average treatment effects (LATE) framework introduced by Imbens and Angrist (1994) allows for causal inference with arbitrary heterogeneity in treatment effects, but in doing so imposes an important form of homogeneity on selection behavior. This homogeneity comes through the LATE *monotonicity* assumption, which is often quite natural to make when the researcher has a single instrumental variable at their disposal. However with multiple instruments, this traditional monotonicity assumption can become hard to justify—a point that has recently been emphasized by Mogstad et al. (2020b).

A natural question is whether causal effects are still identified when monotonicity holds on an instrument-by-instrument basis, what I call *vector monotonicity*. Vector monotonicity (VM) captures the notion that each instrument has an impact on treatment uptake in a direction that is common across units (and typically known ex-ante by the researcher). For example, two instruments for college enrollment might be: i) proximity to a college; and ii) affordability of nearby colleges. It is reasonable to assume that each instrument induces some individuals towards going to college, while discouraging none. This contrasts with traditional LATE monotonicity, which as I describe below requires that either proximity or affordability effectively dominates in selection behavior for all units.

In this paper I provide a simple approach to estimating causal effects under vector monotonicity. I first show that in a setting with a binary treatment and any number of binary instruments satisfying VM, average treatment effects can be point identified for subgroups of the population that satisfy a certain condition. The condition is met by, for example, the group of all units that move into treatment when any fixed subset of the instruments are switched "on". As special cases, this includes for example those units that respond to a movement of a single particular instrument, or those units that have any variation whatsoever in counterfactual treatment status given the available instruments. I show how general discrete instruments can be accommodated by re-expressing them as a larger number of binary instruments, while preserving vector monotonicity. I then propose a simple two-step estimator for the identified causal parameters. The estimator is scalable, involving the same computational burden as 2SLS despite the rapid proliferation of possible selection patterns compatible with VM as the number of instruments increases.

To appreciate the sense in which traditional LATE monotonicity is restrictive with multiple instruments, consider the two instruments for college mentioned above, with each coded as a binary variable ("far"/"close" and "cheap"/"expensive"). LATE monotonicity says that a counterfactual change to the proximity and/or tuition instruments can either move some students into college attendance, or some students out, but not both. In particular, this requires that all units who would go to college when it is far but cheap would also go to college if it was close and expensive, or that the reverse is true. We would generally expect this implication to fail if individuals are heterogeneous in how much each of the instruments "matters" to them: for example, if some students are primarily sensitive to distance and others are primarily sensitive to tuition. Vector monotonicity instead says something quite natural in this context: proximity to a college weakly encourages college attendance, regardless of price, and lower tuition weakly encourages college attendance, regardless of distance.

In a set of papers developed concurrently with this one, Mogstad, Torgovitsky and Walters (2019; 2020a; 2020b) (henceforth MTW) underline the above difficulty for LATE monotonicity with multiple instruments, and introduce a weaker assumption of *partial monotonicity* (PM). PM is similar to VM but allows the direction of "compliance" for each instrument to depend on the values of the other instruments: for instance, college proximity could encourage attendance when nearby colleges are cheap but discourage attendance when they are expensive. Such reversals are empirically testable, and VM is thus not stronger in practice than PM but may be falsified in a given empirical setting.[1] Given a desired target parameter, Mogstad et al. (2020a) develop a marginal treatment effect approach to identification under PM, in which partial identification is the generic case (though parametric assumptions or continuous instruments may aid in obtaining point identification). The present paper takes the reverse approach: in settings where VM holds, what causal parameters are *point* identified without requiring any additional assumptions? I establish that such a class of causal parameters indeed exists, and contains easily interpretable and policy-relevant treatment effect parameters. Further, I argue that VM typically holds given PM, making these identification results empirically relevant.

The estimator proposed in this paper can be seen as an alternative to two-stage-least-squares (2SLS), which has been the typical method to make use of multiple instruments in applied work. 2SLS is known to identify a convex combination of local average treatment effects under the standard LATE assumptions provided that the first stage recovers the propensity score function, but this implication does not hold under VM or PM. By contrast, my estimator is guaranteed to be consistent for the particular chosen parameter of interest. MTW derive additional testable conditions which are sufficient for the 2SLS estimand to deliver a convex combination of treatment effects under PM, though the number of conditions to be verified generally grows combinatorially with the number of instruments. In the Supplemental Material,[2] I consider two special cases in which *linear* 2SLS will uncover averages of causal effects under VM with binary instruments. A sufficient condition for one of the special cases – that the instruments are independent – is straightforward to test empirically. The other special case assumes that each unit is responsive to the value of one instrument only, and is quite restrictive. My main identification result eliminates the need to rely on such additional assumptions.

[1] Nevertheless, VM has additional identifying power: when it holds a larger class of parameters are point identified compared with when PM alone holds. See Proposition 2.9.

[2] Supplemental Material is available here: http://www.columbia.edu/~ltg2111/resources/vm_externalappendix.pdf.

A growing literature has considered extensions to the basic LATE model of Imbens and Angrist (1994), but has typically not emphasized the distinction between separate instruments, when more than one is available. Natural analogs of LATE monotonicity have been studied for treatments that are discrete (Angrist and Imbens, 1995), continuous (Angrist et al., 2000), or unordered (Heckman and Pinto, 2018). Other papers have considered identification under various violations of LATE monotonicity. In the case of a binary treatment, Gautier and Hoderlein (2011), Lewbel and Yang (2016) and Gautier (2020) consider various explicit selection models, while Chaisemartin (2017) shows that a weaker notion than monotonicity can be sufficient to give a causal interpretation to LATE estimands.[3] Lee and Salanié (2018) relax monotonicity in a setting with multivalued treatment and continuous instruments, generalizing results from the local instrumental variables approach of Heckman and Vytlacil (2005). With discrete instruments, Lee and Salanié (2020) show that a notion of particular instrument values "targeting" particular values of a multivalued treatment carries additional identifying power.

In Section 2.2 I discuss the basic setup and definitions. I compare vector monotonicity to the traditional monotonicity assumption and MTW's proposal of partial monotonicity, and discuss examples in the context of a simple choice model. In Section 2.3, I show that like conventional monotonicity, VM partitions the population into well-defined "response groups" that can coexist in arbitrary proportions. I characterize these groups in a setting with any number of binary instruments, nesting a description from MTW of the two-instrument case. In Section 2.4 I use this taxonomy to demonstrate identification of a family of causal parameters, and Section 2.5 proposes corresponding estimators. Section 2.6 reports results from an application to the labor market returns to schooling. In appendices, I consider a generalization of the identification result that relaxes an assumption of rectangular support among the instruments, consider identification with covariates,

[3] LATE monotonicity is also generally not assumed by nonseparable triangular models with endogeneity (e.g. Imbens and Newey 2009, Torgovitsky 2015, D'Haultfœuille and Février 2015, Gunsilius 2020, Feng 2020), which typically impose some version of monotonicity in unobserved heterogeneity.

and additional results regarding the proposed estimator, including a data-driven regularization procedure to improve its performance in small samples. In online Supplemental Material, I also consider some special cases in which linear 2SLS identifies a convex combination of treatment effects under VM, and provide additional examples pertaining to the main text, including a second empirical application to the labor supply effects of family size.

2.2 Setup

Here I fix notation and formalize the basic setup in which a researcher has multiple instrumental variables for a single binary treatment. Within this framework, I contrast the three alternative notions of monotonicity mentioned in the introduction.

Consider a setting with a binary treatment variable D, scalar outcome variable Y, and vector $Z = (Z_1 \dots Z_J)$ of J instrumental variables that can take values in set $\mathcal{Z} \subseteq (\mathcal{Z}_1 \times \mathcal{Z}_2 \times \dots \times \mathcal{Z}_J)$, where $\mathcal{Z}_j$ is the set of values that instrument Z_j can take.[4]

Definition 2.1 (potential outcomes and treatments). *Let $D_i(z)$ denote the treatment status of unit i when their vector of instrumental variables takes value $z \in \mathcal{Z}$, and $Y_i(d, z)$ the realization of the outcome variable that would occur with treatment status $d \in \{0, 1\}$ and instrument value $z \in \mathcal{Z}$.*

The following assumption states that the available instrumental variables are valid:

Assumption 1 (exclusion and independence). *a) $Y_i(d, z) = Y_i(d)$ for all $z' \in \mathcal{Z}, d \in \{0, 1\}$; and b)*

$$(Y_i(1), Y_i(0), \{D_i(z)\}_{z \in \mathcal{Z}}) \perp (Z_{1i}, \dots, Z_{Ji})$$

The first part of Assumption 1 states that the instruments are "excludable" from the outcome function in the sense that potential outcomes do not depend on them once treatment

[4] $\mathcal{Z}$ may be a strict subset of $(\mathcal{Z}_1 \times \mathcal{Z}_2 \times \dots \times \mathcal{Z}_J)$ when certain combinations of instrument values are ruled out on conceptual grounds, e.g. Z_1 indicates a mothers' first two births being girls and Z_2 indicates them both being boys.

status is fixed. The second part of Assumption 1 states that the instruments are independent of potential outcomes and potential treatments. In practice, it is common to maintain a version of this independence assumption that holds only conditional on a set of observed covariates. For ease of exposition, I implicitly condition on any such covariates throughout, then consider incorporating them explicitly in Appendix B.2 and in the empirical application

2.2.1 Notions of monotonicity

It is well-known that when treatment effects are heterogeneous, Assumption 1 alone is not sufficient for instrument variation to identify treatment effects. The seminal LATE model of Imbens and Angrist (1994) introduces the additional assumption of monotonicity:

Assumption IAM (traditional LATE monotonicity). *For all $z, z' \in \mathcal{Z}$: $D_i(z) \geq D_i(z')$ for all i or $D_i(z) \leq D_i(z')$ for all i.*

I follow the terminology of MTW and henceforth refer to this as Assumption IAM, or "Imbens and Angrist monotonicity". As pointed out by Heckman et al. (2006), IAM can be thought of as a type of uniformity assumption: it states that flows of selection into treatment between z in z' move only in one direction, whichever direction that is.[5]

The proposed assumption of *vector monotonicity* captures monotonicity as the notion that "increasing" the value of any instrument weakly encourages (or discourages) all units to take treatment, regardless of the values of the other instruments:

Assumption 2 (vector monotonicity). *There exists an ordering $\geq_j$ on $\mathcal{Z}_j$ for each $j \in \{1 \dots J\}$ such that for all $z, z' \in \mathcal{Z}$, if $z \geq z'$ component-wise according to the $\{\geq_j\}$, then $D_i(z) \geq D_i(z')$ for all i.*

[5]Note that the two instances of "for all i" appearing in IAM can be replaced by "almost surely", without affecting identification results. The definitions given for VM and PM can also be slightly weakened in this way.

Vector monotonicity is referred to as "actual monotonicity" by Mogstad et al. (2020b), when each $\geq_j$ is the standard ordering on real numbers. Mountjoy (2019) imposes a version of VM in a case with a multivalued treatment and continuous instruments.

The partial monotonicity assumption introduced by MTW is weaker than both IAM and VM. Let (z_j, z_{-j}) denote a vector composed of $z_j \in \mathcal{Z}_j$ and $z_{-j} \in \mathcal{Z}_{-j}$, where $\mathcal{Z}_{-j}$ indicates the set of values that the vector of all instruments but Z_j can take.

Assumption PM (partial monotonicity). *For each $j \in \{1 \dots J\}$, $z_j, z'_j \in \mathcal{Z}_j$, and $z_{-j} \in \mathcal{Z}_{-j}$ such that $(z_j, z_{-j}) \in \mathcal{Z}$ and $(z'_j, z_{-j}) \in \mathcal{Z}$, either $D_i(z_j, z_{-j}) \geq D_i(z'_j, z_{-j})$ for all i or $D_i(z_j, z_{-j}) \leq D_i(z'_j, z_{-j})$ for all i.*

Note that under partial monotonicity, there will be a weak ordering on the points in $\mathcal{Z}_j$, for any fixed choice of j and z_{-j}. The crucial restriction made by vector monotonicity beyond partial monotonicity is that under VM, this ordering must be *the same* across all values of $z_{-j} \in \mathcal{Z}_{-j}$ for a given j. Partial monotonicity allows, for instance, a situation in which college proximity encourages attendance when nearby colleges are cheap but discourages attendance when they are expensive—while VM could not.

An alternative characterization of VM makes this relationship to PM more explicit. Call $\mathcal{Z}$ *connected* when for any two $z, z' \in \mathcal{Z}$ there exists a sequence of vectors $z_1, \dots, z_m$ with $z_1 = z$, $z_m = z'$ and each z_m and z_{m-1} differing on only one component, and such that $z_m \in \mathcal{Z}$ for all m.[6]

Proposition 2.1. *Let $\mathcal{Z}$ be connected. Then VM holds iff for each $j \in \{1 \dots J\}$ there is an ordering $\geq_j$ on $\mathcal{Z}_j$ such that for all i: $D_i(z_j, z_{-j}) \geq D_i(z'_j, z_{-j})$ when $z_j \geq_j z'_j$, for all $z_{-j} \in \mathcal{Z}_{-j}$ such that both (z_j, z_{-j}) and (z'_j, z_{-j}) are in $\mathcal{Z}$.*

Proof. See Appendix B.4. □

[6]This rules out cases where $\mathcal{Z}$ is disjoint with respect to such chains of single-instrument switches, for example in a case of two binary instruments if $\mathcal{Z}$ consists only of the points $(0, 0)$ and $(1, 1)$. With this $\mathcal{Z}$, PM and VM are both vacuous.

The additional restriction made by VM over PM is empirically testable, by inspecting the propensity score function:

Proposition 2.2. *Suppose PM and Assumption 1 hold, and $\mathcal{Z}$ is connected. Then VM holds if and only if $\mathbb{E}[D_i|Z_i = z]$ is component-wise monotonic in z, for some fixed ordering $\succeq_j$ on each $\mathcal{Z}_j$.*

Proof. See Appendix B.4. □

By contrast, PM is compatible with any propensity score function. Note that if Assumption 1 holds conditional on covariates X_i, Proposition 2.2 also need only hold with respect to the *conditional* propensity score $\mathbb{E}[D_i|Z_i = z, X_i = x]$ (see Section 2.6).

Since IAM implies PM, it follows as a corollary to Proposition 2.2 that if IAM and Assumption 1 hold and $\mathbb{E}[D_i|Z_i = z]$ is component-wise monotonic in z, then VM holds. This establishes that if a researcher has verified that the propensity score function is monotonic, VM becomes a strictly weaker assumption than IAM. The relationship among Assumptions IAM, VM and PM is depicted graphically in Figure 2.1.

Examples of the points (a)-(e) in Figure 2.1 can be made more concrete by considering a setting of two binary instruments $\mathcal{Z} = \{0,1\} \times \{0,1\}$, with an explicit selection model of the form:

$$D_i(z_1, z_2) = \mathbb{1}(\beta_{0i} + \beta_{1i}z_1 + \beta_{2i}z_2 + \beta_{3i}z_1z_2 \geq 0) \tag{2.1}$$

where $\beta_i = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{3i})' \perp Z_i$ (Assumption 1). Given the binary treatments, this model is general enough to capture all possible selection functions $D_i(z)$.

Equation (2.1) could capture a utility maximization model in which individuals trade off an incentive $\beta_{1i}z_1 + \beta_{2i}z_2 + \beta_{3i}z_1z_2$ produced by the instruments against a net cost $-\beta_{0i}$ of treatment. Table 2.1 discusses restrictions on the support of the components of β_i that illustrate each of the points (a)-(e) in Figure 2.1. In all examples, the cost β_{0i} can be heterogeneous across individuals, but examples (c)-(e) represent threshold crossing

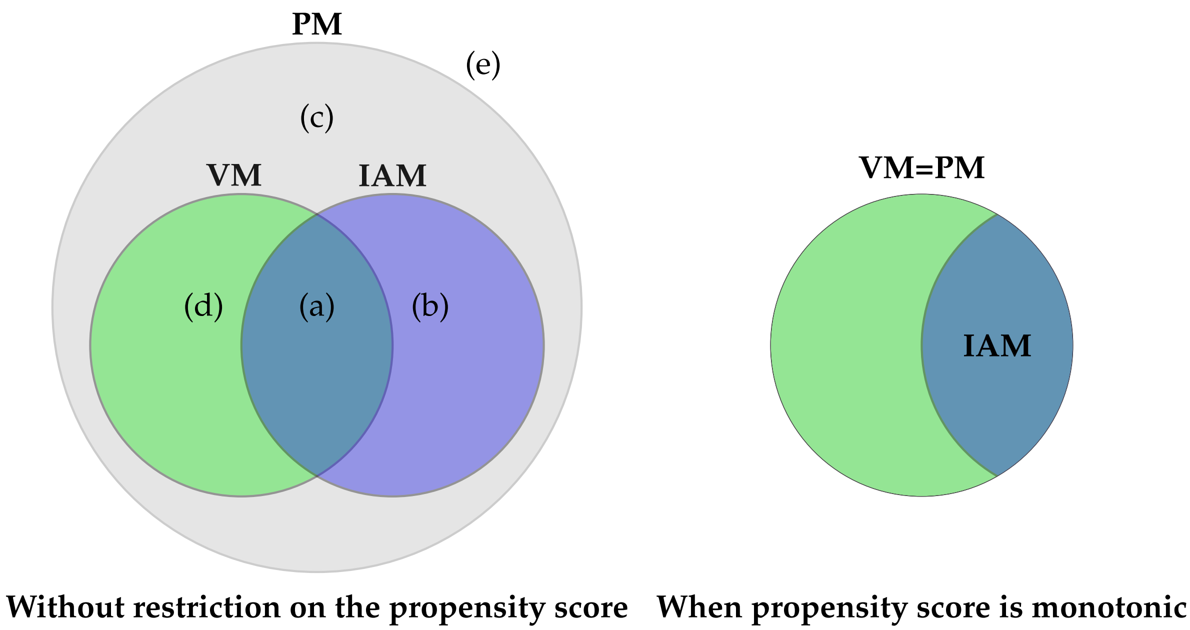

Figure 2.1: Left panel shows ex-ante comparison of Imbens & Angrist monotonicity (IAM), vector monotonicity (VM), and partial monotonicity (PM) before the propensity score function is known. Right panel depicts the relationship when the propensity score is component-wise monotonic: PM and VM become identical, with IAM a special case. Examples for points (a)-(e) are discussed in Table 2.1.

models in which heterogeneity in $D_i(z)$ is *not* linearly separable from z. This is similar to a setup considered by MTW, with a slightly different notation.

Now consider the plausibility of the above cases in the returns to schooling example, with "cheap" and "close" the 1 states of Z_1 and Z_2, respectively. In a utility maximization model β_{0i} might denote the net benefit of attending college when it is far and expensive. If college then became either cheap or close, it is natural to expect this to only increase the net benefit of college, incenting some individuals into enrolling while discouraging none. This motivates making the restrictions $\beta_{1i} \geq 0$ and $\beta_{2i} \geq 0$. If we then imagine changing to $(cheap, close)$ from either $(expensive, close)$ or $(cheap, far)$, it's reasonable to again assume that all students would move weakly towards college, unless there are individuals for whom the interaction coefficient β_{3i} is sufficiently strong and negative.[7]

[7] It is possible to imagine scenarios in which this could happen: for example, suppose there exist students who do not want to live with their parents during college, and feel that they will have to if attending a

Case	Example of support restriction on β's	Implied restrictions on selection
(a)	$\beta_1, \beta_2, \beta_3$ homogeneous; $0 \leq \beta_1 \leq \beta_2$, $\beta_3 = 0$	$D_i(0,0) \leq D_i(1,0) \leq D_i(0,1) \leq D_i(1,1)$
(b)	$\beta_1, \beta_2, \beta_3$ homogeneous; $-\beta_2 \leq \beta_3 \leq -\beta_1 \leq 0$	$D_i(0,0) \leq D_i(1,0) \leq D_i(1,1) \leq D_i(0,1)$
(c)	$\beta_{2i} \geq \beta_{1i} \geq 0, -\beta_{2i} \leq \beta_{3i} \leq -\beta_{1i}$ for all i	$D_i(0,0) \leq D_i(0,1)$; $D_i(0,0) \leq D_i(1,0)$; $D_i(1,0) \leq D_i(1,1)$; $D_i(1,1) \leq D_i(0,1)$
(d)	$\beta_{3i} = 0, \beta_{1i} \geq 0, \beta_{2i} \geq 0$ for all i $P(\beta_{2i} < -\beta_{0i} \leq \beta_{1i}) > 0$, $P(\beta_{1i} < -\beta_{0i} \leq \beta_{2i}) > 0$	$D_i(0,0) \leq D_i(0,1) \leq D_i(1,1)$; $D_i(0,0) \leq D_i(1,0) \leq D_i(1,1)$
e)	a neighborhood of the zero vector in $\mathbb{R}^4$	none

Table 2.1: Illustrative examples of each of the cases (a)-(e) in the random coefficients selection model Eq. (2.1).

Finally, note that a sufficient condition for the restriction from PM to VM is the existence of groups that are sensitive to that instrument alone. For example, suppose Alice only cares about proximity, and Bob only cares about tuition, with:

$$D_{alice}(z_1, z_2) = \mathbb{1}(z_2 = close) \quad \text{and} \quad D_{bob}(z_1, z_2) = \mathbb{1}(z_1 = cheap)$$

Partial monotonicity then requires that the directions of of response that Alice and Bob exhibit (selecting into college based on lower distance and lower tuition, respectively) hold (weakly) for all other units in the population, which then implies VM.[8] Further, the existence of both Alice and Bob imply that IAM is violated.

It is also illustrative to consider an example of this sufficient condition failing to hold. MTW offer an example where PM holds without VM, in which we consider a population of families having two or more kids (following Angrist and Evans 1998), and take as two binary instruments for having a third child indicators for the sex of the first and second child. If selection into a third child is driven uniformly by considerations of having at least one child of each sex, then no parents would respond solely to the sex of one of the

college near their parents' home. Accordingly, some such students might go to college only when it is cheap and far. Note that in this case, PM would then require that there be no other individuals in the population that go to college only if it is both cheap and close. The Supplemental Material provides a taxonomy of such cases that break VM but not PM, as point (c) does, with two binary instruments.

[8]That is, $D_{alice}(1, z_2) > D_{alice}(0, z_2)$ for all $z_2 \in \mathcal{Z}_2$ implies through PM that $D_i(1, z_2) \geq D_i(0, z_2)$ for all $z_2 \in \mathcal{Z}_2$ and i, and similarly Bob implies that $D_i(z_1, 1) \geq D_i(z_1, 0)$ for all $z_1 \in \mathcal{Z}_1$ and i.

first two children alone. This violates VM since whether or not the first child being female encourages or discourages treatment depends on the sex of the second child (and vice versa). However, I note that the instruments in this example can be recoded such that VM holds given the same assumptions about underlying selection behavior (see Supplemental Material for an example).

2.3 Characterizing complier groups under vector monotonicity

In this section I show that the assumption of vector monotonicity partitions the population of interest into a set of well-defined groups that generalize the familiar taxonomy of always-takers, never-takers, and compliers from the case of a single binary instrument. Providing a characterization of the groups will be necessary to state the main identification result in Section 2.4.

To simplify notation, let us define a random variable G_i corresponding to an individual's entire vector of counterfactual treatments $\{D_i(z)\}_{z\in\mathcal{Z}}$. For example, with a single binary instrument $G_i =$ *always-taker* indicates that $D_i(0) = D_i(1) = 1$. We refer to G_i as unit i's "response group", using a term from Lee and Salanié (2020).[9] Response groups partition individuals in the population based on upon their selection behavior over all counterfactual values of the instruments. We will see that all response groups–save for two–correspond to "compliers" of some kind.

Let $\mathcal{G}$ be the support of G_i. We can think of VM as a restriction on which response groups are allowed in the population, or equivalently a restriction on $\mathcal{G}$. As a final bit of notation, we will denote as $\mathcal{D}_g(z)$ the potential treatments function $D_i(z)$ that is common to all units sharing a value g of G_i.

[9] Heckman and Pinto (2018) refer to such groups as *response-types* or *strata*.

2.3.1 With two binary instruments

We first turn to the simplest case of two binary instruments, in which $\mathcal{G}$ can be seen to contain six distinct response groups.

Normalize the instrument value labeled "1" for each instrument to be the direction in which potential treatments are increasing. Table 2.2 describes the six response groups that can occur under VM with two binary instruments, with names introduced for each by MTW. A Z_1 complier, for example, goes to college if and only if college is cheap, regardless of whether it is close. A Z_2 complier, in our example, would go to college if and only if college is close, regardless of whether it is cheap. A reluctant complier is "reluctant" in the sense that they require college to be both cheap and close to attend, while an eager complier goes to college so long as it is either cheap or close. Never and always takers are defined in the same way as they are under IAM: $\max_{z \in \mathcal{Z}} D_i(z) = 0$ and $\min_{z \in \mathcal{Z}} D_i(z) = 1$, respectively.

Name	$\mathbf{D_i(0,0)}$	$\mathbf{D_i(0,1)}$	$\mathbf{D_i(1,0)}$	$\mathbf{D_i(1,1)}$
never takers	N	N	N	N
always takers	T	T	T	T
Z_1 compliers	N	N	T	T
Z_2 compliers	N	T	N	T
eager compliers	N	T	T	T
reluctant compliers	N	N	N	T

Table 2.2: The six response groups under VM with two binary instruments.

A natural question is whether the sizes $p_g := P(G_i = g)$ of the six groups in Table 2.2 can be detected empirically. In general, only two of them are point identified. Let $P(z) := \mathbb{E}[D_i|Z_i = z] = \sum_{g \in \mathcal{G}} p_g \mathcal{D}_g(z)$ be the propensity score function, where the second equality follows from Assumption 1. From the definitions in Table 2.2, it is clear that $p_{n.t} = 1 - P(1,1)$ and $p_{a.t.} = P(0,0)$. For the others, we can identify certain linear combinations of the group occupancies, e.g. $P(1,0) - P(0,0) = p_{Z_1} + p_{eager}$, $P(0,1) - P(0,0) = p_{Z_2} + p_{eager}$, and $P(1,1) - P(0,1) = p_{Z_1} + p_{reluctant}$. This allows us to bound each of the four remaining group sizes, given that each must be positive. For example, $\{P(1,0) - P(0,0)\} - \{P(1,1) - P(0,1)\} \leq p_{eager} \leq \min\{P(0,1) - P(0,0), P(1,0) - P(0,0)\}$. The point identified linear combinations are in fact special cases of the general identification results developed later in Section 2.4.1 (see Corollary 2.2 to Theorem 2.1).

2.3.2 With multiple binary instruments

Now we see how the two-instrument case generalizes to a case where the researcher has any number of binary instruments. While the overall number of response groups explodes combinatorially, we can still keep track of the various groups in a systematic way.

Let there be J binary instruments $Z_1 \ldots Z_J$. I focus on the baseline case in which the space of conceivable instrument values is rectangular: $\mathcal{Z} = \{0,1\}^J$ (see Supplemental Material for some alternatives). We wish to characterize the subset of the 2^{2^J} possible mappings between vectors of instrument values and treatment that satisfy VM, where we continue to normalize the "1" state for each Z_j to be the direction in which potential treatments are weakly increasing.[10] The number of such response groups G_i is equal to the number of isotone boolean functions on J variables, which I denote as Ded_J. The Ded_J follow the so-called *Dedekind sequence*, for which Kisielewicz (1988) derives an analytical expression.[11]

One group that always satisfies VM are those units for whom $D_i(z) = 0$ for all values $z \in \mathcal{Z}$: so-called never-takers. Each of the other groups can be associated with a collection of minimal combinations of instruments that are sufficient for that unit to take treatment.

[10] This "up" value for each instrument will be taken in our results to be known ex ante. In practice, this might follow from a maintained natural hypothesis, such as that lower price encourages rather than discourages college attendance. However, the directions are also empirically identified from the propensity score function (see Proposition 2.2).

[11] The first six numbers in the Dedekind sequence are $3, 6, 20, 168, 7581, 7828354$ (only 8 have been evaluated numerically). While the Dedekind numbers explode quite rapidly, they still do so much more slowly than the total number 2^{2^J} of boolean functions of J variables. For example while $3/4 = 75\%$ of conceivable response groups for $J = 1$ satisfy VM, only $20/256 \approx 7.8\%$ do for $J = 3$, and just $7581/4294967296 \approx 1.7 * 10^{-4}$ do for $J = 5$. The "bite" of VM increases with J, in the sense that it rules out a larger and larger fraction of conceivable selection patterns.

For example, in a setting with three instruments, one response group would be the units that take treatment if either $Z_1 = 1$, or if $Z_2 = Z_3 = 1$. By vector monotonicity, then, any unit in this group must also take treatment if $Z_1 = Z_2 = Z_3 = 1$. However, another group of units might take treatment *only* if $Z_1 = Z_2 = Z_3 = 1$. This group is more "reluctant" than the former. The group of always-takers are the least "reluctant": they require no instruments to equal one in order for them to take treatment.

By this logic, we can associate response groups (aside from never-takers) with families F of subsets $S \subseteq \{1 \dots J\}$ of the instrument labels. However, we need only consider families for which no element S of the family is a subset of some other S': so-called *Sperner families* (see e.g. Kleitman and Milner 1973). Families that are not Sperner would be redundant under VM, since in the example above S' could be dropped without affecting the implied selection function $D_i(z)$.

Definition 2.2 (response group for a Sperner family). *For any Sperner family F, let $g(F)$ denote the response group in which units take treatment if and only ifs $z_j = 1$ for all j in S, for at least one S in F. Denote the Sperner family associated with a response group g as $F(g)$.*

All together, the response groups satisfying VM with J binary instruments are as follows: the never-takers group, along with $Ded_J - 1$ further groups $g(F)$ corresponding to each of the distinct Sperner families F of instrument labels.

In the simplest example of the above, when $J = 1$, vector monotonicity coincides with PM and IAM, and the Sperner families corresponding to this single instrument are simply the null set and the singleton $\{1\}$: corresponding to always-takers and compliers, respectively. Together with never-takers, we have the familiar three groups from LATE analysis with a single binary instrument.

For $J = 2$, the five groups (aside from never takers) described in the previous section map to Sperner families as follows:

F	**name of $\mathbf{G_F}$**
$\emptyset$	"always takers"
$\{1\}$	"Z_1 compliers"
$\{2\}$	"Z_2 compliers"
$\{1\},\{2\}$	"eager compliers"
$\{1,2\}$	"reluctant compliers"

The rapidly expanding richness of selection behavior compatible with VM can be seen with $J = 3$, where there are 19 Sperner families, each indicated within bold brackets:

$$\boldsymbol{\{}\emptyset\boldsymbol{\}}, \boldsymbol{\{}1\boldsymbol{\}}, \boldsymbol{\{}2\boldsymbol{\}}, \boldsymbol{\{}3\boldsymbol{\}},$$

$$\boldsymbol{\{}1,2\boldsymbol{\}}, \boldsymbol{\{}1,3\boldsymbol{\}}, \boldsymbol{\{}2,3\boldsymbol{\}}, \boldsymbol{\{}1,2,3\boldsymbol{\}},$$

$$\boldsymbol{\{}\{1\},\{2\}\boldsymbol{\}}, \boldsymbol{\{}\{2\},\{3\}\boldsymbol{\}}, \boldsymbol{\{}\{1\},\{3\}\boldsymbol{\}}, \boldsymbol{\{}\{1\},\{2\},\{3\}\boldsymbol{\}},$$

$$\boldsymbol{\{}\{1,2\},\{3\}\boldsymbol{\}}, \boldsymbol{\{}\{1,3\},\{2\}\boldsymbol{\}}, \boldsymbol{\{}\{2,3\},\{1\}\boldsymbol{\}},$$

$$\boldsymbol{\{}\{1,2\},\{1,3\}\boldsymbol{\}}, \boldsymbol{\{}\{1,2\},\{2,3\}\boldsymbol{\}}, \boldsymbol{\{}\{1,3\},\{2,3\}\boldsymbol{\}},$$

$$\boldsymbol{\{}\{1,2\},\{1,3\},\{2,3\}\boldsymbol{\}}$$

For instance, an individual with G_i corresponding to $\boldsymbol{\{}\{1,2\},\{1,3\},\{2,3\}\boldsymbol{\}}$ takes treatment so long as any two instruments take the one value.

A central feature of the identification analysis will be that the selection functions corresponding to the various response groups are not all linearly independent from one another. Only 2^J such functions can be independent (though Ded_J is strictly larger for $J > 1$), since any function of binary variables can be written as a polynomial in them. Let

$\mathcal{G}^c := \mathcal{G}/\{a.t., n.t.\}$ denote the set of $\text{Ded}_J - 2$ response groups compatible with Assumption VM that are not never-takers or always takers. All of the groups in $\mathcal{G}^c$ can be thought of as generalized "compliers" of some kind: units that vary treatment uptake in *some* way across possible instrument values.

A natural basis for the set of selection functions $\{\mathcal{D}_g(z)\}_{g\in\mathcal{G}^c}$ can be formed by considering functions that are products over a single subset of the instruments

$$z_S := \prod_{j\in S} z_j = \mathbb{1}\left(z_j = 1 \text{ for all j in S}\right)$$

where $S \subseteq \{1\ldots J\}, S \neq \emptyset$.[12] For a given set S, z_S yields the selection function $\mathcal{D}_{g(S)}(z)$ of the response group $g(S)$ corresponding to the Sperner family consisting only of the set S. I refer to such response groups $g(S)$ as *simple*.

For $J = 2$, the selection functions for the simple response groups are:

$$\mathcal{D}_{Z_1}(z) = z_1 \qquad \mathcal{D}_{Z_2}(z) = z_2 \qquad \mathcal{D}_{reluctant}(z) = z_1 z_2$$

The selection function for the remaining group, eager compliers, can be obtained as:

$$\mathcal{D}_{eager}(z) = z_1 + z_2 - z_1 z_2 = \mathcal{D}_{Z_1}(z) + \mathcal{D}_{Z_2}(z) - \mathcal{D}_{reluctant}(z)$$

We can express this linear dependency by the matrix M_J in the system:

$$\begin{pmatrix} \mathcal{D}_{Z_1}(z) \\ \mathcal{D}_{Z_2}(z) \\ \mathcal{D}_{reluctant}(z) \\ \mathcal{D}_{eager}(z) \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & -1 \end{pmatrix}}_{:=M_2} \begin{pmatrix} \mathcal{D}_{Z_1}(z) \\ \mathcal{D}_{Z_2}(z) \\ \mathcal{D}_{reluctant}(z) \end{pmatrix} \tag{2.2}$$

[12]Note that a similar construction plays a central role in Lee and Salanié, 2018.

For general J, we define the matrix M_J from the analogous system of equations:

$$\{\mathcal{D}_{g(F)}(z)\}_{F:\, g(F)\in\mathcal{G}^c} = M_J\{\mathcal{D}_{g(S)}(z)\}_{S\subseteq\{1\dots J\}, S\neq\emptyset}$$

for all $z \in \mathcal{Z}$. The rows of matrix M_J are indexed by Sperner families (corresponding to the groups in $\mathcal{G}^c$), and the columns by the simple Sperner families for non-null S. The entries of M_J are given by the following expression:[13]

Proposition 2.3. $[M_J]_{F,S'} = \sum_{f\in s(F,S')}(-1)^{|f|+1}$ *where* $s(F,S') := \left\{f \subseteq F : \left(\bigcup_{S\in f} S\right) = S'\right\}$.

Proof. See Appendix B.4. □

2.3.3 Vector monotonicity with discrete instruments

More generally, when the researcher has discrete instrumental variables that satisfy vector monotonicity, they can be re-expressed as a larger number of binary instruments in a way that preserves vector monotonicity. By introducing a binary instrument for every value but one of each discrete instrument, the analysis can be extended to this much more general setting:

Proposition 2.4. *Let Z_1 be a discrete variable with M ordered points of support $z_1 < z_2 < \dots < z_M$, and $Z_2 \dots Z_J$ be other instrumental variables. Let $\tilde{Z}_{mi} := \mathbb{1}(Z_{1i} \geq z_m)$. If the vector $Z = (Z_1, \dots Z_J)$ satisfies Assumption VM on a connected $\mathcal{Z}$ then so does the vector $(\tilde{Z}_2, \dots, \tilde{Z}_M, Z_2, \dots Z_J)$.*

Proof. See Appendix B.4. □

Applying Proposition 2.4 iteratively offers a fairly general recipe for mapping the instruments available in a given empirical setting into the framework of binary instruments.

Note that the mapping in Proposition 2.3.3 introduces restrictions on $\mathcal{Z}$ for the resulting binary instruments, since for example we could not have both $\tilde{Z}_{2i} = 1$ and $\tilde{Z}_{1i} = 0$. As

[13] The matrix M_3, which has $\mathcal{D}_3 - 2 = 18$ rows and $2^3 - 1 = 7$ columns is given explicitly in the Supplemental Material.

a result, not all of the response groups introduced in Section 2.3.2 are necessary to account for, since the possible patterns of instrument variation pool some into equivalent groups. While in the next section I assume full binary instrument support for the baseline results, Appendix B.1 provides the necessary generalizations to make use of Proposition 2.4.

2.4 Parameters of interest and identification

In this section I define and characterize a class of causal parameters, and show that they are point identified under vector monotonicity and a full-support condition on the instruments. This section maintains a setup of J binary instruments with $\mathcal{Z} = \{0,1\}^J$ unless otherwise specified.

2.4.1 Main identification result

My identification analysis considers conditional averages of potential outcomes: for $d \in \{0,1\}$ and an arbitrary function f, let

$$\theta_c^{fd} := \mathbb{E}[f(Y_i(d))|C_i = 1]$$

where $C_i = c(G_i, Z_i)$ is a function $c : \mathcal{G} \times \mathcal{Z} \to \{0,1\}$ of individual i's response group and their realization of the instruments. Intuitively, the event $C_i = 1$ will indicate that unit i belongs to a certain subgroup of generalized "compliers". Most of the discussion will center on the class of average treatment effect parameters:

$$\Delta_c := \mathbb{E}[Y_i(1) - Y_i(0)|C_i = 1] = \theta_c^{y1} - \theta_c^{y0}$$

with $f(y) = y$ the identity function. Treatment effect parameters having the form of Δ_c are familiar both from the LATE (Imbens and Angrist, 1994) and marginal treatment effects (Heckman and Vytlacil, 2005) literatures. For instance, with a single binary instrument the LATE sets $c(g,z) = \mathbb{1}(g = complier)$, independent of z.

The main result is that identification of is possible under VM for certain choices of the function $c(g, z)$. In particular, it will require a condition that I call "Property M":

Definition 2.3 (Property M). *The function $c(g, z)$ satisfies Property M if for all $z \in \mathcal{Z}$: $c(a.t., z) = c(n.t., z) = 0$, while for every $g \in \mathcal{G}^c$:*

$$c(g, z) = \sum_{S \subseteq \{1 \dots J\}, S \neq \emptyset} [M_J]_{F(g),S} \cdot c(g(S), z)$$

where the matrix M_J is defined in Proposition 2.3. I'll also say that θ_c^{fd} or Δ_c "satisfies Property M" if its underlying function $c(g, z)$ does. Intuition for Property M is provided after the statement of the identification result, and an equivalent characterization of Property M and leading examples are given in Section 2.4.2.

Causal parameters that satisfy Property M are identified under VM with binary instruments, provided the instruments provide sufficient independent variation in treatment uptake. This holds when the binary instruments have full (rectangular) support:

Assumption 3 (full support). *$P(Z_i = z) > 0$ for all $z \in \{0, 1\}^J$*

Assumption 3 is stronger than is necessary but simplifies presentation – Appendix B.1 presents a generalization.

An alternative expression of Assumption 3 is useful for writing the identification result explicitly. For an arbitrary ordering of the $k := 2^J - 1$ non-empty subsets $S \subseteq \{1 \dots J\}$, define the random vector $\Gamma_i = (Z_{S_1 i} \dots Z_{S_k i})'$ from products of the Z_{ji} for j within each subset S. That is, each element of Γ_i indicates the treatment status of a particular simple response group, given Z_i. Let Σ be the covariance matrix of Γ_i.

Lemma 2.1. *Assumption 3 holds if and only if Σ has full rank.*

Proof. See Appendix B.4. □

Lemma 2.1 reveals that full support of the instruments is equivalent to there being independent variation in treatment takeup among all of the simple response groups.

We may now state the main result:

Theorem 2.1. *Under Assumptions 1-3 (independence & exclusion, VM, and full support), for any c satisfying Property M and any measurable function $f(Y)$ for each $d \in \{0,1\}$:*

$$\theta_c^{fd} = (-1)^{d+1} \frac{\mathbb{E}[f(Y_i)h(Z_i)\mathbb{1}(D_i = d)]}{\mathbb{E}[h(Z_i)D_i]},$$

provided that $P(C_i = 1) > 0$, where $h(Z_i) = \lambda'\Sigma^{-1}(\Gamma_i - \mathbb{E}[\Gamma_i])$ and

$$\lambda = (\mathbb{E}[c(g(S_1), Z_i)], \ldots \mathbb{E}[c(g(S_k), Z_i)])'$$

Proof. See Appendix B.4. □

It follows immediately from Theorem 2.1 that conditional average treatment effects $\Delta_c = \mathbb{E}[Y_i(1) - Y_i(0)|C_i = 1]$ satisfying Property M are identified as:

$$\Delta_c = \mathbb{E}[h(Z_i)Y_i]/\mathbb{E}[h(Z_i)D_i]$$

Note that as the numerator of Δ_c depends on Z_i and Y_i only and the denominator depends on Z_i and D_i only, identification of Δ_c would hold in a "split-sample" setting where Y_i and D_i are not necessarily linked in the same dataset.

We can also re-express the empirical estimand for Δ_c delivered by Theorem 2.1 in a more illuminating form, directly in terms of conditional expectation functions of each of Y_i and D_i on the instruments:

Corollary 2.1. *Under the Assumptions of Theorem 2.1:*

$$\Delta_c = \frac{\sum_{z \in \mathcal{Z}} \left(\sum_{S \subseteq \{1 \ldots J\}, S \neq \emptyset} \lambda_S A_{S,z} \right) \mathbb{E}[Y_i|Z_i = z]}{\sum_{z \in \mathcal{Z}} \left(\sum_{S \subseteq \{1 \ldots J\}, S \neq \emptyset} \lambda_S A_{S,z} \right) \mathbb{E}[D_i|Z_i = z]}$$

where λ_S is as defined in Theorem 2.1 and $A_{S,z} = \sum_{\substack{f \subseteq z_0 \\ (z_1 \cup f) = S}} (-1)^{|f|}$, with (z_1, z_0) a partition of *the indices $j \in \{1 \dots J\}$ that take a value of zero or one in z, respectively.*

Proof. See Appendix B.4. The proof of Lemma 2.1 gives the explicit form of A for $J = 2$. □

Intuition for Theorem 2.1

The basic logic behind Theorem 2.1 can be appreciated by focusing on the average treatment effect parameters Δ_c, and observing that by Assumption 1 and the law of iterated expectations they can be written as a weighted average over response-group specific average treatment effects $\Delta_g := \mathbb{E}[Y_i(1) - Y_i(0)|G_i = g]$:

$$\Delta_c = \sum_{g \in \mathcal{G}} \left\{ \frac{P(G_i = g)\mathbb{E}[c(g, Z_i)]}{\mathbb{E}[c(G_i, Z_i)]} \right\} \cdot \Delta_g \tag{2.3}$$

where the weights are each proportional to the quantity $\mathbb{E}[c(g, Z_i)]$. Now consider a general type of "2SLS-like" estimand, in which a single scalar instrument $h(Z_i)$ is constructed from the vector of instruments Z_i according to some function h, and then used in a simple linear IV regression.[14]

Proposition 2.5. *Under Assumption 1 (exclusion and independence):*

$$\frac{Cov(Y_i, h(Z_i))}{Cov(D_i, h(Z_i))} = \sum_{g \in \mathcal{G}} \frac{P(G_i = g) \cdot Cov(\mathcal{D}_g(Z_i), h(Z_i))}{\sum_{g' \in \mathcal{G}} P(G_i = g') \cdot Cov(\mathcal{D}_{g'}(Z_i), h(Z_i))} \cdot \Delta_g$$

Proof. See the Supplemental Material for direct proof of this form. □

Proposition 2.5 reveals that such 2SLS-like estimands also uncover a weighted average of the Δ_g, where the weight placed on each response group g is governed by the covariance between $\mathcal{D}_g(Z_i)$ and $h(Z_i)$.

Comparing Equation (2.3) with Equation 2.3, we see that a 2SLS-like estimand can identify Δ_c if the function h is chosen in such a way that $Cov(\mathcal{D}_g(Z_i), h(Z_i)) = \mathbb{E}[c(g, Z_i)]$

[14] Special cases include two stage least squares: $h(z) = \mathbb{E}[D_i|Z_i = z]$, and Wald estimands: $h(z) = \frac{\mathbb{1}(Z_i=z)}{P(Z_i=z)} - \frac{\mathbb{1}(Z_i=z')}{P(Z_i=z')}$.

for all the response groups g. However, since the covariance operator is linear, the linear dependencies examined in Section 2.3.2 translate into a set of linear restrictions among these weights, captured by the matrix M_J. Property M guarantees that the vector of $\mathbb{E}[c(g(F), Z_i)]$ across Sperner families F belongs to the column-space of the matrix M_J, whatever the distribution of Z_i. What remains to secure identification is then simply to tune the covariances for the simple response groups, which is achieved by the construction of $h(Z_i)$ in Theorem 2.1.

The role of Property M in Theorem 2.1 can be thought of as emerging from there being under VM more response groups in $\mathcal{G}^c$ than there are independent pairs of points in the support of the instruments. By contrast, under IAM with J binary instruments both are generally equal to $2^J - 1$, and it is possible to identify the average treatment effect $\Delta_{g'} := \mathbb{E}[Y_i(1) - Y_i(0)|G_i = g']$ within any single such response group g' (and hence also obtain any desired convex combination of the $\Delta_{g'}$). However, under VM the corresponding choice $c(g, z) = \mathbb{1}(g = g')$ fails to satisfy Property M, and we will not be able to identify the Δ_g individually in general.[15] The first requirement in Property M of zero weight on always-takers or never-takers on the other hand is familiar from analysis based on IAM.[16]

2.4.2 Examples from the family of identified parameters

While Property M introduced in Section 2.4 itself is somewhat abstract, the following result shows that it is equivalent to $c(g, z)$ being equal to a sum of selection functions $\mathcal{D}_g(z)$ for all response groups g.

[15]We can see this in a simple example with $J = 2$ and $g = Z_1$ complier. In this case Property M would require that $c(\text{eager complier}, z) = c(Z_1 \text{ complier}, z) + c(Z_2 \text{ complier}, z) - c(\text{reluctant complier}, z)$, i.e. that $0 = 1 + 0 - 1$, cf Eq. (2.2).

[16]Note that $\mathbb{E}[c(g, Z_i)] = 0$ would also be necessary for any additional groups g for whom, given the distribution of Z_i, there is no actual variation in treatment status. In the baseline analysis, such additional groups will be ruled out by Assumption 3.

Proposition 2.6. *A function $c : \mathcal{G} \times \mathcal{Z} \to \{0, 1\}$ satisfies Property M if and only if*

$$c(g, z) = \sum_{k=1}^{K} \{\mathcal{D}_g(u_k(z)) - \mathcal{D}_g(l_k(z))\}$$

for some $K \leq J/2$, where $u_k(\cdot)$ and $l_k(\cdot)$ are functions $\mathcal{Z} \to \mathcal{Z}$ such that $u_k(z) \geq l_k(z)$ component-wise while $l_k(z) \geq u_{k+1}(z)$ component-wise, for all k and $z \in \mathcal{Z}$.

Proof. See Appendix B.4. □

Proposition 2.6 yields a natural interpretation of average treatment effects that satisfy Property M, which is that they can be written as

$$\Delta_c = E\left[Y_i(1) - Y_i(0) \middle| \bigcup_{k=1}^{K} \{D_i(u_k(Z_i)) > D_i(l_k(Z_i))\}\right] \tag{2.4}$$

for some functions u_k and l_k having the properties stated in Proposition 2.6.[17] From Equation 2.4 we see that the types of complier groups that identified parameters can condition on are groups of individuals that are responsive to any of a set of K instrument transitions that each induce only *one-way flows* into treatment. This feature is in fact common to both IAM and VM. Indeed, identified parameters can also be written in this form under IAM (as well as its generalization to Heckman and Pinto (2018)'s concept of unordered monotonicity–see the proof Proposition 2.6 for a discussion).

While the form of Equation 2.4 is somewhat familiar from LATE results under IAM, the additional structure of VM yields new causal parameters that bear economically interesting interpretations. The remainder of this section continues to focus on average treatment effects Δ_c, though θ_c^{fd} parameters can be defined for the analogous groups. Table 2.3 presents some leading examples of Δ_c that satisfy Property M, as can be seen by applying Proposition 2.6. All of the cases presented in Table 2.3 admit the form of Equation (2.4) with a single term ($K = 1$), given in the third column.

[17] This expression is obtained by substituting $C_i = c(G_i, Z_i)$, and noting that $\sum_{k=1}^{K} D_i(u_k(Z_i)) - D_i(l_k(Z_i))$ equals one if and only if $D_i(u_k(Z_i)) > D_i(l_k(Z_i))$ for some k.

Parameter	$\mathbf{c}(\mathbf{g},\mathbf{z})$	**Proposition 2.6 form**
ACL	$\mathbb{1}(g \in \mathcal{G}^c)$	$\mathcal{D}_g(1,1\ldots1) - \mathcal{D}_g(0,0\ldots0)$
$SLATE_{\mathcal{J}}$	$\mathcal{D}_g((1\ldots1),z_{-\mathcal{J}}) - \mathcal{D}_g((0\ldots0),z_{-\mathcal{J}}))$	"
$SLATT_{\mathcal{J}}$	$\mathcal{D}_g(z)\cdot(\mathcal{D}_g((1\ldots1),z_{-\mathcal{J}}) - \mathcal{D}_g((0\ldots0),z_{-\mathcal{J}})))$	$\mathcal{D}_g(z) - \mathcal{D}_g((0\ldots0),z_{-\mathcal{J}})$
$SLATU_{\mathcal{J}}$	$(1-\mathcal{D}_g(z))\cdot(\mathcal{D}_g((1\ldots1),z_{-\mathcal{J}}) - \mathcal{D}_g((0\ldots0),z_{-\mathcal{J}})))$	$\mathcal{D}_g((1\ldots1),z_{-\mathcal{J}}) - \mathcal{D}_g(z)$
$PTE_j(z^*_{-j})$	$\mathcal{D}_g(1,z^*_{-j}) - \mathcal{D}_g(0,z^*_{-j}))$	"

Table 2.3: Leading parameters of interest satisfying Property M, including: the all-compliers LATE, set LATEs, set LATEs on the treated, set LATEs on the untreated, and partial treatment effects (see text for details).

I call the first item in Table 2.3 the "all-compliers LATE" (ACL), which is the average treatment effect among all units who are not always-takers or never-takers. This is the largest subgroup of the population for which treatment effects can be generally point identified from instrument variation.[18] With two instruments, the ACL averages over all units who are Z_1, Z_2, eager or reluctant compliers. In the returns to schooling example, we can equivalently describe the ACL as the average treatment effect among individuals who would go to college were it close and cheap, but would not were it far and expensive.

On the other end of the spectrum, the final row of Table 2.3 gives the most disaggregated type of parameter satisfying Property M, what might be called a *partial treatment effect* $PTE_j(z^*_{-j})$. This is the average treatment effect among individuals that move into treatment when a single instrument j is shifted from zero to one, while the other instrument values are held fixed at some explicit vector of values z^*_{-j}. An example is the average treatment effect among individuals who go to college if it is close and cheap, but not if it is far and cheap. Ultimately, all Δ_c satisfying Property M can be written as convex combinations of such partial treatment effects though the number could be quite large (see Supplemental Material for an explicit expression). However, the PTEs still combine response groups: the example above for instance combines proximity compliers with reluctant compliers.

[18] We may of course still be able to say something about treatment effects for never-takers and always-takers given additional restrictions (see e.g. Section 2.4.3 for bounds on the unconditional ATE when potential outcomes are bounded).

The remaining parameters in Table 2.3 constitute a middle ground between the granular PTE's and the very broad averaging of the ACL. For example, the ACL is a special case of what I call a *set local average treatment effect*, or $SLATE_{\mathcal{J}}$, which captures the average treatment effect among units that move into treatment when all instruments in some fixed set $\mathcal{J}$ are changed from 0 to 1, with the other instruments not in $\mathcal{J}$ fixed at their realized values. The ACL is a special case in which this set is all of the instruments: $\mathcal{J} = \{1, 2, \dots J\}$. When $\mathcal{J}$ contains just one instrument index, SLATE recovers treatment effects among those who would "comply" with variation in that single instrument. For example, $SLATE_{\{2\}}$ is the average treatment effect among individuals who don't go to college if it is far, but do if it is close. This parameter may be of interest to policymakers considering whether to expand a community college to a new campus, for example. The group of individuals included in $SLATE_{\{2\}}$ are Z_2 compliers, eager compliers with high tuition rates ($Z_{1i} = 0$), and reluctant compliers with low tuition rates ($Z_{1i} = 1$).[19]

For a discrete instrumental variable mapped to multiple binary instruments by Proposition 2.4, the LATE among units moved into treatment between any two of its values will also be an example of a SLATE. For example, if Z_1 has support $z_1 < z_2 < z_3 < z_4$, the average treatment effect among individuals for which $D_i(z_4, Z_{-1,i}) > D_i(z_2, Z_{-1,i})$ corresponds to $SLATE_{\mathcal{J}}$ with $\mathcal{J} = \{\tilde{Z}_3, \tilde{Z}_4\}$. SLATE thus allows the practitioner to flexibly condition upon response to individual or joint variation in the instruments.

The treatment effect parameters $SLATT_{\mathcal{J}}$ and $SLATU_{\mathcal{J}}$ in the final two rows of Table 2.3 are similar to $SLATE_{\mathcal{J}}$ but additionally condition on units' realized treatment status. For example $SLATT_{\{1,2\}}$ with our two instruments averages over individuals who do go to college, but wouldn't have were it far and expensive.[20] SLATT and SLATU can also be

[19]Note that a single-instrument SLATE like $SLATE_{\{2\}}$ does not generally correspond to using Z_2 alone as an instrument, since this latter estimand does not control for variation in Z_1 that is correlated with Z_2. If on the other hand the instruments are independent of one another, using 2SLS may be justified, as I show in the Supplemental Material.

[20]Note that with a single binary instrument, $SLATT_{\{1\}}$ coincides with $ACL = SLATE_{\{1\}}$, as $\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1, G_i = \textit{complier}] = \mathbb{E}[Y_i(1) - Y_i(0)|Z_i = 1, \textit{complier}] = \mathbb{E}[Y_i(1) - Y_i(0)|\textit{complier}]$, using Assumption 1. However, when the group $\mathcal{G}^c$ consists of more than one group, the "all-compliers" version of $SLATT$ generally differs from ACL.

used to construct bounds on the average treatment effect among the treated or untreated, when potential outcomes are bounded, following logic for the ATE given in Section 2.4.3.

To construct some further examples of identified parameters from the ones mentioned in Table 2.3, one could make use of a closure property of the set of Δ_c that satisfy Property M. Let $\mathcal{C}$ denote the set of $c : \mathcal{G} \times \mathcal{Z} \to \{0, 1\}$ that satisfy Property M, and let $c_a(g, z)$ and $c_b(g, z)$ be two functions in $\mathcal{C}$. Then it is straightforward to show that $c_a(g, z) - c_b(g, z) \in \mathcal{C}$ if and only if $c_b(g, z) \leq c_a(g, z)$ for all $z \in \mathcal{Z}, g \in \mathcal{G}^c$.[21] We can use this observation to generate parameters that condition on the "complement" of the complier group for Δ_{c_b} within the larger complier group for Δ_{c_a}. For example, with $J = 2$:

$$\mathbb{E}[\Delta_i | G_i \in \mathcal{G}^c - \{D_i(1, Z_{2i}) - D_i(0, Z_{2i})\}]$$

yields the average treatment effect among individuals who are counted in the ACL but not in $SLATE_{\{1\}}$. These individuals would not respond to a counterfactual reduction in college tuition alone, but would respond if both instruments were shifted in concert.

2.4.3 Further results on identification

This section outlines some further results related to identification under VM. I begin with several observations that strengthen or extend the reach of Theorem 2.1.

Consequences and extensions of Theorem 2.1

1) The size of the relevant complier sub-population is identified: The argument used in Theorem 2.1 can be leveraged to show that the proportion of relevant "compliers" associated with any causal parameter satisfying Property M is also identified, and is the denominator of the associated estimand:

[21] This follows from linearity and the definition of Property M, while $c_b(g, z) \leq c_a(g, z)$ is necessary for the image of the new function to remain $\{0, 1\}$.

Corollary 2.2 to Theorem 2.1. *Make Assumptions 1-3. For any c that satisfies Property M, $P(C_i = 1)$ is identified as $\mathbb{E}[h(Z_i)D_i]$, where $h(z)$ is as given in Theorem 2.1.*

Proof. See Appendix B.4. □

2) Property M as a necessary condition. Property M was introduced in this section as part of a set of sufficient conditions for identification of Δ_c. One can show that, loosely speaking, any identified Δ_c must satisfy Property M. In this sense, Property M is also a necessary condition for identification. The simplest form of this result I express in terms of so-called "IV-like estimands" introduced by Mogstad et al. (2018), which are any cross moment $\mathbb{E}[s(D_i, Z_i)Y_i]$ between Y_i and a function of treatment and instruments. Let $\mathcal{P}_{DZ}$ denote the joint distribution of D and Z, which is identified. Then:

Proposition 2.7. *Suppose Δ_c is identified by a finite set of IV-like estimands and $\mathcal{P}_{DZ}$, provided that Assumptions 1-3 hold and $P(C_i = 1) > 0$. Then Δ_c satisfies Property M.*

Proof. See Appendix B.4. □

The result can be strengthened given regularity conditions on the support of potential outcomes:

Proposition 2.8. *Suppose that the support of each potential outcome conditional on $G_i = g$ is independent of all $g \in \mathcal{G}^c$ for which $P(G_i = g) > 0$, and that the density (or p.m.f.) of each $Y_i(d)$ is uniformly bounded and separated from zero over that support, conditional on each such $G_i = g$. Then if Δ_c is point identified from the distribution of (Y_i, D_i, Z_i) whenever Assumptions 1-3 hold and $P(C_i = 1) > 0$, Δ_c must satisfy Property M.*

Proof. See Supplemental Material. □

3) PM alone does not lead to identification. We can demonstrate that the assumption of vector monotonicity does have identifying power in Theorem 2.1, above and beyond that of partial monotonicity. For the $J = 2$ case, it is possible to see by explicit enumeration of the possible response groups that Theorem 2.1 cannot hold under PM only:

Proposition 2.9. *When $J = 2$, if PM holds but neither VM nor IAM hold, the ACL is not point identified from knowledge of any set of IV-like estimands and $\mathcal{P}_{DZ}$.*

Proof. See Appendix B.4. □

4) Linear dependency among the instruments: Assumption 3 is stronger than is strictly necessary for identification, since linear dependencies between products of the instruments may not pose a problem if the corresponding "weights" in Δ_c do not need be tuned independently from one another. In Appendix B.1, I give a version of Assumption 3 and generalization of the identification theorem that can accommodate instrument support restrictions and/or non-rectangular $\mathcal{Z}$ (for instance after applying Proposition 2.4).

5) Conditional distributions of the potential outcomes By choosing $f(Y) = \mathbb{1}(Y \leq y)$ in Theorem 2.1 for some value y in the support of Y_i, we can identify the CDF of each potential outcome at y conditional on $C_i = 1$ as: $F_{Y(d)|C=1}(y) = (-1)^{d+1}\frac{\mathbb{E}[h(Z_i)\mathbb{1}(D_i=d)\mathbb{1}(Y_i \leq y)]}{\mathbb{E}[h(Z_i)D_i]}$ (note that unlike identification of Δ_c this requires observing (Y_i, Z_i, D_i) all in the same sample). This allows for the identification of $C_i = 1$ conditional quantile treatment effects, bounds on the distribution of treatment effects (Fan and Park, 2010), or distributional treatment effects: $F_{Y(1)|C=1}(y) - F_{Y(0)|C=1}(y)$ as $\frac{\mathbb{E}[h(Z_i)\mathbb{1}(Y_i \leq y)]}{\mathbb{E}[h(Z_i)D_i]}$.

6) Covariates. If Assumption 1 holds only conditional on a set of covariates X, and Assumption 3 also holds conditionally, then Theorem 2.1 can be taken to hold within a covariate cell $X_i = x$. In Appendix B.2, I describe how covariates can be accommodated nonparametrically, or parametrically as implemented in Section 2.6.

Identification of the ACL from a single Wald ratio

The population estimand corresponding to the all-compliers LATE takes on a particularly simple form. In particular, the ACL is equal to the following single Wald ratio:

$$\rho_{\bar{Z},\underline{Z}} := \frac{\mathbb{E}[Y_i|Z_i = \bar{Z}] - \mathbb{E}[Y_i|Z_i = \underline{Z}]}{\mathbb{E}[D_i|Z_i = \bar{Z}] - \mathbb{E}[D_i|Z_i = \underline{Z}]} \tag{2.5}$$

where $\bar{Z} = (1, 1, \ldots 1)'$ and $\underline{Z} = (0, 0, \ldots 0)'$, provided that $P(Z_i = \bar{Z}) > 0$ and $P(Z_i = \underline{Z}) > 0$, and the denominator is non-zero.[22] This can be seen by applying the law of iterated expectations over response groups, or using Theorem 2.1. That $\rho_{\bar{Z},\underline{Z}}$ is equivalent to the expression given for ACL by Theorem 2.1 is not obvious, but this can be shown by applying Corollary 2.1 and using properties of the matrix A.[23]

Thus the ACL is identified by a remarkably simple quantity: one can restrict the population to $Z_i \in \{\underline{Z}, \bar{Z}\}$ and use $\mathbb{1}(Z_i = \bar{Z})$ as a single instrument. However, Theorem 2.1 yields identification of a much larger class of parameters than ACL alone, which are not generally equal to a single Wald ratio. Furthermore, as we will see in Section 2.5, the alternative form of Theorem 2.1 suggests a means of improving estimation of the ACL. In particular, when the number of sample observations in $\underline{Z}$ and $\bar{Z}$ is not large, the Wald ratio $\rho_{\bar{Z},\underline{Z}}$ may be difficult to estimate precisely, and the sample analog of Eq. (2.5) can be expected to perform poorly. A regularization procedure based on the expression for Δ_c from Theorem 2.1 can be helpful in such cases, as shown in Appendix B.3.

Identified sets for ATE, ATT, and ATU

One drawback of the identification results presented is that since parameters like Δ_c satisfying Property M exclude never-takers and always-takers by assumption, their definition always depends upon the set of instruments available. This is not ideal unless the

[22] An analogous result holds under IAM as well with finite instruments, where in that case we take any $\bar{Z} \in \text{argmax}_z \mathbb{E}[D_i|Z_i = z]$ and $\underline{Z} \in \text{argmin}_z \mathbb{E}[D_i|Z_i = z]$, and define $\mathcal{G}^c := \{g \in \mathcal{G} : \mathbb{E}[\mathcal{D}_g(Z_i)] \in (0, 1)\}$.

[23] In particular the identity $\sum_{f \subseteq S}(-1)^{|f|} = 0$ for any $S \neq \emptyset$ annihilates all but two of the components of $(0, \lambda')'A$.

complier subpopulation is directly of interest.

When Y_i has bounded support, the parameters identified by Theorem 2.1 can be used to generate sharp worst-case bounds in the spirit of Manski (1990) for the unconditional average treatment effect (ATE), average treatment effect on the treated (ATT), and average treatment effect on the untreated (ATU). Here I show this for the ATE to illustrate – identified sets for the ATT and ATU can be constructed by analogous steps. Suppose that $Y_i(d) \in [\underline{Y}, \bar{Y}]$ with probability one, for each $d \in \{0, 1\}$. Then bounds for the ATE can be constructed by noting that:

1. $ATE := \mathbb{E}[Y_i(1) - Y_i(0)] = p_a \Delta_a + p_n \Delta_n + (1 - p_t - p_a) ACL$
2. $p_n \Delta_n \in [\underline{Y} \cdot p_n - \mathbb{E}[Y_i(1 - D_i)|Z_i = \bar{Z}], \bar{Y} \cdot p_n - \mathbb{E}[Y_i(1 - D_i)|Z_i = \bar{Z}]]$
3. $p_a \Delta_a \in [\mathbb{E}[Y_i D_i|Z_i = \underline{Z}] - p_a \cdot \bar{Y}, \mathbb{E}[Y_i D_i|Z_i = \underline{Z}] - p_a \cdot \underline{Y}]$

where $p_a := P(G_i = a.t.) = \mathbb{E}[D_i|Z_i = \underline{Z}]$ and $p_n := P(G_i = n.t.) = \mathbb{E}[1 - D_i|Z_i = \bar{Z}]$.

Note that under the bounded support condition the ATE can be partially identified whenever its conditional analog is identified for *some* subgroup of the population, and the size of that subgroup is also identified. Using variation in all of the instruments, as the ACL does, for the conditioning event leads to the narrowest possible such bounds.

2.5 Estimation

This section proposes a natural two-step estimator for the family of identified causal parameters introduced in Section 2.4, focusing on the conditional average treatment effects Δ_c. Appendix B.3 discusses its limiting distribution, which is asymptotically normal.

Theorem 2.1 establishes that Δ_c satisfying Property M are equal to a ratio of two population expectations – thus a natural plug-in estimator simply replaces these with their sample counterparts, provided $h(Z_i)$ is a strong enough instrument to avoid any weak identification issues.

Following $h(Z_i) = \lambda'\Sigma^{-1}(\Gamma_i - \mathbb{E}[\Gamma_i])$ from Theorem 2.1, define $\hat{H} = n\tilde{\Gamma}(\tilde{\Gamma}'\tilde{\Gamma})^{-1}\hat{\lambda}$, where $\tilde{\Gamma}$ is a $n \times k$ design matrix with entries $\tilde{\Gamma}_{il} = Z_{S_l i} - \frac{1}{n}\sum_{j=1}^{n} Z_{S_l j}$, where S_l is the l^{th} subset according to some arbitrary ordering of the $k := 2^J - 1$ non-empty subsets $S \subseteq \{1 \dots J\}$. Note that the rows of $\tilde{\Gamma}$ correspond to observations of the vector Γ_i introduced in Section 2.4.1, de-meaned with respect to the sample mean. The vector $\hat{\lambda}$ is a sample estimator of $\lambda = (\mathbb{E}[c(g(S_1), Z_i)], \dots \mathbb{E}[c(g(S_k), Z_i)])'$, given explicitly below for our leading examples.

Given the vector $\hat{H}$ as defined above, consider $\hat{\rho} = (\hat{H}'D)^{-1}(\hat{H}'Y)$, where Y and D are $n \times 1$ vectors of observations of Y_i and D_i, respectively. Noticing that for any vector $V \in \mathbb{R}^n$, $(\tilde{\Gamma}'\tilde{\Gamma})^{-1}\tilde{\Gamma}'V$ is the sample linear projection coefficient vector of V on the de-meaned sample vectors of Z_{Si}, we can re-express it by the Frisch-Waugh-Lovell theorem as $(0, \lambda')(\Gamma'\Gamma)^{-1}\Gamma'V$, where Γ adds a column of ones and skips the demeaning. The estimator can now be written as $\hat{\rho} = \hat{\rho}(\hat{\lambda})$ where

$$\hat{\rho}(\lambda) = \left((0, \lambda')(\Gamma'\Gamma)^{-1}\Gamma'D\right)^{-1}(0, \lambda')(\Gamma'\Gamma)^{-1}\Gamma'Y \tag{2.6}$$

Assume existence of $(\Gamma'\Gamma)^{-1}$ in finite sample, and note that its population analog exists as a consequence of Assumption 3. When Assumption 3 does not hold but identification is still possible (see Appendix B.1), the matrices $\tilde{\Gamma}$ and Γ may be defined in the same way but using only sets S within a smaller collection $\mathcal{F}$. For example, when using construction of Proposition 2.4 that maps discrete to binary instruments, $\mathcal{F}$ can be taken to include all sets of the final binary instruments that do not contain distinct $\tilde{Z}$ from the same original discrete instrument. In all cases, let $\mathcal{F}$ index the elements of Γ_i, where $\mathcal{F} = \{S \subseteq \{1, 2, \dots J\}, S \neq \emptyset\}$ in the baseline setting.

Comparison with 2SLS: Note that the estimator $\hat{\rho}(\lambda)$ in Equation 2.6 is very similar in form to a "fully-saturated" 2SLS estimator that includes an indicator for each value of $Z_i \in \mathcal{Z}$ in the first stage. Indeed, that estimator is $\hat{\rho}_{2sls} = \left(D'\Gamma(\Gamma'\Gamma)^{-1}\Gamma'D\right)^{-1} D'\Gamma(\Gamma'\Gamma)^{-1}\Gamma'Y$.[24] The key difference is that rather than aggregating over linear projection coefficients $(\Gamma'\Gamma)^{-1}\Gamma'V$ for $V \in \{D, Y\}$ using the weights $D'\Gamma$ (which are governed asymptotically by the statistical distribution of D_i and Z_i), $\hat{\rho}(\lambda)$ uses weights $(0, \lambda')$, chosen to match the desired parameter of interest. Relative to 2SLS, $\hat{\rho}(\lambda)$ can be thought of as sacrificing some statistical efficiency in order to guarantee that it recovers a well-defined causal parameter under VM. In Section B.3.1 I discuss regaining some of that lost efficiency through regularization, which is borne out in the simulation in Appendix B.3. It bears emphasizing that with a large number of instruments, $\hat{\rho}$ is no more "expensive" than 2SLS, both involve computing a pair of linear projections with the same number of terms. This is despite the fact that the richness of possible selection behavior is more complex under VM than under IAM, scaling as Ded_J rather than $2^{\mathcal{J}}$.

Under regularity conditions (see Theorem B.1 in Appendix B.3), we will have that for any $\hat{\lambda} \xrightarrow{p} \lambda \in \mathbb{R}^{|\mathcal{F}|}$:

$$\hat{\rho}(\hat{\lambda}) \xrightarrow{p} \sum_{g \in \mathcal{G}^c} \frac{P(G_i = g)[M_J\lambda]_g}{\sum_{g' \in \mathcal{G}^c} P(G_i = g')[M_J\lambda]_{g'}} \cdot \Delta_g$$

Matching the RHS of the above to particular estimands Δ_c that satisfy Property M is achieved by choosing $\hat{\lambda}$. Table 2.4 gives natural sample estimators for ACL, SLATE, SLATT, SLATU and PTE that are consistent. Note that in the case of the ACL $\hat{\lambda}$ does not depend on the data and thus no "first-step" is necessary in estimation.

Regularization: Consider the ACL, and recall from Section 2.4.3 that it is equal to a single Wald ratio. A natural alternative Wald estimator of the ACL is thus:

$$\hat{\rho}_{\bar{Z},\underline{Z}} := \frac{\hat{E}[Y_i|Z_i = \bar{Z}] - \hat{E}[Y_i|Z_i = \underline{z}]}{\hat{E}[D_i|Z_i = \bar{Z}] - \hat{E}[D_i|Z_i = \underline{z}]} \tag{2.7}$$

where recall that under Assumption 3 $\bar{Z} = (1, 1, 1, \ldots, 1)'$ or $\underline{Z} = (0, 0, 0, \ldots, 0)'$. It turns

[24] The proof of Corollary 2.1 gives the basis transformation from a design matrix of indicators to Γ, which cancels in $\hat{\rho}_{2sls}$.

Parameter	Estimator $\hat{\lambda}$ of population λ
ACL	$(1,1,\dots 1)'$
$SLATE_{\mathcal{J}}$	$\hat{\lambda}_S = \mathbb{1}(\mathcal{J} \cap S \neq \emptyset)\hat{P}\left(Z_{S-\mathcal{J},i} = 1\right)$
$SLATT_{\mathcal{J}}$	$\hat{\lambda}_S = \mathbb{1}(\mathcal{J} \cap S \neq \emptyset)\hat{P}\left(Z_{S,i} = 1\right)$
$SLATU_{\mathcal{J}}$	$\hat{\lambda}_S = \mathbb{1}(\mathcal{J} \cap S \neq \emptyset)\hat{P}\left(Z_{S-\mathcal{J},i}(1-Z_{\mathcal{J},i}) = 1\right)$
$PTE_j(z^*_{-j})$	$\hat{\lambda}_S = \mathbb{1}(z^*_{-j,1} \cup j = S)$

Table 2.4: Estimators $\hat{\lambda}$ for the leading parameters of interest. $S - \mathcal{J}$ denotes the set difference $\{j : j \in S, j \notin \mathcal{J}\}$ and $z^*_{-j,1}$ denotes the set of instruments that are equal to one in z^*_{-j}.

out that $\hat{\rho}_{\bar{Z},\underline{Z}}$ and $\hat{\rho}((1,1,\dots 1)')$ in Equation 2.6 are in fact numerically equivalent in finite sample.[25] In situations where there is non-zero but small support on the points $\bar{Z}$ and $\underline{Z}$, we may thus expect that $\hat{\rho}((1,1,\dots 1)')$ may perform quite poorly as an estimator of ACL in small samples, since it effectively ignores all of the data for which $Z_i \notin \{\underline{Z}, \bar{Z}\}$. This issue is mentioned by Frölich (2007) in the context of IAM, in which case $\hat{\rho}_{\bar{Z},\underline{Z}}$ is also consistent for the ACL with finite $\mathcal{Z}$ (see footnote 22). Appendix B.3 develops and investigates the performance of a data-driven regularization procedure to ameliorate this problem, while also showing asymptotic normality of the estimator with or without such regularization. Appendix B.3 also reports a simulation study that shows the regularization procedure can indeed be helpful in practice.

2.6 Revisiting the returns to college

In this section I apply the results to study the labor market returns to college. In the past, this literature has based IV methods on either an assumption of homogeneous treatment effects, or the traditional IAM notion of monotonicity. Using the methods developed in this paper valid under VM, I find evidence of heterogeneous treatment effects across response groups, although statistical precision is an issue due to the small sample. This

[25] To see this, note that the vector H of H_i solves the system of equations $\Gamma' H_i = (1 \dots 1)'$. Among vectors that are in the column space of Γ, H is the unique such solution, given that the design matrix Γ has full column rank. One can readily verify that $\Gamma' H = (1,1,\dots 1)$ with the choice $H_i = \frac{\mathbb{1}(Z_i=(1\dots1))}{\hat{P}(Z_i=(0\dots0))} - \frac{\mathbb{1}(Z_i=(0\dots0))}{\hat{P}(Z_i=(0\dots0))}$, and that this $H = \Gamma\eta$ with $\eta = (1/\hat{P}(Z_i = (1\dots1)), 0, \dots 0, -1/\hat{P}(Z_i = (0\dots0)))$.

complements existing results that find evidence of heterogeneity, but are based upon IAM – a less plausible assumption in this context. For different choice of the instruments than I use, MTW present a test of IAM in this empirical setting and find evidence that it does not hold. In the Supplemental Material, I also present a second empirical application of my methods to the effects of children on labor supply.

2.6.1 Sample and implementation details

I use the dataset from Carneiro, Heckman and Vytlacil (2011) (henceforth CHV) constructed from the 1979 National Longitudinal Survey of Youth. The sample consists of 1,747 white males in the U.S., first interviewed in 1979 at ages that ranged from 14 to 22, and then again annually. The outcome of interest Y_i is the log of individual i's wage in 1991, and treatment $D_i = 1$ indicates i attended at least some college. As in CHV, treatment effects are expressed in roughly per-year equivalents by dividing by four.

CHV consider four separate instruments for schooling. In a baseline setup, I use the two binary instruments from our running example: tuition and proximity. A second setup then adds the remaining two instruments, which capture local labor market conditions when a student is in high school. The first two instruments are defined as follows: $Z_{2i} = 1$ indicates the presence of a public college in i's county of residence at age 14, while $Z_{1i} = 1$ indicates that average tuition rates local to i's residence around age 17 falls below the sample median, which corresponds to about $2,170 in 1993 dollars. This represents one particular choice of how the underlying continuous instrument from CHV can be discretized into a binary variable, but note that the methods in this paper could also be used with tuition recast as a discrete variable with a rich set of tuition levels. The Supplemental Material reports the distribution of the underlying tuition variable, whose definition is described further in CHV.

While VM is a natural assumption for the tuition and proximity instruments, a conditional version of instrument validity is more plausible than Assumption 1. Following

CHV, I assume:

$$\{(Y_i(1), Y_i(0), G_i) \perp Z_i\} \,|X_i \tag{2.8}$$

where X_i is a vector of observed covariates unaffected by treatment. Conditioning on X_i can help control for unobserved heterogeneity that may be correlated with location during teenage years. Appendix B.2 considers extensions of the basic identification and estimation results to include such covariates. The main result of the Appendix is that while conditional average treatment effects $\Delta_c(x) := \mathbb{E}[Y_i(1) - Y_i(0)|C_i = c, X_i = x]$ can be identified for each x in the support of X_i, the unconditional Δ_c turns out to be simpler to estimate, particularly when the two conditional expectation functions $\mathbb{E}[Y_i|Z_i = z, X_i = x]$ and $\mathbb{E}[D_i|Z_i = z, X_i = x]$ are additively separable between z and x. In this case, the only change required to the estimator presented in Section 2.5 is to "control" semiparametrically for X_i in the linear projections of Y_i and D_i onto the instruments. In particular, when

$$\mathbb{E}[Y_i|Z_i = z, X_i = x] = y(z) + w(x) \text{ and } \mathbb{E}[D_i|Z_i = z, X_i = x] = d(z) + v(x)$$

for some functions y, w, d and v, then a causal parameter Δ_c can be estimated as:

$$\hat{\Delta}_c = \frac{\sum_{z\in\mathcal{Z}} \left(\sum_{S\subseteq\{1\ldots J\}, S\neq\emptyset} \hat{\lambda}_S A_{S,z}\right) \hat{y}(z)}{\sum_{z\in\mathcal{Z}} \left(\sum_{S\subseteq\{1\ldots J\}, S\neq\emptyset} \hat{\lambda}_S A_{S,z}\right) \hat{d}(z)} \tag{2.9}$$

where the matrix A is defined in Corollary 2.1 to Theorem 2.1, the estimators $\hat{\lambda}_S$ are as given in Section 2.5, and $\hat{y}(z)$ and $\hat{d}(z)$ are consistent estimators of the functions $y(z)$ and $d(z)$. Note that as the vector Γ_i contains a full set of interactions between the binary instruments, both $y(z)$ and $d(z)$ are automatically linear in Γ_i. If the functions $w(x)$ and $v(x)$ are taken to also be linear in x, Equation 2.9 can be reduced to a simple generalization of the estimator from Section 2.5: $\hat{\Delta}_c = \left((0, \hat{\lambda}')(\Gamma'\mathcal{M}_X\Gamma)^{-1}\Gamma'\mathcal{M}_X D\right)^{-1}(0, \hat{\lambda}')(\Gamma'\mathcal{M}_X\Gamma)^{-1}\Gamma'\mathcal{M}_X Y$ where $\mathcal{M}_X$ is a projection onto the orthogonal complement of the design matrix of X_i. I follow this strategy, computing standard errors by applying the delta method to the system of regression equations (one each for D_i and Y_i, along with a regression on a constant for each component of $\hat{\lambda}$), allowing for heteroscedasticity and cross-correlation between the equations.[26]

I follow CHV and use as control variables a student's corrected Armed Forces Qualification Test score, mother's years of education, number of siblings, "permanent" local earnings in county of residence at 17, mother's years of education, number of siblings, "permanent" unemployment in county of residence at 17, earnings in county of residence in 1991, and unemployment in state of residence in 1991, along with an indicator for urban residence at 17, and cohort dummies. The definition and construction of these variables is described in CHV. Also following CHV, squares of the continuous control variables are included in X_i, relaxing the assumption of strict linearity in each. The above variables represent the union of variables that CHV use in their first stage and outcome equation, with one exception: I drop years of experience in 1991 since it may itself be affected by schooling, as MTW do as well in their empirical application. In the two instrument setup, I also add to X_i the two "unused" instruments from CHV and their squares: long-run local earnings in county of residence at 17 and long run permanent unemployment in state of residence at 17.

2.6.2 Results from baseline setup with two instruments

The left panel of Table 2.5 reports a cross tabulation of the two instruments. As noted, the observations are relatively evenly distributed across the four cells. The instruments are positively correlated, with a Pearson correlation coefficient of about 0.13.

The right panel of Table 2.5 reports the conditional propensity score function $\mathbb{E}[D_i|Z_i = z, X_i = x]$ estimated as described above and averaged over the empirical distribution of

[26]Note that while Appendix B.3 Theorem B.1 provides a variance expression for $\hat{\rho}\hat{\lambda}$, this does not cover the case with covariates, so I do not implement an estimator based upon it here. Also, as the distribution of Z_i is fairly well balanced across the four cells of $\mathcal{Z}$, I do not implement the regularization procedure proposed in Appendix B.3.

Distribution of the instruments

		Z_2="close"	
		0	**1**
Z_1="cheap"	**0**	469	401
	1	361	516

Mean fitted propensity scores

		Z_2	
		far	**close**
Z_1	**expensive**	0.451	0.509
	cheap	0.487	0.530

Table 2.5: Left: number of observations having each pair of values of the instruments, with total sample size $N = 1,747$. Right: fitted propensity scores estimated by linear regression, evaluated at the sample mean of the X_i variables.

X_i (in practice, evaluated at the mean of X_i). This allows us to take the $(expensive, far)$ cell 45.1% as an estimate of the overall proportion of never-takers in the population, while the share of never-takers is estimated to be 47.0%. The remaining roughly 8% of the population are generalized "compliers" consisting of the tuition (Z_1), proximity (Z_2), eager and reluctant compliers. From the table we can also see that $P(D_i(expensive, close, x) > D_i(expensive, far, x)) \approx 5.7\%$, and $P(D_i(cheap, far, x) > D_i(expensive, far, x)) \approx 3.6\%$. Combining these figures and the response group definitions from Section 2.3, we see that between 1.5% and 3.6% of the population are eager compliers, while no more than 2.1% are reluctant compliers. Similarly, no more than 3.6% are tuition compliers, and between 2.1% and 5.7% are proximity compliers. Overall, the data are compatible with a roughly even split between the four groups, but it is also possible that proximity compliers account for more than half of all generalized compliers.

We now turn to treatment effect estimates. Figure 2.2 reports estimates of several of the parameters introduced in Section 2.4, alongside fully-saturated 2SLS for comparison. Consider first the all- compliers LATE (ACL): the point estimate of 0.14 indicates that having attended a year of college increases 1991 wages of all compliers by roughly 14% on average. This estimate is within the range of roughly -0.1 to 0.3 of the marginal treatment effect (MTE) function estimated by CHV under the assumption of IAM, and is similar to their point estimate of the average treatment on the treated under a parametric normal

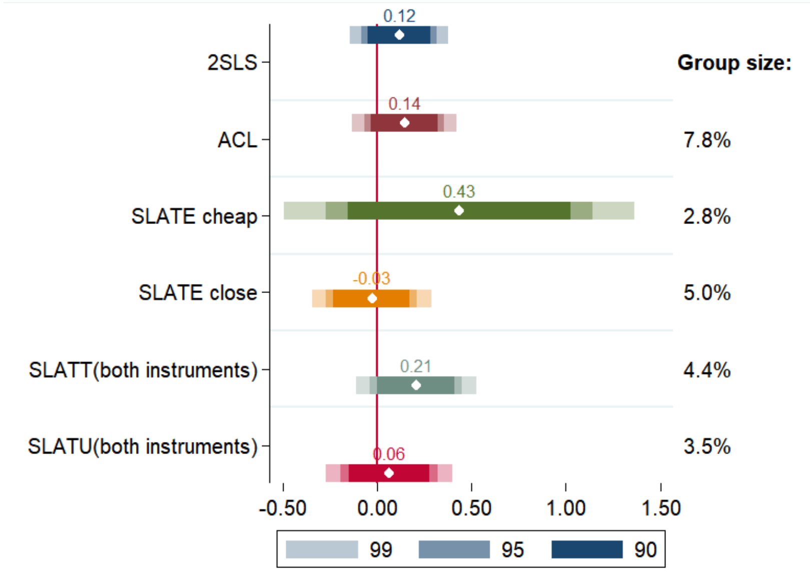

Figure 2.2: Estimates of various causal parameters identified under VM with two instruments, alongside fully-saturated 2SLS for comparison. Bars indicate 95% confidence intervals, and "Group Size" refers to the identified quantity $P(C_i = 1)$ for each parameter

selection model. The 2SLS estimate from Figure 2.2 yields a similar value at 0.12. Note that given the limited sample size none of the estimates are quite significant at even the 90% level. I thus focus discussion on the point estimates for the sake of illustration with this important caveat.

The point estimates from the remaining rows in Figure 2.2 suggest that the ACL aggregates over substantial heterogeneity in the population. For example, the distance SLATE suggests that a year of college has no average effect on the wages of individuals who move into treatment if and only if a college is nearby, given local affordability. Recall that this group includes proximity compliers, eager compliers for whom college is expensive, and reluctant compliers for whom it is cheap. On the other hand, the SLATE for tuition is about three times as large as the ACL. These results are suggestive that the average treatment effect among tuition compliers is larger than it is among proximity compliers, however the sign of the difference is not identified.[27] Note finally that the point estimates

[27]In the Supplemental Material I show in the $J = 2$ case that if Δ_g and corresponding group size p_g is known for one group $g \in \mathcal{G}^c$ ex- ante, then the remaining three group specific treatment effects and group sizes can be identified.

for $SLATU$ and $SLATT$ suggest that among the compliers averaged over by the ACL, those who in fact go to college have greater treatment effects on average than those who do not, which is consistent with some students selecting on the basis of their future gains.

2.6.3 Results with all four instruments

I now add the additional two instruments from CHV, to increase comparability and emphasize the scalability of the proposed methods to multiple instruments.

Accordingly, we let Z_{3i} indicate that local earnings in i's county of residence at 17 is below the sample median, and Z_{4i} indicates that unemployment in i's state of residence at 17 is above the sample 25% percentile. This threshold is chosen as it yields a stronger first stage as compared with the median. The two local labor market variables and their squares are removed from the vector of controls X_i. Vector monotonicity implies that the propensity score is component-wise monotonic in the four instruments, implying 32 linear inequalities among first stage coefficients. Although not reported here, t-statistics are positive for all but six of these hypotheses, and none is rejected at the 10% level.

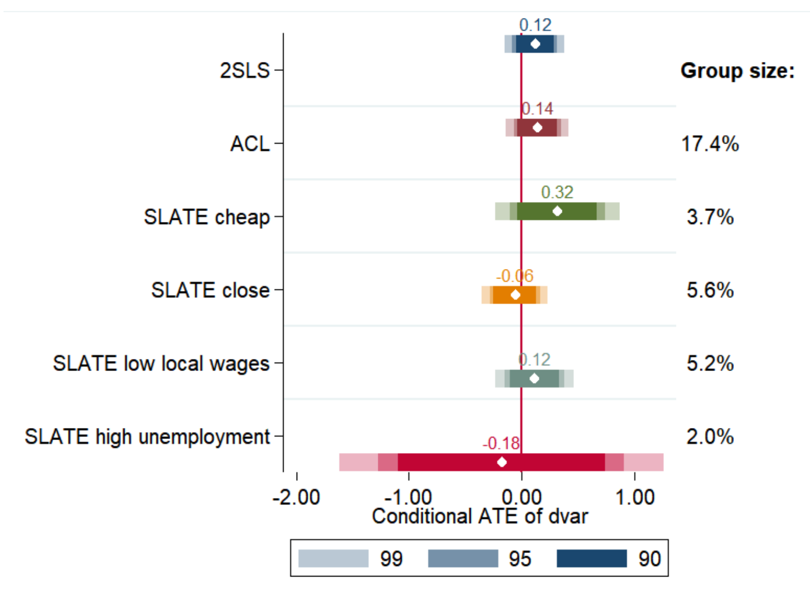

Figure 2.3: Estimates of various causal parameters identified under VM with all four instruments, alongside fully-saturated 2SLS for comparison. Bars indicate 95% confidence intervals, and “Group Size” refers to the identified quantity $P(C_i = 1)$ for each parameter.

Table 2.3 shows that the ACL is not appreciably changed from the case with only two instruments, and we again have that the tuition SLATE is much larger and the proximity SLATE close to zero. The SLATE for low local wages occupies an intermediate value, while the SLATE for high unemployment is estimated to be negative (suggesting that more schooling reduces wages), but with a much larger standard error. The unemployment SLATE is so imprecisely estimated in part because its corresponding complier group is the smallest of the estimands considered: with just 2% of the population.

To compare these results more directly with CHV, recall that the marginal treatment effect function (e.g. Heckman and Vytlacil 2005) is defined as

$$MTE(u, x) := \mathbb{E}[Y_i(1) - Y_i(0)|U_i = u, X_i = x]$$

where U_i is a uniformly distributed heterogeneity parameter that can be thought of as a proclivity against treatment in the selection model $D_i(z, x) = \mathbb{1}(P(z, x) \geq U_i)$, with $P(z, x) := \mathbb{E}[D_i|Z_i = z, X_i = x]$ the propensity score function. CHV estimate the MTE function evaluated at the mean of x to decrease monotonically with u over the unit interval. For each instrument Z_j, call i a "j-responder" if $D_i(1, Z_{-j,i}, X_i) > D_i(0, Z_{-j,i}, X_i)$. In the context of a model in which both IAM and VM hold, the estimates in Figure 2.3 coupled with CHV would thus suggest that tuition responders tend to have the lowest unobserved costs U_i, followed by wage responders, then proximity responders, and then unemployment responders. However, while IAM effectively "flattens" variation in any of the instruments into variation in the scalar parameter $P(Z_i, X_i)$, VM allows flows into treatment to depend in an essential way on which instrument is manipulated when IAM fails. The estimands in Figure 2.3 are directly relevant to hypothetical policies which vary that instrument alone.

The results in Figure 2.3 can also be compared with estimates reported by Mogstad et al. (2020b) that are calculated by 2SLS. While their empirical application focuses on the interpretation of 2SLS under PM or VM, we have seen that in this particular setting 2SLS tends to yield numerical estimates that are close to the ACL. Similarly, the SLATEs for the proximity and low local wage instruments in Figure 2.3 align roughly with 2SLS specifications in MTW in which a single instrument is excluded in the second stage. However this similarity will not hold in all contexts, underlying the importance of methods such as those presented in this paper or in Mogstad et al. (2020a). Appendix B.3 provides simulates a data generating process for example in which 2SLS lies outside of the convex hull of treatment effects in the population.

Finally, observe that in this four instrument setup, there are in principle 167 underlying response groups aside from always- and never-takers, and that together these comprise 17.4% of the population (cf. 7.8% for the four such groups with two instruments). Nevertheless, computing the treatment effect estimates involves regressions with at most 16 terms in addition to the controls, keeping implementation manageable. Note that while the standard errors for the 2SLS estimate are only slightly smaller than for the ACL, this is sufficient for significance at the 95% level even in this small sample. This in part reflects the fact 2SLS weighs across the groups to minimize variance rather than pin down a specific target parameter.

2.7 Conclusion

In both observational and experimental settings, it is natural to expect individuals to vary both in their treatment effects and in how they select into treatment. This latter type of heterogeneity is likely to be particularly pronounced when a researcher is using multiple instrumental variables for a single binary treatment. This paper has shown that causal inference with heterogeneous treatment effects is possible in such settings under a simple restriction on selection that is often motivated by economic theory: what I call vector monotonicity.

In particular, I have defined and characterized a class of interpretable causal parameters that can be point identified under vector monotonicity with discrete instruments, and proposed an estimator that is similar in construction to the familiar method of two stage least squares (2SLS). While the convenience of implementing the two estimators scales similarly with the number of instruments, 2SLS is not guaranteed to recover an interpretable causal parameter under vector monotonicity (though it may in special cases). By contrast, the estimator I propose is always targets a particular well-defined causal parameter. In an application to the labor market returns to college education, I find that estimates based on vector monotonicity suggest that underlying groups in the population that exhibit different selection behavior also have highly heterogeneous treatment effects.

Chapter 3: The Career Impact of First Jobs: Evidence and Labor Market Design Lessons from Randomized Choice Sets

3.1 Introduction

An individual's first job may have important consequences for her career trajectory. This view—common among researchers—appears widely held also among those entering the labor market. New job-seekers may therefore put weight on the expected impact of different jobs on their trajectories when choosing a first job to pursue. Policy, on the other hand—whether centralized mechanisms allocating doctors, teachers, and other groups of workers serving the public to first jobs[1], or the rules and regulations that influence initial worker-job matches in the decentralized labor market—is typically designed without accounting for expected "first job effects" (FJEs).

If FJEs are non-negligible in magnitude and heterogeneous across types of workers, then FJE-responsive policy design could in principle be used to increase welfare. However, even in such a scenario, whether alternative policies *actually* affect initial worker-job matches—and hence realized FJEs—differentially is an empirical question. Unusual types of data and variation are necessary for researchers to be able to identify FJEs and graduates' and policymakers' actual and ideal response. To estimate the causal, long-term effect of an individual's first job, random variation in her match, holding all else constant, is needed. To estimate how individuals' distribution of FJEs across jobs influence their job search choices, causal estimates of the long-term effect of each type of job for each type of

[1] The following is an incomplete list of countries that use centralized mechanisms to assign workers in some (in some of the countries, almost all) public service occupations to first jobs: Australia, Bangladesh, Bhutan, Botswana, Canada, Denmark, France, Ghana, India, Iran, Ireland, Israel, Italy, Japan, Malaysia, Malta, Nepal, Norway, Pakistan, Philipines, Saudi Arabia, Senegal, Singapore, South Africa, South Korea, Taiwan, Tanzania, Uganda, U.K., USA.

individual—and knowledge of individuals' choice set when entering the labor market—are needed.

In this paper we take advantage of Norway's 1997-2013 allocation of doctors' first job—their residency—through a Random Serial Dictatorship (RSD) mechanism[2], and the replacement of the RSD with decentralized job-finding in 2013, to overcome these challenges. We first estimate the consequences for earnings, place of residence, and specialization in the long-term of each type of job characteristic separately for men and women. We do this by exploiting RSD-generated random, individual level variation in new doctors' choice sets over first jobs. In the last part of the paper, we use a 2013 policy change—which replaced the the RSD system with a regular, decentralized job market for new doctors—to assess how total worker welfare, initial worker-job matches, and the associated realized FJEs, differ in a market system relative to RSD.

The unique suitability of doctors' residencies in Norway for studying FJEs is due to the combination of choice sets over jobs being assigned randomly, and the unusually high quality of the registry data available on the universe of Norwegian workers. While our quantitative results may not generalize to other occupations, it is worth noting that (i) the differences between the possible pathways a doctor's career can take share many features with those in other occupations[3], and (ii) the literature generally finds that highly skilled workers are least affected by temporary career shocks (von Wachter and Bender, 2006; Oreopoulos et al., 2012). Most likely our results thus represent a lower bound on FJEs and the associated responses in other occupations.

This paper contributes to the literature on how temporary shocks to a worker's employment status affects her career trajectory. Existing studies have convincingly and care-

[2]A Random Serial Dictatorship mechanism starts with a lottery. The person who draws number 1 then chooses her preferred object freely among all available options. After that, the person who draws number 2 chooses among the remaining objects, and so on.

[3]For example, doctors' jobs are located in many different parts of a country; there is considerable dispersion in employer size and "quality" (which is highly correlated with doctors' earned income); and there are ample opportunities for doctors to undertake horizontal specialization (choice of medical field) as well as vertical specialization (e.g. becoming a specialist as opposed to a General Practitioner).

fully documented the consequences of job displacement (see, among many others, von Wachter and Bender, 2006; Sullivan and von Wachter, 2009; Bender et al., 2009; von Wachter, 2020); exposure to high unemployment rates later in life (Coile et al., 2012); regional labor demand composition (Arellano-Bover, 2020); and, most closely related to this paper, graduating in a weak labor market (Oyer, 2006; Oyer, 2008; Genda et al., 2010; Kahn, 2010; Heisz et al., 2012). These influential studies have shown how cohort and group-level labor market shocks affect individual workers' trajectories.

In addition to taking advantage of an explicit randomization for identification, this paper to our knowledge provides among the first causal evidence on the career consequences of *individual* level shocks to a graduate's first job.[4] The distinction is essential because cohort level studies may not be informative about the career consequences of individual level labor market shocks, which are ubiquitous. When the cohort or group an individual belongs to is hit, for example, by a recession or a mass layoff, then the individual's peers are also affected. Peers' exposure to the shock could adversely affect the trajectory of the individual in question (if for example she now faces more competition for current jobs) or benefit her (if for example she's now competing against less employable other workers for future jobs).[5]

To leverage the RSD to estimate first job effects, we develop an instrumental variables (IV) approach that explicitly allows for the effect of beginning one's career at a given employer to vary by individual. This setting extends beyond standard results in the treatment effects literature, as a worker's first job "treatment" is the identity of their first-job employer. We thus contribute to a recent literature that extends IV research designs to unordered, multivalued treatment variables (**Lee2018a**; Heckman and Pinto, 2018; Lee

[4]To our knowledge, the only existing causal evidence on the long-term effects of individual level shocks to first jobs comes from Angrist (1990)'s seminal study of the Vietnam draft. He shows that being drafted lowered earnings by 15 percent long after the veterans' service ended (see also Angrist, 1998; Angrist and Chen, 2011). Staiger, 2021 use job openings at the employer of young worker's parents to generate individual-level variation in early job opportunities.

[5]Ruhm, 2000 shows that mortality tends to improve during recessions, while Sullivan and Wachter, 2009 show that *own* job displacements increase mortality for U.S. workers.

and Salanié, 2020). Furthermore, relative to recent work that has used other centralized assignment mechanisms for causal inference (Abdulkadiroğlu et al., 2017; Kirkeboen et al., 2017), we do not observe workers preference orderings over outcomes of the matching mechanism. We show that first job effects can still be partially identified, by making use of the observation that variation in mean outcomes across choice sets is informative about heterogeneity in the counterfactual outcomes for a given employer across workers of different unobserved preference types. This builds upon a recent linear programming approach to IV analysis (Mogstad et al., 2018; Kamat, 2020), as well as a recent literature on causal inference with a collection of distinct instrumental variables (Mogstad et al., 2020a; Goff, 2020).

With estimates of FJEs in hand, our final contribution is to assess the impact of the 2013 reform to a decentralized labor market on the welfare of workers. To do so, we decompose preferences over employers into a component that is due to first job effects and another that is due to the "amenity value" workers of a given type associate with employers of a given type,[6]. This leverages lottery draws as a reduced form measure of preferences, coupled with a high-level assumption on the distribution of preferences in the lottery. We show how realized first job effects, amenity values, and overall worker welfare differ, for each group and in total, in a decentralized labor market compared to the RSD system, by examing how wib by examining how worker-employer matches changed after 2013.

The paper is organized as follows. In Section 3.2 we discuss background on the setting and institutional setup. In Section 3.3 we present the datasets used in our empirical analysis. In Section 3.4 we lay out our instrumental variables strategy to estimating first job effects, and present results in Section 3.5. Section 3.6 then applies these results to investigate worker preferences and evaluate the reform to a decentralized labor market from those workers' perspective.

[6]This paper is of course not the first to recognize that workers care about non-income job characteristics and may choose occupations and employers partly based on those preferences. Recently, for example, Sorkin (2018) showed evidence of compensating differentials revealed in workers' job-to-job transitions in the U.S.

3.2 Setting

3.2.1 The Random Serial Dictatorship mechanism for Norwegian doctors

The "turnus" (roster) system that was used to match medical graduates with residency positions in Norway from 1954 to 2013 was a Random Serial Dictatorship (RSD) mechanism. Theorists have shown that, among other important properties, the RSD is incentive-compatible, inducing participants to reveal their true preferences (Abdulkadiroğlu and Sönmez, 1998).

Equitable access to primary healthcare across regions was the main motivation behind the use of a lottery system in Norway. Like other countries, Norway had had trouble filling doctor vacancies in rural areas, and the RSD mechanism was expected to distribute the best doctors more equally across space.[7] In addition, the mechanism appealed to policymakers because it was perceived to be fair to the participating medical graduates.[8]

First, graduating students would enter a lottery, either in February or in August, and be assigned a random draw number. Next, the student with the lowest draw number would choose freely between all available positions. Then, the student with the second lowest draw number would choose from the remaining residency positions. This would continue until the student with the highest draw number remained, who would take whichever spot was available.[9]

Three categories of new doctors received special treatment: couples, who were allowed to draw a shared lottery draw number and to choose residencies simultaneously; doctors with children; and doctors with maternity or health issues. The latter two categories were allowed to choose between positions deemed especially suitable for them before the lot-

[7]Such considerations have been studied in the context of the US National Medical Residency Matching Program, see e.g. (e.g. Roth, 1984; Roth, 1986).

[8]The government also wished to incentivize doctors to work in rural locations in other ways. For instance, doctors who agreed to intern at hospitals in the largely rural counties of Sogn og Fjordande and Finnmark could skip the lottery entirely.

[9]If the number of students exceeded the number of residency positions, the unassigned students would get priority in the next lottery.

tery took place. Since these three types of doctors were not subject to randomization via the lottery, we exclude them from our analysis.

In the late 2000s, the system began to concern the government, because of the growth of the number of applicants and the rise in proportion of students from foreign universities.[10] The number of medical graduates participating in the lottery would routinely exceed the number of training positions available. As a result, it became increasingly difficult for the government to guarantee a six-month maximum waiting time to obtain a residency. In 2013, the Norwegian Health Minister replaced the lottery system with direct qualification after six years of medical school. Medical graduates now apply to residencies directly, as in a regular labor market, and hospital trusts are responsible for selection and recruitment.

3.2.2 Doctors in Norway 1993-2017

This section profiles doctors that worked in Norway during the study period; we go through the data we use in detail in Section 3.3. Medical students in Norway begin their studies in the Fall or Spring semester, and usually take ten semesters to graduate. Starting in the 1950s, the Norwegian government mandated an eighteen month residency period, after which medical school graduates could become fully licensed physicians and practice independently. The first twelve months were to be spent at a hospital, while the remaining six months were to be spent as a General Practitioner (i.e. one who works in Primary Care) within the same county.[11]

Table C.1 summarizes a range of socioeconomic information on doctors including age, proportion born abroad, proportion that studied abroad, family size, field of specialization and income and assets. The last two columns summarize this information separately for men and women. Women comprise over 40 percent of doctors, and tend to be over-represented in fields like gynecology and psychiatry. Male doctors are older and tend to

[10] Norway was compelled by its participation in the EU common labor market system to accept any European medical graduate who could pass a Norwegian language test into the system.

[11] The last six months could be spent at an institution that was disjoint from the hospital of the first twelve months.

be over-represented in fields like surgery and internal medicine. A fifth of all doctors were born abroad, of which an overwhelming proportion are citizens of Denmark and Sweden. Finally, recorded income, asset and debt holdings are higher on average for male doctors.

There are 30 basic medical specialties, and specialization is usually in the form of training on the job.[12] The average length of time required to complete a specialty is five years, but it can take longer with large variations between the specialties. Figure C.4 indicates that there appear to be substantial returns to specializing—job retention rates are higher for specialists, and the salary bump from specialization increases with age.

3.2.3 Hospitals in Norway 1993-2017

The employer-employee database contains information on all registered employers that employ doctors in Norway.[13] Figure C.5 depicts the steady growth in both the number of hospitals and average hospital size (number of doctors employed) since 1995.

Hospitals vary across multiple dimensions. Table C.1 summarizes information on salaries, geographical remoteness, number of doctors and other medical staff, proportion of specialists, as well as the presence of fifteen distinct specialist fields. Most hospitals in Norway are located in urban municipalities; on average, municipalities with hospitals have only 10 per cent of their population living in rural areas. This is noteworthy because the average municipality in Norway had 49 percent of its population living in rural areas.

3.3 Data

We combine information on lottery outcomes with Norwegian administrative data from 1993 to 2017. We obtained information on lottery draw numbers for all lottery participants who were assigned a residency position during 1993-2013 from the Norwegian Registration Authority for Health Personnel (SAFH). This information was linked

[12]The Norwegian Medical Association evaluates whether the candidate has met the requirements to become a specialist. Specialist titles are formally awarded by the Health Directorate.

[13]We define hospitals as employers that hire at least 10 doctors. These account for around 80 per cent of doctor employment.

with the employer-employee registry to match medical graduates to their residency hospitals, as well as employer information in the years following the residency. This data was then linked to administrative registers provided by Statistics Norway, a rich longitudinal database that includes information on medical graduates' socioeconomic information (sex, age, marital status, educational attainment, specialization, income, and gross wealth), geographical identifiers and year-end asset holdings and liabilities (such as real estate, stock holdings, etc) for each year. These data have several valuable attributes. There is no attrition from the sample, and most components of income and wealth are third-party reported without any top or bottom coding.

The final dataset tracks the career path of each graduate, starting with her lottery number and choice of residency hospital. After excluding people belonging to special lottery categories and hospitals with missing information, we end up with a sample of about 9000 individuals and 55 hospitals, which participate in 34 lotteries from 1996 to 2012.[14] Figure C.6 displays the number of individuals and hospitals that participated in each lottery. Figure C.7 splits participants by gender and by birth location. It is evident that there is an increase in the proportion of women and foreign students over time. Most foreign doctors are citizens of the European Economic Area (EEA).

We observe employment outcomes for all doctors up until the year 2017. This allows us to track doctors who graduated in the earliest lotteries (during the 1990s) for over fifteen years, while participants in the last few lotteries (in the 2010s) can only be tracked for a few years. Data from the 2013-2017 period is used to analyze the last cohorts in the lottery system, as well as to observe hospital matches after the 2013 reform. Figure C.8 displays the distribution of the number of times doctors are observed in the years following their residencies. Our sample consists of roughly 4500 individuals five years after their residency, but less than a quarter of these are observed 10 years down the line.

We construct the choice set of hospitals faced by each lottery participant using her lot-

[14] Data is missing for the lottery in January 1998.

tery number and the residency hospitals chosen in that lottery. We know that if a hospital h was chosen by someone with a higher (worse) lottery number than individual i, i must have been given the option of choosing h as well (since hospitals cannot reject applicants). Assuming that no residency spots were left unfilled,[15] we can thus impute the choice sets C_i that were offered to each lottery participant. Most medical graduates have a sizable number of residency options to choose from, as displayed in Figure C.8.

3.4 Identification of first job effects

In this section we describe how we use these RSD-generated choice sets to generate instruments for a doctor's residency hospital. To abstract somewhat from our specific context, we will typically refer to "workers" choosing a first-job "employer", rather than "doctors" choosing a residency "hospital".

The basic intuition behind our approach is that randomization of the RSD lottery ensures that each worker's choice set is is independent of her observable and unobservable characteristics. At the same time, this choice set affects that worker's realized first-job, which is constrained to be within the randomly assigned choice set. This allows us to use the RSD lottery to construct instruments for first job effects (FJEs) that are both exogenous and relevant.

We work in three steps. First, we show in subsection 3.4.2 that FJEs are identified in the RSD without any substantive assumptions on workers' selection behavior—but that this strategy requires conditions on the support of the data that do not hold in our context. We then show in subsection 3.4.3 that FJEs can be identified under weaker conditions on the support of the data—but that this strategy requires strong assumptions on workers' selection patterns. This culminates in subsection 3.4.4, which simultaneously relaxes both assumptions at the expense of yielding bounds rather than a point identification result.

[15] This assumption is reasonable, in part because excess demand was one of the reasons for replacing the turnus system after 2013.

3.4.0 Notation

We begin by establishing some notation that will be used throughout this section. Let h denote employers, $\mathcal{H}$ the set of all employers, and i workers in population $\mathcal{I}$. Let $Y_i(h, c)$ be the potential outcome (e.g. earnings four years later) of worker i if their first job is at employer h, and their choice set from the lottery was c. We assume that choice sets do not effect outcomes except through a doctor's first job employer, that is: $Y_i(h, c) = Y_i(h)$ for all $i \in \mathcal{I}$, $h \in \mathcal{H}$, and $c \subseteq \mathcal{H}$. Since choice sets play the role of instruments, this constitutes the standard IV exclusion restriction in our context.

Let $C_i \subseteq \mathcal{H}$ be a worker's realized choice set, and H_i their realized choice from C_i. Let n be the number of workers and $J = |\mathcal{H}|$ the number of employers. Let $L_i \in \{1 \ldots 34\}$ be an identifier for the lottery among the 34 between 1996 and 2013 in which worker i was allocated their first job. Let $\mathcal{R}_i \in \{1, 2, \ldots\}$ denote worker i's random place-in-line in their lottery, and let $R_i = F_{\mathcal{R}|L=L_i}(\mathcal{R}_i)$ be this lottery draw normalized to the unit interval within each lottery.

We will assume that each worker has a complete preference relation over hospitals, and is indifferent between no two hospitals. We denote by $h \succ_i h'$ if i prefers h to h', and let $\succ_i$ alone denote i's entire preference relation over $\mathcal{H}$. For any choice set $c \subseteq \mathcal{H}$ denote the most-preferred $h \in c$ according to $\succ$ as $H_{\succ}(c)$, and write $H_i(c) = H_{\succ_i}(c)$ for shorthand. Note that θ_i is isomorphic to the vector $\{H_i(c)\}_{c \subseteq \mathcal{H}}$. Similarly, let $\mathcal{Y}_i$ be defined in isomorphism with the vector $\{Y_i(h)\}_{h \in \mathcal{H}}$.

We will often treat $\succ_i$ and C_i as random variables, although a realization of each is a set and a relation on a set, respectively. However, in light of the isomorphism mentioned above, we can view each preference relation $\succ_i$ as element of $\mathbb{Z}^{2^J}$, with integers indicating the index of $H_i(c)$ for each $c \subseteq \mathcal{H}$. Similarly, we can also view any choice set C_i as an element of $\{0, 1\}^{\times J}$, with each component indicating the presence or absence of a hospital $h \in \mathcal{H}$.

3.4.1 Exogeneity of choice sets

The randomization of each lottery ensures that $\mathcal{R}_i$ is independent of potential outcomes $Y_i(h, c)$, conditional on lottery L_i (we will sometimes use the term "cohort" interchangeably). However, with a finite number of workers participating in the lottery, it does not immediately follow that a worker i's probability distribution over possible choice sets C_i is perfectly independent of her characteristics.[16] Rather, any two workers having *the same preferences* $\succ$ will receive a C_i drawn from the same probability distribution within a cohort (Abdulkadiroğlu et al. 2017). Thus choice sets are independent of potential outcomes, conditional on preferences (and lottery). Since preferences are unobserved in our data, we cannot directly control for them.

Let us partition the population of workers into a set of groups $g \in \mathcal{G}$ on the basis of observable demographic variables G_i. These groups will play a central role in our analysis, allowing us to examine observable heterogeneity in first-job effects. We will assume that after conditioning on instance of the lottery and a value of g, choice-sets are as good as randomly assigned:

Assumption 3.1 (independence of choice sets).

$$\{(\mathcal{Y}_i, \succ_i) \perp C_i\} \mid (G_i, L_i)$$

One instance in which Assumption 1 would follow directly from randomization of the lottery is if preferences were perfectly homogeneous within each value of g, e.g. all Norwegian men have the same ranking over employers.[17] In this case, Assumption 3.1 echoes Proposition 1 of Abdulkadiroğlu et al. (2017) and first-job effects could be assessed

[16] To see this, consider a small economy in which there are two workers and two employers, each with one spot available. Worker 1 prefers employer A to B, and Worker 2 prefers B to A. Worker A then has a 50% chance of having $C_i = \{A, B\}$ (if she goes first in the lottery), and a 50% chance of having $C_i = \{A\}$ (if B chooses first). By contrast, Worker B has a 50% chance of having $C_i = \{A, B\}$ and a 50% chance of having $C_i = \{B\}$. The probability distribution over choice sets facing worker i depends on the preferences of each worker *except for i*.

[17] We can formalize randomization of the lottery as the statement that $\{\mathcal{R}_i \perp (\mathcal{Y}_i, \succ_i, G_i)\}|L_i$.

without the use of instrumental variables. Indeed, such preference homogeneity along with Assumption would imply that conditional on G_i, a worker's actual choice of first-job H_i is exogenous, since then $\succ_i$ has a degenerate distribution conditional on G_i and $H_i = H_{\succ_i}(C_i)$.

However, we will not maintain the strong assumption that available observable proxies are sufficient to control for unobserved preferences $\succ_i$. Rather, we offer a second justification for Assumption 3.1. When the number of workers is "large" in comparison with the number of employers, independence will hold approximately even when there is heterogeneity in preferences within groups. As described formally in Appendix C.2, the actual set of n workers is viewed as a sample from an underlying continuum of workers, with each employer accounting for a fixed proportion of the available jobs. In this "continuum economy", choice sets are random unconditionally. The IV estimators we use are then consistent along an asymptotic sequence in which $n \to \infty$ with the share of jobs belonging to each employer fixed. In Appendix C.2 we provide evidence that this asymptotic approximation is a good one in our context, by simulating the lottery many times with a number of workers and employers chosen to match our dataset, and plausible heterogeneity in preferences.

3.4.2 Choice sets as instruments

As a parameter of interest, we are interested in the quantity

$$\mu_{gh} := \mathbb{E}[Y_i(h)|G_i = g]$$

for an individual employer h and demographic group g. The parameter μ_{gh} is the average counterfactual outcome that would occur for a worker in group g if their first job were at employer h, and $\mu_{gh'} - \mu_{gh}$ is the average effect of "moving" workers in g from employer h to employer h'. Note that μ_{gh} generally differs from the observable $E[Y_i|H_i = h, G_i = g] = E[Y_i(h)|H_i = h, G_i = g]$, which unlike μ_{gh} further conditions on the worker's endogenous choice of employer h. Given that randomization of choice sets holds only when conditioning on lottery $L_i = \ell$, we also define a lottery-specific analog of μ_{gh} that will be useful as an intermediate quantity:

$$\mu_{gh\ell} := \mathbb{E}[Y_i(h)|G_i = g, L_i = \ell]$$

With Assumption 3.1 in hand, we seek to use features of a worker's choice set C_i to construct instruments for the causal effect μ_{gh} of her first job. Assuming that a worker will always choose some employer from their choice-set, we can in principle identify first job effects μ_{gh} without placing any further restrictions on selection. The following Proposition is a consequence of Theorem 1 in (Goff, 2020); however it admits of a very simple proof that we present here.

Proposition 3.1 (impractical identification). *Make Assumption 3.1. If for a given h, ℓ: $P(C_i = \{h\}|G_i = g, L_i = \ell) > 0$, then*

$$\mu_{gh\ell} = \mathbb{E}[Y_i|C_i = \{h\}, G_i = g, L_i = \ell]$$

provided that $\forall i$, $H_i(\{h\}) \neq \emptyset$.

Proof. Given that $H_i \in C_i$ and $H_i = H_i(C_i) \neq \emptyset$, we must have $H_i = h$ whenever $C_i = \{h\}$, so: $\mathbb{E}[Y_i|C_i = \{h\}, G_i = g] = \mathbb{E}[Y_i(h)|C_i = \{h\}, G_i = g] = \mu_{gh}$, where the last equality follows from Assumption 3.1. □

Proposition 3.1 shows that if there is some probability that each singleton $\{h\}$ emerges as a worker's choice set, then the μ_{gh} are identified. This requires us to assume that the worker will choose each h over no employer (all hospitals are preferred to the relevant outside option), but requires no other assumptions on workers' choices. For instance, it does not require us to assume that worker's choose rationally, according to well-defined preferences.

In practice however, the event $C_i = \{h\}$ is unlikely to occur for popular employers h, and even for an unpopular h it will only occur at most for the last few workers in a given lottery. As a result, Proposition 3.1 is not directly useful in estimation. In the next section, we thus consider a more practical route to identification under a restriction on heterogeneity, before returning to the general case in a partial identification framework in Section 3.4.4.

3.4.3 Identification with a restriction on treatment effect heterogeneity

The discussion of the last section has shown that first job effects μ_{hg} are identified even with complete heterogeneity of treatment effects, provided that there is a positive probability that some workers will face a choice-set containing only the single employer h. In practice, this result is not immediately useful given our moderately sized sample.

To make estimation tractable, we first impose the following restriction on treatment effect heterogeneity (cf. Kolesar 2015):

Assumption 3.2 (limited selection on gains). *For any h_0, h_1, h, c, g, the quantity:*

$$\mathbb{E}[Y_i(h_1) - Y_i(h_0) | H_i = h, C_i = c, G_i = g, L_i = \ell]$$

depends only on h_1, h_0, and g.

Assumption 3.2 states that for any pair of employers h_0 and h_1, the contrast $Y_i(h_1) - Y_i(h_0)$ is not correlated with actual employer choice H_i within a group and lottery. This rules out selection on unobserved heterogeneity in gains *within* group and choice set—what Heckman et al. (2006) call *essential heterogeneity*. Assumption 3.2 also requires that treatment effects are not correlated with lottery/cohort, however this can be relaxed. However, Assumption 3.2 is strictly weaker than assuming treatment effects are homogenous within each group. It still allows sorting on *levels*, that is that workers choosing $H_i = h$ have a different average value of $Y_i(h)$ than those who do not.

To see this, it is illustrative to write potential outcomes in the two-way-fixed-effects form:

$$Y_i(h) = \alpha_i + \beta_{G_i h} + u_{ih} \tag{3.1}$$

where $\alpha_i := Y_i(h_0)$ with h_0 a fixed reference employer, $\beta_{gh} := \mathbb{E}[Y_i(h) - Y_i(h_0)|G_i = g]$ and $u_{ih} := \{Y_i(h) - Y_i(h_0)\} - \mathbb{E}[Y_i(h) - Y_i(h_0)|G_i]$. Assumption 3.2 implies that idiosyncratic gains u_{ih} are (conditionally) mean independent of first-job choice: $\mathbb{E}[u_{ih}|H_i, G_i = g, L_i = \ell] = 0$, but not that H_i is in any way uncorrelated with the "worker-effects" α_i.[18]

To operationalize the use of choice sets of instruments, it will be convenient to represent a choice set as a vector of indicators Z_{hi} for the presence of each employer h in C_i, where $Z_{hi} = \mathbb{1}(h \in C_i)$. A realization of C_i is equivalent to a realization of the full vector $\mathbf{Z}_i := (Z_{1i}, Z_{2i} \dots Z_{Ji})'$, for some arbitrary ordering of the employers. Similarly, let $\mathbf{D}_i$ be a vector of $D_{hi} := \mathbb{1}(H_i = h)$ across all employers h. Again, the random vector $\mathbf{D}_i$ encodes exactly the same information as H_i. For any group g and lottery ℓ, let $\Sigma_{gl} = \mathbb{E}[\mathbf{Z}_i\mathbf{D}_i'|G_i = g, L_i = \ell]$.

Assumption 3.3 (relevance). *Σ_{gl} has full rank for each $g \in \mathcal{G}$ and $\ell \in \mathcal{L}$.*

Assumption 3.3 imposes the standard IV relevance condition that the J instruments have independent predictive power for the J treatments $h \in \mathcal{H}$, and that this holds within each (g, ℓ) cell.

For any group g, collect the μ_{gh} over all the employers into a vector $\boldsymbol{\mu}_g$. Proposition 3.2 shows that the assumptions given are sufficient to identify this full set of first job effects for each group:

Proposition 3.2 (identification of FJEs). *Make Assumptions 3.1, 3.2 and 3.3. Then for each $g \in \mathcal{G}$ and any ℓ:*

$$\boldsymbol{\mu}_g = \Sigma_{gl}^{-1}\mathbb{E}[\mathbf{Z}_iY_i|G_i = g, L_i = \ell]$$

[18] Note that $\mathbb{E}[u_{ih}|H_i, g, \ell] = 0$ is not sufficient for identification given observations of (Y, H) alone, on account of the unobserved α_i. Unlike "AKM" settings in which the α_i can be differenced out by workers moving between firms (cf. Abowd et al. 1999), a worker by definition has only one actual first-job H_i, yielding the single cross-section of observed earnings: $Y_i = \alpha_i + \beta_{G_iH_i} + u_i$ where $u_i := u_{iH_i}$.

Proof. See Appendix C.4. □

Note that based on Proposition 3.2, the vector $\boldsymbol{\mu}_g$ is in fact over-identified: in principle it can be estimated using the data from any single cohort ℓ. Indeed, the implication of Assumption 3.2 that treatment effects $Y_i(h) - Y_i(h_0)$ are mean independent of L_i could be relaxed to identify FJEs that vary by cohort. However, in practice, it is desirable to pool across lotteries given our limited sample size. The proof of Proposition 3.2 shows that $\boldsymbol{\mu}_g$ can be estimated from a sample that pools across L_i but conditions on $G_i = g$ by two stage least squares (2SLS) with the inclusion of cohort fixed effects, which pick up $\mathbb{E}[Y_i(h_0)|G_i = g, L_i = \ell]$ across the lotteries ℓ.

3.4.4 Partial identification with essential heterogeneity

While the results of the last section yield a straightforward route to identification of FJEs based on random choice set variation, the required assumption that workers do not sort into first jobs on the basis of their idiosyncratic FJE's is restrictive. For instance, when the outcome variable is earnings, it is incompatible with a Roy-type selection model in which there are worker $\times$ employer match effects and workers choose in part on the basis of earnings. Or, if the outcome of interest is mobility after residency, we must believe that workers are not more likely to move away from their residency locations if they ended up in a location that they preferred less during the lottery.

To accommodate violations of Assumption 3.2—*essential heterogeneity*—we develop a partial identification approach based upon the observation that the instruments provide a system of moment conditions that are linear in the FJEs μ_{gh}. This builds upon an existing literature that maps IV identification into a linear programming problem (Mogstad et al. 2018; Kamat 2020).

Let $D_{hi}(c)$ be an indicator for whether worker i would choose first-job employer h given choice set c, i.e. $D_{hi}(c) = \mathbb{1}(H_i(c) = h)$, and recall the notation of D_{hi} as an indicator for i actually choosing h, i.e. $D_{hi} = D_{hi}(C_i)$. Note that $D_{hi}(c)$ only depends on i through

i's preference relation $\succ_i$, and we may instead index the function D by $\succ$ rather than i. For ease of notation let $X_i = (G_i, L_i)$ be a vector composed of demographic group g and lottery ℓ. For any h and z, consider the observable quantity $\mathbb{E}[Y_i D_{hi}|C_i = c, X_i = x]$.

By the law of iterated expectations and Assumption 3.1:

$$\begin{aligned}
\mathbb{E}[Y_i D_{hi}|C_i = c, X_i = x] &= \sum_{\succ} P_{\succ|x} \cdot \mathbb{E}[Y_i D_{hi}| \succ_i = \succ, C_i = c, X_i = x] \\
&= \sum_{\succ} P_{\succ|x} \cdot \mathbb{E}[Y_i(h) D_{h\succ}(c)| \succ_i = \succ, C_i = c, X_i = x] \\
&= \sum_{\succ} D_{h\succ}(c) \cdot \{P_{\succ|x} \cdot \mu_{\succ h|x}\} \qquad (3.2)
\end{aligned}$$

where $P_{\succ|x} := P(\succ_i = \succ \,|X_i = x)$ and $\mu_{\succ h|x} := \mathbb{E}[Y_i(h)| \succ_i = \succ, X_i = x]$ is the average counterfactual outcome that would occur for a worker with preference relation $\succ$ if their first job were at employer h. The above expression reveals that the observable $\mathbb{E}[Y_i D_{hi}|C_i = c, X_i = x]$ identifies a linear combination of the $\mu_{\succ h|x}$ over all preferences $\succ$ under which h is the best choice for the fixed choice set c. Note that a linear combination of the $P_{\succ|x}$ alone with the same weights is also identified by removing Y_i from Eq. (3.2):

$$\mathbb{E}[D_{hi}|C_i = c, X_i = x] = \sum_{\succ} D_{h\succ}(c) \cdot P_{\succ|x} \qquad (3.3)$$

As an example, consider an instance of the lottery in which there three choice sets occur: $\{1,2,3\}$ $\{1,2\}$ and $\{1\}$, and we are interested in the FJE of $h = 2$. The coefficients in the system of linear equations (3.2) or (3.3) can be summarized by Table 3.1. In this

		$\succ_i$					
		$\mathbf{1 \succ 2 \succ 3}$	$\mathbf{1 \succ 3 \succ 2}$	$\mathbf{2 \succ 1 \succ 3}$	$\mathbf{2 \succ 3 \succ 1}$	$\mathbf{3 \succ 1 \succ 2}$	$\mathbf{3 \succ 2 \succ 1}$
	$\mathbf{\{1,2,3\}}$	0	0	1	1	0	0
C_i	$\mathbf{\{1,2\}}$	0	0	1	1	0	1
	$\mathbf{\{1\}}$	0	0	0	0	0	0
	$B^h_{\succ}$	$\{1\}$	$\{1,3\}$	$\emptyset$	$\emptyset$	$\{1,3\}$	$\{3\}$

Table 3.1: Example response matrix $D_{hi}(c)$ with three employers and one lottery.

example $J := |\mathcal{H}| = 3$, and the 6 columns of Table 3.1 represent the $J!$ distinct preference orderings over employers. The first three rows represent the three choice sets observed in the lottery, where first employer 3 becomes unavailable, then employer 2, and for workers with the lowest lottery numbers only employer 1 remains. The entries $\{0, 1\}$ indicate the value of $D_{h\succ}(c)$ for each row and column pair, forming what Heckman and Pinto (2018) call the *response matrix* of the model.

The last row of Table 3.1 summarizes the set of employers that are preferred to h according to the preference relation $\succ$ for that column, which we denote as $B^h_\succ := \{h' \in \mathcal{H} : h' \succ h\}$. Note that any two columns sharing a value of $B^h_\succ$ have an identical response $D_{h\succ}(c)$ for all choice sets c. This is because the event that a worker with preferences $\succ$ chooses h only depends on whether h is their most-preferred employer in c, and not what the relative ordering is between the other available employers c/h. In particular, i will choose h if and only if $h \in c$ and their "better-than-h" set $B^h_i := B^h_{\succ_i}$ does not intersect the choice set c, i.e. $D_{hi}(c) = \mathbb{1}(h \in c \text{ and } B^h_i \cap c = \emptyset)$.

We can thus coarsen the columns of Table 3.1 to combine all preferences $\succ$ that share a value of $B^h_\succ$, by rewriting Equation (3.2) as:

$$\mathbb{E}[Y_i D_{hi}|C_i = c, X_i = x] = \sum_{B \subseteq \mathcal{H}/h} \mathbb{1}(h \in c \text{ and } B \cap c = \emptyset) \cdot \{P_{Bh|x} \cdot \mu_{Bh|x}\} \tag{3.4}$$

where $P_{Bh|x} := P(B^h_i = B|X_i = x) = \sum_{\succ:B^h_\succ = B} P_{\succ|x}$ and $\mu_{Bh|x} = \sum_{\succ:B^h_\succ = B} P_{\succ|x} \cdot \mu_{\succ|x}$. Equation (3.3) can be similarly rewritten as a summation over the B^h rather than $\succ$.

		B^h_i			
		$\emptyset$	$\{1\}$	$\{3\}$	$\{1,3\}$
	$\{1,2,3\}$	1	0	0	0
C_i	$\{1,2\}$	1	0	1	0
	$\{1\}$	0	0	0	0
	$j^{h\ell}_i$	1	0	2	0

Table 3.2: Response matrix from Table 3.1 written in terms of better-than-h sets B^h_i; $h = 2$.

The value of moving from a summation over preferences to a summation over better-than-h sets B^h is that rather than a system of $J!$ unknowns $P_{\succ|x} \cdot \mu_{\succ h|x}$ for each x we now have a system in 2^{J-1} unknowns $P_{B|x} \cdot \mu_{Bh|x}$ for each x. In the example of Tables 3.1 and 3.2 this only reduces the number of columns from $3! = 6$ to $2^2 = 4$; however the gain quickly becomes dramatic for $J > 3$.

Note that there is still redundancy in the columns of Table 3.2: workers who prefer only employer 1 to employer 2 have the same response as workers who prefer both employers 1 and 3 to employer 2, for all choice sets $c \in supp\{C_i\}$. This is because the choice sets $supp\{C_i\}$ have a nesting property arising from the sequential nature of the RSD: once an employer's position has been filled, it never re-enters the choice sets of doctors choosing later in the lottery. This leads to a close connection with Heckman and Pinto (2018), who generalize the notion of "monotonicity" from Angrist and Imbens, 1994 to treatments that take on multiple unordered values.

To appreciate this connection, we introduce some further notation. For a given instance ℓ of the lottery, label the choice sets as $C_{1\ell} \supset C_{2\ell} \supset \cdots \supset C_{J_\ell,\ell}$, where J_ℓ is the number of employers in lottery ℓ. Let J_ℓ^h be the last set along this sequence that contains employer h. Selection behavior $D_{hi}(c)$ towards employer h within a single lottery ℓ can now be characterized by just $J_\ell^h + 1$ distinct groups. The reason is that if $D_{hi}(C_{j\ell}) = 1$ then it must be that $D_{hi}(C_{j'\ell}) = 1$ for all $j \leq j' \leq j_\ell^h$ (if h is chosen from a larger set, it must be chosen from any smaller subset that still contains h). Thus we can infer $D_{hi}(c)$ for all $c \in supp\{C_i | L_i = \ell\}$ from the lowest value of j such that $D_{hi}(C_{j\ell}) = 1$: call this $J_i^{h\ell}$. If alternatively $D_{hi}(C_{J_\ell^h}, \ell) = 0$, i.e. i does not choose h even in the smallest choice set in which it appears, then we can call i an "h-never-taker" in lottery ℓ, and write $j_i^{h\ell} = 0$. The final row in Table 3.2 lists the value of $j_i^{h\ell} = 0$ corresponding to each B_i^h. Note that while there are $2^2 = 4$ better-than-h sets in this example, there are just $J_\ell^h + 1 = 3$ distinct values of $j_i^{h\ell}$.

The analysis from the preceding paragraph reveals that within each lottery ℓ, selection behavior satisfies what Heckman and Pinto, 2018 call *unordered monotonicity*: for each $h \in \mathcal{H}$, there exists an ordering on the points in $c \in supp\{C_i|L_i = \ell\}$ such that $D_{hi}(c)$ is weakly increasing along that order.[19] Heckman and Pinto, 2018 provide point identification results under unordered monotonicity with discrete instruments: in particular their results imply that

$$\mathbb{E}[Y_i(h)|i \text{ is not an h-never-taker}, X_i = x] \tag{3.5}$$

is point identified.[20] In particular, it is equal to $\frac{E[Y_i D_{hi}|C_i = C_{j_\ell^h,\ell}, X_i = x] - E[Y_i D_{hi}|C_i = C_{1,\ell}, X_i = x]}{E[D_{hi}|C_i = C_{j_\ell^h,\ell}, X_i = x] - E[D_{hi}|C_i = C_{1,\ell}, X_i = x]}$, where ℓ is the lottery indicator appearing in $X_i = x$. Indeed, it is also easily shown that the mean of $Y_i(h)$ can be further disaggregated within each of the j_ℓ^h individual complier groups (that occur with positive probability), by considering adjacent values of j in the above ratio.

However parameter (3.5) is by itself not sufficient to identify μ_{gh}, as it does not capture $Y_i(h)$ among workers who would never choose h from any choice set present in their lottery. The proportion of such h-never-takers within a given (ℓ, h) pair can be quite large, leading to wide bounds even if we assume $\mathbb{E}[Y_i(h)|i \text{ is an h-never-taker}, X_i = x]$ belongs to some bounded set $[Y^L, Y^U]$.

By contrast, the approach of Table 3.2 based on better-than-h sets keeps the $j_i^{th} = 0$ group disaggregated into those workers with different better-than-h sets—$\{1\}$ and $\{1, 3\}$ in the example—in a way that is consistent across different lotteries ℓ (unlike the unordered monotonicity approach, in which the meaning of the j_i^{th} groups depend on lottery). This allows us to combine the identifying information of Equation (3.4) across multiple lotteries, that have different sequences of choice sets. In the following example, for

[19]When viewed across all possible choice-sets $c \subseteq \mathcal{H}$, $D_{hi}(c)$ is increasing according to a *partial* order on the c, which depends on h. In particular $D_{hi}(c) \geq D_{hi}(c')$ whenever $c/h \subseteq c'/h$ and $h \in c$ if $h \in c'$. This generalizes the notions of "partial" and "vector" monotonicity analyzed by (Goff, 2020; Mogstad et al., 2020a) to the unordered-treatment case.

[20]Lee and Salanié (2020) also consider identification under unordered monotonicity with discrete instruments. They introduce the concept of particular instrument values *targeting* particular treatments h in such a setting. In their language, we can say that our choice sets $C_{1\ell}$ to $C_{j_\ell^h,\ell}$ *strictly target* employer h (interpreting the event $h \notin C_i$ as endowing h with so low a utility to worker i that they would never choose it).

instance, employer 4 offers jobs only in lottery $\ell = 2$, and the order in which vacancies are filled differs in the two years:

	C_i	$\boldsymbol{B}_i^h$							
		$\emptyset$	$\{1\}$	$\{3\}$	$\{4\}$	$\{1,3\}$	$\{1,4\}$	$\{3,4\}$	$\{1,3,4\}$
	$\mathbf{\{1,2,3\}}$	1	0	0	1	0	0	0	0
$\ell = 1$	$\mathbf{\{1,2\}}$	1	0	1	1	0	0	1	0
	$\mathbf{\{1\}}$	0	0	0	0	0	0	0	0
	$j_i^{h,1}$	1	0	2	1	0	2	1	0
	$\mathbf{\{1,2,3,4\}}$	1	0	0	0	0	0	0	0
	$\mathbf{\{2,3,4\}}$	1	1	0	0	0	0	0	0
$\ell = 2$	$\mathbf{\{2,3\}}$	1	1	0	0	0	1	0	0
	$\mathbf{\{2\}}$	1	1	1	0	1	1	1	1
	$j_i^{h,2}$	1	2	4	0	4	3	4	4

Table 3.3: Response matrix across two lotteries, with $\mathcal{H} = \{1,2,3,4\}$ and $h = 2$.

To make efficient use of our data spanning multiple lotteries, we assume that both first job effects and group sizes are stable across lotteries:

Assumption 3.4 (stability). *Neither $\mu_{Bh|x}$ nor $P_{Bh|x}$ depends on the ℓ component of x (for all $h \in \mathcal{H}$, $B \subseteq \mathcal{H}$ and $g \in \mathcal{G}$)*

Note that Assumption 3.4 does not nest Assumption 3.2, which only requires FJE *differences* to be stable over lotteries.[21] Assumption 3.4 implies that we can write FJEs and group sizes for a given better-than-h set introduced earlier ($\mu_{gh|x}$ and $P_{Bh|x}$) simply as $\mu_{B|g}$ and $P_{B|g}$, depending only on demographic group g and not on lottery ℓ.

With Assumption 3.4 in mind, we now turn to the formal identification analysis. We also make explicit the assumption that workers' choices are made rationally:

[21]Recall that we include lottery fixed effects in the strategy from Section 3.4.3 to absorb the dependence of $\mathbb{E}[Y_i(h_0)|G_i = g, L_i = \ell]$ on ℓ). On the other hand, Assumption 3.4 allows selection on gains, so neither Assumption 3.4 nor Assumption 3.2 nest each other, but emphasize different relaxations of identifying assumptions.

Assumption 3.5 (rationality). *Each worker i has a preference relation $\succ_i$ over $\mathcal{H}$ without indifferences, such that $H_i(c) = \{h \in \mathcal{H} : h \succ_i h', \forall h' \in \mathcal{H}, h' \neq h\}$*

Note that Assumption 3.5 ensures that the better-than-h sets B_i^h are always well-defined. We return to a discussion of Assumption 3.5 below.

Recall that our parameters of interest are the $\mu_{gh} = \mathbb{E}[Y_i(h)|G_i = g]$ for each h and g, and note that under Assumption 3.4 this can be written as

$$\mu_{gh} = \sum_{B \subseteq \mathcal{H}/h} Q_{Bh|g}$$

where we define $Q_{Bh|g} := P_{Bh|g} \cdot \mu_{Bh|g}$. Note that $Q_{Bh|g}$ is precisely the quantity appearing in the summand of each of the linear restrictions (3.4), with the known coefficients $\mathbb{1}(h \in c \text{ and } B \cap c = \emptyset)$. Thus our identification problem involves a linear optimization problem over the $Q_{Bh|g}$, involving a set of linear constraints on them. Proposition 3.3 below makes this precise.

Before stating the result, we introduce one final assumption: that the $\mu_{Bh|g}$ are uniformly bounded by known constants Y^L and Y^U:

Assumption 3.6 (boundedness). $Y^L \leq \mu_{Bh|g} \leq Y^H$ *(for all $h \in \mathcal{H}$, $g \in \mathcal{G}$, $B \subseteq \mathcal{H}/h$)*

Assumption 3.6 is useful because the data will give no direct information about $\mu_{Bh|g}$ for a given B if $P(H_i = h|B_i^h = B, G_i = g) = 0$.[22] Thus, depending on the support of the choice sets C_i, even partial identification of μ_{gh} may require such a auxiliary assumptions, with boundedness being the simplest example. Note that a simple sufficient condition for Assumption 3.6 is that $Y^L \leq Y_i(h) \leq Y^H$ for all doctors i and employers h, that is the bounds hold individually rather than on average.

[22] In particular, this is most likely to happen for $B = \mathcal{H}/h$, for which a doctor will only choose h if their choice set consists only of h. $C_i = \{h\}$ does in fact occur in the example of Table 3.3, and thus every column of the table has at least one entry of 1.

Since $Q_{Bh|g} = P_{Bh|g} \cdot \mu_{Bh|g}$, Assumption 3.6 implies that:

$$Y^L P_{Bh|g} - Q_{Bh|g} \leq 0 \text{ and } Y^U P_{Bh|g} - Q_{Bh|g} \geq 0 \text{ for each } B \subseteq \mathcal{H}/h \tag{3.6}$$

Fix a g and h. Letting $\mathbf{Q}$ be a vector of the $Q_{Bh|g}$ across all $B \subseteq \mathcal{H}/h$, and similarly $\mathbf{P}$ for the $P_{Bh|g}$, we denote by $\mathcal{M}$ the set of all $(\mathbf{Q}, \mathbf{P})$ pairs such that (3.6) holds and that:

$$\sum_{B \subseteq \mathcal{H}/h} P_{Bh|g} = 1 \text{ and } P_{Bh|g} \geq 0 \text{ for each } B \subseteq \mathcal{H}/h \tag{3.7}$$

We may now characterize the identified set of μ_{gh} as an optimization problem over $\mathcal{M}$:

Proposition 3.3. *Under Assumptions 3.1,3.4, 3.5 and 3.6, $\mu_{gh} \in [\theta^L_{gh}, \theta^U_{gh}]$, where*

$$\theta^L_{gh} := \min_{(\mathbf{Q},\mathbf{P}) \in \mathcal{M}} \sum_{B \subseteq \mathcal{H}/h} Q_{Bh|g} \qquad \textit{and} \qquad \theta^U_{gh} := \max_{(\mathbf{Q},\mathbf{P}) \in \mathcal{M}} \sum_{B \subseteq \mathcal{H}/h} Q_{Bh|g}$$

subject to the following restrictions:

$$\sum_{B \subseteq \mathcal{H}/h} \mathbb{1}(h \in c \textit{ and } B \cap c = \emptyset) \cdot Q_{Bh|g} = \mathbb{E}[Y_i D_{hi} | C_i = c, G_i = g, L_i = \ell] \tag{3.8}$$

$$\sum_{B \subseteq \mathcal{H}/h} \mathbb{1}(h \in c \textit{ and } B \cap c = \emptyset) \cdot P_{Bh|g} = \mathbb{E}[D_{hi} | C_i = c, G_i = g, L_i = \ell] \tag{3.9}$$

for each $\ell \in \mathcal{L}$ and $c \in supp\{C_i | L_i = \ell\}$.

Comparison with Mogstad et al. (2018) and Kamat (2020):

Proposition 3.3 is closely related to recent results in Mogstad et al. (2018) and Kamat (2020), who also express IV estimands of solutions to a linear programming problem. Mogstad et al. (2018) develops an approach to identi fying treatment effect parameters that depend on marginal counterfactual means of the form $\mathbb{E}[Y_i(d)|U_i = u]$, in the classic LATE setting in which a binary treatment is driven by a separable threshold crossing model: $D_i(z) = \mathbb{1}(p(z) \geq U_i)$. In their case, the latent groups correspond to the values of U_i, which is uniformly distributed on the unit interval. This avoids the need for a distinction between the vectors $\mathbf{Q}$ and $\mathbf{P}$, which must be both optimized over in our setting, since the analog of $P_{Bh|g}$ for the is simply a uniform measure on $[0, 1]$.

Similar to us, Kamat (2020) also optimizes over latent group probabilities jointly with outcomes. However, they make use of a discrete outcome variable to optimize directly over the full joint distribution of potential outcomes and potential treatments. This keeps the constraints and objective functions linear in parameters without a need to introducethe final set of inequality constraints (3.6). They also consider a richer class of parameters of interest, taking the form $\sum_{B\subseteq\mathcal{H}/h} w_B \cdot Q_{Bh|g} / \sum_{B\subseteq\mathcal{H}/h} w_B \cdot P_{Bh|g}$. This introduces a non-linear objective function, which they handle by introducing an additional variable and re-paramaterizing the problem.

Remark: Note that point identification obtains in Proposition 3.3 if $\sum_{\mathcal{H}/h} Q_{Bh|g} = (1, 1, \ldots, 1)'\mathbf{Q}$ can take on just a single value subject to restrictions (3.8), (3.9) and $(\mathbf{Q}, \mathbf{P}) \in \mathcal{M}$. This holds for example whenever the vector $(1, 1, \ldots, 1)'$ lies in the column space of the response matrix depicted in Tables 3.2 and 3.3, describing the coefficients appearing in expansion (3.8). This yields an alternative way to understand the point identification result of Proposition 3.1; whenever $P(C_i = \{h\}|G_i = g) > 0$, the rows corresponding to this choice set in the response matrix are composed of all ones.

Estimation and Inference:

Given our finite sample of data $\{(Y_i, H_i, C_i, G_i, L_i)\}_{i=1\ldots n}$, the endpoints of the identifed set $\Theta_{gh} = [\theta^L_{gh}, \theta^U_{gh}]$ could be consistently estimated by solving the linear program of Proposition 3.3 upon replacing the expectations appearing in (3.8) and (3.9) with their finite sample analogs $\mathbb{E}_n$, e.g. $\mathbb{E}_n[Y_i] = \frac{1}{n}\sum_{i=1}^n Y_i$. Note however that the resulting interval estimate $\hat{\Theta}_{gh} = [\hat{\theta}^L_{gh}, \hat{\theta}^U_{gh}]$ may be empty even if the the model is correctly specified. For example, in the data we sometimes observe cases where $\mathbb{E}_n[D_{hi}|C_i = c, G_i = g, L_i = \ell] >$

$\mathbb{E}_n[D_{hi}|C_i = c', G_i = g, L_i = \ell]$ where c' is a strict subset of c (and both of which contain h). On its face, this appears to be evidence against the joint hypothesis of choice-set independence (Assumption 3.1) and utility maximization on the part of workers. However, it is also not at all unlikely when the events $(C_i = c \cap L_i = \ell)$ and $(C_i = c' \cap L_i = \ell)$ have only a few observations each, as is often the case in our data.

For the above reason, we solve a relaxation of the linear programs in Proposition 3.3 to obtain point estimates for θ^L_{gh} and θ^U_{gh}, which also forms the basis for our approach to constructing confidence intervals for the underlying parameter μ_{gh}. Firstly, we pool data across lotteries ℓ, replacing moments like $\mathbb{E}_n[D_{hi}|C_i = c, G_i = g, L_i = \ell]$ by $\mathbb{E}_n[D_{hi}|C_i = c, G_i = g]$, and similarly for the moments of $Y_i D_{hi}$. This is justified under Assumption 3.4, which implies by (3.4) and the law of iterated expectations that:[23]

$$\mathbb{E}[Y_i D_{hi}|C_i = c, G_i = g] = \sum_{B \subseteq \mathcal{H}/h} \mathbb{1}(h \in c \text{ and } B \cap c = \emptyset) \cdot \{P_{Bh|g} \cdot \mu_{Bh|g}\} \tag{3.10}$$

An analogous expression holds for Eq. (3.3) with the lottery conditioning removed.

Secondly, we do not require the moment conditions to be satisfied exactly in sample. Define the quantities:

$$s^Y_{hcg} = \left(\sum_{B \subseteq \mathcal{H}/h} \mathbb{1}(h \in c \text{ and } B \cap c = \emptyset) \cdot Q_{Bh|g} - \mathbb{E}_n[Y_i D_{hi}|C_i = c, G_i = g] \right) \tag{3.11}$$

$$s^D_{hcg} = \left(\sum_{B \subseteq \mathcal{H}/h} \mathbb{1}(h \in c \text{ and } B \cap c = \emptyset) \cdot P_{Bh|g} - \mathbb{E}_n[D_{hi}|C_i = c, G_i = g] \right) \tag{3.12}$$

which measure the deviation of a given $(\mathbf{Q}, \mathbf{P})$ pair from the identifying restrictions (3.8) and (3.9). Let $\mathbf{s}$ be a vector of all of the s^V_{hcg} over $V \in \{Y, D\}, c$ and g. Similarly to Mogstad et al. (2018), we consider the smallest deviation $\mathbf{s}$ attainable, in an L^1 norm-sense. In

[23] Note that (3.10) is the same expression we would arrive at after assuming choice-sets exogeneity without conditioning on lottery. However Assumption 3.4 coupled with Assumption 3.1 as stated is still slightly weaker, as it e.g. only requires that the response-group specific conditional means of $Y_i(h)$–rather than their full distributions–do not depend on L_i.

particular, let

$$T_n := \min_{\substack{(\mathbf{Q},\mathbf{P})\in\mathcal{M}\\ \mathbf{s}}} \left(\sum_{\ell\in\mathcal{L}} \quad \sum_{B\subseteq\mathcal{H}/h} \quad \sum_{V\in\{Y,D\}} a^V_{hcg} \cdot \left| s^V_{hcg} \right| \right)$$

subject to (3.11) and (3.12) for each $c \in supp\{C_i|L_i = \ell\}$ and $\ell \in \mathcal{L}$, and where we introduce a set of positive scaling coefficients a. We set the scaling coefficients as $a^D_{hcg} = \sqrt{n_{cg}/Var_n(D_{hi}|G_i = g)}$ and $a^Y_{hcg} = \sqrt{n_{cg}/Var_n(Y_i D_{hi}|G_i = g)}$, where n_{cg} is the number of observations for which $C_i = c$ and $G_i = g$.[24]

Given T_n, now estimate θ^L_{gh} as

$$\hat{\theta}^L_{gh} := \min_{\substack{(\mathbf{Q},\mathbf{P})\in\mathcal{M}\\ \mathbf{s}}} \sum_{B\subseteq\mathcal{H}/h} Q_{Bh|g} \quad \text{s.t.} \quad \left(\sum_{\ell\in\mathcal{L}} \quad \sum_{B\subseteq\mathcal{H}/h} \quad \sum_{V\in\{Y,D\}} a^V_{hcg} \cdot \left| s^V_{hcg} \right| \right) \leq T_n + \kappa_n \tag{3.13}$$

Equation (3.13) finds the smallest value of μ_{gh} among the $(\mathbf{Q}, \mathbf{P})$ that are "closest" to satisfying the identifying restrictions (3.8) and (3.9) in finite sample. The tuning parameter κ_n broadens the notion of "closest" such $(\mathbf{Q}, \mathbf{P})$, and must converge to zero with n for consistency. We report estimates with $\kappa_n = 0$ and $\kappa_n = T_n/10$. The estimate $\hat{\theta}^U_{gh}$ is defined analogously to Eq. (3.13) but with the min operator replaced by a max.

Although the absolute value function is not linear, the objective function defining T_n can be reformulated by adding an additional variable for each component of **s** and reparameterizing the problem slightly. In particular, one can replace each instance of s by $p - n$, and add constraints $a \geq 0, b \geq 0$ to the problem. The simplex algorithm (standard for solving linear programs) will then ensure that these correspond to positive and negative parts of s: $p = \max(s, 0)$ and $n = \max(-s, 0)$, with respect to which the absolute value of s is the linear function $|s| = p + n$. We use this strategy to compute $\hat{\theta}^L_{gh}$ using the mixed integer linear programming package `lpSolveAPI` in R, for each g, h first computing T_n and then evaluating Eq. (3.13).

[24]The goal of this choice is to normalize the sampling variance of each s^V_{hcg} to unity. However, we cannot divide by the conditional sample variance specific to each choice set, because D_{hi} has no variation within some choice sets in which h appears (since no employees in fact choose h). Thus, $Var_n(D_{hi}|C_i = c, G_i = g)$ and $Var_n(Y_i D_{hi}|C_i = c, G_i = g)$ would be zero.

Let us now turn to building confidence intervals for the parameter μ_{gh}. To test the null hypothesis that $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$ for a generic value $\theta_0 \in \mathbb{R}$ we use a test statistic that augments T_n to enforce the null hypothesis, i.e.

$$T_n(\theta_0) := \min_{\substack{(\mathbf{Q},\mathbf{P})\in\mathcal{M} \\ \mathbf{s}}} \left(\sum_{\ell\in\mathcal{L}} \sum_{B\subseteq\mathcal{H}/h} \sum_{V\in\{Y,D\}} a^V_{hcg} \cdot \left| s^V_{hcg} \right| \right) \qquad s.t. \qquad \sum_{B\subseteq\mathcal{H}/h} Q_{Bh|g} = \theta_0 \tag{3.14}$$

and restrictions (3.11) and (3.12) for each $c \in supp\{C_i|L_i = \ell\}$ and $\ell \in \mathcal{L}$. The statistic $T_n(\theta_0)$ can be interpreted as measuring the minimum weighted deviation from constraints (3.8) and (3.9) that is necessary for a $(\mathbf{Q}, \mathbf{P})$ pair to deliver that value of θ_0.

We construct confidence intervals by comparing $T_n(\theta_0)$ to a critical value and collecting those values for which we fail to reject, i.e. $C_n = \{\theta \in \mathbb{R} : T_n(\theta_0\} \leq \hat{c}\}$, where $\hat{c}$ is a critical value estimated from the data. In particular, construct a collection of $\{T^*_{bn}(\theta_0)\}_{b=1\dots B}$ by non-parametric bootstrap, and compute $\hat{c}$ as the $1-\alpha$ quantile of the $T^*_{bn}(\theta_0)$ for $\alpha = 0.05$. Each $T^*_{bn}(\theta_0)$ replaces the moments in (3.11) and (3.12) with bootstrap analogues $\mathbb{E}^*_b$ and re-centers with respect to the "full-sample" estimates, for example:

$$\begin{aligned} s^{*D}_{bhcg\ell} = \sum_{B\subseteq\mathcal{H}/h} & \mathbb{1}(h \in c \text{ and } B \cap c = \emptyset) \cdot (P_{Bh|g} - P^0_{BH|g}) \\ & - (\mathbb{E}^*_l[D_{hi}|C_i = c, G_i = g, L_i = \ell] - \mathbb{E}_n[D_{hi}|C_i = c, G_i = g, L_i = \ell]) \end{aligned}$$

where $P^0_{BH|g}$ denotes the optimizer from (3.14), and $s^{*Y}_{bhcg\ell}$ is defined analogously. We repeat this entire exercise over a grid of θ_0 between the values Y^L and Y^U from Assumption 3.6, and report the maximum and mininum value along that grid for which we fail to reject.

This approach to inference is close to that of Kamat, 2020, who uses a quadratic objective function of the form $\mathbf{s}'\Omega^{-1}\mathbf{s} = ||\Omega^{-1/2}\mathbf{s}||_2$ rather than $||\Omega^{-1/2}\mathbf{s}||_1$, where in our case $\Omega^{-1/2}$ is a diagonal matrix defined by the a coefficients. We use the L^1 norm in order to keep all quantities computable by linear-programming algorithms, which are faster to solve than quadratic programs (computational limitations loom large for us, as we discuss in the next section). Kamat, 2020 also uses subsampling rather than bootstrap to compute the critical values $\hat{c}$, on the basis of results from Kalouptsidi et al., 2020. We choose bootstrap to avoid the need to choose the subset size, which represents an additional tuning parameter. Our setting is also related to methods that treat inference under partial identification in moment equality/inequality models (Chernozhukov et al., 2007; Andrews and Soares, 2010; Chernozhukov et al., 2013), as well as specification tests for random utility models (Kitamura and Stoye, 2018; Smeulders et al., 2021).

Implementation:

In implementing the above methods in our data, we face very real constraints on computational tractability (as well as statistical power). Workers choose among 55 employers in our final sample, with a typical lottery including most of these employers. With $|\mathcal{H}| = 55$, the vectors **Q** and **P** would each contain about 2×10^{-16} entries, which is clearly infeasible from a computational standpoint.[25]

For this reason, we group the employers (hospitals) into a manageable *categories*, and ignore the distinction between hospitals within a category. We define the categories on the basis of employers' overall desirability, as evidenced by the average lottery number R_i among workers who choose it in the RSD, across the study period (recall that R_i is normalized to the unit interval within each lottery). We find that $|\mathcal{H}| = 10$ is about the largest linear program the software will support, yielding roughly 8000 parameters to be optimized over. However for ease of interpretation, we for now use just four categories. Categories 1-3 are defined by terciles of the distribution of $\bar{r}_h$ across hospitals, where $\bar{r}_h := \mathbb{E}_n[R_i|H_i = h]$. To have a well-defined reference group, we for Category 4 use the hospitals in the remote counties of Finnmark and Sogn og Fjordane, the most

[25] Assuming one byte for each entry, simply storing the constraint matrix for the linear program requires about 2 gigabytes for $|\mathcal{H}| = 20$, 2 terabytes for $|\mathcal{H}| = 30$, 2 petabyes for $|\mathcal{H}| = 40$, and so on.

remote regions of northern and western Norway.[26] Table 3.4 reports some observable characteristics of the categories: for example the lower category numbers tend to be larger and more urban.

Grouping the employers together in this way embodies a substantive assumption: to make use of the methods of this section we must be willing to assume that workers are indifferent between hospitals within a single category.[27] A failure of this assumption could explain deviations from the model of the type we previously attributed to sampling variation: e.g. workers being more likely to choose a Category 2 hospital when $C_i = \{1, 2, 3\}$ than when $C_i = \{2, 3\}$.[28] One could explore alternative data-driven ways to define the categories: for example to minimize a statistic like T_n over such choices. An alternative approach may be retain the full set of $|\mathcal{H}| = 55$ employers but find some efficient way select among the 2^{54} columns of the response matrix depicted in Table 3.2.[29] Indeed, only $1,802$ distinct choice sets C_i ever occur across the study period, a tiny fraction of those that are conceptually possible. As a result, the column rank of the response matrix can at most be $1,802$, which is within our maximum manageable number of columns from a computational standpoint.

3.5 Results

This section presents estimates of first job effects obtained by the methods presented in Section 3.4. Throughout, we let the "employers" h correspond to the four Categories

[26]First jobs at employers in these regions are very unappealing to most graduates. Because of this, the government in some years introduced special incentives for residents in Finnmark and Sogn og Fjordane. Despite this, the large majority of candidates that chose employers in these regions drew very high (poor) lottery numbers and therefore had unappealing choice sets. Hospitals in these regions are excluded from the definition of Categories 1-3.

[27]This distinguishes our categories from the related notion of a *filtered* treatment introduced by Lee and Salanié (2020). In the latter case, agents select according to their preferences over a fine set of treatment states, and the researcher observes only coarsened categories of that choice.

[28]Since the order at which hospitals are removed from the available set via the RSD may change year to year, this could happen if doctors tend to prefer the best Category 3 hospitals to the worst Category 2 hospitals, and the best Category 2 hospitals tend to be gone from the RSD before the last hospitals in Category 1 are.

[29]A related idea is pursued in Smeulders et al., 2021.

of hospitals defined above. The first four columns of Table 3.4 report some characteristics of these groups. For simplicity, we also focus throughout this section on just two demographic groups g of workers: male and female.

3.5.1 Results of the point-identification approach with limited selection on gains

Recall that Proposition 3.2 shows that FJEs are point identified under an assumption of no selection on unobserved gains. The final column of Table 3.4 presents estimates based on this result, where the outcome Y_i is chosen to represent earned income four years after a doctor's residency (five years after graduation). The gap of four years chosen to maximize the number of workers that can be included, but later drafts will consider FJEs across various time horizons. To increase power, we first report FJEs that furthermore do not condition on gender, i.e. unconditional counterfactual means $\mathbb{E}[Y_i(h)]$ rather than μ_{gh}.

Differences in $\mathbb{E}[Y_i(h)]$ across residency hospitals h reveal that the location of one's residency affects their earnings *after* they've moved on to their post-residency position. We estimate that, relative to Category 4 employers, a first job in the most-in-demand employer category raises annual earnings five years post-graduation by about $28,000, in 2020. The corresponding estimates for categories 3 and 2 are $38,000 and (an insignificant) $16,000.

Category	Avg. Draw	# Hospitals	Avg. Emp.	Proportion Urban	Earnings FJE diff.
1	0.17	17	1634	0.82	28.21*
					(14.64)
2	0.44	17	1457	0.65	16.03
					(13.00)
3	0.73	16	453	0.50	37.70**
					(18.99)
4	0.89	5	502	0.20	-

Table 3.4: FJEs measure the impact of a first job in each category on earnings 5 years post-graduation, in thousands of 2020 USD. First stage F-statistics for Category 1, 2, and 3 are 322.65, 327.44, and 102.40. $N = 9,049$. Robust standard errors in parentheses. *** p<0.01 ** p<0.05 * p<0.10.

However, statistical significance of the differences $\mu_{gh} - \mu_{g4}$ disappears when we estimate FJEs $\mu_{gh} = \mathbb{E}[Y_i(h)|G_i = g]$ separately by gender. However the point estimates reveal an annual earnings gap five years out of at least $\$20,000$ between men and women—about 13%—across first employers. Appendix Figure C.9 plots the μ_{gh} against average realized earnings $\mathbb{E}[Y_i|H_i = h, G_i = g]$, as a rough indication of the extent of the endogeneity that the IV approach is correcting for. See also Table 3.8 of Section 3.6 for the point estimates.

3.5.2 Results of the general partial identification approach

We now turn to the partial identification approach from Section 3.4.4 that relaxes the assumption of no selection on gains within gender g. Continuing with the earnings outcome variable, Table 3.5 reports estimates of the identified set $[\theta^L_{gh}, \theta^U_{gh}]$ and confidence intervals for μ_{gh}. We set Y^L at 10 thousand dollars in 2020 USD, and Y^U at $\$300,000$ (about 2% of the sample is outside of this range in either direction). For all of our outcome variables, the 95% confidence interval $\mathcal{C}_n$ takes a grid of 20 values of θ_0 across the range $[Y^L, Y^U]$, and uses 200 bootstrap replications.

g	h (category)	$[\hat{\theta}^L_{gh}, \hat{\theta}^U_{gh}]$ $\kappa_n = 0$		$\kappa_n = 10\%$		$\mathcal{C}_n$ (95% CI)	
Women	1	152.49	152.49	129.71	157.91	61.58	190.53
	2	90.93	90.93	89.06	96.31	61.58	138.95
	3	133.31	133.31	130.71	134.12	35.79	164.74
	4	146.47	146.47	133.35	158.06	-	-
Men	1	187.64	187.64	151.65	192.78	61.58	216.32
	2	201.10	201.10	85.10	216.76	-	-
	3	148.61	148.61	132.98	151.38	61.58	216.32
	4	116.71	116.71	110.79	119.26	35.79	190.53

Table 3.5: First job effects $\mu_{gh} = \mathbb{E}[Y_i(h)|G_i = g]$ for earned income four years post residency, in thousands of 2020 USD. Average earnings across the sample are about $\$150,000$. Table reports estimates of the identified set $[\theta^L_{gh}, \theta^U_{gh}]$ and 95% confidence intervals for μ_{gh}.("-" indicates that $C_n = \emptyset$).

As Table 3.5 shows, the $\kappa_n = 0$ point estimates of θ^L_{gh} and θ^U_{gh} in all cases suggest point identification, i.e. $\hat{\theta}^L_{gh} = \hat{\theta}^L_{gh}$. However, this point identification is "spurious", as there can be a unique $(\mathbf{Q}, \mathbf{P})$ that minimizes the sample statistic T_n even when there is no unique $(\mathbf{Q}, \mathbf{P})$ setting it to zero in the population. We thus treat the slightly "nudged" estimates with $\kappa_n = T_n/10$ as preferred. Recall that these seek the largest and smallest values of μ_{gh} compatible with a T_n within 10% of of its minimum value. Overall, the results suggest that Category 1 hospitals cause doctors' earnings to be highest, especially for women. Note that confidence intervals are missing for women in Category 4 and for men in Category 2. In these cases, the null hypothesis not accepted for any θ_0 between Y^L and Y_U. This suggests that the model is rejected in these cases, and warrants further investigation. In the other cases, the confidence intervals are also quite wide, so we focus subsequent attention on the point estimates with $\kappa_n = T_n/10$ in this section.

In Table 3.6, the outcome variable Y_i is whether a doctor ever specializes during their career (Appendix Table C.3 reports results for the *number* of specializations). This outcome variable is bounded by definition, where $Y^L = 0$ and $Y^U = 1$. The results suggest that working at a Category 4 hospital causes the greatest rates of specialization.

g	h	$[\hat{\theta}^L_{gh}, \hat{\theta}^U_{gh}]$				$\mathcal{C}_n$	
	(category)	$\kappa_n = 0$		$\kappa_n = 10\%$		(95% CI)	
Women	1	0.50	0.50	0.49	0.64	0.21	1.00
	2	0.50	0.50	0.50	0.56	0.26	1.00
	3	0.63	0.63	0.62	0.64	0.37	1.00
	4	0.71	0.73	0.71	0.86	0.11	1.00
Men	1	0.58	0.58	0.50	0.68	0.21	1.00
	2	0.50	0.50	0.49	0.66	0.21	1.00
	3	0.83	0.83	0.83	0.87	0.26	1.00
	4	1.00	1.00	1.00	1.00	0.16	1.00

Table 3.6: First job effects $\mu_{gh} = \mathbb{E}[Y_i(h)|G_i = g]$ for whether doctor ever specializes during their career. Overall, about 55% of doctors specialize. Table reports estimates of the identified set $[\theta^L_{gh}, \theta^U_{gh}]$ and 95% confidence intervals for μ_{gh}.

Table 3.7 takes the outcome of interest to be whether the doctor ever moves municipalities, starting with their residency year. Looking at the preferred estimates ($\kappa_n = T_n/10$ column), μ_{gh} is increasing in Category number for both men and women. This is as expected, since lower Category numbers correspond to hospitals that tend to be more sought-after, and thus doctors are more likely to want to stay there longer term.

g	h (category)	$[\hat{\theta}^L_{gh}, \hat{\theta}^U_{gh}]$ $\kappa_n = 0$		$\kappa_n = 10\%$		$\mathcal{C}_n$ (95% CI)	
Women	1	0.07	0.07	0.07	0.13	0.05	0.74
	2	0.17	0.17	0.17	0.20	0.05	1.00
	3	0.14	0.14	0.14	0.16	0.05	1.00
	4	0.16	0.16	0.16	0.25	0.05	1.00
Men	1	0.33	0.33	0.33	0.38	0.05	1.00
	2	0.33	0.33	0.33	0.44	0.05	1.00
	3	0.46	0.51	0.36	0.58	0.05	1.00
	4	0.50	0.50	0.50	0.55	0.05	1.00

Table 3.7: First job effects $\mu_{gh} = \mathbb{E}[Y_i(h)|G_i = g]$ for whether doctor ever changes municipalities after residency. About 13% of doctors in fact move, across the sample. Table reports estimates of the identified set $[\theta^L_{gh}, \theta^U_{gh}]$ and 95% confidence intervals for μ_{gh}.

3.6 Using the FJEs to assess the consequences of decentralization

The last section has presented estimates of first job effects on earnings—among other outcomes variables–based on data from the era in which residencies were allocated by the random serial dictatorship mechanism. We now combine these estimates with data from after the replacement of the RSD with a decentralized labor market in 2013, to understand this reform from the perspective of workers.

Recall that in the RSD era, choice sets were allocated to workers independently of their preferences and potential outcomes. In the market era by contrast, the opportunities available to a worker are likely to be highly correlated with her unobserved ability and preferences. Thus, the reform may have affected average outcomes within each demographic group, by changing the distribution of choice sets the workers in that group face and hence their actual employer matches. Given that we observe the distribution of demographic group $\times$employer matches in both periods, we can calculate the implied welfare changes across groups given a suitable measure of welfare.

We do this by assuming a particular aggregate relationship in the RSD period between a worker's indirect utility at her chosen employer and her lottery number. Given our FJE estimates from the last section, we further decompose this utility into an earnings component and an "amenity" component, allowing us to track changes in both across the reform. To keep the analysis simple, we use FJE estimates based on the point-identification approach from Section 3.4.3, again focusing on the four employer categories described in Section 3.4.4.

3.6.1 Estimating first-job amenity values

The first step of our approach is to define an average "amenity" value for each employer category h. To this end, we take preferences of workers defined over the employer categories h to have the form:

$$U_i(h) = \mu_{G_i h} + A_{G_i h} + \eta_{hi} \tag{3.15}$$

where μ_{gh} is the first job effect of category h for group g, A_{gh} captures the average "amenity" value of employer category h, and $\mathbb{E}[\eta_{hi}|G_i = g] = 0$ for each h and g.[30] The term $\mu_{gh} + A_{gh}$ represents a systematic component of utility for employer h among members of group g, while η_{hi} captures variation in utility arising from individual heterogeneity in preferences. This specification allows "typical" preferences to differ flexibly between genders through the $A_{G_i h}$, and higher moments of η_{ih} beyond the mean may also depend upon G_i (e.g. if men or women have greater variability in preferences).

The form of Equation (3.15) embodies three substantive assumptions. The first is that preferences can be defined at the level of employer categories rather than individual employers, which we require for reasons of statistical power in estimation of $\boldsymbol{\mu}_g$. The second

[30] Equation 3.15 can be obtained from a general additive-in-FJEs form: $U_i(h) = \mu_{G_i h} + \epsilon_{ih}$ with some generic ϵ_{ih} if we define $A_{gh} = \mathbb{E}[\epsilon_{hi}|G_i = g]$ and $\eta_{hi} := \epsilon_{ih} - A_{G_i h}$. Note that Equation (3.15) also nests the canonical conditional logit model (McFadden, 1974), when $\eta_{ih} = \lambda \cdot (u_{ih} - \mathbb{E}[u_{ih}])$ with u_{ih} distributed across h as independent extreme value random variables for all i, and λ a scale parameter.

is quasi-linearity in these first-job-effects, which allows us to separate amenities additively from FJEs. Finally, we take workers to anticipate the mean earnings within their group at a given employer, rather than knowing what their exact outcome will be, so that $\mu_{G_i h} = \mathbb{E}[Y_i(h)|G_i]$ appears in utility rather than $Y_i(h)$ itself. This is consistent with Assumption 3.2, while η_{ih} can still be correlated with $Y_i(h)$ (thus creating endogeneity in FJEs). Both A_{gh} and the distribution of η are taken to be static over the years in which the RSD system was in place.

While the quasi-linearity assumption pins down a unique scale for utility (such that it is measured in dollars), we are also free to fix a location normalization for each i. For an arbitrary fixed employer category h_0, we may define $U_i(h_0) = 0$ for all i. This yields the following interpretation for amenities at any other employer: A_{gh} is the average amount in excess of their expected earnings μ_{gh} at h that workers in group g would be willing to pay to move from h_0 to h. If group g tends to prefer h to h_0 and would be willing to give up *part* of their earnings to stay at h, then $A_{gh} \in [-\mu_{gh}, 0]$. In practice, we choose h_0 to represent Category 4, the hospitals in Finnmark and Sogn og Fjordane.

Let $v_{gh} := \mu_{gh} + A_{gh}$ denote the total systematic component of utility. Define $r_{gh} := \mathbb{E}[R_i|H_i = h, G_i = g]$, where recall that R_i is worker i's random lottery number draw, normalized to the unit interval within each lottery. We make the following assumption:

Assumption 3.7. *$r_{gh} = \alpha_g - \beta \cdot v_{gh}$ for some $\beta > 0$ and α_g.*

Assumption 3.7 formalizes, in a specific way, the intuition that the average lottery number among workers choosing employer h is a proxy for their aggregate preference v_{gh} for that employer. If two employers share a value of r_{gh} (for some g), but differ in their FJEs μ_{gh}, then the difference in amenities at the two employers must offset this difference. This intuition supports assuming that $r_{gh} = \phi_g(v_{gh}, \cdot)$ for some decreasing, possibly non-linear function ϕ_g that itself depends on all of the other v_{hg}, the distribution of (η_{ih}, G_i), and the number of slots available for each h in each run of the lottery. But even parametric assumptions on the η_{hi} (such as the logit model) do not appear to readily imply reduced-form expressions for ϕ_g. Assumption 3.7 reflects the simplest functional form assumption that can reasonably fit the data.[31]

Our goal is to use Assumption 3.7 along with the estimated FJE's and observable r_{gh} to pin down the α_g and β, and hence the amenities A_{gh}. Given the four employer categories, two demographic groups, and utility normalization that implies $A_{gh_0} = -\mu_{gh_0}$ for each g, Assumption 3.7 involves 9 unknowns (six A_{gh}, two α_g, and β), from 8 equations. Thus, one more restriction is needed for identification. Figure 3.1 plots the estimated μ_{gh} against the r_{gh}, which we use to motivate an eighth restriction.

Observe that Assumption 3.7 implies that for any employer category $h \neq h_0$, we can net out the α_g and β parameters to write:

$$\frac{\Delta r_{fh}}{\Delta r_{mh}} = \frac{\Delta A_{fh} + \Delta \mu_{fh}}{\Delta A_{mh} + \Delta \mu_{mh}} \tag{3.16}$$

where for any quantity X, ΔX_{hg} denotes the difference between employer categories h and h_0: $\Delta X_{gh} := X_{gh} - X_{gh_0}$. Comparing categories 1 and 4 in Figure 3.1, we observe that both Δr_{g1} and $\Delta \mu_{g1}$ are nearly identical across genders $g \in \{f, m\}$. By Equation (3.16), this suggests that $\Delta A_{m1} \approx \Delta A_{f1}$, regardless of the values of β and α_g.[32] We thus assume that $\Delta A_{m1} = \Delta A_{f1}$ exactly as a reasonable ninth equation, allowing us to point identify all parameters. Intuitively, this restriction says that men and women exhibit the same willingness to pay—in excess of the earnings difference—for the mostly large, urban employers in Category 1, compared with the smaller, rural employers in Category 4. Given α_g and β we can extract each amenity value A_{gh}, as described in the caption of Figure 3.1.

[31] In particular, the group-specific intercept α_g allows us to reconcile the data with reasonable values of β; although women have significantly lower FJEs for all h, they tend to have similar r_{gh} to men. This is natural: in the limit of a constant earnings gap $\mu_{mh} = \mu_{fh} + \delta$ and no differences in amenities across genders, we would expect that $r_{fh} = r_{mh}$, which requires $\alpha_m = \alpha_f + \beta\delta$. We note however that linearity in Assumption 3.7 can only hold as an approximation for some range of v_{gh}, since R_i only has support on the unit interval. In practice, our r_{gh} range between 0.34 and 0.75.

[32] In principle, ΔA_{m1} and ΔA_{f1} could still be arbitrarily far apart, but the values of the parameters in Assumption 3.7 required to sustain such differences then imply unreasonable values of ΔA_{g1}. We calculated the ΔA_{m1} implied by Equation (3.16) as a function of ΔA_{f1} for $\Delta A_{f1} \in [\$10,000, \$200,000]$. The maximum relative difference $(\Delta A_{m1} - \Delta A_{f1})/\Delta A_{f1}$ is about 2.5% (which occurs for the smallest ΔA_{f1} in that range).

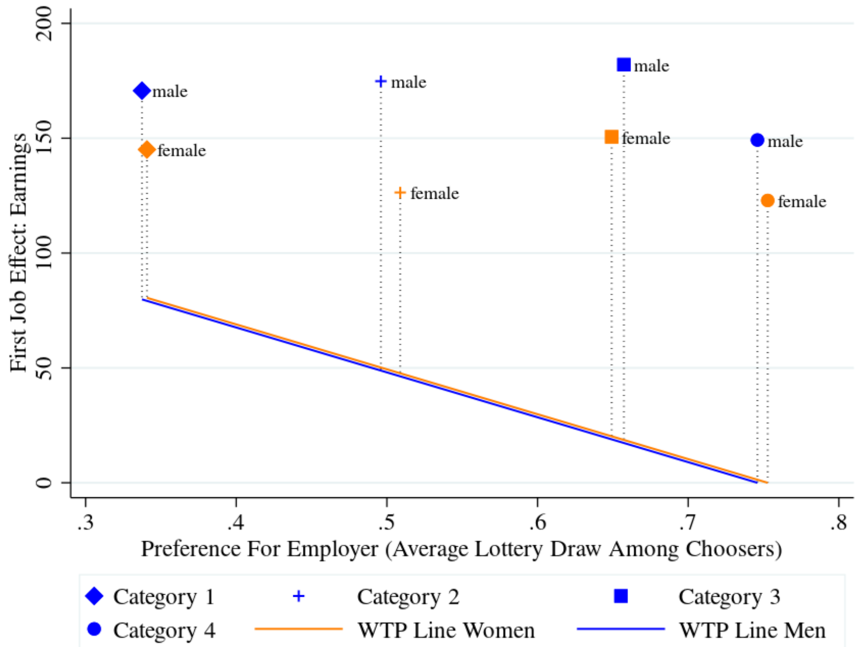

Figure 3.1: Employer amenity values, based on earnings first job effects (thousands 2020 USD) and the average lottery number at which each employer is chosen. The line for each demographic group g yields the systematic component of utility $\mu_{gh} + A_{gh}$, indicating worker's average willingness to pay to move from Category 4, as a function of average lottery number. Amenities A_{gh} are weakly negative for all categories (with magnitudes depicted by vertical dotted lines), reflecting that workers would give up only part of their income to stay at their employer rather than Catgeory 4.

3.6.2 The effects of decentralization

Estimates of amenities A_{gh} now allow us to construct the systematic component of utility at each employer $\mu_{gh} + A_{gh}$ for each group h (in dollar terms), in turn allowing us to approximate average welfare given the new distribution of workers over employers in the post-reform period. And given that we know μ_{gh} and A_{gh} separately, we can decompose this change into changes in earnings and changes in the amenity value of realized employer matches.

Specifically, let $P_{gh} := P(H_i = h|G_i = g)$ be the match probability for group g at employer h, in the pre-reform period, and let $\tilde{P}_{gh}$ be the corresponding probability in the post-reform period. The change in total welfare for group g can be calculated as

$$\sum_h (P_{gh} - \tilde{P}_{gh})(\mu_{gh} + A_{gh})$$

and this change can further be decomposed as a change in earnings $\sum_h (P_{gh} - \tilde{P}_{gh}) \cdot \mu_{gh}$ and an amenity value $\sum_h (P_{gh} - \tilde{P}_{gh}) \cdot A_{gh}$. In addition to assuming first job effects μ_{gh} and average amenity values A_{gh} are stable over time, these calculations consider welfare as captured by these systematic components of utility only. Table 3.8 reports the results.

			Distribution of Workers (%)			
Employer Category	FJEs By Gender	Amenity Values	Pre-Reform (RSD) Women	Pre-Reform (RSD) Men	Post-Reform Women	Post-Reform Men
1	W: 145.07 M: 170.71	W: -68.55 M: -94.87	33.89	33.43	36.89	39.22
2	W: 126.38 M: 174.83	W: -81.11 M: -128.38	37.85	35.38	37.62	34.05
3	W: 150.61 M: 182.04	W: -131.41 M: -165.59	22.49	24.50	19.31	20.94
4	W: 122.89 M: 149.22	W: -122.89 M: -149.22	5.77	6.70	6.18	5.79
Average Predicted Earnings (5 Years Out)			137.96	173.50	137.74	173.24
Average Post-Reform Difference					-0.22	-0.26
Average Predicted Amenity Values			-90.57	-127.69	-88.77	-124.24
Average Post-Reform Difference					1.80	3.45
Total Change in Welfare (Per Worker)					1.58	3.19
Workers	9,049	9,049	4,855	4,194	1,781	1,122

Table 3.8: FJEs measure earnings five year post graduation (four years post residency). Earnings and amenity values in thousands of 2020 USD. Pre/post-reform total welfare$_g$ = $\sum_h prob(h|g)(\mu_{gh} + A_{gh})$ where $prob(h|g)$ = columns 6-7/8-9.

The second column of Table 3.8 reports the earnings FJEs μ_{gh} (also plotted in Figure 3.1) while the third column reports amenitie values A_{gh}. Amenities fall in the range $[-\mu_{gh}, 0]$, indicating that workers would give up some fraction of the earnings at their chosen employer to remain there instead of moving to Category 4. The A_{gh} are generally increasing (decreasing in magnitude) in category popularity while earnings FJEs exhibit a flatter trend. Workers' combined surplus falls at nearly identical rates between men and women as a function of average lottery draw.

Overall, both men and women lose with regards to earnings FJEs, while gaining—to a greater extent—in employer amenities, with the post-reform distribution of workers over employers. The net effect of the decentralized labor market on worker welfare is positive but not large, representing about 4.71% of the pre-reform average of v_{gh}. Men gain more than women in employer amenities. We conclude that, in the setting we study, first jobs affect workers' long-run career trajectories; they do so differentially for men and women; and "market design" policy can affect the aggregate realized effects of workers' first jobs.

References

Abdulkadiroğlu, A., Angrist, J. D., Narita, Y. and Pathak, P. A. (2017). "Research Design Meets Market Design: Using Centralized Assignment For Impact Evaluation". *Econometrica* 85 (5), pp. 1373–1432.

Abdulkadiroğlu, A. and Sönmez, T. (1998). "Random Serial Dictatorship and the Core from Random Endowments in House Allocation Problems". *Econometrica* 66 (3), pp. 689–702.

Abowd, J. M., Kramarz, F. and Margolis, D. N. (1999). "High Wage Workers and High Wage Firms". *Econometrica* 67 (2), pp. 251–333.

Andrews, D. W. K. and Soares, G. (2010). "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection". *Econometrica* 78 (1), pp. 119–157.

Angrist, B. J. D. and Evans, W. N. (1998). "Children and Their Parents ' Labor Supply: Evidence from Exogenous Variation in Family Size". *American Economic Review* 88 (3), pp. 450–477.

Angrist, J. D. (1990). "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records". *American Economic Review* 80 (3), pp. 313–336.

– (1998). "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants". *Econometrica* 66 (2), pp. 249–288.

Angrist, J. D. and Chen, S. H. (2011). "Schooling and the Vietnam-Era GI Bill: Evidence from the Draft Lottery". *American Economic Journal: Applied Economics* 3 (2), pp. 96–118.

Angrist, J. D., Graddy, K. and Imbens, G. W. (2000). "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish". *Review of Economic Studies* 67 (3), pp. 499–527.

Angrist, J. D. and Imbens, G. W. (1994). "Identification and Estimation of Local Average Treatment Effects". *Econometrica* 62 (2), pp. 467–475.

– (1995). "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity". *Journal of the American Statistical Association* 90 (430), pp. 431–442.

Arellano-Bover, J. (2020). "Career Consequences of Firm Heterogeneity for Young Workers: First Job and Firm Size".

Barkume, A. (2010). "The Structure of Labor Costs with Overtime Work in U.S. Jobs". *Industrial and Labor Relations Review* 64 (1).

Bender, S., Schmieder, J. and von Wachter, T. (2009). "The Long-Term Impact of Job Displacement in Germany During the 1982 Recession on Earnings, Income, and Employment".

Bertanha, M., Mccallum, A. H. and Seegert, N. (2020). "Better Bunching , Nicer Notching". *SSRN Working Paper*.

Best, M. C., Brockmeyer, A., Kleven, H. J., Spinnewijn, J. and Waseem, M. (2015). "Production vs Revenue Efficiency With Limited Tax Capacity: Theory and Evidence From Pakistan". *Journal of Political Economy* 123 (6), p. 48.

Bishow, J. L. (2009). "A Look at Supplemental Pay: Overtime Pay, Bonuses, and Shift Differentials". *Monthly Labor Review*.

Blomquist, S., Kumar, A., Liang, C.-Y. and Newey, W. K. (2015). "Individual heterogeneity, nonlinear budget sets and taxable income". *The Institute for Fiscal Studies Working Paper* CWP21/15.

Blomquist, S., Newey, W., Kumar, A. and Liang, C.-Y. (2021). "On Bunching and Identification of the Taxable Income Elasticity". *Journal of Political Economy* (Forthcoming).

Blomquist, S. and Newey, W. (2017). “The Bunching Estimator Cannot Identify the Taxable Income Elasticity”. *The Institute for Fiscal Studies Working Paper* CWP40/17.

Brechling, F. P. R. (1965). “The Relationship Between Output and Employment in British Manufacturing Industries”. *The Review of Economic Studies* 32 (3), p. 187.

Brown, C. and Hamermesh, D. S. (2019). “Wages and Hours Laws: what do we know? what can be done?” *The Russell Sage Foundation Journal of the Social Sciences* 5 (5), pp. 68–87.

Burdett, K. and Mortensen, D. T. (1998). “Wage Differentials, Employer Size, and Unemployment”. *International Economic Review* 39 (2), p. 257.

Caetano, C., Caetano, G. and Nielsen, E. (2020). “Correcting for Endogeneity in Models with Bunching”. *Federal Reserve Board Finance and Economics Discussion Series* 2020-080.

Cahuc, P. and Zylberberg, A. (2014). *Labor economics*. 2nd ed. Cambridge, Mass.: MIT Press.

Carneiro, P., Heckman, J. J. and Vytlacil, E. J. (2011). “Estimating marginal returns to education”. *American Economic Review* 101 (6), pp. 2754–2781.

Cattaneo, M. D., Jansson, M. and Ma, X. (2020). “Simple Local Polynomial Density Estimators”. *Journal of the American Statistical Association* 115 (531), pp. 1449–1455.

Cesarini, D., Lindqvist, E., Notowidigdo, M. J. and Östling, R. (2017). “The Effect of Wealth on Individual and Household Labor Supply: Evidence from Swedish Lotteries”. *American Economic Review* 107 (12), pp. 3917–46.

Chaisemartin, C. de (2017). “Tolerating defiance? Local average treatment effects without monotonicity”. *Quantitative Economics* 8 (2), pp. 367–396.

Chernozhukov, V. and Hansen, C. (2005). "An IV Model of Quantile Treatment Effects". *Econometrica* 73 (1), pp. 245–261.

Chernozhukov, V., Hong, H. and Tamer, E. (2007). "Estimation and Confidence Regions for Parameter Sets in Econometric Models". *Econometrica* 75 (5), pp. 1243–1284.

Chernozhukov, V., Lee, S. and Rosen, A. M. (2013). "Intersection Bounds: Estimation and Inference". *Econometrica* 81 (2), pp. 667–737.

Chetty, R., Friedman, J. N., Olsen, T. and Pistaferri, L. (2011). "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *Quarterly Journal of Economics* 126 (2), pp. 749–804.

Coile, C. C., Levine, P. B. and McKnight, R. (2012). "Recessions, Older Workers, and Longevity: How Long Are Recessions Good For Your Health?" Working Paper Series (18361).

Costa, D. L. (2000). "Hours of Work and the Fair Labor Standards Act: A Study of Retail and Wholesale Trade, 1938–1950". *Industrial and Labor Relations Review*, p. 17.

D'Haultfœuille, X. and Février, P. (2015). "Identification of Nonseparable Triangular Models With Discrete Instruments". *Econometrica* 83 (3), pp. 1199–1210.

Dümbgen, L., Kolesnyk, P. and Wilke, R. A. (2017). "Bi-log-concave distribution functions". *Journal of Statistical Planning and Inference* 184, pp. 1–17.

Dube, A., Manning, A. and Naidu, S. (2020). "Monopsony and Employer Mis-optimization Explain Why Wages Bunch at Round Numbers". *NBER Working Paper* w24991.

Ehrenberg, R. and Schumann, P. (1982). *Longer hours or more jobs? : an investigation of amending hours legislation to create employment*. New York State School of Industrial and Labor Relations, Cornell University.

Ehrenberg, R. G. (1971). “The Impact of the Overtime Premium on Employment and Hours in U . S . Industry”. *Economic Inquiry* 9 (2).

Einav, L., Finkelstein, A. and Schrimpf, P. (2017). “Bunching at the kink: Implications for spending responses to health insurance contracts”. *Journal of Public Economics* 146, pp. 27–40.

Fan, Y. and Park, S. S. (2010). “Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference”. *Econometric Theory* 26 (3), pp. 931–951.

Feng, J. (2020). “Matching Points : Supplementing Instruments with Covariates in Triangular Models”. *Working Paper*.

Frölich, M. (2007). “Nonparametric IV estimation of local average treatment effects with covariates”. *Journal of Econometrics* 139 (1), pp. 35–75.

Gautier, E. (2020). “Relaxing monotonicity in endogenous selection models and application to surveys”. *Working Paper*.

Gautier, E. and Hoderlein, S. (2011). “A triangular treatment effect model with random coefficients in the selection equation”. *Working Paper*.

Gelber, A. M., Jones, D. and Sacks, D. W. (2020). “Estimating Adjustment Frictions Using Nonlinear Budget Sets: Method and Evidence from the Earnings Test”. *American Economic Journal: Applied Economics* 12 (1), pp. 1–31.

Genda, Y., Kondo, A. and Ohta, S. (2010). “Long-Term Effects of a Recession at Labor Market Entry in Japan and the United States”. *Journal of Human Resources* 45 (1), pp. 157–196.

Goff, L. (2020). “A Vector Monotonicity Assumption for Multiple Instruments”. *arXiv* (2009.00553).

Grigsby, J., Hurst, E. and Yildirmaz, A. (2021). “Aggregate Nominal Wage Adjustments: New Evidence from Administrative Payroll Data”. *American Economic Review* 111 (2), pp. 428–71.

Gunsilius, F. F. (2020). "Identifying multivariate models with binary instruments via cyclically monotone dynamics". *Working Paper*.

Hamermesh, D. S. (1996). *Labor demand*. Princeton, NJ: Princeton Univ. Press.

Hamermesh, D. S. and Trejo, S. J. (2000). "The Demand for Hours of Labor : Direct Evidence from California". *The Review of Economics and Statistics* 82 (1), pp. 38–47.

Hart, R. A. (2004). *The economics of overtime working*. Cambridge, UK: Cambridge University Press.

Heckman, J. J. and Pinto, R. (2018). "Unordered Monotonicity". *Econometrica* 86 (1), pp. 1–35.

Heckman, J. J., Urzua, S. and Vytlacil, E. (2006). "Understanding Instrumental Variables in Models with Essential Heterogeneity". *The Review of Economics and Statistics* 88 (3), pp. 389–432.

Heckman, J. J. and Vytlacil, E. (2005). "Structural Equations, Treatment Effects, and Econometric Policy Evaluation". *Econometrica* 73 (3), pp. 669–738.

Heisz, A., Oreopoulos, P. and Wachter, T. M. von (2012). "Short- and Long-Term Career Effects of Graduating in a Recession". *American Economic Journal: Applied Economics* 4 (1), pp. 1–29.

Hjort, J., Li, X. and Sarsons, H. (2020). "Across-Country Wage Compression in Multinationals". *NBER Working Paper* w26788.

Hoerl, A. and Kennard, R. (1970). "Ridge regression : Biased estimation for nonorthogonal problems". *Technometrics* 42 (1), pp. 80–86.

Huang, C.-i. (2008). "Estimating demand for cellular phone service under nonlinear pricing". *Quantitative Marketing and Economics volume* 6, pp. 371–413.

Imbens, B. G. W. and Newey, W. K. (2009). “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity”. *Econometrica* 77 (5), pp. 1481–1512.

Imbens, G. W. and Angrist, J. D. (1994). “Identification and Estimation of Local Average Treatment Effects”. *Econometrica* 62 (2), pp. 467–475.

Imbens, G. W. and Manski, C. F. (2004). “Confidence Intervals for Partially Identified Parameters”. *Econometrica* 72, p. 14.

Imbens, G. W., Rubin, D. B. and Sacerdote, B. I. (2001). “Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players”. *The American Economic Review* 91 (4), p. 25.

Ito, K. and Sallee, J. M. (2017). “The Economics of Attribute-Based Regulation: Theory and Evidence from Fuel-Economy Standards”. *The Review of Economics and Statistics*, pp. 319–336.

Johnson, J. H. (2003). “The Impact of Federal Overtime Legislation on Public Sector Labor Markets”. *Journal of Labor Economics* 21 (1), pp. 43–69.

Kahn, L. B. (2010). “The Long-Term Labor Market Consequences of Graduating from College in a Bad Economy”. *Labour Economics* 17 (2), pp. 303–316.

Kalouptsidi, M., Kitamura, Y., Lima, L. and Souza-Rodrigues, E. A. (2020). “Partial Identification and Inference for Dynamic Models and Counterfactuals”. NBER Working Paper Series (w26761).

Kamat, V. (2020). “Identification of Program Access Effects with an Application to Head Start”.

Kasy, M. (2017). “Who wins, who loses? Identification of the welfare impact of changing wages”. *Working Paper*, pp. 1–26.

Kirkeboen, L., Leuven, E. and Mogstad, M. (2017). “Field of Study, Earnings, and Self-selection”. *Quarterly Journal of Economics* (3).

Kisielewicz, A. (1988). "A solution of Dedekind's problem on the number of isotone Boolean functions." *Journal fur die Reine und Angewandte Mathematik* 386, pp. 139–144.

Kitamura, Y. and Stoye, J. (2018). "Nonparametric Analysis of Random Utility Models". *Econometrica* 86 (6), pp. 1883–1909.

Kleitman, D. J. and Milner, E. C. (1973). "On the average size of the sets in a Sperner family". *Discrete Mathematics* 6 (2), pp. 141–147.

Kleven, H. J. and Waseem, M. (2013a). "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan*". *The Quarterly Journal of Economics* 128 (2), pp. 669–723.

Kleven, H. J. (2016). "Bunching". *Annual Review of Economics* 8 (June), pp. 435–464.

Kleven, H. J. and Waseem, M (2013b). "Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan". *The Quarterly Journal of Economics* 128 (2), pp. 669–723.

Kolesar, M. (2015). "Estimation in an instrumental variables model with treatment effect heterogeneity". (November 2013).

Lee, S. and Salanié, B. (2020). "Filtered and Unfiltered Treatment Effects with Targeting Instruments". *arXiv* (2007.10432).

Lee, S. and Salanié, B. (2018). "Identifying Effects of Multivalued Treatments". *Econometrica* 86 (6), pp. 1939–1963.

Lewbel, A. and Yang, T. T. (2016). "Identifying the average treatment effect in ordered treatment models without unconfoundedness". *Journal of Econometrics* 195 (1), pp. 1–22.

Lewis, H. G. (1969). "Employer Interest in Employee Hours of Work". *Unpublished paper*.

Manning, A. (2003). *Monopsony in Motion: Imperfect Competition in Labor Markets*. Princeton: Princeton University Press.

Manski, C. F. (1990). "Nonparametric Bounds on Treatment Effects". *The American Economic Review* 80 (2), pp. 829–823.

McFadden, D. (1974). "Conditional Logit Analysis of Qualitative Choice Behavior". *Frontiers in Econometrics*. New York: Academic Press, pp. 105–142.

Milgrom, P. and Roberts, J. (1996). "The LeChatelier Principle". *American Economic Review* 1 (86), pp. 173–179.

Mogstad, M., Santos, A. and Torgovitsky, A. (2018). "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters". *Econometrica* 86 (5), pp. 1589–1619.

Mogstad, M., Torgovitsky, A. and Walters, C. (2019). "Identification of Causal Effects with Multiple Instruments: Problems and Some Solutions". *Working Paper*.

– (2020a). "Policy Evaluation with Multiple Instrumental Variables". *Working Paper*.

– (2020b). "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables". *Working Paper*.

Mountjoy, J. (2019). "Community Colleges and Upward Mobility". *Working Paper*, pp. 1–83.

Oreopoulos, P., Wachter, T. von and Heisz, A. (2012). "The Short- and Long-Term Career Effects of Graduating in a Recession". *American Economic Journal: Applied Economics* 4 (1), pp. 1–29.

Oyer, P. (2006). "Initial Labor Market Conditions and Long-Term Outcomes for Economists". *Journal of Economic Perspectives* 20 (3), pp. 143–160.

Oyer, P. (2008). “The Making of an Investment Banker: Macroeconomic Shocks, Career Choice, and Lifetime Income”. *Journal of Finance*.

Quach, S. (2020). “The Labor Market Effects of Expanding Overtime Coverage”. *MPRA Paper* 100613.

Rosen, S. (1968). “Short-Run Employment Variation on Class-I Railroads in the U.S., 1947-1963”. *Econometrica* 36 (3), p. 511.

Roth, A. E. (1984). “The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory”. *Journal of Political Economy* 92 (6), pp. 991–1016.

– (1986). “On the Allocation of Residents to Rural Hospitals: A General Property of Two-Sided Matching Markets”. *Econometrica* 54 (2), pp. 425–427.

Ruhm, C. J. (2000). “Are Recessions Good for Your Health?” *The Quarterly Journal of Economics* 115 (2), pp. 617–650.

Saez, E. (2010). “Do Taxpayers Bunch at Kink Points?” *American Economic Journal: Economic Policy* 2 (3), pp. 180–212.

Sato, R. (1975). “The Most General Class of CES Functions”. *Econometrica* 43 (5), p. 999.

Smeulders, B., Cherchye, L. and De Rock, B. (2021). “Nonparametric Analysis of Random Utility Models: Computational Tools for Statistical Testing”. *Econometrica* 89 (1), pp. 437–455.

Society for Human Resource Management (2018). “National Study of Employers”, p. 79.

Sorkin, I. (2018). “Ranking Firms Using Revealed Preference”. *The Quarterly Journal of Economics* 133 (3), pp. 1331–1393.

Staiger, M. (2021). “The Intergenerational Transmission of Employers and the Earnings of Young Workers”. *Washington Center for Equitable Growth Working Paper*.

Stole, L. A. and Zwiebel, J. (1996). “Intra-Firm Bargaining under Non-Binding Contracts”. *The Review of Economic Studies* 63 (3), pp. 375–410.

Stoye, J. (2009). “More on Confidence Intervals for Partially Identified Parameters”. *Econometrica* 77 (4), pp. 1299–1315.

Sullivan, D. and von Wachter, T. (2009). “Job Displacement and Mortality: An Analysis using Administrative Data”. *Quarterly Journal of Economics*.

Sullivan, D. and Wachter, T. von (2009). “Average Earnings and Long-Term Mortality: Evidence from Administrative Data”. *American Economic Review* 99 (2), pp. 133–38.

Torgovitsky, A. (2015). “Identification of Nonseparable Models Using Instruments With Small Support”. *Econometrica* 83 (3), pp. 1185–1197.

Trejo, S. J. (1991). “The Effects of Overtime Pay Regulation on Worker Compensation”. *American Economic Review* 81 (4), pp. 719–740.

U.S. Department of Labor (2019). “Defining and Delimiting the Exemptions for Executive, Administrative, Professional, Outside Sales and Computer Employees”. *Federal Register* 84 (188).

von Wachter, T. (2020). “The Persistent Effects of Initial Labor Market Conditions for Young Adults and Their Sources”. *The Journal of Economic Perspectives* 34 (4), pp. 168–194.

von Wachter, T. and Bender, S. (2006). “In the Right Place at the Wrong Time: The Role of Firms and Luck in Young Workers’ Careers”. *The American Economic Review* 96 (5), pp. 1679–1705.

Yin, J., Geng, Z., Li, R. and Wang, H. (2010). “Nonparametric covariance model”. *Statistica Sinica* 20 (1), pp. 469–479.

Appendix A: Supplements to Chapter 1

A.1 Identification in a generalized bunching design

This section develops the formal results used in the paper. While the FLSA will provide a running example throughout, I largely abstract from the overtime context to emphasize the wide applicability of the results. To facilitate comparison with the existing literature on bunching at kinks – which has mostly considered cross-sectional data – I throughout this section suppress time indices and use the single index i to refer to each unit of observation (a paycheck in the overtime case).

Further, the "running variable" of the bunching design is denoted throughout this section by Y rather than h. This is done to emphasize the link to the treatment effects literature, while allowing a distinction that can is in some cases necessary (e.g. a model where hours of pay for work differ from actual hours of work).

A.1.1 A generalized bunching-design model

Consider decision-makers i who choose a point $(z, \mathbf{x})$ in some space $\mathcal{X} \subseteq \mathbb{R}^{d+1}$ where z is a scalar and **x** a vector of d components, subject to a constraint of the form:

$$z \geq \max\{B_{0i}(\mathbf{x}), B_{1i}(\mathbf{x})\} \tag{A.1}$$

We require that $B_{0i}(\mathbf{x})$ and $B_{1i}(\mathbf{x})$ are continuous and weakly convex functions of the vector **x**, and that there exist continuous scalar functions $y_i(\mathbf{x})$ and a scalar k such that:

$$B_{0i}(\mathbf{x}) > B_{1i}(\mathbf{x}) \text{ whenever } y_i(\mathbf{x}) < k \quad \text{and} \quad B_{0i}(\mathbf{x}) < B_{1i}(\mathbf{x}) \text{ whenever } y_i(\mathbf{x}) > k$$

The value k is taken to be common to all units i, and is assumed to be known by the researcher.[1] In the overtime setting, $y_i(\mathbf{x})$ represents the hours of work for which a worker is paid in a given week, and $k = 40$. Let X_i be i's realized outcome of $\mathbf{x}$, and $Y_i = y_i(X_i)$. I assume that Y_i is observed by the econometrician, but not that X_i is.

In a typical example, the functions B_{0i}, B_{1i} will represent a schedule of some kind of "cost" as a function of the choice vector $\mathbf{x}$, with two regimes of costs that are separated by the condition $y_i(\mathbf{x}) = k$, characterizing the locus of points at which the two cost functions cross. Let $B_{ki}(\mathbf{x}) := \max\{B_{0i}(\mathbf{x}), B_{1i}(\mathbf{x})\}$. Budget constraints like Eq. $z \geq B_{ki}(\mathbf{x})$ are typically "kinked" because while the function $B_{ki}(\mathbf{x})$ is continuous, it will generally be non-differentiable at the $\mathbf{x}$ for which $y_i(\mathbf{x}) = k$.[2] While the functions B_0, B_1 and y can all depend on i, I will often suppress this dependency for clarity of notation.

In the most common cases from the literature, $\mathbf{x}$ is assumed to be the scalar $y_i(x) = x$, i.e. there is no distinction between the "kink variable" y and underlying choice variables $\mathbf{x}$. For example, the seminal bunching design papers Saez (2010) and Chetty et al. (2011) considered progressive taxation with z being tax liability (or credits), both $y = x$ corresponding to taxable income, and B_0 and B_1 linear tax functions on either side of a threshold y between two adjacent tax/benefit brackets. However, even when the functions B_0 and B_1 only depend on $\mathbf{x}$ through $y_i(\mathbf{x})$, the bunching design is compatible with models in which multiple margins of choice respond to the incentives provided by the kink.[3] In fact, the econometrician may be agnostic as to even what the full set of components of $\mathbf{x}$ are, with $y(\cdot)$, $B_0(\cdot)$ or $B_1(\cdot)$ depending only on various subsets of them. The next section will

[1] This comes at little cost of generality since with heterogeneous k_i this could be subsumed as a constant into the function $y_i(\mathbf{x})$, so long as the k_i are observed by the researcher.

[2] In particular, the subgradient of $\max\{B_{0i}(\mathbf{x}), B_{1i}(\mathbf{x})\}$ will depend on whether one approaches from the $y_i(\mathbf{x}) > k$ or the $y_i(\mathbf{x}) < k$ side. For example with a scalar x and linear B_0 and B_1, the derivative of $B_{ki}(x)$ discontinuously rises when $y_i(x) = k$.

[3] An example from the literature in which a distinction between y and $\mathbf{x}$ cannot be avoided is Best et al. (2015). These authors study firms in Pakistan, who pay either a tax on output or a tax on profit, whichever is higher. The two tax schedules cross when the ratio of profits to output crosses a certain threshold that is pinned down by the two respective tax rates. In this case, the variable y depends both on production and on reported costs, leading to two margins of response to the kink: one from choosing the scale of production and the other from choosing whether and how much to misreport costs.

discuss how the bunching design allows us to conduct causal inference on the variable Y_i, but not directly on the underlying choice variables X_i.

In the overtime context, z corresponds to the cost of a single-worker's labor in a single week, and:

$$B_{0i}(y) := w_{it}y \quad \text{and} \quad B_{1i}(y) := 1.5w_iy - 20w_i \tag{A.2}$$

The functions B_0 and B_1 are depicted in Figure A.1 for a single worker with wage $w_i = w$. B_0 describes a setting in which the worker is paid at their straight-time wage w for all hours, regardless of whether they work more or less than 40. B_1 describes a setting in which the worker is instead paid at their overtime rate $1.5w$ for all hours, but the firm is given a subsidy that keeps them indifferent between the two cost schedules at $y = 40$. With these definitions, we can see that the actual labor cost to the firm of any number of hours h is $B_{ki}(y) := \max\{B_{0i}(y), B_{1i}(y)\}$ for worker i.

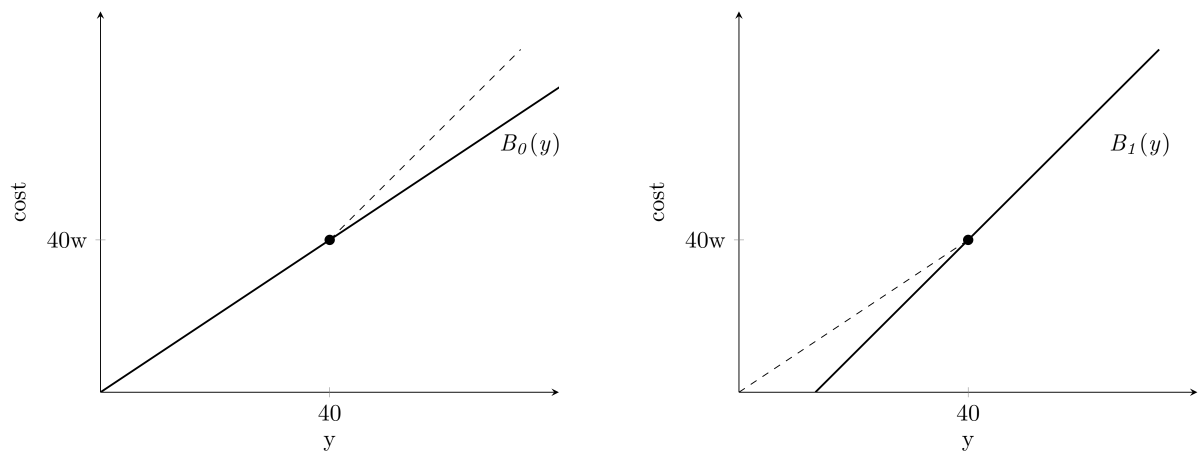

Figure A.1: Definition of counterfactual cost functions B_0 and B_1 that firms could have faced, absent the overtime kink. Dashed lines show the rest of actual cost function in comparison to the counterfactual as a solid line.

A.1.2 Potential outcomes as counterfactual choices

To introduce a notion of treatment effects in the bunching design, I define a pair of potential outcomes as what would occur if the decision-maker faced either of the functions

B_0 or B_1 globally, without the kink:

Definition (potential outcomes). *Let Y_{0i} be the value of $y_i(\mathbf{x})$ that would occur for agent i if they faced the constraint $z \geq B_0(\mathbf{x})$, and let Y_{1i} be the value that would occur under the constraint $z \geq B_1(\mathbf{x})$.*

The above definition requires outcomes y and costs z to be definable at the individual level, but does not require the no-interference condition of the stable unit treatment values assumption (SUTVA). Nevertheless, the interpretation of the treatment effects identified by the bunching design is most straightforward when SUTVA holds. This assumption is standard in the bunching design, though it may be a restrictive one in overtime context where a single firm chooses the hours of multiple workers.[4] I discuss this further in the overtime setting in Section 1.4.4 and Appendix A.3.

To relate these counterfactual outcomes to choices of the decision-maker, we make explicit the assumption that they control the value of $y_i(\mathbf{x})$. For any function B let Y_{Bi} be the outcome that would occur under the choice constraint $z \geq B(\mathbf{x})$, with Y_{0i} and Y_{1i} shorthands for $Y_{B_{0i}i}$ and $Y_{B_{0i}i}$, respectively.[5]

Assumption CHOICE (perfect manipulation of y). *For any function $B(\mathbf{x})$, $Y_{Bi} = y_i(\mathbf{x}_{B_i})$, where $(z_{Bi}, \mathbf{x}_{Bi})$ is the choice that i would make under the constraint $z \geq B(\mathbf{x})$.*

Assumption CHOICE rules out for example optimization error, which could limit the decision-maker's ability to exactly manipulate values of **x** and hence y. It also takes for granted that counterfactual choices are unique, and rules out some kinds of extensive margin effects in which a decision-maker would not choose any value of Y at all under B_1 or B_0. Assumption CHOICE may be relaxed somewhat while still allowing for meaningful causal inference, but I maintain this assumption throughout (however the decision-maker

[4]However I note that SUTVA issues could also occur in canonical bunching designs: for example if spouses choose their labor supply jointly, the introduction of a tax kink may cause one spouse to increase labor supply while the other decreases theirs.

[5]Note that in this notation Assumption CHOICE implies that the actual outcome Y_i observed by the econometrician is equal to $Y_{B_{ki}i}$.

need not always be the firm only; see Appendix A.2). Note that CHOICE here differs from the version given in the main text in that it applies to all functions B, not just B_0, B_1 and B_k (this is useful for Theorem 1.2).

The central behavioral assumption that allows us to reason about the counterfactuals Y_0 and Y_1 is that decision-makers have convex preferences over $(c, \mathbf{x})$ and dislike costs z:

Assumption CONVEX (strictly convex preferences, monotonic in z). *For each agent i and function $B(\mathbf{x})$, choice is $(z_{Bi}, \mathbf{x}_{Bi}) = \text{argmax}_{z,\mathbf{x}}\{u_i(z, \mathbf{x}) : z \geq B(\mathbf{x})\}$ where $u_i(z, \mathbf{x})$ is continuous and strictly quasi-concave in $(z, \mathbf{x})$, and strictly decreasing in z.*

Note that in the overtime setting with firms choosing hours, $u_i(z, \mathbf{x})$ corresponds to the firm's profit function π as a function of the hours of a particular worker (in a particular period), and costs this week for that worker.

A weaker assumption than convexity that will still have identifying power is simply that agents' choices do not violate the weak axiom of revealed preference:

Assumption WARP (rationalizable choices). *Consider two budget functions B and B' and any agent i. If their choice under B' is feasible under B, i.e. $z_{B'i} \geq B(\mathbf{x}_{B'i})$, then $(z_{Bi}, \mathbf{x}_{Bi}) = (z_{B'i}, \mathbf{x}_{B'i})$.*

I make the stronger assumption CONVEX for most of the identification results, but Assumption WARP still allows a version of many of them in which equalities become weak inequalities, indicating a degree of robustness with respect to departures from convexity. Note that the monotonicity assumption in CONVEX implies that choices will always satisfy $z = B(\mathbf{x})$, i.e. agents' choices will lay on their cost functions (despite Eq. A.1 being an inequality, indicating "free-disposal").

In the overtime application, the potential outcomes Y_{0i} and Y_{1i} are the hours that the firm would choose, respectively, in a situation a) in which there was no overtime premium and the firm always had to pay w_i for each hour; and b) a situation in which the firm were to pay $1.5w_i$ for all hours of labor, but receive a subsidy of $20w_i$ that keeps the firm indifferent between B_0 and B_1 when $h = 40$ (cf. Eq. A.2). When firm preferences are quasilinear with respect to wage costs, the choice of hours Y_1 will be the same as what the firm would have chosen without the subsidy of $20w$.

Further notes on the general model

I conclude this section with some further remarks on the generality of Eq. (A.1) given the above assumptions. The first is that the budget functions B_0 and B_1 can depend on a subset of the variables that enter into the function for y, and vice versa. In the former case, this is because the only restriction on the $B_{di}(\mathbf{x})$ for $d \in \{0, 1\}$ is that they are continuous and weakly convex in all components of **x**; thus, having zero dependence on a component of **x** is permissible. This is of particular interest because while the variables entering into the budget functions are generally known from the empirical context generating the kink, the model can allow additional choice variables to enter into the threshold-crossing variable y, that may not even be known to the econometrician. Section 1.4.2 provides some examples of this in the overtime setting.

Suppose that $B_{di}(\mathbf{x}) = B_{di}(\bar{\mathbf{x}})$, where $\bar{\mathbf{x}}$ is a sub-vector of the first m components of **x**, but $y_i(\mathbf{x})$ is still a function of all $m + l$ components of **x**. The values of the remaining l components affect the decision-maker's optimizing choice of y, because they affect the value of y and hence which regime of B_{di} the decision-maker's choice is in. Thus, observed bunching in y can reflect a response along any of these l additional margins, even though they correspond to variables that are unobserved are even unknown to the researcher. This can complicate identification of specific structural elasticities, but does not challenge the credibility of causal inference about y.

A.1.3 Observables in the kink bunching design

Lemma A.1 outlines the core consequence of convexity of preferences for the relationship between observed Y_i and the potential outcomes introduced in the last section:

Lemma A.1 (realized choices as truncated potential outcomes). *Under Assumptions CONVEX and CHOICE:*

$$Y_i = \begin{cases} Y_{0i} & \text{if } \; Y_{0i} < k \\ k & \text{if } \; Y_{1i} \leq k \leq Y_{0i} \\ Y_{1i} & \text{if } \; Y_{1i} > k \end{cases}$$

Proof. See Appendix C.4. □

Lemma A.1 says that the pair of counterfactual outcomes (Y_{0i}, Y_{1i}) is sufficient to pin down actual choice Y_i, which can in fact can be seen as an observation of one or the other potential outcome depending on how they relate to the kink point k. When the Y_{0i} potential outcome is greater than k but the Y_{1i} potential outcome is below – when the potential outcomes "straddle" the kink – the agent will locate choose the corner solution of locating exactly the kink.[6]

Lemma A.1 differs from existing approaches to the bunching design in a basic way by expressing the condition for locating at $Y_i = k$ in terms of the counterfactual choices Y_{0i} and Y_{1i}, rather than primitives of the underlying utility functions $u_i(c, \mathbf{x})$. The typical approach in the literature has been to assume a particular parametric functional form for $u_i(c, \mathbf{x})$, then derive an expression for $\mathcal{B}$ in terms of such parameters (typically an elasticity parameter). Instead, I treat the underlying utility function $u_i(c, \mathbf{x})$ as an intermediate step, only requiring the nonparametric restrictions of convexity and monotonicity. By expressing the bunching event in terms of the "reduced-form" quantity $y_i(\mathbf{x})$, we need only believe that there exists an underlying model of utility satisfying CONVEX, and do not need to know its form explicitly.

Consider a random sample of observations of Y_i. Under i.i.d. sampling of Y_i, the distribution $F(y)$ of Y_i is identified. Let $\mathcal{B} := P(Y_i = k)$ be the observable probability that the agent chooses to locate exactly at $Y = k$. By Lemma A.1, this is equal to the probability

[6] The opposite situation of $Y_{0i} \leq k \leq Y_{1i}$, what we might call "reverse straddling", is ruled out by WARP when it occurs by at least one strict inequality.

of the event $Y_{1i} \leq k \leq Y_{0i}$. With convex preferences, a point mass $\mathcal{B} > 0$ in the distribution of Y_i occurs when the straddling event occurs with positive probability.

Let $\Delta_i = Y_{0i} - Y_{1i}$. This can be thought of as the treatment effect of a counterfactual change from the choice set under B_1 to the choice set under B_0. The straddling event can be expressed in terms of Δ_i as $Y_{0i} \in [k, k + \Delta_i]$. This forms the basic link between the observable quantity $\mathcal{B}$ and treatment effects. **Proposition A.1** states the general result.

Theorem A.1 (relation between bunching and straddling). *a) Under CONVEX and CHOICE: $\mathcal{B} = P(Y_{0i} \in [k, k + \Delta_i])$; b) under WARP and CHOICE: $\mathcal{B} \leq P(Y_{0i} \in [k, k + \Delta_i])$.*

Proof. See Appendix C.4. □

Let $F_1(y) = P(Y_{0i} \leq y)$ be the distribution function of the random variable Y_0, and $F_1(y)$ the distribution function of Y_1. From Lemma A.1 it follows immediately that $F_0(y) = F(y)$ for all $y < k$, and $F_1(y) = F(y)$ for $Y > k$. Thus observations of Y_i are also informative about the marginal distributions of Y_{0i} and Y_{1i}. A weaker version of this also holds under WARP rather than CONVEX:

Corollary (identification of truncated densities). *Suppose that F_0 and F_1 are continuously differentiable with derivatives f_0 and f_1, and that F admits a derivative function $f(y)$ for $y \neq k$. Under WARP and CHOICE: $f_0(y) \leq f(y)$ for $y < k$ and $f_0(k) \leq \lim_{y \uparrow k} f(y)$, while $f_1(y) \leq f(y)$ for $y > k$ and $f_1(k) \leq \lim_{y \downarrow k} f(y)$, with equalities under CONVEX.*

Proof. See Appendix C.4. □

Discussion of treatment effects vs. structural parameters:

The treatment effects Δ_i are "reduced form" in the sense that when the decision-maker has multiple margins of response **x** to the incentives introduced by the kink, these may be bundled together in the treatment effect Δ_i. This clarifies a limitation sometimes levied against the bunching design, while also revealing a perhaps under-appreciated strength. On the one hand, it is not always clear "which elasticity" is elicited by bunching at a kink, complicating efforts to identify a elasticity parameter having a firm structural interpretation.

On the other hand, the bunching design can be useful for ex-post policy evaluation and even forecasting effects of small policy changes (as described in Section 1.4.4), without committing to a tightly parameterized underlying model of choice. The "trick" of Lemma A.1 is to express the observable data in terms of counterfactual choices, rather than of primitives of the utility function. The econometrician need not even know the full vector **x** of choice variables underlying agents' value of y, they simply need to believe that preferences are convex in them, and verify that B_0 and B_1 are convex in a subset of them. This greatly increases the robustness of the method to potential misspecification of the underlying choice model. Appendix A.1 further elucidates some of these issues through an example from the literature.

A.1.4 The buncher LATE when Assumption RANK fails

This section picks up from the discussion in Section 1.4.3, which introduces the buncher LATE Δ_k^* parameter and Assumption RANK, but continues with the notation of this Appendix. When RANK fails (and $p = 0$ for simplicity), the bounds from Theorem 1.1 are still valid for the averaged quantile treatment effect:

$$\frac{1}{\mathcal{B}} \int_{F_0(k)}^{F_1(k)} Q_0(u) - Q_1(u) = \mathbb{E}[Y_{0i}|Y_{0i} \in [k, k+\Delta_0^*]] - \mathbb{E}[Y_{1i}|Y_{1i} \in [k-\Delta_1^*, k]] \tag{A.3}$$

under BLC of Y_0 and Y_1, where we define $\Delta_0^* := Q_0(F_1(k)) - Q_1(F_1(k)) = Q_0(F_1(k)) - k$ and $\Delta_1^* := Q_0(F_0(k)) - Q_1(F_0(k)) = k - Q_1(F_0(k))$. This can be seen to yield a lower bound on the buncher LATE, as described in Figure A.2 below.

Signing the bias when RANK fails

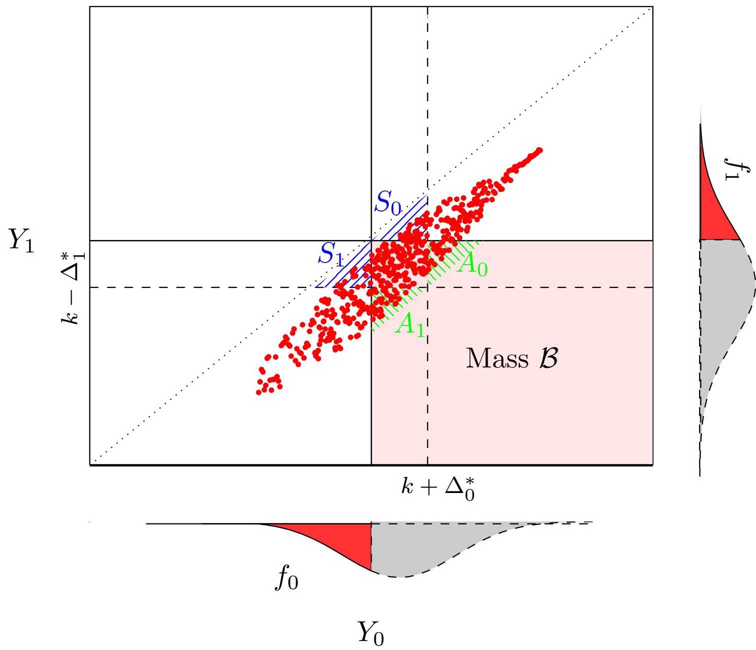

Figure A.2: When Assumption RANK fails, the average $E[Y_{0i}|Y_{0i} \in [k, k+\Delta_0^*]]$ will include the mass in the region S_0, who are not bunchers (blue, NE lines) but will be missing the mass in the region A_0 (green, NW lines) who are. This causes an under-estimate of the desired quantity $E[Y_{0i}|Y_{1i} \leq k \leq Y_{0i}]$. Similarly, $E[Y_{1i}|Y_{1i} \in [k-\Delta_1^*, k]]$ will include the mass in the region S_1, who are not bunchers but will be missing the mass in A_1, who are. This causes an over-estimate of the desired quantity $E[Y_{1i}|Y_{1i} \leq k \leq Y_{0i}]$.

A.1.5 Policy changes in the bunching-design

Consider a bunching design in which the cost functions B_0 and B_1 can be viewed as members of family $B_i(\mathbf{x}; \rho, k)$ parameterized by a continuum of scalars ρ and k, where $B_{0i}(\mathbf{x}) = B_i(\mathbf{x}; \rho_0, k^*)$ and $B_{1i}(\mathbf{x}) = B_i(\mathbf{x}; \rho_1, k^*)$ for some $\rho_1 > \rho_0$ and value k^* of k. In the overtime setting ρ represents a wage-scaling factor, with $\rho = 1$ for straight-time and $\rho = 1.5$ for overtime:

$$B_i(y; \rho, k) = \rho w_i y - k w_i(\rho - 1) \tag{A.4}$$

where work hours y may continue to be a function $y(\mathbf{x})$ of a vector of choice variables to the firm. Here ρ represents an arbitrary wage-scaling factor, while k controls the size of a lump-sum subsidy that keeps $B_i(k;\rho,k)$ invariant across ρ.

Assume that ρ takes values in a convex subset of $\mathbb{R}$ containing ρ_0 and ρ_1, and that for any k and $\rho' > \rho$ the cost functions $B_i(\mathbf{x};\rho,k)$ and $B_i(\mathbf{x};\rho',k)$ satisfy the conditions of the bunching design framework from Section 1.4, with the function $y_i(\mathbf{x})$ fixed across all such values. That is, $B_i(\mathbf{x};\rho',k) > B_i(\mathbf{x};\rho,k)$ iff $y_i(\mathbf{x}) > k$ with equality when $y_i(\mathbf{x}) = k$, the functions $B_i(\cdot;\rho,k)$ are weakly convex and continuous, and $y_i(\cdot)$ is continuous. It is readily verified that Equation (A.4) satisfies these requirements with $y_i(h) = h$.[7]

For any value of ρ, let $Y_i(\rho,k)$ be agent i's realized value of $y_i(\mathbf{x})$ when a choice of $(z,\mathbf{x})$ is made under the constraint $c \geq B_i(\mathbf{x};\rho,k)$. A natural restriction in the overtime setting that is that the function $Y_i(\rho,k)$ does not depend on k, and some of the results below will require this. A sufficient condition for $Y_i(\rho,k) = Y_i(\rho)$ is a family of cost functions that are linearly separable in k, as we have in Equation (A.4), along with quasi-linearity of preferences:

Assumption SEPARABLE (invariance of potential outcomes with respect to k). *For all i,ρ and k, $B_i(\mathbf{x};\rho,k)$ is additively separable between k and $\mathbf{x}$ (e.g. $b_i(\mathbf{x},\rho) + \phi_i(\rho,k)$ for some functions b_i and ϕ_i), and for all i $u_i(z,\mathbf{x})$ can be chosen to be additively separable and linear in z.*

Quasilinearity of preferences is a property of profit-maximizing firms when c represents a cost, thus it is a natural assumption in the overtime setting. However, additive separability of $B_(\mathbf{x};\rho,k)$ in k may be context specific: in the example from Best et al. (2015) described in Appendix A.1, quasi-linearity of preferences is not sufficient since the cost functions are not additively separable in k. To maintain clarity of exposition, I will keep k implicit in $Y_i(\rho)$ throughout the foregoing discussion, but the proofs make it clear when SEPARABLE is being used.

Below I state two intermediate results that allow us to derive expressions for the effects of marginal changes to ρ_1 or k on hours. Lemma A.2 generalizes an existing result from

[7] As an alternative example, I construct in Appendix A.1 functions $B_i(\mathbf{x};\rho,k)$ for the bunching design setting from Best et al. (2015). In that case, ρ parameterizes a smooth transition between an output and a profit tax, where k enters into the rate applied to the tax base for that value of ρ.

Blomquist et al. (2021), and makes use of a regularity condition I introduce in the proof as Assumption SMOOTH.[8] Counterfactual bunchers $K_i^* = 1$ are assumed to stay at k^*, regardless of ρ and k. Let $p(k) = p \cdot \mathbb{1}(k = k^*)$ denote the possible counterfactual mass at the kink as a function of k. Let $f_\rho(y)$ be the density of $Y_i(\rho)$, which exists by SMOOTH and is defined for $y = k^*$ as a limit (see proof).

Lemma A.2 (bunching from marginal responsiveness). *Assume CHOICE, SMOOTH and WARP. Then:*

$$\mathcal{B} - p(k) \leq \int_{\rho_0}^{\rho_1} f_\rho(k)\mathbb{E}\left[-\frac{dY_i(\rho)}{d\rho}\middle| Y_i(\rho) = k\right] d\rho$$

with equality under CONVEX.

Proof. See Appendix C.4. □

Lemma A.2 is particularly useful when combined with a result from Kasy (2017), which considers how the distribution of a generic outcome variable changes as heterogeneous units flow to different values of that variable in response to marginal policy changes.

Lemma A.3 (flows under a small change to ρ). *Under SMOOTH:*

$$\partial_\rho f_\rho(y) = \partial_y \left\{ f_\rho(y)\mathbb{E}\left[-\frac{dY_i(\rho)}{d\rho}\middle| Y_i(\rho) = y, K_i^* = 0\right]\right\}$$

Proof. See Appendix C.4. □

The intuition behind Lemma A.3 comes from fluid dynamics. When ρ changes, a mass of units will "flow" out of a small neighborhood around any y, and this mass is proportional to the density at y and to the average rate at which units move in response to the change. When the magnitude of this net flow varies with y, the change to ρ will lead to a change in the density there.

[8]Blomquist et al. (2021) derive the special case of Lemma A.2 with CONVEX and $p = 0$, in the context of a more restricted model of labor supply under taxation. I establish it here for the general bunching design model where in particular, the $Y_i(\rho)$ may depend on an underlying vector $\mathbf{x}$ which are not observed by the econometrician. I also use different regularity conditions.

With ρ_0 fixed at some value, let us index observed Y_i and bunching $\mathcal{B}$ with the superscript $[k, \rho_1]$ when they occur in a kinked policy environment with cost functions $B_i(\cdot; \rho_0, k)$ and $B_i(\cdot; \rho_1, k)$. Lemmas A.2 and A.3 together imply Theorem 1.2, which I repeat here:

Theorem 2 (marginal comparative statics in the bunching design). *Under Assumptions CHOICE, CONVEX, SMOOTH, and SEPARABLE:*

1. $\partial_k \left\{ \mathcal{B}^{[k,\rho_1]} - p(k) \right\} = f_1(k) - f_0(k)$
2. $\partial_k \mathbb{E}[Y_i^{[k,\rho_1]}] = \mathcal{B}^{[k,\rho_1]} - p(k)$
3. $\partial_{\rho_1} \mathbb{E}[Y_i^{[k,\rho_1]}] = -\int_k^\infty f_{\rho_1}(y) \mathbb{E}\left[\frac{dY_i(\rho_1)}{d\rho} \middle| Y_i(\rho_1) = y \right] dy$

Proof. See Appendix C.4. □

Assumption SEPARABLE is only necessary for Items 1-2 in Theorem 1.2, Item 3 holds without it and with $\frac{\partial Y_i(\rho,k)}{\partial \rho}$ replacing $\frac{dY_i(\rho)}{d\rho}$.

A.1.6 Identification results for existing bunching-design approaches

This section shows how three seminal approaches to the bunching design from the literature can be recast in the framework of Section A.1. Throughout this section, I assume that Y_0 and Y_1 admit a density everywhere so there is no counterfactual bunching at the kink. However, the results in this section can still be applied given a known $p = P(Y_{0i} = Y_{1i} = k)$ by trimming this from the observed bunching and re-normalizing the distribution $F(y)$, as described in Section 1.4.3.

Parametric approaches with constant treatment effects

To generalize the notion of constant treatment effects $\Delta_i = \Delta$, let us for any strictly increasing and differentiable transformation $G(\cdot)$ define for each unit i:

$$\delta_i^G := G(Y_{0i}) - G(Y_{1i})$$

For example, with G equal to the logarithm function, Δ_i^G becomes proportional to a reduced form elasticity measuring the percentage change in $y_i(\mathbf{x})$ when moving from constraint B_{1i} to B_{0i}. This notion of treatment effects facilitates comparison with existing work, because familiar models predict that while Δ_i is heterogeneous δ_i^G is homogeneous when G is the natural logarithm function. For simplicity of notation, let us denote δ_i^G by δ_i when G is the natural logarithm. In particular, in the simplest case of a bunching design in which B_0 and B_1 are linear functions of y with slopes ρ_0 and ρ_1 respectively, if utility follows the iso-elastic quasi-linear form of Equation (1.3), we have that

$$\delta_i = \delta := |\epsilon| \cdot \ln(\rho_1/\rho_0)$$

for all units i.

Note that under CHOICE and CONVEX the result of Lemma A.1 holds with $G(\cdot)$ applied to each of Y_i, Y_{0i}, and Y_{1i} since it is strictly increasing, and thus when Δ_i^G is homogeneous for some G we have that

$$\mathcal{B} = P\left(G(Y_{0i}) \in \left[G(k), G(k) + \delta^G\right]\right)$$

by Proposition A.1. We can also identify the density functions f_0^G of $G(Y_{0i})$ and f_1^G of $G(Y_{1i})$ to the left and right of $G(k)$, respectively. Given that the function $G(\cdot)$ is strictly increasing, we may also write the bunching condition as

$$\mathcal{B} = P(Y_{0i} \in [k, k + \Delta]) \text{ where } \Delta = G^{-1}\left(G(k) + \delta^G\right) - k \tag{A.5}$$

which defines a pseudo-parameter Δ that plays the same role as Δ would in a setup in which we assume a constant treatment effects in levels $\Delta_i = \Delta$. For example, the constant elasticity model motivates $G = \ln$ and hence $\Delta = k(e^\delta - 1)$. Note that if Δ can be pinned down, it will also be possible to identify δ. Nevertheless, it will be important to keep track of the function G when δ_i^G is assumed homogeneous, since for example this implies that $f_0^G(G(k)+\delta) = f_1^G(G(k))$ but not that $f_0(k+\Delta) = f_1(k)$.

Recall from Section A.1.2 that when Δ_i is homogeneous and $f_0(y)$ is locally uniform in the missing region $[k, k+\Delta]$, we have that

$$\Delta = \mathcal{B}/f_0(k) \tag{A.6}$$

and thus with constant effects in logs $\delta_i = \delta$, we can identify δ as $\ln(1+\mathcal{B}/\{kf_0(k)\})$. Taking the approximation $\ln(1+x) \approx x$ and defining $\epsilon = \ln((1-\tau_0)/(1-\tau_1))\delta = -\ln(1-(\tau_1-\tau_0)/(1-\tau_0))\delta$ motivated by the iso-elastic model, we obtain $\epsilon \approx (\tau_1-\tau_0)/(1-\tau_0)\cdot\mathcal{B}/\{kf_0(k)\}$, c.f. Equations (1)-(2) in Kleven (2016).

This represents the simplest and most basic point identification result for the bunching design, and might be motivated by the idea that the kink is small, and a smooth density is locally uniform. Equation A.7 generalizes naturally to a setting with heterogeneous treatment effects, as we shall see in the next section. [9] However, the uniform density assumption/approximation underlying Equation A.6 may be hard to motivate in empirical settings where the kink is not small (e.g. $\tau_0 \not\approx \tau_1$), and the density away from the kink does not appear to be uniform. Thus Saez (2010) instead assumes that $f_(y)$ is linear in the missing region $[k, k+\Delta]$. We can phrase his identification result as a special case of the following:

Proposition A.1 (identification by linear interpolation, à la Saez 2010). *If $\delta_i^G = \delta^G$ for some G, $F_1(y)$ and $F_0(y)$ are continuously differentiable, and $f_0(y)$ is linear on the interval $[k, k+$*

[9]Note that the same "small-kink" approximation might be used to motivate instead the expression:

$$\delta^G = \mathcal{B}/f_0^G(k) = G'(k)\cdot\mathcal{B}/f_0(k), \tag{A.7}$$

which is evidently inconsistent with Equation (A.6) when $G'(k) \neq 1$. This illustrates the point that assuming $f_0(y)$ is constant on the region $[k, k+\Delta]$ is not the same as assuming that $f_0^G(y)$ is constant on $[G(k), G(k)+\delta^G]$ when G is non-linear.

$\Delta]$, *then with CONVEX, CHOICE:*

$$\mathcal{B} \;=\; \frac{1}{2}\left(G^{-1}\left(G(k)+\delta\right)-k\right)\left\{\lim_{y\uparrow k} f(y)+\frac{G'(G^{-1}\left(G(k)+\delta\right))}{G'(k)}\lim_{y\downarrow k} f(y)\right\}$$

Proof. See Section A.8. □

In particular, if we assume the iso-elastic utility model Equation (1.3) then we have:

$$\mathcal{B}=\frac{\Delta}{2}\left\{\lim_{y\uparrow k} f(y)+\frac{k}{k+\Delta}\lim_{y\downarrow k} f(y)\right\}=\frac{k}{2}\left(\left(\frac{\rho_0}{\rho_1}\right)^{\epsilon}-1\right)\left(\lim_{y\uparrow k} f(y)+\left(\frac{\rho_0}{\rho_1}\right)^{-\epsilon}\lim_{y\downarrow k} f(y)\right) \tag{A.8}$$

which can be solved for ϵ by the quadratic formula, and serves as the main estimating equation from Saez (2010). Thus the empirical approach of that paper be seen as applying a result justified in a much more general model than the iso-elastic utility function assumed therein.[10]

While Proposition A.1 constitutes a straightforward solution to the identification problem, the linearity assumption may like uniformity be falsified by visual inspection. For example, if we believe that $f_0(y)$ is continuously differentiable and treatment effects in levels are homogeneous (i.e. G is the identity function), then the linear interpolation used by Proposition A.1 cannot hold unless $\lim_{y\downarrow k} f'(y)=\lim_{y\uparrow k} f'(y)$. Otherwise, f_0 would have to have a kink at one of the endpoints of the missing region. This limitation of Proposition A.1 can be seen as a result of the fact that it only uses information about $f_0(y)$ at two points and ignores it everywhere else. A more popular approach, following Chetty et al. (2011), is to use a global polynomial approximation to $f_0(y)$, which interpolates $f_0(y)$ inwards from both directions across the missing region of unknown width Δ. This technique has the added advantage of accommodating diffuse bunching, for which the relevant $\mathcal{B}$ is the "excess-mass" around k rather than a perfect point mass at k.

[10]Note that if we had instead assumed that $f_0^G(y)$ is linear (on the interval $[G(k),G(k)+\delta^G]$), then we simply replace $f(y)$ by $f^G(y)$ in the above and let G be the identity function, which can be readily solved for δ^G with the simpler expression $\delta^G=\mathcal{B}/\frac{1}{2}\left\{\lim_{y\uparrow k} f^G(y)+\lim_{y\downarrow k} f^G(y)\right\}$.

When bunching is exact, as in the overtime setting, the polynomial approach can be seen as a special case of the following result:

Proposition A.2 (identification from global parametric fit, à la Chetty et al. 2011). *Suppose $f_0(y)$ exists and belongs to a parametric family $g(y;\theta)$, where $f_0(y) = g(y;\theta_0)$ for some $\theta_0 \in \Theta$, and that $\delta_i^G = \delta^G$ for some G and CONVEX and CHOICE hold. Then, provided that*

1. *$g(y;\theta)$ is an analytic function of y on the interval $[k, k+\Delta]$ for all $\theta \in \Theta$, and*

2. *$g(y;\theta_0) > 0$ for all $y \in [k, k+\Delta]$,*

Δ is identified as $\Delta(\theta_0)$, where for any θ, $\Delta(\theta)$ is the unique Δ such that $\mathcal{B} = \int_k^{k+\Delta} g(y;\theta)dy$, and θ_0 satisfies

$$f(y) = \begin{cases} g(y;\theta_0) & y < k \\ g(y+\Delta(\theta_0);\theta_0) & y > k \end{cases} \tag{A.9}$$

Proof. See Section A.8. □

The standard approach of fitting a high-order polynomial to $f_0(y)$ can satisfy the assumptions of Proposition A.2, since polynomial functions are analytic everywhere. Proposition A.2 yields an identification result that can justify an estimation approach similar to one often made in the literature, based on Chetty et al. (2011).[11] However, it requires taking seriously the idea that $f_0(y) = g(y;\theta_0)$, treating the approach as parametric rather than as a series approximation to a nonparametric density $f_0(y)$. This assumption is very strong. Indeed, assuming that $g(y;\theta_0)$ follows a polynomial exactly has even more identifying power than is exploited by Proposition A.2. In particular, if we also have that $f_1(y) = g(y;\theta_1)$ then we could use data on either side of the kink to identify by θ_0 and θ_1, which would allow identification of the average treatment effect with complete treatment effect heterogeneity.

[11] The technique proposed by Chetty et al. (2011) in fact ignores the shift term $\Delta(\theta)$ in Equation (A.9), a limitation discussed by Kleven (2016). A more robust estimation procedure for parametric bunching designs could be based on iterating on Equation (A.9) after updating $\Delta(\theta)$, until convergence. I do not pursue this in the present paper.

A uniform density or "small-kink" approximation

Another argument found in the literature (e.g. Saez 2010 and Kleven and Waseem (2013a)) is to allow heterogeneous treatment effects under a uniform density approximation. If a kink is very small, then this might be justified as an approximation by saying that Δ_i must be small for all individuals, then invoking smoothness assumptions on $f(\Delta, y)$ (see the corollary to Proposition A.3 below). My results in Section 1.4 move beyond the need to approximate the kink as small, however I show here how an analog of this result can be stated in my generalized bunching design framework. The result will make use of the following Lemma, which states that treatment effects must be positive at the kink:

Lemma POS (positive treatment effect at the kink). *Under WARP and CHOICE,* $P(\Delta_i \geq 0 | Y_{0i} = k) = P(\Delta_i \geq 0 | Y_{1i} = k) = 1$.

Proof. Suppose $Y_{0i} = k$ and $\Delta_i < 0$, so that $Y_{1i} > k$. The proof of Proposition A.1 shows that if $Y_{0i} \leq k$ then $Y_i = Y_{0i}$, so we must have that $Y_i = k$. However it also shows that $Y_{1i} \geq k$ implies that $Y_i = Y_{1i}$, so $Y_i > k$, a contradiction. An analogous argument holds when $Y_{1i} = k$. $\square$

Proposition A.3 (identification of a LATE under uniform density approximation). *Let* Δ_i *and* Y_{0i} *admit a joint density* $f(\Delta, y)$ *that is continuous in* y *at* $y = k$. *For each value of* Δ *with support: assume that* $f(\Delta, Y_0) = f(\Delta, k)$ *for all* Y_0 *in the region* $[k, k + \Delta]$. *Under Assumptions WARP and CHOICE*

$$\mathbb{E}\left[\Delta_i | Y_{0i} = k\right] \geq \frac{\mathcal{B}}{\lim_{y \uparrow k} f(y)},$$

with equality under CONVEX.

Proof. Note that

$$\begin{aligned}
\mathcal{B} \leq P(Y_{0i} \in [k, k+\Delta_i]) &= \int_0^\infty d\Delta \int_k^{k+\Delta} dy \cdot f(\Delta, y) = \int_0^\infty f(\Delta, k)\Delta d\Delta \\
&= f_0(k)P(\Delta_i \geq 0 | Y_{0i} = k)\mathbb{E}\left[\Delta | Y_{0i} = k, \Delta \geq 0\right] \\
&\leq \lim_{y \uparrow k} f(y) \cdot \mathbb{E}\left[\Delta | Y_{0i} = k\right]
\end{aligned}$$

using Lemma POS in the last step. The inequalities are equalities under CONVEX. □

Analogous assumptions on the joint distribution of Δ_i and Y_{1i} would justify replacing $\lim_{y\uparrow k} f(y)$ with $\lim_{y\uparrow k} f(y)$ in Proposition A.3. Lemma SMALL in Appendix C.4 formalizes the idea that the uniform density approximation from Proposition A.3 becomes exact in the limit of a "small" kink.

A.2 Incorporating workers that set their own hours

This section considers the robustness of the empirical strategy from Section 1.4 to a case where some workers are able to choose their own hours. In this case, a simple extension of the model leads to the bounds on the buncher LATE remaining valid, but it is only directly informative about the effects of the FLSA among workers who have their hours chosen by the firm. In this section I follow the notation from the main text where h_{it} indicate the hours of worker i in week t.

Suppose that some workers are able to choose their hours each week without restriction ("worker-choosers"), and that for the remaining workers ("firm-choosers") their employers set their hours. In general we can allow who chooses hours for a given worker to depend on the period, so let $W_{it} = 1$ indicate that i is a worker-chooser in period t. Additionally, we continue to allow conterfactual bunchers for whom counterfactual hours satisfy $h_{0it} = h_{1it} = 40$, regardless of who chooses them. This setup is general enough to also allow a stylized bargaining-inspired model in which choices maximize a weighted sum of quasilinear worker and firm utilities.[12]

I replace Assumption CONVEX from Section 1.4 allow agents to either dislike pay (firm-choosers), or like pay (worker-choosers):

Assumption CONVEX* (convex preferences, monotonic in either direction). *For each i, t and function $B(\mathbf{x})$, choice is $(c_{Bi}, \mathbf{x}_{Bi}) = \text{argmax}_{c,\mathbf{x}}\{u_i(c, \mathbf{x}) : c \geq B(\mathbf{x})\}$ where $u_i(c, \mathbf{x})$ is continuous and strictly quasi-concave in $(c, \mathbf{x})$, and*

- *strictly increasing in c, if $W_{it} = 1$*
- *strictly decreasing in c, if $W_{it} = 0$*

In this generalized model, bunching is prima-facie evidence that firm-choosers exist, because there is no prediction of bunching among worker-choosers provided that potential outcomes are continuously distributed (by contrast, k is a "hole" in the worker-chooser hours distribution). Indeed under regularity conditions all of the data local to 40 are from firm-choosers (and counterfactual bunchers). To make this claim precise, we assume that for worker-choosers hours are the only margin of response (i.e. their utility depends on $\mathbf{x}$ only thought $y(\mathbf{x})$), and let $IC_{0it}(y)$ and $IC_{1it}(y)$ be the worker's indifference curves passing through h_{0it} and h_{1it}, respectively. I assume these indifference curves are twice Lipschitz differentiable, with $M_{it} := \sup_y \max\{|IC''_{0it}(y)|, |IC''_{1it}(y)|\}$, where the supremum is taken over the support of hours, and IC'' indicates second derivatives.

[12]In particular, suppose that for any pay schedule $B(h)$:

$$h = \underset{h}{\text{argmax}}\, \beta\,(f(h) - c) + (1-\beta)(c - \nu(h)) \quad with \quad c = B(h) \tag{A.10}$$

where $f(h) - c$ is firm profits with concave production f, $c - \nu(h)$ is worker utility with a convex disutility of labor $\nu(h)$, and $\beta \in [0, 1]$ governs the weight of each party in the negotiation (this corresponds to Nash bargaining in which outside options are strictly inferior to all h for both parties, and utility is log-linear in c). Rearranging the maximand of Equation (A.10) as $(1-2\beta)c + \{\beta f(h) - (1-\beta)\nu(h)\}$, we can observe that this setting delivers outcomes as-if chosen by a single agent with quasi-concave preferences, as $\beta f(h) - (1-\beta)\nu(h)$ is concave. For Assumption CONVEX from Section 1.4 to hold with the assumed direction of monotonicity in costs c, we would require that $\beta > 1/2$ for all worker-firm pairs: informally, that firms have more say than workers do in determining hours. However CONVEX* holds regardless of the distribution of β over worker-firm pairs. If $\beta_{it} < 1/2$, paycheck it will look exactly like a worker-chooser, and if $\beta_{it} > 1/2$ paycheck it will look exactly like a firm-chooser.

Proposition A.4. *Suppose that the joint distribution of h_{0it} and h_{1it} admits a continuous density conditional on $K^*_{it} = 0$, and that for any worker-chooser IC_{0it} and IC_{1it} are differentiable with M_{it}/w_{it} having bounded support. Then, under CHOICE and CONVEX*:*

- $P(h_{it} = k \text{ and } K^*_{it} = 0) = P(h_{1it} \leq k \leq h_{0it} \text{ and } K^*_{it} = 0 \text{ and } W_{it} = 0)$
- $\lim_{h\uparrow k} f(h) = P(W_{it} = 0) \lim_{h\uparrow k} f_{0|W=0}(h)$
- $\lim_{h\downarrow k} f(h) = P(W_{it} = 0) \lim_{h\downarrow k} f_{1|W=0}(h)$

Proof. Omitted for brevity. □

The first bullet of Proposition A.4 says that all active bunchers are also firm-choosers, and have potential outcomes that straddle the kink. The second and third bullets state that the density of the data as hours approach 40 from either direction is composed only of worker-choosers. This result on density limits requires the stated regularity condition, which prevents worker indifference curves from becoming too close to themselves featuring a kink (plus a requirement that straight-time wages w_{it} be bounded away from zero).

Given the first item in Proposition A.4, the buncher LATE introduced in Section 1.4 only includes firm-choosers:

$$\mathbb{E}[h_{0it} - h_{1it}|h_{it} - 40, K^*_{it} = 0] = \mathbb{E}[h_{0it} - h_{1it}|h_{it} - 40, K^*_{it} = 0, W_{it} = 0]$$

Accordingly, I assume rank invariance among the firm-chooser population only:

Assumption RANK* (near rank invariance and counterfactual bunchers). *The following are true:*

(a) $P(h_{0it} = k) = P(h_{1it} = k) = p$

(b) $Y = k$ *iff* $h_0 \in [k, k + \Delta^*_0]$ *and* $W = 0$ *iff* $h_1 \in [k - \Delta^*_1, k]$ *and* $W = 0$*, for some* $\Delta^*_0, .1cm\Delta^*_1$

where p continues to denote $P(K_{it}^* = 1)$.

We may now state a version of Theorem 2 that conditions all quantities on $W = 0$, provided that we assume bi-log concavity of h_0 and h_1 conditional on $W = 0$ and $K = 0$.

Theorem 1* (bi-log-concavity bounds on the buncher LATE, with worker-choosers). *Assume CHOICE, CONVEX* and RANK* hold. If both h_{0it} and h_{1it} are bi-log concave conditional on the event ($W_{it} = 0$ and $K_{it}^* = 0$), then:*

$$\mathbb{E}[h_{0it} - h_{1it} | h_{it} = k, K_{it}^* = 0] \in \left[\Delta_k^L, \Delta_k^U\right]$$

where

$$\Delta_k^L = g(F_{0|W=0,K^*=0}(k), f_{0|W=0,K^*=0}(k), \mathcal{B}^*) + g(1 - F_{1|W=0,K^*=0}(k), f_{1|W=0,K^*=0}(k), \mathcal{B}^*)$$

and

$$\Delta_k^U = -g(1 - F_{0|W=0,K^*=0}(k), f_{0|W=0,K^*=0}(k), -\mathcal{B}^*) - g(F_{1|W=0,K^*=0}(k), f_{1|W=0,K^*=0}(k), -\mathcal{B}^*)$$

where $\mathcal{B}^* = P(h_{it} = k | W_{it} = 0, K_{it}^* = 0)$ and

$$g(a, b, x) = \frac{a}{bx}(a + x)\ln\left(1 + \frac{x}{a}\right) - \frac{a}{b}$$

The bounds are sharp.

Proof. Omitted for brevity. □

Theorem 1* does not immediately yield identification of the buncher-LATE bounds Δ_k^L and Δ_k^U, as we need to estimate each of the arguments to the function g. Using that the function g is homogenous of degree one, the bounds can be rewritten in terms of p, the identified quantities $\mathcal{B}$, $P(W_{it} = 0)\lim_{y\uparrow k} f_{0|W=0}(y)$ and $P(W_{it} = 0)\lim_{y\uparrow k} f_{1|W=0}(y)$, as well as the two probabilities $P(h_{it} < 40 \text{ and } W_{it} =)$ and $P(h_{it} > 40 \text{ and } W_{it} = 0)$ (see proof for details).

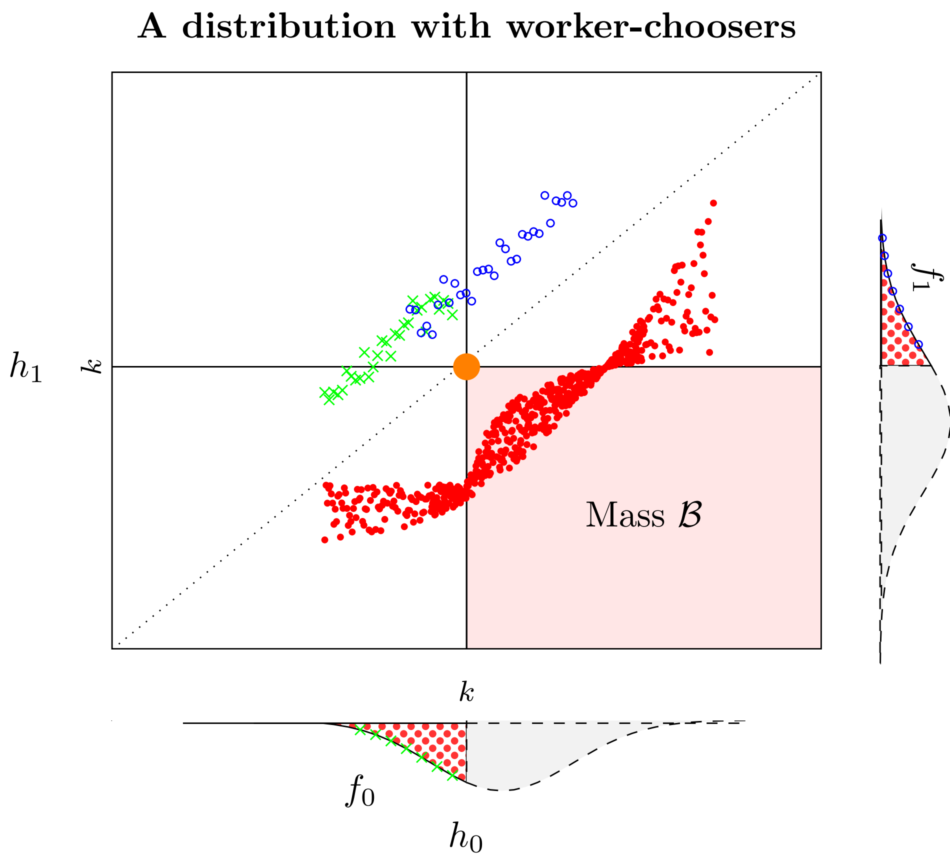

Figure A.3: The joint distribution of (h_{0it}, h_{1it}), for a distribution including worker-choosers and satisfying assumption RANK*, cf. Figure 1.6. See text for description.

Figure A.3 depicts an example of a joint distribution of (h_0, h_1) that includes worker-choosers and satisfies Assumption RANK*. The x-axis is h_0, and the y-axis is h_1, with the solid lines indicating 40 hours and the dotted diagonal line depicting $h_1 = h_0$. The dots show a hypothetical joint-distribution of the potential outcomes, with the (red) cloud south of the 45-degree line being firm-choosers, and the (green and blue) cloud above being worker-choosers. Green x's indicate worker-choosers who choose their value of h_0, while blue circles indicate worker-choosers who choose their value of h_1. The orange dot at $(40, 40)$ represents a mass of counterfactual bunchers.

Observed to the the econometrician is the point mass at 40 as well as the truncated marginal distributions depicted at the bottom and the right of the figure, respectively. The observable $P(h_{it} \leq h)$ for $h < 40$ doesn't exactly identify $P(h_{0it} \leq h)$ because some green x's are missing – these are worker-choosers for whom $h_1 > 40 > h_0$ and choose to work overtime at their h_1 value. Thus they show up in the data at $h > 40$ even thoug they have $h_0 < 40$. Similarly, some blue circles are missing from the data above 40 – these are worker-choosers for whom $h_1 > 40 > h_0$ and choose to work their h_0 value, not working overtime. The probabilities $P(h_{it} < 40$ and $W_{it} =)$ and $P(h_{it} > 40$ and $W_{it} = 0)$ can thus only be estimated with some error, with the size of the error depending on the mass of worker-choosers in the northwest quadrant of Figure A.3. However, this has little impact on the results.[13]

Two further caveats of Theorem 1* are worth mentioning here. First, an evaluation of the FLSA would ideally account for worker-choosers (who are working longer hours as a result of the policy) when averaging treatment effects. However, the proportion of worker-choosers and the size of their hours increases are not identified. Using the buncher LATE to estimate the overall ex-post effect of the FLSA – as described in Section 1.4.4 – may overstate its overall average net hours reduction. Secondly, note that we can no longer directly verify the bi-log concavity assumption of h_0 for $h < k$, and of h_1 for $h > k$, by looking at the data. The reason is that the observed data is a mixture of the firm-chooser and worker-chooser distributions, while our BLC assumption regards the subgroup of firm-choosers. If the proportion of worker-choosers is small, then these caveats should have only a minor impact on the interpretation of the results. The first problem is difficult to avoid: estimating the overall effect of the FLSA based on a subset of firm-choosers is inevitably going to miss the fact that overtime pay increases hours for some workers.

[13]The components of the bounds $\Delta_k^L = L0 + L1$ and $\Delta_k^U = -U0 - U1$ are not sensitive to the values of the CDF inputs $F_{0|W=0,K^*=0}(k)$ and $F_{1|W=0,K^*=0}(k)$, as can be verified numerically (details available upon request). Intuitively, Δ_k^L and Δ_k^U mostly depend on the density estimates and the size of the bunching mass.

A.3 Interdependencies among hours within the firm

In this section I consider the impact that interdependencies among the hours of different units may have on the estimates, reflected in the third term of Equation (1.8) from Section 1.4.4. I develop some structure to guide our intuition of this term, and then present some empirical evidence that it is likely to be small.

The basic issue is as follows: when a single firm chooses hours jointly among mulitple units—either across different workers or across multiple weeks, or both—this term may be nonzero and contribute to the overall effect of the FLSA. This can be thought of as a violation of the stable unit treatment value assumption (SUTVA) in assessing the overall average impact of the FLSA on hours, the effect of which is captured in the third term of Equation (1.8).

To simplify the notation, I'll assume that such SUTVA violations may occur across workers within a firm in a single week, suppressing the time index t and focusing on a single firm. As in Section 1.4.4 let $\mathbf{h_{-i}}$ denote the vector of actual (observed) hours for all workers aside from i within $i's$ firm. These hours are chosen according to the kinked cost schedule introduced by the FLSA. Let $\mathtt{h}_{0i}(\cdot)$ denote the hours that the firm would choose for worker i if they had to pay i' straight-wage w_i for all of i's hours, as a function of the hours profile of the other workers in the firm (suppressing dependence on straight-wages in this section). Define $\mathtt{h}_{1i}(\cdot)$ analogously with $1.5w_i$. In this notation, the potential outcomes from Section 1.4 are $h_{0i} = \mathtt{h}_{0i}(\mathbf{h_{-i}})$ and $h_{1i} = \mathtt{h}_{1i}(\mathbf{h_{-i}})$. As in Section 1.4.4 let $(h_i^*, \mathbf{h^*}_{-i})$ denote the hours profile that would occur absent the FLSA, so that the average ex-post effect of the FLSA is $\mathbb{E}[h_i - h_i^*]$.

For concreteness, we may consider the model introduced in the beginning of Section 1.4 in which hours are chosen to maximize profits with a joint-production function $F(\mathbf{h})$. In this case we have that $(h_i, \mathbf{h}_{-i}) = \text{argmax}\left\{F(\mathbf{h}) - \sum_j B_{kj}(h_j)\right\}$, where the sum is across workers j in the firm and $B_{kj}(h) := w_j h + .5w_j \mathbb{1}(h > 40)(h-40)$. Similarly $(h_i^*, \mathbf{h}_{-i}^*) = \text{argmax}\left\{F(\mathbf{h}) - \sum_j w_j h_j\right\}$ (where for the moment we ignore changes in w_j).

Whether $\mathtt{h}_{0i}(\mathbf{h_{-i}})$ is smaller or larger than h_i^* (with a fixed set of employees) will depend upon whether i's hours are complements or substitutes in production with those of each of their colleagues, and with what strength. It is natural to expect that both cases occur. Consider for example a production function in which workers are divided into groups $\theta_1 \dots \theta_M$ corresponding to different occupations, and:

$$F(\mathbf{h}) = \prod_{m=1}^{M} \left((\sum_{i \in \theta_m} a_i \cdot h_i^{\rho_m})^{1/\rho_m} \right)^{\alpha_m} \tag{A.11}$$

where a_i is an individual productivity parameter for worker i. The hours of workers within an occupation enter as a CES aggregate with substitution parameter ρ_m, which then combine in a Cobb-Douglas form across occupations with exponents α_m. The hours of two workers i and j belonging to different occupations are always complements in production, i.e. $\partial_{h_i} F(\mathbf{h})$ is increasing in h_j. When i and j belong to the same occupation θ_m, it can be shown that worker i and j's hours are substitutes—i.e. $\partial_{h_i} F(\mathbf{h})$ is *decreasing* in h_j—when $\alpha_m \leq \rho_m$.

Thus both substitution and complementarity in hours can plausibly coexist within a firm, and it is difficult to sign theoretically the contribution of interdependencies to θ. Given that occupations or tasks are not observed in the data, it is also difficult to obtain direct evidence with the aid of structural assumptions like Eq. (A.11). I therefore turn to an indirect empirical test of whether these effects are likely to play a significant role in θ.

Figure A.4 shows that in weeks when a worker receives a positive number of sick-pay hours, their individual hours worked for that week decline by about 8 hours on average. Yet I fail to find evidence of a corresponding change in the hours of others in the same firm. This suggests that short term variation in the hours of a worker's colleagues does not tend to translate into contemporaneous changes in their own (for example, if the firm were dividing a fixed number of hours across workers).

Table A.1 shows another piece of evidence: that my overall effect estimates are similar between small, medium, and large firms. If firms were to compensate for overtime hours reductions by "giving" some hours to similar workers who would otherwise be working less than 40, for instance, then we would expect this to play a larger role in firms where there are a large number of substitutable workers–causing a bias that increases with firm size. I cannot reject that my strategy estimates the same parameter value across the three firm size categories, in my preferred specification of estimating p using variation in PTO.

	p=0		p from PTO	
	Bunching	Effect of the kink	Net Bunching	Effect of the kink
Small firms	0.198	[-1.525, -1.455]	0.027	[-0.231, -0.171]
	[0.189, 0.208]	[-1.676, -1.299]	[0.023, 0.031]	[-0.274, -0.139]
Medium firms	0.103	[-1.123, -0.786]	0.030	[-0.337, -0.224]
	[0.095, 0.110]	[-1.237, -0.710]	[0.025, 0.035]	[-0.407, -0.178]
Large firms	0.050	[-0.768, -0.468]	0.024	[-0.371, -0.224]
	[0.047, 0.054]	[-0.861, -0.414]	[0.021, 0.028]	[-0.444, -0.180]

Table A.1: Estimates of the ex-post effect of the kink by firm size. "Small" firms have between 1 and 25 workers in my estimation sample, "Medium" have 26 to 50, and "Large" have more than 50. Note that the estimated net bunching caused by the FLSA is similar across firm sizes (right), despite the raw bunching observed in the data differing by firm size category.

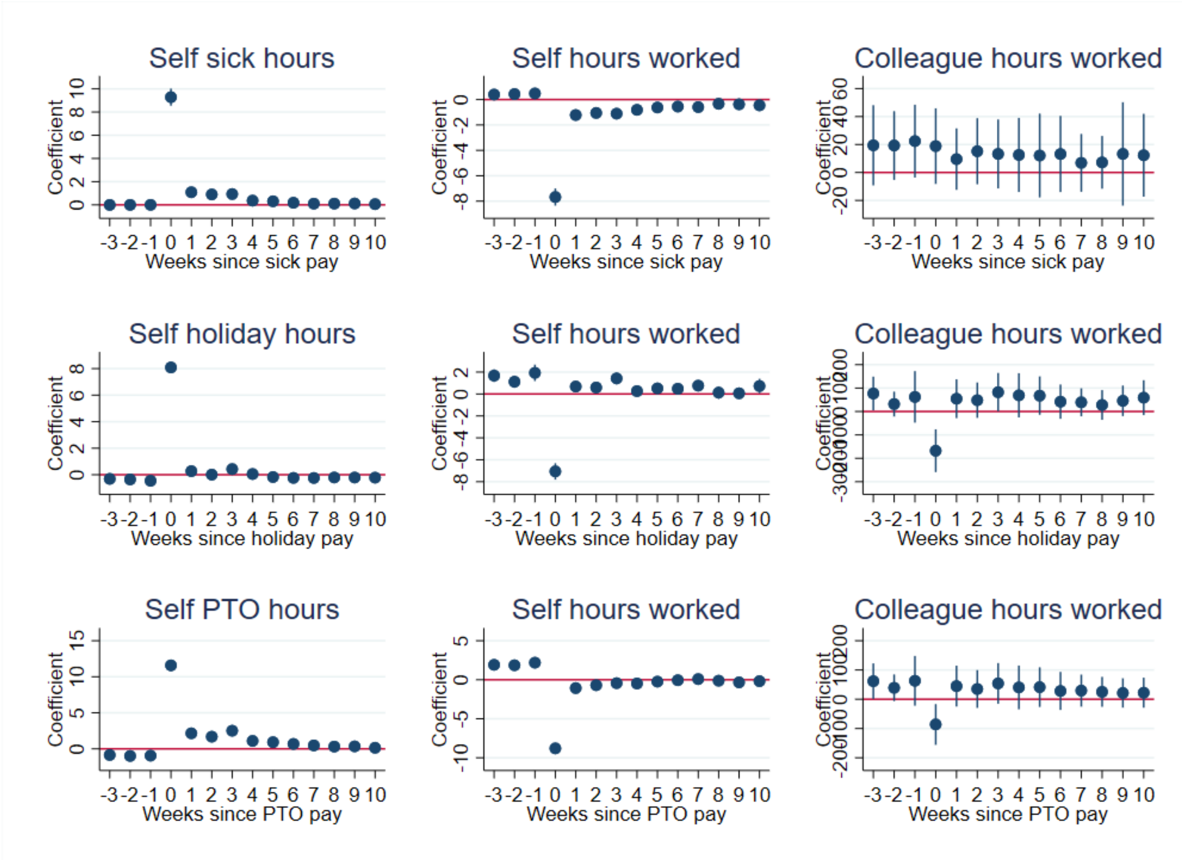

Figure A.4: Event study coefficients β_j and 95% confidence intervals across an instance of a worker receiving pay for non-work hours (either sick pay, holiday pay, or paid time off–'PTO'). Equation is $y_{it} = \mu_t + \lambda_i + \sum_{j=-3}^{10} \beta_j D_{it,j} + u_{it}$, where $D_{it,j} = 1$ if worker i in week t has a positive number of a given type of non-work hours j weeks ago (after a period of at least three weeks in which they did not), λ_i are worker fixed effects, and μ_t are calendar week effects. Rows correspond to choices of the non-work pay type: either sick, holiday, PTO. Columns indicate choices of the outcome y_{it}. "Colleague hours worked" sums the hours of work in t across all workers other than i in i's firm. The timing of holiday and PTO hours appears to be correlated across workers, leading to a decrease in the working hours of i's colleagues in weeks in which i takes either holiday or PTO pay (center-right and bottom-right graphs). However I cannot reject that colleague work hours are unrelated to an instance of sick pay: before, during and after it occurs (top-right). Since i's hours of work reduce by about 8 hours on average during an instance of sick pay (top-center), this suggests that there is no contemporaneous reallocation of i's forgone work hours to their colleagues.

A.4 A simple model of wages and "typical" hours

Each firm faces a labor supply curve $N(z, h)$, indicating the labor force N it can maintain if it offers total compensation z to each of its workers, when they are each expected to work h hours per week. The firm chooses a pair (z^*, h^*) based on the cost-minimization problem:

$$\min_{z,h,K,N} N \cdot (z + \psi) + rK \text{ s.t. } F(Ne(h), K) \geq Q \text{ and } N \leq N(z, h) \tag{A.12}$$

where the labor supply function is increasing in z while decreasing in h, $e(h)$ represents the "effective labor" from a single worker working h hours, and ψ represents non-wage costs per worker. The quantity ψ can include for example recruitment effort and training costs, administrative overhead and benefits that do not depend on h. Concavity of $e(h)$ captures declining productivity at longer hours, for example from fatigue or morale effects. The function F maps total effective labor $Ne(h)$ and capital into level of output or revenue that is required to meet a target Q, and r is the cost of capital. For simplicity, workers within a firm are here identical and all covered by the FLSA.

To understand the properties of the solution to Equation (A.12), let us examine two illustrative special cases.

Special case 1: an exogenous competitive straight-time wage

Much of the literature on hours determination has taken the hourly wage as a fixed input to the choice of hours, and assumed that at that wage the firm can hire any number of workers, regardless of hours. This can be motivated as a special case of Equation (A.12) in which there is perfect competition on the straight-time wage, i.e. $N(z, h) = \bar{N}\mathbb{1}(w_s(z, h) \geq w)$ for some large number $\bar{N}$ and wage w exogenous to the firm. Then

Equation (A.12) reduces to:

$$\min_{N,h,K} N \cdot (hw + \mathbb{1}(h > 40)(w/2)(h - 40) + \psi) + rK \text{ s.t. } F(Ne(h), K) \geq Q \tag{A.13}$$

By limiting the scope of labor supply effects in the firm's decision, Equation (A.13) is well-suited to illustrating the competing forces that shape hours choice on the production side: namely the fixed costs ψ and the concavity of $e(h)$. Were ψ equal to zero with $e(h)$ strictly concave globally, a firm solving Equation (A.13) would always find it cheaper to produce a given level of output with more workers working less hours each. On the other hand, were ψ positive and e weakly convex, it would always be cheapest to hire a single worker to work all of the firm's hours. In general, fixed costs and declining hours productivity introduce a tradeoff that leads to an interior solution for hours.[14]

Equation (A.13) introduces a kink into the firm's costs as a function of hours, much as short-run wage rigidity does in my dynamic analysis. However, the assumption that the firm can demand any number of hours at a set straight-time wage rate is harder to defend when thinking about firms long-run expectations, a point emphasized by Lewis (1969). Equilibrium considerations will also tend to run against the independence of hourly wages and hours - a mechanism explored in Supplemental Appendix A.6.

Special case 2: iso-elastic functional forms

By placing some functional form restrictions on Equation (A.12), we can obtain a closed-form expression for (z^*, h^*). In particular, when labor supply and $e(h)$ are iso-elastic, production is separable between capital and labor and linear in the latter, and firms set the output target Q to maximize profits, Proposition A.5 characterizes the firm's choice of earnings and hours:

[14] In the fixed-wage special case, these two forces along with the wage are in fact sufficient to pin down hours, which do not depend on the production function F or the chosen output level Q. See e.g. Cahuc and Zylberberg (2014) for the case in which $e(h)$ is iso-elastic.

Proposition A.5. *When i) $e(h) = e_0 h^\eta$ and $N(z,h) = N_0 z^{\beta_z} h^{\beta_h}$; ii)$F(L,K) = L + \phi(K)$ for some function ϕ; and iii) Q is chosen to maximize profits, the (z^*, h^*) that solve Equation (A.12) are:*

$$h^* = \left[\frac{\psi}{e_0} \cdot \frac{\beta}{\beta - \eta}\right]^{1/\eta} \qquad \textit{and} \qquad z^* = \psi \cdot \frac{\beta_z}{\beta_z + 1} \frac{\eta}{\beta - \eta}$$

where $\beta := \frac{|\beta_h|}{\beta_z + 1}$, provided that $\psi > 0$, $\eta \in (0, \beta)$, $\beta_h < 0$ and $\beta_z > 0$. Hours and compensation are both decreasing in $|\beta_h|$ and increasing in β_z.

Proof. Omitted for brevity. □

The proposition shows that the hours chosen depend on labor supply via $\beta = \frac{|\beta_h|}{1+\beta_z}$, which gages how elastic labor supply is with respect to hours compared with earnings. The more sensitive labor supply is to a marginal increase in hours as compared with compensation, the higher β will be and lower the optimal number of hours. The proof of Proposition A.5 also shows that unlike Special case 1 of perfect competition on the straight-time wage, when $N(z,h)$ is differentiable the general model can support an interior solution for hours even without fixed costs $\psi = 0$.

Note: Broadly speaking, the function $N(z,h)$ might be viewed as an equilibrium object that reflects both worker preferences over income and leisure and the competitive environment for labor. Thus it is conceivable that equilibrium forces lead to a labor supply function like that of the fixed-wage model, in which the the FLSA has an effect on the hours set at hiring. In Supplemental Appendix A.6, I show that the prediction of the fixed-job model that the FLSA has litte to no effect on h^* or z^* is robust to embedding Equation (A.12) into an extension of the Burdett and Mortensen (1998) model of equilibrium with on-the-job search.[15] In the context of the search model, the only effect of the overtime rule on the distribution of h^* is mediated through the minimum wage, which rules out some of the (z^*, h^*) pairs that would occur in the unregulated equilibrium. In a numerical cal-

[15]This remains true even in the perfectly competitive limit of the model, the basic reason being that workers choose to accept jobs on the basis of their known total earnings z^*, rather than the straight-time wage.

ibration, this effect is quite small, suggesting that equilibrium effects play only a minor role in how the FLSA overtime rule impacts anticipated hours or straight-time wages.

A.5 Additional empirical results

A.5.1 A test of the Trejo (1991) model of straight-time wage adjustment

Another way to assess the role of wage rigidity is to test directly whether straight-time wages and hours are plausibly related according to Equation (1.1). To do this by supposing that some proportion of all paychecks reflect a wage that is determined from the worker's total earnings z_{it} according to Equation (1.1), while the others have wages set in some other way. We indicate those paychecks for which the wage is actively adjusted to this period's hours as $A_{it} = 1$, and let $q(h) = P(A_{it} = 1|h_{it} = h)$. This nests an extreme version of the fixed-job model of Trejo (1991), in which $q(h) = 1$ for all h.

By the law of iterated expectations and some algebra we have that:

$$\mathbb{E}\left[\ln w_{it}|h_{it} = h\right] = q(h)\left\{\mathbb{E}\left[\ln(w_{it})|h_{it} = h, A_{it} = 0\right] - \ln\left(h + 0.5(h-40)\mathbb{1}(h \geq 40)\right)\right\}$$
$$- (1 - q(h))\mathbb{E}\left[\ln w_{it}|h_{it} = h, A_{it} = 1\right]$$

The second term above introduces a kink in the conditional expectation of log wages with respect to hours. If $\mathbb{E}\left[\ln z_{it}|h_{it} = h, A_{it} = 0\right]$, $\mathbb{E}\left[\ln w_{it}|h_{it} = h, A_{it} = 1\right]$ and $q(h)$ are all continuously differentiable in h, then the magnitude of this kink identifies $q(40)$, the proportion of active wage responders local to $h = 40$:[16]

$$\lim_{h \downarrow 40} \frac{d}{dh}\mathbb{E}\left[\ln w_{it}|h_{it} = h\right] - \lim_{h \uparrow 40} \frac{d}{dh}\mathbb{E}\left[\ln w_{it}|h_{it} = h\right] = -\frac{1}{2} \cdot \frac{q(40)}{40}$$

Figure A.5 reports the results of fitting separate local linear functions to the CEF of log

[16] These continuous differentiability assumptions are reasonable, if wage setting according to Equation (1.1) is the only force introducing non-smoothness in the relationship between wages and hours.

wages on either side of $h = 40$. We can reject the hypothesis that the fixed-job model applies to all employees at all times. However, the data appear to be consistent with a proportion $q(40)$ of about 0.25 of all paychecks close to 40 hours reflecting an hours/wage relationship according to Equation (1.1). This is consistent with straight wages being updated intermittently to reflect expected or anticipated hours, which vary in practice between pay periods.

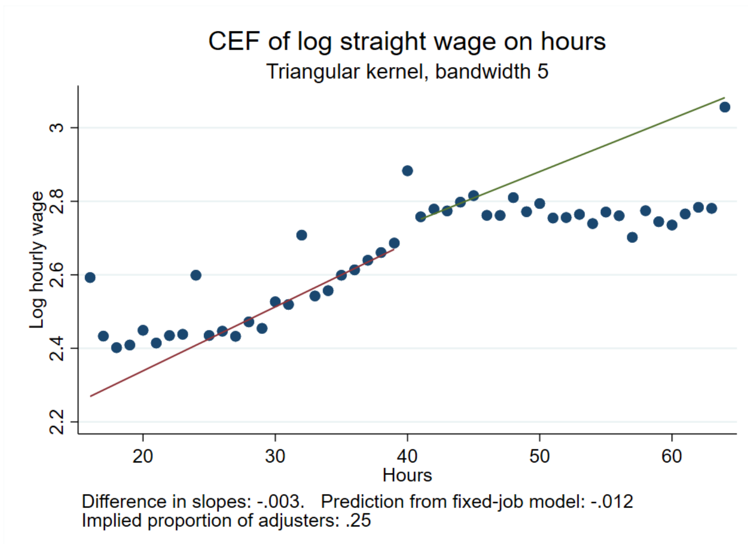

Figure A.5: A kinked-CEF test of the fixed-jobs model presented in Trejo (1991). Regression lines fit on each side with a uniform kernel within 25 hours of the 40.

A.5.2 Further characteristics of the sample

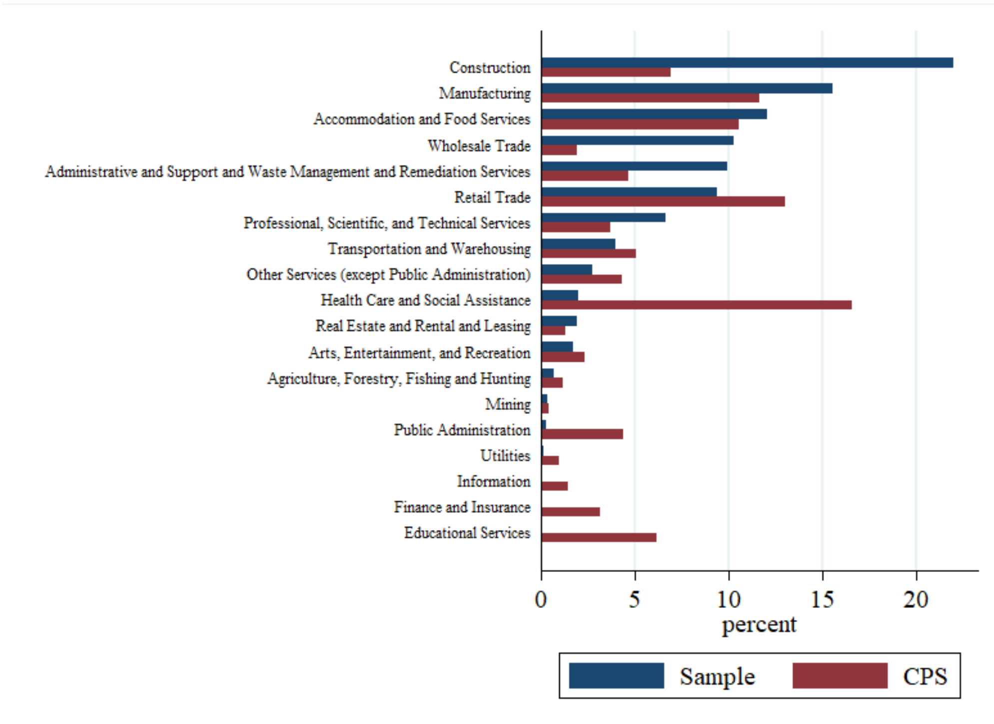

Figure A.6: Industry distribution of estimation sample versus the Current Population Survey sample described in Section 1.3.

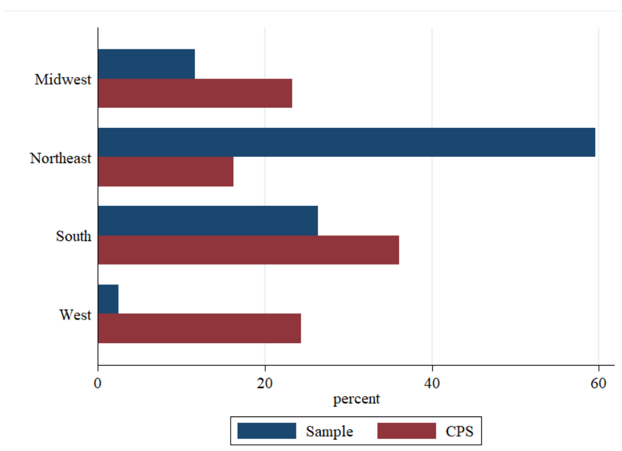

Figure A.7: Geographical distribution of estimation sample versus the Current Population Survey sample described in Section 1.3.

Industry	Avg. OT hours	OT % hours	OT % pay	Industry share
Accommodation and Food Services	2.37	0.06	0.11	0.08
Administrative and Support	5.69	0.13	0.18	0.08
Agriculture, Forestry, Fishing and Hunting	3.76	0.11	0.15	0.00
Arts, Entertainment, and Recreation	3.87	0.10	0.13	0.00
Construction	3.09	0.07	0.10	0.20
Educational Services	1.83	0.05	0.07	0.00
Finance and Insurance	0.31	0.00	0.01	0.00
Health Care and Social Assistance	4.59	0.12	0.12	0.02
Information	1.67	0.04	0.06	0.00
Manufacturing	3.37	0.08	0.11	0.18
Mining	2.26	0.07	0.12	0.00
Other Services	2.61	0.06	0.09	0.02
Professional, Scientific, and Technical Services	2.91	0.07	0.10	0.06
Public Administration	2.36	0.05	0.08	0.00
Real Estate and Rental and Leasing	2.85	0.07	0.09	0.02
Retail Trade	2.83	0.07	0.10	0.08
Transportation and Warehousing	5.24	0.12	0.17	0.04
Utilities	3.80	0.08	0.11	0.00
Wholesale Trade	5.15	0.11	0.14	0.10
Total Sample	3.55	0.08	0.12	0.98

Table A.2: Overtime prevalence by industry in the sample, including average number of OT hours per weekly paycheck, % OT hours among hours worked, % pay for hours work going to OT, and industry share of total hours in sample.

	(1)	(2)	(3)	(4)	(5)
	Work hours=40	OT hours	Total work hours	Work hours=40	OT hours
Tenure	0.000400	0.0515	0.0796		
	(0.95)	(3.95)	(3.31)		
Age	0.000690	0.00266	0.0250		
	(3.82)	(0.74)	(3.25)		
Female	0.0140	-1.322	-1.943		
	(2.08)	(-9.07)	(-6.08)		
Minimum wage worker	0.00121	-1.687	-5.352		
	(0.29)	(-2.39)	(-4.08)		
Firm just hired				-0.00572	0.553
				(-2.95)	(5.78)
Date FE	Yes	Yes	Yes	Yes	Yes
Employer FE	Yes	Yes	Yes		
Worker FE				Yes	Yes
Observations	499619	499619	499619	628449	628449
R squared	0.229	0.264	0.260	0.387	0.515

t statistics in parentheses

Table A.3: Columns (1)-(3) regress hours-related outcome variables on worker characteristics, with fixed effects for the date and employer. Standard errors clustered by firm. Columns (4)-(5) show that bunching and overtime hours among incumbent workers are both responsive to new workers being hired within a firm, even controlling for worker and day fixed effects. "Firm just hired" indicates that at least one new worker appears in payroll at the firm this week, and the new workers are dropped from the regression. "Minimum wage worker" indicates that the worker's straight-time wage is at or below the maximum minimum wage in their state of residence for the quarter. Tenure and age are measured in years, and age is missing for some workers.

	(1)	(2)	(3)
	Total work hours	Total work hours	Total work hours
R squared	0.366	0.499	0.626
Date FE		Yes	
Worker FE		Yes	Yes
Employer x date FE	Yes		Yes
Observations	621011	628449	620854

t statistics in parentheses

Table A.4: Decomposing variation in total hours. Worker fixed effects and employer by day fixed effects explain about 63% of the variation in total hours.

A.5.3 Additional treatment effect estimates and figures

	p=0		p from PTO	
	Bunching	Buncher LATE	Net Bunching	Buncher LATE
Accommodation and Food Services	0.036	[0.937, 0.988]	0.036	[0.937, 0.988]
(N=69427)	[0.029, 0.044]	[0.734, 1.212]	[0.029, 0.044]	[0.734, 1.212]
Administrative and Support	0.062	[1.625, 1.771]	0.009	[0.251, 0.255]
(N=49829)	[0.051, 0.074]	[1.313, 2.136]	[0.005, 0.013]	[0.143, 0.365]
Construction	0.139	[2.759, 3.326]	0.029	[0.612, 0.638]
(N=136815)	[0.128, 0.149]	[2.341, 3.854]	[0.022, 0.035]	[0.442, 0.821]
Health Care and Social Assistance	0.051	[1.412, 1.522]	0.005	[0.146, 0.147]
(N=13951)	[0.034, 0.069]	[0.570, 2.450]	[0.000, 0.010]	[-0.052, 0.348]
Manufacturing	0.137	[2.098, 2.521]	0.018	[0.307, 0.316]
(N=112555)	[0.126, 0.148]	[1.894, 2.785]	[0.016, 0.021]	[0.255, 0.370]
Other Services	0.160	[1.804, 2.240]	0.037	[0.452, 0.478]
(N=19263)	[0.132, 0.188]	[1.243, 2.996]	[0.024, 0.049]	[0.256, 0.693]
Professional, Scientific, Technical	0.136	[2.281, 2.737]	0.010	[0.178, 0.180]
(N=47705)	[0.117, 0.155]	[1.862, 3.297]	[0.003, 0.016]	[0.060, 0.302]
Real Estate and Rental and Leasing	0.187	[3.477, 4.478]	0.097	[1.920, 2.215]
(N=13498)	[0.141, 0.234]	[2.432, 6.053]	[0.060, 0.135]	[1.065, 3.316]
Retail Trade	0.129	[3.694, 4.399]	0.032	[0.969, 1.016]
(N=56403)	[0.112, 0.146]	[2.447, 5.935]	[0.024, 0.040]	[0.550, 1.463]
Transportation and Warehousing	0.091	[2.230, 2.530]	0.015	[0.400, 0.409]
(N=25926)	[0.070, 0.111]	[1.754, 3.127]	[0.009, 0.022]	[0.216, 0.602]
Wholesale Trade	0.126	[2.751, 3.299]	0.046	[1.068, 1.149]
(N=66678)	[0.110, 0.141]	[2.321, 3.848]	[0.037, 0.055]	[0.765, 1.490]
All Industries	0.116	[2.614, 3.054]	0.027	[0.640, 0.666]
(N=630217)	[0.112, 0.121]	[2.483, 3.217]	[0.024, 0.029]	[0.571, 0.740]

Table A.5: Estimates of the buncher LATE by industry, based on $p = 0$ (left) or p estimated from paid time off (right). Estimates are reported only for industries having at least 10,000 observations. 95% bootstrap confidence intervals in gray, clustered by firm.

	p=0		p from PTO	
	Bunching	Effect of the kink	Net Bunching	Effect of the kink
Accommodation and Food Services	0.036	[-0.368, -0.248]	0.036	[-0.368, -0.248]
(N=69427)	[0.029, 0.044]	[-0.450, -0.192]	[0.029, 0.044]	[-0.450, -0.192]
Administrative and Support	0.062	[-1.190, -0.681]	0.009	[-0.178, -0.101]
(N=49829)	[0.051, 0.074]	[-1.424, -0.548]	[0.005, 0.013]	[-0.256, -0.057]
Construction	0.139	[-1.550, -1.121]	0.029	[-0.330, -0.219]
(N=136815)	[0.128, 0.149]	[-1.771, -0.944]	[0.022, 0.035]	[-0.422, -0.157]
Health Care and Social Assistance	0.051	[-0.633, -0.320]	0.005	[-0.065, -0.030]
(N=13951)	[0.034, 0.069]	[-1.020, -0.129]	[0.000, 0.010]	[-0.155, 0.012]
Manufacturing	0.137	[-1.167, -0.850]	0.018	[-0.162, -0.110]
(N=112555)	[0.126, 0.148]	[-1.282, -0.766]	[0.016, 0.021]	[-0.192, -0.090]
Other Services	0.160	[-0.977, -0.811]	0.037	[-0.235, -0.176]
(N=19263)	[0.132, 0.188]	[-1.300, -0.538]	[0.024, 0.049]	[-0.345, -0.095]
Professional, Scientific, Technical	0.136	[-1.192, -0.959]	0.010	[-0.090, -0.063]
(N=47705)	[0.117, 0.155]	[-1.411, -0.767]	[0.003, 0.016]	[-0.150, -0.021]
Real Estate and Rental and Leasing	0.187	[-1.766, -1.466]	0.097	[-0.954, -0.725]
(N=13498)	[0.141, 0.234]	[-2.303, -1.002]	[0.060, 0.135]	[-1.378, -0.392]
Retail Trade	0.129	[-1.685, -1.342]	0.032	[-0.434, -0.308]
(N=56403)	[0.112, 0.146]	[-2.274, -0.908]	[0.024, 0.040]	[-0.626, -0.175]
Transportation and Warehousing	0.091	[-1.590, -0.998]	0.015	[-0.274, -0.166]
(N=25926)	[0.070, 0.111]	[-1.935, -0.783]	[0.009, 0.022]	[-0.406, -0.086]
Wholesale Trade	0.126	[-2.122, -1.297]	0.046	[-0.776, -0.476]
(N=66678)	[0.110, 0.141]	[-2.474, -1.088]	[0.037, 0.055]	[-1.016, -0.333]
All Industries	0.116	[-1.466, -1.026]	0.027	[-0.347, -0.227]
(N=630217)	[0.112, 0.121]	[-1.542, -0.972]	[0.024, 0.029]	[-0.386, -0.202]

Table A.6: Estimates of the hours effect of the FLSA by industry, based on $p = 0$ (left) or p estimated from paid time off (right). Estimates are reported only for industries having at least 10,000 observations. 95% bootstrap confidence intervals in gray, clustered by firm. In the case of Accommodation and Food Services, $P(h_{it} = 40|\eta_{it} > 0) > \mathcal{B}$, so I take the PTO-based estimate to be $p = 0$.

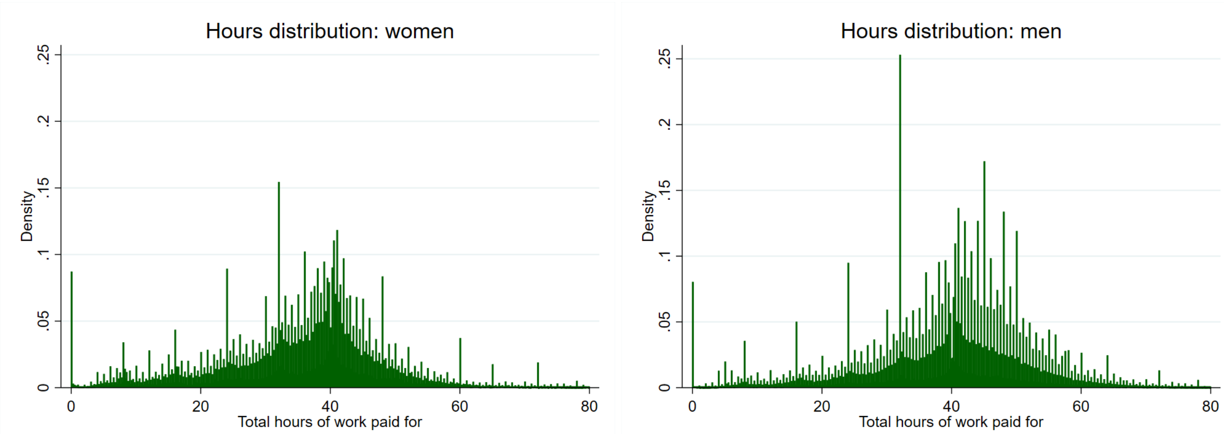

Table A.7: Hours distribution by gender, conditional on different than 40 for visibility (size of point mass at 40 can be read from Figures A.8 and A.9).

	p=0	p from non-changers	p from PTO
Net bunching:	0.090	0.044	0.011
	[0.083, 0.098]	[0.041, 0.048]	[0.009, 0.012]
Buncher LATE	[1.507, 1.709]	[0.763, 0.814]	[0.187, 0.190]
	[1.387, 1.855]	[0.706, 0.877]	[0.150, 0.227]
Buncher LATE as elasticity	[0.093, 0.105]	[0.047, 0.050]	[0.012, 0.012]
	[0.086, 0.114]	[0.044, 0.054]	[0.009, 0.014]
Average effect of kink on hours	[-0.633, -0.489]	[-0.319, -0.231]	[-0.078, -0.054]
	[-0.688, -0.446]	[-0.343, -0.213]	[-0.094, -0.043]
Num observations	147953	147953	147953
Num clusters	352	352	352

Table A.8: Hours distribution and results of the bunching estimator among women.

	p=0	p from non-changers	p from PTO
Net bunching:	0.124	0.060	0.031
	[0.119, 0.129]	[0.058, 0.063]	[0.028, 0.034]
Buncher LATE	[3.074, 3.635]	[1.560, 1.701]	[0.828, 0.868]
	[2.777, 3.991]	[1.407, 1.869]	[0.717, 0.986]
Buncher LATE as elasticity	[0.190, 0.224]	[0.096, 0.105]	[0.051, 0.053]
	[0.171, 0.246]	[0.087, 0.115]	[0.044, 0.061]
Average effect of kink on hours	[-1.867, -1.271]	[-0.921, -0.604]	[-0.482, -0.311]
	[-2.060, -1.149]	[-1.015, -0.545]	[-0.549, -0.269]
Num observations	482264	482264	482264
Num clusters	524	524	524

Table A.9: Hours distribution and results of the bunching estimator among men.

	p=0	p from non-changers	p from PTO
Net bunching:	0.114	0.055	0.027
	[0.109, 0.118]	[0.054, 0.057]	[0.024, 0.029]
Treatment effect			
Linear interpolation	2.621	1.276	0.614
	[2.418, 2.825]	[1.178, 1.374]	[0.541, 0.686]
Monotonicity bounds	[2.320, 3.014]	[1.129, 1.467]	[0.543, 0.705]
	[2.140, 3.201]	[1.034, 1.550]	[0.485, 0.775]
BLC buncher LATE	[2.463, 2.706]	[1.247, 1.309]	[0.612, 0.627]
	[2.311, 2.876]	[1.171, 1.389]	[0.547, 0.695]
Num observations	643720	643720	643720
Num clusters	567	567	567

Table A.10: Treatment effects in levels with comparison to alternative shape constraints.

	p=0	p from non-changers	p from PTO
Net bunching:	0.114	0.055	0.027
	[0.109, 0.118]	[0.054, 0.057]	[0.024, 0.029]
Treatment effect			
Linear interpolation	0.162	0.079	0.038
	[0.150, 0.175]	[0.073, 0.085]	[0.033, 0.042]
Monotonicity bounds	[0.143, 0.186]	[0.070, 0.090]	[0.033, 0.043]
	[0.132, 0.197]	[0.064, 0.096]	[0.030, 0.048]
BLC buncher LATE	[0.152, 0.167]	[0.077, 0.081]	[0.038, 0.039]
	[0.142, 0.177]	[0.072, 0.086]	[0.034, 0.043]
Num observations	643720	643720	643720
Num clusters	567	567	567

Table A.11: Treatment effects as elasticities with comparison to alternative shape constraints.

	p=0	p from non-changers	p from PTO
Buncher LATE as elasticity	[0.161, 0.188]	[0.082, 0.088]	[0.039, 0.041]
	[0.153, 0.198]	[0.077, 0.093]	[0.035, 0.046]
Average effect of FLSA on hours	[-1.466, -1.329]	[-0.727, -0.629]	[-0.347, -0.294]
	[-1.541, -1.260]	[-0.769, -0.593]	[-0.385, -0.262]
Avg. effect among directly affected	[-2.620, -2.375]	[-1.453, -1.258]	[-0.738, -0.624]
	[-2.743, -2.259]	[-1.532, -1.189]	[-0.814, -0.560]
Double-time, average effect on hours	[-2.604, -0.950]	[-1.239, -0.492]	[-0.580, -0.241]
	[-2.716, -0.904]	[-1.293, -0.464]	[-0.639, -0.215]

Table A.12: Estimates of policy effects (replicating Table 1.3) ignoring the potential effects of changes to straight-time wages.

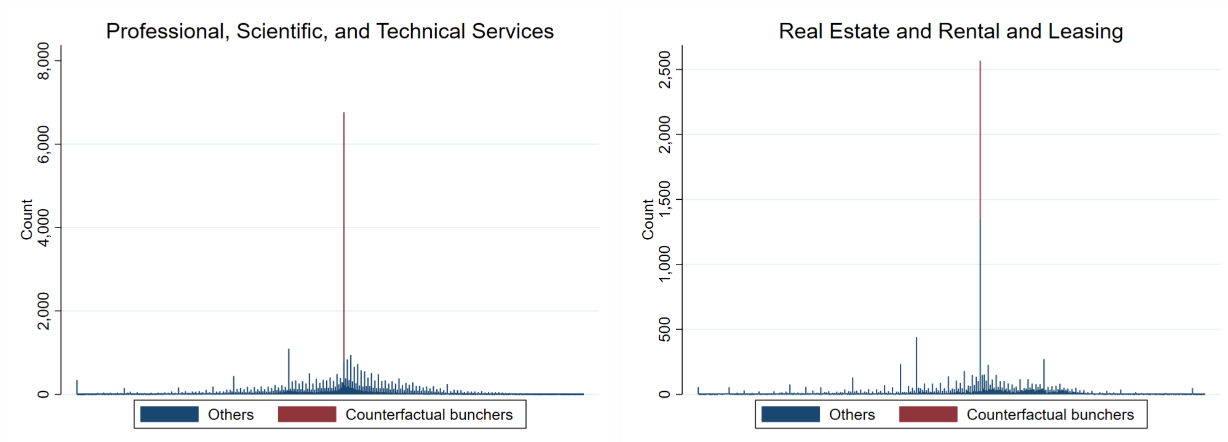

Figure A.8: Hours distribution for an industry with a low treatment effect (left), and a high one (right). Both industries exhibit a comparable amount of raw bunching (14% and 19% respectively, see Table A.6). In Professional, Scientific, and Technical Services, much more of the observable bunching is estimated to be counterfactual bunching, using the PTO-based method. Furthermore, the density of hours is higher just to the right of 40, meaning that the remaining bunching can be explained by a very small responsiveness of hours to the FLSA.

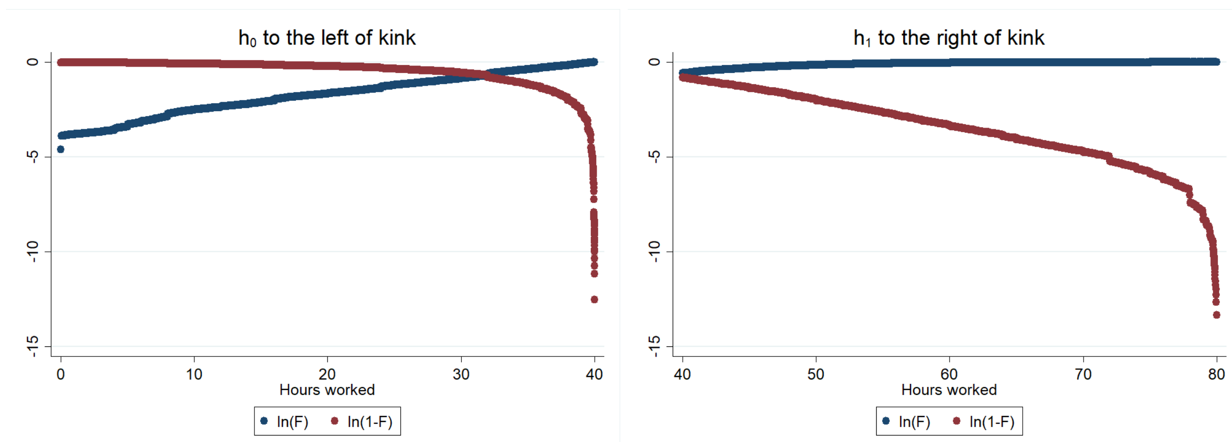

Figure A.9: Validating the assumption of bi-log-concavity away from the kink. The left panel plots estimates of $\ln F_0(h)$ and $\ln(1 - F_0(h))$ for $h < 40$, based on the empirical CDF of observed hours worked. Similarly the right panel plots estimates of $\ln F_1(h)$ and $\ln(1 - F_1(h))$ for $h > k$, where I've conditioned the sample on $Y_i < 80$. Bi-log-concavity requires that the four functions plotted be concave globally.

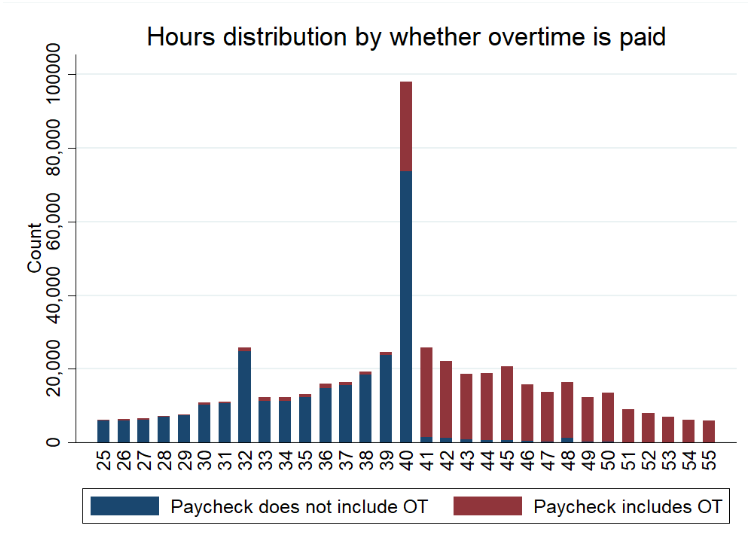

Figure A.10: Histogram of hours worked pooling all paychecks in sample, with one hour bins. Blue mass in the stacks indicate that the paycheck included no overtime pay, while red indicates that the paycheck does include overtime pay.

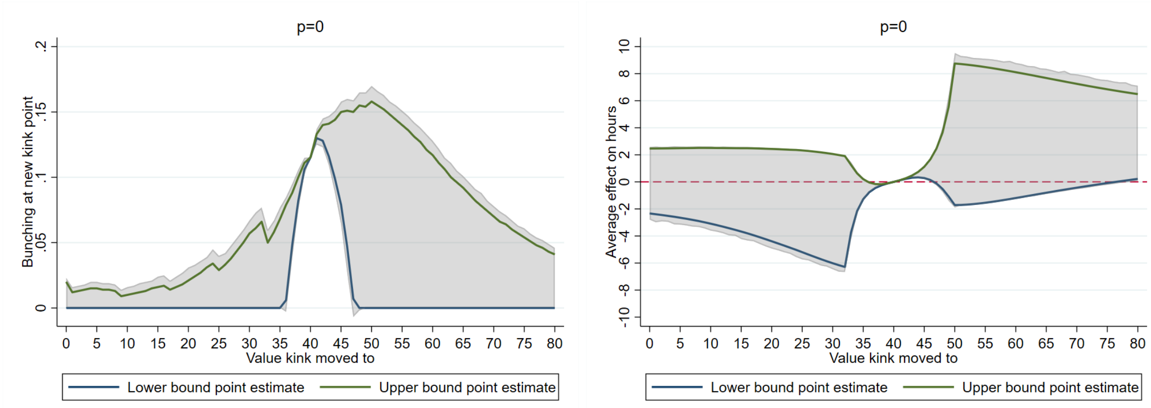

Figure A.11: Estimates of the bunching and average effect on hours were k changed to any value from 0 to 80, assuming $p = 0$. Bounds are not informative far from 40.

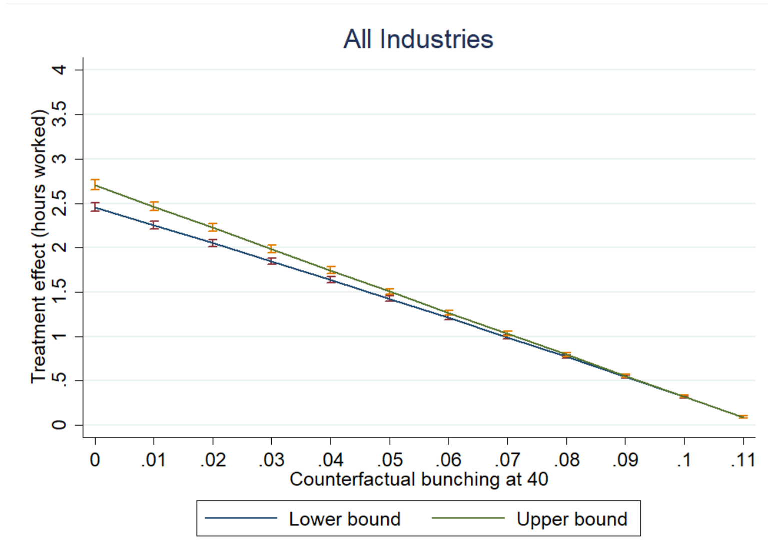

Figure A.12: Treatment effect estimates as a function of assumed counterfactual bunching p at 40, pooling across industries. Confidence intervals depicted here are 95% intervals for each of the bounds separately.

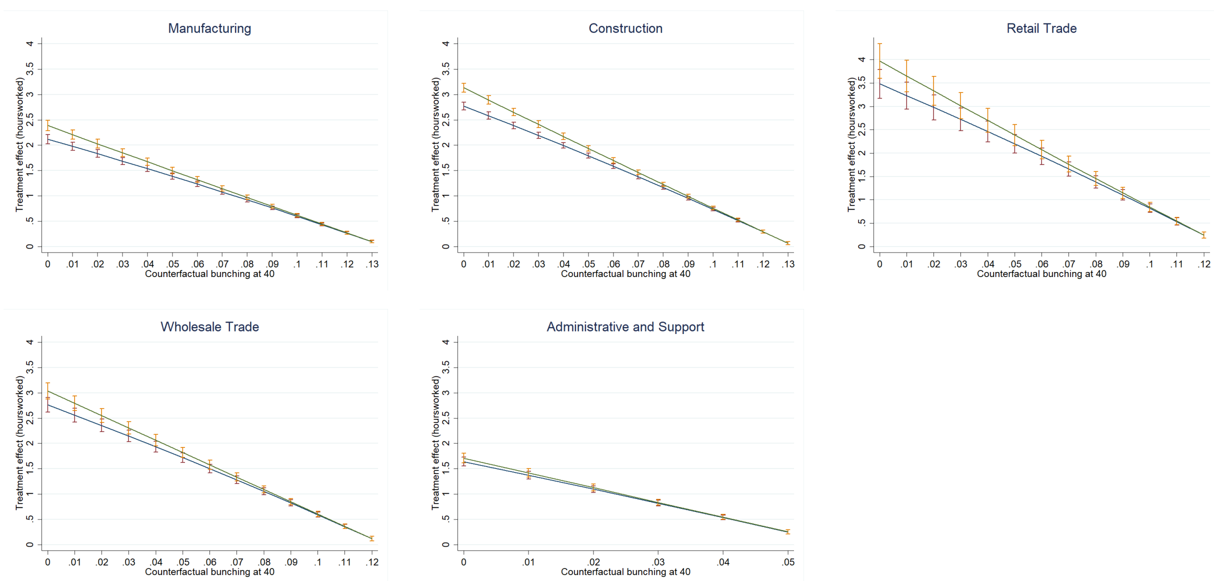

Figure A.13: Treatment effect estimates as a function of p, by each of the largest major industries.

A.5.4 Estimates from the iso-elastic model

This section estimates bounds on ϵ from the iso-elastic model under the assumption that the distribution of $h_{0it} = \eta_{it}^{-\epsilon}$ is bi-log-concave.

If h_{0it} is BLC, bounds on ϵ can be deduced from the fact that

$$F_0(40 \cdot 1.5^{-\epsilon}) = F_0(40) + \mathcal{B} = P(h_{it} \leq 40)$$

where $F_0(h) := P(h_{0it} \leq h)$ and the RHS of the above is observable in the data. $40 \cdot 1.5^{-\epsilon}$ is the location of this "marginal buncher" in the h_0 distribution. In particular,

$$\epsilon = -\ln(Q_0(F_0(40) + \mathcal{B})/40)/(\ln(1.5))$$

where $Q_0 := F_0^{-1}$ is guaranteed to exist by BLC (Dümbgen et al., 2017). In particular:

$$\epsilon \in \left[\frac{\ln\left(1 - \frac{1-F_0(40)}{40f(40)} \ln\left(1 - \frac{\mathcal{B}}{1-F_0(40)}\right)\right)}{-\ln(1.5)}, \frac{\ln\left(1 + \frac{F_0(40)}{40f(40)} \ln\left(1 + \frac{\mathcal{B}}{F_0(40)}\right)\right)}{-\ln(1.5)} \right]$$

where $F_0(k) = \lim_{h \uparrow 40} F(h)$ and $f_0(k) = \lim_{h \uparrow 40} f(h)$ are identified from the data. The bounds on ϵ estimated in this way are $\epsilon \in [-.210, -.167]$ in the full sample.

Since BLC is preserved when the random variable is multiplied by a scalar, BLC of h_{0it} implies BLC of $h_{1it} := \eta_{it}^{-\epsilon} \cdot 1.5^{\epsilon}$ as well. This implication can be checked in the data to the right of 40, since $\eta_{it}^{-\epsilon} \cdot 1.5^{\epsilon}$ is observed there. BLC of h_{1it} implies a second set of bounds on ϵ, because:

$$F_1(40 \cdot 1.5^{\epsilon}) = F_1(40) - \mathcal{B} = P(h_{it} < 40)$$

and the RHS is again observable in the data, where $F_1(h) := P(h_{1it} \leq h)$. Here $40 \cdot 1.5^{\epsilon}$ is the location of a second "marginal buncher" – for which $h_0 = 40$ – in the h_1 distribution. Now we have:

$$\epsilon \in \left[\frac{\ln\left(1 + \frac{F_1(40)}{40f_1(40)} \ln\left(1 - \frac{\mathcal{B}}{F_1(40)}\right)\right)}{\ln(1.5)}, \frac{\ln\left(1 - \frac{1-F_1(40)}{40f_1(40)} \ln\left(1 + \frac{\mathcal{B}}{1-F_1(40)}\right)\right)}{\ln(1.5)} \right]$$

where $F_1(k) = F(k)$ and $f_1(k) := \lim_{h \downarrow 40} f(h)$ are identified from the data. Empirically, these bounds are estimated as $\epsilon \in [-.179, -.141]$. Taking the intersection of these bounds with the range $\epsilon \in [-.210, -.168]$ estimated previously, we have that $\epsilon \in [-.179, -.168]$.[17] The identified set is reduced from a length of $.043$ to $.012$, a factor of nearly 4.

Table A.13 reports estimates broken down by industry, as well as estimates that allow counterfactual bunching at the kink to be estimated from PTO (see Section 3.5).

[17]Note that this interval differs slightly from the identified set of the buncher LATE as elasticity for $p = 0$ in Table 1.3. The latter quantity averages the effect in levels over bunchers and rescales: $\frac{1}{40\ln(1.5)}\mathbb{E}[h_{0it}(1 - 1.5^{\epsilon})|h_{it} = 40]$, but the two are approximately equal under $1.5^{\epsilon} \approx 1 + .5\epsilon$ and $\ln(1.5) \approx .5$.

	p=0		p from PTO	
	Bunching	Elasticity	Net Bunching	Elasticity
Accommodation and Food Services	0.036	[-0.059, -0.060]	0.036	[-0.059, -0.060]
(N=69427)	[0.029, 0.044]	[-0.073, -0.073]	[0.029, 0.044]	[-0.073, -0.073]
Administrative and Support	0.062	[-0.102, -0.106]	0.009	[-0.014, -0.017]
(N=49829)	[0.051, 0.074]	[-0.125, -0.125]	[0.005, 0.013]	[-0.020, -0.020]
Construction	0.139	[-0.190, -0.180]	0.029	[-0.034, -0.043]
(N=136815)	[0.128, 0.149]	[-0.218, -0.218]	[0.022, 0.035]	[-0.043, -0.043]
Health Care and Social Assistance	0.051	[-0.085, -0.095]	0.005	[-0.008, -0.010]
(N=13951)	[0.034, 0.069]	[-0.135, -0.135]	[0.000, 0.010]	[-0.018, -0.018]
Manufacturing	0.137	[-0.158, -0.127]	0.018	[-0.018, -0.020]
(N=112555)	[0.126, 0.148]	[-0.177, -0.177]	[0.016, 0.021]	[-0.022, -0.022]
Other Services	0.160	[-0.120, -0.123]	0.037	[-0.024, -0.033]
(N=19263)	[0.132, 0.188]	[-0.167, -0.167]	[0.024, 0.049]	[-0.034, -0.034]
Professional, Scientific, Technical	0.136	[-0.140, -0.160]	0.010	[-0.009, -0.013]
(N=47705)	[0.117, 0.155]	[-0.175, -0.175]	[0.003, 0.016]	[-0.014, -0.014]
Real Estate and Rental and Leasing	0.187	[-0.250, -0.230]	0.097	[-0.115, -0.133]
(N=13498)	[0.141, 0.234]	[-0.355, -0.355]	[0.060, 0.135]	[-0.177, -0.177]
Retail Trade	0.129	[-0.256, -0.238]	0.032	[-0.055, -0.066]
(N=56403)	[0.112, 0.146]	[-0.359, -0.359]	[0.024, 0.040]	[-0.084, -0.084]
Transportation and Warehousing	0.091	[-0.124, -0.161]	0.015	[-0.019, -0.031]
(N=25926)	[0.070, 0.111]	[-0.167, -0.167]	[0.009, 0.022]	[-0.029, -0.029]
Wholesale Trade	0.126	[-0.212, -0.163]	0.046	[-0.067, -0.068]
(N=66678)	[0.110, 0.141]	[-0.248, -0.248]	[0.037, 0.055]	[-0.088, -0.088]
All Industries	0.116	[-0.179, -0.168]	0.027	[-0.037, -0.043]
(N=630217)	[0.112, 0.121]	[-0.190, -0.190]	[0.024, 0.029]	[-0.041, -0.041]

Table A.13: Estimates of ϵ in the iso-elastic model based on assuming $h_{0it} = \eta_{it}^{-\epsilon}$ is bi-log-concave, by industry. 95% bootstrap confidence intervals in gray brackets, clustered by firm.

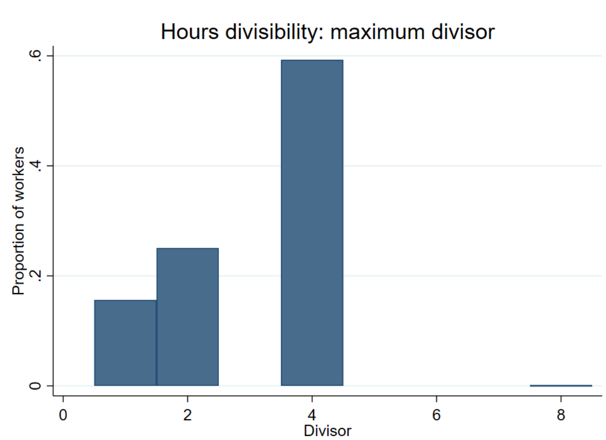

Figure A.14: Distribution of the largest integer $m = 1 \ldots 10$ that maximizes the proportion of worker i's paychecks for which hours are divisible by m. This can be thought of as the granularity of hours reporting for worker i.

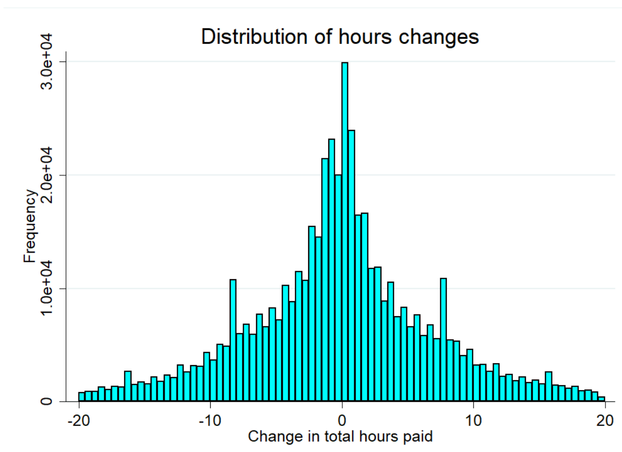

Figure A.15: Distribution of changes in total hours between subsequent pay periods (truncated at -20 and 20).

A.6 An equilibrium search model of hours and wages

A.6.1 The model

I focus on a minimal extension of Burdett and Mortensen (1998) that takes firms to be homogeneous in their technology and workers to be homogeneous in their tastes over the tradeoff between income and working hours. Let there be a large number N_w of workers and large number N_f of firms, and define $m = N_w/N_f$.[18] Formally, we model this as a continuum of workers with mass m, and continuum of firms with unit mass. Firms choose a value of pay z and hours h to apply to all of their workers. Each period, there is an exogenous probability λ that any given worker receives a job offer, drawn uniformly from the set of all firms. Employed workers accept a job offer when they receive an earnings-hours package that they prefer to the one they currently hold, where preferences are captured by a utility function $u(z, h)$ taken to be homogeneous across workers and strictly quasiconcave, where $u_z > 0$ and $u_h < 0$. If a worker is not currently employed, they leave unemployment for a job offer if $u(z, h) \geq u(b, 0)$, where b represents a reservation earnings level required to incent a worker to enter employment. Workers leave the labor market with probability δ each period, and an equal number enters the non-employed labor force.

Before we turn to earnings-hours posting decision of firms, we can already derive several relationships that must hold for the earnings-hours distribution in a steady state equilibrium. First note that the share unemployed v of the workforce must be $v = \frac{\delta}{\delta+\lambda}$, since mass $m(1-v)\delta$ enters unemployment each period, and $m\lambda v$ leaves (we take for granted here that firms only post job offers that are preferred to unemployment, which will indeed be a feature of the actual equilibrium). Let's say that job (z, h) is "inferior" to (z', h') when $u(z, h) \leq u(z', h')$. Let P_{ZH} be the firm-level distribution over offers (Z_j, H_j), and define

$$F(z, h) := P_{ZH}(u(Z_j, H_j) \leq u(z, h)) \tag{A.14}$$

to be the fraction of firms offering inferior job packages to (z, h). The separation rate of workers at a firm choosing (z, h) is thus: $s(z, h) = \delta + \lambda(1 - F(z, h))$. To derive the recruitment of new workers to a given firm each period, we define the related quantity $G(z, h)$ – the fraction of employed workers that are at inferior firms to (z, h). In a steady

[18] Here we largely follow the notation of the presentation of the Burdett & Mortensen model by Manning (2003).

state, note that $G(z,h)$ must satisfy

$$\underbrace{m(1-v)\cdot G(z,h)(\delta+\lambda(1-F(z,h))}_{\text{mass of workers leaving set of inferior firms}} = \underbrace{mv\lambda F(z,h)}_{\text{mass of workers entering set of inferior firms}}$$

since the number of workers at firms inferior to (z,h) is assumed to stay constant. To get the RHS of the above, note that workers only enter the set of firms inferior to (z,h) from unemployment, and not from firms that they prefer. This expression allows us to obtain the recruitment function $R(z,h)$ to a firm offering (z,h). Recruits will come from inferior firms and from unemployment, so that

$$\begin{aligned}
R(z,h) &= \lambda m\left((1-v)G(z,h)+v\right) \\
&= \lambda mv\left(\frac{\lambda F(z,h)}{\delta+\lambda(1-F(z,h))}+1\right) \\
&= m\left(\frac{\delta\lambda}{\delta+\lambda(1-F(z,h))}\right)
\end{aligned}$$

Combining with the separation rate, we can derive the steady-state labor supply function facing each firm:

$$N(z,h) = R(z,h)/s(z,h) = \frac{m\delta\lambda}{\left(\delta+\lambda(1-F(z,h)\right)^2} \tag{A.15}$$

Eq. (A.15) is analogous to the baseline Burdett and Mortensen model, with the quantity $F(z,h)$ playing the role of the firm-level CDF of wages in the baseline model.

Now we turn to how the form of $F(z,h)$ in general equilibrium. We take the profits of firms to be

$$\pi(z,h) = N(z,h)(p(h)-z) = m\delta\lambda\cdot\frac{p(h)-z}{\left(\delta+\lambda(1-F(z,h)\right)^2} \tag{A.16}$$

where the function $p(h)$ corresponds to $e(h)-\psi$, with $e(h)$ being a weakly concave and increasing "effective labor" function with $e(0)=0$, and z recurring non-wage costs per worker. To simplify some of the exposition, we will emphasize the simplest case of $p(h) = p \cdot h$, such that worker hours are perfectly substitutable across workers.

In equilibrium, the identical firms each playing a best response to $F(z, h)$, and thus all choices of (z, h) in the support of P_{ZH} must yield the same level of profits π^*. This gives an expression for $F(z, h)$ over all (z, h) in the support of P_{ZH}, in terms of π^*:

$$F(z, h) = 1 + \frac{\delta}{\lambda} - \sqrt{\frac{m\delta}{\lambda} \cdot \frac{p(h) - z}{\pi^*}} \tag{A.17}$$

where we subtract the positive square root since the negative square root cannot deliver a real number less than or equal to unity for $F(z, h)$. Note that Eq. (A.17) only needs to hold at (z, h) that are actually chosen by firms in equilibrium

It follows from Eqs. (A.17) and (A.15) that we can rank firms in equilibrium by $F(z, h)$ and by size according to the quantity $z - p(h)$. Note that since Eq. (A.15) is continuously differentiable in (z, h), we can rule out mass points in P_{ZH} by an argument paralleling that in Burdett and Mortensen (1998). Suppose $P_{ZH}(z, h) = \delta > 0$ for some (z, h). Then any firm located at (z, h) and earning positive profits could increase their profits further by offering a sufficiently small increase in compensation (or reduction in hours, or a combination of both). Since $F(z + \delta_z, h) = F(z, h) + \delta$ to first order, there exists a small enough δ_z such that $\pi(z + \delta_t, h) > \pi(z, h)$ by Eq. (A.16).

To fully characterize the equilibrium P_{ZH}, we begin by arguing that for a strictly quasiconcave utility function u, workers cannot be indifferent between more than two points that (z, h) share a value of $z - p(h)$. This implies that offers in the support of P_{ZH} lie along a one dimensional path through $\mathbb{R}^2$. Consider for example the case of perfect hours substitutability: $p(h) = ph$, and imagine moving continuously along a line that that keeps $z - ph$ constant from a given point (z, h) in the support of P_{ZH}. Since $F(z, h)$ is constant along this line, we must have from the definition of $F(z, h)$ that either utility is constant or that P_{ZH} has no additional mass along the line. However, we cannot be moving along an indifference curve, as strict convexity of preferences implies that the marginal rate of substitution between compensation and hours can equal p (or more generally $p'(h)$, which is non-increasing) at no more than a single point for a single level of utility. Thus, P_{ZH} puts a positive density on at most one point along each isoquant of $z - p(h)$, and must have positive density on each isoquant within some connected interval. Given this, we can parametrize the points in support of P_{ZH} by a single scalar $t \in [0,1]$, such that $\text{supp}(P_{ZH}) = \{(z(t), h(t))\}_{t\in[0,1]}$ and $t = F(z(t), h(t))$.

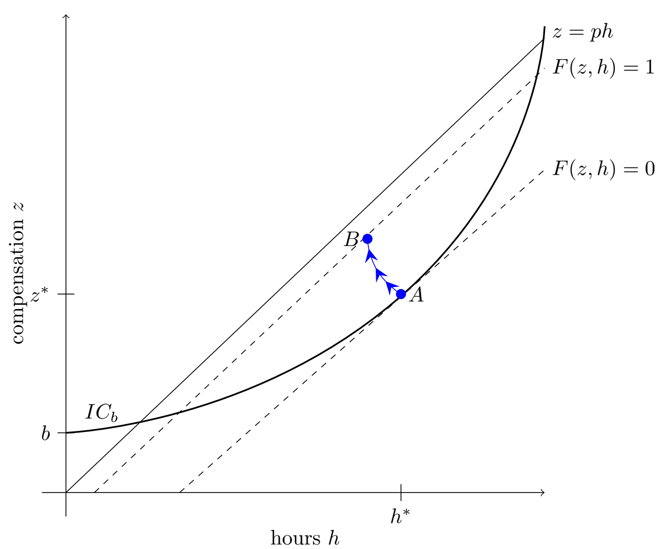

Figure A.16: The support of the equilibrium distribution of compensation-hours offers (z, h) lies along the arrowed (blue) curve AB. Figure shows the case of perfect hours substitutability $p(h) = ph$. Plain curve IC_b is the indifference curve passing through the unemployment point $(b, 0)$. The least desirable firm in the economy lies at the pair (z^*, h^*), labeled by A, where IC_b has a slope of p. The other points chosen by firms are found by beginning at point A and moving in the direction of higher utility, while maintaining a marginal rate of substitution of p between hours and earnings. This path intersects the line of solutions to $F(z, h) = 1$ given Eq. (A.17) at point B. Note that this line still lies below the zero profit line $z = ph$, as firms make positive profit. Curve AB shown for a general non-quasilinear, non-homothetic utility function (see text for details).

Now observe that each $(z(t), h(t))$ must pick out the point along its respective isoquant of $z - p(h)$ which delivers the highest utility to workers, i.e.:

$$(z(t), h(t)) = \text{argmax}_{z,h} u(z, h) \text{ s.t. } z - p(h) = F^{-1}(t)$$

where $F^{-1}(t) = F(z(t), h(t))$, viewed as a function of t. Suppose instead that $u(z(t), h(t)) < \max_{(z,h):z-p(h)=F^{-1}(t)} u(z, h)$. Then any firm located at $(z(t), h(t))$ could profitably deviate to the argmax while keeping profits per worker constant but increasing their labor supply by attracting workers from $(z(t), h(t))$. The first order condition for this problem implies that $(z(t), h(t))$ maintains a marginal rate of substitution of $p'(h(t))$ (p in the baseline case) between compensation and hours at all t, while the slope of the path $(z(t), h(t))$ can be derived from the implicit function theorem:

$$\frac{z'(t)}{h'(t)} = -\left.\frac{u_{hh}(z,h) + p''(h)u_z(z,h) + p'(h)u_{zh}(z,h)}{p'(h)u_{zz}(z,h) + u_{zh}(z,h)}\right|_{(z,h)=(z(t),h(t))}$$

The curve AB shown in Figure A.16 depicts the path $\{(z(t), h(t))\}_{t\in[0,1]}$ for a case in which preferences are neither homothetic nor quasilinear, for example: $u(z, h) = \frac{z^{1-\gamma}}{1-\gamma} - \beta\frac{h^{1+1/\epsilon}}{1+1/\epsilon}$. If preferences were instead homothetic then AB would be a straight line pointing to the north-west from A. This will be the case in the numerical calibration, in which we take preferences to follow the Stone-Geary functional form.[19] If preferences were quasilinear in income (for example the above with $\gamma = 0$), then AB would be a vertical line rising from point A and there would be no hours dispersion in equilibrium.

To pin down the initial point A, we note that it must lie on the indifference curve passing through the unemployment point $(b, 0)$, labeled as IC_b in Figure A.16. If it were to the northwest of the IC_b curve, a firm located there could increase profits by offering a lower value of $z - p(h)$, since given that $F(z(0), h(0)) = 0$ their steady state labor supply already only recruits from unemployment. However, they cannot offer a pair that lies to the southeast of IC_b, since they could never attract workers from unemployment to have positive employment. We assume that the marginal rate of substitution between compensation and hours is less than $p'(0)$ at $(z, h) = (b, 0)$ (such that there are gains from

[19] A CES generalization of Stone-Geary preferences would also result in a straight line AB: $u(z, h) = [\theta(z - \gamma_z)^\lambda + (1-\theta)(\gamma_h - h)\lambda]^{1/\lambda}$. It is also possible to obtain a non-linear path AB while maintaining constant elasticity of substitution between earnings and leisure. The work of Sato (1975) on production functions suggests utility functions satisfying $\frac{u_z(z,h)}{u_h(z,h)} = \left(\frac{z-\gamma_z}{h-\gamma_h}\right)^{\frac{1}{1-\lambda}} \phi(u(c, h))$ where ϕ is any positive function.

trade) and increases continuously with h, eventually passing $p'(h)$ at some point h^*. This point is unique given strict quasiconcavity of $u(\cdot)$. Then, let z^* be the earnings value such that workers are indifferent between (z^*, h^*) and unemployment $(b, 0)$, which represents a reservation level of utility required to enter employment.

Finally, we can also express $F(z,h)$ as a function of $(z^*, h^*) = (z(0), h(0))$ in order to derive an expression for the $F(z,h) = 1$ line, representing the most desired firms in equilibrium. Using that $\pi^* = \pi(z^*, h^*)$, we can rewrite Equation (A.17) as:

$$F(z,h) = \frac{\lambda + \delta}{\lambda} \cdot \left[1 - \sqrt{\frac{p(h) - z}{p(h^*) - z^*}}\right]$$

The firms at point B in Figure A.16 thus solve $z - p(h) = \left(\frac{\delta}{\delta+\lambda}\right)^2 (z^* - p(h^*))$. Equilibrium profits are

$$\pi^* = m(p(h^*) - z^*) \cdot \frac{\lambda/\delta}{(1 + \lambda/\delta)^2}$$

Note that in equilibrium, there exists dispersion not only in both earnings and in hours (provided preferences are not quasi-linear), but also in effective hourly wages, as the ratio $z(t)/h(t)$ is also strictly increasing with t. Note that π^* goes to zero in the limit that $\lambda/\delta \to \infty$. In this limit dispersion over hours, earnings, and hourly earnings all disappear as the line AB collapses to a single point on the zero profit line $z = p(h)$.[20]

A.6.2 Effects of FLSA policies

Now consider the introduction of a minimum wage, which introduces a floor on the hourly wage $w := y/h$. We assume that the point (z^*, h^*) does not satisfy the minimum wage, so that the minimum wage binds and rules out part of the unregulated support of P_{ZH}. The left panel of Figure A.17 depicts the resulting equilibrium, in which the initial

[20] Note that there is no contradiction here as the argument that point A must be on IC_b relies on $F(z(0), h(0)) = 0$, which is implied by no mass points in P_{ZH}, in turn implied by firms making positive profit.

point $(z(0), h(0))$ moves to the point marked A', at which the marginal rate of substitution between compensation and hours is $p'(h)$, but the compensation-hours pair just meets the minimum wage. This compresses the distribution P_{ZH} compared with the unregulated equilibrium from Figure A.16, which now follows a subset of the original path AB. In a stochastic dominance sense, all jobs see a reduction in hours and an increase in total compensation (and hence a compounded effect on hourly wages) when a minimum wage is introduced or increased.

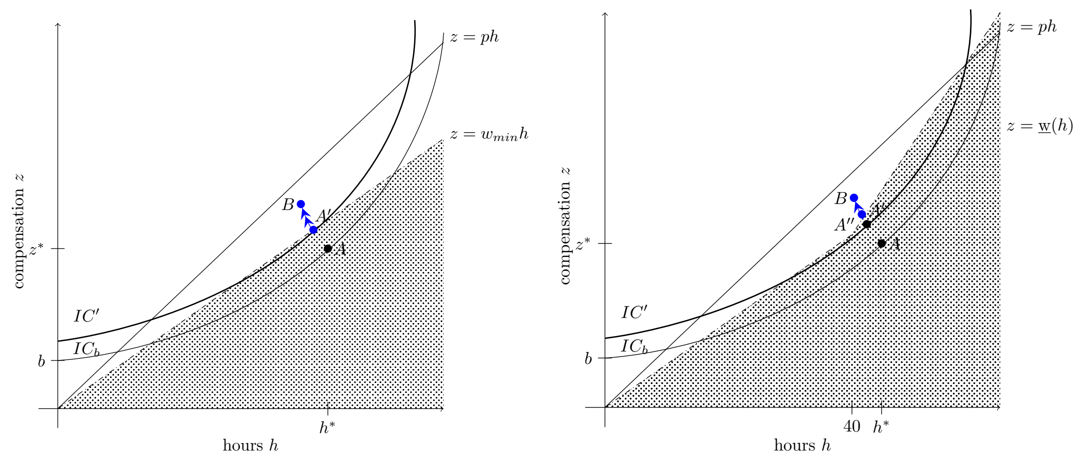

Figure A.17: Left panel shows the support of the equilibrium distribution of compensation-hours offers (z, h) under a binding minimum wage. The compensation hours pairs that do meet $\underline{w}$ are indicated by the shaded region. The lowest-wage offer in the economy moves from point A in the unregulated equilibrium to the point A' on the minimum wage line $z = \underline{w}h$ at which the marginal rate of substitution between compensation and hours equals p. This is equal to the point at which curve AB from Figure A.16 crosses the minimum wage line. Curve $A'B$ traces the remainder of curve AB. The compensation-hours offers are thus more compressed and the new distribution of earnings stochastically dominates the distribution from the unregulated equilibrium, while the opposite is true of hours. Right panel shows how this effect is augmented when overtime premium pay for hours in excess of 40 is required, and A' lies at greater than 40 hours. In this case the support of P_{ZH} begins at point A'', which lies on the kinked minimum wage function $\underline{w}(h)$.

The right panel of Figure A.17 shows how equilibrium is further affected if in addition to a binding minimum wage, premium pay is required at a higher minimum wage $1.5\underline{\mathrm{w}}$ for hours in excess of 40, provided that the point A' lies at an hours value that is greater than 40. In this case, $(z(0), h(0))$ will lie at point A'', at which the marginal rate of substitution between compensation and hours is equal to h', and compensation is equal to the minimum-compensation function under both the minimum wage and overtime policies: $\underline{\mathrm{w}}(h) := \underline{\mathrm{w}}h + 1(h > 40)(h - 40)\underline{\mathrm{w}}/2$.

A.6.3 Calibration

To allow wealth effects in worker utility while facilitating calibration based on existing empirical studies, we assume worker utility is Stone-Geary:

$$u(z, h) = \beta \log(z - \gamma_z) + (1 - \beta)\log(\gamma_h - h)$$

This simple specification allows a closed form solution to the path $(z(t), h(t))$, given by the following Proposition. Using this result, we calibrate the model to consider the effects of FLSA policies on earnings and hours.

Proposition. *Under Stone-Geary preferences and linear production $p(h) = ph - z$, the equilibrium offer distribution is a uniform distribution over $\{(z(t), h(t))\}_{t\in[0,1]}$, where:*

$$\begin{pmatrix} z(t) \\ h(t) \end{pmatrix} = \begin{pmatrix} p\beta\gamma_h + (1-\beta)\gamma_z - \beta z - \beta\left(1 - \frac{t}{1+\frac{\delta}{\lambda}}\right)^2 \cdot (ph(0) - z - z(0)) \\ \beta\gamma_h + \frac{1-\beta}{p}(\gamma_z + z) + \frac{(1-\beta)}{p}\left(1 - \frac{t}{1+\frac{\delta}{\lambda}}\right)^2 \cdot (ph(0) - z - z(0)) \end{pmatrix}$$

The initial point $(z(0), h(0))$ is

1. $h(0) = \gamma_h - \left(\frac{(b-\gamma_c)(1-\beta)}{p\beta}\right)^{\beta}\gamma_h^{1-\beta}$ *and* $z(0) = z^* = \gamma_z + \left(\frac{p\beta\gamma_h}{1-\beta}\right)^{1-\beta}((b-\gamma_c)(1-\beta))^{\beta}$ *in the unregulated equilibrium*

2. $h(0) = (\frac{p\beta}{1-\beta}\gamma_h + \gamma_z)(\underline{w} - \frac{p\beta}{1-\beta})^{-1}$ *and* $z(0) = \underline{w}h(0)$ *with a binding minimum wage of* $\underline{w}$ *(binding in the sense that* $z^* < \underline{w}h^*$*)*

3. $h(0) = (\frac{p\beta}{1-\beta}\gamma_h + \gamma_z + 20\underline{w})(1.5\underline{w} - \frac{p\beta}{1-\beta})^{-1}$ *and* $z(0) = 1.5\underline{w}h(0) - 20\underline{w}$ *with a minimum*

wage of $\underline{w}$ and time-and-a-half overtime pay after 40 hours, if the expression for $h(0)$ in item 2. is greater than 40

Moments with respect to the worker distribution can be evaluated for any measurable function $\phi(z,h)$ as:

$$E_{workers}[\phi(Z_i,H_i)] = \left(1+\frac{\lambda}{\delta}\right)\int_0^1 \phi(z(t),h(t))\cdot\left(1+\frac{\lambda}{\delta}(1-t)\right)^{-2} dt$$

We calibrate the model focusing on a lower-wage labor market where productivity is a constant $p = \$15$. We allow non-wage costs of $z = \$100$ a week, with the value based on estimates of benefit costs in the low-wage labor market.[21] We take $b = \$250$ corresponding to unemployment benefits that can be accrued at zero weekly hours of work.[22]. We calibrate the factor λ/δ using estimates from Manning (2003) using the proportion of recruits from unemployment. Using Manning's estimates from the US in 1990 of about 55% of recruits coming from unemployment, this implies a value of $\lambda/\delta \approx 3$ in the baseline Burdett and Mortensen model.

To calibrate the preference parameters, we first pin down β from estimates of the marginal propensity to reduce earnings after random lottery wins (Imbens et al. 2001; Cesarini et al. 2017). Both of these studies report a value in the neighborhood of $\beta = 0.85$. We take a value of $\gamma_z = \$200$ as the level of consumption at which the marginal willingness to work is infinite, and take $\gamma_h = 50$ hours of work per week. We choose this value according to a rule-of-thumb as the average hours among full-time workers in the US (42.5), divided by β.[23] The value of γ_h plays a central role in setting the location of the hours distribution that we focus on. Again, the model should be interpreted as for a

[21]Specifically, I take a benefit cost of \$2.43 per hour worked for the 10th percentile of wages in 2019: BLS ECEC, multiplied by the average weekly hours worked of 42.5 from the 2018 CPS summary, which results in $102.425 \approx 100$.

[22]We use the UI replacement rate for single adults 2 months after unemployment from the OECD. Taking this for individuals at 2/3 of average income (the lowest available in this table), and then use a BLS figure for average income at the 10% percentile of $22,880$, we have $b \approx \$22,880 \cdot 0.6/52.25 = \$263$

[23] Cesarini et al. (2017) point out that when γ_c and no-unearned income, optimal hours choice is $\beta\gamma_h$. By comparison, these authors calibrate γ_h to be about 35 hours in the Swedish labor market.

specific homogeneous labor market, which we take here to be full-time low wage workers in the US. We ignore taxation in the calibration.

Given these values, we can compute moments of functions of the joint distribution of compensation and hours using the Proposition and numerical evaluation of the integrals. Table A.14 reports worker-level means of hours, weekly compensation, and the hourly wage z/h, as well as employment and profits per worker averaged across the firm distribution. In the unregulated equilibrium, the lowest-compensated workers work about 49 hours a week earning about \$$300$, while the highest-compensated workers work about 46 hours and earn more than \$$550$. This equates to a more than doubling of the hourly wage, which is about \$$6$ for the $t = 0$ workers and over \$$12$ for the $t = 1$ workers. For each of the first three variables, the mean is much closer to the $t = 1$ value than the $t = 0$ value, which follows from the higher-t firms having more employees. The convexity of the labor supply function across values of t is apparent from the firm size row: the largest firm is about 16 times as large as the smallest, while the average firm size is four times larger than the $t = 0$ firms. The final row reports weekly profits per worker: the average worker captures more than half of the employer surplus for the $t = 0$ worker, whose weekly compensation is comparable to the employer's profit for that worker.

	Unregulated equilibrium			$\underline{w} = 7.25$	$\underline{w} = 7.25$ & *OT*	$\underline{w} = 12$ & *OT*
	t=0	t=1	mean	mean	mean	mean
weekly hours	48.85	45.71	46.34	46.18	46.11	45.51
weekly earnings	297.36	564.68	511.22	524.31	530.93	581.78
hourly wage	6.09	12.35	11.06	11.37	11.53	12.78
firm size / smallest	1.00	16.00	4.00	4.00	4.00	4.00
weekly profit per worker	335.46	20.97	146.76	119.81	106.18	1.49

Table A.14: Results from the calibration. The parameter $t \in [0, 1]$ indicates firm rank in desirability from the perspective of workers. Means for weekly hours, weekly earnings, and hourly wages are computed with respect to the worker distribution, while firm size and profits per worker is averaged with respect to the firm distribution.

The third column of Table A.14 adds a minimum wage set at the current federal rate of \$$7.25$. This provides a small increase of about 30 cents to the average hourly wage, which now begins at \$7.25 for $t = 0$ rather than \$6.06. Note that the minimum wage provides spillovers by reallocating firm mass up the entire wage ladder, beyond the mechanical effect of increasing the wages of those previously below 7.25. Average hours worked are decreased slightly due to the minimum wage, by about ten minutes per week. As this effect is mediated by a wealth effect in labor supply, we can expect it to be small unless worker preferences deviate significantly from quasi-linearity with respect to income. With $\beta = .85$, this effect is reasonably modest but non-negligible. In the fourth column, we see that the combination of the minimum wage and overtime premium has little effect beyond the direct effect of the minimum wage: hourly earnings increase another 15 cents and hours worked go down by another 0.07. Finally, we see that increasing the minimum wage to \$12 has much larger effects: adding another dollar to average wages and reducing working time by a bit more than half an hour per week. Given the fixed costs assumed in this calibration, a \$12 minimum wage causes employers to run on extremely thin margins for these workers (who have \$15 an hour productivity). However, note that in this model a minimum wage causes neither an increase nor decrease in aggregate non-employment, as this is governed in the steady state only by λ/δ. Thus, the average absolute firm size is unchanged across the policy environments.

A.7 Main Proofs

A.7.1 Proof of Lemma A.1

For any convex budget function $B(\mathbf{x})$, $(z_{Bi}, \mathbf{x}_{Bi}) = \text{argmax}_{z,\mathbf{x}} \{u_i(z, \mathbf{x}) \text{ s.t. } z \geq B(\mathbf{x})\}$ exists and is unique since it maximizes the strictly quasi-concave function $u_i(z, \mathbf{x})$ over the convex domain $\{(z, \mathbf{x}) : z \geq B(\mathbf{x})\}$. Furthermore, by monotonicity of $u(z, \mathbf{x})$ in z we may substitute in the constraint $z = B(\mathbf{x})$ and write

$$\mathbf{x}_{Bi} = \text{argmax}_{\mathbf{x}} u_i(B(\mathbf{x}), \mathbf{x})$$

Consider any $\mathbf{x} \neq \mathbf{x}_{Bi}$, and let $\tilde{\mathbf{x}} = \theta\mathbf{x} + (1-\theta)\mathbf{x}^*$ where $\mathbf{x}^* = \mathbf{x}_{Bi}$ and $\theta \in (0,1)$. Since $B(\mathbf{x})$ is convex in $\mathbf{x}$ and $u_i(z, \mathbf{x})$ is weakly decreasing in z:

$$u_i(B(\tilde{\mathbf{x}}), \tilde{\mathbf{x}}) \geq u_i(\theta B(\mathbf{x}) + (1-\theta)B(\mathbf{x}^*), \tilde{\mathbf{x}}) > \min\{u_i(B(\mathbf{x}), \mathbf{x}), u_i(B(\mathbf{x}^*), \mathbf{x}^*)\} = u_i(B(\mathbf{x}), \mathbf{x}) \tag{A.18}$$

where I have used strict quasi-concavity of $u_i(z, \mathbf{x})$ in the second step, and that $\mathbf{x}^*$ is a maximizer in the third. This result implies that for any $\mathbf{x} \neq \mathbf{x}^*$, if one draws a line between $\mathbf{x}$ and $\mathbf{x}^*$, the function $u_i(B(\mathbf{x}), \mathbf{x})$ is strictly increasing as one moves towards $\mathbf{x}^*$. When $\mathbf{x}$ is a scalar, this argument is used by Blomquist et al. (2015) (see Lemma A2 therein) to show that $u_i(B(\mathbf{x}), \mathbf{x})$ is strictly increasing to the left of $\mathbf{x}^*$, and strictly decreasing to the right of $\mathbf{x}^*$. Note that for any (binding) linear budget constraint $B(\mathbf{x})$, the result holds without monotonicity of $u_i(z, \mathbf{x})$ in z. This is useful for Theorem 2* in which some workers choose their hours.

Let $\mathcal{X}_{0i} = \{\mathbf{x} : y_i(\mathbf{x}) \leq k\}$ and $\mathcal{X}_{1i} = \{\mathbf{x} : y_i(\mathbf{x}) \geq k\}$. For any function B, let $u_{Bi}(\mathbf{x}) = u_i(B(\mathbf{x}), \mathbf{x})$, and note that

$$u_{B_k i}(\mathbf{x}) = \begin{cases} u_{B_0 i}(\mathbf{x}) & \text{if} \quad \mathbf{x} \in \mathcal{X}_{0i} \\ u_{B_1 i}(\mathbf{x}) & \text{if} \quad \mathbf{x} \in \mathcal{X}_{1i} \end{cases}$$

Let $\mathbf{x}_{ki}$ be the unique maximizer of $u_{B_k i}(\mathbf{x})$, where $Y_i = y_i(\mathbf{x}_{ki})$. Suppose that $Y_i < k$. By continuity of $y_i(\mathbf{x})$, $\mathcal{X}_{0i}$ is a closed set and $\mathbf{x}_{ki}$ belongs to the interior of $\mathcal{X}_{0i}$. Suppose furthermore that $Y_{0i} \neq Y_i$, with $\mathbf{x}_{0i}$ the maximizer of $u_{B_0 i}(\mathbf{x})$. If this were the case, then there would exist a point $\tilde{\mathbf{x}} \in \mathcal{X}_{0i}$ along the line from $\mathbf{x}_{0i}$ to $\mathbf{x}_{ki}$. By Eq. (A.18) with $B = B_k$, we must have $u_{B_k i}(\tilde{\mathbf{x}}) > u_{B_k i}(\mathbf{x}_{0i})$. Since $u_{B_0 i}(\mathbf{x}) = u_{B_k i}(\mathbf{x})$ in $\mathcal{X}_{0i}$ this means that $u_{B_0 i}(\tilde{\mathbf{x}}) > u_{B_0 i}(\mathbf{x}0i)$, contradicting the premise that $\mathbf{x}_{0i}$ maximizes $u_{B_0 i}(\mathbf{x})$. Figure A.18 depicts the logic visually. Thus, $Y_i < k$ implies $Y_i = Y_{0i}$. We can similarly show that $Y_i > k$ implies $Y_i = Y_{1i}$. Taking the contrapositive of each of these, we have that $Y_{1i} \leq k \leq Y_{0i}$

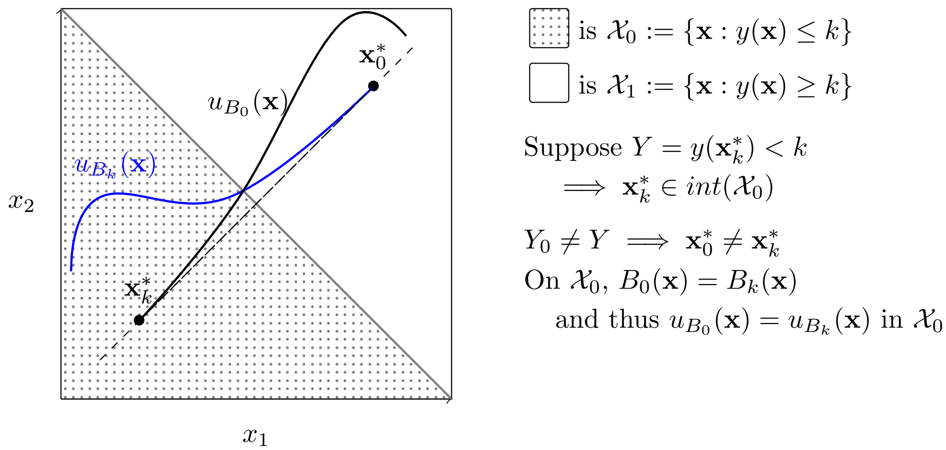

Figure A.18: Depiction of the step establishing $(Y < k) \implies (Y = Y_0)$ in the proof of Lemma A.1. In this example $z = (x_1, x_2)$ and $y(\mathbf{x}) = x_1 + x_2$. We suppress indices i for clarity. Proof is by contradiction. If $Y_0 \neq Y$, then $\mathbf{x}_k^* \neq \mathbf{x}_0^*$, where $\mathbf{x}_k^*$ and $\mathbf{x}_0^*$ are the unique maximizers of $u_{B_k}(\mathbf{x})$ and $u_{B_0}(\mathbf{x})$, respectively. By Equation A.18, we have that the function $u_{B_0}(\mathbf{x})$, depicted heuristically as a solid black curve, is strictly increasing as one moves along the dotted line from $\mathbf{x}_k^*$ towards $\mathbf{x}_0^*$. Similarly, the function $u_{B_0}(\mathbf{x})$, depicted as a solid blue curve, is strictly increasing as one moves in the opposite direction along the same line, from $\mathbf{x}_0^*$ towards z_k^*. By the assumption that $Y < k$, then using continuity of $y(\mathbf{x})$ it must be the case that $\mathbf{x}_k^*$ lies in the interior of $\mathcal{X}_0$, the set of $\mathbf{x}$'s that make $y(\mathbf{x}) \leq k$. This means that there is some interval of the dotted line that is within $\mathcal{X}_0$ (regardless of whether z_0^* is also within $\mathcal{X}_0$, or it is not, as depicted). On this interval, the functions B_0 and B_k are equal, and thus so must be the functions u_{B_k} and u_{B_0}. Since the same function cannot be both strictly increasing and strictly decreasing, we have obtained a contradiction.

implies that $Y_i = k$.

It is easily demonstrated under WARP alone (see the proof of Theorem A.1 below) that $Y_{0i} \leq k$ implies that $Y_i = Y_{0i}$ and that $Y_{1i} \geq k$ implies that $Y_i = Y_{1i}$. Note that together these imply that $Y_{0i} < k \leq Y_{1i}$ and $Y_{0i} \leq k < Y_{1i}$ are both impossible (since we would then have both that $Y_i < k$ and $Y_i \geq k$ or that both that $Y_i \leq k$ and $Y_i > k$). Thus, we can summarize the relationship between observable Y_i and potential outcomes in the remaining three cases as:

$$Y_i = \begin{cases} Y_{0i} & \text{if } \; Y_{0i} < k \\ k & \text{if } \; Y_{1i} \leq k \leq Y_{0i} \\ Y_{1i} & \text{if } \; Y_{1i} > k \end{cases}$$

A.7.2 Proof of Theorem A.1

We first prove the statement in b). If $Y_{0i} \leq k$, then by CHOICE $\mathbf{x}_{B_0}$ is in $\mathcal{X}_0$, where $\mathcal{X}_0$ is defined in the proof of Lemma A.1. Since $B_k(\mathbf{x}) = B_0(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_0$, it follows that $z_{B_0 i} \geq B_k(\mathbf{x}_{B_0 i})$, i.e. Y_{0i} is feasible under B_k. Note that $B_{ki}(\mathbf{x}) \geq B_{0i}(\mathbf{x})$ for all $\mathbf{x}$. By WARP then $(z_{B_k i}, \mathbf{x}_{B_k i}) = (z_{B_0 i}, \mathbf{x}_{B_0 i})$. Thus $Y_i = y_i(\mathbf{x}_{B_k}) = y_i(\mathbf{x}_{B_0}) = Y_{0i}$. So $Y_{0i} \leq k \implies Y_i = Y_{0i}$. As an implication we have that $Y_{0i} < k \implies Y_i < k$.

By the same logic we can show that $Y_{1i} \geq k \implies Y_i = Y_{1i}$ and thusly that $Y_{1i} > k \implies Y_i > k$. Taking the contrapositives, we see that $Y_i = k \iff Y_i \leq k \;\&\; Y_i \geq k$ implies $Y_{1i} \leq k$ and $Y_{0i} \geq k$. Thus $Y_i = k$ implies $Y_{1i} \leq k \leq Y_{0i}$ and hence $\mathcal{B} \leq P(Y_{1i} \leq k \leq Y_{0i})$.

This holds under CONVEX or WARP since CONVEX implies WARP. However under CONVEX we also have from Lemma A.1 that $Y_{1i} \leq k \leq Y_{0i}$ implies $Y_i = k$, and thus $\mathcal{B} \geq P(Y_{1i} \leq k \leq Y_{0i})$. Together we have that both $\mathcal{B} \leq P(Y_{1i} \leq k \leq Y_{0i})$ and $\mathcal{B} \geq P(Y_{1i} \leq k \leq Y_{0i})$ and hence $\mathcal{B} = P(Y_{1i} \leq k \leq Y_{0i})$ under CONVEX.

A.7.3 Proof of the Corollary to Theorem A.1

In the proof of Theorem A.1 I showed that under WARP and CHOICE, $Y_{0i} \leq k \implies Y_i = Y_{0i}$. Thus, for any $\delta > 0$ and $y < k$: $Y_{0i} \in [y - \delta, y]$ implies that $Y_i \in [y - \delta, y]$ and hence $P(Y_{0i} \in [y-\delta, y]) - P(Y_i \in [y - \delta, y])$ is negative. This implies that $f_0(y) - f(y) = \lim_{\delta \downarrow 0} \frac{P(Y_{0i} \in [y-\delta,y]) - P(Y_i \in [y-\delta,y])}{\delta} \leq 0$, i.e. that $f(y) \geq f_0(y)$. An analogous argument holds for Y_1, where we consider the event $Y_{1i} \in [y, y+\delta]$ any $y > k$. Under CONVEX instead of WARP, the inequalities are all equalities, by Lemma A.1.

A.7.4 Proof of Theorem 1.1

By Theorem 1 of Dümbgen et al. (2017): for $d \in \{0, 1\}$ and any t, bi-log concavity implies that:

$$1 - (1 - F_{d|K^*=0}(k))e^{-\frac{f_{d|K^*=0}(k)}{1-F_{d|K^*=0}(k)}t} \leq F_{d|K^*=0}(k + t) \leq F_{d|K^*=0}(k)e^{\frac{f_{d|K^*=0}(k)}{F_{d|K^*=0}(k)}t}$$

Defining $u = F_{0|K^*=0}(k + t)$, we can use the substitution $t = Q_{0|K^*=0}(u) - k$ to translate the above into bounds on the conditional quantile function of Y_{0i}, evaluated at u:

$$\frac{F_{0|K^*=0}(k)}{f_{0|K^*=0}(k)} \cdot \ln\left(\frac{u}{F_{0|K^*=0}(k)}\right) \leq Q_{0|K^*=0}(u) - k \leq -\frac{1 - F_{0|K^*=0}(k)}{f_{0|K^*=0}(k)} \cdot \ln\left(\frac{1-u}{1 - F_{0|K^*=0}(k)}\right)$$

And similarly for Y_1, letting $v = F_{1|K^*=0}(k - t)$:

$$\frac{1 - F_{1|K^*=0}(k)}{f_{1|K^*=0}(k)} \cdot \ln\left(\frac{1-v}{1 - F_{1|K^*=0}(k)}\right) \leq k - Q_{1|K^*=0}(v) \leq -\frac{F_{1|K^*=0}(k)}{f_{1|K^*=0}(k)} \cdot \ln\left(\frac{v}{F_{1|K^*=0}(k)}\right)$$

Note that:

$$\begin{aligned} E[Y_{0i} - Y_{1i}|Y_i = k, K_i^* = 0] &= \frac{1}{\mathcal{B}^*}\int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k)+\mathcal{B}^*} \{Q_{0|K^*=0}(u) - Q_{0|K^*=0}(u)\}du \\ &= \frac{1}{\mathcal{B}^*}\int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k)+\mathcal{B}^*} \{Q_{0|K^*=0}(u) - k\}du + \frac{1}{\mathcal{B}^*}\int_{F_{1|K^*=0}(k)-\mathcal{B}^*}^{F_{1|K^*=0}(k)} \{k - Q_{1|K^*=0}(v)\}dv \end{aligned}$$

where $\mathcal{B}^* := P(Y_i = k|K^* = 0)$. A lower bound for $E[Y_{0i} - Y_{1i}|Y_i = k, K_i^* = 0]$ is thus:

$$\begin{aligned} &\frac{F_{0|K^*=0}(k)}{f_{0|K^*=0}(k)(\mathcal{B}^*)}\int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k)+\mathcal{B}^*} \ln\left(\frac{u}{F_{0|K^*=0}(k)}\right)du + \frac{1 - F_{1|K^*=0}(k)}{f_{1|K^*=0}(k)(\mathcal{B}^*)}\int_{F_{1|K^*=0}(k)-(\mathcal{B}^*)}^{F_{1|K^*=0}(k)} \ln\left(\frac{1-v}{1 - F_{1|K^*=0}(k)}\right)dv \\ &= g(F_{0|K^*=0}(k), f_{0|K^*=0}(k), \mathcal{B}^*) + h(F_{1|K^*=0}(k), f_{1|K^*=0}(k), \mathcal{B}^*) \end{aligned}$$

where

$$
\begin{aligned}
g(a,b,x) := \frac{a}{bx}\int_a^{a+x} \ln\left(\frac{u}{a}\right) du &= \frac{a^2}{bx}\int_1^{1+\frac{x}{a}} \ln(u)\, du \\
&= \frac{a^2}{bx}\left\{u\ln(u) - u\right\}\Big|_1^{1+\frac{x}{a}} \\
&= \frac{a^2}{bx}\left\{\left(1+\frac{x}{a}\right)\ln\left(1+\frac{x}{a}\right) - \frac{x}{a}\right\} \\
&= \frac{a}{bx}(a+x)\ln\left(1+\frac{x}{a}\right) - \frac{a}{b}
\end{aligned}
$$

and

$$
h(a,b,x) := \frac{1-a}{bx}\int_{a-x}^{a} \ln\left(\frac{1-v}{1-a}\right) dv = \frac{(1-a)^2}{bx}\int_1^{1+\frac{x}{1-a}} \ln(u)\, du = g(1-a,b,x)
$$

Similarly, an upper bound is:

$$
\begin{aligned}
&-\frac{1-F_{0|K^*=0}(k)}{f_{0|K^*=0}(k)(\mathcal{B}^*)}\int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k)+\mathcal{B}^*} \ln\left(\frac{1-u}{1-F_{0|K^*=0}(k)}\right) du \\
&\qquad\qquad - \frac{F_{1|K^*=0}(k)}{f_{1|K^*=0}(k)(\mathcal{B}^*)}\int_{F_{1|K^*=0}(k)-(\mathcal{B}^*)}^{F_{1|K^*=0}(k)} \ln\left(\frac{v}{F_{1|K^*=0}(k)}\right) dv \\
&= g'(F_{0|K^*=0}(k), f_{0|K^*=0}(k), \mathcal{B}^*) + h'(F_{1|K^*=0}(k), f_{1|K^*=0}(k), \mathcal{B}^*)
\end{aligned}
$$

where

$$
\begin{aligned}
g'(a,b,x) := -\frac{1-a}{bx}\int_a^{a+x} \ln\left(\frac{1-u}{1-a}\right) du &= -\frac{(1-a)^2}{bx}\int_{1-\frac{x}{1-a}}^{1} \ln(u)\, du \\
&= \frac{(1-a)^2}{bx}\left\{u - u\ln(u)\right\}\Big|_{1-\frac{x}{1-a}}^{1} \\
&= \frac{1-a}{b} + \frac{1-a}{bx}(1-a-x)\ln\left(1-\frac{x}{1-a}\right) \\
&= -g(1-a,b,-x)
\end{aligned}
$$

and

$$h'(a,b,x) := -\frac{a}{bx}\int_{a-x}^{a} \ln\left(\frac{v}{a}\right) dv = -\frac{a^2}{bx}\int_{1-\frac{x}{a}}^{1} \ln(u)\, du = g'(1-a,b,x) = -g(a,b,-x)$$

This $\Delta_k^* \in [\Delta_k^L, \Delta_k^U :]$ were

$$\Delta_k^L := g\left(F_-(k), f_-(k), \mathcal{B}-p\right) + g\left(1-F(k), f_+(k), \mathcal{B}-p\right)$$

and

$$\Delta_k^U := -g\left(1-p-F_-(k), f_-(k), p-\mathcal{B}\right) - g\left(F(k)-p, f_+(k), p-\mathcal{B}\right)$$

The bounds are sharp as CHOICE, CONVEX and RANK imply no further restrictions on the marginal potential outcome distributions. To obtain the final result, note then that

$$F_{0|K^*=0}(k) = \frac{F_0(k)-p}{1-p} \quad \text{and} \quad F_{1|K^*=0}(k) = \frac{F_1(k)-p}{1-p}$$

$$f_{0|K^*=0}(k) = \frac{f_0(k)}{1-p} \quad \text{and} \quad f_{1|K^*=0}(k) = \frac{f_1(k)}{1-p}$$

$$\mathcal{B}^* := P(Y_i = k | K_i^* = 0) = \frac{\mathcal{B}-p}{1-p}$$

and finally that the function $g(a,b,x)$ is homogeneous of degree zero. As shown by Dümbgen et al. (2017), BLC implies the existence of a continuous density function, which assures that these density limits exist and are equal to the corresponding potential outcome densities above.

A.7.5 Proof of Lemma A.2

Let $\Delta_i^k(\rho,\rho') := Y_i(\rho,k) - Y_i(\rho',k)$ for any $\rho,\rho' \in [\rho_0,\rho_1]$ and value of k.

Assumption SMOOTH (regularity conditions). *The following hold:*

1. $P(\Delta_i^k(\rho,\rho') \leq \Delta, Y_i(\rho,k) \leq y)$ *is twice continuously differentiable at all* $(\Delta,y) \neq (0,k^*)$,

for any $\rho, \rho' \in [\rho_0, \rho_1]$ and k.

2. *$Y_i(\rho, k) = Y(\rho, k, \epsilon_i)$, where ϵ_i has compact support $E \subset \mathbb{R}^m$ for some m. $Y(\cdot, k, \cdot)$ is continuously differentiable on all of $[\rho_0, \rho_1] \times E$, for every k.*

3. *there possibly exists a set $\mathcal{K}^* \subset E$ such that $Y(\rho, k, \epsilon) = k^*$ for all $\rho \in [\rho_0, \rho_1]$ and $\epsilon \in \mathcal{K}^*$. The quantity $\mathbb{E}\left[\frac{\partial Y_i(\rho,k)}{\partial \rho}\middle| Y_i(\rho, k) = y, \epsilon_i \notin \mathcal{K}^*\right]$ is continuously differentiable in y for all y including k^*.*

In the remainder of this proof I keep k be implicit in the functions $Y_i(\rho, k)$ and $\Delta_i^k(\rho, \rho')$, as it will remained fixed. Item 1 of SMOOTH excludes the point $(0, k^*)$ on the basis that we may expect point masses at $Y_i(\rho) = k^*$, as in the overtime setting. Following Section 1.4, item 3 imposes that all such "counterfactual bunchers" have zero treatment effects, while also introducing a further condition that will be used later in Lemma A.3. Let K_i^* be an indicator for $\epsilon_i \in \mathcal{K}^*$ and denote $p = P(K_i^* = 1)$. Item 1 implies that the density $f_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, y)$ is continuous in y whenever $y \neq k^*$ or $\Delta \neq 0$, so I define $f_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, k^*) = \lim_{y\to k^*} f_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, y)$ for any ρ, ρ' and Δ. Similarly, we can define the marginal density $f_\rho(y)$ of $Y_i(\rho)$ at k^* to be $\lim_{y\to k^*} f_\rho(y)$ for any ρ.

The main tool in the proof of Lemma A.2 will be the following Lemma, which shows that the uniform density approximation of Theorem A.3 becomes exact in the limit that the two cost functions approach one another.

Lemma SMALL (small kink limit). *Assume CHOICE*, WARP, and SMOOTH. Then:*

$$\lim_{\rho' \downarrow \rho} \frac{P(Y_i(\rho) \leq k \leq Y_i(\rho')) - p(k)}{\rho' - \rho} = -f_\rho(k)\mathbb{E}\left[\frac{dY_i(\rho)}{d\rho}\middle| Y_i(\rho) = k\right]$$

Proof. Throughout this proof we let f_W denote the density of a generic random variable or random vector W_i, if it exists. Write $\Delta_i(\rho, \rho') = \Delta_i(\rho, \rho', \epsilon_i)$ where $\Delta_i(\rho, \rho', \epsilon) := Y(\rho, \epsilon) -$

$Y(\rho', \epsilon)$.

$$\begin{aligned}
\lim_{\rho' \downarrow \rho} \frac{P(Y_i(\rho) \leq k \leq Y_i(\rho')) - p(k)}{\rho' - \rho} &= \lim_{\rho' \downarrow \rho} \frac{P(Y_i(\rho) \in [k, k + \Delta(\rho, \rho')_i]) - p(k)}{\rho' - \rho} \\
&= \lim_{\rho' \downarrow \rho} \frac{P(Y_i(\rho) \in (k, k + \Delta(\rho, \rho')_i])}{\rho' - \rho} \\
&= \lim_{\rho' \downarrow \rho} \frac{1}{\rho' - \rho} \int_0^\infty d\Delta \int_k^{k+\Delta} dy \cdot f_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, y) \\
&= \lim_{\rho' \downarrow \rho} \int_0^\infty d\Delta \int_k^{k+\Delta} dy \cdot \frac{f_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, k) + (y-k) r_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, k, y)}{\rho' - \rho}
\end{aligned} \tag{A.19}$$

where we have used that by item 1 the joint density of $\Delta_i(\rho, \rho')$ and $Y_i(\rho)$ exists for any ρ, ρ' and is differentiable and $r_{\Delta(\rho,\rho'),Y(\rho)}$ is a first-order Taylor remainder term satisfying

$$\lim_{y \downarrow k} |r_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, y)| = |r_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, k)| = 0$$

for any Δ.

I now show that the whole term corresponding to this remainder is zero. First, note that:

$$\begin{aligned}
\left| \lim_{\rho' \downarrow \rho} \int_0^\infty d\Delta \int_k^{k+\Delta} dy \cdot \frac{(y-k) r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta, y)}{\rho' - \rho} \right| &= \lim_{\rho' \downarrow \rho} \left| \int_0^\infty d\Delta \int_k^{k+\Delta} dy \cdot \frac{(y-k) r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta, y)}{\rho' - \rho} \right| \\
&\leq \lim_{\rho' \downarrow \rho} \int_0^\infty d\Delta \int_k^{k+\Delta} dy \cdot \left| \frac{(y-k) r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta, y)}{\rho' - \rho} \right| \\
&\leq \lim_{\rho' \downarrow \rho} \int_0^\infty d\Delta \frac{\Delta}{\rho' - \rho} \int_k^{k+\Delta} dy \cdot \left| r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta, y) \right|
\end{aligned}$$

where I've used continuity of the absolute value function and the Minkowski inequality. Define $\xi(\rho, \rho') = \sup_{\epsilon \in E} \Delta(\rho, \rho', \epsilon)$. The strategy will be show that $\lim_{\rho' \downarrow \rho} \xi(\rho, \rho') = 0$, and then since $r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta, y) = 0$ for any $\Delta > \xi(\rho, \rho')$ and all y (since the marginal density $f_{\Delta(\rho,\rho')}(\Delta)$ would be zero for such Δ). With $\xi(\rho, \rho')$ so-defined:

$$\begin{aligned}
\text{RHS of above} &\leq \lim_{\rho' \downarrow \rho} \int_0^{\xi(\rho,\rho')} d\Delta \frac{\xi(\rho, \rho')}{\rho' - \rho} \int_k^{k+\xi(\rho,\rho')} dy \cdot \left| r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta, y) \right| \\
&= \lim_{\rho' \downarrow \rho} \frac{\xi(\rho, \rho')}{\rho' - \rho} \cdot \lim_{\rho' \downarrow \rho} \int_0^{\xi(\rho,\rho')} d\Delta \int_0^{\xi(\rho,\rho')} dy \cdot \left| r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta, k + y) \right|
\end{aligned} \tag{A.20}$$

where in the second step I have assumed that each limit exists (this will be demonstrated below). Let us first consider the inner integral of the above: $\int_k^{k+\xi(\rho,\rho')} dy \cdot \left| r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta, y) \right|$, for any Δ. Supposing that $\lim_{\rho'\downarrow\rho} \xi(\rho,\rho') = 0$, it follows that this inner integral evaluates to zero, by the Leibniz rule and using that $r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta, k) = 0$. Thus the entire second limit is equal to zero.

Now I prove that $\lim_{\rho'\downarrow\rho} \xi(\rho,\rho') = 0$ and that $\lim_{\rho'\downarrow\rho} \frac{\xi(\rho,\rho')}{\rho'-\rho}$ exists. First, note that continuous differentiability of $Y(\rho,\epsilon_i)$ implies $Y_i(\rho)$ is continuous for each i so $\lim_{\rho'\downarrow\rho} \Delta_i(\rho,\rho') = 0$ point-wise in ϵ. We seek to turn this point-wise convergence into uniform convergence over ϵ, i.e. that $\lim_{\rho'\downarrow\rho} \sup_{\epsilon\in E} \Delta(\rho,\rho',\epsilon) = \sup_{\epsilon\in E} \lim_{\rho'\downarrow\rho} \Delta(\rho,\rho',\epsilon) = \sup_{\epsilon\in E} 0 = 0$. The strategy will be to use equicontinuity of the sequence and compactness of E. Consider any such sequence $\rho_n \overset{n}{\to} \rho$ from above, and let $f_n(\epsilon) := Y(\rho,\epsilon) - Y(\rho_n,\epsilon)$ and $f(\epsilon) = \lim_{n\to\infty} f_n(\epsilon) = 0$. Equicontinuity of the sequence $f_n(\epsilon)$ says that for any $\epsilon, \epsilon' \in E$ and $e > 0$, there exists a $\delta > 0$ such that $||\epsilon - \epsilon'|| < \delta \implies |f_n(\epsilon) - f_n(\epsilon')| < e$.

This follows from continuous differentiability of $Y(\rho,\epsilon)$. Let $M = \sup_{\rho\in[\rho_0,\rho_1],\epsilon\in E} |\nabla_{\rho,\epsilon} Y(\rho,\epsilon)|$. M exists and is finite given continuity of the gradient and compactness of $[\rho_0,\rho_1] \times E$. Then, for any two points $\epsilon, \epsilon' \in E$ and any $\rho \in [\rho_0,\rho_1]$:

$$|Y(\rho,\epsilon) - Y(\rho,\epsilon')| = \left| \int_{\epsilon'}^{\epsilon} \nabla_\epsilon Y(\rho,\epsilon) \cdot \mathbf{d}\epsilon \right| \le \int_{\epsilon'}^{\epsilon} |\nabla_\epsilon Y(\rho,\epsilon) \cdot \mathbf{d}\epsilon| \le M \int_{\epsilon'}^{\epsilon} ||\mathbf{d}\epsilon|| \le M||\epsilon - \epsilon'||$$

where $\mathbf{d}\epsilon$ is any path from ϵ to ϵ' and I have used the definition of M and Cauchy-Schwarz in the second inequality. The existence of a uniform Lipschitz constant M for $Y(\rho,\epsilon)$ implies a uniform equicontinuity of $Y(\rho,\epsilon)$ of the form that for any $e > 0$ and $\epsilon, \epsilon' \in E$, there exists a $\delta > 0$ such that $||\epsilon - \epsilon'|| < \delta \implies \sup_{\rho\in[\rho_0,\rho_1]} |Y(\rho,\epsilon) - Y(\rho,\epsilon')| < e/2$, since we can simply take $\delta = e/(2M)$. This in turn implies that whenever $||\epsilon - \epsilon'|| < \delta$:

$$\begin{aligned} |Y(\rho,\epsilon) - Y(\rho_n,\epsilon) - \left\{Y(\rho,\epsilon') - Y(\rho_n,\epsilon')\right\}| &= |Y(\rho,\epsilon) - Y(\rho,\epsilon') - \left\{Y(\rho_n,\epsilon) - Y(\rho_n,\epsilon')\right\}| \\ &\le |Y(\rho,\epsilon) - Y(\rho,\epsilon')| + |Y(\rho_n,\epsilon) - Y(\rho_n,\epsilon')| \le e, \end{aligned}$$

our desired result. Together with compactness of E, equicontinuity implies that $\lim_{n\to\infty}\sup_{\epsilon\in E} f_n(\epsilon) = \sup_{\epsilon\in E}\lim_{n\to\infty} f_n(\epsilon) = 0$.

We apply an analogous argument for $\lim_{\rho'\downarrow\rho}\frac{\xi(\rho,\rho')}{\rho'-\rho}$, where now $f_n(\epsilon) = \frac{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)}{\rho_n-\rho}$. For this case it's easier to work directly with the function $\frac{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)}{\rho_n-\rho}$, showing that it is Lipschitz in deviations of ϵ uniformly over $n\in\mathbb{N}, \epsilon\in E$.

$$\begin{aligned}
\left|\frac{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)}{\rho_n-\rho} - \frac{Y(\rho,\epsilon')-Y(\rho_n,\epsilon')}{\rho_n-\rho}\right| &= \frac{1}{\rho_n-\rho}\left|\int_{\epsilon'}^{\epsilon}\nabla_\epsilon Y(\rho,\epsilon)\cdot\mathbf{d}\epsilon - \int_{\epsilon'}^{\epsilon}\nabla_\epsilon Y(\rho_n,\epsilon)\cdot\mathbf{d}\epsilon\right| \\
&\leq \frac{1}{\rho_n-\rho}\left(\left|\int_{\epsilon'}^{\epsilon}\nabla_\epsilon Y(\rho,\epsilon)\cdot\mathbf{d}\epsilon\right| + \left|\int_{\epsilon'}^{\epsilon}\nabla_\epsilon Y(\rho_n,\epsilon)\cdot\mathbf{d}\epsilon\right|\right) \\
&\leq \frac{2M}{\rho_n-\rho}\int_{\epsilon'}^{\epsilon}||\mathbf{d}\epsilon|| \leq \frac{2M}{\rho_n-\rho}||\epsilon-\epsilon'||
\end{aligned}$$

This implies equicontinuity of $\frac{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)}{\rho_n-\rho}$with the choice $\delta = e(\rho_n-\rho)/(2M)$. As before, equicontinuity and compactness of E allow us to interchange the limit and the supremum, and thus:

$$\begin{aligned}
\lim_{n\to\infty}\frac{\xi(\rho,\rho_n)}{\rho_n-\rho} &= \lim_{n\to\infty}\frac{\sup_{\epsilon\in E}\{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)\}}{\rho_n-\rho} = \lim_{n\to\infty}\sup_{\epsilon\in E}\frac{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)}{\rho_n-\rho} \\
&= \sup_{\epsilon\in E}\lim_{n\to\infty}\frac{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)}{\rho_n-\rho} = \sup_{\epsilon\in E}\frac{\partial Y(\rho,\epsilon)}{\partial\rho} := M' < \infty
\end{aligned}$$

where finiteness of M' follows from it being defined as the supremum of a continuous function over a compact set. This establishes that the first limit in Eq. (A.20) exists and is finite, completing the proof that it evaluates to zero.

Given that the second term in Eq. (A.19) is zero, we can simplify the remaining term as:

$$
\begin{aligned}
\lim_{\rho'\downarrow\rho} \frac{P(Y_i(\rho) \le k \le Y_i(\rho')) - p(k)}{\rho' - \rho} &= \lim_{\rho'\downarrow\rho} \frac{1}{\rho' - \rho} \int_0^\infty f_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, k)\Delta d\Delta \\
&= f_\rho(k) \lim_{\rho'\downarrow\rho} \frac{1}{\rho' - \rho} P(\Delta_i(\rho,\rho') \ge 0 | Y_i(\rho) = k) \\
&\qquad \cdot \mathbb{E}\left[\Delta_i(\rho,\rho') | Y_i(\rho) = k, \Delta_i(\rho,\rho') \ge 0\right] \\
&= f_\rho(k)(k) \lim_{\rho'\downarrow\rho} \frac{1}{\rho' - \rho} \mathbb{E}\left[\Delta_i(\rho,\rho') | Y_i(\rho) = k, \Delta_i(\rho,\rho')\right] \\
&= f_\rho(k)(k) \mathbb{E}\left[\lim_{\rho'\downarrow\rho} \frac{\Delta_i(\rho,\rho')}{\rho' - \rho} \middle| Y_i(\rho) = k\right] \\
&= f_\rho(k) \mathbb{E}\left[-\frac{Y_i(\rho)}{d\rho} \middle| Y_i(\rho) = k\right]
\end{aligned}
$$

where I have used Lemma POS and then finally the dominated convergence theorem. To see that we may use the latter, note that $\frac{dY_i(\rho)}{d\rho} = \frac{\partial Y(\rho,\epsilon_i)}{\partial\rho} < M$ uniformly over all $\epsilon_i \in E$, and $\mathbb{E}\left[M | Y_i(\rho) = k\right] = M < \infty$. □

Now we return to the proof of Lemma A.2. By item 1 of Assumption SMOOTH, the marginal $F_\rho(y) := P(Y_i(\rho) \le y)$ is differentiable away from $y = k$ with derivative $f_\rho(y)$. From the proof of Theorem A.1 it follows that $\mathcal{B} \le F_{\rho_1}(k) - F_{\rho_0}(k) + p(k)$ with equality under CONVEX, and thus:

$$
\begin{aligned}
\mathcal{B} - p(k) &\le F_{\rho_1}(k) - F_{\rho_0}(k) \\
&= \int_{\rho_0}^{\rho_1} \frac{d}{d\rho} F_\rho(k) d\rho \\
&= \int_{\rho_0}^{\rho_1} \lim_{\delta\downarrow 0} \frac{F_{\rho+\delta}(k) - F_\rho(k)}{\delta} d\rho \\
&= \int_{\rho_1}^{\rho_0} \lim_{\delta\downarrow 0} \frac{F_\rho(k) - F_{\rho+\delta}(k)}{\delta} d\rho \\
&= \int_{\rho_1}^{\rho_0} \lim_{\delta\downarrow 0} \frac{P(Y_i(\rho) \le k \le Y_i(\rho+\delta)) - p(k)}{\delta} d\rho \\
&= \int_{\rho_1}^{\rho_0} f_\rho(k) \mathbb{E}\left[\frac{Y_i(\rho)}{d\rho} \middle| Y_i(\rho) = k\right] d\rho
\end{aligned}
$$

where the fourth equality has applied the identity $1 = P(Y_{0i} \le k) + P(Y_i(\rho) \le k \le$

$Y_i(\rho+\delta)) + P(Y_{1i} > k)$ under CHOICE and WARP to the pair of choice constraints $B(\rho)$ and $B(\rho+\delta)$, noting that $P(Y_i(\rho) < k) = F_\rho(k) - p(k)$.

A.7.6 Proof of Lemma A.3

This mostly follows the proof in Kasy (2017) adapted to our setting in which y is one-dimensional. As in the proof of Lemma A.2 I leave k implicit in the functions $Y_i(\rho, k)$ and $Y(\rho, k, \epsilon)$, as k remains fixed throughout. One additional subtlety concerns the possibility of a point mass in the distribution of each $Y_i(\rho)$ at k^*. Note that Assumption SMOOTH implies a continuous density $f_\rho(y)$ for all $\rho \in [\rho_0, \rho_1]$ and $y \neq k^*$, which is also continuously differentiable in ρ. We define $f_\rho(k^*) = \lim_{y\to k} f_\rho(y)$ in the case that $p > 0$.

Consider any bounded differentiable function $a(y)$ having the property that $a(k^*) = 0$, and note that we may write $A(y) := \frac{d}{d\rho}\mathbb{E}[a(Y_i(\rho))]$ in two separate ways. Firstly:

$$A(y) = \frac{d}{d\rho}\int dy \cdot f_\rho(y) \cdot a(y) = \int dy \cdot a(y) \cdot \frac{d}{d\rho} f_\rho(y) \tag{A.21}$$

and secondly:

$$A(y) = \frac{d}{d\rho}\mathbb{E}[a(Y_i(\rho, \epsilon_i))] = \int dF_\epsilon(\epsilon)\frac{d}{d\rho}a(Y(\rho,\epsilon)) = \int dF_\epsilon(\epsilon) a'(Y(\rho,\epsilon)) \cdot \partial_\rho Y(\rho, \epsilon) \tag{A.22}$$

The first representation integrates over the distribution of $Y_i(\rho)$, while the second integrates over the distribution of the underlying heterogeneity ϵ_i. In both cases we are justified in swapping the integral and derivative by boundedness of $a(y)$.

Continuing with Eq. (A.22), we may apply the law of iterated expectations over values of $Y(\rho, \epsilon)$, and then integrate by parts:

$$
\begin{aligned}
A(y) &= \int dy f_\rho(y) a'(y) \int dF_{\epsilon|Y(\rho,\epsilon)=y} \partial_\rho Y(\rho, \epsilon) \\
&= \int dy f_\rho(y) a'(y) \cdot \mathbb{E}\left[\frac{\partial Y(\rho,\epsilon)}{\partial \rho} \middle| Y(\rho, \epsilon) = y \right] \\
&= -\int dy \cdot a(y) \cdot \frac{\partial}{\partial y} \left\{ f_\rho(y) \cdot \mathbb{E}\left[\frac{\partial Y(\rho,\epsilon)}{\partial \rho} \middle| Y(\rho, \epsilon) = y \right] \right\}
\end{aligned}
$$

where we've assumed the density $f_\rho(y)$ vanishes at the limits of y. Comparing with Eq. (A.21), we see that for this to be true of any bounded differentiable function a (satisfying $a(k^*) = 0$, we must have

$$
\frac{d}{d\rho} f_\rho(y) = -\frac{\partial}{\partial y} \left\{ f_\rho(y) \cdot \mathbb{E}\left[\frac{\partial Y(\rho,\epsilon)}{\partial \rho} \middle| Y(\rho, \epsilon) = y \right] \right\}
$$

point-wise for all $y \neq k^*$.

Now consider $y = k^*$. First note that

$$
\frac{d}{d\rho} f_\rho(k^*) = \frac{d}{d\rho} \lim_{y \to k^*} f_\rho(y) = \lim_{y \to k^*} \frac{d}{d\rho} f_\rho(y) = -\lim_{y \to k^*} \frac{\partial}{\partial y} \left\{ f_\rho(y) \mathbb{E}\left[\frac{\partial Y(\rho,\epsilon)}{\partial \rho} \middle| Y(\rho, \epsilon) = y \right] \right\}
$$

where we can interchange the limit and derivative by the Moore-Osgood theorem, since $\frac{d}{d\rho} f_\rho(y)$ is uniformly bounded over $\rho \in [\rho_1, \rho_0]$ by Assumption SMOOTH. Furthermore, for all $y \neq k^*$: $\mathbb{E}\left[\frac{\partial Y(\rho,\epsilon)}{\partial \rho} \middle| Y(\rho, \epsilon) = y \right] = \mathbb{E}\left[\frac{\partial Y(\rho,\epsilon)}{\partial \rho} \middle| Y(\rho, \epsilon) = y, K_i^* = 0 \right]$, and the latter of these is continuously differentiable at all y (including $y = k^*$) by item 3 of Assumption SMOOTH. Thus:

$$
\frac{d}{d\rho} f_\rho(k^*) = -\frac{\partial}{\partial y} \left\{ f_\rho(k^*) \cdot \mathbb{E}\left[\frac{\partial Y(\rho,\epsilon)}{\partial \rho} \middle| Y(\rho, \epsilon) = k^*, K_i^* = 0 \right] \right\}
$$

since $f_\rho(y)$ is also continuously differentiable at $y = k^*$, by SMOOTH and the definition of $f_\rho(k^*)$ as $\lim_{y \to k^*} f_\rho(y)$.

A.7.7 Proof of Theorem 1.2

This proof follows the notation of Appendix A.1. Throughout this proof we let $Y_i(\rho, k) = Y_i(\rho)$, given Assumption SEPARABLE.

First, consider the effect of changing k on the bunching probability:

$$\begin{aligned}\partial_k \{\mathcal{B} - p(k)\} &= -\frac{\partial}{\partial k}\int_{\rho_0}^{\rho_1} f_\rho(k)\mathbb{E}\left[\left.\frac{Y_i(\rho)}{d\rho}\right| Y_i(\rho) = k\right] d\rho \\ &= -\int_{\rho_0}^{\rho_1} \frac{\partial}{\partial k}\left\{f_\rho(k)\mathbb{E}\left[\left.\frac{Y_i(\rho)}{d\rho}\right| Y_i(\rho) = k\right]\right\} d\rho \\ &= \int_{\rho_0}^{\rho_1} \partial_\rho f_\rho(k) d\rho = f_1(k) - f_0(k)\end{aligned}$$

I turn now to the total effect on average hours.

$$\begin{aligned}\partial_k E[Y_i^{[k,\rho_1]}] &= \partial_k \{P(Y_i(\rho_0) < k)\mathbb{E}[Y_i(\rho_0)|Y_i(\rho_0) < k]\} + k\partial_k\left(\mathcal{B}^{[k,\rho_1]} - p(k)\right) + \mathcal{B}^{[k,\rho_1]} - p(k) \\ &\qquad + \partial_k \{P(Y_i(\rho_1) > k)\mathbb{E}[Y_i(\rho_1)|Y_i(\rho_1) > k]\} \\ &= \partial_k \int_{-\infty}^{k} y \cdot f_{\rho_0}(y)\cdot dy + k\left(f_0(k) - f_1(k)\right) + \mathcal{B}^{[k,\rho_1]} - p(k) + \partial_k \int_k^\infty y \cdot f_{\rho_1}(y) \cdot dy \\ &= \cancel{kf_0(k)} + \cancel{k\left(f_1(k) - f_0(k)\right)} + \mathcal{B}^{[k,\rho_1]} - p(k) - \cancel{kf_1(k)}\end{aligned}$$

Meanwhile:

$$\begin{aligned}&\partial_{\rho_1} E[Y_i^{[k,\rho_1]}] = k\partial_{\rho_1}\mathcal{B}^{[k,\rho_1]} + \partial_{\rho_1}\{P(Y_i(\rho_1) > k)\mathbb{E}[Y_i(\rho_1)|Y_i(\rho_1) > k]\} \\ &= k\partial_{\rho_1}\mathcal{B}^{[k,\rho_1]} + \int_k^\infty y \cdot \partial_{\rho_1} f_{\rho_1}(y)\cdot dy \\ &= -kf_{\rho_1}(k)\mathbb{E}\left[\left.\frac{Y_i(\rho_1)}{d\rho}\right| Y_i(\rho_1) = k\right] - \int_k^\infty y \cdot \partial_y \left\{f_{\rho_1}(y)\mathbb{E}\left[\left.\frac{dY_i(\rho_1)}{d\rho}\right| Y_i(\rho_1) = y\right]\right\} dy \\ &= \cancel{-kf_{\rho_1}(k)\mathbb{E}\left[\left.\frac{Y_i(\rho_1)}{d\rho}\right| Y_i(\rho_1) = k\right]} + \cancel{yf_{\rho_1}(y)\mathbb{E}\left[\left.\frac{dY_i(\rho_1)}{d\rho}\right| Y_i(\rho_1) = y\right]\Bigg|_\infty^k} \\ &\qquad\qquad - \int_k^\infty f_{\rho_1}(y)\mathbb{E}\left[\left.\frac{dY_i(\rho_1)}{d\rho}\right| Y_i(\rho_1) = y\right] dy\end{aligned}$$

where I have used Theorem A.7.5 and Lemma A.3, and then integration by parts along with the boundary condition that $\lim_{y\to\infty} y \cdot f_{\rho_1}(y) = 0.$, implied by Assumption SMOOTH.

A.8 Secondary proofs

Proof of Proposition A.1

By constant treatment effects, $f_1^G(y) = f_0^G(y+\delta)$ and note that both $f_0^G(k)$ and $f_1^G(k)$ are identified from the data. These can be transformed into densities for Y_{0i} and Y_{1i} via $f_d(y) = G'(y) f_d^G(G(y))$ for $d \in \{0,1\}$. With $f_0(y)$ linear on the interval $[k, k+\Delta]$, the integral $\int_k^{k+\Delta} f_0(y)dy$ evaluates to $\mathcal{B} = \frac{\Delta}{2}\left(f_0(k) + f_0(k+\Delta)\right)$. Although $f_0(k) = \lim_{y\uparrow k} f(y)$ by CONT, $f_0(k+\Delta)$ is not immediately observable. However:

$$f_0(k+\Delta) = f_0\left(G^{-1}\left(G(k)+\delta\right)\right) = G'(k+\Delta) f_0^G(G(k)+\delta)$$

and furthermore by constant treatment effects:

$$f_0^G(G(k)+\delta) = f_1^G(G(k)) = (G'(k))^{-1} f_1(k) = (G'(k))^{-1} \lim_{y\downarrow k} f(y)$$

Combining these equations, we have the result.

Proof of Proposition A.2

We seek a Δ such that for some θ_0:

$$\mathcal{B} = \int_{\tilde{k}}^{k+\Delta} g(y;\theta_0)dy \tag{A.23}$$

and

$$f(y) = \begin{cases} g(y;\theta_0) & y < k \\ g(y+\Delta;\theta_0) & y > k \end{cases} \tag{A.24}$$

and

$$g(y;\theta_0) > 0 \text{ for all } y \in [k, k+\Delta] \tag{A.25}$$

Recall from Equation (A.5) that $\Delta = G^{-1}(G(k)+\delta) - k$ and hence $\delta = G(k+\Delta) - G(k)$. Thus if we find a unique Δ satisfying the two equations, we have found a unique value of δ: the true value of the homogenous effect δ^G.

Suppose we have two candidate values $\Delta' > \Delta$. For them to both satisfy (B.3), we would need $\Delta' = \Delta(\theta')$ and $\Delta = \Delta(\theta)$, where $\theta, \theta' \in \Theta$ and $\Delta(\theta_0)$ is the unique Δ satsifying Eq. (B.3) for a given θ_0, which is unique for each permissible value θ_0 by the positivity condition (B.5). To satisfy (B.4), we would also need

$$g(y;\theta) = \begin{cases} f(y) & y < k \\ f(y-\Delta(\theta)) & y > k+\Delta(\theta) \end{cases} \qquad g(y;\theta') = \begin{cases} f(y) & y < k \\ f(y-\Delta(\theta')) & y > k+\Delta(\theta') \end{cases} \tag{A.26}$$

Since $g(y;\theta)$ is a real analytic function for any $\theta \in \Theta$, the function $h_{\theta\theta'}(y) := g(y;\theta) - g(y;\theta')$ is real analytic. An implication of this is that if $h_{\theta\theta'}(y)$ vanishes on the interval $[0,\tilde{k}]$, as it must by Equation (B.6), it must vanish everywhere on $\mathbb{R}$. Thus for any $y > k+\Delta(\theta)$:

$$g(y+\Delta(\theta')-\Delta(\theta);\theta) = g(y+\Delta(\theta')-\Delta(\theta);\theta') = g(y;\theta)$$

So $g(y;\theta)$ is periodic with period $\Delta(\theta')-\Delta(\theta)$. Since g is non-negative, it cannot integrate to unity globally, and thus cannot be the same function as $f_0(y)$.

Details of calculations for policy estimates

Ex-post evaluation of time-and-a-half after 40

$$\mathbb{E}[Y_{0i}-Y_i] = (\mathcal{B}-p)\mathbb{E}[Y_{0i}-k|Y_i=k, K_i^*=0] + p\cdot 0 + P(Y_{1i}>k)\mathbb{E}[Y_{0i}-Y_{1i}|Y_i>k]$$

Consider the first term

$$(\mathcal{B}-p)E[Y_{0i}-k|Y_i=k,K_i^*=0]=(1-p)\mathcal{B}^*\cdot\frac{1}{\mathcal{B}^*}\int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k)+\mathcal{B}^*}\{Q_{0|K^*=0}(u)-k\}du$$

where $\mathcal{B}^* := P(Y_i = k|K^* = 0) = \frac{\mathcal{B}-p}{1-p}$. Bounds for the rightmost quantity are given by bi-log-concavity of Y_{0i}, just as in Theorem 1.1. In particular:

$$\begin{aligned}(\mathcal{B}-p)E[Y_{0i}-k|Y_i=k,K_i^*=0] &\geq (1-p)\mathcal{B}^*\cdot\frac{F_{0|K^*=0}(k)}{f_{0|K^*=0}(k)(\mathcal{B}^*)}\int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k)+\mathcal{B}^*}\ln\left(\frac{u}{F_{0|K^*=0}(k)}\right)du\\ &=(1-p)\mathcal{B}^*\cdot g(F_{0|K^*=0}(k),f_{0|K^*=0}(k),\mathcal{B}^*)\\ &=(\mathcal{B}-p)\cdot g(F_-,f_-,\mathcal{B}-p)\end{aligned}$$

and

$$\begin{aligned}(\mathcal{B}-p)E[Y_{0i}-k|Y_i=k,K_i^*=0] &\leq -(1-p)\mathcal{B}^*\cdot\frac{1-F_{0|K^*=0}(k)}{f_{0|K^*=0}(k)(\mathcal{B}^*)}\int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k)+\mathcal{B}^*}\ln\left(\frac{1-u}{1-F_{0|K^*=0}(k)}\right)du\\ &=(1-p)\mathcal{B}^*\cdot g'(F_{0|K^*=0}(k),f_{0|K^*=0}(k),\mathcal{B}^*)\\ &=-(\mathcal{B}-p)\cdot g(1-p-F_-,f_+,p-\mathcal{B})\end{aligned}$$

where as before $g(a,b,x)=\frac{a}{bx}\,(a+x)\ln\left(1+\frac{x}{a}\right)-\frac{a}{b}$ and $g'(a,b,x)=-g(1-a,b,-x)$.

Now consider the second term of $\mathbb{E}[Y_{0i}-Y_i]$: $P(Y_{1i}>k)\mathbb{E}[Y_{0i}-Y_{1i}|Y_i>k]$. Taking as a lower bound an assumption of constant treatment effects in levels: $P(Y_{1i}>k)\mathbb{E}[Y_{0i}-Y_{1i}|Y_i>k]\geq P(Y_{1i}>k)\Delta_k^L$.

For an upper bound, we assume that $\mathbb{E}\left[\frac{dY_i(\rho)}{d\rho}\frac{\rho}{Y_i(\rho)}\middle| Y_i(\rho')=y,K_i^*=0\right]=\mathcal{E}$ for all ρ, ρ' and y. Consider then the buncher LATE in logs:

$$\begin{aligned}\mathbb{E}\left[\ln Y_{0i}-\ln Y_{1i}|Y_i=k,K_i^*=0\right] &= \mathbb{E}\left[\ln Y_{0i}-\ln Y_{1i}|Y_{0i}\in[k,Q_{0|K^*=0}(F_{1|K^*=0})],K_i^*=0\right]\\ &=\int_{\rho_0}^{\rho_1}d\rho\cdot\mathbb{E}\left[\frac{dY_i(\rho)}{d\rho}\frac{1}{Y_i(\rho)}\middle|Y_{0i}\in[k,k+\Delta_0^*],K_i^*=0\right]\\ &=\int_{\rho_0}^{\rho_1}d\ln\rho\cdot\frac{1}{\mathcal{B}^*}\int_k^{k+\Delta_0^*}dy\cdot f_0(y)\cdot\mathbb{E}\left[\frac{dY_i(\rho)}{d\rho}\frac{\rho}{Y_i(\rho)}\middle|Y_{0i}=y,K_i^*=0\right]\\ &=\mathcal{E}\int_{\rho_0}^{\rho_1}d\ln\rho=\mathcal{E}\ln(\rho_1/\rho_0)\end{aligned}$$

with the notation that $\Delta_0^* := Q_{0|K^*=0}(F_{1|K^*=0}) - k$. Moreover:

$$\begin{aligned}
\mathbb{E}[Y_{0i} - Y_{1i}|Y_i > k] &= \int_{\rho_0}^{\rho_1} d\rho \cdot \mathbb{E}\left[\left.\frac{dY_i(\rho)}{d\rho}\right| Y_{1i} > k, K_i^* = 0\right] \\
&= P(Y_{1i} > k)^{-1} \int_{\rho_0}^{\rho_1} d\ln\rho \cdot \int_k^{\infty} y \cdot f_1(y) \cdot \mathbb{E}\left[\left.\frac{dY_i(\rho)}{d\rho}\frac{\rho}{Y_i(\rho)}\right| Y_{1i} = y, K_i^* = 0\right] dy \\
&= \mathcal{E} \cdot \mathbb{E}[Y_{1i}|Y_{1i} > k] \int_{\rho_0}^{\rho_1} d\ln\rho = \mathcal{E}\ln(\rho_1/\rho_0) \cdot \mathbb{E}[Y_{1i}|Y_{1i} > k]
\end{aligned}$$

Thus in the isoelastic model

$$E[Y_{0i} - Y_i] = (\mathcal{B} - p)E[Y_{0i} - k|Y_i = k, K_i^* = 0] + \mathbb{E}[Y_{1i}|Y_{1i} > k] \cdot P(Y_{1i} > k)\mathbb{E}\left[\ln Y_{0i} - \ln Y_{1i}|Y_i = k, K_i^* = 0\right]$$

and an upper bound is

$$\delta_k^U \cdot E[Y_i|Y_i > k] - (\mathcal{B} - p) \cdot g(1 - p - F_-, f_+, p - \mathcal{B})$$

where δ_k^U is an upper bound to the buncher LATE in logs $\mathbb{E}\left[\ln Y_{0i} - \ln Y_{1i}|Y_i = k, K_i^* = 0\right]$.

Moving to double time

I make use of the first step deriving the expression for $\partial_{\rho_1} E[Y_i^{[k,\rho_1]}]$ in Theorem 1.2, namely that:

$$\partial_{\rho_1} E[Y_i^{[k,\rho_1]}] = k\partial_{\rho_1}\mathcal{B}^{[k,\rho_1]} + \partial_{\rho_1}\left\{P(Y_i(\rho_1) > k)\mathbb{E}[Y_i(\rho_1)|Y_i(\rho_1) > k]\right\}$$

Thus:

$$
\begin{aligned}
E[Y_i^{[k,\rho_1]}] - E[Y_i^{[k,\bar{\rho}_1]}] &= -\int_{\rho_1}^{\bar{\rho}_1} \partial_\rho E[Y_i^{[k,\rho]}] d\rho = -\int_{\rho_1}^{\bar{\rho}_1} \left\{ k\partial_\rho \mathcal{B}^{[k,\rho]} + \partial_\rho \left\{ P(Y_i(\rho) > k)\mathbb{E}[Y_i(\rho)|Y_i(\rho) > k] \right\} \right\} d\rho \\
&= -k(\mathcal{B}^{[k,\bar{\rho}_1]} - \mathcal{B}^{[k,\rho_1]}) + P(Y_i(\rho_1) > k)\mathbb{E}[Y_i(\rho_1)|Y_i(\rho_1) > k] - P(Y_i(\bar{\rho}_1) > k)\mathbb{E}[Y_i(\bar{\rho}_1)|Y_i(\bar{\rho}_1) > k] \\
&= -k(\mathcal{B}^{[k,\bar{\rho}_1]} - \mathcal{B}^{[k,\rho_1]}) + \{P(Y_i(\rho_1) > k) - P(Y_i(\bar{\rho}_1) > k)\} \cdot \mathbb{E}[Y_i(\bar{\rho}_1)|Y_i(\bar{\rho}_1) > k] \\
&\qquad + P(Y_i(\rho_1) > k) \left(\mathbb{E}[Y_i(\rho_1)|Y_i(\rho_1) > k] - \mathbb{E}[Y_i(\bar{\rho}_1)|Y_i(\bar{\rho}_1) > k] \right) \\
&= \left(\mathbb{E}[Y_{1i}|Y_{1i} > k] - k \right) (\mathcal{B}^{[k,\bar{\rho}_1]} - \mathcal{B}^{[k,\rho_1]}) + P(Y_{1i} > k) \left(\mathbb{E}[Y_{1i}|Y_{1i} > k] - \mathbb{E}[Y_i(\bar{\rho}_1)|Y_i(\bar{\rho}_1) > k] \right) \\
&\leq \left(\mathbb{E}[Y_i(\bar{\rho}_1)|Y_i(\bar{\rho}_1) > k] - k \right) (\mathcal{B}^{[k,\bar{\rho}_1]} - \mathcal{B}^{[k,\rho_1]}) + P(Y_{1i} > k)\mathbb{E}[Y_i(\rho_1) - Y_i(\bar{\rho}_1)|Y_{1i} > k] \\
&\leq \left(\mathbb{E}[Y_i(\bar{\rho}_1)|Y_i(\bar{\rho}_1) > k] - k \right) (\mathcal{B}^{[k,\rho_1]} - p) + P(Y_{1i} > k)\mathbb{E}[Y_i(\rho_1) - Y_i(\bar{\rho}_1)|Y_{1i} > k] \\
&\leq \left(\mathbb{E}[Y_i(\bar{\rho}_1)|Y_i(\bar{\rho}_1) > k] - k \right) (\mathcal{B}^{[k,\rho_1]} - p) + P(Y_{1i} > k)\mathbb{E}[Y_{0i} - Y_{1i}|Y_{1i} > k] \\
&\approx \left(\mathbb{E}[Y_{1i}|Y_{1i} > k] - k \right) (\mathcal{B}^{[k,\rho_1]} - p) + P(Y_{1i} > k)\mathbb{E}[Y_{0i} - Y_{1i}|Y_{1i} > k] \\
&\leq \left(\mathbb{E}[Y_{1i}|Y_{1i} > k] - k \right) (\mathcal{B}^{[k,\rho_1]} - p) + P(Y_{1i} > k)E[Y_i|Y_i > k] \cdot \delta_k^U
\end{aligned}
$$

In the iso-elastic model, making use instead of the final expression for $\partial_{\rho_1} E[Y_i^{[k,\rho_1]}]$ in Theorem 1.2:

$$
\begin{aligned}
&E[Y_i^{[k,\rho_1]}] - E[Y_i^{[k,\bar{\rho}_1]}] = -\int_{\rho_1}^{\bar{\rho}_1} \partial_\rho E[Y_i^{[k,\rho_1]}] d\rho = \int_{\rho_1}^{\bar{\rho}_1} d\rho \int_k^\infty f_\rho(y) \mathbb{E}\left[\left. \frac{dY_i(\rho)}{d\rho} \right| Y_i(\rho) = y \right] dy \\
&= \int_{\rho_1}^{\bar{\rho}_1} d\ln\rho \int_k^\infty f_\rho(y) y \cdot \mathbb{E}\left[\left. \frac{dY_i(\rho)}{d\rho} \frac{\rho}{Y_i(\rho)} \right| Y_i(\rho) = y \right] dy \\
&\geq \mathcal{E} \int_{\rho_1}^{\bar{\rho}_1} d\ln\rho \int_k^\infty f_\rho(y) y \cdot dy \\
&= \mathcal{E} \int_{\rho_1}^{\bar{\rho}_1} d\ln\rho \cdot P(Y_i(\rho) > k)\mathbb{E}[Y_i(\rho)|Y_i(\rho) > k] \\
&\geq \mathcal{E} \ln(\bar{\rho}_1/\rho_1) \cdot P(Y_i(\bar{\rho}_1) > k)\mathbb{E}[Y_i(\bar{\rho}_1)|Y_i(\bar{\rho}_1) > k] \\
&= \mathcal{E} \ln(\bar{\rho}_1/\rho_1) \cdot \{ P(Y_{1i} > k)\mathbb{E}[Y_{1i}|Y_{1i} > k] + (P(Y_i(\bar{\rho}_1) > k)\mathbb{E}[Y_i(\bar{\rho}_1)|Y_i(\bar{\rho}_1) > k] - P(Y_{1i} > k)\mathbb{E}[Y_{1i}|Y_{1i} > k]) \} \\
&= \mathcal{E} \ln(\bar{\rho}_1/\rho_1) \cdot \left\{ P(Y_{1i} > k)\mathbb{E}[Y_{1i}|Y_{1i} > k] - \left(E[Y_i^{[k,\rho_1]}] - E[Y_i^{[k,\bar{\rho}_1]}] \right) + k(\mathcal{B}^{[k,\bar{\rho}_1]} - \mathcal{B}^{[k,\rho_1]}) \right\}
\end{aligned}
$$

where in the fourth step I've used that $Y_i(\rho)$ is decreasing in ρ with probability one, which follows from SEPARABLE and CONVEX. So

$$
\begin{aligned}
E[Y_i^{[k,\rho_1]}] - E[Y_i^{[k,\bar{\rho}_1]}] &\geq \frac{\mathcal{E}\ln(\bar{\rho}_1/\rho_1)}{1 + \mathcal{E}\ln(\bar{\rho}_1/\rho_1)} \cdot \left\{ P(Y_{1i} > k)\mathbb{E}[Y_{1i}|Y_{1i} > k] + k(\mathcal{B}^{[k,\bar{\rho}_1]} - \mathcal{B}^{[k,\rho_1]}) \right\} \\
&\geq \frac{\mathcal{E}\ln(\bar{\rho}_1/\rho_1)}{1 + \mathcal{E}\ln(\bar{\rho}_1/\rho_1)} \cdot P(Y_{1i} > k)\mathbb{E}[Y_{1i}|Y_{1i} > k]
\end{aligned}
$$

Effect of a change to the kink point on bunching

Using that $p(k^*) = p$ and $p(k') = 0$:

$$\begin{aligned}
\mathcal{B}^{[k',\rho_1]} - \mathcal{B}^{[k^*,\rho_1]} &= \left(\mathcal{B}^{[k',\rho_1]} - p(k')\right) - \left(\mathcal{B}^{[k^*,\rho_1]} - p(k^*)\right) - p = -p + \int_{k^*}^{k'} dk \cdot \partial_k \left(\mathcal{B}^{[k',\rho_1]} - p(k)\right) \\
&= -p + \int_{k^*}^{k'} dk \cdot (f_1(k) - f_0(k)) = -p + F_1(k') - F_1(k^*) - F_0(k') + F_0(k^*) \\
&= P(k^* < Y_{1i} \leq k') - P(k^* < Y_{0i} \leq k') - p \\
&= P(k^* < Y_i \leq k') - P(k^* < Y_{0i} \leq k') - p
\end{aligned}$$

if $k' > k^*$.

Similarly, if $k' < k^*$:

$$\begin{aligned}
\mathcal{B}^{[k',\rho_1]} - \mathcal{B}^{[k^*,\rho_1]} &= P(k' \leq Y_{0i} < k^*) - P(k' \leq Y_{1i} < k^*) - p \\
&= P(k' \leq Y_i < k^*) - P(k' \leq Y_{1i} < k^*) - p
\end{aligned}$$

The Lemma in the next section gives identified bounds on the potential outcome probability in either case.

Average effect of a change to the kink point on hours

$$\begin{aligned}
E[Y_i^{[k',\rho_1]}] - E[Y_i^{[k^*,\rho_1]}] &= \int_{k^*}^{k'} \partial_k E[Y_i^{[k,\rho_1]}] dk = \int_{k^*}^{k'} \left\{\mathcal{B}^{[k,\rho_1]} - p(k)\right\} dk \\
&= k\left(\mathcal{B}^{[k,\rho_1]} - p(k)\right)\Big|_{k^*}^{k'} - \int_{k^*}^{k'} k \cdot \partial_k \left\{\mathcal{B}^{[k,\rho_1]} - p(k)\right\} dk \\
&= k'\mathcal{B}^{[k',\rho_1]} - k^*(\mathcal{B} - p) - \int_{k^*}^{k'} y\,(f_1(y) - f_0(y))\,dy \\
&= (k' - k^*)\mathcal{B}^{[k',\rho_1]} + k^*\left(\mathcal{B}^{[k',\rho_1]} - \mathcal{B}\right) + pk^* - \int_{k^*}^{k'} y\,(f_1(y) - f_0(y))\,dy
\end{aligned}$$

For $k' > k^*$, this is equal to

$$
\begin{aligned}
&(k'-k^*)\mathcal{B}^{[k',\rho_1]} + k^*\left(\mathcal{B}^{[k',\rho_1]} - (\mathcal{B}-k)\right) + P(k^* < Y_{0i} \leq k')(\mathbb{E}[Y_{0i}|k^* < Y_{0i} \leq k'] \\
&\qquad\qquad\qquad\qquad\qquad\qquad - P(k^* < Y_{1i} \leq k')(\mathbb{E}[Y_{1i}|k^* < Y_{1i} \leq k'] \\
&= (k'-k^*)\mathcal{B}^{[k',\rho_1]} + P(k^* < Y_{0i} \leq k')(\mathbb{E}[Y_{0i}|k^* < Y_{0i} \leq k'] - k^*) - P(k^* < Y_{1i} \leq k')(\mathbb{E}[Y_{1i}|k^* < Y_{1i} \leq k'] - k^*) \\
&= (k'-k^*)\mathcal{B}^{[k',\rho_1]} + P(k^* < Y_{0i} \leq k')(\mathbb{E}[Y_{0i}|k^* < Y_{0i} \leq k'] - k^*) - P(k^* < Y_{i} \leq k')(\mathbb{E}[Y_{i}|k^* < Y_{i} \leq k'] - k^*)
\end{aligned}
$$

The first term represents the mechanical effect from the bunching mass under k' being transported from k^* to k', and can be bounded given the bounds for $\mathcal{B}^{[k',\rho_1]} - \mathcal{B}^{[k^*,\rho_1]}$ in the last section. The last term is point identified from the data, while the middle term can be bounded using bi-log concavity of Y_{0i} conditional on $K^* = 0$. Similarly, when $k' < k^*$, the effect on hours is:

$$
\begin{aligned}
&(k'-k^*)\mathcal{B}^{[k',\rho_1]} + P(k' \leq Y_{0i} < k^*)(k^* - \mathbb{E}[Y_{0i}|k' \leq Y_{0i} < k^*]) - P(k' \leq Y_{1i} < k^*)(k^* - \mathbb{E}[Y_{1i}|k' \leq Y_{1i} < k^*]) \\
&= (k'-k^*)\mathcal{B}^{[k',\rho_1]} + P(k' \leq Y_{i} < k^*)(k^* - \mathbb{E}[Y_{i}|k' \leq Y_{i} < k^*]) - P(k' \leq Y_{1i} < k^*)(k^* - \mathbb{E}[Y_{1i}|k' \leq Y_{1i} < k^*])
\end{aligned}
$$

with the middle term point identified from the data and last term bounded by bi-log concavity of Y_{1i} conditional on $K^* = 0$. The analytic bounds implied by BLC in each case are given by the Lemma below.

Lemma. *Suppose Y_i is a bi-log concave random variable with CDF $F(y)$. Let $F_0 := F(y_0)$ and $f_0 = f(y_0)$ be the CDF and density, respectively, evaluated at a fixed y_0.*

For any $y' > y_0$:

$$A \leq P(y_0 \leq Y_i \leq y')\left(\mathbb{E}[Y_i|y_0 \leq Y_i \leq y'] - y_0\right) \leq B$$

and for any $y' < y_0$:

$$B \leq P(y' \leq Y_i \leq y_0)\left(y_0 - \mathbb{E}[Y_i|y' \leq Y_i \leq y_0]\right) \leq A$$

where $A = g(F_0, f_0, F_L(y'))$ *and* $B = g(1 - F_0, f_0, 1 - F_U(y'))$*, with*

$$F_L(y') = 1 - (1 - F_0)e^{-\frac{f_0}{1-F_0}(y-y_0)}, \qquad\qquad F_U(y') = F_0 e^{\frac{f_0}{F_0}(y'-y_0)}$$

and

$$g(a,b,c) = \begin{cases} \frac{ac}{b}\left(\ln\left(\frac{c}{a}\right) - 1\right) + \frac{a^2}{b} & \text{if } c > 0 \\ \frac{a^2}{b} & \text{if } c \leq 0 \end{cases}$$

In either of the two cases $\max\{0, F_L(y')\} \leq F(y') \leq \min\{1, F_U(y')\}$.

Proof. As shown by Dümbgen et al., 2017, bi-log concavity of Y_i implies not only that $f(y)$ exists, but that it is strictly positive, and we may then define a quantile function $Q = F^{-1}$ such that $Q(F(y)) = y$ and $y = Q(F(y))$. Theorem 1 of Dümbgen et al., 2017 also shows that for any y':

$$\underbrace{1 - (1 - F_0)e^{-\frac{f_0}{1-F_0}(y-y_0)}}_{:=F_L(y')} \leq F(y') \leq \underbrace{F_0 e^{\frac{f_0}{F_0}(y'-y_0)}}_{:=F_U(y')}$$

We can re-express this as bounds on the quantile function evaluated at any $u' \in [0,1]$:

$$\underbrace{y_0 + \frac{F_0}{f_0}\ln\left(\frac{u}{F_0}\right)}_{Q_L(u')} \leq Q(u') \leq \underbrace{y_0 - \frac{1-F_0}{f_0}\ln\left(\frac{1-u}{1-F_0}\right)}_{Q_U(u')}$$

Write the quantity of interest as:

$$P(y_0 \leq Y_i \leq y')\left(\mathbb{E}[Y_i | y_0 \leq Y_i \leq y'] - y_0\right) = \int_{y_0}^{y'} (y - y_0) f(y) dy = \int_{F_0}^{F(y')} (Q(u) - y_0) du$$

Given that $Q(u) \geq y_0$, the integral is increasing in $F(y')$. Thus an upper bound is:

$$
\begin{aligned}
P(y_0 \leq Y_i \leq y')\left(\mathbb{E}[Y_i|y_0 \leq Y_i \leq y'] - y_0\right) &\leq \int_{F_0}^{F_U(y')} (Q_U(u) - y_0)du \\
&= -\frac{1-F_0}{f_0}\int_{F_0}^{F_U(y')} \ln\left(\frac{1-u}{1-F_0}\right) du \\
&= \frac{(1-F_0)^2}{f_0}\int_1^{\frac{1-F_U(y')}{1-F_0}} \ln(v)\, dv \\
&= \frac{(1-F_0)(1-F_U(y'))}{f_0}\left(\ln\left(\frac{1-F_U(y')}{1-F_0}\right) - 1\right) + \frac{(1-F_0)^2}{f_0}
\end{aligned}
$$

where we've made the substitution $v = \frac{1-u}{1-F_0}$ and used that $\int \ln(v)dv = v(\ln(v)-1))$. Inspection of the formulas for F_U and F_L reveal that $F_U \in (0, \infty)$ and $F_L \in (-\infty, 1)$. In the event that $F_U(y') \geq 1$, the above expression is undefined but we can replace $F_U(y')$ with one and still obtain valid bounds:

$$
P(y_0 \leq Y_i \leq y')\left(\mathbb{E}[Y_i|y_0 \leq Y_i \leq y'] - y_0\right) \leq -\frac{(1-F_0)^2}{f_0}\int_0^1 \ln(v)\, dv = \frac{(1-F_0)^2}{f_0}
$$

where we've used that $\int_0^1 \ln(v)dv = -1$.

Similarly, a lower bound is:

$$
\begin{aligned}
P(y_0 \leq Y_i \leq y')\left(\mathbb{E}[Y_i|y_0 \leq Y_i \leq y'] - y_0\right) &\geq \int_{F_0}^{F_L(y')} (Q_L(u) - y_0)du = \frac{F_0}{f_0}\int_{F_0}^{F_L(y')} \ln\left(\frac{u}{F_0}\right) du \\
&= \frac{F_0^2}{f_0}\int_1^{F_L(y')/F_0} \ln(v)\, du \\
&= \frac{F_0 F_L(y')}{f_0}\left(\ln\left(\frac{F_L(y')}{F_0}\right) - 1\right) + \frac{F_0^2}{f_0}
\end{aligned}
$$

where we've made the substitution $v = \frac{u}{F_0}$. If $F_L(y') \leq 0$, then we replace with zero to obtain

$$
P(y_0 \leq Y_i \leq y')\left(\mathbb{E}[Y_i|y_0 \leq Y_i \leq y'] - y_0\right) \geq -\frac{F_0^2}{f_0}\int_0 1 \ln(v)\, du = \frac{F_0^2}{f_0}
$$

When $y' < y$, write the quantity of interest as:

$$P(y' \leq Y_i \leq y_0)\left(y_0 - \mathbb{E}[Y_i|y' \leq Y_i \leq y_0]\right) = \int_{y'}^{y_0} (y_0 - y) f(y) dy = \int_{F(y')}^{F_0} (y_0 - Q(u)) du$$

This integral is decreasing in $F(y')$, so an upper bound is:

$$\begin{aligned} P(y' \leq Y_i \leq y_0)\left(y_0 - \mathbb{E}[Y_i|y' \leq Y_i \leq y_0]\right) &\leq \int_{F_L(y')}^{F_0} (y_0 - Q_L(u)) du = -\frac{F_0}{f_0} \int_{F_L(y')}^{F_0} \ln\left(\frac{u}{F_0}\right) du \\ &= -\frac{F_0^2}{f_0} \int_{F_L(y')/F_0}^{1} \ln(v)\, du \\ &= \frac{F_0 F_L(y')}{f_0}\left(\ln\left(\frac{F_L(y')}{F_0}\right) - 1\right) + \frac{F_0^2}{f_0} \end{aligned}$$

or simply F_0^2/f_0 when $F_L(y') \leq 0$, and a lower bound is:

$$\begin{aligned} P(y' \leq Y_i \leq y_0)\left(y_0 - \mathbb{E}[Y_i|y' \leq Y_i \leq y_0]\right) &\geq \int_{F_U(y')}^{F_0} (y_0 - Q_U(u)) du \\ &= \frac{1 - F_0}{f_0} \int_{F_U(y')}^{F_0} \ln\left(\frac{1-u}{1-F_0}\right) du \\ &= -\frac{(1-F_0)^2}{f_0} \int_{\frac{1-F_U(y')}{1-F_0}}^{1} \ln(v)\, dv \\ &= \frac{(1-F_0)(1-F_U(y'))}{f_0}\left(\ln\left(\frac{1-F_U(y')}{1-F_0}\right) - 1\right) + \frac{(1-F_0)^2}{f_0} \end{aligned}$$

or simply $(1-F_0)^2/f_0$ when $F_U(y') \geq 1$. □

Wage correction terms

For the ex-post effect of the kink

Suppose that straight-time wages w^* are set according to Equation (1.1) for all workers, where h^* are their anticipated hours. The straight-wages that would exist absent the FLSA w_0^*, yield the same total earnings z^*, so:

$$w_0^* h^* = w^*(h^* + (\rho_1 - 1)(h^* - k)\mathbb{1}(h^* > k))$$

where $k = 40$ and $\rho_1 = 1.5$. The percentage change is thus

$$(w_0^* - w^*)/w^* = \frac{(\rho_1 - 1)(h^* - k)\mathbb{1}(h^* > k)}{h^* + (\rho_1 - 1)(h^* - k)\mathbb{1}(h^* > k)}$$

If h_{0i} is constant elasticity in the wage with elasticity $\mathcal{E}$, then we would expect

$$\frac{h_{0it} - h_{0it}^*}{h_{0it}} = 1 - \left(1 + \frac{(\rho_1 - 1)(h^* - k)\mathbb{1}(h^* > k)}{h^* + (\rho_1 - 1)(h^* - k)\mathbb{1}(h^* > k)}\right)^{\mathcal{E}}$$

Taking $h_{0it} \approx h_{1it} \approx h^*$ and integrating along the distribution of h_{1it}, we have:

$$\mathbb{E}[h_{0it} - h_{0it}^*] \approx \mathbb{E}\left[\mathbb{1}(h_{it} > k)h_{it}\left(1 - \left(1 + \frac{(\rho_1 - 1)(h_{it} - k)}{h_{it} + (\rho_1 - 1)(h_{it} - k)}\right)^{\mathcal{E}}\right)]\right]$$

which will be negative provided that $\mathcal{E} < 0$.The total ex-post effect of the kink is:

$$\mathbb{E}[h_{it} - h_{0it}^*] = \mathbb{E}[h_{it} - h_{0it}] + \mathbb{E}[h_{0it} - h_{0it}^*]$$

For a move to double-time

The straight-wages w_2^* that would exist with double time, for workers with $h^* > k$, that yield the same total earnings z^* as the actual straight wages w^* satisfy:

$$w_2^*(k + (\bar{\rho}_1 - 1)(h^* - k)) = w^*(k + (\rho_1 - 1)(h^* - k))$$

where $\bar{\rho}_1 = 2$. The percentage change is thus

$$(w_2^* - w^*)/w^* = \frac{k + (\rho_1 - 1)(h^* - k)}{k + (\bar{\rho}_1 - 1)(h^* - k)} - 1$$

Let $\bar{h}_{0i}$ be hours under a straight-time wage of w_2^*. By a similar calculation thus:

$$\mathbb{E}[\bar{h}_i^{[\bar{\rho}_1,k]} - h_{it}^{[\bar{\rho}_1,k]}] \approx \mathbb{E}\left[\mathbb{1}(h_{it} > k)h_{it}\left(\left(\frac{k + (\rho_1 - 1)(h^* - k)}{k + (\bar{\rho}_1 - 1)(h^* - k)}\right)^{\mathcal{E}} - 1\right)]\right]$$

The total effect of a move to double-time is:

$$\mathbb{E}[\bar{h}_{it}^{[\bar{\rho}_1,k]} - h_{it}] = \mathbb{E}[\bar{h}_{it}^{[\bar{\rho}_1,k]} - h_{it}^{[\bar{\rho}_1,k]}] + \mathbb{E}[h_{it}^{[\bar{\rho}_1,k]} - h_{it}]$$

The above definitions are depicted visually in Figure A.19 below.

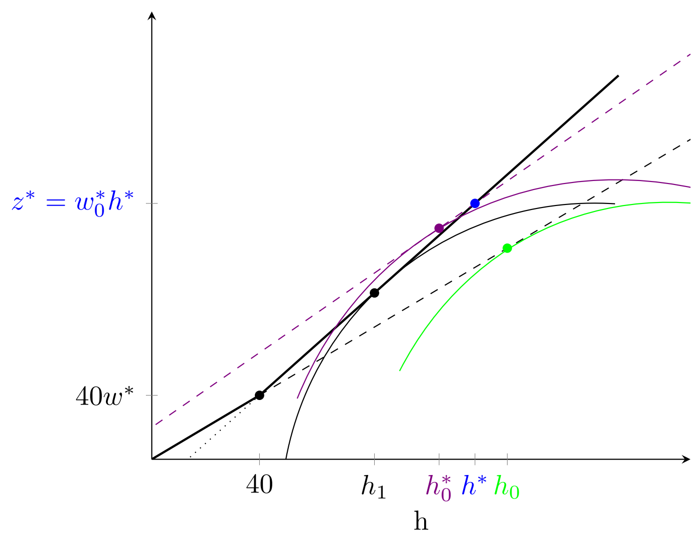

Figure A.19: Depiction of h^*, h_0, h_0^* and h_1 for a single fixed worker that works overtime at h_1 hours this week. Their realized wage w^* has been set to yield earnings z^* based on anticipated hours h^* given the FLSA kink. In a world without the FLSA, the worker's wage would instead be $w_0^* = z^*/h^*$, and this week the firm would have chosen h_0^* hours, where the worker's marginal productivity this week is w_0^* (in the benchmark model). *Note:* while (z^*, h^*) is chosen jointly with employment and on the basis of anticipated productivity, choice of h_0^* is instead constrained by the contracted purple pay schedule (with the worker already hired) and on the basis of updated productivity. h_1 may differ from h^* for this same reason. In the numerical calculation h^* is approximated by h_1 – which corresponds to productivity variation being small and h^* being a credible choice given the FLSA. If credibility (the firm not wanting to renege too far on hours after hiring) were a constraint on the choice of (z^*, h^*) in the no-FLSA counterfactual, then h^* would be smaller without the FLSA, but I consider this "second-order" and do not attempt a correction here.

Changing the location of the kink

Let $\mathcal{B}_w^{[k]}$ denote bunching with the kink at location k and (a distribution of) wages denoted by w. Then the effect of moving k on bunching is

$$\mathcal{B}_{w'}^{[k']} - \mathcal{B}_w^{[k^*]} = \left(\mathcal{B}_w^{[k']} - \mathcal{B}_w^{[k^*]}\right) + \left(\mathcal{B}_{w'}^{[k']} - \mathcal{B}_w^{[k']}\right)$$

where w' are the wages that would occur with bunching at the new kink point k'. The first term has been estimated by the methods described above, with the second term representing a correction due to wage adjustment. Taking $Y_{0i} \approx Y_{1i} \approx h^*$, the straight-time wages w^* set according to Equation (1.1) that would change are those between k' and k^*. Consider the case $k' < k^*$. We expect wages to fall, as the overtime policy becomes more stringent, and $\left(\mathcal{B}_{w'}^{[k']} - \mathcal{B}_w^{[k']}\right)$ is only nonzero to the extent that the increase in Y_0 and Y_1 changes the mass of each in the range $[k', k^*]$. With the range $[k', k^*]$ to the left of the mode of Y_{0i}, it is most plausible that this mass will decrease. Similarly, for Y_{1i}, it is most likely that this mass will decrease, making the overall sign of $\left(\mathcal{B}_{w'}^{[k']} - \mathcal{B}_w^{[k']}\right)$ ambiguous However, since most of the adjustment should occur for workers who are typically found between k and k', we would not expect either term to be very different from zero.

Now consider the effect on average hours:

$$\mathbb{E}[Y_{w'}^{[k']} - Y_w^{[k^*]}] = \mathbb{E}[Y_w^{[k']} - Y_w^{[k^*]}] + \mathbb{E}[Y_{w'}^{[k']} - Y_w^{[k']}]$$

For a reduction in k, we would expect wages w' to be lower with $k = k'$ and hence the second term positive. This will attenuate the effects that are bounded by the methods above, holding the wages fixed at their realized levels.

Consider first the case of $k' < k^*$. Let w' be wages under the new kink point k', and assuming they adjust to keep total earnings z^* constant, wages w' will change if w^* is between k and k' as:

$$w'(k' + 0.5(h^* - k')) = w^* h^*$$

And the percentage change for these workers is thus

$$(w' - w^*)/w^* = \frac{h^*}{k' + 0.5(h^* - k')} - 1$$

$$\mathbb{E}[Y_{w'}^{[k']} - Y_{w}^{[k']}] \approx \mathbb{E}\left[\mathbb{1}(k' < Y_i < k^*)Y_i \left(\left(\frac{Y_i}{k' + 0.5(Y_i - k')}\right)^{\mathcal{E}} - 1\right)]\right]$$

In the case of $k' > k^*$, we will have wages change as:

$$w' h^* = w^*(k^* + 0.5(h^* - k^*))$$

w^* is between k and k'. And the percentage change for these workers is thus

$$(w' - w^*)/w^* = \frac{k^* + 0.5(h^* - k^*)}{h^*} - 1$$

$$\mathbb{E}[Y_{w'}^{[k']} - Y_{w}^{[k']}] \approx \mathbb{E}\left[\mathbb{1}(k^* < Y_i < k')Y_i \left(\left(\frac{k^* + 0.5(Y_i - k^*)}{Y_i}\right)^{\mathcal{E}} - 1\right)\right]$$

Appendix B: Supplements to Chapter 2

B.1 Identification result without rectangular support

This section provides an extension of Theorem 1 for cases when the support $\mathcal{Z}$ of the instruments is not rectangular (i.e. $supp(Z_i) \neq (\mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_J)$), and there may be perfect linear dependencies between the instruments (of the form that would arise from the mapping from discrete to binary instruments presented in Section 2.3.3).

A weaker version of Assumption 3 is comprised of the following two conditions, with the definition that $Z_{\emptyset i}$ is a degenerate random variable that takes the value of one with probability one:

Assumption 3a* (existence of instruments). *There exists a family $\mathcal{F}$ of subsets of the instruments $S \subseteq \{1 \dots J\}$, where $\emptyset \in \mathcal{F}$ and $|\mathcal{F}| > 1$, such that random variables Z_{Si} for all $S \in \mathcal{F}$ are linearly independent, i.e. $P\left(\sum_{S \in \mathcal{F}} \omega_S Z_{Si} = 0\right) < 1$ for all vectors $\omega \in \mathbb{R}^{|\mathcal{F}|}/\mathbf{0}$.*

Assumption 3b* (non-degenerate subsets generate the response groups). *There exists a family $\mathcal{F}$ satisfying Assumption 3a*, such that for any $S \notin \mathcal{F}$, $g(F) \notin \mathcal{G}$ for all Sperner families F that contain S.*

Assumption 3a* is in itself very weak, requiring only that there exists some product of the instruments that has strictly positive variance. Assumption 3b* is much more restrictive: it says that all response groups aside from never-takers can be generated from members of a family of linearly independent subsets of the instruments.

The construction in Proposition 2.4 mapping discrete instruments to binary instruments yields a case where Assumption 3* will hold, given rectangular support of the original discrete instruments.

Proposition. *Let each Z_j have M_j ordered points of support $z_1^j < z_2^j \cdots < z_{M_j}^j$ and let $\tilde{Z}_m^j = \mathbb{1}(Z_{ji} \geq z_m^j)$. If $P(Z_i = z) > 0$ for $z \in (\mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_J)$, then Assumption 3* holds with $\mathcal{F}$ the family of all subsets of $\mathcal{M} := \{\tilde{Z}_m^j\}_{\substack{j\in\{1\dots J\} \\ m=2\dots m_j}}$ containing at most one Z_m^j for any given $j \in \{1 \dots J\}$.*

Proof. See Appendix B.4. □

The above proposition allows us to make use of Assumption 3* in cases where discrete instruments are mapped to binary instruments via Proposition 2.4. To illustrate, consider a case with a single discrete instrument Z_1 having three levels $z_1 < z_2 < z_3$ and instruments $2 - J$ binary. Proposition 2.4 shows that if $Z_1 \dots Z_J$ satisfies VM then so does the set of $J+1$ instruments $\tilde{Z}_2, \tilde{Z}_3, Z_2, \dots Z_J$ where $\tilde{Z}_2 = \mathbb{1}(Z_1 \geq z_2)$ and $\tilde{Z}_3 = \mathbb{1}(Z_1 \geq z_3)$. In this case there are 2^{J-1} "redundant" simple response groups vis-a-vis Assumption 3, since for any $S \subseteq \{2 \dots J\}$: $\tilde{Z}_{2i}\tilde{Z}_{3i}Z_{Si} = \tilde{Z}_{3i}Z_{Si}$.

In this example, the vector Γ_i would contain all non-null subsets of $\{\tilde{Z}_2, \tilde{Z}_3, Z_2, \dots Z_J\}$ that do not contain both of $\tilde{Z}_2$ and $\tilde{Z}_3$. In general, $\mathcal{F}$ can be constructed by considering all subsets of the instruments, and for each subset considering all possible assignments of a value to each instrument, with one fixed value for each instrument omitted from consideration throughout. Provided rectangular support on the original instruments, Assumption 3* then follows with this choice of $\mathcal{F}$, for which a generalized version of Theorem 2.1 can be stated:

Theorem 1*. *The results of Theorem 2.1 holds under Assumption 3* replacing Assumption 3, where now $\Gamma_i := \{Z_{Si}\}_{S\in\mathcal{F},S\neq\emptyset}$, $\lambda := \{\mathbb{E}[c(S), Z_i)]\}_{S\in\mathcal{F},S\neq\emptyset}$ and again $h(Z_i) = \lambda'\Sigma^{-1}(\Gamma_i - \mathbb{E}[\Gamma_i])$ with $\Sigma := Var(\Gamma_i)$, for any family $\mathcal{F}$ satisfying Assumption 3*.*

Proof. Identical to that of Theorem 2.1, except as noted therein. □

Theorem 1* may also be useful in other cases in which the practitioner has auxiliary knowledge that some of the response groups are not present in the population. In such cases, Assumption 3* may hold even without rectangular support among the instruments.

B.2 Identification with covariates

This section discusses how one can accommodate, in a nonparametric way, covariates that need to be conditioned on for the instruments to be valid. In practice, it is often easier to justify a conditional version of Assumption 1:

$$\{(Y_i(1), Y_i(0), G_i) \perp Z_i\} \,|X_i$$

where X are a set of observed covariates unaffected by treatment. In this section I discuss identification and considerations for estimation in such a setting. I maintain that vector monotonicity continues to hold for a set of binary instruments, as VM is expressed in Assumption 2. This implies that the direction of response is the same regardless of X_i, since the condition in Assumption 2 holds with probability one.

If Assumption 3 and Property M each hold conditional on $X_i = x$, then Theorem 2.1 implies that we can identify $\Delta_c(x) := \mathbb{E}[\Delta_i|C_i = 1, X_i = x]$ for Δ_c satisfying Property M, from the distribution of $(Y_i, Z_i, D_i)|X_i = x$. In particular, the function $h(z)$ from Theorem 2.1 will now depend on the conditioning value of X_i:

$$h(Z_i, x) = \lambda(x)' Var(\Gamma_i|X_i = x)^{-1} \left(\Gamma_i - \mathbb{E}[\Gamma_i|X_i = x]\right)$$

for each $x \in \mathbb{X}$, where recall that Γ_i is a vector of products Γ_{Si} of Z_{ji} within subsets of the instruments, where S indexes such subsets. Here we define $\lambda(x)_S = \mathbb{E}[c(g(S), Z_i)|X_i = x]$ – which is identified – for each simple response group $g(S)$. Under these assumptions, we have that $\Delta_c(x) = \mathbb{E}[h(Z_i, x)Y_i|X_i = x]/\mathbb{E}[h(Z_i, x)D_i|X_i = x]$.

If the support of X_i corresponds to a small number of "covariate-cells", it might be feasible to repeat the entire estimation on fixed-covariate subsamples, to estimate $\Delta_c(x)$ for each $x \in \mathbb{X}$. If the number of groups is large, or if X_i includes continuous variables, estimation of $\Delta_c(x)$ could still in principle be implemented by nonparametric regression of each component of Γ_i on X_i as well as nonparametrically estimating the conditional variance-covariance matrix $Var(\Gamma_i|X_i = x)$ (Yin et al. (2010) describe a kernel-based method for this). The vector $\lambda(x)$ can also be computed via nonparametric regression.

Furthermore, when the object of interest is simply the unconditional version of Δ_c, the conditional quantities become nuisance parameters. Notably, they can be integrated over separately in the numerator and the denominator of the empirical estimand. To see that this, write:

$$\begin{aligned}\Delta_c &= \mathbb{E}[\Delta_i|C_i = 1] = \int dF_{X|C}(x|1)\Delta_c(x) \\ &= \int dF_{X|C}(x|1)\frac{\mathbb{E}[h(Z_i,x)Y_i|X_i = x]}{\mathbb{E}[h(Z_i,x)Y_i|X_i = x]} = \int dF_{X|C}(x|1)\frac{\mathbb{E}[h(Z_i,x)Y_i|X_i = x]}{P(C_i = 1|X_i = x)} \\ &= \frac{1}{P(C_i = 1)}\int dF_X(x)\mathbb{E}[h(Z_i,X_i)Y_i|X_i = x] = \frac{\mathbb{E}[h(Z_i,X_i)Y_i]}{\mathbb{E}[h(Z_i,X_i)D_i]}\end{aligned}$$

where we have used Bayes' rule and that $P(C_i = 1|X_i = x) = \mathbb{E}[h(Z_i,x)D_i|X_i = x]$ (and hence $P(C_i = 1) = \mathbb{E}[h(Z_i,X_i)D_i]$ as well). This provides a VM analog to a similar result that holds under IAM. In that context, Frölich (2007) shows that this fact can deliver $\sqrt{n}$-consistency of a nonparametric analog of the Wald ratio.

Note that by the conditional version of Corollary 2.1 we have that:

$$\Delta_c = \frac{\mathbb{E}[\tilde{\lambda}(X_i)'A\{\mathbb{E}[Y_i|Z_i = z, X_i]\}]}{\mathbb{E}[\tilde{\lambda}(X_i)'A\{\mathbb{E}[D_i|Z_i = z, X_i]\}]}$$

if we define $\tilde{\lambda}(x)$ to have component $\lambda(x)$ for any $S \subseteq \{1 \dots J\}, S \neq \emptyset$ and 0 for $S = \emptyset$, and we let $\{\cdot\}$ indicate vector representations of functions over $z \in \mathcal{Z}$. If the CEFs of Y and D happen to both be separable between Z and X, i.e $\mathbb{E}[Y_i|Z_i = z, X_i = x] = y(z) + w(x)$ and $\mathbb{E}[D_i|Z_i = z, X_i = x] = d(z) + v(x)$, then the expression simplifies:

$$\Delta_c = \frac{\mathbb{E}[\tilde{\lambda}(X_i)'A\{y(z)\} + w(X_i)\tilde{\lambda}(X_i)'A\mathbf{1}]}{\mathbb{E}[\tilde{\lambda}(X_i)'A\{d(z)\} + v(X_i)\tilde{\lambda}(X_i)'A\mathbf{1}]} = \frac{\mathbb{E}[\tilde{\lambda}(X_i)'A\{y(z)\}]}{\mathbb{E}[\tilde{\lambda}(X_i)'A\{d(z)\}]}$$

where $\mathbf{1}$ is a vector of ones and we have used that $\tilde{\lambda}(x)'A\mathbf{1} = 0$ for any x. This follows from the definition of the entries: $A_{S,z} = \sum_{\substack{f \subseteq z_0 \\ (z_1 \cup f) = S}} (-1)^{|f|}$ where z_0 is the set of components of z that are equal to zero. For any $S \neq \emptyset$,the identity $\sum_{f \subseteq S} (-1)^{|f|} = 0$ implies that $[A\mathbf{1}]_S = \sum_{z_1 \subseteq S} \sum_{f \subseteq (S - z_1)} (-1)^{|f|} = 0$. The first component of $A\mathbf{1}$, corresponding to $S = \emptyset$, does not contribute since the first component of $\tilde{\lambda}(x)$ is always zero, by construction.

Now, since each $\lambda_S(x)$ is defined as $\mathbb{E}[C_i = 1 | G_i = g(S), X_i = x]$, its expectation delivers the unconditional analog: $\lambda_S := \mathbb{E}[C_i = 1 | G_i = g(S)] = \mathbb{E}[\lambda(X_i)_S]$. Thus we can write $\Delta_c = \frac{\lambda' A \{y(z)\}}{\lambda' A \{d(z)\}}$. This shows that in this separable case the estimand that identifies Δ_c is essentially unchanged from the baseline case without covariates, aside from the need to control semiparametrically for X_i to obtain the functions $y(z)$ and $d(z)$. The estimates reported in Section 2.6 use this result, with $w(x)$ and $v(x)$ taken to each be linear.

B.3 Regularization and asymptotic distribution

In this section I propose a regularization procedure for the estimator, to improve its performance in small samples. I then show asymptotically normality of the regularized estimator and give an expression for the variance, based on a result from Imbens and Angrist, 1994.

B.3.1 Regularization of the estimator

Recall from Section 2.5 that the simple plug-in estimator of the all-compliers LATE in fact only uses data at two points in $\mathcal{Z}$. This issue can be seen as a near collinearity problem: when there are few observations in the points $\bar{Z}$ and $\underline{Z}$, the $n \times |\mathcal{F}|$ design matrix Γ will have singular values that are close to zero (to see this, note that $\Gamma'\Gamma = A'^{-1} n \cdot diag\{\hat{P}(Z_i = z)\} A^{-1}$). This observation suggests that the issue might be mitigated by employing a ridge-type shrinkage estimator (see e.g. Hoerl and Kennard, 1970). Accordingly, we allow a sequence of regularization parameters α_n:

$$\hat{\rho}(\hat{\lambda}, \alpha) = \left((0, \hat{\lambda}')(\Gamma'\Gamma + \alpha I)^{-1} \Gamma' D\right)^{-1} (0, \hat{\lambda}')(\Gamma'\Gamma + \alpha I)^{-1} \Gamma' Y \tag{B.1}$$

The estimator $\hat{\rho}(\hat{\lambda}, \alpha)$ with a choice of $\alpha > 0$ establishes a floor on the singular values of the matrix Γ.

In the case of the ACL, Corollary 2.1 can be leveraged to show that $\alpha > 0$ allows the estimator to make use of the full support of Z_i, rather than just the two points $\bar{Z}$ and $\underline{Z}$. But ridge regression comes at the expense of some bias. Proposition B.1 below yields a means of navigating this trade-off to choose α in practice. In particular, I propose choosing α to minimize a feasible estimator of the conditional MSE $\mathbb{E}[(\hat{\rho}(\lambda, \alpha) - \Delta_c)^2 | Z_1 \dots Z_n]$.

Proposition B.1. *Under the assumptions of Theorem 2.1, $\mathbb{E}[(\hat{\rho}(\lambda, \alpha) - \Delta_c)^2 | Z_1 \dots Z_n]$ is, up to second order in estimation error and a positive constant of proportionality:*

$$\tilde{\lambda}'(\Gamma'\Gamma + \alpha I)^{-1} \left\{ \Gamma'(\Omega_Y + \Delta_c^2 \Omega_D - 2\Delta_c \Omega_{YD})\Gamma \right.$$
$$\left. + \alpha^2(\beta_Y \beta_Y' + \Delta_c^2 \beta_D \beta_D' - 2\Delta_c \beta_Y \beta_D') \right\} (\Gamma'\Gamma + \alpha I)^{-1} \tilde{\lambda} \quad \text{(B.2)}$$

where $\tilde{\lambda} := (0, \lambda')'$, $\beta_Y := \mathbb{E}[\Gamma_i \Gamma_i']^{-1} \mathbb{E}[\Gamma_i Y_i]$, $\beta_D := \mathbb{E}[\Gamma_i \Gamma_i']^{-1} \mathbb{E}[\Gamma_i D_i]$, and $\Omega_{VW} = \mathbb{E}[(V - \beta_V \Gamma)(W - \beta_W \Gamma)' | \Gamma]$ for $V, W \in \{Y, D\}$, and all expectations are assumed to exist.

Furthermore, if $\hat{\alpha}_{mse}$ is chosen as the smallest positive local minimizer of the following estimate of the above:

$$\hat{M}(\alpha) := (0, \hat{\lambda}')(\Gamma'\Gamma + \alpha I)^{-1} \left\{ n\hat{\Pi} + \alpha^2(\hat{\beta}\hat{\beta}') \right\} (\Gamma'\Gamma + \alpha I)^{-1} (0, \hat{\lambda}')'$$

with $\hat{\beta}_V := (\Gamma'\Gamma)^{-1}\Gamma' V$ for each $V \in \{Y, D\}$, $\hat{\Pi} := \frac{1}{n}\sum_i (Y_i - \hat{\beta}_Y \Gamma_i - \frac{(0,\hat{\lambda}')\hat{\beta}_Y}{(0,\hat{\lambda}')\hat{\beta}_D}(D_i - \hat{\beta}_D \Gamma_i))^2 \Gamma_i \Gamma_i'$ and $\hat{\beta} := \hat{\beta}_Y - \frac{(0,\hat{\lambda}')\hat{\beta}_Y}{(0,\hat{\lambda}')\hat{\beta}_D} \hat{\beta}_D$ then

$$\hat{\alpha}_{mse} / \sqrt{n} \xrightarrow{p} 0$$

provided that $\hat{\lambda} \xrightarrow{p} \lambda$, $(0, \lambda')\Sigma^{-1}(\beta_Y + \Delta_c \beta_D) \neq 0$.

Proof. See Appendix B.4. □

The proposed data-driven choice $\hat{\alpha}_{mse}$ estimates the unknown quantities in Eq. (B.2) based on an initial guess of $\alpha = 0$, and then minimizes with respect to α. This can be seen as a "one-step" version of a more general iterative algorithm in which a value α_t is used to compute the function $\hat{M}(\alpha)$, which is then minimized to find α_{t+1} and so on until convergence. I implement the single-step version in Appendix B.3, and find that it indeed improves estimation error considerably for the simulation DGPs considered.

The reason that my proposed rule evaluates $\hat{\alpha}_{mse}$ as a local minimizer of $\hat{M}(\alpha)$ rather than a global minimizer, is that the function $\hat{M}(\alpha)$ is always positive but approaches zero as $\alpha \to \infty$. This stands in contrast with the standard case of ridge regression in which regularization bias always grows with α, eventually dominating any efficiency gains from increasing it further. In the present case, the vector $\hat{\beta}$ as defined above and $(0, \hat{\lambda}')'$ are orthogonal (in sample as well as in the population limit), and thus the "(squared) bias" term vanishes as $\alpha \to \infty$, along with the variance of the regularized estimator (this is roughly analogous to ridge regularizing a vector of regression coefficients when their true values are all zero). Nevertheless, the function $\hat{M}(\alpha)$ does have a well-defined local minimum that achieves a lower value than $\hat{M}(0)$ at some strictly positive α (see Appendix B.4 for details), and this local minimum is shown to provide a helpful guide to choosing α in the simulations of Appendix B.3. Note that the condition $(0, \lambda')\Sigma^{-1}(\beta_Y + \Delta_c \beta_D) \neq 0$ in Proposition B.1 rules out a knife-edge case in which the Hessian of $\hat{M}(\alpha)$ is zero when the other arguments of $\hat{M}$ are evaluated at their probability limits.

B.3.2 Asymptotic distribution

Consistency and asymptotic normality of the estimator $\hat{\rho}(\hat{\lambda}, \alpha)$ follows in a straightforward way from the results thus far. In particular, with $\alpha = 0$ the asymptotic variance can be computed as a special case of Theorem 3 in Imbens and Angrist (1994). In our setting, we can view estimation of $h(z)$ as a parametric problem $h(z) = g(z, \theta)$ where the parameter vector θ is the mean and variance of Γ_i, along with the vector λ:

$$\theta = (\mu_\Gamma, \Sigma, \lambda)' = (\{\mu_{\Gamma,l}\}_l, \{\Sigma_{lm}\}_{l \leq m}, \{\lambda\}_l)' \text{ with } l, m \in \{1 \ldots |\mathcal{F}|\}$$

Then $\hat{\rho}(\lambda, \alpha) = \widehat{Cov}(g(Z_i, \hat{\theta}), Y_i)/\widehat{Cov}(g(Z_i, \hat{\theta}), D_i)$, where $\hat{\theta}$ solves a set of moment conditions $\sum_{i=1}^N \psi(Z_i, \hat{\theta}) = 0$ given explicitly in the theorem below.

Theorem B.1 below allows $\alpha_n > 0$ provided that the sequence converges in probability to zero at a sufficient rate. By Proposition B.1, we obtain this rate for the "one-step" minimizer of the feasible MSE estimate given in Eq. (B.2).

Theorem B.1. *Under the Assumptions of Theorem 2.1, if $\alpha_n = o_p(\sqrt{n})$ then*

$$\sqrt{n}(\hat{\rho}(\hat{\lambda}, \alpha_n) - \Delta_c) \xrightarrow{d} N(0, V)$$

where $V = \mathbf{e_1}'\Pi^{-1}\Omega(\Pi')^{-1}\mathbf{e_1}$ (i.e. the top-left element of $\Pi^{-1}\Omega(\Pi')^{-1}$) with:

$$\Omega = \begin{pmatrix} -\mathbb{E}[D_i g(Z_i, \theta)] & -\mathbb{E}[g(Z_i, \theta)] & \mathbb{E}[U_i d_\theta g(Z_i, \theta)] \\ -\mathbb{E}[D_i] & -1 & 0 \\ 0 & 0 & \mathbb{E}[d_{\theta'} \psi(Z_i, \theta)] \end{pmatrix}$$

$$\Pi = \begin{pmatrix} \mathbb{E}[g(Z_i, \theta)^2] & \mathbb{E}[g(Z_i, \theta) U_i] & \mathbb{E}[g(Z_i, \theta)\psi(Z_i, \theta)]' \\ \mathbb{E}[g(Z_i, \theta) U_i] & \mathbb{E}[U_i^2] & \mathbb{E}[U_i \psi(Z_i, \theta)]' \\ \mathbb{E}[g(Z_i, \theta) U_i \psi(Z_i, \theta)] & \mathbb{E}[U_i \psi(Z_i, \theta)] & \mathbb{E}[\psi(Z_i, \theta)\psi(Z_i, \theta)'] \end{pmatrix}$$

so long as Ω and Π are finite and Π has full rank, with the definitions:

$$U_i := Y_i - \mathbb{E}[Y_i] - \Delta_c(D_i - \mathbb{E}[D_i])$$

$$\theta = (\mu_\Gamma, \Sigma, \lambda)' = (\{\mu_{\Gamma,l}\}_l, \{\Sigma_{lm}\}_{l \leq m}, \{\lambda\}_l)'$$

$$g(z, \theta) = \lambda' \Sigma^{-1}(\Gamma(Z_i) - \mu_\Gamma)$$

$$\psi(Z_i,\theta) = ((\Gamma(Z_i) - \mu_\Gamma)', \{(\Gamma_l(Z_i) - \mu_{\Gamma,l})(\Gamma_m(Z_i) - \mu_{\Gamma,m}) - \Sigma_{lm}\}_{l\le m}, \{c_l(Z_i) - \lambda_l\}_l)'$$

Here $\Gamma(Z_i) = (\Gamma_1(Z_i)\dots\Gamma_{|\mathcal{F}|}(Z_i))'$ where $\Gamma(Z_i)_l = Z_{S_l,i}$ for some arbitrary ordering S_l of the sets in $\mathcal{F}$, and $c_l(z) = c(g(S_l), z)$ (and thus $P(C_i = 1|G_i = g(S_l)) = \mathbb{E}[c_l(Z_i)]$).

Proof. See Appendix B.4. □

B.3.3 Simulation study

This section reports a Monte Carlo experiment in which the regularized estimator proposed above is compared against its unregularized version and 2SLS. I proceed in two steps. In a first simulation involving three binary instruments, I demonstrate the practical importance of regularization. A second simulation with two binary instruments highlights the potential dangers of using 2SLS.

Three instrument DGP:

We first let $J = 3$, and put equal weight $P(G_i = g) = .05$ over each of the 20 response groups. To introduce endogeneity, I let $Y_i(0) = G_i \cdot U_i$ where the G_i are numbered arbitrarily from one to 20 and $U_i \sim Unif[0,1]$. The treatment effect within each group g is chosen to be constant and equal to g, so that

$$Y_i(1) = Y_i(0) + G_i + V_i$$

with $V_i \sim Unif[0,1]$. With this setup, $ACL = 10$.

For the joint distribution of the instruments, I consider two alternatives, meant to capture different extremes regarding statistical dependence among the instruments:

1. (Z_{1i}, Z_{2i}, Z_{3i}) generated as uncorrelated coin tosses

2. (1) followed by the following transformation: if $Z_{2i} = 1$ set $Z_{3i} = 0$ with probability 95%

I let the sample size be $n = 1000$, and perform one thousand simulations. Our primary goal is to compare the estimator $\hat{\rho}(1, 1, \ldots, 1, \alpha)$, where α chosen by the feasible approximate MSE minimizing procedure described in Section 2.5, to the simple Wald estimator of ACL $(\hat{E}[Y_i|Z_i = (111)] - \hat{E}[Y_i|Z_i = (000)])/(\hat{E}[D_i|Z_i = (111)] - \hat{E}[D_i|Z_i = (000)])$, which is equal to $\hat{\rho}(1 \ldots 1, \alpha = 0)$. I also benchmark both estimators against fully saturated 2SLS. I stress that 2SLS is not generally consistent for the ACL (or any convex combination of treatment effects) under vector monotonicity. Nevertheless, given the popularity of 2SLS and its desirable properties under traditional LATE monotonicity, it is important to know if and when the proposed estimator $\hat{\rho}(\lambda, \alpha)$ outperforms 2SLS in practice.

Figure B.1 shows the results for the first DGP, where the Z_j are independent Bernoulli random variables with mean 1/2. We see that with the good overlap of the points $\bar{Z} = (1, 1, 1)$ and $\bar{Z} = (0, 0, 0)$ (which are each equal to 1/8), the Wald estimator performs well. For this DGP, the procedure to choose $\hat{\alpha}_{mse}$, minimizing MSE, results in small values with high probability. Hence the regularized estimator $\hat{\rho}((1, 1, \ldots 1)', \hat{\alpha}_{mse})$ according to Proposition B.1 is very close to the Wald estimator (recall that they are identical when $\alpha = 0$). However, my estimator does deliver a slightly smaller RMSE, as expected, at the cost of some bias. Fully saturated 2SLS happens to also perform well for this DGP.

Figure B.2 shows the results for the second DGP, where I modify the joint distribution of (Z_1, Z_2, Z_3) to impose $E(Z_{3i}|Z_{2i} = 1) = 0.05$. In this case, the Wald estimator performs comparatively poorly. We see that regularizing the estimator to use the full sample rather than just the points $\bar{Z} = (1, 1, 1)$ and $\bar{Z} = (0, 0, 0)$ can help considerably.

Two instrument DGP:

Note that in both Figures B.1 and B.2, fully saturated 2SLS (regression on the propensity score) performs well, in the latter case actually outperforming both of the alternative esti-

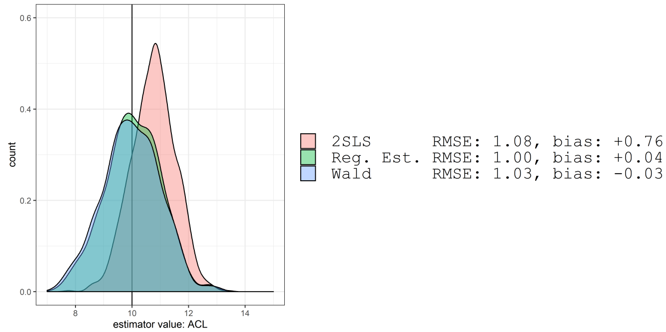

Figure B.1: Monte Carlo distributions of estimators, for the first DGP (Z uncorrelated coin tosses) with three binary instruments. "Reg. Est." indicates $\hat{\rho}(1, \ldots, 1, \hat{\alpha}_{mse})$. The vertical line shows the true value of ACL.

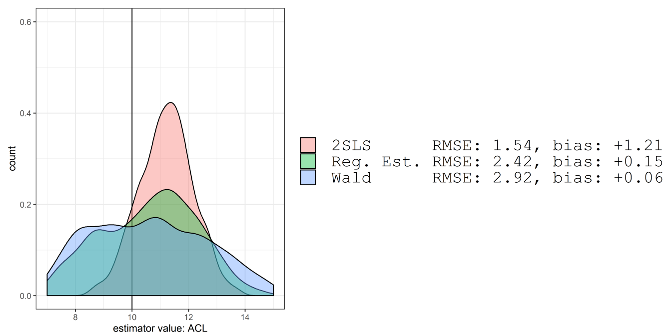

Figure B.2: Monte Carlo distributions of estimators, for the first DGP ($P(Z_{3i}|Z_{2i} = 1) = 0.05$) with three binary instruments. "Reg. Est." indicates $\hat{\rho}(1, \ldots, 1, \hat{\alpha}_{mse})$. The vertical line shows the true value of ACL.

mators. This is despite the fact that it is not consistent for the ACL, and is in general not even guaranteed to be consistent for Δ_c for any choice of the function $c(g, z)$. To demonstrate that 2SLS can in practice perform very poorly under vector monotonicity, I below report results from an additional simulation in which $J = 2$.

For this simulation, the DGP is as follows. Among the six possible response groups under vector monotonicity, I give units a 90% chance of being Z_1 complier and a 10% chance of Z_2 complier. The treatment effect is set to 2 for Z_1 compliers, and -8 for Z_2 compliers, resulting in a ACL of unity. I generate negatively correlated binary instruments (with correlation of about $-.1$) from a multivariate normal. In particular, with

$$\begin{pmatrix} Z_1^* \\ Z_2^* \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & -.8 \\ -.8 & 1 \end{bmatrix} \right]$$

I set $Z_{1i} = 1$ when Z_{1i}^* is over its median and $Z_{2i} = 1$ when Z_{2i}^* is over its median. I again let the sample size be $n = 1000$, and perform a thousand simulations.

Figure B.3 shows that in this case, 2SLS is indeed outside of the convex hull of treatment effects, despite having high precision. The proposed regularized estimator clearly outperforms both of the alternatives for this DGP.

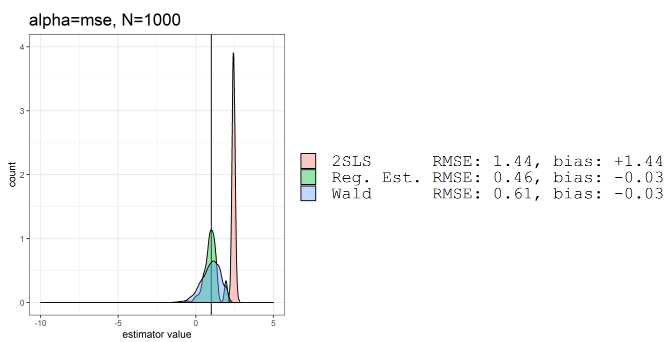

Figure B.3: Monte Carlo distributions of estimators, for the second two-instrument DGP. "Reg. Est." indicates $\hat{\rho}(1, \ldots, 1, \hat{\alpha}_{mse})$. The vertical line shows the true value of ACL.

B.4 Proofs

This section provides proofs for the formal results presented in the body of the paper.

B.4.1 Proof of Proposition 2.1

To simplify notation take each ordering $\geq_j$ to be the ordering on the natural numbers $\geq$, without loss. The two versions of VM are:

Assumption VM (vector monotonicity). *For $z, z' \in \mathcal{Z}$, if $z \geq z'$ component-wise, then $D_i(z) \geq D_i(z')$ for all i.*

Assumption VM′ (alternative characterization). *$D_i(z_j, z_{-j}) \geq D_i(z'_j, z_{-j})$ for all i when $z_j \geq z'_j$ and both (z_j, z_{-j}) and $(z'_j, z_{-j}) \in \mathcal{Z}$*

The claim is that $VM \iff VM'$.

- **VM** $\implies$ **VM′** : immediate, since $(z_j, z_{-j}) \geq (z'_j, z_{-j})$ in a vector sense when $z_j \geq z'_j$
- **VM′** $\implies$ **VM** : consider $z, z' \in \mathcal{Z}$ such that $z \geq z'$ in a vector sense, i.e. $z_j \geq z'_j$ for all $j \in \{1 \dots J\}$. Then by VM' and connectedness of $\mathcal{Z}$, then for some ordering of the instrument labels $1 \dots J$:

$$D_i \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_J \end{pmatrix} \geq D_i \begin{pmatrix} z'_1 \\ z_2 \\ \vdots \\ z_J \end{pmatrix} \forall i, \quad D_i \begin{pmatrix} z'_1 \\ z_2 \\ \vdots \\ z_J \end{pmatrix} \geq D_i \begin{pmatrix} z'_1 \\ z'_2 \\ \vdots \\ z_J \end{pmatrix} \forall i, \quad etc \dots$$

  and thus:

$$D_i \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_J \end{pmatrix} \geq D_i \begin{pmatrix} z'_1 \\ z_2 \\ \vdots \\ z_J \end{pmatrix} \geq D_i \begin{pmatrix} z'_1 \\ z'_2 \\ \vdots \\ z_J \end{pmatrix} \geq \dots \geq D_i \begin{pmatrix} z'_1 \\ z'_2 \\ \vdots \\ z'_J \end{pmatrix} \text{ for all } i$$

B.4.2 Proof of Proposition 2.2

Let $P(z) := \mathbb{E}[D_i|Z_i = z]$ be the propensity score function. By the law of iterated expectations and Assumption 1:

$$P(z) = \sum_{g \in \mathcal{G}} P(G_i = g|Z_i = z)\mathbb{E}[D_i(Z_i)|G_i = g, Z_i = z] = \sum_{g \in \mathcal{G}} P(G_i = g)\mathcal{D}_g(z)$$

By VM, $\mathcal{D}_g(z)$ is component-wise monotonic for any g in the support of G_i. As a convex combination of component-wise monotonic functions, $P(z)$ will thus also be component-wise monotonic.

In the other direction, note that by PM if $P(z_j, z_{-j}) > P(z_j', z_{-j})$, then we must have that $D_i(z_j, z_{-j}) \geq D_i(z_j', z_{-j})$ rather than $D_i(z_j, z_{-j}) \leq D_i(z_j', z_{-j})$. Thus component-wise monotonicity of $P(z)$ with respect to some collection of orderings $\{\geq_j\}_{j \in \{1 \dots J\}}$ implies $D_i(z_j, z_{-j}) \geq D_i(z_j', z_{-j})$ for all choices of $j \in \{1 \dots J\}$, $z_j \geq_j z_j'$, and $z_{-j} \in \mathcal{Z}_{-j}$ (and all i). This is the equivalent form of VM stated in Proposition 2.1.

B.4.3 Proof of Proposition 2.4

Let $\tilde{\mathcal{Z}}$ be the set of possible values for the new set of instruments $(\tilde{Z}_2, \dots \tilde{Z}_m, Z_{-1})$. Since $P(\tilde{Z}_{mi} = 0 \,\&\, \tilde{Z}_{ni} = 1) = 0$ for any $m > n$, we can take $\tilde{\mathcal{Z}}$ to only consist of cases where for all m: $\tilde{Z}_{-m}$ is composed of all zeros for the first $m - 1$ entries, and then ones for $m + 1 \dots M$. Note that fixing Z_1 is equivalent to fixing $\tilde{Z}_2 \dots \tilde{Z}_M$.

If $\mathcal{Z}$ is connected, then the $\tilde{\mathcal{Z}}$ given above is connected. Then, by Proposition 2.1, we simply need to show that for any $Z_{-1} = (Z_2, \dots, Z_J)$ and $\tilde{Z}_{-m} = (\tilde{Z}_2, \dots, \tilde{Z}_m, \tilde{Z}_{m+1}, \dots, \tilde{Z}_M)$ such that $(0, \tilde{Z}_{-m}, Z_{-1}) \in \mathcal{Z}$ and $(1, \tilde{Z}_{-m}, Z_{-1}) \in \mathcal{Z}$:

$$D_i(1, \tilde{Z}_{-m}; Z_{-1}) \geq D_i(0, \tilde{Z}_{-m}; Z_{-1})$$

where the notation $D_i(a, b; c)$ is understood as $D_i(d, c)$ where d is the value of Z_1 corresponding to $\tilde{Z}$ with value a for $\tilde{Z}_m$ and b for $\tilde{Z}_{-m}$. For any $\tilde{Z}_{-m}$ satisfying $(0, \tilde{Z}_{-m}, Z_{-1}) \in \mathcal{Z}$ and $(1, \tilde{Z}_{-m}, Z_{-1}) \in \mathcal{Z}$, switching $\tilde{Z}_m$ from zero to ones corresponds to switching Z_1 from value z_{m-1} to value z_m. Since

$$D_i(1, \tilde{Z}_{-m}; Z_{-1}) - D_i(0, \tilde{Z}_{-m}; Z_{-1}) = D_i(z_m, Z_{-1}) - D_i(z_{m-1}, Z_{-1}) \geq 0$$

by vector monotonicity on the original vector $(Z_1 \dots Z_J)$, the result now follows.

B.4.4 Proof of Proposition 2.3

For any fixed z, write the condition $\mathcal{D}_{g(F)}(z) = 1$ as

$$\left\{\mathcal{D}_{g(F)}(z) = 1\right\} \iff \left\{\bigcup_{S \in F} \left\{\mathcal{D}_{g(S)}(z) = 1\right\}\right\} \iff \text{ not } \left\{\bigcap_{S \in F} \left\{\mathcal{D}_{g(S)}(z) = 0\right\}\right\}$$

which can be written as

$$\mathcal{D}_g(z) = 1 - \prod_{S \in F} \left(1 - \mathcal{D}_{g(S)}(z)\right) = \sum_{f \subseteq F : f \neq \emptyset} (-1)^{|f|+1} \prod_{S \in F} \mathcal{D}_{g(S)}(z)$$

Let $\mathbf{z}(z) = \{j \in \{1 \dots J\} : z_j = 1\}$ represent z as the subset of instrument indices for which the associated instrument takes the value of one. Then, using that for a simple response group $\mathcal{D}_{g(S)}(z) = \mathbb{1}(S \subseteq \mathbf{z}(z))$:

$$
\begin{aligned}
\mathcal{D}_g(z) &= \sum_{f \subseteq F : f \neq \emptyset} (-1)^{|f|+1} \prod_{s \in F} \mathcal{D}_{g(S)}(z) \\
&= \sum_{f \subseteq F : f \neq \emptyset} (-1)^{|f|+1} \cdot \mathcal{D}_{g\left(\left(\bigcup_{S \in f} S\right)\right)}(z) \\
&= \sum_{f \subseteq F : f \neq \emptyset} (-1)^{|f|+1} \cdot \mathbb{1}\left(\left(\bigcup_{S \in f} S\right) \subseteq \mathbf{z}(z)\right) \\
&= \sum_{\substack{\emptyset \subset f \subseteq F: \\ \left(\bigcup_{S \in f} S\right) \subseteq \mathbf{z}(z)}} (-1)^{|f|+1} = \sum_{S' \subseteq \mathbf{z}(z)} \sum_{\substack{\emptyset \subset f \subseteq F: \\ \left(\bigcup_{S \in f} S\right) = S'}} (-1)^{|f|+1} \\
&= \sum_{S' \subseteq \{1 \dots J\}} \mathbb{1}\left(S' \subseteq \mathbf{z}(z)\right) \sum_{\substack{\emptyset \subset f \subseteq F: \\ \left(\bigcup_{S \in f} S\right) = S'}} (-1)^{|f|+1} \\
&= \sum_{S' \subseteq \{1 \dots J\}} \left[\sum_{\substack{\emptyset \subset f \subseteq F: \\ \left(\bigcup_{S \in f} S\right) = S'}} (-1)^{|f|+1} \right] \mathcal{D}_{g(S')}(z) = \sum_{\emptyset \subset S' \subseteq \{1 \dots J\}} \left[\sum_{\substack{f \subseteq F: \\ \left(\bigcup_{S \in f} S\right) = S'}} (-1)^{|f|+1} \right] \mathcal{D}_{g(S')}(z)
\end{aligned}
$$

Thus, letting $s(F, S') := \left\{ f \subseteq F : \left(\bigcup_{S \in f} S\right) = S' \right\}$, we have $\mathcal{D}_{g(F)}(z) = \sum_{S'} [M_J]_{F,S'} \mathcal{D}_{g(S)}(z)$, where the sum ranges over non-null subsets of the instruments $\emptyset \subset S' \subseteq \{1 \dots J\}$ and $[M_J]_{F,S'} = \sum_{f \in s(F,S')} (-1)^{|f|+1}$.

B.4.5 Proof of Lemma 2.1

Any indicator $\mathbb{1}(Z_i = z)$ for a value $z \in \{0,1\}^J$ can be expanded out as a polynomial in the instrument indicators as $\mathbb{1}(Z_i = z) = \prod_{j \in z_1} Z_{ji} \prod_{j \in z_0} (1 - Z_{ji}) = \sum_{f \subseteq z_0} (-1)^{|f|} Z_{(z_1 \cup f),i}$, where (z_1, z_0) is a partition of the indices $j \in \{1 \dots J\}$ that take a value of zero or one in z, respectively. With $J = 2$ for example,

$$
((1 - Z_{1i})(1 - Z_{2i}), Z_{1i}(1 - Z_{2i}), Z_{2i}(1 - Z_{1i}), Z_{1i}Z_{2i}) = (1, Z_{1i}, Z_{2i}, Z_{1i}Z_{2i})A = (1, \Gamma_i')A
$$

where $A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 1 \end{pmatrix}$. Denote the random vector of such indicators $\mathfrak{z}_i$. Then $(1, \Gamma_i')A = \mathfrak{z}_i'$, with the matrix of coefficients $A_{S,z} = \sum_{\substack{f \subseteq z_0 \\ (z_1 \cup f) = S}} (-1)^{|f|}$. The matrix A so defined must be invertible, because any product of the instruments Z_{Si} for $S \subseteq \{1 \dots J\}$ can similarly be expressed as a linear combination of the components of $\mathfrak{z}_i$, where we define $Z_{\emptyset i} = 1$. Specifically, $Z_{Si} = \sum_{z \in \mathcal{Z}} \mathbb{1}\left(\forall_{j \in S}, z_j = 1\right) \mathbb{1}(Z_i = z)$.

Consider the matrix

$$\Sigma^* := \mathbb{E}[(1, \Gamma_i')'(1, \Gamma_i')] = A'^{-1}\mathbb{E}[\mathfrak{z}_i \mathfrak{z}_i']A^{-1} = A'^{-1} diag\{P(Z_i = z)\}A^{-1}$$

where $\mathbb{E}[\mathfrak{z}_i \mathfrak{z}_i']$ is diagonal since the events that Z_i take on two different values are exclusive. Since A^{-1} exists, the rank of Σ^* must be equal to the rank of $diag\{P(Z_i = z)\}$, which is in turn equal to the cardinality of $\mathcal{Z}$. Assumption 3 thus holds if and only if Σ^* has full rank of 2^J. Note that although A^{-1} diagonalizes the matrix Σ^*, it does not provide its eigen-decomposition, as $A^{-1} \neq A'$ (A is not orthogonal).

Now we prove that Σ^* has full rank whenever Σ has full rank, and vice versa. Note that $\Sigma = Var(\Gamma_i)$ has full rank if and only if $\omega'\mathbb{E}[(\Gamma_i - E\Gamma_i)(\Gamma_i - E\Gamma_i)]\omega = \mathbb{E}[\omega'(\Gamma_i - E\Gamma_i)(\Gamma_i - E\Gamma_i)\omega] > 0$, i.e. $P(\omega'(\Gamma_i - E\Gamma_i) = 0) < 1$ for any $\omega \in \mathbb{R}^{2^J - 1}/\mathbf{0}$. Similarly Σ^* has full rank if $P((\omega_0, \omega)'((1, \Gamma_i) = 0) < 1$ for any $\omega_0 \in \mathbb{R}, \omega \in \mathbb{R}^{2^J - 1}$ where (ω_0, ω) is not the zero vector in $\mathbb{R}^{2^J}$. But if for some ω, $\omega'(\Gamma_i - E\Gamma_i) = 0$ w.p.1., then we also have $(\omega_0, \omega)'(1, \Gamma_i) = 0$ w.p.1. by choosing $\omega_0 = -\omega'\mathbb{E}[\Gamma_i]$. In the other direction, note that $(\omega_0, \omega)'(1, \Gamma_i) = 0$ w.p.1. implies that $\omega'\Gamma_i = -\omega_0$ and hence $\omega'(\Gamma_i - E\Gamma_i) = -\omega_0 - \omega' E\Gamma_i = -\omega_0 - \mathbb{E}[\omega'\Gamma_i] = -\omega_0 + \omega_0 = 0$.

B.4.6 Proof of the Appendix B.1 Proposition

Introduce the notation that $\sqcup$ indicates inclusion of a new set among a family of sets (while $\cup$ continues to indicate taking the union of elements across sets).

For any $S \subseteq \mathcal{M}$ that contains both Z^j_m and $Z^j_{m'}$ for some j and $m < m'$, $g(F \sqcup S)$ and $g(F \sqcup S/\{Z^j_m\})$ generate the same selection behavior for any Sperner family F on all of $\mathcal{Z}$ (this can be seen by mapping the implied selection behavior back to the original discrete instrument Z_j). Thus, we can take $\mathcal{G}$ to exclude such S without loss of generality.

Now, consider the family $\mathcal{F}$ of all $S \subset \mathcal{M}$ that contain at most one Z^j_m for any given j. By the above, this choice of $\mathcal{F}$ satisfies Assumption 3b*. Suppose it did not satisfy Assumption 3a*. Then, there would need to exist a non-zero vector ω such that $P\left(\sum_{S \in \mathcal{F}} \omega_S Z_{Si} = 0\right) = 1$ with $Z_{Si} := \prod_{(j,m) \in S} \tilde{Z}^j_m$. This would imply non-invertibility of $\Sigma^* := \mathbb{E}[(1, \Gamma_i)(1, \Gamma_i)']$, where $\Gamma_i := \{Z_{Si}\}_{S \in \mathcal{F}, S \neq \emptyset}$ by the same argument as in the proof of Lemma 2.1 (Γ_i and a vector of indicators for all $z \in \mathcal{Z}$ are each related by an invertible linear map), which in turn contradicts the assumption of full support. Note that invertibility of Σ^* is again equivalent to invertibility of $Var(\Gamma_i)$ as before.

B.4.7 Proof of Theorem 2.1

We first note that any measurable function $f(Y)$ preserves Assumption 1, that is

$$(f(Y_i(1)), f(Y_i(0)), G_i) \perp Z_i$$

and Assumptions 2-3 are unaffected by such a transformation to the outcome variable. Thus, we continue without loss with Y_i, $Y_i(1)$ and $Y_i(0)$ possibly redefined as $f(Y_i)$, $f(Y_i(1))$ and $f(Y_i(0))$ respectively.

Note that the function $h(\cdot)$ given in Theorem 2.1 has the property that $\mathbb{E}[h(Z_i)] = 0$, for any distribution of the instruments. Consider the quantity $\mathbb{E}[Y_i D_i h(Z_i)]$ for a function h having this property. By the law of iterated expectations, and the independence assumption:

$$\begin{aligned}\mathbb{E}[Y_i D_i h(Z_i)] &= \sum_g P(G_i = g)\mathbb{E}[Y_i D_i h(Z_i)|G_i = g] \\ &= \sum_g P(G_i = g)\mathbb{E}[Y_i(1)\mathcal{D}_g(Z_i)h(Z_i)|G_i = g] \\ &= \sum_g P(G_i = g)\mathbb{E}[Y_i(1)|G_i = g]\mathbb{E}[\mathcal{D}_g(Z_i)h(Z_i)] \end{aligned} \tag{B.3}$$

where $\mathcal{D}_g(z)$ denotes the selection function for response group g. Similarly,

$$\begin{aligned}\mathbb{E}[Y_i(1 - D_i)h(Z_i)] &= \sum_g P(G_i = g)\mathbb{E}[Y_i(0)(1 - D_i)h(Z_i)|G_i = g] \\ &= \sum_g P(G_i = g)\left\{\mathbb{E}[Y_i(0)|G_i = g]\mathbb{E}[h(Z_i)]\right. \\ &\qquad\qquad \left. - \mathbb{E}[Y_i(0)|G_i = g]\mathbb{E}[\mathcal{D}_g(Z_i)h(Z_i)]\right\} \\ &= \sum_g -P(G_i = g)\mathbb{E}[Y_i(0)|G_i = g]\mathbb{E}[\mathcal{D}_g(Z_i)h(Z_i)] \end{aligned} \tag{B.4}$$

where we have used that $Z_i \perp (Y_i(0), Z_i)$ and $\mathbb{E}[h(Z_i)] = 0$.

Combining these two results:

$$\mathbb{E}[Y_i h(Z_i)] = \mathbb{E}[Y_i D_i h(Z_i)] + \mathbb{E}[Y_i(1 - D_i)h(Z_i)] = \sum_g P(G_i = g)\mathbb{E}[\mathcal{D}_g(Z_i)h(Z_i)]\Delta_g \tag{B.5}$$

where $\Delta_g := \mathbb{E}[Y_i(1) - Y_i(0)|G_i = g]$. By the law of iterated expectations, we also have that

$$\mathbb{E}[D_i h(Z_i)] = \sum_g P(G_i = g)\mathbb{E}[\mathcal{D}_g(Z_i)h(Z_i)] \tag{B.6}$$

Note that in all of Equations (B.3), (B.4) and (B.5), the weighing over various groups g is governed by the quantity $\mathbb{E}[\mathcal{D}_g(Z_i)h(Z_i)]$. It can be seen that never takers and always takers receive no weight, since $\mathbb{E}[\mathcal{D}_{n.t}(Z_i)h(Z_i)] = \mathbb{E}[0] = 0$ and since $\mathbb{E}[\mathcal{D}_{a.t}(Z_i)h(Z_i)] = \mathbb{E}[h(Z_i)] = 0$.

Let $\mathcal{F}$ denote the set of non-empty subsets of the instrument indices: $\mathcal{F} := \{S \subseteq \{1, 2, \dots J\}, S \neq \emptyset\}$, and recall that these correspond each to a simple response group $g(S)$, where $\mathcal{D}_{g(S)}(Z_i) = Z_{Si}$. I first show that for any $\lambda \in \mathbb{R}^{|\mathcal{F}|}$, Assumption 3 allows us to define an $h(Z_i)$ such that $\mathbb{E}[\mathcal{D}_{g(S)}(Z_i)h(Z_i)] = \mathbb{E}[Z_{Si}h(Z_i)] = \lambda_S$. Note that since $\mathbb{E}[h(Z_i)] = 0$, this is the same as tuning each covariance $Cov(Z_{Si}, h(Z_i))$ to λ_S (c.f. Proposition 2.5). In particular, consider the choice $h(Z_i) = (\Gamma_i - \mathbb{E}[\Gamma_i])'\Sigma^{-1}\lambda$, where recall that Γ_i is a vector of Z_{Si} for each $S \in \mathcal{F}$.

$$\begin{aligned}(\mathbb{E}[h(Z_i)_i, \Gamma_{i1}], \mathbb{E}[h(Z_i), \Gamma_{i2}], \dots, \mathbb{E}[h(Z_i), \Gamma_{ik}])' &= \mathbb{E}[(\Gamma_i - \mathbb{E}[\Gamma_i])h(Z_i)] \\ &= \mathbb{E}[(\Gamma_i - \mathbb{E}[\Gamma_i])(\Gamma_i - \mathbb{E}[\Gamma_i])']\Sigma^{-1}\lambda \\ &= \Sigma\Sigma^{-1}\lambda = \lambda\end{aligned}$$

We can understand the algebra of this result as follows. Let $V = span(\{Z_{Si} - \mathbb{E}[Z_{Si}]\}_{S\in\mathcal{F}})$. V is a subspace of the vector space $\mathcal{V}$ of random variables on $\mathcal{Z}$, with the zero vector being a degenerate random variable equal to zero. Since the matrix Σ is positive semidefinite by construction, Assumption 3 is equivalent to the statement that for all $\omega \in \mathbb{R}^{|\mathcal{F}|}/\mathbf{0}$, $\omega'\mathbb{E}[(\Gamma_i - \mathbb{E}[\Gamma_i])(\Gamma_i - \mathbb{E}[\Gamma_i])']\omega = \mathbb{E}[|\omega'(\Gamma_i - \mathbb{E}[\Gamma_i])|^2] > 0$: i.e. $P\left(\sum_{S\in\mathcal{F}} \omega_S(Z_{Si} - \mathbb{E}[Z_{Si}])) = 0\right) < 1$ for all $\omega \in \mathbb{R}^{|\mathcal{F}|}/\mathbf{0}$. In other words, the random variables $(Z_{Si} - \mathbb{E}[Z_{Si}])$ for $S \in \mathcal{F}$ are linearly independent, and hence form a basis of V. Since V is finite dimensional, there exists an orthonormal basis of random vectors of the same cardinality, $|\mathcal{F}|$, where orthonormality is defined with respect to the expectation inner product: $\langle A, B\rangle := \mathbb{E}[A_iB_i]$. It is this orthogonalized version of the Z_{Si} that affords the ability to separately tune each of the $\mathbb{E}[h(Z_i)Z_{Si}]$ to the desired value λ_S, without disrupting the others.

Note that under Assumption 1:

$$\Delta_c = \sum_{g\in\mathcal{G}} \left\{\frac{P(G_i = g)P(C_i = 1|G_i = g)}{P(C_i = 1)}\right\} \cdot \Delta_g = \frac{\sum_{g\in\mathcal{G}} P(G_i = g)P(C_i = 1|G_i = g) \cdot \Delta_g}{\sum_{g\in\mathcal{G}} P(G_i = g)P(C_i = 1|G_i = g)}$$

Comparing with Equations (B.5) and (B.6), the equality $\Delta_c = \mathbb{E}[h(Z_i)Y_i]/\mathbb{E}[D_i h(Z_i)]$ follows (provided that $P(C_i = 1) > 0$) if the coefficients match. That is: $\mathbb{E}[\mathcal{D}_g(Z_i)h(Z_i)] = P(C_i = 1|G_i = g)$, for all $g \in \mathcal{G}^c$. By the above, this is guaranteed under Property M if we choose $\lambda_S = P(C_i = 1|G_i = g(S)) = \mathbb{E}[c(g(S), Z_i)]$, since the quantity $\mathbb{E}[\mathcal{D}_g(Z_i)h(Z_i)]$ appearing in Eq. (B.5) is linear in $\mathcal{D}_g(Z_i)$. The same logic follows for causal parameters of the form $\mathbb{E}[Y_i(d)|C_i = 1]$ for $d \in \{0, 1\}$, using Equations (B.3) and (B.4) and

$$\begin{aligned}
\mathbb{E}[Y_i(d)|C_i = 1] &= \sum_{g \in \mathcal{G}} P(G_i = g|C_i = 1)\mathbb{E}[Y_i(d)|G_i = g, c(g, Z_i) = 1] \\
&= P(C_i = 1)^{-1} \sum_{g \in \mathcal{G}} P(G_i = g)P(C_i = 1|G_i = g)\mathbb{E}[Y_i(d)|G_i = g]
\end{aligned}$$

by independence. Note that the quantity λ_S for each S can be computed from the observed distribution of Z.

To replace Assumption 3 with Assumption 3* from Appendix B.1, simply replace $\mathcal{F}$ as defined here with a maximal $\mathcal{F}$ from Assumption 3a*.

B.4.8 Proof of Corollary 2.1 to Theorem 2.1

The proof of Lemma 2.1 shows that $(1, \Gamma_i')A$ is a vector of indicators $\mathfrak{z}_i'$ for values of Z, where A is the matrix with entries given in Corollary 2.1, which is invertible, and $\mathfrak{z}_i$ is a vector of indicators $\mathbb{1}(Z_i = z)$ for each of the values $z \in \mathcal{Z}$. We can thus write $h(Z_i)$ from Theorem 2.1 as

$$\begin{aligned}
h(Z_i) = \lambda'\Sigma^{-1}(\Gamma_i - \mathbb{E}[\Gamma_i]) &= (0, \lambda')\mathbb{E}[(1, \Gamma_i')'(1, \Gamma_i')]^{-1}(1, \Gamma_i')' \\
&= (0, \lambda')\mathbb{E}[A'^{-1}A'(1, \Gamma_i')'(1, \Gamma_i')AA^{-1}]^{-1}A'^{-1}\mathfrak{z}_i \\
&= (0, \lambda')A\mathbb{E}[\mathfrak{z}_i\mathfrak{z}_i']^{-1}\mathfrak{z}_i
\end{aligned}$$

This is useful because $\mathbb{E}[\mathfrak{Z}_i\mathfrak{Z}_i']$ is diagonal, since the events that Z_i take on two different values are exclusive: $\mathbb{E}[\mathfrak{Z}_i\mathfrak{Z}_i'] = diag\{P(Z_i = z)\}_{z\in\mathcal{Z}}$.

Now, for $V \in \{Y, D\}$, $\mathbb{E}[h(z)V_i] = (0, \lambda')Adiag\{P(Z_i = z)\}^{-1}_{z\in\mathcal{Z}}\{\mathbb{E}[\mathbb{1}(Z_i = z)V_i]\}_{z\in\mathcal{Z}} = (0, \lambda')A\{\mathbb{E}[V_i|Z_i = z]\}_{z\in\mathcal{Z}}$. Thus $(0, \lambda')A$ describes the coefficients in an expansion of $\mathbb{E}[h(z)V_i]$ into CEFs of V_i across the support of Z_i.

B.4.9 Proof of Proposition 2.6

VM case

The if direction is most straightforward. From Proposition 2.3 we have that for any $z \in \mathcal{Z}$ and $g \in \mathcal{G}^c$:

$$\mathcal{D}_g(z) = \sum_{S\subseteq\{1\ldots J\}, S\neq\emptyset} [M_J]_{F(g),S} \cdot \mathcal{D}_{g(S)}(z)$$

Thus, for any such $c(g, z)$:

$$\begin{aligned}
c(g, z) &= \sum_{k=1}^{K} \sum_{S\subseteq\{1\ldots J\}, S\neq\emptyset} [M_J]_{F(g),S} \cdot \mathcal{D}_{g(S)}(h_k(z))) - \sum_{S\subseteq\{1\ldots J\}, S\neq\emptyset} [M_J]_{F(g),S} \cdot \mathcal{D}_{g(S)}(l_k(z))) \\
&= \sum_{S\subseteq\{1\ldots J\}, S\neq\emptyset} [M_J]_{F(g),S} \cdot \left\{\sum_{k=1}^{K} \mathcal{D}_{g(S)}(h_k(z))) - \mathcal{D}_{g(S)}(l_k(z)))\right\} \\
&= \sum_{S\subseteq\{1\ldots J\}, S\neq\emptyset} [M_J]_{F(g),S} \cdot c(g(S), z)
\end{aligned}$$

for any $z \in \mathcal{Z}$. To finish verifying Property M, we need only observe that $c(a.t., z) = c(n.t., z) = 0$ for all z since $\mathcal{D}_g(h_k(z)) = \mathcal{D}_g(l_k(z))$ for any h_k, l_k when $g \in \{a.t., n.t.\}$.

Now we turn to the other implication of the Proposition, that any c satisfying Property M has a representation like the above. For shorthand, let $c^{-1}(z)$ indicate the family of $S \subseteq \{1 \ldots J\}$ such that $c(g(S), z) = 1$. The following Lemma establishes that the family $c^{-1}(z)$ and its complement are each closed under unions:

Lemma. *Let c be a function from $\mathcal{G} \times \mathcal{Z}$ to $\{0, 1\}$ satisfies Property M. If $A \in c^{-1}(z)$ and $B \in c^{-1}(z)$, then $A \cup B \in c^{-1}(z)$, and if $A \notin c^{-1}(z)$ and $B \notin c^{-1}(z)$, then $A \cup B \notin c^{-1}(z)$.*

Proof. If the sets A and B are nested, then the result follows trivially. Now suppose neither set contains the other, and consider the Sperner family $A \sqcup B$ constructed of the two sets A and B. By Property M and using Proposition 2.3:

$$\begin{aligned} c(g(A \sqcup B), z) &= \sum_{\emptyset \subset S' \subseteq \{1 \ldots J\}} \left[\sum_{\substack{f \subseteq \{A,B\}: \\ \left(\bigcup_{S \in f} S\right) = S'}} (-1)^{|f|+1} \right] c\left(\bigcup_{S \in f} S, z \right) \\ &= \sum_{\emptyset \subset f \subseteq \{A,B\}} c\left(\bigcup_{S \in f} S, z \right) \\ &= c(g(A), z) + c(g(B), z) - c(g(A \cup B), z) \end{aligned}$$

In the first case, if both A and B are in $c^{-1}(z)$, then we must have $c(g(A \cup B), z) = 1$ to prevent $c(g(A \sqcup B), z)$ from evaluating to 2, which contradicts the assumption that c takes values in $\{0, 1\}$. In the second case, when both $c(g(A), z)$ and $c(g(B), z)$ are zero, we must have $c(g(A \cup B), z) = 1$ to prevent $c(g(A \sqcup B), z)$ from evaluating to -1. $\square$

As a consequence of the Lemma, since $c^{-1}(z)$ is a finite set, there exists a member $S_1(z)$ of $c^{-1}(z)$ that satisfies $S_1(z) = \bigcup_{S \in c^{-1}(z)} S$ (similarly, there exists a $S_0(z) = \bigcup_{S \notin c^{-1}(z)} S$ with $S_0(z) \notin c^{-1}(z)$). All members of the family $c^{-1}(z)$ are subsets of $S_1(z)$, and all $S \subseteq \{1 \ldots J\}$ that are not in $c^{-1}(z)$ are subsets of $S_0(z)$.

Let z take some fixed value, and beginning with the set $S_1 = S_1(z)$, define a sequence of sets $\{S_1, S_2, S_3, \ldots\}$ as follows:

$$S_{2k} = \bigcup_{\substack{S' \subseteq S_{2k-1}: \\ S' \notin c^{-1}(z)}} S' \qquad \text{and} \qquad S_{2k+1} = \bigcup_{\substack{S' \subseteq S_{2k}: \\ S' \in c^{-1}(z)}} S'$$

where we take $\bigcup_{S' \in \emptyset} S'$ to evaluate to the empty set. This sequence provides a characterization of the family $c^{-1}(z)$ as follows. For any $\emptyset \subset S \subseteq \{1 \dots J\}$:

$$
\begin{aligned}
c(g(S), z) &= \mathbb{1}(S \in c^{-1}(z)) \\
&= \mathbb{1}(S \subseteq S_1 : S \in c^{-1}(z)) \\
&= \mathbb{1}(S \subseteq S_1) - \mathbb{1}(S \subseteq S_1 : S \notin c^{-1}(z)) \\
&= \mathbb{1}(S \subseteq S_1) - \big(\mathbb{1}(S \subseteq S_2) - \mathbb{1}(S \subseteq S_2 : S \in c^{-1}(z))\big) \\
&= \mathbb{1}(S \subseteq S_1) - \mathbb{1}(S \subseteq S_2) + \big(\mathbb{1}(S \subseteq S_3) - \mathbb{1}(S \subseteq S_3 : S \notin c^{-1}(z))\big) \\
&= \dots \\
&= \sum_{n=1}^{N} (-1)^{n+1} \cdot \mathbb{1}(S \subseteq S_n) + (-1)^N \cdot \begin{cases} \mathbb{1}(S \subseteq S_N : S \in c^{-1}(z)) & \text{if } N \text{ even} \\ \mathbb{1}(S \subseteq S_N : S \notin c^{-1}(z)) & \text{if } N \text{ odd} \end{cases}
\end{aligned}
$$

for any natural number N.

Think of the power set of S_1 as a "first-order" approximation to the family $c^{-1}(z)$. However, in most cases this family is too large, as there will be subsets of S_1 that are not found in $c^{-1}(z)$. Define S_2 to be the union of all such offending sets. The power set of S_2 now provides a possible "overestimate" of the family of offending sets (since they are all in 2^{S_2}) and hence removing all subsets of S_2 as a correction to be applied to 2^{S_1} as an estimate of $c^{-2}(z)$ will overcompensate: we will have removed some sets which are indeed in $c^{-1}(z)$. We thus define S_3 analogously, whose power set provides an approximation to the error in S_2 as an approximation to the error in S_1, and so on.

Does this process of over-correction eventually terminate, so that the final remainder term is zero? Note that for any n: $S_n \subseteq S_{n-1}$. If $S_n = S_{n-1} \neq \emptyset$, then we have a fixed point S where $\bigcup_{S' \subseteq S : S' \in c^{-1}(z)} S' = \bigcup_{S' \subseteq S : S' \notin c^{-1}(z)} S'$. But by the Lemma, this would imply that S is a member both of $\{S' \subseteq S : S' \in c^{-1}(z)\}$ and of $\{S' \subseteq S : S' \notin c^{-1}(z)\}$, and therefore that both $c(g(S), z) = 1$ and $c(g(S), z) = 0$, a contradiction. Thus, $S_n \subset S_{n-1}$, and $|S_n|$ is a decreasing sequence of non-negative integers that is strictly decreasing so long as $|S_n| > 0$. It must thus converge to zero in at most $|S_1|$ iterations, so that $S_n = \emptyset$ for all $n \geq |S_1|$.

Without loss, we can terminate the sequence on an even term, since $\mathbb{1}(S \subseteq \emptyset) = 0$ for any $S \supset \emptyset$. Let $2K$ denote the smallest even number such that $S_n = \emptyset$ for all $n > 2K$, for a fixed z. Thus, we have for any $\emptyset \subset S \subseteq \{1 \dots J\}$:

$$c(g(S), z) = \sum_{n=1}^{2K} (-1)^{n+1} \cdot \mathcal{D}_{g(S)}(S_n) = \sum_{k=1}^{K} \mathcal{D}_{g(S)}(S_{2k-1}) - \mathcal{D}_{g(S)}(S_{2k})$$

where $2K \leq |S_1| \leq J$, and we have used that $\mathcal{D}_{g(S)}(S') = \mathbb{1}(S \subset S')$ for any S'.

Now recall that we have left the dependence of each of the sets S_n (as well as the integer K) on z implicit, and have also adopted the notational convention of $\mathcal{D}_g(S)$ as a shorthand for $\mathcal{D}_g(z)$ where z is a point in $\mathcal{Z}$ that takes a value of one for exactly the instruments in the set S. To obtain the notation of the final result, define for each $k = 1 \dots K$ the point $u_k(z) \in \mathcal{Z}$ to have a value of one exactly for the elements in S_{2k-1} for that value of z, and $l_k(z) \in \mathcal{Z}$ to have a value of one exactly for the elements in S_{2k} for that value of z. We may thus write, for any $\emptyset \subset S \subseteq \{1 \dots J\}$ and any $z \in \mathcal{Z}$:

$$c(g(S), z) = \sum_{k=1}^{K(z)} \mathcal{D}_{g(S)}(u_k(z)) - \mathcal{D}_{g(S)}(l_k(z)) = \sum_{k=1}^{K} \mathcal{D}_{g(S)}(u_k(z)) - \mathcal{D}_{g(S)}(l_k(z))$$

where we let K be the maximum of $K(z)$ over the finite set $\mathcal{Z}$, and we define $u_k(z)$ and $l_k(z)$ to each be a vector of zeros whenever $k > K(z)$. For each z, the relations $u_k(z) \geq l_k(z)$ and $l_k(z) \geq u_{k+1}(z)$ component-wise now follow from $S_n \subseteq S_{n+1}$.

Now we may apply Property M to construct $c(g, z)$ for any of the non-simple response groups as well. Recall that Property M says that $c(g(F), z) = \sum_{\emptyset \subset S \subseteq \{1 \dots J\}} [M_J]_{F,S} \cdot$

$c(g(S), z)$ for all z, for any Sperner family F. Thus:

$$
\begin{aligned}
c(g(F), z) &= \sum_{\emptyset \subset S \subseteq \{1 \dots J\}} [M_J]_{F,S} \cdot \sum_{k=1}^{K} \{\mathcal{D}_{g(S)}(u_k(z)) - \mathcal{D}_{g(S)}(l_k(z))\} \\
&= \sum_{k=1}^{K} \{ \sum_{\emptyset \subset S \subseteq \{1 \dots J\}} [M_J]_{F,S} \cdot \mathcal{D}_{g(S)}(u_k(z))\} - \{ \sum_{\emptyset \subset S \subseteq \{1 \dots J\}} [M_J]_{F,S} \cdot \mathcal{D}_{g(S)}(l_k(z))\} \\
&= \sum_{k=1}^{K} \mathcal{D}_{g(F)}(u_k(z)) - \mathcal{D}_{g(F)}(l_k(z))
\end{aligned}
$$

Finally, note that $\mathcal{D}_g(u_k(z)) = \mathcal{D}_g(l_k(z))$ for any $g \in \{a.t., n.t.\}$ so the following expression holds for all $g \in \mathcal{G}$:

$$
c(g, z) = \sum_{k=1}^{K} \mathcal{D}_g(u_k(z)) - \mathcal{D}_g(l_k(z))
$$

IAM case

Now I prove that representation from Proposition 2.6 also holds under IAM. Note that under IAM Property M places no restriction beyond $c(a.t., z) = c(n.t., z) = 0$ since there is no perfect linear dependency between the functions $\mathcal{D}_g(z)$ to worry about. Under IAM, each $g \in \mathcal{G}^c$ can be associated with an integer $m =\in \{1, 2 \dots 2^J - 1\}$ and characterized directly $\mathbb{1}(g = m) = \mathcal{D}_g(z_{m+1}) - \mathcal{D}_g(z'_m)$, where $z_1, z_2, \dots, z_{2^J}$ is any fixed ordering of the points that is weakly increasing according to the propensity score $\mathbb{E}[D_i|Z_i = z_m]$. m is simply the "first" point in $\mathcal{Z}$ along this sequence for which individuals of response type g take treatment.

Thus, for any function $g : \mathcal{G} \times \mathcal{Z} \to \{0, 1\}$ such that $c(a.t., z) = c(n.t., z) = 0$:

$$
\begin{aligned}
c(g, z) &= \sum_{m=1}^{2^J - 1} c(m, z) \cdot (\mathcal{D}_g(z_{m+1}) - \mathcal{D}_g(z'_m)) \\
&= \sum_{k=1}^{K} \mathcal{D}_g(u_k(z)) - \mathcal{D}_g(l_k(z))
\end{aligned} \tag{B.7}
$$

with $K = 2^J - 1$ where for each z we let $l_k(z) = z_m$ and we let $u_k(z) = \begin{cases} z_k & \text{if } c(k,z) = 0 \\ z_{k+1} & \text{if } c(k,z) = 1 \end{cases}$.
Note that if any set of consecutive $c(k,z) = c(k+1,z)\dots c(k+T,z)$ are all equal to one, then one can drop $T-1$ of these terms as the inner terms will all cancel leaving $\mathcal{D}_g(u_{k+T}(z)) - \mathcal{D}_g(l_k(z))$. Thus we may take without loss $K \leq 2^J/2 = 2^{J-1}$ (corresponding to the case where $c(1,z) = 1$, $c(2,z) = 0$, $c(3,z) = 1$ etc.).

Finally, we discuss how the above analysis reveals some common structure to Theorem T-6 of Heckman and Pinto (2018) (HP), which extends IAM-type identification to settings with more than two treatment states. Their analysis assumes a condition they call *unordered monotonicity*, which reduces to IAM when treatment takes on just two values, as we consider in our paper.

To establish the connection, let us extend the above mapping between points in $\mathcal{Z}$ and response groups $g \in \mathcal{G}$ under IAM to associate $m = 0$ with always-takers, and $m = |\mathcal{Z}|$ with never-takers (the fact that $|\mathcal{Z}| = 2^J$ for J binary instruments with rectangular support is not important here). Recall that $\mathbb{1}(g = m) = \mathcal{D}_g(z_{m+1}) - \mathcal{D}_g(z'_m)$ for the other groups. Then, HP's Theorem T-6 applied to a binary treatment implies that $\mathbb{E}[Y_i(1)|G_i = m]$ is identified for any $m < |\mathcal{Z}|$, and $\mathbb{E}[Y_i(0)|G_i = m]$ is identified for any $m > 0$ under IAM. We focus here on $\mathbb{E}[Y_i(1)|G_i = m]$, as the argument is symmetric under redefinition of treatment and control.

Note that $\mathbb{E}[Y_i(1)|G_i = m]$ corresponds to the choice $c(g,z) = \mathbb{1}(g = m)$. Since we've defined Property M in a way that rules out the $m = 0$ case, we focus on $m > 0$ to keep the discussion simple. Let b be a row vector of $c(g,z)$ across the $g = 1\dots|\mathcal{Z}|$, noting that this choice of $c(g,z)$ does not depend on instrument value z. The key step in the proof of HP's T-6 is the intermediate result that $b = b[B^\dagger B]$, where B is a matrix with generic element $B_{zg} = \mathcal{D}_g(z)$ and $B^\dagger$ is its Moore-Penrose pseudo-inverse. Given our ordering of $\mathcal{Z}$, B is simply a lower triangular matrix of ones, appended to the right by a single column of zeros (for the never-takers). It can then be verified that rows $r = 2\dots(|\mathcal{Z}| - 1)$

of $B^\dagger$ are of the form $(0, \ldots, -1, 1, \ldots 0)'$ with $r-2$ zeroes on the left (while the first row is composed of a single 1 in the first column, and the last row is all zeros). Since the columns of B are vectors of the $\mathcal{D}_g(z)$ across $z \in \mathcal{Z}$, it follows that for any $m = 1 \ldots (|\mathcal{Z}|-1)$: $\mathcal{D}_g(z_{m+1}) - \mathcal{D}_g(z'_m)$ is simply the m, g component of $B^\dagger B$, which we've seen is also equal to $\mathbb{1}(g = m)$. For these values of m, the equation $b = b[B^\dagger B]$ is then identical to Equation (B.7), viewed as a row vector over g.

B.4.10 Proof of Corollary 2.2 to Theorem 2.1

Using independence and Property M:

$$\begin{aligned}
\mathbb{E}[h(Z_i)D_i] &= \sum_g P(G_i = g)\mathbb{E}[h(Z_i)\mathcal{D}_g(Z_i)] \\
&= \sum_g P(G_i = g)E\left[h(Z_i)\left\{\sum_S [M_J]_{F(g),S}\mathcal{D}_{g(s)}(Z_i)\right\}\right]) \\
&= \sum_g P(G_i = g)\sum_S [M_J]_{F(g),S} P(C_i = 1|\mathcal{D}_{g(s)}(Z_i)) \\
&= \sum_g P(G_i = g)P(C_i = 1|G_i = g) \\
&= P(C_i = 1)
\end{aligned}$$

B.4.11 An Equivalence Result

The proofs of Proposition 2.7 and 2.9 will make use of the following equivalence result:

Proposition B.2. *Let the support $\mathcal{Z}$ of the instruments be discrete and finite. Fix a function $c(g, z)$. Let $\mathcal{P}_{DZ}$ denote the joint distribution of D_i and Z_i. Then the following are equivalent:*

1. *Δ_c is (point) identified by $\mathcal{P}_{DZ}$ and $\{\beta_s\}_{s\in\mathcal{S}}$, for some finite set $\mathcal{S}$ of known or identified (from $\mathcal{P}_{DZ}$) measurable functions $s(d, z)$, and $\beta_s := \mathbb{E}[s(D_i, Z_i)Y_i]$*

2. *$\Delta_c = \beta_s$ for a single such $s(d, z)$*

3. *$\Delta_c = \mathbb{E}[t(D_i, Z_i, Y_i)]$ with $t(d, z, y)$ a known or identified (from $\mathcal{P}_{DZ}$) measurable function*

4. Δ_c *is identified from the set of CEFs* $\{\mathbb{E}[Y_i|D_i = d, Z_i = z]\}$ *for* $d \in \{0,1\}$, $z \in \mathcal{Z}$ *along with the joint distribution* $\mathcal{P}_{DZ}$

Proof. See Supplemental Material. □

In saying that a parameter θ is *identified* by some set of empirical estimands, I mean that the set of values of θ that are compatible with the empirical estimands is a singleton, regardless of the distribution of the latent variables $(G_i, Y_i(1), Y_i(0))$ – for all $\mathcal{P}_{DZ}$ within some class (note that the marginal distribution of G_i must also be compatible with $\mathcal{P}_{DZ}$). For example, by writing the estimand of Theorem 2.1 $\sum_{z\in\mathcal{Z}} \frac{P(Z_i=z)h(z;\mathcal{P}_{DZ})}{\mathbb{E}[h(Z_i;\mathcal{P}_{DZ})D_i]} \cdot \mathbb{E}[Y_i|Z_i = z]$, where we make explicit that the function h depends on $\mathcal{P}_{DZ}$, it is clear that for any Δ_c satisfying Property M and under Assumptions 1-2, Δ_c is identified in the sense of item 4., for all $\mathcal{P}_{DZ}$ with the properties: i) the marginal distribution of Z_i satisfies Assumption 3; and ii) $\mathbb{E}[h(Z_i;\mathcal{P}_{DZ})D_i] > 0$.

B.4.12 Proof of Proposition 2.7

By Proposition B.2, we know that if Δ_c is identified from a finite set of IV-like estimands and $\mathcal{P}_{DZ}$, it can be written as a single one: $\Delta_c = \beta_s$ with $s(d,z)$ an identified functional of $\mathcal{P}_{\mathcal{DZ}}$. Now, using that $Y_i = Y_i(0) + D_i\Delta_i$ where $\Delta_i := Y_i(1) - Y_i(0)$:

$$\begin{aligned}
\Delta_c = \beta_s &= \{\mathbb{E}[s(D_i,Z_i)Y_i(0)] + \mathbb{E}[s(D_i,Z_i)D_i\Delta_i]\} \\
&= \sum_g P(G_i = g)\left\{\mathbb{E}[s(\mathcal{D}_g(Z_i),Z_i)Y_i(0)|G_i = g] + \mathbb{E}[s(\mathcal{D}_g(Z_i),Z_i)\mathcal{D}_g(Z_i)\Delta_i|G_i = g]\right\} \\
&= \sum_g P(G_i = g)\left(\cancel{\mathbb{E}[s(\mathcal{D}_g(Z_i),Z_i)]}\right)\mathbb{E}[Y_i(0)|G_i = g] \\
&\qquad\qquad + \sum_g P(G_i = g)\left(\mathbb{E}[s(\mathcal{D}_g(Z_i),Z_i)\mathcal{D}_g(Z_i)]\right)\mathbb{E}[\Delta_i|G_i = g] \\
&= \sum_g P(G_i = g)\left(\mathbb{E}[s(1,Z_i)\mathcal{D}_g(Z_i)]\right)\Delta_g
\end{aligned}$$

where we've used independence, and that the crossed out term must be equal to zero for every g by the assumption that $\beta_s = \Delta_c$ for every joint distribution of response groups and potential outcomes compatible with $\mathcal{P}_{DZ}$ in some class (it is always possible to translate the support of the distribution of $Y_i(0)$ and $Y_i(1)$ by the same constant without affecting Δ_i). Finally, $s(\mathcal{D}_g(Z_i), Z_i)\mathcal{D}_g(Z_i) = s(1, Z_i)\mathcal{D}_g(Z_i)$ with probability one, establishing the final equality.

Recall that from Equation (2.3) that Δ_c can also be written as a weighted average of group-specific average treatment effects $\Delta_g = \mathbb{E}[Y_i(1) - Y_i(0)|G_i = g]$ as:

$$\Delta_c = \frac{1}{P(C_i = 1)} \sum_g P(G_i = g)\mathbb{E}[c(g, Z_i)] \cdot \Delta_g$$

Since $\beta_s = \Delta_c$ holds for any vector of $\{\Delta_g\}$ across all of the g for which $P(G_i = g) > 0$ is compatible with $\mathcal{P}_{DZ}$, we can match coefficients within this group to establish that $\mathbb{E}[c(g, Z_i)] = P(C_i = 1)\mathbb{E}[s(1, Z_i)\mathcal{D}_g(Z_i)]$. This set of weights satisfies Property M, since for any $g \in \mathcal{G}^c$:

$$\begin{aligned}
\mathbb{E}[c(g, Z_i)] &= P(C_i = 1)\mathbb{E}[s(1, Z_i) \sum_S [M_J]_{F(g),S}\mathcal{D}_{g(S)}(Z_i)] \\
&= \sum_S [M_J]_{F(g),S} \left(P(C_i = 1)\mathbb{E}[s(1, Z_i)\mathcal{D}_{g(S)}(Z_i)]\right) \\
&= \sum_S [M_J]_{F(g),S} \cdot \mathbb{E}[c(Z_i, g(S))]
\end{aligned}$$

If this holds for any distribution of Z_i satisfying Assumption 3, then we must have $c(g, z) = \sum_S [M_J]_{F(g),S} \cdot c(g(S), z)$ for all $z \in \mathcal{Z}$. To see this, consider a sequence of distributions for Z_i that converges point-wise to a degenerate distribution at any single point z, but satisfies Assumption 3 for each term in the sequence. Applying the dominated convergence theorem to $\mathbb{E}[c(g, Z_i)] - \sum_S [M_J]_{F(g),S} \cdot \mathbb{E}[c(g(S), Z_i)] = 0$ along this sequence, we have that $c(g, z) = \sum_S [M_J]_{F(g),S} \cdot c(g(S), z)$. We can apply a similar argument to establish that $c(a.t., z) = c(n.t., z) = 0$ for all $z \in \mathcal{Z}$ given that $\mathbb{E}[c(g, Z_i)] = P(C_i = 1)\mathbb{E}[s(1, Z_i)\mathcal{D}_g(Z_i)]$ and $\mathbb{E}[s(1, Z_i)] = 0$.

B.4.13 Proof of Proposition 2.9

In the Supplemental Material, I show that with two binary instruments, if PM holds but not VM or IAM, then $\mathcal{G}$ consists of seven response groups, whose definitions are given in the Supplemental Material. We suppose that all 7 groups are possibly present, and the practitioner has knowledge of $\mathbb{E}[Y_i|D_i = d, Z_i = z]$ for all eight combinations of (d, z), as well as the joint distribution of D_i and Z_i. This is equivalent to knowledge of $\mathbb{E}[Y_i D_i|Z_i = z]$ and $\mathbb{E}[Y_i(1 - D_i)|Z_i = z]$ for all $z \in \mathcal{Z}$ and the joint distribution of (D_i, Z_i). Point identification from these moments is in turn equivalent to point identification from a finite set of IV-like estimands, by Proposition B.2.

Using Supplemental Material Table 2, these eight moments can be written in matrix form as

$$\left(\begin{array}{c} \mathbb{E}[Y_i D_i|Z_i = (0,0)] \\ \mathbb{E}[Y_i D_i|Z_i = (0,1)] \\ \mathbb{E}[Y_i D_i|Z_i = (1,0)] \\ \mathbb{E}[Y_i D_i|Z_i = (1,1)] \\ \hline \mathbb{E}[Y_i(1-D_i)|Z_i = (0,0)] \\ \mathbb{E}[Y_i(1-D_i)|Z_i = (0,1)] \\ \mathbb{E}[Y_i(1-D_i)|Z_i = (1,0)] \\ \mathbb{E}[Y_i(1-D_i)|Z_i = (1,1)] \end{array}\right) = \left[\begin{array}{ccccccc|ccccccc} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \end{array}\right] \left(\begin{array}{c} p_{odd} \cdot \mathbb{E}[Y_i(1)|G_i = odd] \\ p_{eager} \cdot \mathbb{E}[Y_i(1)|G_i = eager] \\ p_{reluct.} \cdot \mathbb{E}[Y_i(1)|G_i = reluct.] \\ p_1 \cdot \mathbb{E}[Y_i(1)|G_i = 1only] \\ p_2 \cdot \mathbb{E}[Y_i(1)|G_i = 2only] \\ p_a \cdot \mathbb{E}[Y_i(1)|G_i = a.t.] \\ p_n \cdot \mathbb{E}[Y_i(1)|G_i = n.t.] \\ \hline p_{odd} \cdot \mathbb{E}[Y_i(0)|G_i = odd] \\ p_{eager} \cdot \mathbb{E}[Y_i(0)|G_i = eager] \\ p_{reluct.} \cdot \mathbb{E}[Y_i(0)|G_i = reluct.] \\ p_1 \cdot \mathbb{E}[Y_i(0)|G_i = 1only] \\ p_2 \cdot \mathbb{E}[Y_i(0)|G_i = 2only] \\ p_a \cdot \mathbb{E}[Y_i(0)|G_i = a.t.] \\ p_n \cdot \mathbb{E}[Y_i(0)|G_i = n.t.] \end{array}\right),$$

for some labeling of the instrument values, where the groups "reluctant defiers" and "odd compliers" are defined in the Supplemental Material. If this equation is written as $b = Ax$, where b is the 8×1 vector of identified quantities, and x the 14×1 unknown vector of potential outcome moments (note the matrix A here is not the same as the matrix A defined in Corollary 2.1), then ACL can be written as

$$ACL = \frac{1}{1-p_a-p_n} \cdot \underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & 0 & 0 \end{pmatrix}'}_{:=\lambda} x \tag{B.8}$$

ACL is identified only if the vector λ is in the row space of matrix A (the column space of A'), which follows from the proof of **4** $\rightarrow$ **2** in Proposition B.2. This can be readily verified not to hold, since

$$A'(AA')^{-1}A\lambda \approx \begin{pmatrix} 1.45 & .82 & .82 & .73 & .73 & .18 & 0 & -1.45 & -.73 & -.73 & -.82 & -.82 & 0 \end{pmatrix}$$

where $A'(AA')^{-1}A$ is the orthogonal projector into the row space of A (which has full row rank). Since the RHS of the above is not equal to λ (given explicitly in Eq. B.8), λ is not in the row space of A.

B.4.14 Proof of Proposition B.1

Write the parameter of interest Δ_c as θ_Y/θ_D, where for $V \in \{Y, D\}$, $\theta_V = \tilde{\lambda}'\beta_V$ with $\beta_V := \mathbb{E}[\Gamma_i\Gamma_i']^{-1}\mathbb{E}[\Gamma_i'V_i]$ and $\tilde{\lambda} = (0, \lambda')'$. Denote the estimator $\hat{\rho}(\hat{\lambda}, \alpha)$ as $\hat{\Delta}_c$ for shorthand. It takes the form $\hat{\Delta}_c = \hat{\theta}_Y/\hat{\theta}_D$, where $\hat{\theta}_V := (0, \hat{\lambda}')'(\Gamma'\Gamma + K)^{-1}\Gamma'V$, and $K = \alpha I$. I keep the notation in terms of K as the first part of the argument below will go through with any diagonal matrix of positive entries, allowing a different regularization parameter corresponding to each singular vector of $\Gamma'\Gamma$. Write each $\hat{\theta}_V := (0, \hat{\lambda}')'\hat{\beta}_V^*$ where $\hat{\beta}_V^*$ is the ridge-regression estimate of β_V, and let $\hat{\beta}_V = (\Gamma'\Gamma)^{-1}\Gamma'V$ be the unregularized regression coefficient estimator.

Consider the conditional MSE $M = \mathbb{E}[(\hat{\Delta}_c - \Delta_c)^2|\Gamma]$. It can be rearranged as:

$$\begin{aligned} M &= E\left[\left(\frac{\hat{\theta}_Y}{\hat{\theta}_D} - \frac{\theta_Y}{\theta_D}\right)^2 \middle| \Gamma\right] = \frac{1}{\theta_D^2} E\left[\left((\hat{\theta}_Y - \theta_Y) - \hat{\Delta}_c(\hat{\theta}_D - \theta_D)\right)^2 \middle| \Gamma\right] \\ &= \frac{1}{\theta_D^2} E\left[(\hat{\theta}_Y - \theta_Y)^2 + \hat{\Delta}_c^2(\hat{\theta}_D - \theta_D)^2 - 2\hat{\Delta}_c(\hat{\theta}_Y - \theta_Y)(\hat{\theta}_D - \theta_D) \middle| \Gamma\right] \end{aligned} \tag{B.9}$$

For any $V, W \in \{Y, D\}$, and $m \geq 1$:

$$\begin{aligned} E\left[(\hat{\Delta}_c)^m(\hat{\theta}_V - \theta_V)(\hat{\theta}_W - \theta_W)\middle|\,\Gamma\right] &= E\left[(\hat{\Delta}_c)^m(0,\hat{\lambda})'(\hat{\beta}^*_V - \beta_V)(\hat{\beta}^*_W - \beta_W)'(0,\hat{\lambda})'\middle|\,\Gamma\right] \\ &= (\Delta_c)^m\tilde{\lambda}'E\left[(\hat{\beta}^*_V - \beta_V)(\hat{\beta}^*_W - \beta_W)'\middle|\,\Gamma\right]\tilde{\lambda} + R^m_n \end{aligned}$$

where the first term in the above is viewed as an approximation that ignores terms that are of third or higher order in estimation errors. The asymptotic rate at which the approximation error captured by the R^m_n converges to zero is considered explicitly at the end of this section.

Let $Z = (\Gamma'\Gamma + K)^{-1}\Gamma'\Gamma$ and notice that $\hat{\beta}^*_V = Z\hat{\beta}_V$. Using that $\mathbb{E}[\hat{\beta}_V|\Gamma] = \beta_V$ (as Γ_i includes all products of the instruments the CEF must be linear) for $V \in \{Y, D\}$:

$$\begin{aligned} E\left[(\hat{\beta}^*_V - \beta_V)(\hat{\beta}^*_W - \beta_W)'\middle|\,\Gamma\right] &= ZE\left[(\hat{\beta}_V - \beta_V)(\hat{\beta}_W - \beta_W)'\middle|\,\Gamma\right]Z' + (Z - I)\beta_V\beta_W'(Z - I)' \\ &= (\Gamma'\Gamma + K)^{-1}(\Gamma'\Omega_{VW}\Gamma + K\beta_V\beta_W'K)(\Gamma'\Gamma + K)^{-1} \end{aligned}$$

where we define the $n \times 1$ vector $U_V = V - \Gamma\beta_V$ and $\Omega_{VW} = \mathbb{E}[U_V U_W'|\Gamma]$. Thus, total conditional MSE is, by Equation (B.9):

$$\begin{aligned} M \approx \frac{1}{\theta_D^2}\tilde{\lambda}'(\Gamma'\Gamma + K)^{-1}\big\{&\Gamma'(\Omega_Y + \Delta_c^2\Omega_D - 2\Delta_c\Omega_{YD})\Gamma \\ &+K(\beta_Y\beta_Y' + \Delta_c^2\beta_D\beta_D' - 2\Delta_c\beta_Y\beta_D')K\big\}(\Gamma'\Gamma + K)^{-1}\tilde{\lambda} \end{aligned}$$

This development follows and generalizes that of Hoerl and Kennard (1970), who consider MSE optimal regularization via ridge regression for estimating a single regression vector, under homoscedasticity. Our case targets the ratio $\hat{\theta}_Y/\hat{\theta}_D$ rather than a vector of regression coefficients, and also allows for heteroscedasticity.

We now prove that $\alpha/\sqrt{n} \xrightarrow{p} 0$ if α is chosen to minimize the following "single-step" estimator of the MSE (ignoring the positive factor of θ_D^{-2} that does not depend on K):

$$\hat{M} := \tilde{\lambda}'(\Gamma'\Gamma + K)^{-1}\left\{\Gamma'\left(\hat{\Omega}_Y + \left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)^2\hat{\Omega}_D - 2\left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)\hat{\Omega}_{YD}\right)\Gamma + K\left(\hat{\beta}_Y\hat{\beta}_Y' + \left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)^2\hat{\beta}_D\hat{\beta}_D' - 2\left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)\hat{\beta}_Y\hat{\beta}_D'\right)K\right\}(\Gamma'\Gamma + K)^{-1}\tilde{\lambda}$$

where $\left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)$ is the un-regularized estimator of Δ_c. The problem can be re parameterized as a choice of $b := \alpha/n$, where

$$\begin{aligned}\hat{M}(b) &:= \tilde{\lambda}'\left(\frac{\Gamma'\Gamma}{n} + bI\right)^{-1}\left\{\frac{1}{n}\frac{\Gamma'\left(\hat{\Omega}_Y + \left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)^2\hat{\Omega}_D - 2\left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)\hat{\Omega}_{YD}\right)\Gamma}{n} + b^2\left(\hat{\beta}_Y - \left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)\hat{\beta}_D\right)\left(\hat{\beta}_Y - \left(\frac{\hat{\theta}_Y}{\hat{\theta}_D}\right)\hat{\beta}_D\right)'\right\}\left(\frac{\Gamma'\Gamma}{n} + bI\right)^{-1}\tilde{\lambda} \\ &:= m(b, \hat{\Pi}, \hat{\beta}, \hat{\Sigma}, \hat{\lambda})\end{aligned}$$

where $\hat{\Pi} := \frac{1}{n}\sum_i(\hat{U}_{Yi} - \hat{\theta}_Y/\hat{\theta}_D\hat{U}_{Di})^2\Gamma_i\Gamma_i'$, $\hat{\beta} := (\hat{\beta}_Y - \hat{\theta}_Y/\hat{\theta}_D\hat{\beta}_D)$, and $\hat{\Sigma^*} := \frac{1}{n}\sum_i\Gamma_i\Gamma_i'$. Note that $\hat{\beta} \xrightarrow{p} \beta := \beta_Y - \Delta_c\beta_D$, $\hat{\Sigma}^* \xrightarrow{p} \Sigma^* := \mathbb{E}[(1,\Gamma_i')'(1,\Gamma_i')]$, $\sqrt{n}\left(\hat{\Pi} - \Pi\right) \xrightarrow{d} N(0,V)$ for some V provided that the variance of $(\hat{U}_{Yi} - \hat{\theta}_Y/\hat{\theta}_D\hat{U}_{Di})^2\Gamma_i\Gamma_i'$ exists, where $\Pi := \mathbb{E}[(\hat{U}_{Yi} - \hat{\theta}_Y/\hat{\theta}_D U_{Di})^2\Gamma_i\Gamma_i']$. The function m is

$$m(b, \Pi/n, \beta, \Sigma^*, \lambda) = (0,\lambda')\left(\Sigma^* + bI\right)^{-1}\left\{\Pi/n + b^2\beta\beta'\right\}\left(\Sigma^* + bI\right)^{-1}(0,\lambda')'$$

We wish to show that $\sqrt{n}b = \alpha/\sqrt{n} \xrightarrow{p} 0$, when b is chosen as the smallest positive minimizer of $m(\cdot, \hat{\Pi}/n, \hat{\beta}, \hat{\Sigma}, \hat{\lambda})$. The strategy will be to show that $nb \xrightarrow{p} X$ where X is a finite degenerate random variable. Since Π and $\beta\beta'$ are positive definite, it is clear that $m(b, \Pi/n, \beta, \Sigma^*, \lambda)$ is weakly positive for any choice of b. Further, $m(b, \Pi/n, \beta, \Sigma^*, \lambda)$ is typically strictly positive at $b = 0$, and it can also be seen that $\lim_{b\to\infty} m(b, \Pi/n, \beta, \Sigma^*, \lambda) = 0$ (see Section B.3.1 for discussion). However, m is generally not monotonically decreasing in between, as we shall see below.

Observe that $b = 0$ minimizes $m(b, \mathbf{0}, \beta, \Sigma^*, \lambda)$ with respect to b regardless of the values of β, Σ^*, λ, where **0** is a $k \times k$ matrix of zeros (the dimension of Π), since $m(\cdot)$ is always positive and when its second argument vanishes can be made equal to zero by choosing $b = 0$. Furthermore, $b = 0$ is a local minimizer when $\Pi/n = \mathbf{0}$, since m_b vanishes when evaluated at $(0, \mathbf{0}, \beta, \Sigma^*, \lambda)$–see below, while the second derivative of m with respect to b, evaluated at $(0, \mathbf{0}, \beta, \Sigma^*, \lambda)$, is equal to

$$(0, \lambda')\Sigma^{*-1}\beta\beta'\Sigma^{*-1}\lambda = \Big((0, \lambda')\Sigma^{*-1}\beta\Big)^2$$

up to a strictly positive constant. We have assumed that the quantity in parenthesis is non-zero. By the implicit function theorem, there then exists a unique function $g(\Pi/n; \beta, \Sigma^*, \lambda)$ such that $g(\mathbf{0}; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) = 0$ and $m_b(g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}), \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) = 0$, in a neighborhood $\mathcal{N}$ of the probability limits $(\mathbf{0}, \beta, \Sigma^*, \lambda)$ of $(\hat{\Pi}/n, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})$, and this function is continuously differentiable with respect to all parameters, (including, in particular, the elements of Π). Since the second derivative of m is strictly positive at $(0, \mathbf{0}, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})$ and continuous with respect to all arguments, $\mathcal{N}$ can furthermore be chosen such that the critical point at $(g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}), \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})$ is always a local minimum within $\mathcal{N}$.

Since for any realization of $\hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}$:

$$m_b(0, \mathbf{0}, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) = 2\tilde{\lambda}'(\hat{\Sigma}^* + bI)^{-1}\left\{bI - b^2(\hat{\Sigma}^* + bI)^{-1}\right\}\hat{\beta}\hat{\beta}'(\hat{\Sigma}^* + bI)^{-1}\tilde{\lambda}\Big|_{b=0} = 0$$

we see that m has a critical point at $b = 0$ for values $(\mathbf{0}, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})$ of the other arguments. By uniqueness of the function $g(\Pi/n; \beta, \Sigma^*, \lambda)$, this implies then that $g(\mathbf{0}, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) = 0$. By the mean value theorem, we can write

$$g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) = g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) - g(\mathbf{0}, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})$$
$$= \frac{\partial}{\partial x} g(vec(cn^{-1}\hat{\Pi}); \hat{\beta}, \hat{\Sigma}^*, \hat{\lambda}) \cdot \frac{vec(\hat{\Pi})}{n}$$

for some $c \in [0,1]$, where $vec(\Pi)$ denotes the vectorization x of the matrix Π, and we let $\frac{\partial}{\partial x} g(x; \beta, \Sigma^*, \lambda)$ denote a gradient of g with respect to that vector (recall that existence of the derivative is a consequence of the implicit function theorem). By continuity of $\frac{\partial}{\partial x} g(x; \beta, \Sigma^*, \lambda)$ and the continuous mapping theorem then,

$$n \cdot g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}) \xrightarrow{p} \frac{\partial}{\partial x} g(\mathbf{0}, \beta, \Sigma^*, \lambda) vec(\Pi) \tag{B.10}$$

which is a finite scalar.

To complete the proof, we now simply note that with probability approaching unity, $(\hat{\Pi}/n, \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})$ is within the neighborhood $\mathcal{N}$, and thus if b is chosen as the smallest positive local minimizer of $m(b, \hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})$ we have that $b = g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})$. We have now established the result, since for any $B > 0$:

$$P(|\alpha/\sqrt{n}| > B) \leq P(|\alpha/\sqrt{n}| > B \text{ and } b = g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})) + P(b \neq g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}))$$
$$= P(|n \cdot g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda})| > \sqrt{n}B) + P(b \neq g(\hat{\Pi}/n; \hat{\beta}, \hat{\Sigma}^*, \tilde{\lambda}))$$
$$\xrightarrow{n} 0 + 0$$

Finally, I consider the error involved in the approximation made to Equation (B.9). Write this as:

$$
\begin{aligned}
R_n &:= R_n^m + R_n^m = \\
&= \frac{1}{\theta_D^2}\tilde{\lambda}'(\Gamma'\Gamma + K)^{-1}\left\{(\hat{\Delta}_c^2 - \Delta_c^2)(\Gamma'\Omega_D\Gamma + K\beta_D\beta_D'K)\right. \\
&\qquad\qquad\qquad\qquad \left.-2(\hat{\Delta}_c - \Delta_c)(\Gamma'\Omega_{YD}\Gamma + K\beta_Y\beta_D'K)\right\}(\Gamma'\Gamma + K)^{-1}\tilde{\lambda} \\
&= \frac{1}{\theta_D^2\cdot n^{3/2}}\cdot\tilde{\lambda}'\left(\frac{\Gamma'\Gamma}{n} + \frac{K}{n}\right)^{-1}\left\{\sqrt{n}(\hat{\Delta}_c^2 - \Delta_c^2)\left(\frac{\Gamma'\Omega_D\Gamma}{n} + \frac{K}{\sqrt{n}}\beta_D\beta_D'\frac{K}{\sqrt{n}}\right)\right. \\
&\qquad\qquad \left.-2\sqrt{n}(\hat{\Delta}_c - \Delta_c)\left(\frac{\Gamma'\Omega_{YD}\Gamma}{n} + \frac{K}{\sqrt{n}}\beta_Y\beta_D'\frac{K}{\sqrt{n}}\right)\right\}\left(\frac{\Gamma'\Gamma}{n} + \frac{K}{n}\right)^{-1}\tilde{\lambda}
\end{aligned}
$$

Provided that $\alpha/\sqrt{n} \xrightarrow{p} 0$ as above, we will show in Theorem B.1 that $\hat{\Delta}_c$ is $\sqrt{n}$-consistent for Δ_c. In this case, the approximation error term is $O_p(n^{-3/2})$.

B.4.15 Proof of Theorem B.1

When $\alpha_n = 0$, the result follows from Theorem 3 of Imbens and Angrist (1994). To see that $o_p(\sqrt{n})$ regularization has no asymptotic effect, note that

$$
\begin{aligned}
(0,\hat{\lambda}')'(\Gamma'\Gamma + \alpha I)^{-1}\Gamma'Y &= (0,\hat{\lambda}')'(\Gamma'\Gamma + \alpha I)^{-1}(\Gamma'\Gamma + \alpha I - \alpha I)(\Gamma'\Gamma)^{-1}\Gamma'Y \\
&= (0,\hat{\lambda}')'(\Gamma'\Gamma)^{-1}\Gamma'Y - \alpha(0,\hat{\lambda}')'(\Gamma'\Gamma + \alpha I)^{-1}(\Gamma'\Gamma)^{-1}\Gamma'Y
\end{aligned}
$$

and similarly for D, thus:

$$
\begin{aligned}
\rho(\hat{\lambda},\alpha) &= \frac{(0,\hat{\lambda}')'(\Gamma'\Gamma)^{-1}\Gamma'Y - \alpha(0,\hat{\lambda}')'(\Gamma'\Gamma + \alpha I)^{-1}(\Gamma'\Gamma)^{-1}\Gamma'Y}{(0,\hat{\lambda}')'(\Gamma'\Gamma)^{-1}\Gamma'D - \alpha(0,\hat{\lambda}')'(\Gamma'\Gamma + \alpha I)^{-1}(\Gamma'\Gamma)^{-1}\Gamma'D} \\
&= \frac{\widehat{Cov}(g(Z_i,\hat{\theta}),Y_i) - \frac{\alpha}{n}(0,\hat{\lambda}')'(\frac{1}{n}\Gamma'\Gamma + \frac{\alpha}{n}I)^{-1}(\frac{1}{n}\Gamma'\Gamma)^{-1}\frac{1}{n}\Gamma'Y}{\widehat{Cov}(g(Z_i,\hat{\theta}),D_i) - \frac{\alpha}{n}(0,\hat{\lambda}')'(\frac{1}{n}\Gamma'\Gamma + \frac{\alpha}{n}I)^{-1}(\frac{1}{n}\Gamma'\Gamma)^{-1}\frac{1}{n}\Gamma'D} \\
&= \frac{\widehat{Cov}(g(Z_i,\hat{\theta}),Y_i)}{\widehat{Cov}(g(Z_i,\hat{\theta}),D_i)} + \frac{\alpha}{n}\cdot\frac{(0,\hat{\lambda}')'(\frac{1}{n}\Gamma'\Gamma + \frac{\alpha}{n}I)^{-1}(\frac{1}{n}\Gamma'\Gamma)^{-1}\left\{\frac{1}{n}\Gamma'D\cdot\frac{\widehat{Cov}(g(Z_i,\hat{\theta}),Y_i)}{\widehat{Cov}(g(Z_i,\hat{\theta}),D_i)} - \frac{1}{n}\Gamma'Y\right\}}{\widehat{Cov}(g(Z_i,\hat{\theta}),D_i) - \frac{\alpha}{n}(0,\hat{\lambda}')'(\frac{1}{n}\Gamma'\Gamma + \frac{\alpha}{n}I)^{-1}(\frac{1}{n}\Gamma'\Gamma)^{-1}\frac{1}{n}\Gamma'D}
\end{aligned}
$$

and thus the asymptotic distribution of $\sqrt{n}(\hat{\rho}(\hat{\lambda},0) - \Delta_c)$ is the same as that of $\sqrt{n}\left(\frac{\widehat{Cov}(g(Z_i,\hat{\theta}),Y_i)}{\widehat{Cov}(g(Z_i,\hat{\theta}),D_i)} - \Delta_c\right)$, provided that $\alpha_n/\sqrt{n} \xrightarrow{p} 0$ (in which case the second term above is $o_p(n^{-1/2})$).

Appendix C: Supplements to Chapter 3

C.1 Formal treatment of the RSD mechanism

In this appendix we provide a formal definition of the Random Serial Dictatorship Mechanism in order to motivate our instrumental variables analysis. We do so by adopting the notion of a continuum economy from Abdulkadiroğlu et al. (2017) that provides an approximation in large samples. Throughout this section we focus on a single instance of the lottery, and suppress the lottery indicator L_i.

C.1.1 RSD mechanism

Recall the notation from Section 3.4 that we let i index individual doctors within some population $\mathcal{I}$. We now consider two cases, one in which $\mathcal{I}$ is a finite set of n doctors (referred to as the *finite economy*), and another in which $\mathcal{I}$ is considered to be the unit interval (referred to as the *continuum economy*). In either case, we assume that there is a fixed set $\mathcal{H}$ of hospitals, and that doctors are indifferent between (residency) jobs within a hospital. We assume each doctor i has a well-defined preference ordering $\succ_i$ over hospitals $h \in \mathcal{H}$. Each hospital each can accommodate proportion q_h of the doctors in that year. That is, in the finite case hospital h has $Q_h = nq_h$ positions available, and in the continuum case hospital h can accomodate proportion q_h of the unit measure of doctors. In our actual data, we have ~ 250 doctors per year across ~ 60 hospitals, so a typical q_h can be thought of as being in the vicinity of $1/60$.

RSD begins by allocating each doctor a lottery number R_i, taken to lie in the unit interval $[0,1]$ (see Section 3.4). While in our setting the RSD mechanism is decentralized (doctors indicate a selection from their choice set in real time before the procedure moves to the next doctor), it delivers the same outcome as one in which all doctors submit their full preference ordering $\succeq_i$ and allocations are made centrally, under the assumption that doctors choose according to well-defined and stable such preferences. In this framework, RSD is a special case of the deferred acceptance (DA) mechanism, in which hospitals have no priorities over doctors beyond lottery number. Abdulkadiroğlu et al. (2017), characterize DA in terms of a set of cutoffs $\tau = \{\tau_h\}_{h \in \mathcal{H}}$ such that hospital h is available to i iff $R_i \leq \tau_h$ (and thus i would be centrally assigned to h if they prefer h to any other h' such that $R_i \leq \tau_{h'}$).

Let the *type* θ_i of doctor i be the tuple of their demographic group, potential outcomes and preferences: $\theta_i = (G_i, \mathcal{Y}_i, \succ_i)$, and let Θ be the possible values of θ_i. In a finite economy, the cutoffs τ arising from RSD are determined by the set of pairs $\{(\theta_i, R_i)\}_{i=1\dots n}$. Given a fixed realization of the lottery, we may equivalently represent this set by a discrete uniform distribution over the pairs (θ_i, R_i) in the economy. Denote this probability distribution as F_n. Taking $\mathcal{I}$ to be the underlying sample space, we write $F_n(\mathcal{I}_0) = |\mathcal{I}_0|/n$ for any set $\mathcal{I}_0 \subseteq I$ of individuals. In the continuum economy, we again work in terms of a distribution over pairs (θ_i, R_i), denoted as F_0. To construct F_0 begin with an underlying "population" distribution F_0^θ over types, and take the product measure with a uniform $U[0, 1]$ measure for the lottery draws. This allows for a unified probability space both over individuals and over lottery draws, while maintaining independence between θ_i and R_i.

As described in Abdulkadiroğlu et al. (2017), the cutoffs can be expressed as $\tau_h = \lim_{t\to\infty} \tau_h^t$ where we imagine a set of "rounds" t in which initially all hospitals are available: $\tau_h^0 = 1$ for all h, and in subsequent rounds the thresholds are lowered for hospitals that were "over-subscribed" given last rounds' thresholds. With F equal to either F_n or F_0, we can write this as:

$$\tau_h^{t+1} = \begin{cases} 1 & \text{if } F(Q_h(\tau^t)) < q_h \\ \max\{t \in [0, 1] : F(Q_h(\tau^t) \cap \{i : R_i \leq t\}) \leq q_h\} & \text{if } F(Q_h(\tau^t)) \geq q_h \end{cases} \tag{C.1}$$

where

$$Q_h(\tau) = \{i : R_i \leq \tau_h \text{ and } h \succeq_i h' \text{ for all } h' \text{ s.t. } R_i \leq \tau_{h'}\}$$

is the set of doctors who prefer hospital h from their choice set.[1] Intuitively, Equation (C.1) reduces the threshold for each oversubscribed hospital to the largest value t such that it is no longer over-capacity, ignoring indirect effects of this change from space in other hospitals being made available by the doctors who will now newly choose h. We can write the final choice set C_i for doctor i as a function of their lottery number and the final vector of cutoffs: $C_i = \{h \in \mathcal{H} : R_i \leq \tau_h\}$.

C.1.2 Asymptotics

We will motivate choices of instruments and treatment based on observations about the continuum economy, which can be expected to provide an accurate approximation to finite economies of sufficient size n. Recall that in either case, the thresholds characterizing the outcome of RSD are determined by the function F (either F_n or F_0, depending on the case).

Fix a continuum economy with joint distribution F_0 over types and lottery numbers, and vector of hospital capacities q. Formally, we will view a finite economy as a random sample $\{\theta_i, R_i\}_{i=1\ldots n}$ of n individuals from this fixed continuum economy. Let F_n be the empirical distribution over (θ_i, R_i), noting that this coincides with the definition of F_n given in the last section. Asymptotic arguments will consider a sequence of such $\{F_n\}$ with increasing sample size. By the Glivenko-Cantelli theorem, $F_n \to F_0$ almost surely. This will provide a basis for consistent estimation of "population" quantities defined with respect to the continuum economy.

For a given sample, the econometrician observes a realized outcome Y_i, doctor i's

[1] Abdulkadiroğlu et al. (2017) define the function F over sets taking the form $\mathcal{I}_0 = I(\Theta_o, r_0) := \{i \in I : \theta_i \in \Theta_0, R_i \leq r_0\}$ for any subset $\Theta_0 \subset \Theta$. $F(\mathcal{I}_0)$ is again defined as $|\mathcal{I}_0|/n$ in the finite case, and as $P(\theta_i \in \Theta_0) \cdot r_0$ in the continuum case. However, as the set $Q_h(\tau)$ does not take this form, we do not pursue this definition here.

choice of job H_i and rank R_i according to the lottery. Their choice set can be imputed as

$$C_i = \{h \in \mathcal{H} \text{ such that } |\{j : H_j = h \text{ and } R_j < R_i\}| < Q_h\}$$

where Q_h is the actual number of job openings at hospital h (the continuum economy is defined to have $q_h = Q_h/n$ for the actual sample size n, in our case ~ 250). Let $R_i = R_i/n$ be lottery numbers normalized to the unit interval. The econometrician can compute empirical cutoffs as the largest lottery number such that hospital h is available:

$$\hat{\tau}_h = \max\{R_i : h \in C_i\}_{i=1\ldots n}$$

Let $\hat{\tau}$ be the vector of cutoffs for a given sample of size n. Lemma 3 of Abdulkadiroğlu et al. (2017) shows that $\hat{\tau} \overset{a.s.}{\to} \tau$, where τ is the set of cutoffs arising from the continuum economy F. This property will prove useful in the following section.

C.1.3 Independence of choice sets

Recall that Assumption 3.1 from Section 3.4 does generally not hold exactly in a finite economy (c.f. footnote 16). However, we can justify it by appealing to the continuum economy. One way to think about the example given in footnote 16 is that in a finite economy, the probability distribution over C_i depends on $\mathcal{P}_i = \{\succeq_j\}_{j \in I, j \neq i}$, the set of preferences of all individuals in the economy that are not i. Each of these other individuals has the opportunity to choose before i for some realizations of the lottery, thus having an impact on the choices remaining for i. As n gets large, it is reasonable to expect the magnitude of this effect to attenuate, as $\mathcal{P}_i$ and $\mathcal{P}_{i'}$ become nearly the same "overall" for any $i \neq i'$. We can now formalize this notion, based on the asymptotic sequence of economies introduced in the last section.

Let us now consider a binary "instrument" Z_i that indicates a particular value of i's choice set: $Z_i = \mathbb{1}(C_i = c)$ for some fixed $c \subseteq \mathcal{H}$. For any economy (whether finite or continuum) with cutoff vector τ, we can write $Z_i = f(R_i, \tau)$ where

$$f(R_i, \tau) := \mathbb{1}(\forall h \in c : R_i \leq \tau_h \text{ and } \forall h \notin c : R_i > \tau_h)$$

For each n, let τ_n be the cutoffs according to F_n. Let $Z_i^n = f(R_i, \tau_n)$ denote the instrument defined with respect to the finite economy's cutoffs τ_n, and let $Z_i = f(R_i, \tau)$ represent the "population" analog defined with respect to the continuum limiting cutoffs τ.

Proposition C.1 independence in the continuum economy. $Z_i \perp \theta_i$.

Proof. Immediate, since with τ fixed, Z_i is a measurable function of R_i, and $R_i \perp \theta_i$. $\square$

Since the above holds for any c, and the events $C_i = c$ and $C_i = c'$ are exclusive for $c \neq c'$, Proposition C.1 implies that $C_i \perp \theta_i$, with C_i interpreted with respect to the continuum economy. Thus Assumption, 3.1 as stated in Section 3.4, since $\theta_i = (G_i, \succ_i, \mathcal{Y}_i)$ and thus Proposition C.1 implies that $\{(\mathcal{Y}_i, \succ_i) \perp C_i\} | G_i$ (recall that lottery L_i is conditioned on implicitly in this section).

Intuitively, Proposition C.1 makes use of the notion that with a continuum of doctors and a continuum of positions available at each hospital, any two doctors A and B share the same function that maps lottery numbers to choice sets. With any single doctor a measure zero set from a continuum of doctors, $\mathcal{P}_i$ defined above does not differ between doctors. Note that we do not have an analog of Proposition C.1 for the finite economy, since the finite economy cutoffs τ_n depend on F_n, itself a random quantity. As F_n is not independent of θ_i for any fixed i in the realized sample, we cannot expect $Z_i^n = f(R_i, \tau_n)$ to be exactly, except in special cases.

Furthermore, even Z_i^n itself is not directly observed. In a finite sample, we can only compute the estimate $\hat{Z}_i^n = f(R_i, \hat{\tau})$ which uses the empirical cutoffs $\hat{\tau}$ observed in the sample. Nevertheless, $\hat{Z}_i^n$ becomes close to the unobserved ideal instrument Z_i for large n, as the empirical cutoffs τ_n approach their continuum analogs τ. To establish consistency of standard IV estimators, it is sufficient to apply the following Lemma to each of the random variables $V_i \in \{Y_i, D_{hi}\}$:

Proposition C.2. *For any V_i such that $E[V_i^2] < \infty$, $\frac{1}{n}\sum_{i=1}^{n} V_i \hat{Z}_i^n \xrightarrow{p} P(C_i = c)E[V_i|C_i = c]$*

Proof. See Appendix C.4. □

Note that Proposition C.2 also implies that $\frac{1}{n}\sum_{i=1}^{n} V_i(1 - f(R_i, \hat{Z}_i^n) \xrightarrow{p} P(C_i \neq c)E[V_i|C_i \neq c]$, since $\frac{1}{n}\sum_{i=1}^{n} V_i(1 - \hat{Z}_i^n) = \frac{1}{n}\sum_{i=1}^{n} V_i - \frac{1}{n}\sum_{i=1}^{n} V_i \hat{Z}_i^n$ and $\frac{1}{n}\sum_{i=1}^{n} V_i \xrightarrow{p} E[V_i]$ by the weak law of large numbers. Then apply the law of iterated expectations over Z_i.

C.2 Simulation evidence on the random choice-set approximation

In this section we present simulation evidence that the asymptotic approximation in Section is a reasonable one in our context.

The simulation DGP constructs an environment with 235 doctors and 60 hospitals, roughly matching a typical year from our data. Hospitals have between 2 and 6 spots available, with a distribution reported in Figure C.1, again intended to match the empirical setting. The total number of spots is 258, allowing each doctor to be placed.

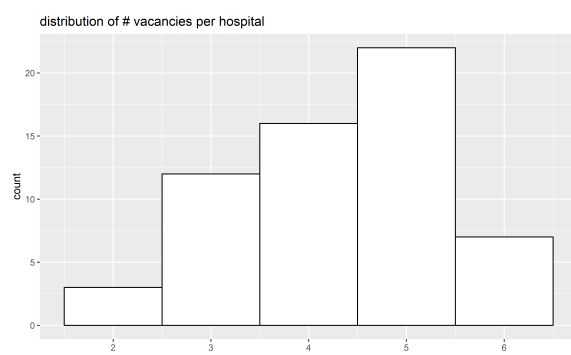

Figure C.1: Simulation distribution of number of spots Q_h in hospital h, across the 60 hospitals.

Preferences over hospitals for the 235 doctors are constructed by first introducing two types of hospitals to create a structured basis for preference heterogeneity. 80% of hospitals (47 in total) are "urban", and the remaining 20% "rural". There are two broad types of doctors, those who typically prefer urban hospitals, and those who typically prefer rural ones. 90% of the doctors (199 in total) are urbanites. Within each doctor type, we introduce a "typical" ordering over hospitals, which places all urban hospitals ahead of any rural hospitals, or vice versa. This is meant to reflect a standard ranking over which hospitals are a good place to live/work. Doctors are indifferent between spots in the same hospital.

For 75% of doctors of each type, we start with the archetype ordering for that type and perturb it by performing a series of random swaps of adjacent hospitals in the ordering. Swaps occur with an increasing probability further down the list, reflecting the notion that there is the most agreement among the most desired hospitals, and more heterogeneity among less desired options.[2] Since swaps may permute urban with non-urban hospitals, this procedure softens slightly the constraint that all urbanite doctors prefer all urban hospitals to all rural hospitals. The remaining 25% of doctors within each type receive a completely random preference ordering within hospital type, then ordered lexicographically across urban/rural.

The simulation then proceeds by running the RSD lottery 500,000 times, and allocating doctors to hospitals based upon their preferences. Since there are enough jobs for all of the doctors, and none prefer an outside option (by construction), all doctors receive a position. Figures C.2 and C.3 compare the distribution across simulation runs of features of the choice sets facing two doctors: "Doctor 1" and "Doctor 2". Doctor 1 is a single randomly chosen urbanite doctor, and Doctor 2 is a single ranomly chosen non-urbanite doctor. If choice sets are unconditionally random, then the doctors should face the same probability distribution over any function of their choice set. In both cases, statistical tests reject this null-hypothesis. However, the figures reveal that the differences are quite minor, almost imperceptible without close inspection. This provides evidence that the asymptotic approximation of choice-set independence is likely to be quite reasonable in

[2]Specifically, a vector of 30 "swap positions" are introduced through the preference list, with a CDF increasing as the square of number in the list. For each swap position j, a random draw determines whether the hospital in that position is swapped with the one below, the one above, or no change is made.

our context.

Figure C.2 compares the distributions facing the two doctors over lottery draws of the proportion of hospitals in their choice set that are urban. If these distributions were substantially different from one another, it would cast doubt on using something like the proportion urban of one's choice set as an instrument for choosing an urban hospital. In particular, it would suggest that this instrument is not indpeendent of preferences, since the only difference between Doctors 1 and 2 in the simulation are their preferences (recall that Doctor 1 prefers urban hospitals and Doctor 2 rural ones). A two sample Kolmogorov-Smirnov test strongly rejects the null that the two distributions are identical, which is not surprising given the large number of simulation draws. Statistics of the distribution also differ: for example the average proportion urban for Doctor 1 is 68.0% vs. 67.3% for Doctor 2. Nevertheless, the differences are quite small in practical terms, as evident in the histograms. Across all of the 235 Doctors, the minimum value of the average proportion urban is 67.3%, and its maximum is 68.0%.

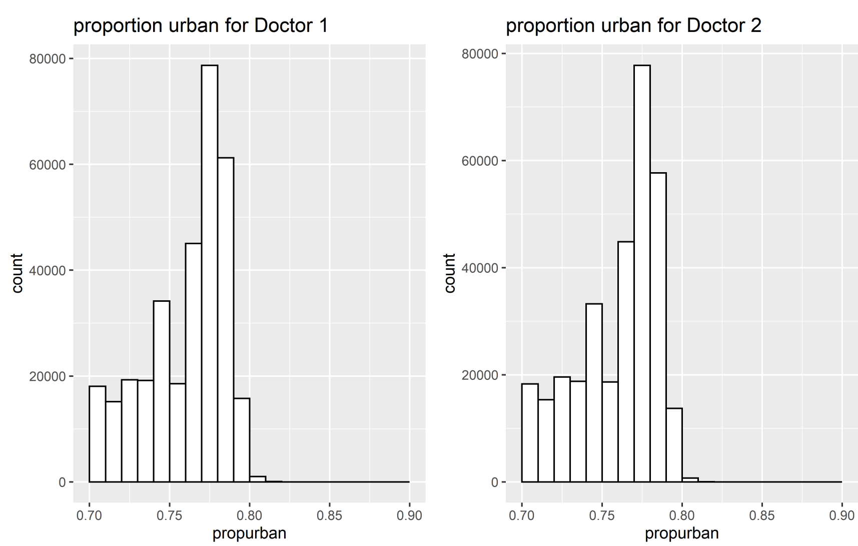

Figure C.2: Simulation distribution over the proportion of one's choice set that is urban hospitals, between Doctors 1 and 2)

As a benchmark, Figure C.3 compares the distribution over number of distinct hospitals present in each of the two doctors' choice sets. Given that there is no statistical relationship between hospital size Q_h and whether h is urban or rural, we might expect $|C_i|$ to be independent of θ_i in this case, even in a finite sample. A chi-squares test also strongly rejects this null hypothesis, however the distributions appear nearly identical in practical terms.

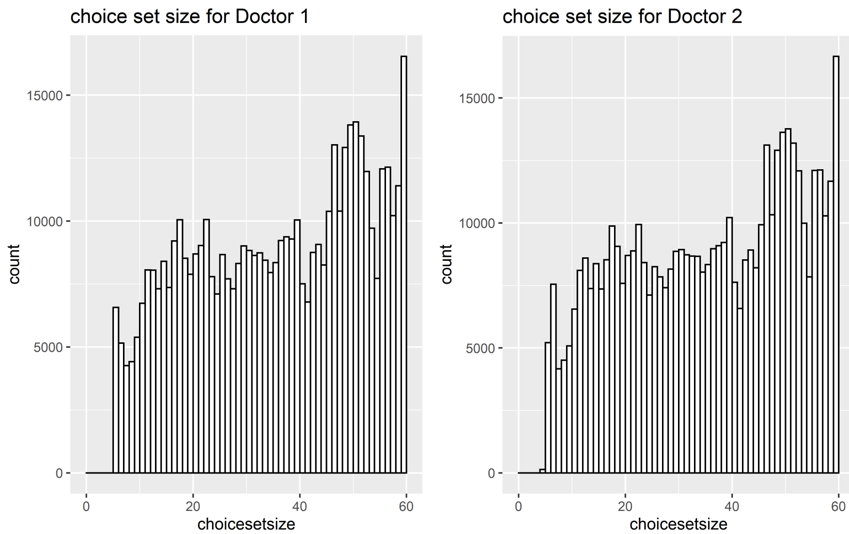

Figure C.3: Comparison of the simulation distribution over number of hospitals in one's choice set, between Doctors 1 and 2)

C.3 Additional tables and figures

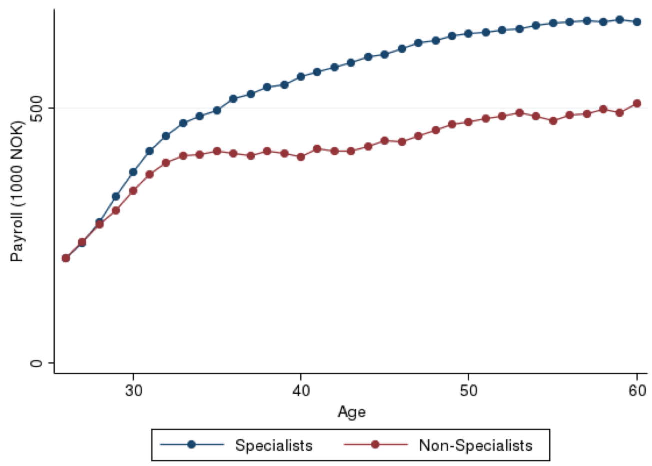

Figure C.4: Average income by age among specialists and non-specialists.

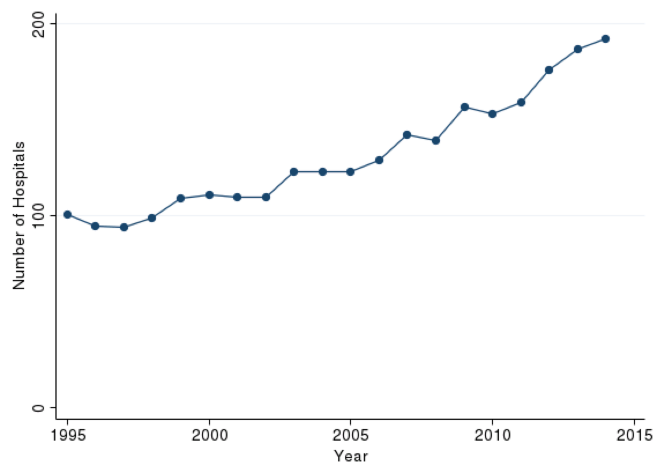

Figure C.5: Number of hospitals by year.

Hospital Characteristics	Good Hospitals		Other Hospitals	
	Mean	SD	Mean	SD
Number of doctors	186.92	198.88	47.53	47.13
Mean doctor experience (years)	12.07	2.28	10.88	3.36
Number of specialists	90.46	104.63	78.21	88.24
Proportion of doctors who are specialists	0.54	0.10	0.51	0.15
Mean doctor income	161.47	30.05	163.55	34.76
Average doctor age	43.75	2.04	44.80	4.41
Proportion of male doctors	0.67	0.10	0.72	0.15
Proportion of foreign doctors	0.27	0.14	0.38	0.21
Observations	703		698	

Doctor Characteristics	Female		Male	
	Mean	SD	Mean	SD
Age	39.19	9.80	45.73	11.40
Cohabit	0.63	0.48	0.75	0.43
Number of Children	1.05	1.14	0.97	1.18
Born Abroad	0.23	0.42	0.20	0.40
After-Tax Income	71.51	49.83	90.91	135.37
Real Estate	30.68	46.74	48.49	55.24
Debt	133.98	163.10	209.00	255.96
Specialization				
Specialist	0.34	0.47	0.56	0.50
General Practice	0.09	0.28	0.12	0.32
Internal Medicine	0.13	0.34	0.18	0.39
Surgery	0.05	0.22	0.18	0.39
Observations	80,025		134,458	

Table C.1: This table presents summary statistics using annual data on hospitals and doctors in Norway during 1995-2011. Hospitals (and doctors) that are observed twice will count as two separate observations, since observable characteristics may change over time. Rural location is defined as the proportion of population in the municipality that lives in rural areas. Doctor income is in thousands in 2011 USD.

Individual Characteristic	ω_C	t-stat	R^2	N
Male	0.003	0.561	0.082	9828
Age	-0.001	-1.370	0.082	9828
Rural Residence at Age 15	-0.011	-0.846	0.085	8267
Born Abroad	-0.005	-0.655	0.082	9828
Study Abroad	-0.0002	-0.021	0.095	2871

Table C.2: This table presents evidence that the lottery number was not influenced by doctor characteristics. Each row presents estimates from a separate regression, where the dependent variable is the lottery draw number normalized to lie between 0 and 1, and the independent variable is an individual (doctor) characteristic. Regressions include lottery fixed effects to allow for demographic changes in the participant pool over time. The number of observations is much lower for the last row because data on study location is only available for the last few years of the sample.

g	h	$[\hat{\theta}^L_{gh}, \hat{\theta}^U_{gh}]$				$\mathcal{C}_n$	
	(category)	$\kappa_n = 0$		$\kappa_n = 10\%$		(95% CI)	
Women	1	0.50	0.50	0.49	0.60	0.26	3.95
	2	1.00	1.00	0.99	1.10	0.26	4.74
	3	0.63	0.71	0.62	0.78	0.53	4.74
	4	0.86	0.88	0.85	1.12	0.26	5.00
Men	1	0.79	0.79	0.74	0.91	0.26	4.74
	2	0.50	0.50	0.49	0.68	0.53	3.68
	3	1.00	1.03	1.00	1.13	0.53	3.95
	4	1.50	1.50	1.50	1.56	0.26	5.00

Table C.3: First job effects $\mu_{gh} = \mathbb{E}[Y_i(h)|G_i = g]$ for career number of specializations. Table reports estimates of the identified set $[\theta^L_{gh}, \theta^U_{gh}]$ and 95% confidence intervals for μ_{gh}.

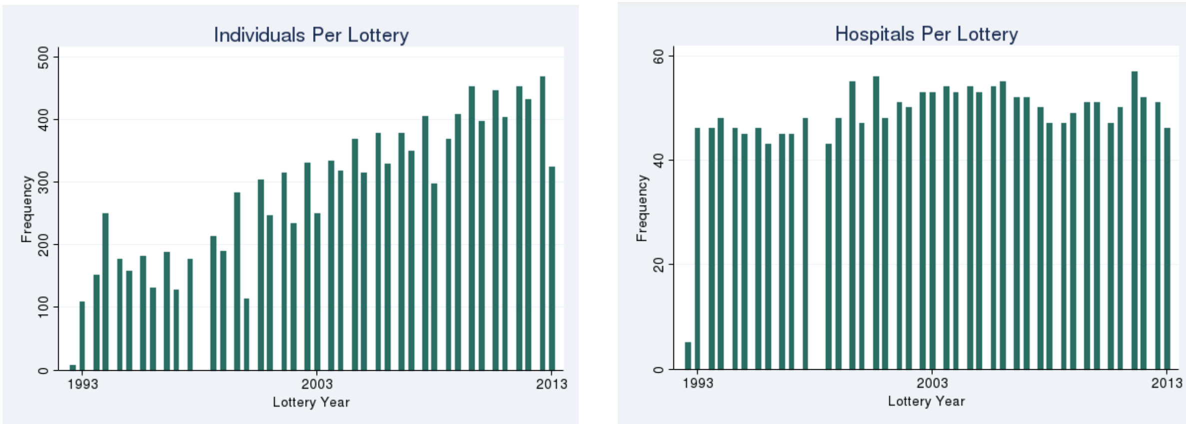

Figure C.6: Number of individuals and hospitals by lottery.

Figure C.7: Number of individuals and hospitals by lottery.

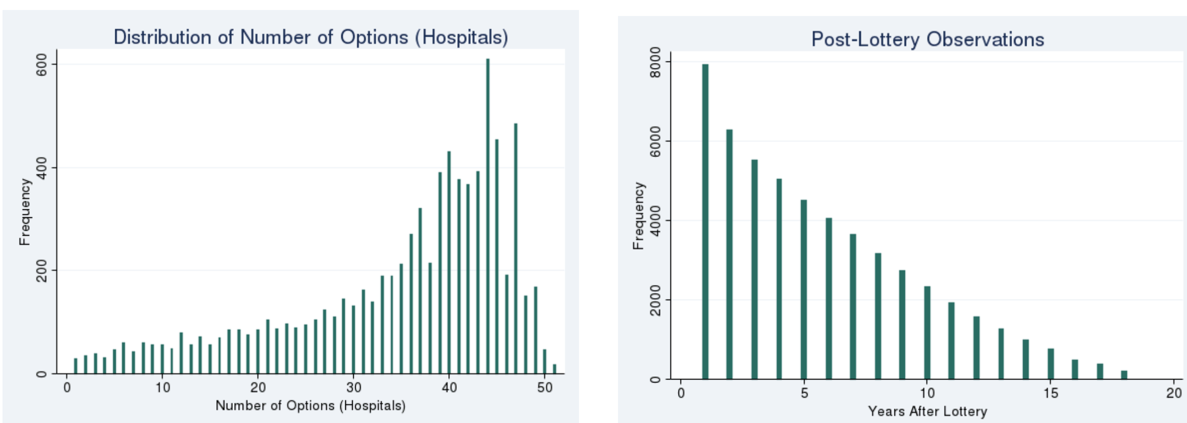

Figure C.8: Distribution of the number of choices for residency hospital $|C_i|$ (left), and the number of observations post-residence for a given doctor (right).

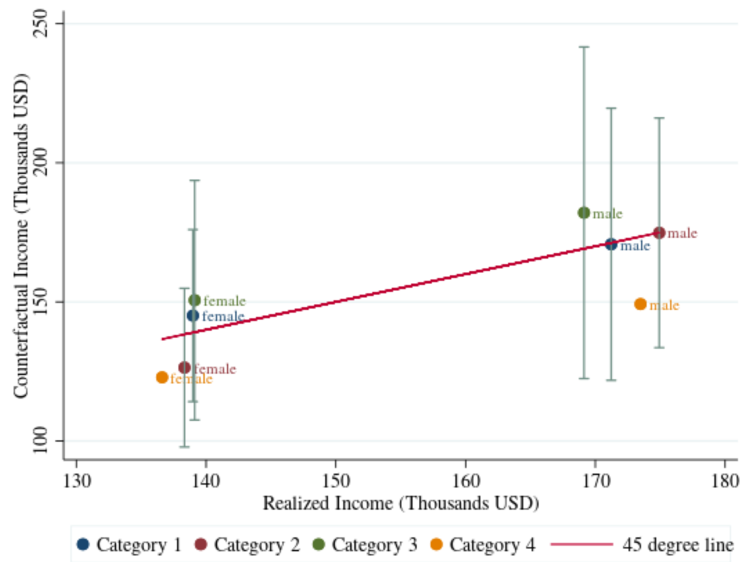

Figure C.9: Earnings FJE's μ_{gh} vs. average realized earnings $\mathbb{E}[Y_i|H_i = h, G_i = g]$, with 45 degree line in red. Brackets represent 95% intervals on the *difference* between μ_{gh} and μ_{g4} for Category 4.

C.4 Proofs

C.4.1 Proof of Proposition 3.2

This proof follows the logic of Kolesar (2015) from the binary treatment case. Assumptions 3.1 implies that

$$(Y_i(h) \perp C_i) \, |G_i, L_i \tag{C.2}$$

where L_i is the lottery/cohort of doctor i. Assumption 3.2 says that

$$\mathbb{E}[Y_i(h) - Y_i(h_0)|H_i = h', \mathcal{C}_i = c, L_i = \ell, G_i = g] = \beta_{hg} \tag{C.3}$$

where β_{hg} is a number that does not depend on h', h_0, or x (or c). Fix an arbitrary choice of h_0, which will serve as a comparison hospital throughout intermediate steps of the proof.

By the law of iterated expectations over H_i and C_i, β_{hg} must be equal to $\mu_{hg} - \mu_{h_0 g}$.

Now substituting $h' = h$ into Equation (C.3), we have:

$$\mathbb{E}[Y_i(h) - Y_i(h_0)|H_i = h, \mathcal{C}_i = c, L_i = \ell, G_i = g] = \beta_{hg} \text{ for all } h, c.$$

Collect the β_{hg} across h for a fixed g into a vector $\boldsymbol{\beta}_g$. Then we can rewrite this as:

$$\mathbb{E}[Y_i - Y_i(h_0) - \boldsymbol{\beta}_g'\mathbf{D}_i|\mathbf{D}_i, \mathbf{Z}_i = z, L_i = \ell, G_i = g] = 0$$

for any z, which implies by the law of iterated expectations over $\mathbf{D}_i$ that

$$\mathbb{E}[Y_i - Y_i(h_0) - \boldsymbol{\beta}_g'\mathbf{D}_i|\mathbf{Z}_i = z, L_i = \ell, G_i = g] \tag{C.4}$$

Consider first the case in which there is a single cohort, and we can thus ignore the conditioning on L_i. Then Assumption 3.1 implies that $\mathbb{E}[Y_i(h_0)|\mathbf{Z}_i = z, G_i = g] = \mu_{h_0g}$. and we can thus write:

$$\mathbb{E}[Y_i - \mu_{h_0g} - \boldsymbol{\beta}_g'\mathbf{D}_i|\mathbf{Z}_i = z, L_i = \ell, G_i = g] = \mathbb{E}[Y_i - \boldsymbol{\mu}_g'\mathbf{D}_i|\mathbf{Z}_i = z, L_i = \ell, G_i = g] = 0$$

This implies in particular that $\mathbb{E}[\mathbf{Z}_iY_i|G_i = g] = \mathbb{E}[\mathbf{Z}_i\mathbf{D}_i'|G_i = g]\boldsymbol{\mu}_g$. $\mathbb{E}[\mathbf{Z}_i\mathbf{D}_i'|G_i = g]$ is invertible by Assumption 3.3, yielding the result as stated in Proposition 2.

In actual estimation, we pool over cohorts with cohort fixed effects. To see that this is valid under Equations C.2 and C.3, let $\mathbf{L}_i$ be a vector of indicators for each value of L_i. Note that $\mathbb{E}[Y_i(h_0)|L_i = \ell, G_i = g]$ must be linear in $\mathbf{L}_i$ for each g. Let

$$\mathbb{E}[Y_i(h_0)|\mathbf{Z}_i = z, L_i = \ell, G_i = g] = \boldsymbol{\delta}_g'\mathbf{L_i}$$

Thus, picking up from Equation C.4:

$$\mathbb{E}[Y_i - \boldsymbol{\delta}_g'\mathbf{L_i} - \boldsymbol{\beta}_g'\mathbf{D}_i|\mathbf{Z}_i = z, \mathbf{L}_i, G_i = g] = 0$$

which gives us moment conditions to identify $\boldsymbol{\beta}_g$ and $\boldsymbol{\delta}_g$ for each g. The FJE's are now recoverable as:

$$\boldsymbol{\mu}_{hg} = \mathbb{E}[Y_i(h)|G_i = g] = \mathbb{E}[Y_i(h_0)|G_i = g] + \beta_{gh} = \mathbb{E}[\mathbf{L_i}']\boldsymbol{\delta}_g + \beta_{gh}$$

C.4.2 Proof of Lemma C.2

Recall that $Z_i = \mathbb{1}(C_i = c) = f(R_i, \tau)$ and $\hat{Z}_i^n = f(R_i, \hat{\tau})$. We show that

$$plim\left(\frac{1}{n}\sum_{i=1}^{n} V_i f(R_i, \hat{\tau}) - \frac{1}{n}\sum_{i=1}^{n} V_i Z_i\right) = 0.$$

This suffices to prove the Lemma since $plim\left(\frac{1}{n}\sum_{i=1}^{n} V_i Z_i\right) = P(C_i = c)E[V_i|C_i = c]$ by the weak law of large numbers.

For a given value r, the function $f(r, \hat{\tau})$ may be discontinuous at values of $\hat{\tau}$ such that $\hat{\tau}_h = r$ for some $h \in S$. However, the continuous mapping theorem nevertheless implies that $f(r, \hat{\tau}) \stackrel{a.s.}{\to} f(r, \tau)$ pointwise for all r such that $r \neq \tau_h$ for all h. To simplify notation, let $G = \{r \in [0,1] : r \neq \tau_h \text{ for all } h\}$. By the Severini-Egorov theorem, for any $\tilde{\epsilon}' > 0$, there exists a set $J \subset [0,1]$ of Lebesque measure smaller than $\tilde{\epsilon}'$ such that for any $\tilde{\delta} > 0$: $P\left(\sup_{r\in G\setminus J} |f(r, \hat{\tau}) - f(r, \tau)| > \tilde{\delta}\right) \stackrel{n}{\to} 0$.

Fix any $\delta > 0$ and $\epsilon > 0$, and for now, consider an aribrary set J. Expanding over the two cases:

$$\begin{aligned}
P\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_i \;\; f(R_i, \hat{\tau}) - \frac{1}{n}\sum_{i=1}^{n} V_i Z_i\right| \geq \delta\right) &= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_i(f(R_i, \hat{\tau})) - Z_i)\right| \geq \delta\right) \\
&\leq P\left(\left|\frac{1}{n}\sum_{i:R_i\in G\setminus J} V_i(f(R_i, \hat{\tau})) - Z_i) + \frac{1}{n}\sum_{i:R_i\notin G\setminus E} V_i(f(R_i, \hat{\tau})) - Z_i)\right| \geq \delta\right) \\
&\leq P\left(\left|\frac{1}{n}\sum_{i:R_i\in G\setminus J} V_i(f(R_i, \hat{\tau})) - Z_i)\right| \geq \delta/2\right) + P\left(\left|\frac{1}{n}\sum_{i:R_i\notin G\setminus J} V_i(f(R_i, \hat{\tau})) - Z_i)\right| \geq \delta/2\right)
\end{aligned}$$

Considering the first term, and applying the Markov and Cauchy-Schwarz inequalities:

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{i:R_i\in G\setminus J} V_i(f(R_i,\hat{\tau}))-Z_i)\right|\geq \delta/2\right) &\leq P\left(\frac{1}{n}\sum_{i:R_i\in G\setminus K}|V_i(f(R_i,\hat{\tau}))-f(R_i,\tau))|\geq \delta/2\right) \\ &\leq \frac{2}{\delta}E\left[\mathbb{1}(R_i\in G\setminus J)\cdot|V_i(f(R_i,\hat{\tau}))-f(R_i,\tau))|\right] \\ &\leq \frac{2}{\delta}E\left[|V_i|\cdot\mathbb{1}(R_i\in G\setminus J)\cdot|(f(R_i,\hat{\tau}))-f(R_i,\tau))|\right] \\ &\leq \frac{2}{\delta}\sqrt{E[V_i^2]}\cdot\sqrt{E\left[\mathbb{1}(R_i\in G\setminus J)\cdot|f(R_i,\hat{\tau}))-f(R_i,\tau)|^2\right]} \\ &\leq \frac{2}{\delta}\sqrt{E[V_i^2]}\cdot\sqrt{P(R_i\in G\setminus J))\cdot E\left[(f(R_i,\hat{\tau}))-f(R_i,\tau))^2|R_i\in G\setminus J\right]} \\ &\leq \frac{2}{\delta}\sqrt{E[V_i^2]}\cdot\sqrt{E\left[(f(R_i,\hat{\tau}))-f(R_i,\tau))^2|R_i\in G\setminus J\right]} \\ &\leq \frac{2}{\delta}\sqrt{E[V_i^2]}\cdot\sqrt{E\left[\{\sup_{r\in G\setminus J}|f(r,\hat{\tau})-f(r,\tau)|\}^2\right]} \end{aligned}$$

For any $\tilde{\epsilon}>0$, there exists an N_1 such that for all $n\geq N_1$, $P\left(\sup_{r\in G\setminus J}|f(r,\hat{\tau})-f(r,\tau)|>\tilde{\delta}\right)<\tilde{\epsilon}$. Given that $|f(r,\hat{\tau})-f(r,\tau)|\leq 1$ for any r and $\hat{\tau}$, it then follows that for $n\geq N_1$:

$$E\left[\{\sup_{r\in G\setminus J}|f(r,\hat{\tau})-f(r,\tau)|\}^2\right]\leq\tilde{\delta}^2(1-\tilde{\epsilon})+\tilde{\epsilon}$$

By choosing $\tilde{\epsilon}$ and $\tilde{\delta}$ such that $\tilde{\delta}^2(1-\tilde{\epsilon})+\tilde{\epsilon}\leq\frac{\epsilon^2\delta^2}{36E[V_i^2]}$, we will have

$$P\left(\left|\frac{1}{n}\sum_{i:R_i\in G\setminus J} V_i(f(R_i,\hat{\tau}))-Z_i)\right|\geq\delta/2\right)<\epsilon/3$$

In particular, we can choose $\tilde{\epsilon}=1/2\cdot\min\{\frac{\epsilon^2\delta^2}{36E[V_i^2]},1\}$ and $\tilde{\delta}=\frac{\tilde{\epsilon}}{1-\tilde{\epsilon}}$.

Now we turn to the second term.

$$
\begin{aligned}
P\left(\left|\frac{1}{n}\sum_{i:R_i\notin G\backslash J} V_i(f(R_i,\hat{\tau}))-Z_i)\right|\geq \delta/2\right) &\leq P\left(\frac{1}{n}\sum_{i:R_i\notin G\backslash J}|V_i(f(R_i,\hat{\tau}))-Z_i)|\geq \delta/2\right)\\
&\leq P\left(\frac{1}{n}\sum_i \mathbb{1}(R_i\notin G\backslash J)\cdot|V_i|\geq\delta/2\right)\\
&\leq \frac{2}{\delta}\sqrt{E[V_i^2]}\cdot\sqrt{P(R_i\notin G\backslash J)}\\
&\leq \frac{2}{\delta}\sqrt{E[V_i^2]}\cdot\sqrt{P(R_i\notin G)+P(R_i\in J)}\\
&\leq \frac{2}{\delta}\sqrt{E[V_i^2]}\cdot\left(\sqrt{P(R_i\notin G)}+\sqrt{P(R_i\in J)}\right)
\end{aligned}
$$

by similar steps as above. Firstly, we choose J such that $P(R_i\in J)=\tilde{\epsilon}'=\frac{\delta\epsilon^2}{36E[V_i^2]}$. Secondly, note that since $R_i\xrightarrow{d}U[0,1]$ and G is a finite set of points in $[0,1]$, $P\left(R_i\in G\right)\xrightarrow{n}0$: that is, given any $\tilde{\epsilon}''>0$ there exists a N_2 such that $P\left(R_i\in G\right)>1-\tilde{\epsilon}''$ for all $n\geq N_2$. In particular, choose $\tilde{\epsilon}''=\frac{\delta\epsilon^2}{36E[V_i^2]}$.

All together, for $n\geq\max\{N_1,N_2\}$ we have that

$$
P\left(\left|\frac{1}{n}\sum_{i=1}^{n}V_i(f(R_i,\hat{\tau}))-Z_i)\right|\geq\delta\right)\leq\epsilon/3+\epsilon/3+\epsilon/3=\epsilon
$$

and thus $plim\left(\frac{1}{n}\sum_{i=1}^{n}V_if(R_i,\hat{\tau})-\frac{1}{n}\sum_{i=1}^{n}V_iZ_i\right)=0$.