The Effect of Maternal Labor Supply on Children: Evidence from Bunching*

Carolina Caetano[1], Gregorio Caetano[1], Eric Nielsen[2], and Viviane Sanfelice[3]

[1]University of Georgia
[2]Federal Reserve Board
[3]Temple University

December 2022

Abstract

We study the effect of maternal labor supply in the first three years of life on early childhood cognitive skills. We pay particular attention to heterogeneous effects by the skill of the mother, by the intensity of her labor supply, and by her pre-birth wages. We correct for selection using a control function approach which uses the fact that many mothers are bunched at zero working hours – skill variation in the children of these bunched mothers is informative about the effect of unobservables on skills. We find that maternal labor supply typically has a significant, negative effect on children's early cognitive skills with more negative effects for higher-skill mothers. By contrast, we do not find significant heterogeneity depending on the pre-birth wage rate of the mother. These findings suggest that there may be more scope to avoid short-term, unintended consequences of maternal labor supply through policies that promote more flexible work arrangements rather than through policies that increase the financial rewards to working.

JEL Codes: D13, I21, I2, J01, J22, C24. Keywords: cognitive skills, bunching, maternal labor supply, early childhood, skill development

1 Introduction

This paper estimates the effect of mothers working longer hours during the first three years of a child's life on that child's cognitive skills around age 6. We use data on maternal work histories in the National Longitudinal Surveys of Youth 1979 (NLSY79) linked to childhood skill measures from the Children of the National Longitudinal Surveys (CNLSY), focusing on mothers whose children were born between 1979 and 2008. We aim to understand an important aspect of the trade-off mothers may face when deciding how much to work. On the one hand, maternal labor supply may be detrimental to children's skills because time spent at work is time not spent with children. Indeed, there is a wealth of evidence suggesting that an enriching environment with high-quality

*We would like to thank Joseph Altonji, Andrew Goodman-Bacon, Joseph Hotz, Josh Kinsler, Rodrigo Pinto, Christopher Ruhm, David Slichter, Christopher Taber, Hao Teng, and seminar participants at various institutions. The analysis and conclusions set forth here are those of the authors and do not indicate concurrence by other members of the research staff, the Board of Governors, or the Federal Reserve System.

parent/child interactions in early childhood is important for subsequent skill development (e.g., Todd and Wolpin 2007, Del Boca, Flinn, and Wiswall 2014, Hsin and Felfe 2014, Bono, Francesconi, Kelly, and Sacker 2016). On the other hand, additional work hours will bring in additional income, which may itself have a direct, positive impact on skills (Blau 1999, Milligan and Stabile 2011, Dahl and Lochner 2012, Løken, Mogstad, and Wiswall 2012).

The trade-off between time working and time at home has become increasingly salient as maternal labor supply has increased in recent decades (Eckstein and Lifshitz 2011, Fogli and Veldkamp 2011). Understanding the sign and magnitude of the total effect of maternal labor supply on childhood skill development is critical both for understanding the sources of childhood skill differences and as an input into various policy-relevant analyses. For instance, many public policies – including child allowances and tax credits, subsidized child care, or even the progressivity of the tax code – will alter mothers' labor supply choices.[1] Such policies may have unintended consequences on childhood skill development, with important implications for intergenerational mobility and inequality in general (Blau and Currie 2006, Currie and Almond 2011, Flood, McMurry, Sojourner, and Wiswall 2022).

This paper takes a distinct approach in focusing on heterogeneous effects by the skills of mothers and by the quantity of their labor supply. Both of these dimensions should alter the intensity of the trade-off between maternal labor supply and time at home. More skilled mothers tend to earn higher wages, but their time not working may also be more valuable (in terms of skill production) to their children. It is unclear whether the additional resources (financial or otherwise) earned by skilled working mothers can better offset any detrimental effect of working.[2] Moreover, on the margin, this trade-off is likely to change depending on whether the mother works longer hours (Ettinger, Riley, and Price 2018). These two dimensions may interact, as higher-skilled mothers tend to work longer hours (Cortes and Tessada 2011, Adda, Dustmann, and Stevens 2017, Chen, Grove, and Hussey 2017).

Estimating these effects is challenging because maternal labor supply may be correlated with unobservables that are themselves inputs in childhood skill production. Prior research has addressed this endogeneity using standard approaches including family fixed effects and instrumental variables (IVs). We discuss this prior related work in greater detail in Section 2. Methodologically, we add to the literature by using a novel control function approach that does not require IVs, leveraging instead the fact that maternal labor supply is bunched at zero (Caetano, Caetano, and Nielsen

[1] The literature on tax credits effects on female labor supply is large; for some references, see work by Eissa and Liebman 1996, Averett, Peters, and Waldman 1997, Meyer and Rosenbaum 2001, Grogger 2003, Bosch and Van der Klaauw 2012, Blundell, Costa Dias, Meghir, and Shaw 2016, Bick and Fuchs-Schündeln 2017. For work on child care effects on maternal labor supply at early ages (less than 3 years old) see Baker, Gruber, and Milligan 2008, Goux and Maurin 2010, Givord and Marbot 2015, Carta and Rizzica 2018, Yamaguchi, Asai, and Kambayashi 2018, Gathmann and Sass 2018, and Andresen and Havnes 2019.

[2] On the one hand, their spouse or others in her network might be more available to the child (Kalenkoski, Ribar, and Stratton, 2009; Sayer and Gornick, 2012), they may be able to afford higher-quality childcare (Blau and Hagy, 1998; Flood et al., 2022), or they may be better able to substitute market-purchased goods for their own time (Anderson and Levine, 1999). On the other hand, the time of a higher-skilled mother may be less substitutable (from the perspective of the child) with any of these options (Ruhm, 2009; Carneiro, Meghir, and Parey, 2013; Polachek, Das, and Thamma-Apiroam, 2015).

2021). We argue that mothers bunch at zero because they are at a corner solution: the constraint that hours worked cannot be negative is binding for them. Bunched mothers have different levels of the unobservable confounder, but they all choose the same amount of working hours (namely, zero). Because the treatment (working hours) is zero for the bunched mothers, skill variation among these mothers is driven by the effect of the confounder on skills. We use this skill variation for the bunched mothers to uncover the confounder's effect, which then allows us to correct for the endogeneity bias in our main estimate of interest.

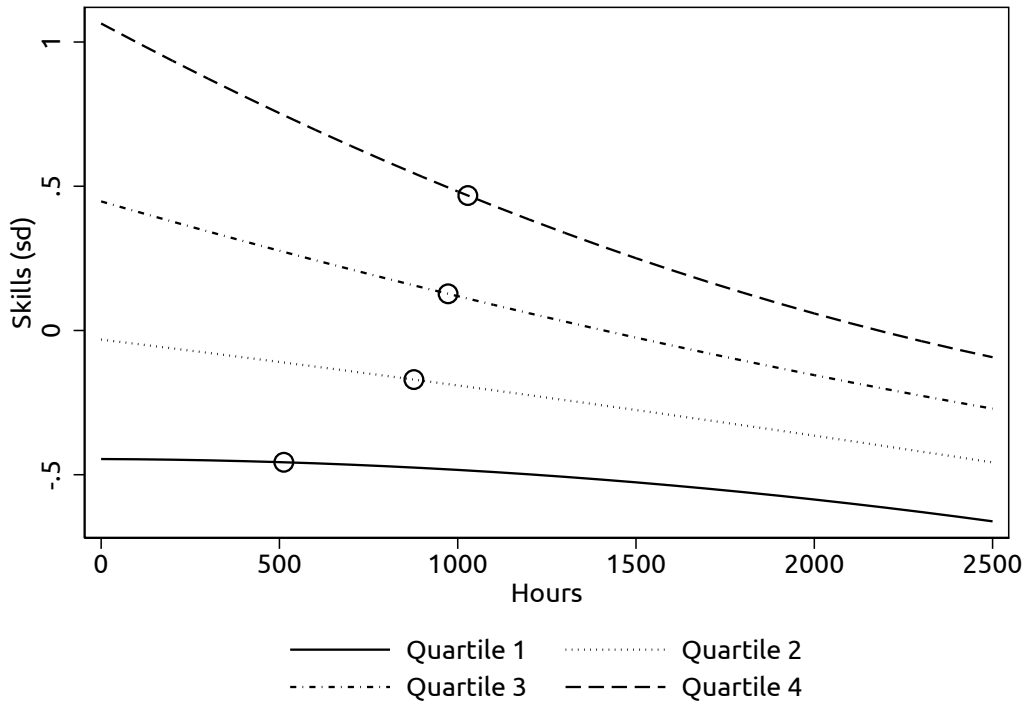
This control function approach adds value to the understanding of the effects of maternal labor supply on children because it offers a sensible identification strategy that allows us to use the full NLSY79/CNLSY sample – we do not need to restrict our analysis to families with siblings (as in fixed effects models) or to families for whom a particular instrumental variable is available.[3] This enables us to uncover effects broken down along important dimensions of heterogeneity, which enrich our understanding of the relationship between maternal labor supply and child development. It also allows us to recover average (homogeneous) treatment effects using an entirely new source of variation relative to prior literature. We provide many robustness checks that speak to the validity of the identification strategy in our empirical context.

We find that maternal labor supply has, on average, negative effects on children's cognitive skills in the short-run. Our estimates imply that an additional 10 hours of maternal labor per week during the first three years of a child's life lowers the child's cognitive skills at age six by about 10% of a standard deviation (s.d.). We also find substantial heterogeneity in these effects by the mother's skill, measured by the Armed Forces Qualifying Test (AFQT) score, and some heterogeneity by the total number of hours worked. These heterogeneous results are shown in Figure 1. The hollow circles show the average observed childhood cognitive skills and post-birth maternal work hours for each quartile of the maternal AFQT distribution. The lines show the counterfactual skills of children observed at each of these hollow points if their mother worked a different number of hours above or below the average value in the data. First, we find evidence that the schedule is nonlinear, with the degree of concavity/convexity changing depending on the skill level of the mother. However, although statistically significant, this curvature is economically not very important – the linearity assumption that is typically made in this literature seems to be a good approximation for the range of hours and AFQT scores observed in the data. Second, we find substantial heterogeneity depending on the skill of the mother. Labor supplied by higher-skilled mothers tends to have more negative effects, while for lower-skilled mothers the effects are closer to zero.

Of course, maternal labor supply has many other positive effects which may justify the implementation of work-promoting policies. Specifically, the additional income the mother earns may be beneficial to the child in the long run via several channels – better schools and social networks, support for college admissions, reduced levels of stress, etc. Furthermore, the additional income will

[3] Another potential concern with understanding heterogeneous effects using IVs is that the "compliers" of a given IV – who cannot be directly observed – may disproportionately have a certain skill level, making it difficult to compare the estimates across skill levels in a meaningful way. By contrast, we demonstrate empirically that mothers of all skill levels are well-represented in the bunched-at-zero group.

Figure 1: Children's Cognitive Skills – Quartiles of Maternal AFQT



Note: Total effects based on estimates from Table 3. The hollow circles represent the average skills and working hours of all observations in the corresponding quartile of the maternal AFQT distribution.

generally affect all family members, including the mother herself, in various positive ways. There are also important considerations with respect to career timing. Although a mother may have liked to cut back her hours during her child's early years, she might prefer to remain in the labor force full time, even if doing so is detrimental to the child in the short-run, because of potential negative long-term effects on her career. Better career growth, and the higher resources that come with it, could in turn benefit all family members, including the child, in the long run. Thus, it would be valuable to investigate further whether it is feasible to design and implement work-promoting policies that mitigate the scope for this negative unintended consequence for higher-skilled mothers.

With these considerations in mind, we investigate why working longer hours seems to have particularly detrimental short-run effects on the cognitive skills of the children of higher-skilled mothers. A natural candidate explanation is that the last hour worked may be particularly costly for the children of mothers working longer hours, as higher-skill mothers disproportionately work longer hours. However, this potential explanation is ruled out by our finding that the effect is approximately linear, as shown in Figure 1. In fact, to the extent that we find some evidence of nonlinearity, Figure 1 indicates that the last hour might be particularly costly for children of mothers working longer hours only for mothers who have sufficiently low skills.

Another potential explanation is that the additional money earned by these higher-skilled mothers may not be enough to offset their opportunity cost of working, at least in the short-run. It is therefore valuable to consider heterogeneity of the effects by another dimension beyond maternal

skills: the mother's wage rate. If the negative effects accrue mostly to the children of high-skilled, *low-wage* mothers, then policies that provide financial support to low-wage mothers could be effective by allowing their families to pay for goods and services to offset their absence. However, if the negative effects accrue also to the families of high-skilled, *high-wage* mothers, suggesting that close substitutes for high-skilled maternal time are unavailable even for this high-earning group, then providing financial support may be ineffective. In this case, it may make more sense to focus on policies aimed at promoting flexible location and work schedules for mothers, thus allowing mothers to maintain (or increase) their labor supply while not reducing their interactions with their children.

In order to investigate these issues, we study the heterogeneity of the effects of maternal work hours (beyond skill and number of hours) by the hourly pre-birth wage rate of the mother. Allowing for this third dimension of heterogeneity leads to more noisy estimates, as expected. Nonetheless, this exercise is still informative. We find some evidence that mothers with higher pre-birth wages may be able to mitigate some of the detrimental effect of their absence, although the degree of mitigation appears to be modest at best. It appears that current institutions and norms do not provide sufficient scope for families with higher-skill working mothers to mitigate any detrimental effects of the mother working on the child's cognitive skills in the short-run, even if the mother is also a high-earner.[4]

These findings yield important insights about the effects of maternal work during the first years of the children. First, using a new identification strategy leveraging different features of the data, we confirm what has typically been found in the prior literature (see Section 2): maternal labor supply on average has a negative short-run effect on children's cognitive skills. Second, we provide new evidence that this short-run unintended consequence tends to be small for low-skill mothers, even those who work long hours. Third, our analysis gives us a clue about how work-promoting policies could potentially avoid short-run unintended consequences of maternal labor supply for the children of higher-skill mothers. It may be safer (from the perspective of mitigating any such unintended consequences) to focus on policies that encourage greater flexibility in work arrangements for mothers, rather than focusing on policies that provide financial support to working mothers. Policies that increase flexibility in the schedule and location of jobs could allow mothers to spend more time with their children without sacrificing their work hours. Additionally, enhancing spouses' flexibility along the same lines might be complementary, further mitigating any potential cost to children.[5] Some of these policies, including remote and hybrid work, have recently expanded dramatically due to the necessity of social distancing during the COVID-19 pandemic, especially for higher skilled workers (Bartik, Cullen, Glaeser, Luca, and Stanton, 2020; Bick, Blandin, and Mertens, 2020; Dingel and Neiman, 2020). It would be valuable to understand the long-run impact

[4]Equivalently, the mother's (non-)working time appears to enter the production function of children from high skilled mothers in such a way that current institutions and norms cannot provide scope for a full compensation in the short-run, even if the mother is a high-earner.

[5]It is worth emphasizing that while our results pertain to the effects of maternal labor supply, similar considerations and concerns would in principle apply to the labor supply choices of any parent or caregiver. Our focus on mothers is purely pragmatic – data sources such as the NLSY79/CNLSY do not allow one to connect paternal labor supply to measures of childhood skills.

of these changes.

The rest of this paper is organized as follows. Section 2 relates this paper to previous work in the literature, while Section 3 presents our data. Section 4 discusses our empirical approach. Section 5 presents our main empirical findings, while Section 6 provides a detailed sensitivity analysis and various assessments of our key identifying assumptions. Section 7 concludes. The appendix contains additional material that helps provide further context to our study.

2 Related literature

The vast majority of studies on this topic use the NLSY79/CNLSY data and focus on estimating the impact of maternal hours worked during the three first years of a child's life on the child's skills at an early age, as we do. To overcome the endogeneity of maternal labor supply, these studies use either (i) a considerable set of control variables (Desai, Chase-Lansdale, and Michael, 1989; Baydar and Brooks-Gunn, 1991; Vandell and Ramanan, 1992; Parcel and Menaghan, 1994; Hill and O'Neill, 1994; Waldfogel, Han, and Brooks-Gunn, 2002; Baum II, 2003; Ruhm, 2004, 2009), (ii) local labor market conditions as an instrumental variable (Blau, Grossberg, et al., 1992; James-Burdumy, 2005), (iii) family fixed effects (Waldfogel et al., 2002; James-Burdumy, 2005), or (iv) dynamic choice models that simultaneously consider a mother's choice to work and invest in the child's cognitive skill (Bernal, 2008).

The results in these studies vary widely, making it difficult to draw a clear conclusion about the magnitude of the effect of maternal employment. Nonetheless, on balance, this literature finds that maternal labor supply has either a null or a detrimental effect on children's cognitive skills. Our results are consistent with these findings. For instance, Ruhm (2004), which adopts a selection-on-observables approach, finds that each additional twenty hours worked per week during the first three years of life is associated with a 0.11 standard deviation decrease on the reading assessment and a 0.08 standard deviation decrease in the mathematics assessment. Similarly, Bernal (2008) finds that working full-time and using childcare for one year is associated with a 0.13 standard deviation reduction in test scores. Other papers finding negative effects include Desai et al. (1989), Baydar and Brooks-Gunn (1991), Hill and O'Neill (1994), and Baum II (2003). Using fixed-effects models, James-Burdumy (2005) finds null effects in some cases and negative effects in others. Parcel and Menaghan (1994) similarly find null effects, while Blau et al. (1992) and Waldfogel et al. (2002) find negative effects in the first year of the child's life and offsetting, positive effects subsequently. Finally, Vandell and Ramanan (1992) reports positive effects of early maternal employment on math achievement for children from low-income families, which is consistent with our heterogeneous results.

Our analysis matches the context of this literature: we also focus on the impact of maternal labor supply in the first three years of the child on the child's early outcomes, and we also use the NLSY79/CNLSY data. Because of the similar context, we complement the main findings in this literature in many ways: (a) We confirm the main findings of negative effects with a different

approach to control for confounders; (b) We confirm that the linearity assumption made in this literature is a good approximation for the range of hours and skills in the data; (c) We provide new results about heterogeneity by skills; (d) We investigate the direct vs. income-mediated channel of the effect, providing further context to the findings of this literature while shedding light on the potential impacts of different policies.

We are not the first study to investigate the direct vs. the income-mediated channel of maternal labor supply. Two recent papers investigate such effects, but in contexts different than those in the literature discussed above. Agostinelli and Sorrenti (2021) use the NLSY79/CNLSY to estimate time and income effects of maternal labor supply when children are 4-16 years old on the children's contemporaneous outcomes, instrumenting for maternal labor supply with local labor market conditions and for family income with Earned Income Tax Credit (EITC) expansions. They find negative direct effect of maternal hours worked and positive income effect that are not fully offsetting, as we do. Using Norwegian registry data, Nicoletti, Salvanes, and Tominey (2020) estimates the direct and income-mediated effects of maternal labor supply during the first five years of the child on test scores at ages 11 and 15. To handle the endogeneity of maternal work hours and family income, the authors construct instruments for each based on the characteristics of the peers of the parental peers. They find a negative direct effect of maternal labor supply on test scores and a positive income effect that fully offsets the negative direct effect.

This paper is also related to the early literature on empirical models of female labor supply, which paid particular attention to how different economic and statistical specifications affect labor supply elasticities with respect to wage, years of experience, and partner's income (see, e.g., Heckman, 1974; Cogan, 1980; Mroz, 1987; Zabel, 1993). In addition to being concerned mainly with the estimation of different quantities (wage elasticities), these papers differ from ours in a number of other important dimensions. First, these papers face a censoring problem: wages are only observed for working mothers. By contrast, we do not face a censoring problem because we observe childhood skills and maternal work hours for everyone in our sample. Second, these papers often explicitly model both an intensive and an extensive margin of labor supply, with fixed costs being a relevant, modeled feature. As we discuss in detail in Remark 4.1, the treatment variable used in our paper and the whole related literature discussed above is aggregated at the year level. Following Cogan (1980), we argue that fixed costs of labor supply are more likely to be empirically relevant when working hours are defined at shorter time scales.

3 Data

We use data from two linked surveys: the NLSY79, which gives us information about mothers, and the CNLSY, which gives us information about their children. The NLSY79 follows a cohort of young adults aged 14-22 from 1980 through the present, while the CNLSY follows the children born to the women in the NLSY79 sample.[6] Linked together, these surveys provide a unique source of

[6]The NLSY79 interviews are annual from 1979-1994 and biennial thereafter. The CNLSY interviews are biennial starting in 1986.

information on children and their parents, including detailed information on maternal labor supply, childhood cognitive development, and household characteristics. Our final sample is a cross-sectional data set of children born from 1979 to 2008 for whom information on cognitive measures, maternal labor supply, and family characteristics are available. Children who were reported not to be living with their mother in the first years of life are dropped from our sample. We also drop observations who report working exactly 40 hours per week for all 52 weeks during each of the three years, as this lack of variation across weeks suggests that these reported hours do not reflect the actual working hours of the mother. However, replicating our analysis using these observations yields nearly the same results in all instances.

Following the economic literature in child development, we measure cognitive skills using the reading recognition and math tests from the Peabody Individual Achievement Test (PIAT). The reading recognition test is designed to measure reading comprehension based on a child's ability to recognize and pronounce words. The math test assesses attainment in mathematics beginning with early skills, such as recognizing numerals, and progressing to advanced concepts in geometry and trigonometry. The PIAT was administered to all children over the age of 5 in each CNLSY wave. Because our focus is on early childhood skill development, we adopt as our outcome a unified score for childhood cognitive skills constructed by applying factor analysis to the age-standardized math and reading PIAT scores from the first time each child in the CNLSY is assessed, which happens around age 6.[7] Throughout the analysis, we measure skills in standard-deviation (s.d.) units.

We measure our primary variable of interest, maternal labor supply, using the average number of hours worked annually by the mother in the first three years of the child's life. The NLSY79 collects extensive weekly information on employment status and hours worked. This allows us to construct a weekly work history for each mother after giving birth. Some mothers may report that they are working shortly after giving birth when they are actually on paid maternity leave (Baum II, 2003). We can only distinguish these two possibilities – working after birth versus paid maternity leave – in the survey waves from 1988 onward. To avoid losing a large portion of our sample and yet to avoid measurement error due to maternity leave, we begin to measure hours worked in the fourth month following the month of birth.[8] For instance, for a child born in July, we compute hours worked by the mother starting in the first week of November. For this child, maternal labor supply in the first year of life would be computed from the first week of November of the year of birth until last week of October in the following year. We continue this yearly computation for the next two years in order to measure hours worked by the mother in the second and third year of the child's life. Finally, our treatment variable is computed by taking the average of annual number of hours worked by the mother in these three years.[9]

[7] These age-specific scores are based on a nationally representative sample of children and are normalized to have a mean of 100 and a standard deviation of 15.

[8] The findings in this paper do not change if we start counting hours in the month immediately after the month the child is born.

[9] For some years, the NLSY79 reports weekly employment information over 53 weeks instead of 52 weeks. In order to avoid this type of measurement error, we discard information about hours worked in the 53rd week of a year, if any. In practice, this change turns out to be immaterial for the results.

A key explanatory variable in this study is the mother's cognitive skill, which we measure using the Armed Forces Qualifying Test (AFQT). The AFQT was administered to almost all NLSY79 respondents in the base year of the survey. The AFQT is a general measure of achievement in math and reading and is a primary eligibility criterion for service and placement in the United States Armed Forces. Because of its use in U.S. military personnel decisions, the AFQT has undergone extensive vetting and has been used in numerous prior economic studies as a proxy for cognitive skill or human capital (Neal and Johnson, 1996; Hirsch and Schumacher, 1998; Arcidiacono, Bayer, and Hizmo, 2010).[10]

In addition to maternal AFQT, we construct a number of other control variables based on the child, mother, and household characteristics. Unless otherwise specified, control variables such as the mother's education and marital status are computed at the year of birth. We opt for this approach in order to keep our control variables pre-determined.[11]

Table 1 presents summary statistics of our sample. The table first shows the mean and standard deviation of each element used to generate the children cognitive skill measure. These variables are normalized by age and follow a nationally representative sample with a mean of 100 and standard deviation 15. On average, children in our sample score above the national average on the PIAT reading recognition, and marginally below the average on math.

Next, the table reports statistics about maternal employment status and hours worked in the three first years of the child's life. The average annual number of hours worked in the three years following birth is 848 hours (approximately 16 hours per week) with substantial variation across children. One quarter of children in our sample have mothers who do not work during the three first years following their birth.

Turning to maternal skill, on average mothers in our sample scored 38 out of 100 in the AFQT. Since this test is set to have mean 50 and standard deviation 10 in the overall population, the mother of the average child in the sample is about one standard deviation below the national average. We also note that the AFQT scores vary notably across mothers. For our analysis, we standardize AFQT within our sample, so that it has mean zero and standard deviation one.

The remainder of the table displays summary statistics for our control variables. Most children have mothers who had completed high school and were at least 25 years old at the time of birth. Children were about 75 months old (6 years old) when they took the PIAT. The sample of children is equally balanced on gender, and is composed of 21% of Hispanic and 29% Black children. Finally, in 60% of the cases, the mother's spouse is present in the household at the time of birth, and the average child is born to a family of about three other members.

[10]The AFQT is based on a subset of tests from the Armed Services Vocational Aptitude Battery (ASVAB). Throughout, we use the current (post-1989 renormalization) definition of AFQT math as the sum of the arithmetic reasoning and mathematics knowledge subscores of the ASVAB.

[11]For children born after 1994 in odd years, the survey was not conducted in their year of birth. In these cases, we measure control variables in the year before birth, except family size which is measured at the year after birth in order for the child itself to be counted as part of the family.

Table 1: Summary Statistics

	Mean	Std.Dev.
Outcome variables		
PIAT Reading Recognition	105.33	14.04
PIAT Math	99.72	14.03
Treatment variable		
Mother's average hours worked in 3 first years	847.64	838.18
Bunching variables		
Mother worked 0 hours in 3 first years	0.25	0.44
Control variables		
Mother's AFQT score	38.20	28.21
Mother's wage year prior to the birth of the child	14.69	11.04
Mother's education less than high school	0.23	0.42
Mother's education completed high school	0.43	0.50
Mother's education some college	0.19	0.40
Mother's education completed college	0.10	0.30
Mother's education more than college	0.04	0.20
Mother's age less than 20 years old	0.11	0.32
Mother's age 20 to 24 years old	0.33	0.47
Mother's age 25 to 29 years old	0.28	0.45
Mother's age 30 to 34 years old	0.18	0.39
Mother's age 35 years old or more	0.09	0.29
Mother's spouse present	0.60	0.49
Mother's spouse highest grade	12.83	2.69
Child's age at test (in months)	75.07	14.13
Sex of child (male=1, female=0)	0.51	0.50
Birth order of child	2.06	1.18
Child is Hispanic	0.21	0.40
Child is Black	0.29	0.45
Family size	3.85	1.91
Lives in north region	0.15	0.36
Lives in north-central region	0.23	0.42
Lives in south region	0.35	0.48
Lives in west region	0.19	0.39
Observations	6924	

Note: Unless specified, control variables are measured at the child's year of birth. For children born in odd years after 1994 (years that the survey is not conducted), control variables are measured at the year before birth, except family size which is measured at the year after birth. Among the control variables we also include indicator variables for the year the child took the PIAT test. The mother's wage variable is conditional on being greater than zero and it is measured per hour in 2019 dollars.

List of Controls

Here we detail the complete list of controls used in the analysis. For the mother, we use variables meant to capture her human capital: AFQT, AFQT squared, and indicators for completed education at birth: less than high school, high school only, some college, college, and more than college. We also control for her age at birth by including indicators for whether her age was ≤ 19, $\in [20, 24]$, $\in [25, 29]$, $\in [30, 34]$, or ≥ 35. As household-level controls, we include an indicator for whether the mother's spouse is present in the household, the spouse's education at birth, and the natural log of total family size. Finally, as child-level controls, we include indicators for the child's sex, race, birth order, Census geographic region, age in months at the time of the cognitive assessment, and indicators for the year of cognitive assessment.

4 Empirical Strategy

4.1 A Selection-on-Unobservables Framework

Let L_i be our treatment variable, the average yearly working hours of mother i over the first three years of her child's life. While L_i is her chosen number of working hours, we denote by L_i^* her *desired* choice of the treatment in a decision problem that would allow her to choose negative work hours if she so desires. The details of this decision problem are irrelevant here – the only thing that matters is that the only difference between L and L^* is the constraint that the actually chosen L cannot be negative.[12] Mothers have different characteristics, some observable and some unobservable by the econometrician. These different characteristics will lead mothers to have different desired levels of hours, L^*. Thus, we can understand L_i^* as an index of the "type" of the mother, which includes both observed and unobserved confounding factors. We denote as X the vector of observed factors, and as η the index of the remaining (unobserved) factor, so that two mothers with the same level of both X and η will have the same type L^*.

Consider the comparison of the average childhood skills S among two groups of mothers with the same level of observed covariates X: those working $l_0 > 0$ hours and those working $l_1 = l_0 + 1$ hours. We can decompose this observed comparison into the component we wish to identify and a selection bias term:

$$\begin{aligned} &\underbrace{\mathbb{E}[S|L = l_1, L^* = l_1, X] - \mathbb{E}[S|L = l_0, L^* = l_0, X]}_{\text{what we observe}} = \\ &\underbrace{\mathbb{E}[S|L = l_1, L^* = l_1, X] - \mathbb{E}[S|L = l_0, L^* = l_1, X]}_{\text{treatment effect}} + \\ &\underbrace{\mathbb{E}[S|L = l_0, L^* = l_1, X] - \mathbb{E}[S|L = l_0, L^* = l_0, X]}_{\text{selection bias}}. \end{aligned} \tag{1}$$

[12] Thus, if the choice of L involves an optimization under additional constraints (e.g., budget constraints), then these constraints also affect L^*.

In words, we can observe the difference in the average S of children whose mothers choose $L = l_1$ versus those whose mothers choose $L = l_0$, but these two groups of mothers are different from each other in unobserved ways: the first group consists of the type of mothers who desire to work l_1 hours, while the second group consists of the type of mothers who desire to work l_0 hours. By adding and subtracting the term $\mathbb{E}[S|L = l_0, L^* = l_1, X]$, we obtain the right-hand-side of equation (1). The first term (labeled "treatment effect") now makes an appropriate causal comparison because it compares the average skill of children from mothers choosing l_1 versus choosing l_0 for the same type of mothers, namely those who desire l_1 hours and who have the same observables X. The term labeled "selection bias" is the bias representing the average difference in S solely due to the different type of the mothers in these two groups, i.e., the effect of the unobserved confounder η, which is the remaining source of variation in L^* beyond X.

The typical approach to identify causal effects is to use a source of variation that shuts off the selection bias term, leaving only the treatment effect term. In this paper, we go the other way: we use a source of variation that shuts off the treatment effect term, allowing us to identify the selection bias term. This will enable us to identify the treatment effect term indirectly because the left-hand-side of equation (1) is observed.

Intuitively, we make use of the discontinuous variation in the outcome S as we approach $L = 0$ from above, which we know to be due solely to η (and thus only to L^*). This allows us to isolate variation in L^*, giving us the effect of mothers being of different types on childhood skills, S – the term labeled "selection bias" in equation (1). To see how this can be achieved in our setting, consider the comparison of the average childhood skills S between mothers at $L = 0$ and mothers at $L = l$ for a marginally positive value of $l > 0$. The difference in average S between these two groups cannot be attributed to the term "treatment effect" because both groups of mothers choose almost the same quantity of hours. However, the average type of these mothers may differ substantially: while the mothers at $L = l > 0$ are all of type $L^* = l$, the mothers at $L = 0$ may not all be of type $L^* = 0$. Indeed, while some of the mothers choosing $L = 0$ may be exactly indifferent between working and not working (i.e., they are of type $L^* = 0$), it is likely that other mothers at $L = 0$ are farther from indifference (i.e., they have types strictly less than zero). The average type among mothers at $L = 0$ will therefore be strictly negative in this case. Moreover, there would also be a discontinuity in the average type at $L = 0$ as this average will be strictly positive for small values of $L > 0$ but strictly negative at $L = 0$. Any discontinuity in the average of S at $L = 0$ can thus be attributed to this discontinuous difference in average types. We next show evidence that at least some mothers at $L = 0$ must have types $L^* < 0$, and thus that the average type does vary discontinuously at $L = 0$.

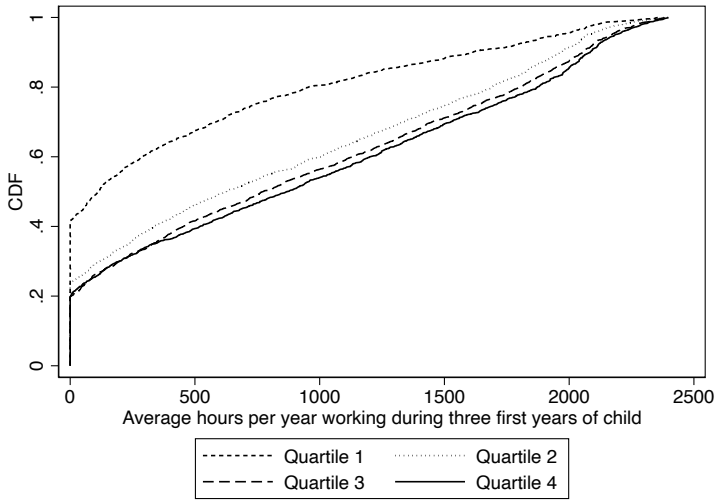
4.2 Evidence of Bunching and Selection

Evidence of Bunching

We begin by showing in Figure 2 that our treatment variable L has a notable bunching point at $L = 0$ for mothers of all skill levels. Specifically, the curves show the cumulative distribution of

mothers' average working hours, L, for each quartile of the maternal AFQT distribution.

Figure 2: Evidence of Bunching by Quartile of the Maternal AFQT Distribution



Note: This figure shows the cumulative density function (CDF) of the average yearly hours mothers have worked in the three years following the birth of their child for each quartile of the maternal AFQT distribution.

About 40% of the mothers in the lowest quartile choose to work exactly zero hours, while only a small proportion of them choose to work a few hours per year. The degree of bunching at $L = 0$ tends to vary with the AFQT score of the mother, as expected. Nonetheless, we find a substantial amount of bunching at zero hours for all quartiles. This is evidence that many mothers, irrespective of their skill level, find themselves at a corner solution when deciding how many hours to work: some mothers are not indifferent between working and not working – their type is $L^* < 0$. We next show more direct evidence that many of the bunched mothers are not close to indifference.

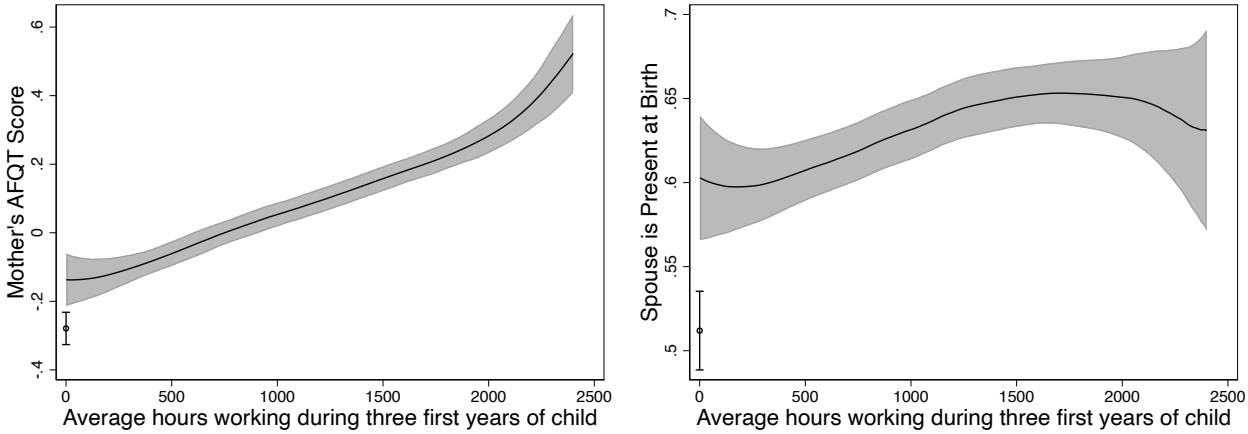
Evidence of Selection

It is of course possible that the bunching shown in Figure 2 simply reflects a mass of mothers whose type makes them exactly indifferent between working and not working. However, we have strong reasons to believe that this is not the case: while some mothers at $L = 0$ might be of type $L^* = 0$, others are likely to be of type $L^* < 0$. If it were the case that all mothers at $L = 0$ were exactly indifferent between working and not working, $L^* = 0$, then they should be comparable to those mothers that are working just a tiny amount of hours, $L^* = l$ for marginally positive $l > 0$, since these two groups of mothers would be almost of the same type. In that scenario, we would also expect the observable characteristics of mothers at $L = 0$ to be similar on average to the observable characteristics of mothers at $L = l$ for marginally positive $l > 0$.

This is not what we find in Figure 3. The figure shows local linear regression fits for key covariates on L, estimated on the $L > 0$ sample. We also show the average values of the covariates for the subsample of mothers bunched at $L = 0$. The left panel shows that the fit for maternal AFQT is very smooth for positive values of L, suggesting that mothers who work similar positive hours tend to have similar AFQT scores. However, mothers who work exactly zero hours tend to have sharply

lower AFQT scores than those who work a small, positive number of hours. This discontinuity suggests that among the mothers at $L = 0$ are some with $L^* < 0$, as those with $L^* = 0$ should have on average similar AFQT scores to mothers choosing $L = l$ for small $l > 0$. The right panel shows a similar pattern for the proportion of mothers at each level of $L = l$ whose spouse was present in the household in the year the child was born: mothers at $L = 0$ are discontinuously less likely to have their spouse present at birth than mothers with marginally positive hours $L = l > 0$.

Figure 3: Mothers at $L = 0$ are discontinuously different from working mothers



Note: This figure shows the local linear regression of two key observed covariates on L (average hours working per year during the first 3 years of the child) along with the 95% confidence interval. The bandwidth is 400 hours. At $L = 0$, the average along with the 95% confidence interval is also shown.

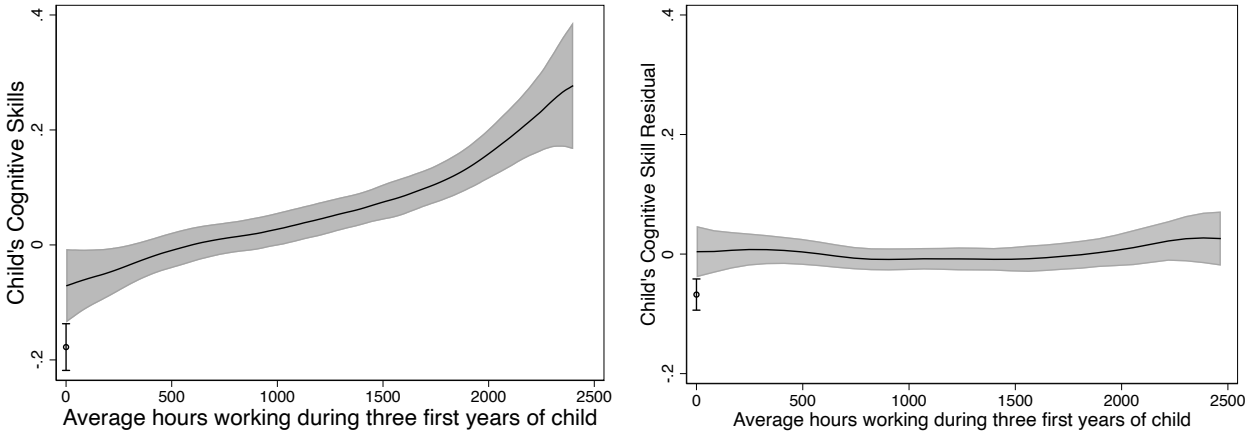
We find similar discontinuities for many observed covariates. Moreover, the sign of these discontinuities all tend to suggest positive selection of maternal working hours near $L = 0$. That is, mothers who work a small, positive number of hours per year tend to have discontinuously higher levels of covariates that are themselves positively correlated with children's cognitive skills, relative to those who work zero hours.

Direct evidence of this positive selection is shown in Figure 4. The left panel is analogous to those in Figure 3, but for the outcome variable (childhood cognitive skills S) in the vertical axis instead of a key covariate. A significant positive discontinuity is evident, confirming the intuition that, on balance, the covariates (observed or unobserved) that are discontinuously higher as L increases from $L = 0$ tend to be positively correlated to S. The discontinuity in the left panel could come from two possible sources. First, as Figure 3 demonstrates, the mothers at $L = 0$ and the mothers at marginally positive $l > 0$ differ on many observable dimensions relevant for childhood skills. Second, the two groups of mothers likely differ on their unobservable type L^*, with the average L^* for the bunching mothers likely being strictly less than zero.

It is in principle possible that observed covariates X are enough to control for the positive selection evident in the left panel of Figure 4. However, this hypothesis turns out to not be true: When we test the selection-on-observables assumption with Caetano (2015)'s discontinuity test based on the bunching at $L = 0$, we strongly reject it. This can be seen in the right panel of the

figure, where we plot the residual of a regression of S on our full list of controls X.[13] Because controls are held constant, the significant discontinuity at $L = 0$ now reflects only the discontinuity in the component of unobservables that was not absorbed by covariates, i.e., the confounding component beyond any selection on observables, which we denote as η. This positive discontinuity at $L = 0$ suggests that an OLS estimator of the coefficient of L on a regression of S on L and controls X would be positively biased, thus motivating our selection-on-unobservables approach described in Section 4.3.

Figure 4: Evidence of Positive Selection



Note: This figure is analogous to Figure 3, but instead of covariates the variables in the vertical axis represent the outcome variable S (children's cognitive skills) in the left panel, and the "residualized" outcome S in the right panel. See Footnote 13 for details.

Remark 4.1. *Implicit in the discussion in Section 4.2 are two assumptions similar to the ones made in regression discontinuity designs (RDDs): (a) L is a continuously distributed variable for $L > 0$, so we can arrive arbitrarily close to $L = 0$ from positive values of L; and (b) treatment effects are continuous at $L = 0$. These two assumptions are plausible in our context. For (a), note from Figure 2 that the slope of the CDF (which is the probability density function, PDF) is positive for small positive values of L around zero; indeed we observe several mothers working just a few hours per year on average during the first three years of the child's life, which can be directly seen in the histogram in Figure 13 in Appendix C.*[14] *For (b), it is plausible that mothers working on average just a few hours per year during the first three years of the child's life should not generate a sharp*

[13] The covariates X enter nonparametrically in this regression, as described in Section 4.4. See Caetano (2015) for more details on how this test is implemented.

[14] One might expect that in practice there are fixed costs to supplying labor, which would imply there will be no mothers choosing $L > 0$ for sufficiently small number of hours. Although fixed costs may be relevant for the decision of working on any given day, they seem not to be as relevant for the treatment variable used in this paper (and in much of the related literature): the average number of yearly hours worked during the first three years of the child's life. In our data, we observe that the mothers who work on average just a few hours per year tend to concentrate their hours in the same week, which is consistent with the fixed cost hypothesis. We are not the first to note that a variable cost model fits the data better than a fixed cost model when hours are at the yearly scale, even if the true model at a high-frequency timescale includes substantial fixed costs. For instance, Cogan (1980) states, "It may be argued that the major sources of the costs of work are more properly treated as variable costs with respect to annual hours. This is true, especially if most of the variation in annual hours worked was the result of variations in days worked per year."

causal effect on the cognitive skills of the child at age 6, relative to those working zero hours per year. Thus, discontinuities in $\mathbb{E}[S|L = l]$ as l approaches 0 from the positive side cannot generate any discontinuity because of its causal effect, so it must be because selection varies discontinuously. As this discussion suggests, there is a close parallel between our approach and RDDs, which is discussed and clarified further in Appendix B.

4.3 Control Function Approach

In this section, we first formalize the ideas discussed above. We then show how to leverage bunching with a control function approach that corrects for selection on unobservables.

A Model of Constrained Labor Supply

We start by specifying the child's cognitive skill S as

$$S = f(L, X; \boldsymbol{\beta}) + g(X) + \epsilon, \tag{2}$$

where g is nonparametric, L is the average number of maternal work hours in the first three years of the child's life, X is a vector of pre-determined controls, and ϵ is the unobservable error term.

We will specify the parametric function $f(\cdot; \boldsymbol{\beta})$ in alternative ways in order to improve our understanding of heterogeneity in the effects of L on S. In its simplest form, $f(L, X; \boldsymbol{\beta}) = \beta L$ when we aim to identify the average treatment effect across all mothers. Regardless of the specification of the parametric function $f(\cdot; \boldsymbol{\beta})$, our goal is to identify the vector of parameters $\boldsymbol{\beta}$. The challenge is that L is endogenous – L and ϵ are correlated to each other conditional on X. Thus, a (nonparametric) regression of S on L and X will yield a biased estimate of $\boldsymbol{\beta}$, as we have shown in Figure 4.

Mothers face a constrained optimization problem when deciding how many hours to work because the actual number of work hours chosen is constrained to be non-negative. Specifically, we write the *desired* number of work hours, L^*, as

$$L^* = h(X) + \eta, \tag{3}$$

where h is nonparametric, and

$$L = \max\{0, L^*\}, \quad \text{with } \mathbb{P}(L^* < 0) > 0. \tag{4}$$

Equation (3) is written without loss of generality, as η is simply defined as the remainder of L^* after $h(X)$ is controlled for. Equation (4) is motivated by the findings discussed in Section 4.2. Intuitively, some mothers are at a corner solution: they are not exactly indifferent between working and not working.

Next, we add some structure. We open the error term in equation (2) so that $\epsilon = \delta(X)\eta + \varepsilon$, and assume that L and ε are uncorrelated to each other conditional on covariates X and the unobservable η:

$$S = f(L, X; \boldsymbol{\beta}) + g(X) + \delta(X)\eta + \varepsilon, \quad \text{with } \mathbb{E}[\varepsilon|L, X, \eta] = 0, \tag{5}$$

This is a selection-on-unobservables assumption, since we allow for L and η to be correlated to each other, but, conditional on η and on covariates X, the remaining error is uncorrelated to L. For concreteness, throughout the paper, we call this a linearity assumption, as it effectively assumes that the unobserved confounder term can be written as a linear function of η, $\delta(X)\eta$. The slope may change nonparametrically depending on the value of the covariates X.

We exploit bunching to identify $\boldsymbol{\beta}$ under the model specified by equations (3), (4) and (5). To see how bunching allows us to do this, consider the children of mothers bunched at $L = 0$. All such children have the same treatment $L = 0$, and remaining differences in their observables X can be controlled for. Thus, any systematic differences in S between children with the same value of $L = 0$ and the same observables X must occur because of differences in η, per equation (5). This allows us to isolate the effect of η on S, which we can use to build a control function approach to identify $\boldsymbol{\beta}$. Next, we detail how we implement this idea.

Building the Control Function

The model defined by equations (3), (4) and (5) implies[15]

$$\mathbb{E}[S|L, X] = f(L, X; \boldsymbol{\beta}) + \underbrace{g(X) - \delta(X)h(X)}_{m(X)} + \delta(X)[L + \mathbb{E}[L^*|L = 0, X]\mathbf{1}(L = 0)]. \tag{6}$$

If we could identify $\mathbb{E}[L^*|L = 0, X]$, then we could add the term $L + \mathbb{E}[L^*|L = 0, X]\mathbf{1}(L = 0)$ to the equation as another control, allowing us to identify $f(\cdot)$, $m(\cdot)$ and $\delta(\cdot)$. Thus, the identification of $\boldsymbol{\beta}$ can be achieved by first identifying $\mathbb{E}[L^*|L = 0, X]$. In words, $\mathbb{E}[L^*|L = 0, X]$ refers to the average type among mothers at $L = 0$ with observed covariates X. Because of the evidence shown in Section 4.2, we know that at least some of the mothers at $L = 0$ are of type $L^* < 0$. However, we do not know exactly how close to indifference (i.e., $L^* = 0$) is the average mother at $L = 0$, and it is precisely this average that is needed for identification. Fortunately, it turns out that our main findings about $\boldsymbol{\beta}$ do not change for all plausible values of $\mathbb{E}[L^*|L = 0, X]$, as we will show in detail in Section 6.1. Nonetheless, in order to achieve point identification of $\boldsymbol{\beta}$, we will make an additional assumption, beyond linearity, on the distribution of $\eta|X$.

Assumption 1. *(Nonparametric Tail Symmetry) For all censored quantiles q_0,*

$$\eta|X \text{ has symmetric tails below } q_0 \text{ and above } 1 - q_0,$$

In order to provide more context about the robustness of our results to distributional assumptions, we also show results under stronger but more standard assumptions:

Assumption 1′. *(Semiparametric Normal)*

$$\eta|X \sim \mathcal{N}(l(X), \sigma^2(X)),$$

[15] This equation is obtained by first substituting η from equation (3) into equation (5), and then noting that equation (4) implies $L^* = L + L^* \cdot \mathbf{1}(L = 0)$.

Assumption 1″. *(Semiparametric Uniform)*

$$\eta|X \sim U[\kappa(X), \mu(X)].$$

While the normal distribution is a standard choice, we also consider the uniform distribution because it seems to fit the data well for most values of X. We also consider nonparametric full symmetry, which, while stronger than tail symmetry, is testable.[16] We are almost never able to reject the full symmetry assumption in our data.[17] We thus take as our headline estimates those estimated under nonparametric tail symmetry (Assumption 1) because it is the weakest assumption we make that achieves point identification. Additionally, throughout the paper, we only focus on findings that hold irrespective of the distributional assumption.

Summarizing, our control function approach achieves point identification under two assumptions beyond the standard rank condition:[18] (a) the linearity assumption in equation (5), and (b) the nonparametric tail symmetry assumption discussed above. Importantly, the approach does not rely on Instrumental Variables (IVs), which are difficult to find in this context. Intuitively, we circumvent the need for IVs because of the constraint in the choice of L. Due to the binding restriction for some mothers, $\mathbb{E}[L^*|L = 0, X]$ is negative, which makes the last term of equation (6) discontinuous at $L = 0$. This allows us to identify $f(\cdot)$ and $\delta(\cdot)$ separately using the discontinuity in $\mathbb{E}[S|L, X]$ at $L = 0$. Controls X play no special role in this identification strategy, beyond controlling for further endogenous variation, thus weakening the assumptions on the unobservable η. Remarks 4.2 and 4.3 below discuss the role of controls in more detail.

4.4 Estimation Details

Section 3 details the list of controls X, some elements of which are continuous. As discussed in Caetano et al. (2021), in order to maintain the predominantly nonparametric structure of the model (rather than resorting to linear structures due to the high dimensionality of X) it is a good idea to "discretize" X before estimating the expectation $\mathbb{E}[L^*|L = 0, X]$. Let $\{\hat{\mathcal{C}}_1, \ldots, \hat{\mathcal{C}}_K\}$ be a finite partition of the support of X into sets, which we call clusters, and let $\hat{C}_K = (\mathbf{1}(X \in \hat{\mathcal{C}}_1), \ldots, \mathbf{1}(X \in \hat{\mathcal{C}}_K))'$ be the vector of cluster indicators. In the estimation of the expectation, we substitute X with $\hat{C}_K$, which has finite support. The estimator $\hat{\mathbb{E}}[L^*|L = 0, X] = \hat{\mathbb{E}}[L^*|L = 0, \hat{C}_K]$ is thus constructed using a two-step procedure in which first X is discretized and then one of the

[16] To see how this assumption is testable, consider an example where for a given value of X there is 30% of bunching at $L = 0$. Then we can compare the empirical distribution of L between percentiles 30 and 50 with the symmetric fit, which is the mirror image of the empirical distribution between percentiles 50 and 70 of the data.

[17] Specifically, Kolmogorov-Smirnov tests reject full symmetry for only 1 out of 50 clusters at a familywise α of 10%, while 2 additional clusters are not testable because they have more than 50% censoring. Similarly, we rarely or never reject full symmetry using smaller numbers of clusters where high censoring rates are less of an issue. Finally, we also rarely or never reject t-tests comparing the means of the observed, below-median distributions and the mirror images of the above-median distributions.

[18] The rank condition that allows for the linear independence between the first and last terms of equation (6) follows trivially from bunching, $\mathbb{P}(L^* < 0) > 0$, and from $f(\cdot)$ being continuous at $L = 0$ (see Remark 4.1).

distributional assumptions is applied separately for each cluster.[19]

The clusters are not arbitrary – rather, we employ a clustering algorithm in which two observations will be clustered together if they have similar Xs. As K grows the observations within the same cluster have increasingly more similar values of X.[20] Thus, if $\mathbb{E}[L^*|L=0,X]$ is continuous, then $\hat{\mathbb{E}}[L^*|L=0,\hat{C}_k]$ will approximate $\mathbb{E}[L^*|L=0,X]$ more and more closely as K grows.

Following a similar logic, we also use the same clusters to make sure the function $m(X)$ in equation (6) approximates a nonparametric function of X. Specifically, we specify $m(X) = X'\tau + \sum_{k=1}^{K} \alpha_k \mathbf{1}(X \in \hat{\mathcal{C}}_k)$, so the cluster indicators control nonparametrically for differences across clusters, while the within-cluster differences due to X are controlled linearly. As the number of clusters increases, the nonparametric match improves, leaving less unexplained variation within cluster. We show that the main estimates do not change appreciably when the total number of clusters grows, thus suggesting that our approach of conditioning on these indicators well approximates conditioning on a nonparametric function of X. As a robustness check in Section 6.2.3, we also consider a similar specification of $\delta(X)$ with K_δ clusters.

Summarizing, our estimation approach consists of the following steps. First, we create K clusters using our vector of pre-determined controls X. Next, separately for each cluster, we estimate $\mathbb{E}[L^*|L=0,X\in\hat{\mathcal{C}}_k]$ using the methods outlined above, and we construct the variable $(L+\hat{\mathbb{E}}[L^*|L=0,\hat{C}_K]\mathbf{1}(L=0))$. Finally, we estimate our model of interest including this variable as an additional control. For example, in our homogeneous model in which both $f(\cdot)$ and $\delta(\cdot)$ are constant for different values of X, we estimate via OLS the following regression:

$$S = \beta L + X'\tau + \sum_{k=1}^{K} \alpha_k \mathbf{1}(X \in \hat{\mathcal{C}}_k) + \delta[L + \hat{\mathbb{E}}[L^*|L=0,\hat{C}_K]\mathbf{1}(L=0)].$$

Remark 4.2. ***Controls and the linearity assumption.*** *As discussed in Caetano et al. 2021,*

[19] The estimation of $\mathbb{E}[L^*|L=0,\hat{C}_K]$ under each distributional assumption is straightforward. In each case, one conducts the estimation separately by cluster. Thus, for a given cluster k, under Assumption 1 (Nonparametric Tail Symmetry),

$$\hat{\mathbb{E}}[L^*|L=0,X\in\hat{\mathcal{C}}_k] = \hat{F}^{-1}_{L|\hat{\mathcal{C}}_k}(1-\hat{F}_{L|\hat{\mathcal{C}}_k}(0)) - \hat{\mathbb{E}}[L|L\geq \hat{F}^{-1}_{L|\hat{\mathcal{C}}_k}(1-\hat{F}_{L|\hat{\mathcal{C}}_k}(0)), X\in\hat{\mathcal{C}}_k],$$

where $\hat{F}_{L|\hat{\mathcal{C}}_k}(0)$ is just the fraction of bunched (at zero) observations in cluster k, $\hat{F}^{-1}_{L|\hat{\mathcal{C}}_k}(1-\hat{F}_{L|\hat{\mathcal{C}}_k}(0))$ is estimated by substituting the empirical quantile of the distribution of L among observations in cluster k, and the final expectation term is estimated simply as the sample average in cluster k of those observations with L greater than this empirical quantile. This is implemented for clusters k such that $\hat{F}_{L|\hat{\mathcal{C}}_k}(0)\leq 0.5$, and Assumption 1′ is made for other clusters, if any. Under Assumption 1′ (Semiparametric Normality), one simply runs a Tobit regression of L on a constant using only the observations in cluster k. Letting $\hat{\alpha}_k$ and $\hat{\sigma}_k$ be the estimated coefficient (intercept) and standard deviation from this Tobit model, $\hat{\mathbb{E}}[L^*|L=0,X\in\hat{\mathcal{C}}_k] = -\hat{\alpha}_k - \hat{\sigma}_k\lambda\left(-\hat{\alpha}_k/\hat{\sigma}_k\right)$, where λ is the inverse Mill's ratio. Finally, under Assumption 1″ (Semiparametric Uniform),

$$\hat{\mathbb{E}}[L^*|L=0,X\in\hat{\mathcal{C}}_k] = -\hat{\mathbb{E}}[L|L>0,X\in\hat{\mathcal{C}}_k]\left(\hat{F}_{L|\hat{\mathcal{C}}_k}(0)/(1-\hat{F}_{L|\hat{\mathcal{C}}_k}(0))\right),$$

where $\hat{\mathbb{E}}[L|L>0,X\in\hat{\mathcal{C}}_k]$ is just the average of L among the non-bunched mothers and $\hat{F}_{L|\hat{\mathcal{C}}_k}(0)$ is just the fraction of bunched mothers in cluster k. These cases and others are covered in detail in Caetano et al. (2021).

[20] We show results using hierarchical clustering (with the Gower measure of distance and Ward's linkage) for its simplicity, stability, and ease of interpretation as we vary the number of clusters (Hastie, Tibshirani, and Friedman 2009).

observed controls are not necessary for this identification strategy. However, as we add controls, we tend to weaken the linearity assumption for two different reasons. First, only η needs to have a linear effect on the outcome S. So we still allow for L^* to have a non-linear effect as long as it is through controls X (see equation (3)). Indeed, to see how this happens in practice, compare the extent to which the left panel of Figure 4 (which shows how S changes with L unconditionally) is more non-linear relative to the right panel of that figure, which shows how the residual of S changes with L after partialling out $m(X)$ using $K = 50$ clusters.[21] *Second, controls allows us to weaken the linearity assumption because we also may allow for δ to change with X, as we discuss in detail in Section 6.2.*

Moreover, controls also allow us to perform further tests of the linearity assumption, as discussed in Section 6.2.2.

Remark 4.3. ***Controls and the distributional assumption.*** *Adding controls diminishes the sources of variation of L^* that will be attributed to η, as is clear from equation (3). This has important implications for the credibility of our distributional assumptions on η. For instance, consider Assumption 1, which states that η is symmetric on the tails with a distribution that depends on X. With no controls, this assumption would restrict η from all observations to be drawn from the same distribution. As we add controls, this assumption changes, requiring that η be drawn from the same distribution only for observations within the same cluster.*

In Section 6.3, we change the number of clusters from $K = 1$ all the way to $K = 100$ when estimating the by-cluster expectations. When $K = 1$, X contains only a constant, so η is simply the demeaned version of L^*. Next, when $K = 2$, we allow for L^* to be distributed differently for observations from different clusters. Further, the more clusters we add, the more we allow for L^* from different observations to be distributed differently from each other, thus effectively changing our distributional assumption. Thus, checking how our results change when we change the number of clusters of X from $K = 1$ to $K = 100$ is informative about whether our key findings are an artifact of the distributional assumption.

5 Results

In this section, we present our estimates of $\boldsymbol{\beta}$ for different specifications of the function $f(L, X; \boldsymbol{\beta})$. In order to facilitate the interpretation of our results, we also specify $\delta(X)$ to be simple functions of the controls X, as these functions make clear how selection varies with some key observables. We show in Section 6.2.3 that our estimates of $\boldsymbol{\beta}$ are robust to specifications that allow $\delta(X)$ to depend nonparametrically on X.

[21] Moreover, Section 6.3 shows further what happens with our main estimates when the total number of clusters used to estimate $m(X)$ changes from $K = 1$ to $K = 100$. As K grows, $m(X)$ is specified more flexibly, which may partially absorb any remaining non-linearity. It is therefore informative to check if the main results change as K grows.

5.1 Homogeneous, Linear Results

We start from the most parsimonious specification of $f(L, X; \boldsymbol{\beta})$,

$$f(L, X; \boldsymbol{\beta}) = \beta L. \tag{7}$$

This specification of $f(L, X; \boldsymbol{\beta})$ is comparable to the specifications in the prior literature.

Table 2 presents these homogeneous, linear estimates. Column (i) presents the results of simple regressions of skills on maternal working hours with no additional controls. Cognitive skills are strongly positively associated with maternal work hours. However, column (ii), which adds observable controls, $m(X)$, to the specification in column (i), shows that pre-determined observables remove most of this positive relationship – the residual regression coefficient is close to zero and insignificant.

Table 2: The Effect of Maternal Hours Worked on Early Childhood Cognitive Skills

	(i) Uncorrected No Controls	(ii) Uncorrected w/ Controls	(iii) Het. Tobit	(iv) Het. Uniform	(v) Het. Symmetric
β	0.014**	0.000	-0.016**	-0.019**	-0.019**
	(0.001)	(0.001)	(0.005)	(0.006)	(0.005)
δ			0.014**	0.017**	0.017**
			(0.004)	(0.005)	(0.005)

Note: This table shows estimates of the effect of an additional 100 hours per year working in the three years following the child's birth on the child's early cognitive skills. $N = 6,924$. Bootstrapped standard errors in parentheses (1,000 bootstrap samples). This is the list of controls X: mother's AFQT and AFQT squared, mother's completed education at birth (<HS, HS, some college, college, >college), mother's age (≤ 19, $\in [20, 24]$, $\in [25, 29]$, $\in [30, 34]$, or ≥ 35), indicators for whether the mother's spouse is present in the household and the spouse's education at birth, ln(total family size), child's sex, race, birth order, Census geographic region, age at assessment date, and indicators for the year of the assessment. We specify $m(X)$ to enter both linearly and as indicators of each cluster of X, with a total of $K = 50$ clusters. $\mathbb{E}[L^*|L = 0, X]$ is also estimated with $K = 50$ clusters, and $\delta(X)$ is estimated with $K_\delta = 1$ cluster. See Figure 11 and Table 5 for the analogous results for different values of K and K_δ, respectively. ** p<0.05, * p<0.1.

Columns (iii)-(v) present our estimates correcting for endogeneity using the control function approach outlined in Section 4. Each column differs only in the assumption made on the distribution of $\eta|X$. Column (iii) supposes that $\eta|X$ is normally distributed, with mean and variance that depend on the cluster to which X belongs. Column (iv) supposes that $\eta|X$ is uniformly distributed with the upper and lower limits of the support depending on the cluster of X. Finally, column (v) supposes that $\eta|X$ is symmetric in the tails, so that the unobserved portion of $\eta|X$ below 0 is the mirror image of the corresponding tail above 0, conditional on the cluster of X. As discussed in Section 4, these are our preferred point estimates, as they rely on our weakest distributional assumption.

Table 2 reveals that maternal labor supply has quantitatively large and statistically significant negative effects on cognitive skills. Column (v) suggests that an additional 100 hours of maternal labor supply annually over the first three years of a child's life lowers the child's cognitive skills around age 6 by 0.019 standard deviations (s.d) on average. This effect is economically large given

the sample variance in maternal labor supply – a one s.d. increase in maternal labor supply over the first three years of life (an increase in 838 hours per year, see Table 1) would translate to reductions in cognitive skills of about 0.16 s.d. The estimates assuming normal or uniform distributions for η are quite similar to the symmetric case.

Table 2 also shows the estimates of δ, the average effect of the confounder η on the outcome Y. These estimates are positive and significant, consistent with what we find in Section 4: mothers who work more hours tend to be positively selected relative to those who work fewer hours.

Selection on Observables vs. Selection on Unobservables

To provide more context to our findings, it is useful to note what happens to the estimates of β in Table 2 as the controls are added. The estimate of β, which was large and positive without controls (column (i)), goes down to zero when controls are added (column (ii)). Thus, our negative results in columns (iii)-(v) simply show that when we further control for unobservables, the estimate of β goes down further, indicating that selection on unobservables tends to be in the same direction as selection on observables. In Appendix A, we formalize this idea by implementing the method developed in Oster (2019) to show that our main estimate from column (v) is consistent with the degree of selection on unobservables being smaller (but in the same direction) than the degree of selection on observables, which is generally considered plausible (Altonji, Elder, and Taber 2005). Importantly, Oster (2019)'s method relies on completely different assumptions and does not use bunching of the treatment variable in any way. We thus view these results as independent evidence that our findings are plausible and not merely an artifact of our method.

5.2 Heterogeneity by Maternal AFQT and Labor Supply

We now turn to assessing heterogeneity in the effect of maternal labor supply by the skill (AFQT score) of the mother and the intensity of her labor supply. We specify $f(L, X; \boldsymbol{\beta})$ as a quadratic function of hours, with the slope and convexity allowed to be heterogeneous by the normalized AFQT score of the mother, A:

$$f(L, X; \boldsymbol{\beta}) = (\beta + \beta_A A + \beta_L L + \beta_{AL} AL)L. \tag{8}$$

Table 3 presents the heterogeneous results. As before, we discuss the results of our preferred point estimates under nonparametric tail symmetry. The other distributional assumptions yield qualitatively and quantitatively similar estimates.

Table 3 reveals a number of interesting patterns. First, $\beta < 0$ – the estimated effect on cognitive skills of the first 100 working hours for a mother with average skills ($A = 0$) is negative. Second, $\beta_A < 0$ – the effect of work hours on cognitive skills is more negative the higher is the skill of the mother. This effect is statistically significant and quite large; each additional standard deviation of mother's AFQT lowers β by 0.026 s.d. Third, $\beta_L > 0$ – the effect of work hours for a mother with average AFQT skills become larger (or less negative) the more total hours are worked, although the

Table 3: The Effect of Maternal Hours Worked on Early Childhood Cognitive Skills by the AFQT Score and Labor Supply of the Mother

	(i) Uncorrected No Controls	(ii) Uncorrected w/ Controls	(iii) Het. Tobit	(iv) Het. Uniform	(v) Het. Symmetric
β	0.018**	0.003	-0.023**	-0.026**	-0.030**
	(0.004)	(0.003)	(0.008)	(0.010)	(0.009)
β_A	0.047**	-0.010**	-0.017**	-0.021**	-0.026**
	(0.003)	(0.004)	(0.008)	(0.009)	(0.009)
β_L ($\times$1000)	-0.004**	-0.001	0.002	0.001	0.002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
β_{AL} ($\times$1000)	-0.015**	0.003**	0.004**	0.004**	0.004**
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
δ			0.016**	0.022**	0.023**
			(0.005)	(0.007)	(0.006)
δ_A			0.005	0.009	0.013**
			(0.005)	(0.006)	(0.006)

Note: This table shows estimates of the effect of an additional 100 hours per year working in the three years following the child's birth on the child's early cognitive skills. $N = 6,924$. Bootstrapped standard errors in parentheses (1,000 bootstrap samples). We specify $m(X)$ to enter both linearly and as indicators of each cluster of X, with a total of $K = 50$ clusters. $\mathbb{E}[L^*|L = 0, X]$ is also estimated with $K = 50$ clusters, and $\delta(X)$ is estimated with $K_\delta = 1$ cluster. See Figure 14 and Table 7 for the analogous results for different values of K and K_δ, respectively. ** p<0.05, * p<0.1.

degree of convexity is small and not significant.[22] Fourth, $\beta_{AL} > 0$ – the effects of working hours are more convex for higher skill mothers. These results are illustrated in Figure 1 in the Introduction for four hypothetical children who are representative of each quartile of the distribution of skill of their mothers. Each child's skill and corresponding maternal working hours is represented by a hollow circle in the figure, which together illustrate that mothers of higher skill tend to work longer hours. As their mothers works longer hours, the children's skills go down, but more quickly for the children with higher skilled mothers.[23] The figure also shows that, for the range of hours in the data, the curves are approximately linear. On average, the estimates in Table 3 imply more negative βs than the homogeneous, linear estimate in Table 2: the average estimated effect in the sample is -0.027 compared to -0.19 in the homogeneous case.

The estimates of $\delta(X)$ are also intuitive. While we continue to find evidence of positive selection (as in the homogeneous results), here we also find evidence that the positive selection is more intense for mothers with higher skills (that is, $\delta_A > 0$).

[22]Note that the terms multiplying both β_L and β_{AL} have L^2 in them, a very large number for most working mothers. Thus, the coefficient estimates are much smaller than the other estimates. For readability, we present these estimates multiplied by 1,000.

[23]For some children in our sample whose mothers have sufficiently low skills, Table 3 actually implies positive effects. In particular, of the 6,924 children in our sample, 592 have positive estimated effects of maternal hours on skills. These positive effects are quite small, however, with an average of just 0.002 and a maximum of 0.005.

5.3 Further Heterogeneity by Pre-Birth Wages

The results so far suggest a misalignment between the work incentives facing mothers and the short-term benefits this work yields for their children's skills. The mothers who work the most are on average those whose work has the most negative short-term consequences for their children. Moreover, it does not appear that working longer hours is the reason why maternal work is particularly detrimental to the skills of the children from higher-skilled mothers. One explanation for this pattern is that the additional money earned by these higher-skilled mothers might not be enough to offset the higher opportunity cost of working in terms of the foregone interactions with their child. In order to further investigate this hypothesis, we expand equation (8) by allowing for a third dimension of heterogeneity, the pre-birth wage rate of the mother, W:[24]

$$f(L, X; \boldsymbol{\beta}) = (\beta + \beta_A A + \beta_W W + \beta_L L + \beta_{AL} AL + \beta_{WL} WL)L. \tag{9}$$

We allow for the slope and convexity of the effect of hours to depend on maternal skills and wage rates in a separable way.[25] Intuitively, we want to separately identify the effect in skills due to the loss of interaction between the child and the mother because she is working and any potential gains to skills flowing from the additional money received by the mother due to this work.[26] We assume that the heterogeneous effects via A (AFQT skills) holding W (pre-birth wages) constant tend to incorporate mostly the loss of interaction between the mother and the child, while the heterogeneous effects via W holding A constant tend to incorporate mostly the potential for additional earnings to offset this loss, which may happen via higher-quality child care, increased goods purchases (better food, more books, etc.), a reduction in parental stress, or in many other ways that may be difficult to observe. Indeed, it is plausible that the skills that are valued in the job market (affecting wages) aside from AFQT may have only a small effect on the quality of the interaction between the mother and the child during the first three years of the life of the child.

Table 4 presents the estimates of the function $f(\cdot)$ as specified in equation (9). For this table, we must restrict the sample to women who were working prior to giving birth and we must include the pre-birth wage rate and its interaction with AFQT score as additional controls. The results from the more restricted functions $f(L, X; \boldsymbol{\beta})$ estimated in Tables 2 and 3 are similar when we impose this sample restriction. As expected, the estimates are more noisy once this third dimension of heterogeneity is included. However, this exercise is still informative. Comparing our preferred symmetric estimates to the analogous estimates in Table 3 reveals a number of noteworthy results. First, as expected, the heterogeneity in the effect by the mother's skill (β_A) becomes more intense,

[24] Specifically, W denotes a residualized wage measure in which the effect of the age of the mother at birth and year fixed effects have been removed. We standardize this measure so that it is in standard deviation units like our AFQT measure. The estimates of β_W and β_{WL} should thus be comparable to the estimates of β_A and β_{AL}.

[25] We also experimented with allowing for further heterogeneity by also including the interaction $A \cdot W$. Including this additional interaction barely changes the main estimates, while the interaction itself is imprecisely estimated.

[26] Using other variables in the data to more directly conduct this analysis is infeasible. There are two potential variables from the NLSY that speak more directly to whether the mother needs to distance herself from the child when working: whether the job has flexible working hours and how many hours per week the person works from home. Unfortunately, these variables have too many missing observations to be useful for our purposes.

Table 4: The Effect of Maternal Hours Worked on Early Childhood Cognitive Skills by the AFQT Score, Pre-Birth Wages, and Labor Supply of the Mother

	(i) Uncorrected No Controls	(ii) Uncorrected w/ Controls	(iii) Het. Tobit	(iv) Het. Uniform	(v) Het. Symmetric
β	-0.004	-0.002	-0.009	-0.013	-0.017
	(0.005)	(0.004)	(0.014)	(0.026)	(0.017)
β_A	0.039**	-0.016**	-0.027**	-0.047**	-0.035**
	(0.004)	(0.004)	(0.014)	(0.023)	(0.017)
β_W	0.007*	0.001	0.014	0.029	0.018
	(0.004)	(0.005)	(0.016)	(0.026)	(0.019)
β_L ($\times$1000)	0.002	0.001	0.001	0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
β_{AL} ($\times$1000)	-0.012**	0.005**	0.006**	0.007**	0.007**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
β_{WL} ($\times$1000)	-0.002	0.001	-0.000	-0.001	-0.000
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
δ			0.006	0.010	0.013
			(0.010)	(0.022)	(0.013)
δ_A			0.008	0.027	0.015
			(0.009)	(0.020)	(0.013)
δ_W			-0.009	-0.024	-0.014
			(0.010)	(0.021)	(0.014)

Note: This table shows estimates of the effect of an additional 100 hours per year working in the three years following the child's birth on the child's early cognitive skills. $N = 3,994$. Bootstrapped standard errors in parentheses (1,000 bootstrap samples). In addition to the controls mentioned in Tables 2 and 3, these models also include pre-birth wages (W) and the interaction of pre-birth wages and maternal AFQT ($W \cdot A$) as elements of X. We specify $m(X)$ to enter both linearly and as indicators of each cluster of X, with a total of $K = 50$ clusters. $\mathbb{E}[L^*|L = 0, X]$ is also estimated with $K = 50$ clusters, and $\delta(X)$ is estimated with $K_\delta = 1$ cluster. See Figure 15 and Table 8 for the analogous results for different values of K and K_δ, respectively. ** p<0.05, * p<0.1.

since now it does not incorporate the offsetting income mechanism. Second, there is some evidence of a positive offsetting effect due to wages (β_W), although this is not significant at standard levels. Note also that there is a bit more evidence of nonlinearity by mother's skill (β_{AL}), and there is no evidence of nonlinearity by wages (β_{WL}). Still, the linearity assumption made in the literature continues to be a reasonable approximation for the range of hours, skills and wages in the sample.

Because of the noisy results, it is difficult to rule out some positive value for β_W. A positive value for β_W is intuitive: mothers with higher pre-birth wages are likely paid more for each hour they work post-birth, and this additional income could be beneficial to their child's skills through a variety of mechanisms already discussed. Nonetheless, it seems that there is much less heterogeneity across different values of W than across different values of A.[27] To see this, consider the point estimates

[27]We also estimate specifications analogous to Table 3 but with pre-birth wages taking the place of AFQT, also

of $\beta_A = -0.035$ and $\beta_W = 0.018$ from the table, and note that a mother who is 1 s.d. above the average in terms of A tends to be only one quarter of a s.d. above the average in terms of W. While the child of such a mother is expected to lose 0.035 s.d. in cognitive skills for each additional hour she works (relative to the average mother), only $0.018/4 = 0.0045$ s.d. would be expected to be offset by her higher earnings. In other words, a mother who is 1 s.d. above the average in skills would have to earn roughly 8 times the salary that she would be expected to earn given her skill to fully offset the short-run impact on the cognitive skills of her child.

Although more noisy, the estimates of $\delta(X)$ continue to suggest some positive selection, and disproportionately positive selection for higher-skilled mothers, as before. The table also shows some evidence that the selection might be less positive for high-skill, high-wage mothers relative to high-skill, low-wage mothers.

6 Sensitivity and Robustness

In this section, we show that our key findings in Section 5 are robust to violations of the identifying assumptions we make. Recall that we make two identifying assumptions: (a) a distributional assumtion on $\eta|X$ (Assumption 1, 1′ or 1″), and (b) the linearity assumption in equation (5). We consider violations of each of these assumptions in turn.

6.1 The Distributional Assumption

We consider what happens with our estimates of $\boldsymbol{\beta}$ if our estimator of $\mathbb{E}[L^*|L=0,X]$ was biased, which would happen if our distributional assumption (Assumption 1, 1′ or 1″) was invalid. We begin with our homogeneous results from Section 5.1. Let $\tilde{\beta}$ and $\tilde{\mathbb{E}}$ be the quantities identified by our approach, which may be different from the true quantities β and $\mathbb{E}$, respectively. Caetano et al. (2021) shows that the bias in the identification of β ($\mathbb{B}_\beta = \tilde{\beta} - \beta$) can be written as a function of the bias in the identification of the expectation ($\mathbb{B}_\mathbb{E} = \tilde{\mathbb{E}} - \mathbb{E}$) as[28]

$$\mathbb{B}_\beta = -\frac{\mathbb{B}_\mathbb{E}}{\tilde{\mathbb{E}}} \cdot \delta. \tag{10}$$

This formula reveals an asymmetry: for a given value of δ, and for a given magnitude of the bias in the expectation, $|\mathbb{B}_\mathbb{E}|$, it is preferable to err towards $\tilde{\mathbb{E}}$ being larger in magnitude than $\mathbb{E}$. This means that, all else the same, it is generally less consequential for our estimator of β if we err towards our estimator of $\mathbb{E}[L^*|L=0,X]$ being too negative rather than not negative enough.

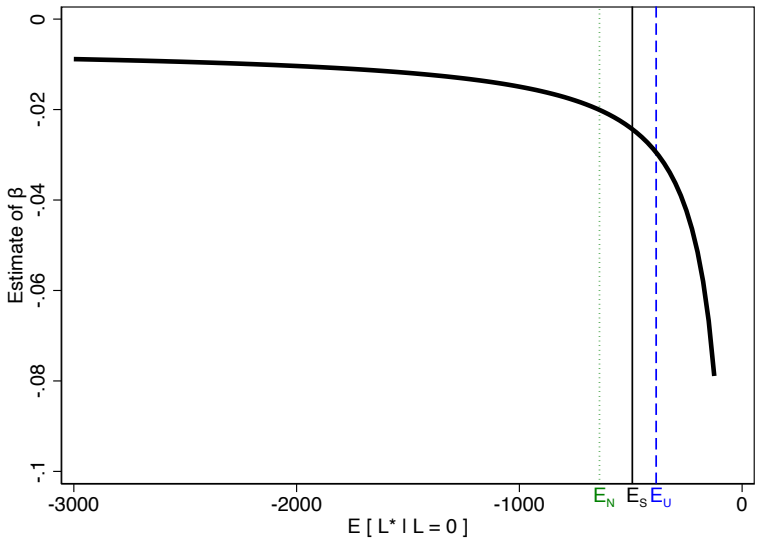
This asymmetry is confirmed in Figure 5, which shows as a thick black curve the estimates of β that we would have obtained for each possible counterfactual value of the expectation $\tilde{\mathbb{E}}[L^*|L=$

obtaining little evidence for heterogeneous effects by the mother's wage even when AFQT skills are not explicitly specified in the function $f(L,X;\boldsymbol{\beta})$.

[28] This is the formula without controls. The general formula is more cumbersome and can be seen in Remark 2.3 in that paper, but, essentially, the main take away is that controls tend to absorb some of the bias from the expectation, which tends to reduce $\mathbb{B}_\beta$ for each given value of $\mathbb{B}_\mathbb{E}$, δ, and $\tilde{\mathbb{E}}$.

$0, X]$ that we could have plugged in equation (6).[29] The vertical lines show the actual estimated expectations based on Assumption 1′ (Semiparametric Normal, dotted green line), Assumption 1″ (Semiparametric Uniform, dashed blue line) and Assumption 1 (Nonparametric Tail Symmetry, solid black line), so that the implied β estimates in the vertical axis of the panel represent the estimates analogous to the ones reported on columns (iii)-(v) from Table 2. Two main conclusions

Figure 5: Model from Section 5.1: $\hat{\beta}$ for each counterfactual value of $\tilde{\mathbb{E}}[L^*|L=0]$



Note: The thick black curve shows, for different counterfactual values of $\tilde{\mathbb{E}}[L^*|L=0]$, what would be $\hat{\beta}$ obtained from regression equation (6) in the homogeneous case of Section 5.1. The vertical lines represent the weighted average of the estimates of $\mathbb{E}[L^*|L=0,X]$ across all $K=50$ clusters, obtained from Assumptions 1 (solid black line), 1′ (dotted green line), and 1″ (dashed blue line). $N=6,924$.

stand out. First, the qualitative finding that $\beta<0$ does not depend on the value of $\tilde{\mathbb{E}}$, and thus does not depend on the distributional assumption we make. Second, there is substantially more scope for our estimates from columns (iii)-(v) from Table 2 to be underestimating the magnitude of β than be overestimating it. Indeed, if we had used a value of $\tilde{\mathbb{E}}$ that was more negative than what we in fact estimated and used, then our estimate of β would have changed little, while if we had used instead a value of $\tilde{\mathbb{E}}$ that was less negative than what we estimated, then our estimate of β would have become more negative than what we estimated, potentially substantially so.

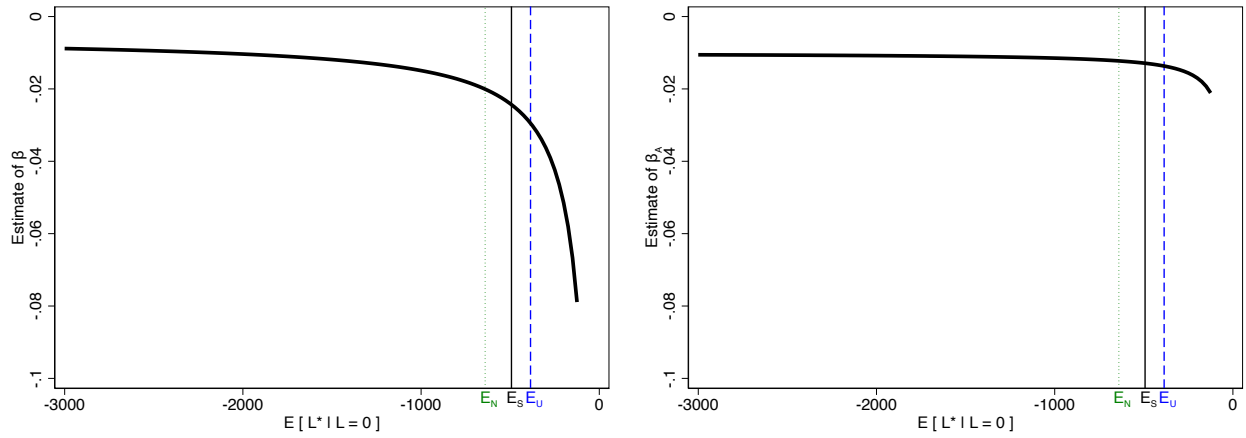
The overall take-away is that the qualitative conclusions from Section 5.1 would be the same for any distributional assumption we could have made, while the quantitative conclusions may change depending on the particular distributional assumption.

Analogous results are shown in Figure 6 for the main parameters of the heterogeneous model from Section 5.2. It is clear that our key qualitative conclusions that $\beta<0$ and that $\beta_A<0$ do not

[29] In order to be able to show the results of this analysis in one plot only, for simplicity we consider the case where $\tilde{\mathbb{E}}[L^*|L=0,X]$ is the same for all $K=50$ clusters of X, which yields slightly different estimates from when we allow for $\tilde{\mathbb{E}}[L^*|L=0,X]$ to vary arbitrarily per cluster. We considered more general cases that are more cumbersome to report, but found similar conclusions. Indeed, it is easy to see that this simplification should not be consequential to our conclusions, since our estimates of β for $K=1$ are very similar to our estimates of β for $K=50$ (Figure 11 in Appendix C).

depend on the estimated value of the expectation. Moreover, note that our quantitative conclusions about β_A depend less on the estimated expectation than our quantitative conclusions about β.

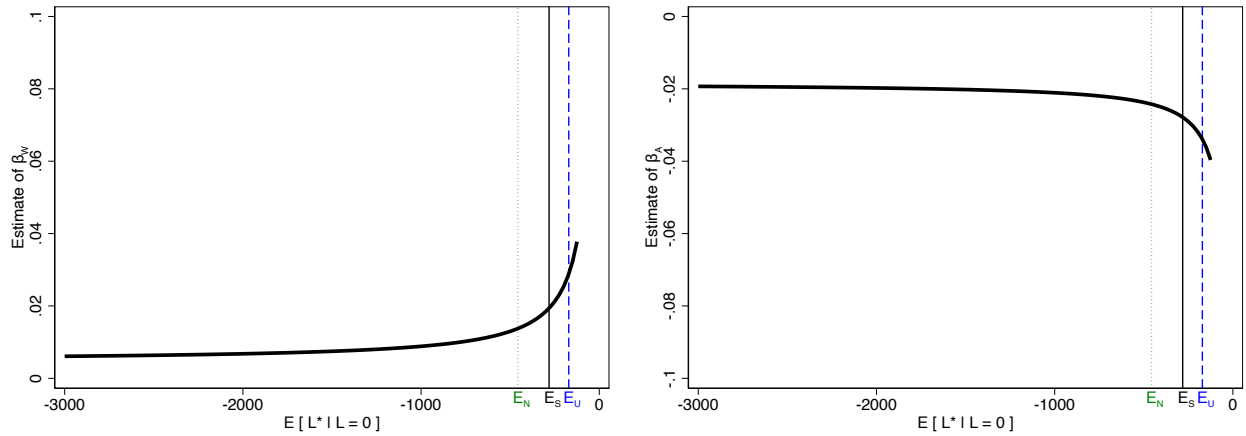
Figure 6: Model from Section 5.2: $\hat{\beta}$ and $\hat{\beta}_A$ for each counterfactual value of $\tilde{\mathbb{E}}[L^*|L=0]$



Note: This figure is analogous to Figure 5, but for β and β_A from Section 5.2 rather than β from Section 5.1. Note that the value of $\tilde{\mathbb{E}}[L^*|L=0]$, whatever it may be, must be the same in both panels of the figure. $N = 6,924$.

Finally, Figure 7 shows analogous results for the main parameters of the heterogeneous model from Section 5.3.[30] It is clear that our key qualitative conclusions that $\beta_W \geq 0$, $\beta_A < 0$ and that $|\beta_A| > |\beta_W|$ again do not depend on the estimated value of the expectation. Moreover, our estimates of β_W and β_A are similarly sensitive to different values of $\tilde{\mathbb{E}}$.

Figure 7: Model from Section 5.3: $\hat{\beta}_W$ and $\hat{\beta}_A$ for each counterfactual value of $\tilde{\mathbb{E}}[L^*|L=0]$



Note: This figure is analogous to Figure 6, but for β_W and β_A from Section 5.3 rather than β and β_A from Section 5.2. Note that the value of $\tilde{\mathbb{E}}[L^*|L=0]$, whatever it may be, must be the same in both panels of the figure. $N = 3,994$.

In sum, our four main qualitative conclusions in the paper: (a) $\beta < 0$, (b) $\beta_A < 0$, (c) $\beta_W \geq 0$ and (d) $|\beta_A| > |\beta_W|$ do not depend on the particular distributional assumption we make. Regarding

[30] Note that the vertical lines in Figure 7 are closer to zero, relative to the vertical lines from Figures 5 and 6. This is intuitive, as Figure 7 restricts the sample to mothers who worked prior to the birth of the child. These working mothers who choose $L = 0$ after birth are likely on average closer to indifference between working and not working after birth, $L^* = 0$, than the full sample of all mothers choosing $L = 0$ from Figures 5 and 6.

our quantitative conclusions, our estimate of β is more sensitive to the particular distributional assumption than our estimates of β_A and β_W.

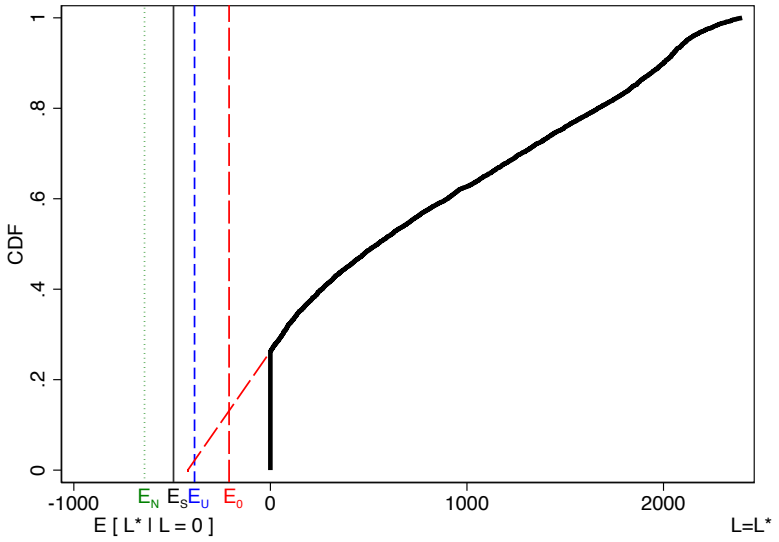
Remark 6.1. *To make further progress on understanding whether our findings about β are quantitatively robust, it is necessary to restrict the range of possible values of $\mathbb{E}[L^*|L=0]$.*[31] *Recall that a mother of type $L^*=0$ is exactly indifferent between working and not working. Thus, $\mathbb{E}[L^*|L=0]$ represents how far from indifference is the average mother who chooses $L=0$.*

Figure 8 presents a useful benchmark method of estimating $\mathbb{E}[L^*|L=0]$ which gives insight into the quantitative robustness of any conclusions regarding β. The right side of the figure shows as a thick black curve the cumulative density function (CDF) of L, with bunching of 25% of mothers at $L=0$. The right side of the figure shows different vertical lines representing different estimates of the average type of these bunched mothers. Besides the Normal, Tail Symmetry and Uniform estimates which were reported in Figure 5, we also show a new estimate of the expectation as a long-dashed vertical red line, $\mathbb{E}_0$. This estimate is obtained by extrapolating the slope of the CDF (black thick curve) as it approaches $L=0$ from positive values of L, as shown by the long-dashed red line.[32] *This estimation method provides a useful benchmark: it assumes that all negative types near indifference between working and not working (i.e., $L^*<0$ near $L^*=0$) are as common as the type who is exactly indifferent between working and not working, $L^*=0$. Because all mothers of types $L^*\leq 0$ need to add up to 25% of the sample, the method implicitly assumes that there are no mothers with type L^* lower than the intersection of the long-dashed red line with the horizontal axis, which in the figure occurs at a value between $\mathbb{E}_S$ and $\mathbb{E}_U$. If one is willing to assume that the long-dashed vertical red line is conservative, in the sense that it underestimates how close to indifference between working and not working is the average mother at $L=0$, then $\mathbb{E}_0$ would be an upper bound of $\mathbb{E}$. Plugging in this upper bound of $\mathbb{E}$ into the schedule depicted in Figure 5, this would allow us to conclude that the β estimate in that figure would not be more negative than -0.06. The corresponding implied estimates for β_A and β_W in Figure 7 are respectively -0.04 and 0.04.*

[31] We abstract from controls X here for simplicity. The argument of course could be made separately by X.

[32] See Caetano, Caetano, and Nielsen (2022) for estimation details.

Figure 8: A useful benchmark for $\mathbb{E}$



Note: The right side of the figure shows the CDF as a black thick curve, while the left side shows as vertical lines the different estimates of $\mathbb{E}[L^*|L=0]$ from Figure 5 plus a new benchmark estimate, $\mathbb{E}_0$ (long-dashed red vertical line). This benchmark estimate comes from the assumption that mothers with negative types around $L^*=0$ are as common as mothers of type $L^*=0$ (Caetano et al. 2022). The implied distribution of types of the mothers at $L=0$ under this assumption is also shown as a long-dashed red line. $N=6,924$.

6.2 The Linearity Assumption

The linearity assumption states that the effect of the confounder η on outcome S, which we are able to identify at $L=0$, can be extrapolated to be the same for other values of L (see Remark B.1 in Appendix B for further details).

In this section, we study how our main findings would change under violations of this linearity assumption. We consider three sensitivity analyses that are complementary to each other. First, we extrapolate the effect of L only locally, for smaller values of $L>0$, in order to understand whether the estimates we obtain are sufficiently different from the estimates we obtain when the extrapolation is non-local. Second, we directly test for whether the effect of the confounder η on the outcome S, δ, is different for $L=0$ than for $L>0$. Third, we check how our main estimates change as we allow for non-linearities by allowing for the effect of the confounder on the outcome, δ, to change across observations with different values of L.

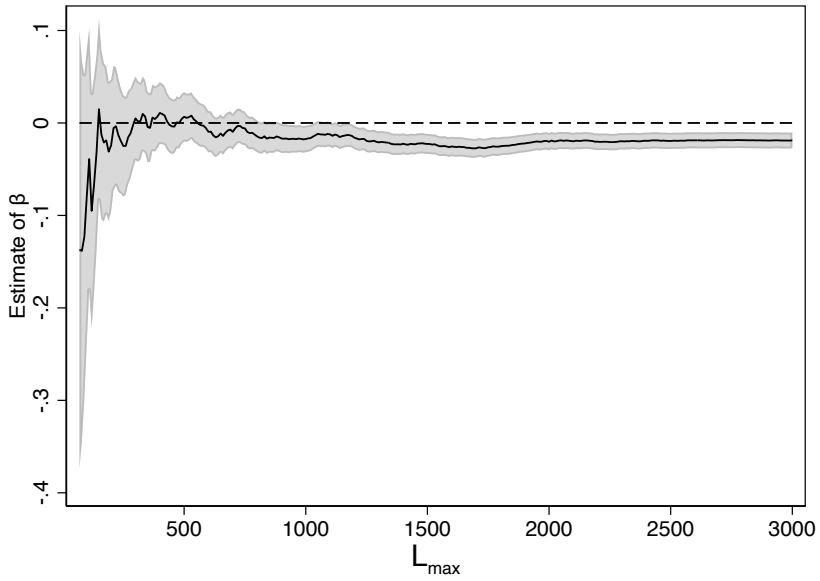
6.2.1 Local vs. non-local extrapolation

In the first sensitivity analysis, we restrict the sample to $L \leq L_{\max}$ for progressively larger values of $L_{\max}$. If our main conclusions are an artifact of the erroneous non-local extrapolation of the effect of η estimated at $L=0$, then we should expect that $\hat{\boldsymbol{\beta}}$ for lower values of $L_{\max}$ would lead to different conclusions than $\hat{\boldsymbol{\beta}}$ for higher values of $L_{\max}$.[33]

[33] We keep the expectation the same as we vary $L_{\max}$, so any change $\hat{\boldsymbol{\beta}}$ as $L_{\max}$ varies can only be attributed to non-linearities as we go from a local to a non-local extrapolation.

Figure 9 plots the homogeneous estimates of β corresponding to column (v) of Table 2 for different values of $L_{\max}$. Overall, the truncated estimates are generally indistinguishable from the full-sample estimates (which are the ones in the far right of the plot, when the restriction $L \leq L_{\max}$ is not binding). For local extrapolations around zero (i.e., for low values of $L_{\max}$), the point estimates for β are generally more negative than the full-sample estimate, although they are also imprecise. As the extrapolation becomes more and more non-local, the estimates become more precise and less negative, converging to the estimate shown in Table 2. Thus, Figure 9 provides evidence that our non-local linear extrapolation does not affect our qualitative conclusions regarding the sign of the average causal effect, but it is not capable of assessing whether the non-local extrapolation affects our quantitative findings, since the estimates for local extrapolations around zero are too imprecise.

Figure 9: Model from Section 5.1: Does the non-local linear extrapolation matter?



Note: Figure plots as a solid line estimates of β for restricted samples with $L \leq L_{\max}$. Bootstrapped 95% confidence intervals based on 250 iterations shown in gray. Estimates shown for $L_{\max} \geq 70$.

Figures 16 and 17 in Appendix C show analogous results for the main estimates from Sections 5.2 and 5.3, respectively. Although the estimates are very noisy for small values of $L_{\max}$, by $L_{\max} \approx 400$, when confidence intervals become narrower and some signal starts to emerge, it is clear that our conclusions about β_A and β_W do not seem to depend on the linearity assumption.

6.2.2 Testing for Non-linearities

Consider equation (5) for the homogeneous model from Section 5.1, with one important generalization,

$$S = \beta L + g(X) + \delta\eta + \delta^{+}\eta\mathbf{1}(L > 0) + \varepsilon, \quad \mathbb{E}[\varepsilon|L, X, \eta] = 0, \tag{11}$$

where we now allow for confounders at $L > 0$ to have a different effect on the outcome S than at $L = 0$, thus allowing for a violation of the non-linearity assumption, modulated by the value of δ^+. If $\delta^+ = 0$, we are in our baseline model where linearity is assumed for each value of L. Otherwise, non-linearities of the effect of η are allowed to exist between $L = 0$ and $L > 0$, but we still do not model any non-linearity for different values of $L > 0$.

We want to consider this specific source of heterogeneity in the effect of confounders, around $L = 0$, as the labor supply decision on the extensive and intensive margins might be completely different in nature. If indeed the confounders operating on these two margins are different from each other, then their average effect on the outcome may be different too, which would correspond to $\delta^+ \neq 0$.[34]

Re-writing equation (6) in this more general case yields:

$$\begin{aligned}\mathbb{E}[S|L,X] =&(\beta+\delta^+)L + \underbrace{g(X)-\delta h(X)}_{m(X)} - \underbrace{\delta^+ h(X)}_{\tilde{\delta}^+(X)}\mathbf{1}(L>0) \\ &+ \delta(L + \mathbb{E}[L^*|L=0,X]\mathbf{1}(L=0)).\end{aligned} \tag{12}$$

Two conclusions can be drawn from this equation. First, our estimate of β will be biased if $\delta^+ \neq 0$. Second, we can test whether $\delta^+ \neq 0$. Indeed, we can identify a new vector of coefficients, $\tilde{\delta}^+(X)$, which is scaled by δ^+.[35] Consider the test for whether $\delta^+ = 0$ (i.e., $H_0 : \delta^+ = 0$, $H_1 : \delta^+ \neq 0$ where H_0 and H_1 refer to the null and alternative hypotheses). Under the maintained assumption that $h(X) \neq 0$ among observations such that $L > 0$,[36] we can perform this test with a standard F-statistic by assessing whether $\delta_k^+ = 0$ for all clusters $k = 1, ..., K$, where $\tilde{\delta}^+(X) = \sum_{k=1}^K \delta_k^+ \mathbf{1}(X \in \hat{\mathcal{C}}_k)$.[37] Implementing this test in our data, we are unable to reject the null hypothesis that $\delta^+ = 0$ at standard levels of significance ($F = 0.95$, p-value $= 0.56$). We also fail to reject the null hypotheses for the heterogeneous models from Sections 5.2 and 5.3.

We conjecture that one of the reasons why the confounders at the extensive and intensive margins do not have detectable, different effects is because our treatment variable L is aggregated to the three-year level, where this distinction between extensive and intensive margins is plausibly less relevant. For further details, see the discussion in Footnote 14.

6.2.3 Allowing for non-linearities

Finally, we allow for non-linearities indirectly by allowing for the effect of the confounder, δ in equation (5), to vary arbitrarily with the cluster of X. Because observations in different clusters

[34] Note that η is simply the remainder in equation (3), so there is no assumption about the nature of this confounder in that equation. In particular, η can be different for each value of L^* and X in a flexible way, as $h(X)$ is a nonparametric function. Any assumption about confounders is embedded directly in equation (11).

[35] $\tilde{\delta}^+(X)$ can be separately identified from $m(X)$ provided $\mathbb{E}[L^*|L=0,X] \neq h(X)$, which is plausible since otherwise equation (3) would imply $\mathbb{E}[\eta|L=0,X] = 0$, i.e., X would be exogenous with respect to η among observations at $L = 0$.

[36] As it is clear from equation (3), this maintained assumption is trivially true in our case, as L^* and X are correlated to each other for observations such that $L > 0$.

[37] More specifically, if we specify $h(X) = \sum_{k=1}^K \gamma_k \mathbf{1}(X \in \hat{\mathcal{C}}_k)$, then $\delta_k^+ = \delta^+ \gamma_k$, since $\tilde{\delta}^+(X) = \delta^+ h(X)$.

tend to have different values of L, this approach indirectly allows for the effect of the confounder to vary with L, at least to some extent. If the confounders do in fact have non-linear effects, thereby creating bias in our estimator of β, then our β estimates should change as we allow for $\delta(\cdot)$ to be more heterogeneous across the clusters of X. Indeed, a more flexible specification for $\delta(\cdot)$ should allow us to indirectly improve our fit of any non-linearities in the effect of the confounder η.

To clarify, ideally we would like to have written equation (5) more generally as $\tilde{\delta}(X, L)$ instead of $\delta(X)$, but we cannot. Thus, the question is whether the restricted function $\delta(X)$ can partially absorb some of the non-linearity in case $\tilde{\delta}(X, L)$ is in fact non-linear in L. We therefore specify $\delta(X)$ as $\hat{C}_{K_\delta}(X)\boldsymbol{\delta}$, where $\hat{C}_{K_\delta}(X)$ is the vector of indicators for each of the K_δ cluster of X, and $\boldsymbol{\delta}$ is a vector of dimension K_δ (See Section 4.4 for details). Next, we show that our estimates of β do not change when we change this specification from $K_\delta = 1$ (i.e., $\delta(X)$ is specified as a constant) to $K_\delta = 50$ (i.e., the specification allows for unrestricted heterogeneity of δ across all 50 clusters of X). If it were the case that the true function $\tilde{\delta}(X, L)$ varies with L, then some of this non-linearity would be absorbed by $\hat{C}_{50}(X)\boldsymbol{\delta}$. Table 5 shows that the estimates of β when $K_\delta = 50$ are very similar to the estimates of β for $K_\delta = 1$ reported in Table 2. This suggests that allowing for non-linearities in L, at least the components that get projected on X, does not seem to change our results.

Table 5: Model from Section 5.1: Allowing for $\delta(X)$ to vary by cluster

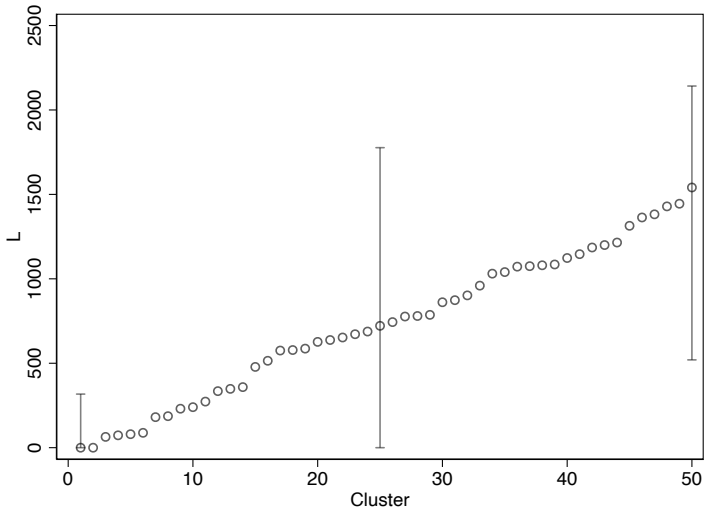
	(i) Uncorrected No Controls	(ii) Uncorrected w/ Controls	(iii) Het. Tobit	(iv) Het. Uniform	(v) Het. Symmetric
β	0.014**	0.000	-0.014**	-0.019**	-0.017**
	(0.001)	(0.001)	(0.005)	(0.007)	(0.006)

Note: This table shows estimates analogous to the estimate from column (v) from Table 2, but with $K_\delta = 50$ instead of $K_\delta = 1$. N=6,924. Bootstrapped standard errors in parentheses (1,000 bootstrap samples). ** p<0.05, * p<0.1.

This discussion and analysis presupposes that L does tend to vary across clusters. Figure 10 shows that different clusters of X do in fact have very different values of L. The figure shows the median value of L within each of the 50 clusters, sorted from the lowest to the highest median. The interquartile ranges for a few select clusters are also shown. Higher clusters tend to have a large proportion of mothers working long hours, while lower clusters tend to have a large proportion of mothers spending little or no time working. Thus, when we allow for δ to change arbitrarily across clusters, our specified function $\hat{C}_K(X)\boldsymbol{\delta}$ can fit some of the non-linearity on L. For instance, the δ for cluster 1 (the first element of the vector $\boldsymbol{\delta}$) will be a weighted average of the effects of the confounder for mothers with similar values of X, but also those that disproportionately work little or no hours L. Conversely, the δ for cluster 50 (the last element of the vector $\boldsymbol{\delta}$) will be a weighted average of the effects of the confounder for similar other values of X, but also those who disproportionately work long hours L. In other words, the heterogeneity in δ is not only picking up heterogeneity in the effect of η across different values of X, it is also picking up some heterogeneity in the effects

of η across different values of the treatment variable L. Yet is is precisely this heterogeneity that would be present if our linearity assumption were violated. Thus, the fact that our estimates do not change when we change the number of clusters in our specification suggests that the linearity assumption is a reasonable approximation in our context.

Figure 10: $\delta(X)$ picks up potential non-linearities on the effects of η across different values of L



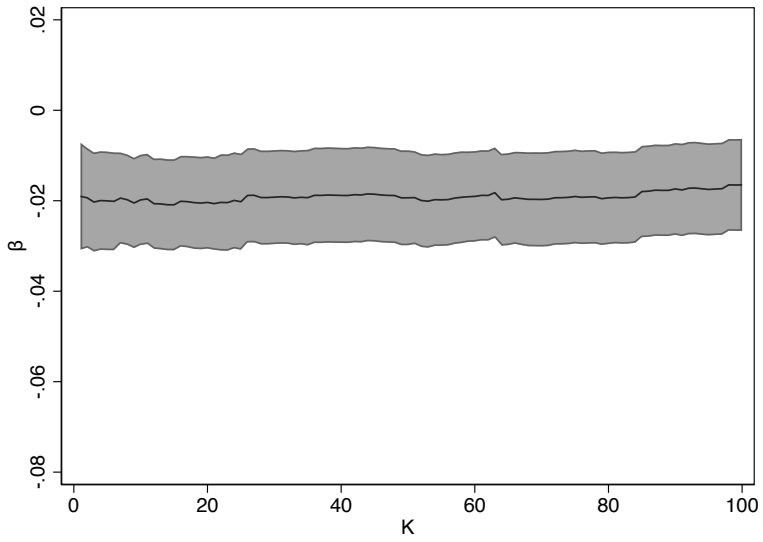
Note: This plot shows the median within cluster for each of the clusters in the case of $K = 50$, where clusters are sorted from left to right based on the median of L within cluster. The interquartile ranges within cluster for selected clusters are also shown.

Tables 7 and 8 in Appendix C show the analogous results for the heterogeneous analyses from Sections 5.2 and 5.3. The main estimates when we allow for δ to vary across $K_\delta = 50$ clusters are very similar to the estimates when $K_\delta = 1$ (Tables 3 and Tables 4, respectively). Our quantitative results seem to be robust to violations of our linearity assumption.

6.3 Controlling for X flexibly

The results from columns (iii)-(v) in Table 2 use $K = 50$ clusters to better approximate the non-parametric function $m(X)$ and the expectation $\mathbb{E}[L^*|L = 0, X]$ (see Section 4.4 for details). In Figure 11, we replicate the analysis in column (v) of Table 2 for $K = 1, \ldots, 100$. The estimates of β are very similar for all values of K, and remarkably so for K ranging from $K = 50$ all the way to $K = 100$. Since as K grows we are better able to approximate $m(X)$ and $\mathbb{E}[L^*|L = 0, X]$, this gives us confidence that $K = 50$ is sufficient for such approximations. Moreover, as discussed in Remark 4.3 in Section 4.4, the relative stability of our main estimate as we are effectively changing our expectation estimate provides further evidence that our conclusions do not seem to be dependent on our distributional assumption. The analogous findings for our results from Tables 3 and 4 are shown in Figures 14 and 15 in Appendix C.

Figure 11: Model from Section 5.1: Different Cluster Numbers (K)



Note: Estimates correspond to the estimate in column (v) from Table 2 (based on Assumption 1), but with different numbers of clusters used in the analysis. 95% confidence intervals based on 1,000 bootstrap iterations. $N = 6,924$.

7 Conclusion

In this paper, we estimate the effect of maternal hours worked in the first three years of life on early childhood cognitive skills. We correct for the endogeneity of labor supply using a control function approach that leverages the bunching of some mothers at zero hours worked. We use this novel identification strategy to estimate the average treatment effect of maternal labor supply for our entire sample. We also allow for heterogeneity in these effects by maternal skill, hours worked, and maternal pre-birth wages, which together improve our understanding of the trade-offs mothers and their families face when deciding whether and how many hours to work.

We find that working longer hours has a negative effect on the cognitive skills of the child, particularly for higher-skill mothers. Our results suggest that the presence of high-skill mothers in the home is especially valuable for childhood skill accumulation, at least in the short run.

The effects we estimate could operate through a number of distinct channels. First, the hours worked by the mother in the first three years of the child's life might have a direct effect on the child's skills measured around age 6. Second, the mother's labor supply in these early years might have a causal effect on her subsequent labor supply in years 3-6, which in turn might have a direct effect on the child's skills. Our estimates blend these two channels together – we plan to address each separately in future work.

Our results provide some useful insights for policies aimed at incentivizing mothers to work in the first years of their children's lives. We confirm the typical finding from the previous literature that there is some scope for such policies to generate short-run, negative effects on children's skills. Our main contribution lies in providing a detailed analysis of the heterogeneity in these effects along several dimensions, allowing us to assess for whom these unintended consequences are more likely

to be important. We find that there is little scope for such unintended consequences among low-skilled mothers, even for those who work long hours. However, we also find clear evidence of such unintended consequence among high-skilled mothers, even for those that are high earners, suggesting that additional earnings are not enough to compensate for their absence. On the one hand, our finding that these unintended consequences tend to be concentrated towards families with higher skilled mothers may be reassuring, since these are the families who are likely to have more resources to mitigate any potential downsides as the child grows. On the other hand, skill development is a dynamic process, and skill losses early in life may be particularly consequential. Understanding the long-run consequences of maternal labor supply and their implications for intergenerational inequality is an important research topic left for future work.

We conclude that policies aimed at increasing the flexibility of work arrangements are more likely to avoid these unintended negative consequences of maternal labor supply in early childhood. Such policies might allow mothers to maintain their work hours and career development while also allowing them to spend valuable time with their children. Flexible work schedules may also allow their partners to be a better substitute for their absence. The need for social distancing during the Covid-19 pandemic has led to marked increases in the share of workers operating under flexible work arrangements (e.g., work from home). In future work, it would be interesting to study the childhood development consequences of these recent changes, particularly if t hey prove persistent.

References

Adda, J., C. Dustmann, and K. Stevens (2017). The career costs of children. *Journal of Political Economy 125*(2), 293–337.

Agostinelli, F. and G. Sorrenti (2021). Money vs. time: Family income, maternal labor supply, and child development. *Working Paper* (273).

Altonji, J. G., T. E. Elder, and C. R. Taber (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy 113*(1), 151–184.

Anderson, P. M. and P. B. Levine (1999). Child care and mothers' employment decisions. Technical report, National bureau of economic research.

Andresen, M. E. and T. Havnes (2019). Child care, parental labor supply and tax revenue. *Labour Economics 61*, 101762.

Angrist, J. D. and M. Rokkanen (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association 110*(512), 1331–1344.

Arcidiacono, P., P. Bayer, and A. Hizmo (2010). Beyond signaling and human capital: Education and the revelation of ability. *American Economic Journal: Applied Economics 2*(4), 76–104.

Averett, S. L., H. E. Peters, and D. M. Waldman (1997). Tax credits, labor supply, and child care. *Review of Economics and Statistics 79*(1), 125–135.

Baker, M., J. Gruber, and K. Milligan (2008). Universal child care, maternal labor supply, and family well-being. *Journal of Political Economy 116*(4), 709–745.

Bartik, A. W., Z. B. Cullen, E. L. Glaeser, M. Luca, and C. T. Stanton (2020). What jobs are being done at home during the covid-19 crisis? evidence from firm-level surveys. Technical report, National Bureau of Economic Research.

Baum II, C. L. (2003). Does early maternal employment harm child development? an analysis of the potential benefits of leave taking. *Journal of Labor Economics 21*(2), 409–448.

Baydar, N. and J. Brooks-Gunn (1991). Effects of maternal employment and child-care arrangements on preschoolers' cognitive and behavioral outcomes: Evidence from the children of the national longitudinal survey of youth. *27*(6), 932.

Bernal, R. (2008). The effect of maternal employment and child care on children's cognitive development. *International Economic Review 49*(4), 1173–1209.

Bick, A., A. Blandin, and K. Mertens (2020). Work from home after the covid-19 outbreak. *CEPR Discussion Paper No. DP15000*.

Bick, A. and N. Fuchs-Schündeln (2017). Quantifying the disincentive effects of joint taxation on married women's labor supply. *American Economic Review 107*(5), 100–104.

Blau, D. and J. Currie (2006). Pre-school, day care, and after-school care: who's minding the kids? *Handbook of the Economics of Education 2*, 1163–1278.

Blau, D. M. (1999). The effect of income on child development. *Review of Economics and Statistics 81*(2), 261–276.

Blau, D. M. and A. P. Hagy (1998). The demand for quality in child care. *Journal of Political Economy 106*(1), 104–146.

Blau, F. D., A. J. Grossberg, et al. (1992). Maternal labor supply and children's cognitive development. *The Review of Economics and Statistics 74*(3), 474–481.

Blundell, R., M. Costa Dias, C. Meghir, and J. Shaw (2016). Female labor supply, human capital, and welfare reform. *Econometrica 84*(5), 1705–1753.

Bono, E. D., M. Francesconi, Y. Kelly, and A. Sacker (2016). Early maternal time investment and early child outcomes. *The Economic Journal 126*(596), F96–F135.

Bosch, N. and B. Van der Klaauw (2012). Analyzing female labor supply—evidence from a dutch tax reform. *Labour Economics 19*(3), 271–280.

Caetano, C. (2015). A Test of Exogeneity Without Instrumental Variables in Models With Bunching. *Econometrica 83*(4), 1581–1600. Available here.

Caetano, C., G. Caetano, and E. Nielsen (2021). Correcting Endogeneity Bias in Models with Bunching. *Working Paper*. Available here.

Caetano, C., G. Caetano, and E. Nielsen (2022). Partial Identification of Treatment Effects Using Bunching. *Working Paper*.

Carneiro, P., C. Meghir, and M. Parey (2013). Maternal education, home environments, and the development of children and adolescents. *Journal of the European Economic Association 11*(suppl_1), 123–160.

Carta, F. and L. Rizzica (2018). Early kindergarten, maternal labor supply and children's outcomes: evidence from italy. *Journal of Public Economics 158*, 79–102.

Chen, W., W. A. Grove, and A. Hussey (2017). The role of confidence and noncognitive skills for post-baccalaureate academic and labor market outcomes. *Journal of Economic Behavior & Organization 138*, 10–29.

Cogan, J. (1980). Labor supply with costs of labor market entry. In *Female labor supply*, pp. 327–364. Princeton University Press.

Cortes, P. and J. Tessada (2011). Low-skilled immigration and the labor supply of highly skilled women. *American Economic Journal: Applied Economics 3*(3), 88–123.

Currie, J. and D. Almond (2011). Human capital development before age five. In *Handbook of labor economics*, Volume 4, pp. 1315–1486. Elsevier.

Dahl, G. B. and L. Lochner (2012, May). The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit. *American Economic Review 102*(5), 1927–56.

Del Boca, D., C. Flinn, and M. Wiswall (2014). Household choices and child development. *Review of Economic Studies 81*(1), 137–185.

Desai, S., P. L. Chase-Lansdale, and R. T. Michael (1989). Mother or market? effects of maternal employment on the intellectual ability of 4-year-old children. *Demography 26*(4), 545–561.

Dingel, J. I. and B. Neiman (2020). How many jobs can be done at home? *Journal of Public Economics 189*, 104–235.

Eckstein, Z. and O. Lifshitz (2011). Dynamic female labor supply. *Econometrica 79*(6), 1675–1726.

Eissa, N. and J. B. Liebman (1996). Labor supply response to the earned income tax credit. *The Quarterly Journal of Economics 111*(2), 605–637.

Ettinger, A. K., A. W. Riley, and C. E. Price (2018). Increasing maternal employment influences child overweight/obesity among ethnically diverse families. *Journal of Family Issues 39*(10), 2836–2861.

Flood, S., J. McMurry, A. Sojourner, and M. Wiswall (2022). Inequality in early care experienced by us children. *Journal of Economic Perspectives 36*(2), 199–222.

Fogli, A. and L. Veldkamp (2011). Nature or nurture? learning and the geography of female labor force participation. *Econometrica 79*(4), 1103–1138.

Gathmann, C. and B. Sass (2018). Taxing childcare: Effects on childcare choices, family labor supply, and children. *Journal of Labor Economics 36*(3), 665–709.

Givord, P. and C. Marbot (2015). Does the cost of child care affect female labor market participation? an evaluation of a french reform of childcare subsidies. *Labour Economics 36*, 99–111.

Goux, D. and E. Maurin (2010). Public school availability for two-year olds and mothers' labour supply. *Labour Economics 17*(6), 951–962.

Grogger, J. (2003). The effects of time limits, the eitc, and other policy changes on welfare use, work, and income among female-headed families. *Review of Economics and Statistics 85*(2), 394–408.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica: Journal of the Econometric Society*, 679–694.

Hill, M. A. and J. O'Neill (1994). Family endowments and the achievement of young children with special reference to the underclass. *Journal of Human Resources*, 1064–1100.

Hirsch, B. T. and E. J. Schumacher (1998). Unions, wages, and skills. *Journal of Human Resources*, 201–219.

Hsin, A. and C. Felfe (2014). When does time matter? maternal employment, children's time with parents, and child development. *Demography 51*(5), 1867–1894.

James-Burdumy, S. (2005). The effect of maternal labor force participation on child development. *Journal of Labor Economics 23*(1), 177–211.

Kalenkoski, C. M., D. C. Ribar, and L. S. Stratton (2009). The influence of wages on parents' allocations of time to child care and market work in the united kingdom. *Journal of Population Economics 22*(2), 399–419.

Løken, K. V., M. Mogstad, and M. Wiswall (2012). What linear estimators miss: The effects of family income on child outcomes. *American Economic Journal: Applied Economics 4*(2), 1–35.

Meyer, B. D. and D. T. Rosenbaum (2001). Welfare, the earned income tax credit, and the labor supply of single mothers. *The Quarterly Journal of Economics 116*(3), 1063–1114.

Milligan, K. and M. Stabile (2011). Do child tax benefits affect the well-being of children? evidence from canadian child benefit expansions. *American Economic Journal: Economic Policy 3*(3), 175–205.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica 55*(4), 765–799.

Neal, D. A. and W. R. Johnson (1996). The Role of Premarket Factors in Black-White Wage Differences. *The Journal of Political Economy 104*, 869–895.

Nicoletti, C., K. G. Salvanes, and E. Tominey (2020). Mothers working during preschool years and child skills. does income compensate? *Forthcoming at Journal of Labor Economics, https://www.journals.uchicago.edu/doi/10.1086/719688*.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics 37*(2), 187–204.

Parcel, T. L. and E. G. Menaghan (1994). Early parental work, family social capital, and early childhood outcomes. *American Journal of Sociology 99*(4), 972–1009.

Polachek, S. W., T. Das, and R. Thamma-Apiroam (2015). Micro-and macroeconomic implications of heterogeneity in the production of human capital. *Journal of Political Economy 123*(6), 1410–1455.

Ruhm, C. J. (2004). Parental employment and child cognitive development. *Journal of Human Resources 39*(1), 155–192.

Ruhm, C. J. (2009). Maternal employment and child development. *Handbook of families and work*, 331–354.

Sayer, L. C. and J. C. Gornick (2012). Cross-national variation in the influence of employment hours on child care time. *European Sociological Review 28*(4), 421–442.

Todd, P. and K. Wolpin (2007). The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps. *Journal of Human Capital 1*(1), 91–136.

Vandell, D. L. and J. Ramanan (1992). Effects of early and recent maternal employment on children from low-income families. *Child development 63*(4), 938–949.

Waldfogel, J., W.-J. Han, and J. Brooks-Gunn (2002). The effects of early maternal employment on child cognitive development. *Demography 39*(2), 369–392.

Yamaguchi, S., Y. Asai, and R. Kambayashi (2018). Effects of subsidized childcare on mothers' labor supply under a rationing mechanism. *Labour Economics 55*, 1–17.

Zabel, J. E. (1993). The relationship between hours of work and labor force participation in four models of labor supply behavior. *Journal of Labor Economics 11*(2), 387–416.

A Sensitivity Analysis Based on Oster (2019)

The estimates of δ in Table 2 are positive and significant, implying positive selection – a higher value of η increases both childhood skills and maternal labor supply. Positive selection is intuitive and is consistent with the discontinuities presented in Figures 3 and 4 in which the mothers bunched at zero hours were shown to have discontinuously lower levels of variables that are known to be positively correlated with childhood cognitive skills.

In this appendix, we provide additional evidence that the degree of selection implied by our estimates is plausible. We do this by implementing the method proposed in Oster (2019), which itself builds on Altonji et al. (2005). The method requires as inputs the estimates of β and the R^2 from the regressions we ran in columns (i) and (ii) from Table 2. Given these inputs, under the assumptions from Oster (2019), we can infer the amount of selection on unobservables relative to selection on observables that are implied by the true value of β being identical to the estimate of β from column (v) in Table 2.

We report this implied ratio of selection on unobservables by selection on observables, which we denote δ_{Oster}, in Table 6. Following Oster (2019), we show these results for different potential values of R_{max}, which is the R-squared of a hypothetical regression of S on L, our observable controls, and all unobservable confounders (including some that we may have not have controlled for with our control function approach). For all possible values of R_{max}, our conclusions from Table 2 imply that selection on unobservables would be less pronounced than selection on observables, sometimes substantially less. For instance, if R_{max} is 0.70, then δ_{Oster} implies that our main estimates from Table 2 are compatible with selection on unobservables being about two-thirds as intense as selection on observables. If $R_{max} = 1$, which corresponds to the value of R_{max} suggested in Altonji et al. (2005), selection on unobservables need to be only about 40% as intense as selection on observables.

These results suggest that one does not need particularly strong selection on unobservables to rationalize our results. Importantly, Oster (2019)'s method requires completely different assumptions than ours. In particular, it does not use bunching in L and it does not make the distributional assumption we make. Thus, we view the results in Table 6 as providing independent confirmation of the plausibility of our corrected estimates.

Table 6: Proportional selection of observables and unobservables, Oster (2019).

True $\beta = -0.019$						
R_{max}	0.50	0.60	0.70	0.80	0.90	1.00
δ_{Oster}	1.12	0.82	0.65	0.54	0.46	0.40

Note: The table shows the values of δ_{Oster} as in Oster (2019) for different values of R_{max} when the true effect is $\beta = -0.019$, our estimate from column (v) of Table 2. δ_{Oster} can be interpreted as the degree of selection on unobservables relative to observables.

B Relationship to Regression Discontinuity Designs

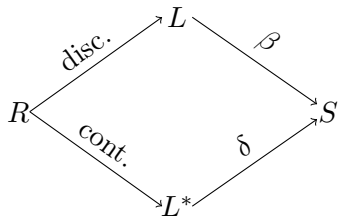
In this Appendix, we clarify the relationship between our method and regression discontinuity designs (RDDs). To simplify the discussion, we consider the case without controls X. In that case, equations (3) and (5) simplify to

$$S = m_0 + \beta L + \delta L^* + \varepsilon, \tag{13}$$

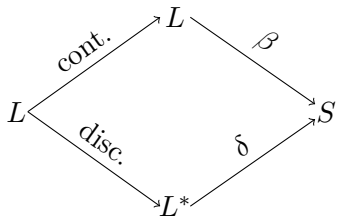
where $m_0 = g_0 - \delta h_0$ denotes the constant, with g_0 denoting the constant from $g(X)$ in equation (5) and h_0 denoting the constant from $h(X)$ in equation (3). Figure 12 provides a graphical representation of the parallel between these two methods using this simple linear model without controls, where the arrows represent causal relationships. We are interested in the effect of L on S, denoted β, and if $\delta \neq 0$, we have an endogeneity problem, since L^* includes all potential confounders.

Figure 12: Relationship between RDD and Our Approach

(a) RDD



(b) Our Approach



The left panel of this figure shows the RDD case. Typically, in an RDD there is a "running variable" R such that the treatment variable L varies discontinuously with R at a known threshold, while the confounder L^* varies continuously with R at the same threshold. Thus, whenever the outcome variable S varies discontinuously with R at the threshold, we can infer this discontinuity happened through the L channel and not through the L^* channel. We can identify β by simply dividing the intention-to-treat (reduced-form) effect, which is the discontinuity in S as R crosses the threshold, by the first stage, which is the discontinuity in L as R crosses the threshold.

In contrast, our approach is shown in the right panel of the figure. It can be seen as an "upside down" RDD that aims to identify the effect of L^*, rather than the effect of L. To see this, consider L as also the "running variable", using the language of RDD. The treatment L varies continuously with the running variable L, as they are the same variable. However, as discussed in Section 4.2, L^* varies discontinuously with the running variable L at the threshold, $L = 0$. Thus, any discontinuity in the outcome S at $L = 0$ can be attributed solely to the L^* channel. More formally,

this "intention-to-treat" or "reduced-form" discontinuity is[38]

$$ITT := \lim_{l \to 0^+} \mathbb{E}[S|L = l] - \mathbb{E}[S|L = 0] =$$
$$= \lim_{l \to 0^+} \mathbb{E}[m_0 + \beta L + \delta L^*|L = l] - \mathbb{E}[m_0 + \beta L + \delta L^*|L = 0] = -\delta \mathbb{E}[L^*|L = 0]$$

as $\lim_{l \to 0^+} \mathbb{E}[\beta L + \delta L^*|L = l] = 0$, since $L^* = L$ for $L > 0$. Moreover, the "first stage" discontinuity in this case is

$$FS := \lim_{l \to 0^+} \mathbb{E}[L^*|L = l] - \mathbb{E}[L^*|L = 0] = -\mathbb{E}[L^*|L = 0].$$

Unlike an actual RDD, we cannot directly identify the "first stage" discontinuity, since L^* is not observed. This is why we need to make a distributional assumption to identify $\mathbb{E}[L^*|L = 0]$. Once this "first stage" is obtained, we can identify δ simply by dividing the "intention-to-treat" effect by the "first stage" effect, as we do in RDDs. Once the confounder effect is identified, we can easily obtain the treatment effect by subtracting this selection bias term from the observed naive difference in outcomes, as discusssed in equation (1).

Remark B.1. ***Why do we need to make the Linearity Assumption?***
The parallels between our method and RDD also clarify why we need the linearity assumption in equation (5). Without the linearity assumption, we can still identify a meaningful treatment effect: the average treatment effect of L for those at $L = 0$. The reason we need to make the linearity assumption is that we want to identify the average treatment effect not only among those at $L = 0$. Thus, we need to extrapolate by assuming that the effect of L^* we identify at $L = 0$ is the same as the effect of L^* for $L > 0$. This extrapolation is where the parallel with the standard RDD ends: RDD studies most often focus on identifying the effect only at the threshold. RDD studies that attempt to extrapolate in order to identify non-local effects typically make similar assumptions. For instance, Angrist and Rokkanen (2015) identify effects away from the threshold by using an assumption of ignorability of the running variable conditional on other predictors of the outcome. This ignorability assumption is closely related to the linearity assumption we make: our assumption is that, conditional on covariates X, the outcome does not get affected by the "running variable" L beyond its treatment effect component ($f(L, X; \boldsymbol{\beta})$ in terms of equation (6)) and its (linearly assumed) endogeneity effect component ($\delta(X)[L + \mathbb{E}[L^*|L = 0, X]\mathbf{1}(L = 0)]$ in terms of equation (6)). For more context, Section 6.2.1 shows different estimates of β when we only make a linearity assumption locally, in order to estimate the effect around the bunching point. This is analogous to RDDs, which typically focus on estimating the effect around the cutoff with a local linear polynomial.

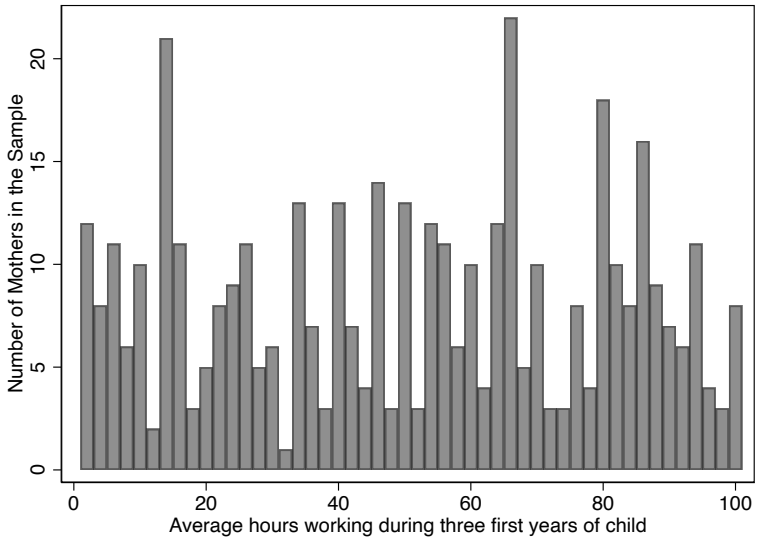
Remark B.2. ***Controls*** *Another important difference between RDDs and our approach is the use of controls. Typically, in RDDs controls X are added in the analysis only for efficiency purposes.*

[38] Similarly to the RDD literature, we use the notation $\lim_{l \to 0^+} \mathbb{E}[V|L = l]$ to refer to the limit of the expectation of a generic random variable V as l approaches 0 from the positive side. Unlike RDDs, the threshold is at the extreme of the support of the distribution of the running variable, at $L = 0$, so we do not need to take the limit from the other side.

In contrast, as discussed in Remarks 4.2 and 4.3, in our approach controls are useful to weaken the linearity assumption, and to test both the linearity and the distributional assumptions. Covariates are also useful for both methods to provide support for the identification strategy. In RDDs, one typically indirectly argues that unobservables vary continuously at the cutoff by showing that observables vary continuously at the cutoff. Analogously, we indirectly argue that unobservables vary discontinuously at the bunching point by showing in Figure 3 that observables vary discontinuously at the bunching point.

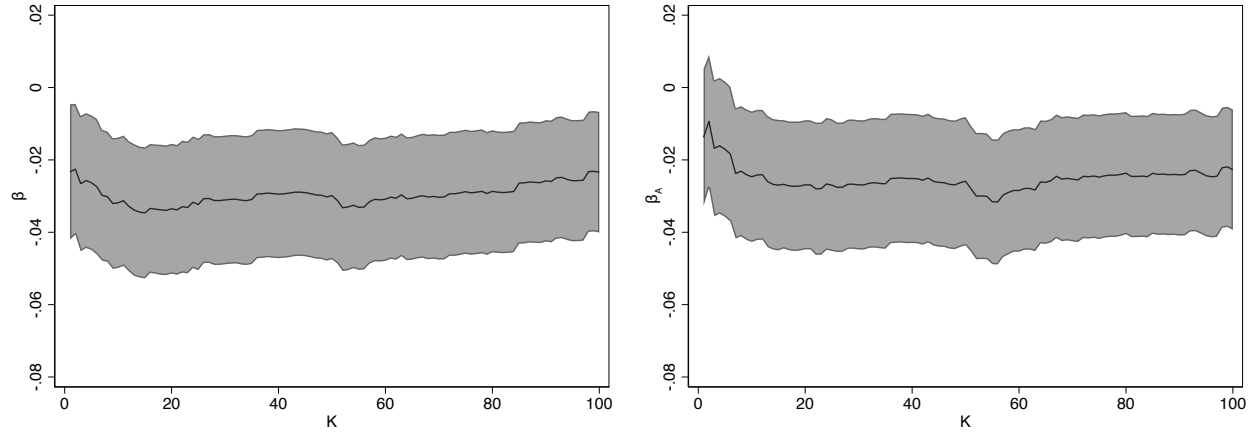
C Other Tables and Figures

Figure 13: L can be understood as a continuous variable around $L = 0$



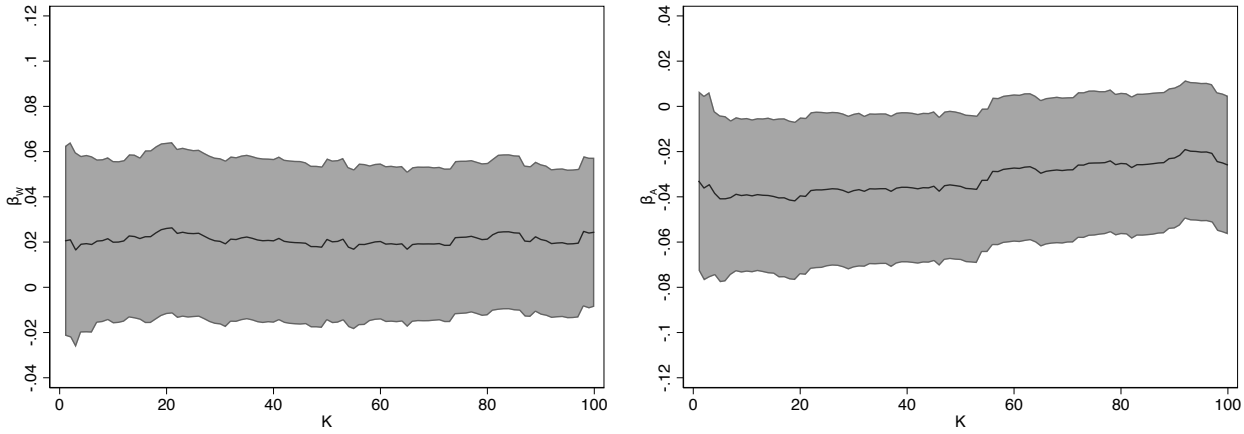
Note: This Figure shows the histogram of $0 < L \leq 100$ for the full sample, using a bandwidth equals to 2 hours.

Figure 14: Model from Section 5.2: Different Cluster Numbers (K)



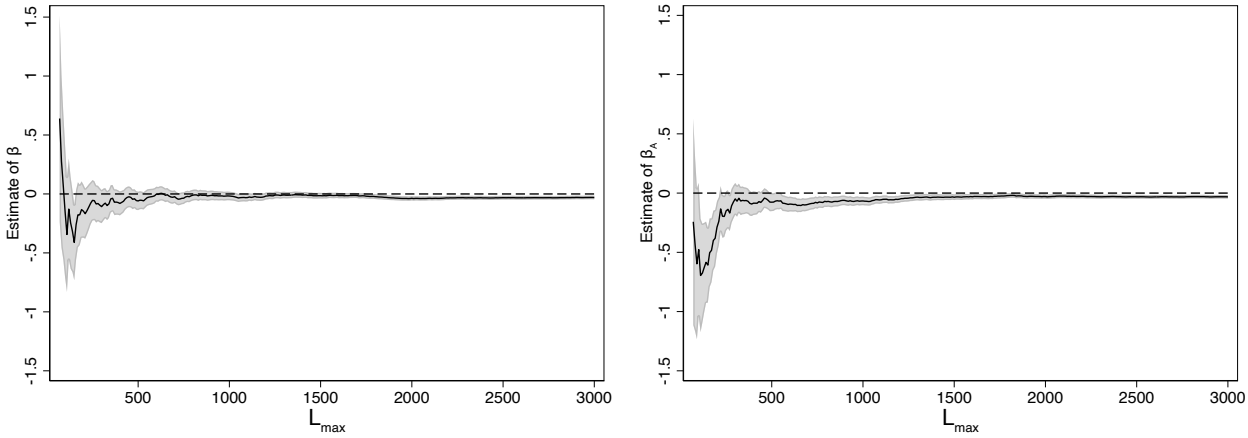
Note: Estimates correspond to the estimate in column (v) from Table 3 (based on Assumption 1), but with different numbers of clusters used in the analysis. 95% confidence intervals based on 1,000 bootstrap iterations. $N = 6,924$.

Figure 15: Model from Section 5.3: Different Cluster Numbers (K)



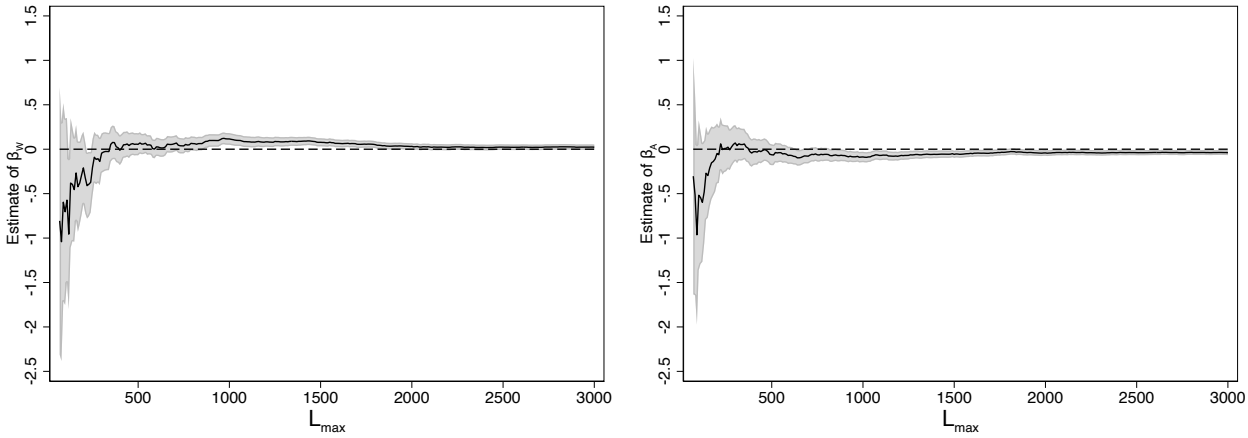
Note: Estimates correspond to the estimate in column (v) from Table 4 (based on Assumption 1), but with different numbers of clusters used in the analysis. 95% confidence intervals based on 1,000 bootstrap iterations. $N = 3,994$.

Figure 16: Model from Section 5.2: Does the non-local linear extrapolation matter?



Note: Figure plots as a solid line estimates of β (left panel) and β_A (right panel) for restricted samples with $L \leq L_{\max}$. Bootstrapped 95% confidence intervals based on 250 iterations shown in gray. Estimates shown for $L_{\max} \geq 70$.

Figure 17: Model from Section 5.3: Does the non-local linear extrapolation matter?



Note: Figure plots as a solid line estimates of β_W (left panel) and β_A (right panel) for restricted samples with $L \leq L_{\max}$. Bootstrapped 95% confidence intervals based on 250 iterations shown in gray. Estimates shown for $L_{\max} \geq 70$.

Table 7: Model from Section 5.2: Allowing for $\delta(X)$ to vary by cluster

	(i) Uncorrected No Controls	(ii) Uncorrected w/ Controls	(iii) Het. Tobit	(iv) Het. Uniform	(v) Het. Symmetric
β	0.018**	0.003	-0.023**	-0.031**	-0.030**
	(0.004)	(0.003)	(0.009)	(0.011)	(0.010)
β_A	0.047**	-0.010**	-0.019**	-0.024**	-0.026**
	(0.003)	(0.004)	(0.009)	(0.010)	(0.010)
β_L ($\times$1000)	-0.004**	-0.001	0.002	0.002	0.002
	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
β_{AL} ($\times$1000)	-0.015**	0.003**	0.004**	0.004**	0.004**
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
δ_A			0.008	0.013*	0.014**
			(0.005)	(0.007)	(0.006)

Note: This table shows estimates analogous to the estimate from column (v) from Table 3, but with $K_\delta = 50$ instead of $K_\delta = 1$. N=6,924. Bootstrapped standard errors in parentheses (1,000 bootstrap samples). ** p<0.05, * p<0.1.

Table 8: Model from Section 5.3: Allowing for $\delta(X)$ to vary by cluster

	(i) Uncorrected No Controls	(ii) Uncorrected w/ Controls	(iii) Het. Tobit	(iv) Het. Uniform	(v) Het. Symmetric
β	-0.004	-0.002	-0.010	-0.021	-0.020
	(0.005)	(0.004)	(0.014)	(0.027)	(0.019)
β_A	0.039**	-0.016**	-0.024*	-0.040	-0.026
	(0.004)	(0.004)	(0.014)	(0.025)	(0.017)
β_W	0.007*	0.001	0.016	0.029	0.024
	(0.004)	(0.005)	(0.016)	(0.026)	(0.019)
β_L	0.002	0.001	0.001	0.001	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
β_{AL} ($\times$1000)	-0.012**	0.005**	0.006**	0.006**	0.006**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
β_{WL} ($\times$1000)	-0.002	0.001	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
δ_A			0.006	0.021	0.009
			(0.010)	(0.022)	(0.014)
δ_W			-0.010	-0.024	-0.018
			(0.010)	(0.021)	(0.014)

Note: This table shows estimates analogous to the estimate from column (v) from Table 4, but with $K_\delta = 50$ instead of $K_\delta = 1$. N=3,994. Bootstrapped standard errors in parentheses (1,000 bootstrap samples). ** p<0.05, * p<0.1.