PARENTHOOD TIMING AND GENDER INEQUALITY

JULIUS ILCIUKAS*

Amsterdam School of Economics, University of Amsterdam

February 17, 2025

Abstract

I study how parenthood affects women's labor market outcomes and gender inequality, using quasi-experimental variation in the success of assisted conception procedures. To account for births following an initially failed procedure, I develop a method to quantify treatment effects in quasi-experimental settings with dynamic non-compliance, where individuals may opt to undergo assignment multiple times. Using administrative data from the Netherlands, I find that parenthood persistently reduces women's work hours and income by 9 to 27 percent. Despite these substantial effects, I find that at least half of the observed within-couple gender inequality in these outcomes after childbirth cannot be attributed to parenthood. I propose a unified framework to disentangle and quantify the bias in conventional estimators arising from selective parenthood timing and timing-dependent effects, demonstrating that these factors are the key to reconciling conflicting findings in the literature.

Keywords: Parenthood, gender inequality, treatment effects, dynamic non-compliance

*I thank Jérôme Adda, Francesco Agostinelli, Monique de Haan, Christian Dustmann, Phillip Heiler, Christine Ho, Artūras Juodis, Jura Liaukonyte, Hessel Oosterbeek, Erik Plug, Benjamin Scuderi, Arthur Seibold, Giuseppe Sorrenti, Mel Stephens, Bas van der Klaauw, Yun Xiao, Basit Zafar, Alminas Zaldokas, Lina Zhang, Yang Zhong, conference participants at AFEPOP, the Berlin School of Economics Gender Workshop, COMPIE, EALE, EEA-ESEM, ESPE, the Luxembourg Gender and Economics Workshop, the Warwick Economics PhD Conference, and seminar participants at the Chinese University of Hong Kong, Monash University, Peking University HSBC Business School, Singapore Management University, the University of Lausanne, the University of Michigan, and the University of Pennsylvania. The data used in this paper is available through the Microdata services of Statistics Netherlands. All URLs accessed February 17, 2025.

1 Introduction

The differential impact of parenthood on the careers of women and men is widely considered a major contributor to gender disparities in the labor market (Goldin, 2014; Bertrand, 2020; Cortés & Pan, 2023; Kleven et al., 2024). Quantifying this impact is central to understanding gender inequality and informing policy.

Assessing the career impact of parenthood is challenging because of selection and dynamic effects. Selection arises when fertility decisions—or the timing of parenthood—correlate with labor market outcomes independent of parenthood, such as when women with greater career potential delay childbearing. Dynamic effects arise when the impact varies by timing, such as when early parenthood permanently hinders career progression or when delaying parenthood results in missed opportunities at the peak of one's career.

Two reduced-form approaches are central to the debate, each addressing either selection or dynamic effects but not both, yielding conflicting results. The event study approach (Kleven et al., 2019, 2024) accounts for dynamic effects but assumes fertility timing is not selective. The approach by Lundborg et al. (2017) addresses selection using in vitro fertilization (IVF) success as a quasi-experiment in an instrumental variable framework. Because most women eventually conceive after initial IVF failure, it assumes effects are independent of timing. In Denmark, IV estimates suggest that parenthood has a minimal effect on gender inequality in earnings, whereas event study estimates—both in the general population and the IVF sample—attribute most of the inequality to parenthood (Lundborg et al., 2024).

In the first part of this paper, I propose a new method to quantify the effects of parenthood by leveraging assisted conception procedures (ACP), such as IVF or artificial insemination, while simultaneously addressing selection and dynamic effects. The method compares labor market outcomes between women who conceive during their first ACP and those who remain childless after its failure. A key challenge is that women who remain childless after a failed first ACP may systematically differ from those who conceive. The key innovation in addressing this challenge is to leverage quasi-experimental variation across women's entire ACP histories. This is complex because the decision to pursue additional ACPs may be selective and because some

women may conceive without ACPs. I first demonstrate that a specific weighting scheme—assigning higher weight to childless women with more failed ACPs—is sufficient to address selection into additional ACPs. Then, I use a bounding procedure to account for non-ACP births, assuming the most extreme selection consistent with the data.

The method does not rely on assumptions about heterogeneity across individuals or time, nor about selection into subsequent ACPs or non-ACP conception. The only core assumption is that the success of each ACP is as good as random, conditional on observables—an assumption used in existing work that relies on first-procedure success but here extended to subsequent procedures. The identified bounds are sharp, meaning that no effect within them can be ruled out without additional assumptions or data. To tighten the bounds, I assume that women who have non-ACP children after a successful ACP would have had at least one child if ACPs had failed, reflecting a determination to have at least one child over having additional children.

I apply my approach to a novel Dutch administrative dataset linking detailed labor market information from tax records with comprehensive hospital records on ACPs. My analysis focuses on couples trying to conceive their first child through intrauterine insemination, also known as artificial insemination. I find that parenthood persistently reduces women's annual work hours by 10%–24% and income by 6%–32%. These effects last at least seven years after first birth. For men, the bounds are similar in length but centered near zero. Over this period, parenthood causes 36%–54% of within-couple gender inequality in work hours and up to 46% in income.

In the second part of this paper, I focus on the conflicting findings between leading methods. Because Danish event study estimates are nearly identical for ACP and general samples but differ substantially from IV estimates (Lundborg et al., 2024), understanding why the methods diverge is crucial for assessing the generalizability of ACP-based findings. While differences in estimates within the same sample may indicate bias in either method, other factors may also contribute. First, the methods estimate effects for different moments of becoming a parent. Second, they consider different counterfactual scenarios for not having children—ranging from not attempting to conceive to trying but failing—which may have distinct mental health and relationship implications. Third, they estimate effects for different subpopulations.

If these factors substantially influence estimates even within the ACP sample, ACP-based estimates may offer limited insight into effects in the general population.

I begin by replicating IV and event study results using Dutch data, revealing substantial differences consistent with Danish findings. Then, I show that differences in the moment of becoming a parent and severe mental health or relationship consequences following failed conception have little explanatory power for the discrepancy. Afterward, I assess selective timing and dynamic effects within a consistent subpopulation. I start by bounding the effects of delaying parenthood, showing that it may have minimal impact or cause IV estimates to substantially understate motherhood's career costs. Next, I use the timing of failed ACPs as a proxy for fertility decisions, comparing women's actual childless career trajectories with those of women who postponed motherhood—mirroring the event study approach. I find that early mothers are negatively selected, while early fathers are positively selected, leading the event study estimates to overstate parenthood's role in gender inequality. This bias alone is substantial enough to reconcile the conflicting results across methods and samples, supporting the extrapolation of ACP-based estimates to the general population.

My work is linked to the large literature on the effects of children on gender inequality in the labor market (see Bertrand (2011), Blau & Kahn (2017), and Olivetti et al. (2024) for an overview).[1] It is most closely related to two recent working papers that exploit women's first in vitro fertilization procedure and carefully address dynamic effects: Bensnes et al. (2023) and Gallen et al. (2023). These studies rely on assumptions about effect heterogeneity across individuals and time. These assumptions are closely related to wave-ignorability assumptions used by Ferman & Tecchio (2023) and Angrist et al. (2024) in methodological work on dynamic non-compliance. The key advantage of my method is that it does not require such assumptions. Us-

[1]This includes studies that focus on the extensive fertility margin (Rosenzweig & Wolpin, 1980; Bronars & Grogger, 1994; Angrist & Evans, 1996; Jacobsen et al., 1999; Iacovou, 2001; Cruces & Galiani, 2007; Maurin & Moschion, 2009; Hirvonen, 2009; Vere, 2011); studies that restrict dynamic effects but address selection using various quasi-experiments (Hotz et al., 2005; Agüero & Marks, 2008; Cristia, 2008; Miller, 2011; Brooks & Zohar, 2021; Gallen et al., 2023); and studies that address dynamic effects but restrict selection, including those using timing differences (Fitzenberger et al., 2013; Angelov et al., 2016; Chung et al., 2017; Bütikofer et al., 2018; Eichmeyer & Kent, 2022; Melentyeva & Riedel, 2023) and structural methods (Adda et al., 2017).

ing my identification results, I develop a test for homogeneity assumptions and find they are unsupported in my application, introducing substantial bias.[2] Additionally, compared to these studies, a key focus of my analysis is assessing how various factors contribute to differences across methods—including those that these studies restrict by assumption.

My primary empirical contribution is to provide estimates of the career impact of parenthood that simultaneously account for selective fertility and dynamic effects. I demonstrate that these factors may substantially bias leading estimators and are the key to reconciling the main conflicting findings in the literature. My secondary empirical contribution concerns improved external relevance compared to existing studies leveraging ACPs, all of which focus on IVF and Scandinavian data. By focusing on intrauterine insemination, which is less costly, less invasive, and more accessible, I mitigate concerns about sample selectivity of IVF couples and procedure side effects. Additionally, using data from the Netherlands, where family policies align with the OECD average, makes my findings more relevant for common policy settings.

Methodologically, my approach builds on ideas from two branches of literature. The first step of my approach, which accounts for selection into parenthood via subsequent ACPs, leverages insights from the extensive biostatistics literature on dynamically assigned treatments (see Hernán & Robins (2020) for an overview). In economics, it is most closely related to a procedure developed by Van den Berg & Vikström (2022), which explicitly incorporates treatment assignment eligibility. The second step of my approach, which addresses selection into parenthood through non-ACP means, relates to the extensive literature on bounds for treatment effects, beginning with Manski (1989, 1990). It is most closely related to a procedure typically used to account for sample selection, introduced by Zhang & Rubin (2003) and further developed by Lee (2009). I present a detailed discussion of how my approach relates to and differs from these methods in Section 3.4.

My primary methodological contribution is an approach to bound treatment effects in settings with quasi-experimental assignment and dynamic non-compliance.

[2]Another advantage of my method is that it does not rely on a longitudinal data structure, allowing to estimate impacts on outcomes observed irregularly or only once.

Particularly, when individuals obtain treatment by undergoing multiple assignments or through entirely selective pathways. In addition to relaxing assumptions employed by the IV approach, the method quantifies effects for a broader group—those who comply with either the initial or subsequent treatment assignment, rather than only those who comply with the initial assignment. My method also makes it possible to assess effects over time for a stable group, which is not possible with the IV approach. Examples of other potential applications include educational programs with multiple admission cycles, job training programs where unassigned individuals can reapply, legal settings where individuals are assigned to judges with varying propensities to sanction and where unsanctioned individuals may reoffend and experience future sanctions, and clinical trials in the extension phases where participants can enroll in other trials or pursue alternative therapies.

The remainder of the paper is structured as follows. Section 2 introduces the model. Section 3 demonstrates the identification challenge, presents intuition for the approach, states the formal results, discusses relations to existing methodological literature, and outlines estimation. Section 4 describes the institutions, ACPs, and the data, and presents support for the assumptions. Section 5 presents the main estimates of the effects of parenthood on women's labor market outcomes and gender inequality. Section 6 covers generalizability, comparisons with existing methods, and concerns about mental health and relationship stability. Section 7 concludes.

2 Model

The model adapts the local average treatment effect (LATE) framework (Angrist & Imbens, 1995), incorporating a standard extension to a dynamic setting. The key innovation is formalizing how treatment—or parenthood status—depends not only on the initial treatment assignment (or the outcome of a woman's first ACP) but also on the decision to pursue additional assignments (or initiate subsequent ACPs) and the outcomes of those assignments (or procedure success).

All women start ACP for their first child at time $t = 1$, ticking up to $\bar{t}$. D_t is the treatment indicator, representing whether a woman has any children in period t. Treatment is an absorbing state. Women may have multiple children; I focus on

whether they have any. $Y_t(0)$ is the potential outcome in period t if a woman remains childless, which I refer to as the *control outcome*. For $k > 0$, $Y_t(k)$ is the potential outcome in period t if a woman becomes a mother in period k. I refer to $Y_t(1)$ as the *treated outcome*. A woman's realized labor market outcome in period t is Y_t, and the relationship between potential and realized outcomes is given by:

$$Y_t = Y_t(0)1_{\{D_T=0\}} + Y_t(1)1_{\{D_1=1\}} + \sum_{k=2}^{\bar{t}} Y_t(k)1_{\{D_k=1, D_{k-1}=0\}}.$$

I characterize each woman by two unobserved variables. First, $W_t \in \{1, \ldots, \overline{w}\}$ is the total number of ACPs a woman would undergo for her first child up to period t if all previous ACPs failed. I refer to W_t as the willingness to undergo ACPs, although it only describes women's behavior in the scenario where all ACPs fail and does not require any economic interpretation. Second, $R_t \in \{0, 1\}$ indicates whether a woman would remain childless up to period t if all W_t ACPs failed. I refer to R_t as the reliance on ACPs. I refer to women with $R_t = 1$ as *reliers*, reliant on ACPs to have children, and those with $R_t = 0$ as *non-reliers*, who would have children even if all ACPs failed. W_t and R_t may relate to potential outcomes and to each other and may be known or unknown to women ex-ante.

Reliers are the focus of this paper. They are the most general group of women whose parenthood status depends on ACP success.[3] They are closely related to compliers in the LATE framework—women who would remain childless if their first ACP failed ($C_t = 1$). However, reliers are a more general group, meaning compliers are a subset of reliers. Reliers additionally include women who would conceive through a subsequent ACP if the first ACP failed but would remain childless if all ACPs failed; such women are always-takers in the LATE framework ($C_t = 0$). There are no never-takers or defiers because few women who conceive via ACPs remain childless, but the approach can extend to a setting with never-takers.[4]

The observed indicator for the success of ACP j is Z_j. It takes the value 1 if the

[3]When outcomes have a bounded support, less informative bounds for all women can be obtained using Horowitz & Manski (2000).

[4]The simplest extension is using Zhang & Rubin (2003) bounds for two-side non-compliance in the second step.

ACP succeeded, and 0 either if the ACP failed or if a woman did not undergo ACP j. Note that the index for success of ACP j is different from the time index t, meaning that insofar there are no restrictions on how many ACPs a woman can undergo in a given period. To simplify notation, I only count ACPs that occur before the first child, meaning $Z_j = 0$ for all j such that $Z_l = 1$ for some $l < j$.

The realized number of ACPs a woman undergoes by period t is A_t. A woman undergoes ACPs either until one succeeds or until reaching her maximum willingness. Formally: $A_t = \min(\{j : Z_j = 1, j \leq W_t\} \cup \{W_t\})$. The outcome of a woman's last ACP up to period t is Z_{A_t}. A woman has a child in period t either if an ACP has succeeded or if she is a non-relier who would have a child even if all ACPs failed: $D_t = Z_{A_t} + (1 - Z_{A_t})(1 - R_t)$.

The individual-level treatment effect in period t is $\tau(t) = Y_t(1) - Y_t(0)$. The average treatment effect in period t is $\tau_{ATE}(t) = \mathbb{E}[\tau(t)]$. The average treatment effect in period t for reliers in period t is $\tau_{ATR}(t) = \mathbb{E}[\tau(t) \mid R_t = 1]$. The local average treatment effect in period t is $\tau_{LATE}(t) = \mathbb{E}[\tau(t) \mid C_t = 1]$. I discuss other effects under extensions.

3 Method

In this section, I first outline the limitations of the IV approach, introduce the intuition behind my method, and present formal results. I then relate my approach to the existing methodological literature and discuss estimation.

To demonstrate the intuition, I leverage the unconditional sequential unconfoundedness assumption:

Assumption 1 (Sequential unconfoundedness). $(Y_t(k), R_t, W_t) \perp\!\!\!\perp Z_j \mid A_t \geq j$, *for all* j, k, t.

It states that, among women who undergo ACP j, the outcome of ACP j is effectively random—independent of potential outcomes and type. This aligns with the standard unconfoundedness assumption in previous studies using IVF: among women undergoing embryo insertion into the uterus, pregnancy resulting from the procedure is essentially random. Unlike prior studies, this assumption covers not only the first,

but also subsequent embryo insertions women undergo. It is worth highlighting that this assumption does not concern the decision to undergo additional procedures, just the success of each individual procedure after it takes place. It also does not restrict how success rates vary across procedures. To simplify exposition, I do not distinguish between IVF and intrauterine insemination when discussing the intuition. The main method in Section 3.3 accounts for selection into procedure type, procedure-specific success rates, and other observed factors that influence success, such as age at the time of the procedure. I discuss empirical support for this assumption in Section 4.3.

I also impose a standard no-anticipation assumption used by existing methods:

Assumption 2 (No anticipation). $Y_t(k) = Y_t(0)$ *for all* $k > t$.

It states that outcomes before becoming a mother do not depend on having children in the future. This assumption is plausible for conception, as future success is unknown, but less so for adoption, which may be anticipated; however, adoptions are rare in my application.[5] I also assume that SUTVA holds for all potential outcomes and types. Finally, to ensure that results are independent of period definitions, I assume that all births following a failed first ACP occur after the first period: $\Pr(D_1 = 1|Z_1 = 0) = 0$.

3.1 Instrumental Variable Bias

The IV approach uses the success of women's first ACP as an instrument for parenthood. It starts with the reduced form, which is the difference in average outcomes between those whose first ACP succeeded and those whose first ACP failed. This means a group of women who conceived at their first ACP is compared to a mixed group consisting of childless women (the compliers) and women who had children later (the always-takers). Under unconfoundedness, the reduced form identifies a combination of two effects: the average treatment effect for compliers and the timing effect for always-takers. Assuming that all conceptions after the first ACP fails occur

[5]Each year, about 40 domestic adoptions occur in the Netherlands, and foreign-born children make up less than 1% of my sample, including those whose mothers were abroad at childbirth.

in the second period, for simplicity of exposition, for any $t > 1$:

$$\begin{aligned}\mathbb{E}[Y_t|Z_1 = 1] - \mathbb{E}[Y_t|Z_1 = 0] = &\mathbb{E}[Y_t(1) - Y_t(0)|D_t = 0, Z_1 = 0] \Pr(D_t = 0|Z_1 = 0) \\ &+ \mathbb{E}[Y_t(1) - Y_t(2)|D_t = 1, Z_1 = 0] \Pr(D_t = 1|Z_1 = 0).\end{aligned}$$

Scaling the reduced form by the difference in the share of mothers between the two groups—the first stage—yields:

$$\frac{\mathbb{E}[Y_t|Z_1 = 1] - \mathbb{E}[Y_t|Z_1 = 0]}{\mathbb{E}[D_t|Z_1 = 1] - \mathbb{E}[D_t|Z_1 = 0]} = \tau_{LATE}(t) + \mathbb{E}[Y_t(1) - Y_t(2)|C_t = 0]\frac{\Pr(C_t = 0)}{\Pr(C_t = 1)}.$$

When the outcomes do not depend on the moment of becoming a mother, meaning $Y_t(1) = Y_t(2)$, the second term drops out, and τ_{LATE} is identified. In the standard Rubin (1974) model with only one motherhood outcome, this assumption is covered by the no-multiple-versions-of-treatment (SUTVA). Otherwise, the second term biases the IV estimator of τ_{LATE}.[6]

In the context of parenthood, the bias direction is ambiguous. It may lead to underestimation of career costs if women who have children later face high care demands at the peak of their careers, or to overestimation if early motherhood permanently hinders career progression. Because 75% of women whose first ACP fails eventually have children, even small timing effects can introduce sizable bias.

3.2 Intuition

In this section, I present the intuition behind my bounding approach. I separately explain how I identify the relier average control outcome and bound their average treated outcome, and how I use additional information to tighten the bounds.

[6]Another way to describe this bias is using the negative weights terminology popularized by the recent difference-in-differences literature (see Roth et al. (2023) for an overview). With an always-taker-to-complier ratio of 3, the IV estimator assigns a weight of 4 to τ_{ATE} and -3 to the average effect of delayed parenthood for always-takers, $\mathbb{E}[Y_t(2) - Y_t(0)|C_t = 0]$. Difference-in-differences methods are inapplicable here because parenthood timing may be selective.

3.2.1 Control Outcome

To demonstrate how the relier average control outcome can be identified, I first express it as a weighted average of childless outcomes among reliers with different willingness to undergo ACPs, and then explain how each term in this expression is identified:

$$\mathbb{E}[Y_t(0)|R_t = 1] = \sum_{w=1}^{\overline{w}} \mathbb{E}[Y_t(0)|R_t = 1, W_t = w] \Pr(W_t = w|R_t = 1).$$

The average outcome among women who underwent exactly w ACPs and remained childless identifies the average control outcome for reliers willing to undergo exactly w ACPs:

$$\mathbb{E}[Y_t|A_t = w, D_t = 0] = \mathbb{E}[Y_t(0)|W_t = w, R_t = 1].$$

This holds because, first, control outcomes are realized among childless women. Second, women who underwent exactly w ACPs and remained childless form an as-good-as-random subsample of reliers willing to undergo exactly w ACPs. The latter follows from two key observations. First, only reliers who are willing to undergo exactly w ACPs can remain childless after undergoing exactly w unsuccessful ACPs, since non-reliers would have children, and those willing to undergo more than w ACPs would have done so. Second, conditional on being a relier willing to undergo exactly w ACPs, remaining childless is effectively random, determined solely by whether any ACP up to w succeeds, with each ACP outcome being as good as random.

The shares of different types can be identified following a similar argument. First, women who experience at least w failed ACPs form a random subsample of those willing to undergo at least w ACPs. Thus, the share of these women initiating a subsequent ACP identifies the share willing to undergo at least $w + 1$ ACPs in this group:

$$\Pr(A_t \geq w + 1 \mid A_t \geq w, Z_w = 0) = \Pr(W_t \geq w + 1 \mid W_t \geq w).$$

Second, women who do not undergo an additional ACP after w failed ACPs form a random subsample of those willing to undergo exactly w ACPs. Thus, the share of

these women who remain childless identifies the share of women reliant on ACPs in this group:

$$\Pr(D_t = 0 \mid A_t = w, Z_w = 0) = \Pr(R_t = 1 \mid W = w).$$

Combining these conditional probabilities allows me to construct $\Pr(W_t = w, R_t = 1)$ for all w, meaning that the shares of all types are identified.

When all women conceive solely through ACPs, making everyone a relier, the average control outcome is identified. Combined with the average treated outcome, this allows identification of τ_{ATE}. However, if some women conceive naturally, τ_{ATE} cannot be identified without additional assumptions, as non-relier control outcomes are unobservable. Furthermore, τ_{ATR} also cannot be identified, since treated outcomes are observed only for those whose first ACP succeeds, and the identity of reliers among them is unknown. I next describe how I bound the relier average treated outcome to bound τ_{ATR}.

3.2.2 Treated Outcome

I bound the relier average treated outcome using the distribution of outcomes among women whose first ACP succeeded. Since the success of the first procedure is as good as random, this distribution reflects the treated outcomes of all women entering ACPs. Combined with the relier share identified in the previous step, this allows me to construct worst-case bounds for the relier average treated outcome by assuming they either have the lowest or highest treated outcomes.

To illustrate the intuition, suppose there are 100 women whose first ACP succeeded, and the first step identifies that 80% of women are reliers. Then, by unconfoundedness, there are approximately 80 reliers among the 100 women, and their expected outcome is the same as the relier average treated outcome. While it is not known which 80 out of the 100 women are the reliers, the upper bound for their average treated outcome can be constructed by selecting the 80 women with the highest outcomes. The argument for the lower bound is symmetric.

3.2.3 Narrowing Bounds with Covariates

When ACP success is as good as random conditional on pre-ACP covariates, these covariates can help refine the bounds. For instance, suppose we identify that 80% of women in both high- and low-education groups are reliers. The most conservative starting point is to construct the lower bound by excluding the 20% of first-ACP mothers with the highest outcomes, disregarding education. However, if all of these 20% are high-educated, this selection procedure becomes inconsistent with the education-conditional relier shares, as we know that some non-reliers are low-educated. Since this was the most conservative starting point, any other selection can only produce the same or a higher lower bound. The new bounds are instead constructed using the lowest and highest outcomes within each education group.

3.2.4 Narrowing Bounds with Assumptions

The bounds can be narrowed further with additional assumptions on which women whose first ACP succeeded are (or are not) reliers. For instance, it could be reasonable to assume that women who had non-ACP children after their first ACP succeeded would have also had at least one non-ACP child if all ACPs had failed, ensuring they are not reliers. This assumption aligns with the economic idea that families are more determined to have their first child than to have additional children.

To illustrate why this helps, suppose there are 100 women whose first ACP succeeded. Further suppose that in addition to identifying that 80% of the 100 women are reliers, as before, 10 are observed to have an additional non-ACP child. The assumption implies that these 10 women are not reliers and they can be excluded before selecting the 80 potential reliers to construct the bounds. Selecting 80 women with either the highest or lowest outcomes from a subset of 90 women can only result in tighter bounds compared to selecting from a set of 100 women.

Formally, R_t^+ is an indicator for a woman's reliance on ACPs for additional children after becoming a mother through her first ACP. Specifically, R_t^+ takes the value 1 if, in the case that her first ACP succeeds, a woman would have only ACP children, and 0 otherwise. I refer to women who rely on ACPs for all subsequent children as *subsequent reliers*. D_t^+ is an indicator for having at least one non-ACP child, defined

as: $D_t^+ = Z_{A_t}(1 - R_t^+) + (1 - Z_{A_t})(1 - R_t)$.[7] In words, a woman has at least one non-ACP child either if an ACP succeeded and she is not a subsequent relier, or if all ACPs failed and she is not a relier.[8]

Assumption 3 (Monotonicity). $\Pr(R_t^+ \geq R_t) = 1$.

The monotonicity assumption states that women who are not reliant on ACPs for additional children are also not reliant on ACPs for their first child. This auxiliary assumption concerns behavior and is arguably stronger than previous assumptions. It does not imply a preference for multiple children over one or vice versa, nor does it rule out sterility. It states only that couples who conceive a child naturally after having an ACP child would have also done so if ACPs had failed. From a fertility choice perspective, it excludes only couples who prefer more children over fewer but would rather have none than just one or a few. Consequently, a sufficient condition is that couples prefer having at least one child over none—a reasonable assumption given that the sample consists of couples who chose to pursue assisted conception while likely aware that they might be unable to have multiple children.

Nonetheless, given the uncertainty of conception, the assumption may be violated if the success of the first ACP influences relationship stability or mental health—such as preventing separation or reducing depression—thereby increasing effort to conceive naturally and leading to non-ACP births that would not have occurred otherwise. I relax the assumption to allow for such violations in Section 6.4 and provide empirical support for both versions of the assumption afterward. Remaining theoretical results assume monotonicity.[9]

3.3 Relier Average Treatment Effect

In this section, I formalize and combine ideas introduced in Section 3.2 to bound τ_{ATR}. Since the method is effectively cross-sectional, I simplify notation by omitting the time

[7] For brevity, I do not distinguish between reliance on ACPs for subsequent children after becoming a mother via the first and subsequent ACPs; only former is relevant.

[8] It implies that women with at least one non-ACP child also have at least one child ($D \geq D^+$), and if ACPs fail, having a child is equivalent to having a non-ACP child ($D = D^+ \mid Z_A = 0$).

[9] Setting $R^+ = 1$, $D^+ = 0$ if $Z_1 = 1$ is equivalent to the case without monotonicity.

index t, with all unindexed variables and functions implicitly indexed by t. Before stating the formal results, I introduce the conditional sequential unconfoundedness assumption:

Assumption 4 (Conditional sequential unconfoundedness).
$(Y(k), R^+, R, W) \perp\!\!\!\perp Z_j \mid X_j$ *for all* j, k, *and* $X_j \in \mathcal{X}_j^1 = \{x \in \mathcal{X}_j : 1_{\{A \geq j\}} = 1\}$.

Where X_j are covariates at the time of ACP j, with support $\mathcal{X}_j$. To simplify notation, they include an indicator for whether the woman has undergone at least j ACPs, $1_{\{A \geq j\}}$. Covariates specific to ACP j are set to 0 if the woman does not undergo ACP j.[10] In words, the success of ACP j is independent of potential outcomes and type, conditional on undergoing at least j ACPs and covariates at the time of ACP j. The next assumption provides regularity conditions. Let $e_j(x) = \Pr(Z_j = 1 \mid X_j = x)$.[11]

Assumption 5 (Regularity).

1. $0 < \underline{e} < e_j(x) < \overline{e} < 1$ *for all* j *and* $x \in \mathcal{X}_j^1$, *for some fixed* $\underline{e}$ *and* $\overline{e}$.

2. Y *has a probability density function for* $Z_1 = 1$, $D^+ = 0$, *and all* $x \in \mathcal{X}_1$.

It contains two parts. First, the probability of ACP success conditional on undergoing the procedure and covariates at the time differs from 0 and 1. Second, Y is a continuous random variable conditional on the first ACP succeeding, having only ACP children, and any value of X_1. In practice, adding a negligible amount of continuously distributed noise to Y is sufficient to avoid ties in trimming without meaningful bias.

The bounding procedure begins with identifying several nuisance functions involved in the trimming step. First, the covariate-conditional relier share is identified using the weighted share of women without children among those whose ACPs failed:

$$r(x) = \mathbb{E}\left[\frac{(1 - D^+)\Pi_{j=1}^{\overline{w}}(1 - Z_j)}{\Pi_{j=1}^{\overline{w}}(1 - e_j(X_j))}\middle| X_1 = x\right].$$

Since $e_j(x_j)$ takes values above zero only for women who undergo ACP j, larger weights are given to women who underwent more ACPs. This accounts for the fact

[10] For $j > 1$, X_j also includes covariates from previous ACPs.

[11] An ACP cannot succeed unless a woman initiates the procedure: given j, $e_j(x) = 0$ for all $x \in \mathcal{X}_j \setminus \mathcal{X}_j^1$.

that women willing to undergo more ACPs are less likely to not experience ACP success, making them underrepresented in this group. Next, the covariate-conditional share of subsequent reliers is identified from the share of women having only ACP children among those whose first ACP succeeded $r^+(x) = \mathbb{E}\left[1 - D^+ \mid Z_1 = 1, X_1 = x\right]$. Under monotonicity, the covariate-conditional share of reliers among subsequent reliers is then $p(x) = r(x)/r^+(x)$.

The covariate-conditional quantile function of the treated outcome distribution among subsequent reliers is identified from the outcome distribution of women whose first ACP succeeded and who have only ACP children:

$$q(u, x) = \inf\left\{q : u \leq \Pr(Y \leq q \mid X_1 = x, Z_1 = 1, D^+ = 0)\right\}.$$

Finally, $q(p(x), x)$ and $q(1-p(x), x)$ identify the covariate-conditional $p(x)$-th and $1-p(x)$-th quantiles of the treated outcome distribution among subsequent reliers. These quantiles will be used to trim the tails of the outcome distribution and select reliers in the scenarios where they have either the lowest or the highest treated outcomes.

The nuisance functions are combined with the data to construct the moments:

$$m^L(G, \eta^0) = Y(1 - D^+)1_{\{Y < q(p(X_1), X_1)\}}\frac{Z_1}{e_1(X_1)} - Y(1 - D^+)\Pi_{j=1}^{\overline{w}}\frac{(1 - Z_j)}{(1 - e_j(X_j))}$$
$$m^U(G, \eta^0) = Y(1 - D^+)1_{\{Y > q(1-p(X_1), X_1)\}}\frac{Z_1}{e_1(X_1)} - Y(1 - D^+)\Pi_{j=1}^{\overline{w}}\frac{(1 - Z_j)}{(1 - e_j(X_j))},$$

where G is a vector containing all observed variables and η^0 contains the nuisance functions:

$$\eta^0(x_1, \ldots, x_{\overline{w}}) = \{r^+(x_1), r(x_1), q(p(x_1), x_1), q(1 - p(x_1), x_1), e_1(x_1), \ldots, e_{\overline{w}}(x_{\overline{w}})\}.$$

The first term in $m^L(G, \eta^0)$ is used to bound the relier average treated outcome. It assigns positive weights to women whose first ACP succeeded, who have only ACP children, and whose outcomes fall below the covariate-conditional trimming threshold $q(p(x), x)$. The second term, used to identify the relier average control outcome, assigns positive weights to outcomes of childless women. Larger weights are given to women who underwent more ACPs to account for the fact that reliers willing

to undergo more ACPs are less likely to not experience ACP success, making them underrepresented in this group. $m^U(G, \eta^0)$ mirrors this for the scenario where reliers have the highest treated outcomes. The moments are then scaled by the relier share.

Theorem. *Under Assumptions 2, 3, 4, and 5, sharp lower and upper bounds on* τ_{ATR} *are given by* $\theta_L = \mathbb{E}[m^L(G, \eta^0)]/\mathbb{E}[r(X_1)]$ *and* $\theta_U = \mathbb{E}[m^U(G, \eta^0)]/\mathbb{E}[r(X_1)]$.

3.4 Relation to Methodological Literature

My method integrates ideas from two branches of methodological literature. The first step, addressing selection via subsequent ACPs, draws on the literature for evaluating time-varying treatments (see Hernán & Robins (2020) for an overview). These methods, designed for settings with quasi-random assignment to sequential treatments, are most suitable for controlled experiments. In my setting, treatment assignment corresponds to conceiving via ACPs, but the possibility of non-ACP motherhood, or treatment without assignment, makes these methods unsuitable.

Even without selective treatment, a key distinction lies in the treatment assignment mechanism. In my model, the decision to initiate each subsequent ACP may be selective, with quasi-random treatment assignment occurring only upon initiation. Women who do not initiate an additional ACP cannot be assigned treatment. This differs from settings where all individuals have a non-zero probability of assignment to different regimes. In this respect, my model is related to Van den Berg & Vikström (2022), where individuals have a positive likelihood of being assigned treatment each period until they receive it or permanently exit eligibility. In my setting, however, treatment is not assigned at specific moments but depends on when and how often individuals choose to pursue it (e.g., undergo ACPs). As a result, selection cannot be captured by a duration variable but is determined by the number of procedures, their timing, and covariates at the time of each procedure.

The second step of my approach, which addresses selection via non-ACP means, is methodologically related to the Zhang & Rubin (2003) and Lee (2009) procedure (henceforth, ZRL) for handling unobserved outcomes in quasi-experimental settings. While dynamic effects are conceptually distinct from unobserved outcomes, the ZRL

approach can be adapted to bound τ_{LATE}.[12]

The key distinction from the adapted ZRL approach is that my method leverages women's entire ACP histories rather than just the first procedure to bound effects for reliers rather than compliers. This is advantageous not only because reliers are a more general group, but also because it guarantees narrower bounds. In application, I demonstrate that the ZRL bounds are several times wider, making them uninformative.[13]

3.5 Estimation

The bounds on τ_{ATR} can be estimated using sample averages of m^L and m^U after plugging in an estimate of the nuisance parameter η^0. With a few discrete covariates, asymptotic normality can be demonstrated building on Lee (2009). However, incorporating continuous covariates may be crucial for tighter bounds, requiring nonparametric estimation of the nuisance parameter, which complicates inference. To justify inference, I build on the estimation approach for the ZRL procedure introduced by Semenova (2023), leveraging orthogonalization and sample splitting. I modify it by incorporating the first step of my identification approach and present the new orthogonal moments and estimator in Appendix A2.[14] In Appendix A3, I discuss alternative estimation methods, all yielding results consistent with my main estimates.

[12] Treat outcomes of mothers whose first ACP fails as unobserved.

[13] Another distinction of my approach is the use of the monotonicity assumption. ZRL assumes treatment has a non-negative effect on being observed, I use non-ACP conceptions of additional children as an auxiliary variable to help identify non-reliers.

[14] I use 3-fold cross-fitting, estimating propensity scores via logistic regressions with quadratic terms for each partner's age at the procedure, interacted with procedure type and education dummies, using women initiating the respective ACP. I use the first ten ACPs, treating conceptions via other ACPs as natural. Following Heiler (2024), I estimate remaining nuisance functions using Generalized Random Forests (Athey et al., 2019), incorporating all propensity score covariates up to the current ACP, along with pre-ACP income and work hours for both partners. Confidence intervals for the bounds follow Stoye (2020).

4 Institutions, Procedures, and Data

In this section, I first describe Dutch family policies and the labor market context. Then, I discuss IVF and intrauterine insemination, and the differences between them. Afterward, I overview the data, provide empirical support for the sequential unconfoundedness assumption, and compare the ACP sample to the general population.

4.1 Family Policies in the Netherlands

Dutch women are entitled to 4 to 6 weeks of pregnancy leave before the due date and at least 10 weeks of maternity leave after birth, totaling at least 16 weeks.[15] During this period, they receive full wage replacement from the unemployment insurance agency (up to a daily limit). Fathers receive one week of fully paid leave within the first four weeks, covered by the employer.[16]

Children can enroll in private daycare from three months old. In 2022, 72% of children under two attended formal child care, averaging 20 hours per week (OECD, 2023a). After turning four and starting elementary school, they become eligible for out-of-school care. In 2023, families using child care paid an average of 8,950 euros, with 64% reimbursed, translating to a net cost equivalent to 10% of median disposable household income.[17]

The Netherlands has average family policies compared to other OECD countries. Paternity and maternity leave durations are slightly below the OECD averages of 2.5 and 21 weeks, respectively (OECD, 2023c). While formal child care enrollment for children under two is the highest among OECD countries, average time spent in care is the lowest (OECD, 2023a). After age four, enrollment rates and out-of-school care hours align with OECD averages (OECD, 2022).

While employment rates for mothers, fathers, and non-parents in the Netherlands

[15] For multiple births, women receive 20 weeks of leave; this has negligible impact on the results.

[16] A 2020 reform granted fathers up to six weeks of leave; most births in the data occurred earlier.

[17] www.cbs.nl/nl-nl/nieuws/2024/30/ouders-betaalden-gemiddeld-3-210-euro-aan-kinderopvang-in-2023, longreads.cbs.nl/materiele-welvaart-in-nederland-2024/inkomen-van-huishoudens/.

exceed their respective OECD averages, part-time work is far more common, making average hours worked comparable to the OECD average (OECD, 2023b). In 2021, the maternal employment rate was 80%, compared to the OECD average of 71%. However, in 2023, 52% of women and 18% of men worked part-time (less than 30 hours per week), more than twice the respective OECD averages (OECD, 2023d). Among two-parent families, only 14% had both parents working full-time, 52% had a full-time working father and a part-time working mother, and 12% had both parents working part-time.[18]

4.2 Assisted Conception Procedures

I use two types of ACPs: IVF, previously used to study the career impact of parenthood in Denmark and Sweden (Lundborg et al., 2017; Bensnes et al., 2023; Gallen et al., 2023; Lundborg et al., 2024), and intrauterine insemination (IUI), which has not been used for this purpose. Both procedures may begin with cycle tracking and hormonal stimulation to enhance egg production. IVF is a surgical procedure where eggs are retrieved through the vaginal wall, fertilized in the lab, and transferred as embryos into the uterus. It is relatively invasive, performed under sedation or anesthesia, and has a success rate of about 25% per embryo transfer. IUI involves injecting sperm directly into the uterus via a catheter. With a lower success rate of about 10%, IUI is significantly less invasive—lasting about five minutes and generally painless—and is the first-line infertility treatment in most countries. In the Netherlands, couples without a specific infertility diagnosis must typically undergo six IUI cycles before attempting IVF. Compulsory health insurance covers unlimited IUI and up to three IVF procedures. In 2022, each additional IVF cycle costs 4,000 euros, but since multiple embryos can be frozen per cycle, subsequent transfers may cost 1,000 or less.

4.3 Data

I use administrative data from Statistics Netherlands, covering all residents. ACP data span 2012-2017 and come from the Diagnosis-Treatment Combination system,

[18] www.cbs.nl/en-gb/news/2024/10/fewer-and-fewer-families-in-which-only-the-father-works

which Dutch hospitals are required to report to. My main variables are the procedure type—IVF or IUI—and the date of sperm or embryo insertion. ACP success is defined as having a child born within 10 months of insertion with no subsequent insertions, a definition validated against medical records by Lundborg et al. (2017).

Labor market data span 2011–2023, with annual work hours and gross labor income derived from tax records. Work hours include maternity leave, and income includes maternity pay. While leave pay accurately reflects women's financial situation, incorporating leave duration complicates the interpretation of work hours. To address this, I define maximum-leave-adjusted hours, scaling reported hours during each childbirth year by 36/52 to account for up to 16 weeks of leave. In my main analyses, I estimate upper bounds using reported hours and lower bounds using adjusted hours, ensuring that the effects on actual work hours fall within these bounds. Since existing methods do not naturally accommodate such adjustments, I use leave-adjusted hours in secondary analyses. The choice of measure only affects results in the first year of motherhood and does not impact the method comparisons or bias estimates.

I use several demographic variables, including an indicator for completing higher education, number of children, birth dates, and cohabitation status. My main sample consists of cohabiting opposite-sex couples undergoing intrauterine insemination for their first child. To ensure the first observed ACP is their actual first, I follow Lundborg et al. (2017) and exclude individuals whose first observed procedure occurred in the first data year, as they likely had prior ACPs. I also exclude those whose first ACP occurred in the last data year to prevent misattributing births from unobserved ACPs in the following year to failed ACPs. These restrictions have negligible impact on my results. My main sample includes 15,523 couples. For comparison with the general population, I use 376,157 women who were cohabiting with a male partner when they conceived their first child between 2013 and 2017, without prior ACPs. All analyses use the full samples, regardless of employment status.

Table 1 compares average characteristics of couples whose first ACP succeeded (column 1) and those whose first ACP failed (column 2). Labor market outcomes are measured in the year preceding each woman's first ACP. The two groups had similar average annual income, but women whose first ACP succeeded worked 30

Table 1: First ACP Outcomes and Descriptives

	Success (1)	Fail (2)	Dif. (1)-(2)	Cond. dif. (1)-(2) cond.	Rep. (5)	Suc. vs rep. (1)-(5)
Work (W)	0.882	0.863	0.019	0.008	0.801	0.080
	[0.323]	[0.344]	(0.009)	(0.009)	[0.001]	(0.010)
Work (P)	0.884	0.865	0.019	0.013	0.783	0.101
	[0.320]	[0.342]	(0.009)	(0.009)	[0.001]	(0.010)
Hours (W)	1240.315	1207.860	32.455	18.702	1076.204	164.111
	[604.666]	[635.194]	(16.183)	(16.560)	[1.135]	(16.856)
Hours (P)	1474.530	1438.590	35.940	18.579	1250.948	223.582
	[658.231]	[695.692]	(17.713)	(17.870)	[1.294]	(19.211)
Income 1000s € (W)	28.065	27.418	0.647	0.745	21.362	6.703
	[19.559]	[20.219]	(0.516)	(0.546)	[0.030]	(0.444)
Income 1000s € (P)	37.205	36.952	0.252	0.364	28.107	9.098
	[26.482]	[29.452]	(0.746)	(0.730)	[0.047]	(0.704)
Bachelor deg. (W)	0.480	0.451	0.029		0.411	0.069
	[0.500]	[0.498]	(0.013)		[0.001]	(0.012)
Bachelor deg. (P)	0.394	0.381	0.013		0.345	0.049
	[0.489]	[0.486]	(0.012)		[0.001]	(0.012)
Age (W)	31.638	32.388	-0.750		28.713	2.926
	[4.015]	[4.383]	(0.111)		[0.008]	(0.113)
Age (P)	34.675	35.461	-0.786		31.686	2.989
	[5.513]	[5.996]	(0.152)		[0.009]	(0.139)
Observations	1,714	13,809			376,152	
Joint p-val.			0.000	0.928		0.000

Note: Labor market outcomes measured in the year preceding first ACP. (W) - woman, (P) - partner, cond. dif. - conditional difference, rep. - representative, suc. - success. Last column uses inverse probability weights for the first ACP that follow the main specification. Standard deviations in brackets. Standard errors in parentheses.

more hours per year, were nearly 2 percentage points more likely to be employed, and were slightly more educated. A similar education gradient in IVF success has been documented in Denmark (Groes et al., 2024). Partner characteristics follow a similar pattern. Notably, both women and their partners whose first ACP succeeded were nearly nine months younger, consistent with age being the key factor in ACP success.

Following Lundborg et al. (2024), the last column in Table 1 reports differences between the two groups after adjusting for education and age, making remaining gaps negligible. Excluding education has no effect. Appendix A4 presents equivalent results for subsequent ACPs. Since women have limited control over ACP outcomes, the main threat to conditional sequential unconfoundedness is ACP success depending

on health factors that also influence labor market outcomes. As such factors likely also affect pre-ACP outcomes, balance on these outcomes provides relative strong support for the assumption. Remaining analyses accounts for differences in success rate by age at the time of each procedure, education, and procedure type.

Table 1 also compares the main sample to the representative sample of mothers, weighted to match the birth year distribution of first children among those whose first ACP succeeded (column 5). Before motherhood, women in the representative sample were less likely to work, had lower income, worked fewer hours, and were less educated. Differences for fathers follow a similar pattern. Notably, the relative gender gaps in work hours and income are remarkably similar between the two samples, which is crucial for the generalizability of my estimates. The largest difference between samples is age, with first ACP mothers and fathers conceiving three years later on average. This is partly mechanical, as Dutch couples, like those in most countries, are required to try conceiving naturally for at least a year before seeking medical assistance, and intrauterine insemination is not initiated immediately. After parenthood, both groups have similar completed fertility (1.8 children on average). Women whose first ACP succeeded are more likely to have twins (7% vs 1.5% in the representative sample), though multiple births remain uncommon in both groups.

Another reason sequential unconfoundedness may fail is that women pursue additional ACPs based on information suggesting a higher likelihood of success, introducing a correlation between the willingness to undergo ACPs and procedure outcomes. A similar concern applies to the IV approach, which assumes complier status is independent of the first ACP success, though compliance may depend on the success of subsequent procedures. To investigate this, I examine how success rates vary across procedures. If women willing to pursue more ACPs are more likely to succeed, we might expect higher success rates at later procedures. Since success likelihood declines with age, potentially obscuring any pattern, I fix age at the first-procedure average. Figure 1 shows that intrauterine insemination success rates are similar at the first procedure, which includes the full sample, and later procedures, which only include women willing to undergo additional procedures. This suggests a limited relationship between the willingness to undergo ACPs and success likelihood.

Women whose first ACP fails undergo an average of 4.1 additional procedures.

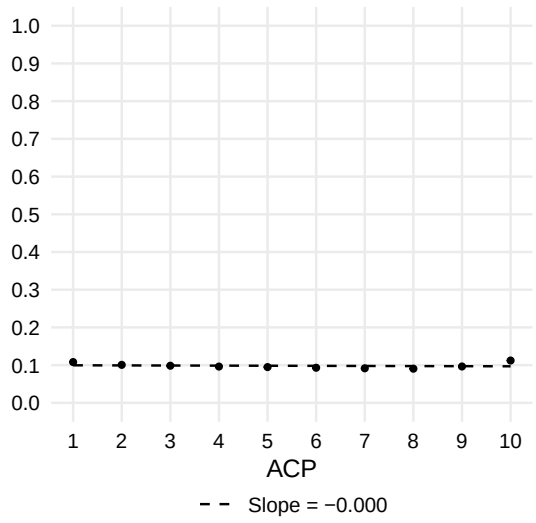

Figure 1: Estimated Success Probabilities

The estimated average willingness to undergo ACPs in case of failure, W, is 7.3. Three years after the first ACP, the estimated relier share is 0.8, decreasing to 0.45 in year seven. For compliers, these shares are 0.4 and 0.25, respectively. The estimated correlation between reliance and willingness is close to zero (although it is not required for the bounding approach). Appendix A4 presents further details.

5 Results

Figure 2 presents effects on women's annual work hours and income. In the conception year, bounds indicate a reduction in work hours of 10 to 130 hours, or 1% to 11% relative to the point-identified relier average control outcome. The impact on income is negligible. Wider bounds for hours in the second year reflect uncertainty due to potential maternity leave. From year 3 to 7, bounds for work hours remain stable, indicating reductions of 90 to 290 hours, or 8% to 26%. Bounds for income widen slightly over time, with reductions of 1,500 to 10,800 euros, or 5% to 34%, in year 7.

Figure 3 presents effects on men's outcomes. The bounds are similar in width to those for women but are centered near zero. Seven years into parenthood, the bounds rule reductions in work hours over 4% and reductions in income over 16%.

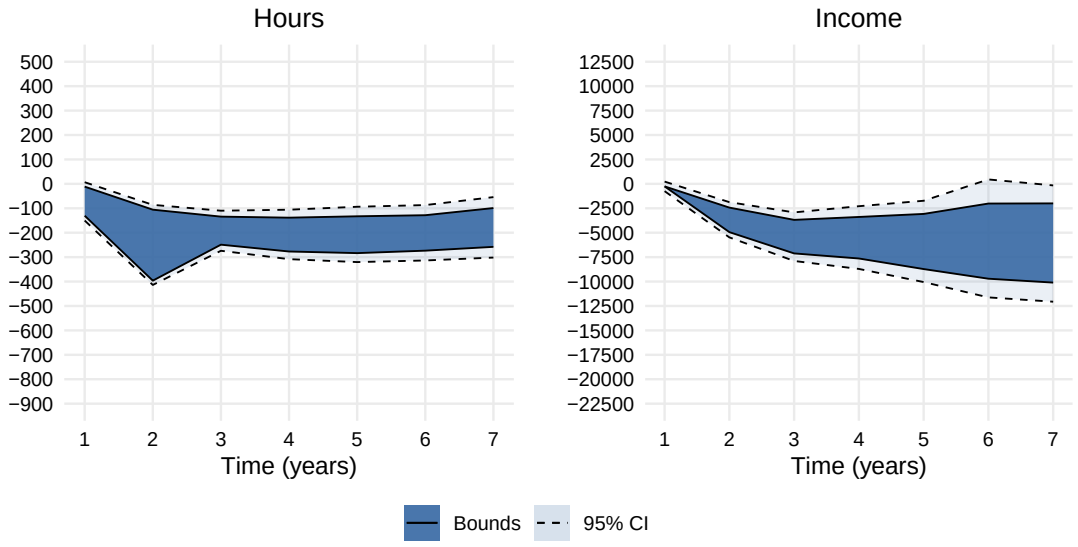

Figure 2: Effects on Women

Figure 4 plots the share of gender inequality caused by parenthood—that is, the effects on the gaps in outcomes between men and women relative to the gaps in their average treated outcomes.[19] Between years 3 and 7, parenthood caused 26–60% of gender inequality in annual work hours and up to 50% in annual income, with upper bounds stable over this period. Aggregating bounds across periods yields non-sharp bounds for cumulative effects, as per-period bounds ignore within-woman and within-couple outcome relationships over time. Using cumulative hours and income over the seven years as outcomes, I find that parenthood caused 36–54% of inequality in work hours and 5–46% of inequality in income during this period.

I present extensions and sensitivity analyses in Appendix A3. They include using an alternative monotonicity assumption following Semenova (2023), applying a GMM estimator that does not rely on orthogonalization and sample splitting, adjusting for the age difference between partners when estimating the share of gender inequality caused by parenthood, and ensuring estimates in each period cover the same subpopulation. All results remain similar. Appendix A3 also presents bounds that do

[19]Using differences between male and female outcomes ensures sharpness (Semenova, 2023). The ratios are calculated as $1 - a/b$, where a is the identified control outcome and b is the lower or upper bound for the treated outcome, estimated using orthogonal moments. Inference based on Delta method.

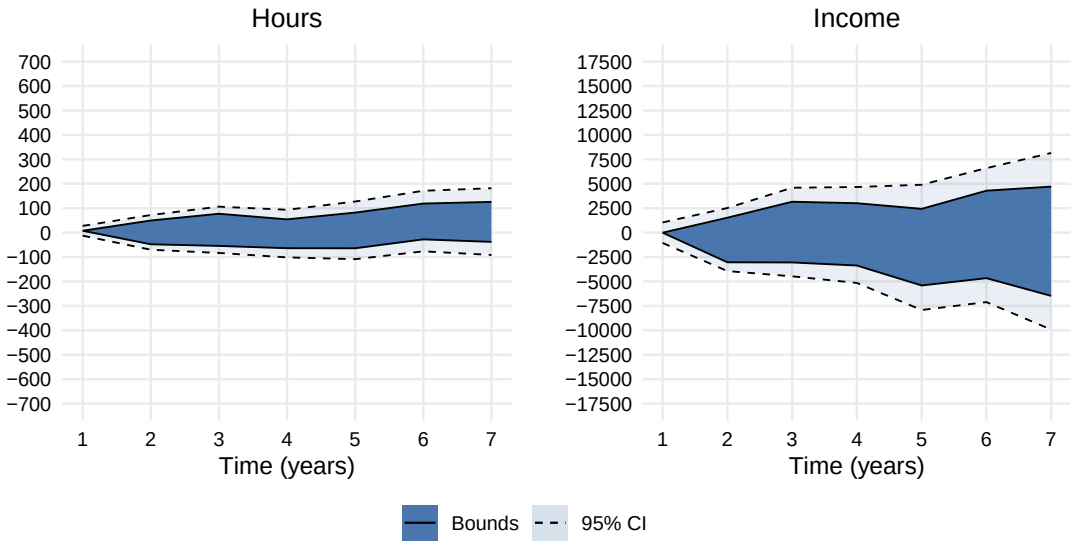

Figure 3: Effects on Men

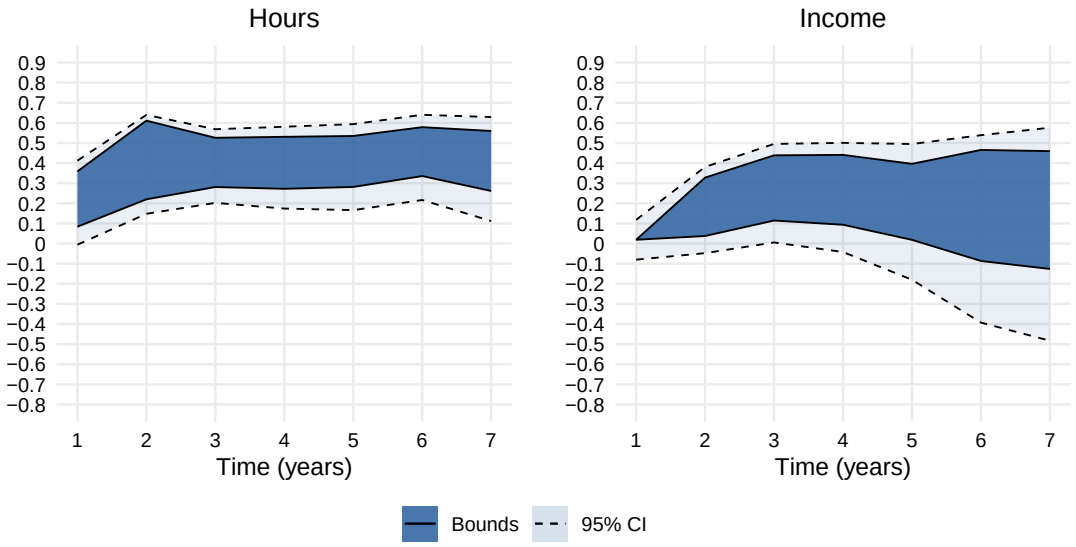

Figure 4: Share of Gender Inequality Caused by Parenthood

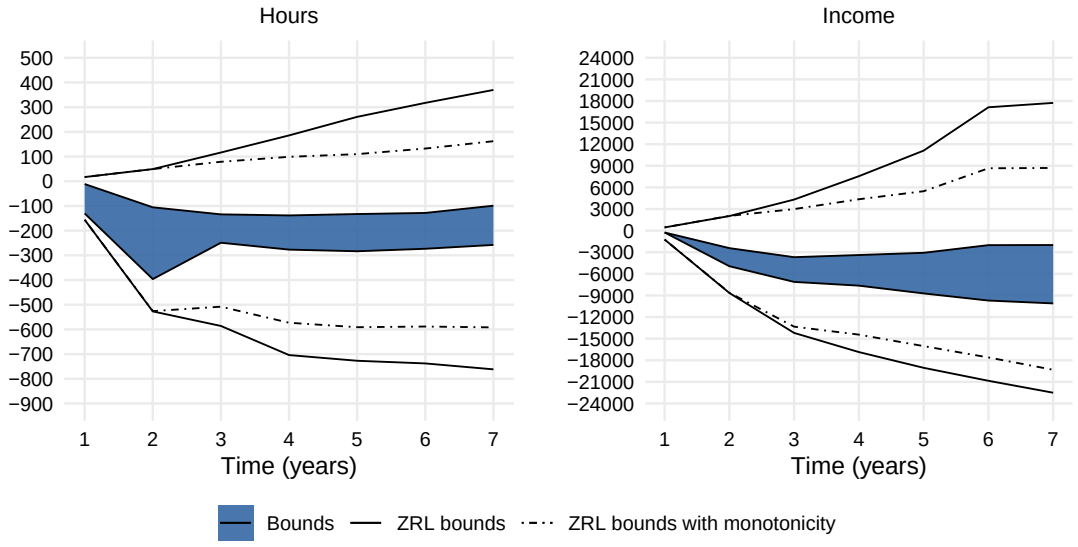

Figure 5: Comparison with ZRL Bounds for Effects on Women

not use the monotonicity assumption. They are similar to the baseline bounds in the first three years but widen thereafter. Under the most extreme selection, the share of gender inequality caused by parenthood over the sample period increases to 64% for hours and 70% for income.

Before turning to existing methods and generalizability, I compare my bounds with those relying solely on women's first ACP, equivalent to the Zhang & Rubin (2003) and Lee (2009) bounds. Figure 5 presents the effects on women's labor market outcomes.[20] The ZRL bounds are considerably wider, unable to rule out large positive or negative effects on women's income and work hours in both the short and medium run. Even leveraging the new monotonicity assumption, the bounds in year 7 are 5 and 3.5 times wider than mine for hours and income, respectively.

6 Existing Methods and Generalizability

Two reduced-form approaches are central to the debate on the career consequences of parenthood: the IV approach, described in Section 3.1, and the event study (ES) approach, which compares women who are one year away from pregnancy with those

[20]Estimation uses the baseline approach, ignoring all ACPs after the first.

who already have children, controlling for age. Since Danish ES estimates are nearly identical for ACP and general samples but differ substantially from the IV estimates (Lundborg et al., 2024), understanding why these methods diverge is crucial for assessing the generalizability of ACP-based findings.

Section 6.1 discusses why the two methods may yield different results, presents a baseline comparison using Dutch data, and assesses if differences in birth timing can reconcile the discrepancies. Sections 6.2 and 6.3 evaluate the extent of dynamic effects and selection, respectively, as well as their explanatory power. Finally, Section 6.4 addresses concerns related to mental health and relationship breakdowns.

6.1 Baseline Comparison and Birth Timing

Even with the same sample, the bias in the ES and IV estimators cannot be assessed by comparing their estimates to the bounds for several reasons. First, they measure effects for different moments of becoming a parent: the IV and the bounds consider the effect of conceiving at the first ACP, while ES also includes conceptions that follow an initial ACP failure. Second, they target different subpopulations: the IV focuses on compliers, the bounds cover reliers, while ES also includes non-reliers. Third, they consider different childlessness scenarios: while all methods assume no anticipation, the IV and the bounds use women trying but failing to conceive, whereas ES may additionally include those who have not yet attempted conception, introducing potential differences in mental health and relationships.

Figure 6 compares baseline estimates from different methods.[21] The ES estimates using the ACP sample suggest a large negative impact on women's labor market outcomes, while IV estimates indicate a smaller effect. The ES estimates from the representative sample closely match those from the ACP sample, consistent with findings from Denmark. Compared to the bounds, the ES estimates suggest a higher cost of motherhood, while IV estimates generally fall within the bounds but suggest larger reductions in hours and income in the fourth year. In the medium run, the bounds include substantially larger negative effects than those suggested by the IV

[21] Hours refer to leave-adjusted work hours. Implementation details for IV and ES are in Appendix A5, they follow Lundborg et al. (2017) and Kleven et al. (2019).

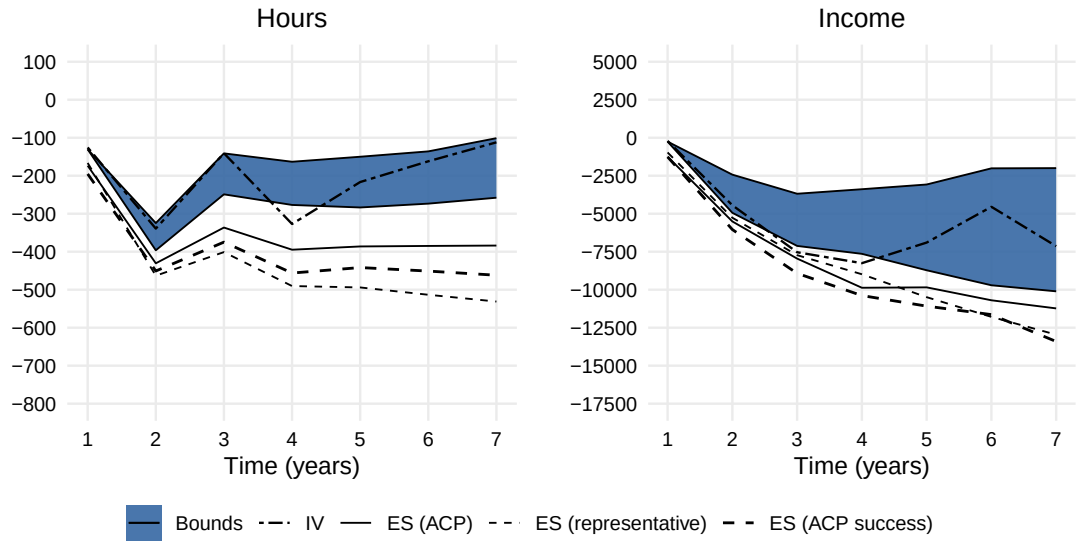

Figure 6: Comparison of Different Methods

estimates.[22]

Figure 6 also presents ES estimates using only women whose first ACP succeeded, which trivially aligns the targeted parenthood moment between the IV, ES, and the bounds to the first ACP. The results remain nearly identical to the baseline ES estimates, suggesting that differences in the moment of becoming a parent do not explain the discrepancy.

6.2 Instrumental Variable and Delayed Parenthood

The IV estimator may be biased when the effect of conceiving at the first ACP differs from that of conceiving later. To evaluate the extent of dynamic effects, I bound the effect of conceiving after ACPs fail relative to conceiving at the first ACP for non-reliers, who are a subset of always-takers driving the potential bias of the IV estimator. This impact is not only relevant for assessing bias but also interesting in its own right, as it sheds light on the career consequences of delaying parenthood.

[22]While my method addresses the bias, it does so at the expense of point identification—a disadvantage to researchers willing to impose stronger assumptions regarding selection or dynamic effects. However, my method compensates with enhanced precision. I discuss this in Appendix A6, where I demonstrate that the 95% confidence intervals for my bounds closely match those for the IV.

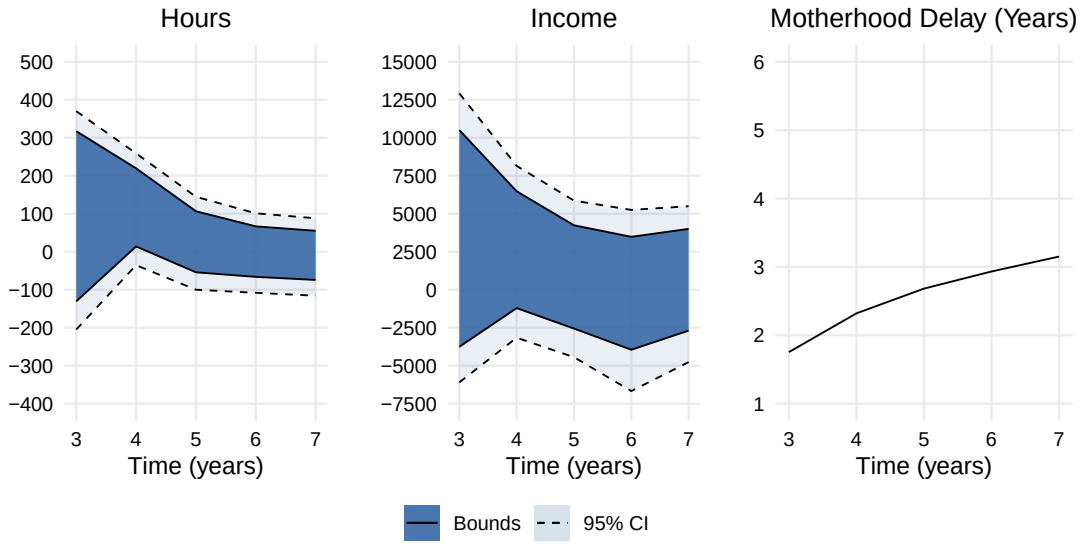

Figure 7: Effects of Delaying Motherhood

This bounding procedure mirrors that for τ_{ATR}.[23]

Figure 7 presents the results. The last column shows the effects on parenthood timing: for women who would conceive within four years regardless of ACP failure, failure delays fertility by an average of 2.3 years, while for those conceiving within seven years, the delay averages 3.1 years. The first two columns report effects on labor market outcomes. Four years after the first ACP, delaying parenthood slightly increases women's work hours. By year seven, the bounds for both income and hours are narrow and centered near zero.

The positive short run effect of delaying parenthood is unexpected. A younger first child typically increases care demands. Thus, IV estimates are often thought to understate the career cost of motherhood (Lundborg et al., 2024), but my findings suggest otherwise. A possible explanation lies in total fertility, as earlier mothers tend

[23]First, I identify the average later-treated outcome for non-reliers with a specific willingness to undergo ACP from women who have their first child without ACP after undergoing a specific number of ACPs: $\mathbb{E}[Y_t \mid A_t = w, Z_A = 0, D_t = 1] = \mathbb{E}[Y_t(R^*) \mid W_t = w, R_t = 0]$, where $Y_t(R^*)$ is the potential outcome for non-reliers in period t in case they have a child after all ACPs fail. Then, I bound the average treated non-relier outcome by trimming the outcome distribution among women whose first ACP succeeded using the identified non-relier share. Monotonicity is leveraged by always including women who have non-ACP children after the first ACP succeeds.

to have more children, meaning that delayed motherhood may reduce care needs and improve labor market outcomes. The data support this explanation—women whose first ACP succeeds have, on average, 0.2 more children than those who conceive after failure. Although small, the positive short-run effect of delaying parenthood is enough to reconcile the gap between the IV estimates and the bounds in Section 6.1.

While the bounds suggest at most a modest medium-run effect of delaying motherhood, the large always-taker share—reaching 75%—may introduce substantial bias in the IV estimator. For example, the IV estimates may understate the career impact of motherhood on work hours and income in year seven by up to 70%. This aligns with the baseline comparisons between the bounds and IV estimates in Section 6.1.

When the effects of parenthood are static, τ_{ATR} is point-identified.[24] This provides a basis for testing homogeneity assumptions leveraged in methods that use short-run IV estimates to address long-run bias (Bensnes et al., 2023; Gallen et al., 2023). If heterogeneity is limited, short-run estimates of τ_{ATR} or τ_{LATE} should yield similar bias-corrected estimates in the long run. Appendix A7 provides the formal argument and empirical results, revealing significant violations of the homogeneity assumption.

6.3 Event Study and Selective Fertility Timing

Next, I quantify the extent of selective fertility among reliers. For brevity, I focus on the intuition and present the formal argument in Appendix A8. The ES approach compares women who have been mothers for t years to those one year away from pregnancy, controlling for age. This captures two factors: the effect of motherhood and differences in career trajectories between those who already have children and those who will later, in the absence of children—selection. If fertility timing is random, selection plays no role, and the comparison isolates the effect of parenthood.

To assess selection, I start with childless reliers career trajectories identified using the baseline approach. Using the timing of women's first ACP as a proxy for fertility

[24] Under static effects, $\mathbb{E}[Y_t(R^*) \mid R_t = 0] = \mathbb{E}[Y_t(1) \mid R_t = 0]$, and since $\mathbb{E}[Y_t(R^*) \mid R_t = 0]$ is identified similarly to $\mathbb{E}[Y_t(0) \mid R_t = 1]$, the relier average treated outcome can be backed out: $\mathbb{E}[Y_t(1) \mid R_t = 1] = \mathbb{E}[Y_t(1)] - \Pr(R_t = 0)\mathbb{E}[Y_t(R^*) \mid R_t = 0]/\Pr(R_t = 1)$. Then, it can be compared to $\mathbb{E}[Y_t(0) \mid R_t = 0]$, identified using the baseline procedure.

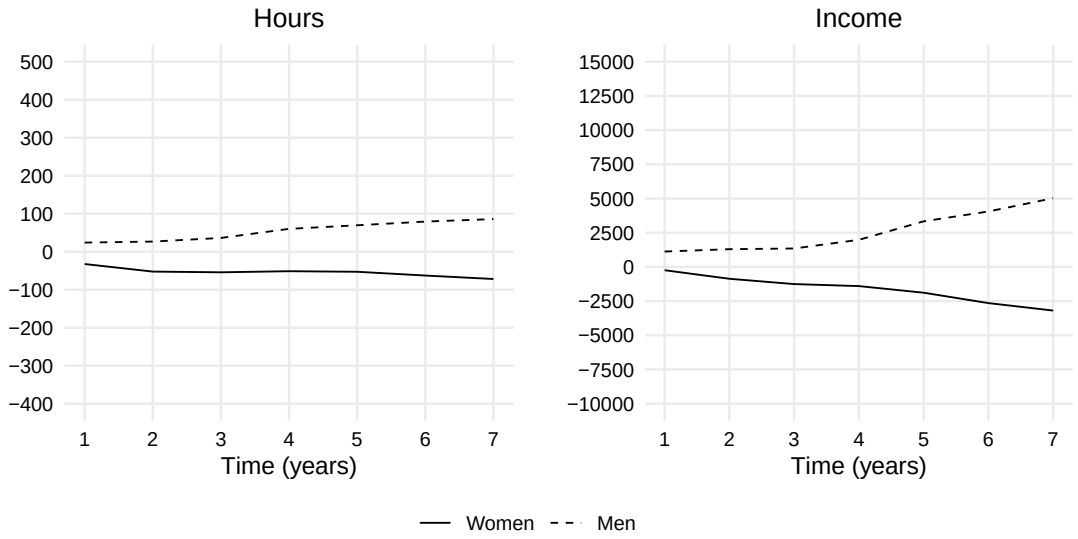

Figure 8: Career Progression Differences by Parenthood Timing in the Absence of Children

decisions, I compare labor market outcomes in the absence of children between similarly aged reliers who planned to have children in the past and those who plan to have children in a year, as in the ES approach. This captures how relier labor market outcomes vary by parenthood timing in the absence of children, which measures the extent of selection in the ES context.[25]

Figure 8 presents separate estimates for men and women. A negative estimate in year t means that women who became mothers t years ago would have had worse labor market outcomes than those about to become mothers, even without children. While initial differences are minor, they grow over time, indicating substantial negative selection among early mothers. For men, selection is reversed: early fathers show better career progression than later fathers, even without children. This positive association between fatherhood and labor market outcomes aligns with evidence of a "fatherhood premium" (Lundberg & Rose, 2000), observed globally and in the Netherlands (Kleven et al., 2024).

The main results in Figure 9 show the extent to which gender inequality among

[25]The implementation involves using a standard event study specification (Kleven et al., 2019), but applied to childless women in the ACP sample, where the event is their first ACP. Women who underwent more ACPs are weighted higher to correct for the underrepresentation of reliers with greater willingness to undergo ACPs.

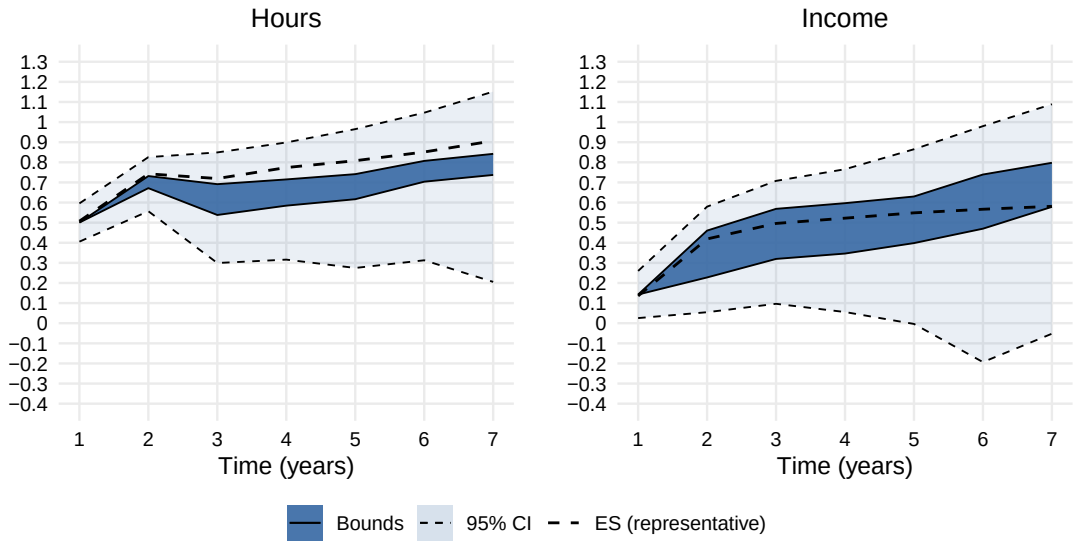

Figure 9: Share of Gender Inequality Explained by Effects of Parenthood and Selection

reliers is jointly explained by the causal effect of parenthood (from Figure 2) and selective fertility timing (from Figure 8). By year 7, these factors account for 70% to 85% of work-hour inequality and 50% to 80% of income inequality.[26]

Figure 9 also presents ES estimates for the share of gender inequality caused by parenthood from the representative sample, which closely follow the bounds. This reconciles conflicting findings on the career impact of parenthood between the ACP sample, based on a procedure that accounts for selection, and estimates from the general population, based on a procedure that does not. It supports extrapolating ACP-based estimates to the general population, suggesting that at least 34% of gender inequality in work hours and 42% of inequality in income associated with parenthood arise from selective fertility timing rather than the causal effect of parenthood.

In my model, these results imply that selective fertility biases the ES estimator. However, my model does not distinguish between failing to conceive and choosing not to have children. These scenarios may differ, as conception failure, especially via ACPs, could affect mental health and relationship stability, worsening labor market outcomes. If substantial, these effects could explain the gap between the bounds or the IV estimates, where the counterfactual is failing to conceive, and the ES estimates,

[26]The share increases further after accounting for age differences between partners.

where the counterfactual may involve not having attempted yet. This difference, rather than bias from selective fertility, may be what the selection ES estimates capture. I address this in the next section.

6.4 Mental Health and Relationship Stability

Women who remain childless after ACP failure may experience negative mental health outcomes and/or relationship breakdowns, which could also affect labor market outcomes. This is important for two reasons. First, it may explain differences between the ES estimates and the IV estimates or the bounds, even without selection. Second, it may undermine the monotonicity assumption used to narrow the bounds. I first discuss the two concerns and then introduce a procedure to address them simultaneously.

The extent to which mental health issues and relationship breakdowns following ACP failure affect the interpretation and comparison of different methods depends on the counterfactual of interest. From an economic perspective, the most relevant counterfactual for parenthood is arguably the scenario in which couples want children but choose not to have them.

If mental health issues and relationship breakdowns stem from unmet fertility goals, they are not a concern for the IV and the bounds, as these factors act as mechanisms through which parenthood influences labor market outcomes. Meanwhile, ES estimates, which compare mothers to women who may not yet want children, could miss these effects, leading to biased estimates relative to the counterfactual of interest.[27]

By contrast, if these issues arise from failed conceptions or ACP side effects rather than the absence of children, they pose a concern for the IV and the bounds. Couples would not have experienced these issues had they chosen not to undergo ACPs, making the ES comparison group—women who have not yet undergone these procedures—more relevant. This concern has been raised in studies using IVF due to its invasive nature (Bögl et al., 2024). While focusing on IUI, a less invasive procedure,

[27]Nonetheless, extrapolation to non-ACP families remains a concern, as this impact may be stronger in ACP families due to their high desire for children. Additionally, breakups complicate the interpretation of within-couple inequality measures.

mitigates these concerns, they remain only partially addressed, as about one-third of women whose first IUI fails eventually undergo IVF.

Finally, mental health side effects and relationship breakdowns may threaten the monotonicity assumption. Improved mental health and stability after a successful conception may increase natural conception attempts, leading to non-ACP births that would not have occurred if ACPs had failed.

Before formally addressing these concerns, it is worth noting that the empirical relevance of mental health side effects may be limited. Lundborg et al. (2024) use conservative back-of-the-envelope calculations, drawing on research on health shocks and medical evidence on ACP side effects, and conclude that such effects, even if sizable in relative terms, are unlikely to meaningfully influence women's career trajectories in absolute terms. Consistent with this, the estimated effect of ACP failure on antidepressant uptake (Appendix A9) is precise and indistinguishable from zero.[28]

To address concerns about mental health and relationship breakdowns formally, I adapt my approach to bound the effects for reliers who, in the event of ACP failure, would remain with their partners and avoid severe mental health issues, proxied by the onset of antidepressant use. These bounds, mechanically wider than baseline, include the effect of parenthood specifically for women who would not face severe consequences of failing to conceive. If they remain comparable to baseline bounds, this suggests that mental health and relationship breakdowns following failed conception have limited impact on the estimates.

Formally, let $S_t = 1$ if, in period t, in the scenario where ACPs fail, a woman would not uptake antidepressants and would remain cohabiting with her partner from the time of the first ACP; otherwise, $S_t = 0$. I refer to this group as *resilient* ($S_t = 1$). I bound the average treatment effect for resilient reliers $\mathbb{E}[\tau(t) \mid R_t = 1, S_t = 1]$. The procedure accommodates a relaxed monotonicity assumption:

Assumption 6 (Partial monotonicity). $\Pr(R_t^+ \geq R_t \mid S_t = 1) = 1$.

It states that monotonicity holds for resilient women, allowing for monotonicity violations among women who uptake antidepressants or separate from their partner

[28] This is only suggestive: the counterfactual is having children instead of not undergoing ACPs.

following ACP failure. I discuss empirical support for the original and partial monotonicity assumptions in Appendix A10. Specifically, I demonstrate that the estimated subsequent relier share is at least as large as the relier share, as implied by monotonicity. The formal identification result for the resilient relier bounds amounts to treating childless women who uptake antidepressants or separate from their original partner as having children, or being non-reliers.[29]

Figure 10 presents the results. In the first few years, the bounds are similar to those from the baseline approach but widen in later years.[30] Nonetheless, although this procedure for addressing mental health and relationship side effects is conservative, its impact on the results is minimal: even in the most extreme scenario, the share of gender inequality in work hours and income attributed to parenthood increases by no more than 10 percentage points.[31]

The key takeaway from these results is that severe mental health issues and relationship breakdowns, regardless of their source, do no explain the differences in estimates across methods. Since differences in fertility timing also fail to account for the discrepancies, these findings suggest that bias from dynamic effects and selective fertility is the key driver of differences between the IV and ES estimates, and the bounds. Since ES estimates are nearly identical for ACP and representative samples, these results support the generalizability of my estimates to the general population. Moreover, as these results indicate that my estimates are not driven by families facing severe consequences from unsuccessful conception, they support generalizing to families with a weaker desire for children, where such effects may be less pronounced.

[29]Setting D^+ to 1 when a woman has no children but either takes antidepressants or separates from her partner, the proof follows the same steps as for the theorem.

[30]Until year four, over 90% of reliers are resilient, declining to 85% by year seven.

[31]The procedure I use to address mental health and relationship stability concerns is conservative for two reasons. First, it excludes from the comparison women who would experience poor mental health or relationship breakdowns regardless of fertility or attempts to conceive. This makes the bounds wider than if such women were included. Second, it tackles both mental health and relationship breakdowns simultaneously. The bounds presented in Appendix A10, which address mental health separately, are even closer to the baseline estimates.

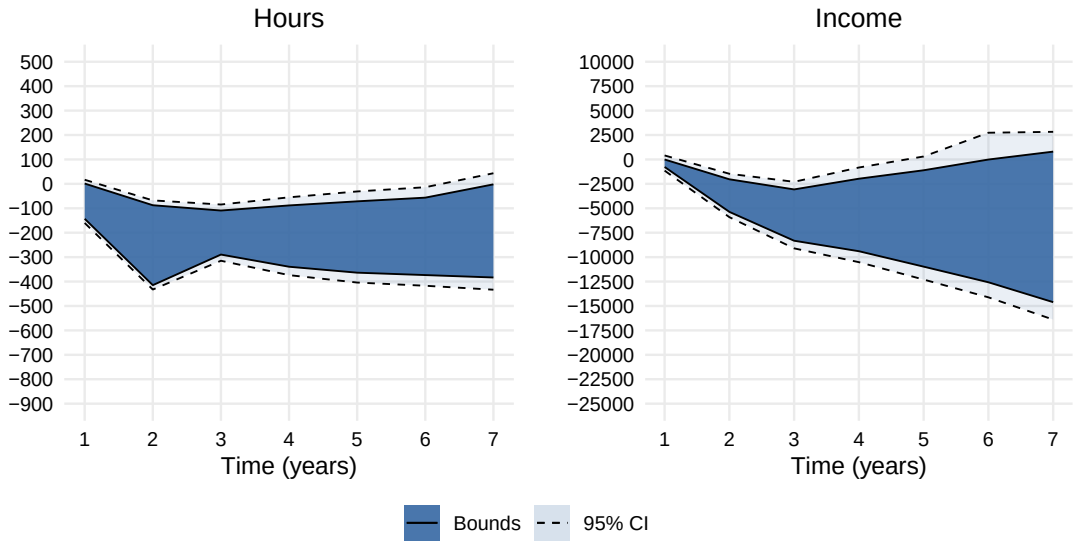

Figure 10: Effects on Resilient Women

7 Conclusion

Parenthood explains much of the gender inequality in Western labor markets (Kleven et al., 2024), but identifying its causal effects remains challenging. This paper introduces a method that leverages assisted conception procedures to bound the effects of parenthood while accounting for selective fertility timing and dynamic effects. The procedure can also be applied to other quasi-experimental settings where individuals are assigned to treatment or control but may transition between states through repeated assignments or entirely selective pathways.

Using data from the Netherlands, I find that parenthood persistently reduces women's yearly work hours by 10%-25% and income by 9%–29%. However, at least half of the within-couple gender inequality in these outcomes after childbirth is not caused by parenthood. I also provide evidence that event study estimates may overstate the impact due to selective fertility timing, and IV estimates may understate it due to dynamic effects. While event study and IV estimates may differ for several reasons, some limiting the generalizability of estimates using assisted conception success, I show that this bias is key to reconciling the conflicting results, with other factors playing a minor role. These findings support generalizing my results from the assisted conception sample to the broader population.

My results have implications for understanding gender inequality in the labor market, identifying potential remedies, and guiding future research. While leading narratives either attribute nearly all gender inequality to parenthood or suggest it has little impact, my findings offer a more nuanced perspective: parenthood has substantial effects but is not the sole driver of gender inequality. From a research perspective, this underscores the importance of considering additional factors contributing to gender inequality. From a policy perspective, my results help explain why family-friendly policies may have limited direct effects on gender inequality. However, such policies could still reduce inequality by shaping behavior in anticipation of parenthood, which is not captured by the methods considered in this paper.

References

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

Agüero, J. M., & Marks, M. S. (2008). Motherhood and female labor force participation: evidence from infertility shocks. *American Economic Review*, *98*(2), 500–504.

Angelov, N., Johansson, P., & Lindahl, E. (2016). Parenthood and the gender gap in pay. *Journal of Labor Economics*, *34*(3), 545–579.

Angrist, J., & Evans, W. N. (1996). Children and their parents' labor supply: Evidence from exogenous variation in family size. *National Bureau of Economic Research*.

Angrist, J., Ferman, B., Gao, C., Hull, P., Tecchio, O. L., & Yeh, R. W. (2024). Instrumental variables with time-varying exposure: New estimates of revascularization effects on quality of life. *National Bureau of Economic Research*.

Angrist, J., & Imbens, G. (1995). Identification and estimation of local average treatment effects. *National Bureau of Economic Research*.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *arXiv preprint arXiv:1610.01271*.

Bensnes, S., Huitfeldt, I., & Leuven, E. (2023). Reconciling estimates of the long-term earnings effect of fertility. *IZA Discussion Paper*.

Bertrand, M. (2011). New perspectives on gender. *Handbook of Labor Economics*, *4*, 1543–1590.

Bertrand, M. (2020). Gender in the twenty-first century. *AEA Papers and Proceedings*, *110*, 1–24.

Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, *55*(3), 789–865.

Bögl, S., Moshfegh, J., Persson, P., & Polyakova, M. (2024). The economics of infertility: Evidence from reproductive medicine. *National Bureau of Economic Research*.

Bronars, S. G., & Grogger, J. (1994). The economic consequences of unwed motherhood: Using twin births as a natural experiment. *American Economic Review*, *84*(5), 1141–1156.

Brooks, N., & Zohar, T. (2021). Out of labor and into the labor force? The role of abortion access, social stigma, and financial constraints. *CEMFI Working Paper No. 2111*.

Bütikofer, A., Jensen, S., & Salvanes, K. G. (2018). The role of parenthood on the gender gap among top earners. *European Economic Review*, *109*, 103–123.

Chernozhukov, V., Chetverikov, D., & Kato, K. (2019). Inference on causal and structural parameters using many moment inequalities. *The Review of Economic Studies*, *86*(5), 1867–1900.

Chung, Y., Downs, B., Sandler, D. H., Sienkiewicz, R., et al. (2017). The parental gender earnings gap in the United States. *Unpublished manuscript*.

Cortés, P., & Pan, J. (2023). Children and the remaining gender gaps in the labor market. *Journal of Economic Literature*, *61*(4), 1359–1409.

Cristia, J. P. (2008). The effect of a first child on female labor supply: Evidence from women seeking fertility services. *Journal of Human Resources*, *43*(3), 487–510.

Cruces, G., & Galiani, S. (2007). Fertility and female labor supply in Latin America: New causal evidence. *Labour Economics*, *14*(3), 565–573.

Eichmeyer, S., & Kent, C. (2022). Parenthood in poverty. *Centre for Economic Policy Research*.

Ferman, B., & Tecchio, O. (2023). Identifying dynamic lates with a static instrument. *arXiv preprint arXiv:2305.18114*.

Fitzenberger, B., Sommerfeld, K., & Steffes, S. (2013). Causal effects on employment after first birth—a dynamic treatment approach. *Labour Economics*, *25*, 49–62.

Gallen, Y., Joensen, J. S., Johansen, E. R., & Veramendi, G. F. (2023). The labor market returns to delaying pregnancy. *Available at SSRN 4554407*.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, *104*(4), 1091–1119.

Groes, F., Houštecká, A., Iorio, D., & Santaeulàlia-Llopis, R. (2024). The unequal battle against infertility: Theory and evidence from IVF success. *Centre for Economic Policy Research*.

Heiler, P. (2024). Heterogeneous treatment effect bounds under sample selection with an application to the effects of social media on political polarization. *Journal of Econometrics*, *244*(1), 105856.

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC.

Hirvonen, L. (2009). The effect of children on earnings using exogenous variation in family size: Swedish evidence. *Swedish Institute for Social Research Working Paper*.

Horowitz, J. L., & Manski, C. F. (1995). Identification and robustness with contaminated and corrupted data. *Econometrica*, 281–302.

Horowitz, J. L., & Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, *95*(449), 77–84.

Hotz, V. J., McElroy, S. W., & Sanders, S. G. (2005). Teenage childbearing and its life cycle consequences: Exploiting a natural experiment. *Journal of Human Resources*, *40*(3), 683–715.

Iacovou, M. (2001). Fertility and female labour supply. *ISER Working Paper Series*.

Jacobsen, J. P., Pearce III, J. W., & Rosenbloom, J. L. (1999). The effects of childbearing on married women's labor supply and earnings: Using twin births as a natural experiment. *Journal of Human Resources*, 449–474.

Kleven, H., Landais, C., & Leite-Mariante, G. (2024). The child penalty atlas. *The Review of Economic Studies*, rdae104.

Kleven, H., Landais, C., & Søgaard, J. E. (2019). Children and gender inequality: Evidence from Denmark. *American Economic Journal: Applied Economics*, *11*(4), 181–209.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, *76*(3), 1071–1102.

Lundberg, S., & Rose, E. (2000). Parenthood and the earnings of married men and women. *Labour Economics*, *7*(6), 689–710.

Lundborg, P., Plug, E., & Rasmussen, A. W. (2017). Can women have children and a career? IV evidence from IVF treatments. *American Economic Review*, *107*(6), 1611–37.

Lundborg, P., Plug, E., & Rasmussen, A. W. (2024). Is there really a child penalty in the long run? New evidence from IVF treatments. *IZA Discussion Paper*.

Manski, C. F. (1989). Anatomy of the selection problem. *Journal of Human resources*, 343–360.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review*, *80*(2), 319–323.

Maurin, E., & Moschion, J. (2009). The social multiplier and labor market participation of mothers. *American Economic Journal: Applied Economics*, *1*(1), 251–272.

Melentyeva, V., & Riedel, L. (2023). Child penalty estimation and mothers' age at first birth. *ECONtribute Discussion Paper*.

Miller, A. R. (2011). The effects of motherhood timing on career path. *Journal of Population Economics*, *24*, 1071–1100.

OECD. (2022). *Out-of-school-hours services*.

OECD. (2023a). *Enrolment in childcare and pre-school*.

OECD. (2023b). *OECD employment database*. Retrieved from https://stats.oecd.org/Index.aspx?DatasetCode=AVE_HRS

OECD. (2023c). *Parental leave system*.

OECD. (2023d). *Part-time employment rate (indicator)*. Retrieved from https://www.oecd.org/en/data/indicators/part-time-employment-rate.html

Olivetti, C., Pan, J., & Petrongolo, B. (2024). The evolution of gender in the labor market. *Handbook of Labor Economics*, *5*, 619–677.

Olma, T. (2021). Nonparametric estimation of truncated conditional expectation functions. *arXiv preprint arXiv:2109.06150*.

Rosenzweig, M. R., & Wolpin, K. I. (1980). Life-cycle labor supply and fertility: Causal inferences from household models. *Journal of Political Economy*, *88*(2), 328–348.

Roth, J., Sant'Anna, P. H., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, *235*(2), 2218–2244.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.

Semenova, V. (2023). Generalized Lee bounds. *arXiv preprint arXiv:2008.12720v3*.

Stoye, J. (2020). A simple, short, but never-empty confidence interval for partially identified parameters. *arXiv preprint arXiv:2010.10484*.

Van den Berg, G. J., & Vikström, J. (2022). Long-run effects of dynamically assigned treatments: A new methodology and an evaluation of training effects on earnings. *Econometrica*, *90*(3), 1337–1354.

Vere, J. P. (2011). Fertility and parents' labour supply: new evidence from US census data: Winner of the OEP prize for best paper on women and work. *Oxford Economic Papers*, *63*(2), 211–231.

Zhang, J. L., & Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, *28*(4), 353–368.

A1 Proof of Theorem

Corollary. *Under Assumption 4:*

$$(Y(1), Y(0), R^+, R, W, Z_1, \ldots, Z_{j-1}, X_1, \ldots, X_{j-1}) \perp\!\!\!\perp Z_j | X_j \text{ for all } j > 1.$$

Proof of Corollary. X_j includes $1_{\{A \geq j\}}$, and when $A < j$, we have $Z_j = 0$, which covers the cases when $X_j \in \mathcal{X}_j \setminus \mathcal{X}_j^1$. The remainder follows from Assumption 4, since given $X_j \in \mathcal{X}_j$, $1_{\{A \geq j\}}$, $Z_1, \ldots, Z_{j-1}$, and $X_1, \ldots, X_{j-1}$ are known. □

Lemma. *For any for l s.t. $1 \leq l \leq \overline{w}$ and any measurable function $g(M_l)$, where $M_l = (Y(1), Y(0), R^+, R, W, Z_1, \ldots, Z_l, X_1, \ldots, X_l)$, under Assumptions 4 and 5:* $\mathbb{E}\left[g(M_l)\Pi_{j=l+1}^{\overline{w}} \frac{(1-Z_j)}{(1-e_j(X_j))}\middle| X_l\right] = \mathbb{E}\left[g(M_l)|X_l\right]$.

Proof of Lemma. Assumption 5 ensures that for any j, $1 - e_j(x_j) > 0$ for all $x_j \in \mathcal{X}_j$. Then, w.l.o.g. for some l s.t. $l < \overline{w}$:

$$\mathbb{E}\left[g(M_l)\Pi_{j=l+1}^{\overline{w}} \frac{(1-Z_j)}{(1-e_j(X_j))}\middle| X_l\right] = \mathbb{E}\left[\mathbb{E}\left[g(M_l)\Pi_{j=l+1}^{\overline{w}} \frac{(1-Z_j)}{(1-e_j(X_j))}\middle| X_{\overline{w}}\right]\middle| X_l\right] \tag{1}$$

$$= \mathbb{E}\left[g(M_l)\Pi_{j=l+1}^{\overline{w}-1} \frac{(1-Z_j)}{(1-e_j(X_j))}\,\mathbb{E}\left[\frac{1-Z_{\overline{w}}}{1-e_{\overline{w}}(X_{\overline{w}})}\middle| X_{\overline{w}}\right]\middle| X_l\right] \tag{2}$$

$$= \mathbb{E}\left[g(M_l)\Pi_{j=l+1}^{\overline{w}-1} \frac{(1-Z_j)}{(1-e_j(X_j))}\middle| X_l\right] \tag{3}$$

$$= \mathbb{E}\left[g(M_l)|X_l\right], \tag{4}$$

where (1) holds by law of iterated expectations and because X_j includes X_l for $j \geq l$, (2) holds by the Corollary, (3) holds because: $\mathbb{E}\left[\frac{1-Z_{\overline{w}}}{1-e_{\overline{w}}(X_{\overline{w}})}\middle| X_{\overline{w}}\right] = 1$, and where (4) follows from steps similar to (1) through (3) for X_j for j s.t. $l < j < \overline{w}$. □

Proof of theorem. I demonstrate the result for the upper bound, the result for the lower bound is symmetric. First, I demonstrate that $\mathbb{E}\left[Y(1-D^+)\Pi_{j=1}^{\overline{w}} \frac{(1-Z_j)}{(1-e_j(X_j))}\right] / \mathbb{E}[r(X_1)] = \mathbb{E}[Y(0)|R=1]$. Using that $D^+ = D|Z_A = 0$, $1 - D = R|Z_A = 0$, and $Y = Y(0)|D = 0$ (by Assumption 2):

$$\mathbb{E}\left[Y(1-D^+)\Pi_{j=1}^{\overline{w}} \frac{(1-Z_j)}{(1-e_j(X_j))}\right] = \mathbb{E}\left[\mathbb{E}\left[Y(0)R\Pi_{j=1}^{\overline{w}} \frac{(1-Z_j)}{(1-e_j(X_j))}\middle| X_1\right]\right] \tag{5}$$

$$= \mathbb{E}\left[\mathbb{E}\left[Y(0)R(1-Z_1)/(1-e_1(X_1))|X_1\right]\right] \tag{6}$$

$$= \mathbb{E}\left[\mathbb{E}\left[Y(0)R|X_1\right]\mathbb{E}\left[(1-Z_1)/(1-e_1(X_1))|X_1\right]\right] \tag{7}$$

$$= \mathbb{E}\left[Y(0)R|X_1\right] \tag{8}$$

$$= \mathbb{E}\left[Y(0)|R=1\right]\Pr(R=1), \tag{9}$$

where (5) holds by law of iterated expectations, (6) holds by Lemma, and (7) and (8)

hold by Assumption 4. Moreover since $1-D+ = R|Z_A = 0$ and $1-Z_A = \Pi_{j=1}^{\overline{w}}(1-Z_j)$:

$$\mathbb{E}\left[(1-D^+)\Pi_{j=1}^{\overline{w}}\frac{(1-Z_j)}{(1-e_j(X_j))}\middle|X_1\right] = \mathbb{E}\left[R\Pi_{j=1}^{\overline{w}}\frac{(1-Z_j)}{(1-e_j(X_j))}\middle|X_1\right] \tag{10}$$

$$= \mathbb{E}\left[R(1-Z_1)/(1-e_1(X_1))|X_1\right] \tag{11}$$

$$= \Pr(R=1|X_1), \tag{12}$$

where (11) holds by Lemma and (12) holds by Assumption 4. Since $\mathbb{E}[\Pr(R=1|X_1 = x)] = \Pr(R=1)$, the result holds.

Remains to show that $\mathbb{E}\left[Y(1-D^+)1_{\{Y>q(1-p(X_1),X_1)\}}\frac{Z_1}{e_1(X_1)}\right]/\mathbb{E}[r(X_1)]$ is a sharp upper bound for $\mathbb{E}[Y(1)|R=1]$. I first demonstrate that $p(x) = \Pr(R=1|D^+ = 0, Z_1 = 1, X_1 = x)$. Assumption 4 together with $D^+ = 1 - R^+|Z_1 = 1$ implies that $r^+(x) = \Pr(R^+ = 1|X_1 = x)$. Under Assumption 3, $\Pr(R=1|X_1 = x) = \Pr(R=1, R^+ = 1|X_1 = x)$. Applying the definition of conditional probability gives $p(x) = \Pr(R=1|R^+ = 1, X_1 = x)$. Assumption 4 together with $D^+ = 1 - R^+|Z_1 = 1$ gives $\Pr(R=1|D^+ = 0, Z_1 = 1, X_1 = x) = \Pr(R=1|R^+ = 1, X_1 = x)$, which implies the result.

The remainder of the proof is similar to Lee (2009). Let $\gamma_x = \mathbb{E}[Y|Z_1 = 1, D^+ = 0, Y \geq q(1-p(x), x), X_1 = x]$. I next demonstrate that γ_x is a sharp upper bound for $\mathbb{E}[Y(1)|X_1 = x, R=1]$. Using that $p(x) = \Pr(R=1|D^+ = 0, Z_1 = 1, X_1 = x)$, Corollary 4.1 in Horowitz & Manski (1995) gives, $\gamma_x \geq \mathbb{E}[Y|Z_1 = 1, D^+ = 0, R = 1, X_1 = x]$. Using that $D^+ = 0|R=1$ and $Y = Y(1)|Z_1 = 1$ and by Assumption 4, $\mathbb{E}[Y|Z_1 = 1, D^+ = 0, R = 1, X_1 = x] = \mathbb{E}[Y(1)|X_1 = x, R=1]$, meaning that γ_x is an upper bound for $\mathbb{E}[Y(1)|X_1 = x, R=1]$. Since $p(x)$ is identified and $Y(1)$ is observed only among those whose first ACP succeeded (because $\Pr(D_1 = 1|Z_1 = 0) = 0$) Corollary 4.1 in Horowitz & Manski (1995) implies sharpness.

Let $f_{x|R=1}(x)$ be the p.d.f. of X_1 conditional on $R=1$. Applying Bayes rule for densities to $\Pr(R=1|X_1 = x)$ identified by $r(x)$ and p.d.f. of X_1 identified directly identifies $f_{x|R=1}(x)$, making $\int_{\mathcal{X}_1}\gamma_x f_{x|R=1}(x)dx$ the sharp upper bound for $\mathbb{E}[Y(1)|R=1]$.

The last step is to show that:

$$\int_{\mathcal{X}_1} \gamma_x f_{x|R=1}(x)dx = \mathbb{E}\left[Y(1-D^+)1_{\{Y>q(1-p(X_1),X_1)\}}\frac{Z_1}{e_1(X_1)}\right] / \mathbb{E}[r(X_1)].$$

By the law of iterated expectations:

$$\begin{aligned}&\mathbb{E}\left[Y(1-D^+)1_{\{Y>q(1-p(X_1),X_1)\}}\frac{Z_1}{e_1(X_1)}\right]\\&= \mathbb{E}\left[\frac{1}{e_1(X_1)}\,\mathbb{E}[Y(1-D^+)1_{\{Y>q(1-p(X_1),X_1)\}}Z_1|X_1]\right].\end{aligned}$$

Applying the definition of conditional probability:

$$\begin{aligned}&\mathbb{E}[Y(1-D^+)1_{\{Y>q(1-p(X_1),X_1)\}}Z_1|X_1] =\\&\qquad\qquad \mathbb{E}[\gamma_{X_1}|X_1]\Pr(D^+=0,Z_1=1,Y>q(1-p(X_1),X_1)|X_1).\end{aligned}$$

Applying the definition of conditional probability twice:

$$\begin{aligned}&\Pr(D^+=0,Z_1=1,Y>q(1-p(X_1),X_1)) =\\&\Pr(Y>q(1-p(X_1),X_1)|D^+=0,Z_1=1,X_1)\Pr(D^+=0|Z_1=1,X_1)\Pr(Z_1=1|X_1).\end{aligned}$$

Using the definitions of $p(X_1)$, $r^+(X_1)$, and $e_1(X_1)$, the term on the right-hand side is $p(X_1)r^+(X_1)e_1(X_1)$, and from definition of $p(X_1)$ it simplifies to $r(X_1)e_1(X_1)$, giving:

$$\begin{aligned}\mathbb{E}\left[Y(1-D^+)1_{\{Y>q(1-p(X_1),X_1)\}}\frac{Z_1}{e_1(X_1)}\right] &= \mathbb{E}\left[\frac{1}{e_1(X_1)}\,\mathbb{E}[\gamma_{X_1}|X_1]r(X_1)e_1(X_1)\right]\\&= \mathbb{E}[\gamma_{X_1}r(X_1)].\end{aligned}$$

Applying Bayes rule for densities: $\mathbb{E}[\gamma_{X_1}r(X_1)] = \int_{\mathcal{X}_1}\gamma_x f_{x|R=1}(x)dx\Pr(R=1)$. Since $\mathbb{E}[r(X_1)] = \Pr(R=1)$, the statement holds.

□

A2 Estimating the Bounds

Section A2.1 presents the estimator. Section A2.2 demonstrates orthogonality. Section A2.3 describes the implementation.

A2.1 Estimator

The method of Semenova (2023) for the ZRL procedure consists of two key components: orthogonalization and sample splitting. Orthogonalization modifies the baseline moments by adding terms that preserve their expectations at the true nuisance parameter while eliminating sensitivity to small estimation errors in the nuisance parameter. Sample splitting means that the nuisance parameter for each observation is estimated without using that observation. These components together allow inference using standard methods, ensuring the estimation of the nuisance parameter does not affect the asymptotic distribution of the averaged moments for a broad class of nonparametric estimators.

I modify the moments in Semenova (2023) to incorporate the first step of my identification approach. Specifically, I replace the terms for the complier control outcome and share with corresponding terms for the relier control outcome and share. These modifications align with the way the two parameters are identified in the Theorem, using outcomes and parenthood indicators among women who never experience ACP success, weighted by the number of ACPs and the propensity score at each ACP, $e_j(X_j)$. Since this makes the moments sensitive to the propensity scores, I include additional terms to correct for this sensitivity. These terms account for the role of the scores in estimating the relier average control outcome and the relier share.

The moment functions are given in Table A1. The new moments identify the same parameters as the baseline moments:

$$\mathbb{E}[\psi^{L+}(G,\xi^0)] = \mathbb{E}[m^L(G,\eta^0)],\ \mathbb{E}[\psi^{U+}(G,\xi^0)] = \mathbb{E}[m^U(G,\eta^0)].$$

However, the original moments are sensitive to small errors in the nuisance parameter, whereas the new moments are not. For example, for some j, let $\widehat{e}_j(x_j)$ be an estimate of the propensity score $e_j(x_j)$ such that $\widehat{e}_j(x_j) \neq e_j(x_j)$ for $x_j \in \mathcal{X}_j^1$. Define $r \in$

Table A1: Orthogonal Moment Functions

	Moment functions
$\psi^{L+}(G,\xi^0)$	$Y(1-D^+)1_{\{Y<q(p(X_1),X_1)\}}\frac{Z_1}{e_1(X_1)} - Y(1-D^+)\Pi_{j=1}^{\overline{w}}\frac{(1-Z_j)}{(1-e_j(X_j))}$ $+q(p(X_1),X_1)\Big[\Pi_{j=1}^{\overline{w}}\frac{(1-Z_j)}{(1-e_j(X_j))}(1-D^+-r_1(X_1))$ $-\frac{Z_1}{e_1(X_1)}p(X_1)(1-D^+-r^+(X_1)) - \frac{Z_1}{e_1(X_1)}(1-D^+)(1_{\{Y<q(p(X_1),X_1)\}}-p(X_1))\Big]$ $-\frac{Z_1-e_1(X_1)}{e_1(X_1)}z^{L+}(1,X_1)r_1(X_1) + \sum_{k=1}^{\overline{w}}1_{\{A\geq k\}}\Pi_{j=1}^{k-1}\frac{(1-Z_j)}{(1-e_j(X_j))}\frac{e_k(X_k)-Z_k}{1-e_k(X_k)}[r_k(X_k)\beta_k(X_k)$ $+q(p(X_1),X_1)(r_1(X_1)-r_k(X_k))]$
$\psi^{U+}(G,\xi^0)$	$Y(1-D^+)1_{\{Y>q(1-p(X_1),X_1)\}}\frac{Z_1}{e_1(X_1)} - Y(1-D^+)\Pi_{j=1}^{\overline{w}}\frac{(1-Z_j)}{(1-e_j(X_j))}$ $+q(1-p(X_1),X_1)\Big[\Pi_{j=1}^{\overline{w}}\frac{(1-Z_j)}{(1-e_j(X_j))}(1-D^+-r_1(X_1))$ $-\frac{Z_1}{e_1(X_1)}p(X_1)(1-D^+-r^+(X_1)) - \frac{Z_1}{e_1(X_1)}(1-D^+)(1_{\{Y>q(1-p(X_1),X_1)\}}-p(X_1))\Big]$ $-\frac{Z_1-e_1(X_1)}{e_1(X_1)}z^{U+}(1,X_1)r_1(X_1) + \sum_{k=1}^{\overline{w}}1_{\{A\geq k\}}\Pi_{j=1}^{k-1}\frac{(1-Z_j)}{(1-e_j(X_j))}\frac{e_k(X_k)-Z_k}{1-e_k(X_k)}[r_k(X_k)\beta_k(X_k)$ $+q(1-p(X_1),X_1)(r_1(X_1)-r_k(X_k))]$
$\psi^{-}(G,\xi^0)$	$Y(1-D^+)\frac{Z_1}{e_1(X_1)}p(X_1) - Y(1-D^+)\Pi_{j=1}^{\overline{w}}\frac{(1-Z_j)}{(1-e_j(X_j))}$ $-\beta^+(X_1)\Big[\frac{Z_1}{e_1(X_1)}\frac{(1-D^+-r^+(X_1))}{r^+(X_1)}r_1(X_1) - \Pi_{j=1}^{\overline{w}}\frac{(1-Z_j)}{(1-e_j(X_j))}(1-D^+-r_1(X_1))\Big]$ $-\frac{Z_1-e_1(X_1)}{e_1(X_1)}\beta^+(X_1)r_1(X_1) + \sum_{k=1}^{\overline{w}}1_{\{A\geq k\}}\Pi_{j=1}^{k-1}\frac{(1-Z_j)}{(1-e_j(X_j))}\frac{e_k(X_k)-Z_k}{1-e_k(X_k)}\big[r_k(X_k)\beta_k(X_k)$ $+\beta^+(X_1)(r_1(X_1)-r_k(X_k))\big]$
$\psi^{R}(G,\xi^0)$	$r_1(X_1) + (1-D^+-r_1(X_1))\Pi_{j=1}^{\overline{w}}\frac{(1-Z_j)}{(1-e_j(X_j))}$ $+\Sigma_{k=1}^{\overline{w}}1_{\{A\geq k\}}\Pi_{j=1}^{k-1}\frac{(1-Z_j)}{1-e_j(X_j)}\frac{(e_k(X_k)-Z_k)}{1-e_k(X_k)}[r_1(X_1)-r_k(X_k)]$
	Nuisance functions
$\xi^0(x_1,\ldots,x_{\overline{w}})$	$\{e_1(x_1),\ldots,e_{\overline{w}}(x_{\overline{w}}), r_1(x_1),\ldots,r_{\overline{w}}(x_{\overline{w}}), r^+(x_1), q(p(x_1),x_1), q(1-p(x_1),x_1),$ $\beta_1(x_1),\ldots,\beta_{\overline{w}}(x_{\overline{w}}), \beta^+(x_1), z^{U+}(x_1), z^{L+}(x_1)\}$
$r_k(x)$	$\mathbb{E}[(1-D^+)/(\Pi_{j=k+1}^{A}(1-e_j(X_j))) \mid X_k=x, Z_A=0]$ $\mathbb{E}[\Pi_{j=k+1}^{A}(1-e_j(X_j)) \mid X_k=x, Z_A=0]$
$\beta_k(x)$	$\mathbb{E}[Y/(\Pi_{j=k+1}^{A}(1-e_j(X_j))) \mid X_k=x, D=0]$ $\mathbb{E}[\Pi_{j=k+1}^{A}(1-e_j(X_j)) \mid X_k=x, D=0]$
$\beta^+(x)$	$\mathbb{E}[Y \mid X_1=x, Z_1=1, D^+=0]$
$z^{U+}(x)$	$\mathbb{E}[Y \mid X_1=x, Z_1=1, D^+=0, Y\geq q(1-p(x),x)]$
$z^{L+}(x)$	$\mathbb{E}[Y \mid X_1=x, Z_1=1, D^+=0, Y\leq q(p(x),x)]$

$[0,1) \to \psi^{U+}(G,r) \equiv \psi^{U+}(G,\xi_r)$, where:

$$\xi_r = \{e_1(x_1),\ldots,e_l(x_l,r),\ldots,e_{\overline{w}}(x_{\overline{w}}), r_1(x_1),\ldots,r_{\overline{w}}(x_{\overline{w}}), r^+(x_1), q(p(x_1),x_1),$$
$$q(1-p(x_1),x_1), \beta_1(x_1),\ldots,\beta_{\overline{w}}(x_{\overline{w}}), \beta^+(x_1), z^{U+}(x_1), z^{L+}(x_1)\},$$

and where $e_l(x_l,r) = e_l(x_l) + r(\widehat{e}_l(x_l) - e_l(x_l))$, meaning that $e_l(x_l,0) = e_l(x_l)$. Then,

for the new moment:

$$\partial_r \, \mathbb{E}[\psi^{U+}(G, \xi_r)|X_l]|_{r=0} = 0 \; a.s.,$$

while for the original moment:

$$\partial_r \, \mathbb{E}[m^{U}(G, \eta_r)|X_l]|_{r=0} \neq 0 \; a.s.,$$

meaning that the original moment is sensitive to estimation errors in $\widehat{e}_j(X_j)$, whereas the new moment is not.

The estimator for θ_L is:

$$\widehat{\theta_L} = \left(\sum_i \left(\psi^{L+}(G_i, \widehat{\xi_i}) 1_{\{p(X_1) \leq 1\}} + \psi^{-}(G_i, \widehat{\xi_i}) 1_{\{p(X_1) > 1\}} \right) \right) / \left(\sum_i \psi^{R}(G_i, \widehat{\xi_i}) \right)$$

where G_i is the data for observation i and $\widehat{\xi_i}$ is the nuisance parameter for observation i, estimated on a subsample that excludes observation i. The estimator for θ_U is symmetric.

A2.2 Orthogonality

I demonstrate orthogonality of $\psi^{U+}(G, \xi^0)$ with respect to one of the propensity scores $e_l(x_l)$ for l s.t. $1 < l < \overline{w}$. The arguments for other parameters involve first applying the Lemma to eliminate the dependence of the conditional expectation of the moment function on propensity scores $e_j(x_j)$ for $j > 1$. Afterward, the steps are similar to those in Semenova (2023). The approach for other moments follows a similar process.

I demonstrate that $\partial_r \, \mathbb{E}[\psi^{U+}(G, \xi_r)|X_l]|_{r=0} = 0$, a.s.. Since the moment does not depend on r when $A < l$ (because $1_{\{A \geq l\}} = 0$ and because $e_l(X_l, r) = e_l(X_l)$ in such cases) it is sufficient to show that $\partial_r \, \mathbb{E}[\psi^{U+}(G, \xi_r)|X_l]|_{r=0} = 0$ for values of X_l s.t. $A \geq l$; the rest of the argument assumes X_l satisfies this condition.

For $k \geq l$ define $S_k \equiv \{1, \ldots, k\} \setminus \{l\}$. Using that $Z_j = 0, e_j(X_j) = 0 | A < l$ $\mathbb{E}[\psi^{U+}(G, \xi_r)|X_l]$ simplifies to:

$$\mathbb{E}[\psi^{U+}(G, \xi_r)|X_l] = \mathbb{E}\left[- Y(0) R \Pi_{j \in S_{\overline{w}}} \frac{(1 - Z_j)}{(1 - e_j(X_j))} \frac{(1 - Z_l)}{(1 - e_l(X_l, r))} \right.$$

$$
\begin{aligned}
&+ q(p(X_1), X_1)[\Pi_{j \in S_{\overline{w}}} \frac{(1 - Z_j)}{(1 - e_j(X_j))} \frac{(1 - Z_l)}{(1 - e_l(X_l, r))} (R - r_1(X_1))] \\
&+ \sum_{k=l+1}^{\overline{w}} 1_{\{A \geq k\}} \Pi_{j \in S_{k-1}} \frac{(1 - Z_j)}{(1 - e_j(X_j))} \frac{(1 - Z_l)}{(1 - e_l(X_l, r))} \frac{e_k(X_k) - Z_k}{1 - e_k(X_k)} [r_k(X_k) \beta_k(X_k) \\
&+ q(p(X_1), X_1)(r_1(X_1) - r_k(X_k))] \\
&+ 1_{\{A \geq k\}} \Pi_{j \in S_l} \frac{(1 - Z_j)}{(1 - e_j(X_j))} \frac{e_l(X_l, r) - Z_l}{1 - e_l(X_l, r)} [r_k(X_k) \beta_k(X_k) \\
&+ q(p(X_1), X_1)(r_1(X_1) - r_k(X_k))] \Big| X_l \Big].
\end{aligned}
$$

Define:

$$
\begin{aligned}
f_k^l(X_k) \equiv 1_{\{A \geq k\}} \Pi_{j \in S_{k-1}} &\frac{(1 - Z_j)}{(1 - e_j(X_j))} \frac{1 - Z_l}{(1 - e_l(X_l, r))} [r_k(X_k) \beta_k(X_k) \\
&+ q(p(X_1), X_1)(r_1(X_1) - r_k(X_k))].
\end{aligned}
$$

For $k > l$:

$$
\begin{aligned}
&\mathbb{E}\left[f_k^l(X_k) \frac{e_k(X_k) - Z_k}{1 - e_k(X_k)} \middle| X_l \right] = \mathbb{E}\left[\mathbb{E}\left[f_k^l(X_k) \frac{e_k(X_k) - Z_k}{1 - e_k(X_k)} \middle| X_k, X_l \right] \middle| X_l \right] && (13) \\
&= \mathbb{E}\left[f_k^l(X_k)\, \mathbb{E}\left[\frac{e_k(X_k) - Z_k}{1 - e_k(X_k)} \middle| X_k \right] \middle| X_l \right] && (14) \\
&= 0, && (15)
\end{aligned}
$$

where (13) holds by law of iterated expectations, (14) holds because X_k contains X_l, and (15) holds because: $\mathbb{E}\left[\frac{e_k(X_k) - Z_k}{1 - e_k(X_k)} \middle| X_k \right] = 0$.

Moreover, defining $\mathbb{E}\left[\frac{(1 - Z_l)}{(1 - e_l(X_l, r))} \middle| X_l \right] \equiv h_l^r(X_l, Z_l)$, by the Corollary:

$$
\begin{aligned}
&\mathbb{E}\Bigg[- Y(0) R \Pi_{j \in S_{\overline{w}}} \frac{(1 - Z_j)}{(1 - e_j(X_j))} \frac{(1 - Z_l)}{(1 - e_l(X_l, r))} \\
&+ q(p(X_1), X_1)[\Pi_{j \in S_{\overline{w}}} \frac{(1 - Z_j)}{(1 - e_j(X_j))} \frac{(1 - Z_l)}{(1 - e_l(X_l, r))} (R - r_1(X_1))] \Big| X_l \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\Bigg[- Y(0)R\Pi_{j\in S_l}\frac{(1-Z_j)}{(1-e_j(X_j))}\frac{(1-Z_l)}{(1-e_l(X_l,r))} \\
&+ q(p(X_1),X_1)[\Pi_{j\in S_l}\frac{(1-Z_j)}{(1-e_j(X_j))}\frac{(1-Z_l)}{(1-e_l(X_l,r))}(R-r_1(X_1))]\Big| X_l\Bigg] \\
&= \Pi_{j\in S_l}\frac{(1-Z_j)}{(1-e_j(X_j))}\,\mathbb{E}[-Y(0)R + q(p(X_1),X_1)[(R-r_1(X_1))]|X_l]h_l^r(X_l,Z_l) \qquad (16)\\
&= \Pi_{j\in S_l}\frac{(1-Z_j)}{(1-e_j(X_j))}[-\beta_l(X_l)r_l(X_l) + q(p(X_1),X_1)(r_l(X_l)-r_1(X_1))]h_l^r(X_l,Z_l) \\
&\qquad (17)\\
&\equiv \mu_l(X_l)h_l^r(X_l,Z_l), \qquad (18)
\end{aligned}
$$

where (16) also holds by the Corollary, and where (17) holds by applying the Corollary to the definitions of $\beta_l(.)$ and $r_l(.)$, after noting that we are considering values of X_l s.t. $1_{\{A>l\}} = 1$.

Similarly, applying the Corollary yields:

$$
\begin{aligned}
&\mathbb{E}\Bigg[\Pi_{j\in S_l}\frac{(1-Z_j)}{(1-e_j(X_j))}\frac{e_l(X_l,r)-Z_l}{(1-e_l(X_l,r))}[r_l(X_l)\beta_l(X_l) \\
&+ q(p(X_1),X_1)(r_1(X_1)-r_l(X_l))]\Big| X_l\Bigg] \\
&= \Pi_{j\in S_l}\frac{(1-Z_j)}{(1-e_j(X_j))}\,\mathbb{E}\left[\frac{e_l(X_l,r)-Z_l}{(1-e_l(X_l,r))}\Big| X_l\right][r_l(X_l)\beta_l(X_l) \\
&+ q(p(X_1),X_1)(r_1(X_1)-r_l(X_l))].
\end{aligned}
$$

Combining the above, $\mathbb{E}[\psi^{U+}(G,\xi_r)|X_l]$ simplifies to:

$$
\begin{aligned}
\mathbb{E}[\psi^{U+}(G,\xi_r)|X_l] &= \mu_l(X_l)\left[\mathbb{E}\left[\frac{(1-Z_l)}{(1-e_l(X_l,r))}\Big| X_l\right] - \mathbb{E}\left[\frac{e_l(X_l,r)-Z_l}{(1-e_l(X_l,r))}\Big| X_l\right]\right] \\
&= \mu_l(X_l)\left[\frac{1-e_l(X_l)}{1-e_l(X_l,r)} - \frac{e_l(X_l,l)-e_l(X_l)}{1-e_l(X_l,r)}\right] \\
&= \mu_l(X_l),
\end{aligned}
$$

meaning that $\partial_r \mathbb{E}[\psi^{U+}(G, \xi_r)|X_l]|_{r=0} = 0$ a.s.. Meanwhile, for the baseline moment:

$$\begin{aligned}\partial_r \mathbb{E}[m^U(G, \eta_r)|X_l]|_{r=0} &= \partial_r \mu_l(X_l) \frac{1 - e_l(X_l)}{1 - e_l(X_l, r)}\bigg|_{r=0} \\ &= \mu_l(X_l) \frac{1 - e_l(X_l)}{(1 - e_l(X_l, r))^2} (\widehat{e}_l(X_l) - e_l(X_l)).\end{aligned}$$

meaning that $\partial_r \mathbb{E}[m^U(G, \eta_r)|X_l]|_{r=0} \neq 0$ a.s..

A2.3 Implementation

I use 3-fold cross-fitting, meaning that in each sample split 2/3 of the observations are used to estimate the nuisance parameter. Because I assume that the propensity scores only include a few discrete covariates—age in years, a dummy for higher education, and procedure type—they could be estimated nonparametrically using saturated fixed effects regressions. However, later propensity scores need to be estimated on small samples, and including many fixed effects makes them susceptible to outliers. This is especially undesirable because these scores are also used as weights to estimate other nuisance functions. Instead, in my main specification, I estimate them using logistic regressions. Specifically, for each ACP, I regress the outcome among women who entered that ACP on second-order polynomials of women's and partners' ages at the time of the procedure, interacted with treatment-type dummies (IUI or ACP), and separate dummies for each partner having at least a bachelor's degree.[32] To further avoid outlier weights, I only use the first 10 ACPs women undergo and treat conceptions through later ACPs as natural; only 7% of women reach the tenth ACP. This means that, in my application, reliers are women who would remain childless in the scenario where their first 10 ACPs fail. Including up to 15 ACPs has little impact on my estimates. The remaining nuisance functions are estimated using Generalized Random Forests for conditional expectations and quantiles (Athey et al., 2019).[33]

[32] Using age-fixed effects and/or excluding the education dummies has little impact.

[33] I estimate z_t^{U+} and z_t^{L+} by trimming data above or below the estimated quantiles and estimating conditional expectations. Ideally, a nonparametric estimator for truncated conditional expectations, such as Olma (2021), should be used, but none are implemented at the time of writing.

The covariates in X_1 include the woman's and their partner's income and work hours measured in the year before the woman's first ACP, and other covariates included in the first propensity score. The covariates in X_k additionally include those from the propensity scores at all ACPs up to and including ACP k. I modify work hours and income outcomes by adding a small amount of continuously distributed noise to ensure the new outcomes are continuous $u \sim U(0, 0.001)$.[34] Following Heiler (2024), I based my confidence intervals for the bounds on Stoye (2020).

Confidence intervals for the bounds on the effect scaled by the treated mean are also based on Stoye (2020), with covariance matrices estimated using delta method in the following steps: (1) estimate $\widehat{\xi}_i$ using cross-fitting, (2) construct separate sample moments for the control mean and the upper and lower bounds for the treated mean evaluated at $\widehat{\xi}_i$ (m_1, m_2, and m_3, respectively), (3) compute the joint covariance matrix for the three sample moments, (4) obtain the joint covariance matrix for $(m_2 - m_1)/m_2$ and $(m_3 - m_1)/m_3$ using delta method.

A3 Extensions for Bounds

Section A3.1 introduces an alternative monotonicity assumption following Semenova (2023) and presents the corresponding estimates. Section A3.2 proposes a method to leverage continuous covariates without relying on orthogonalization and sample splitting and reports the results. Section A3.3 discusses the age differences between partners, their implications for calculating the share of gender inequality caused by parenthood, and provides relevant robustness checks. Section A3.4 presents an extension to bound effects over time for a stable group. Section A3.5 presents bounds that do not leverage monotonicity.

A3.1 Relaxing Monotonicity Following Semenova (2023)

A challenge in implementing the bounds under a non-trivial monotonicity assumption arises when, for some values of X_1, such as x_1^*, the estimated relier share exceeds the subsequent relier share. This may indicate a violation of monotonicity but can also

[34] The procedure requires continuously distributed outcomes only to avoid ties in trimming; adding continuously distributed noise resolves the issue.

occur due to estimation noise, as the two shares are computed from different groups. In either case, it results in the estimated trimming share $p(x_1^*)$ exceeding one, making the quantile function $q(p(x_1^*), x_1^*)$ ill-defined. To address a similar issue in Lee (2009), Semenova (2023) relaxes the monotonicity assumption, allowing its direction to vary with X_1. In my setting, this implies that all women with certain pre-ACP covariates who had a non-ACP child after ACP failure would have also had a non-ACP child if their first ACP had succeeded. This assumption is harder to justify economically and the adapted approach is challenging to implement, as it requires estimating weighted quantile functions on small groups of women entering subsequent ACPs. Because of this, my main specification retains the original monotonicity assumption, treating cases where the estimated relier share exceeds the subsequent relier share as if the two were equal. If this reversal occurs because the true shares are very close, this method and the approach that follows Semenova (2023) should yield nearly identical results. Under sequential unconfoundedness, the expectation of the moment for treating them as equal, $\psi^-(G, \xi^0)$, identifies the difference between the conditional subsequent relier average treated outcome and the conditional relier average control outcome:

$$\frac{\mathbb{E}[\psi^-(G, \xi^0) \mid X_1 = x_1^*]}{\mathbb{E}[r(X_1) \mid X_1 = x_1^*]} = \mathbb{E}[Y(1) \mid R^+ = 1, X_1 = x_1^*] - \mathbb{E}[Y(0) \mid R = 1, X_1 = x_1^*].$$

When the shares of the two types are equal, monotonicity implies that the two groups are the same, and the difference between the two terms is $\mathbb{E}[\tau \mid R = 1, X_1 = x_1^*]$.

To test the sensitivity of my result, I allow for the direction of monotonicity to vary with covariates following Semenova (2023). Define $\mathcal{X}_{help} \equiv \{x : r^+(x) \geq r(x)\}$ and $\mathcal{X}_{hurt} \equiv \mathcal{X}_1 \setminus \mathcal{X}_{help}$. The relaxed monotonicity assumption is that $\forall x \in \mathcal{X}_{help}$ $R^+ \geq R$ *a.s.*, and $\forall x \in \mathcal{X}_{hurt}$ $R^+ < R$ *a.s.*. Table A2 presents the moments for the case when $X_1 \in \mathcal{X}_{hurt}$. The new estimator of the lower bound is:

$$\frac{\sum_i \left(\psi^{L+}(G_i, \widehat{\zeta}_i) 1_{\{p(X_1) \leq 1\}} + \psi^{L-}(G_i, \widehat{\zeta}_i) 1_{\{p(X_1) > 1\}}\right)}{\sum_i \left(\psi^{R}(G_i, \widehat{\zeta}_i) 1_{\{p(X_1) \leq 1\}} + \psi^{R+}(G_i, \widehat{\zeta}_i) 1_{\{p(X_1) > 1\}}\right)}.$$

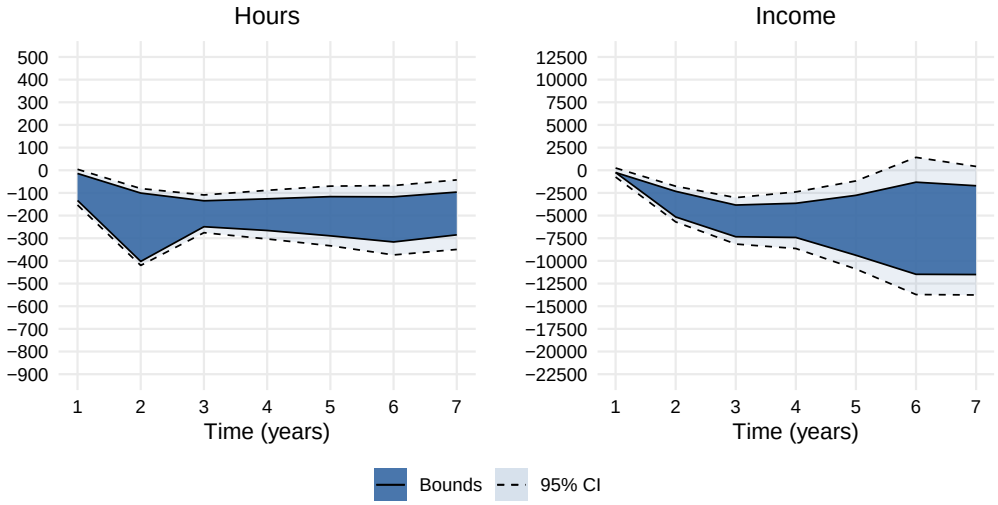

Figure A11: Effect on Women Under Relaxed Monotonicity Following Semenova (2023)

The new estimator of the upper bound is:

$$\frac{\sum_i \left(\psi^{U+}(G_i, \widehat{\zeta}_i) 1_{\{p(X_1)\leq 1\}} + \psi^{U-}(G_i, \widehat{\zeta}_i) 1_{\{p(X_1)>1\}}\right)}{\sum_i \left(\psi^{R}(G_i, \widehat{\zeta}_i) 1_{\{p(X_1)\leq 1\}} + \psi^{R+}(G_i, \widehat{\zeta}_i) 1_{\{p(X_1)>1\}}\right)}.$$

I implement it following the baseline approach. Since a weighted generalized quantile forests estimator is unavailable, I estimate all nuisance functions involving expectations and quantiles using OLS and quantile regressions, respectively. Using regression for the quantile function has little impact on the estimates. Figure A11 shows the results for women's outcomes, which closely resemble the baseline estimates.

A3.2 Alternative Approach to Continuous Covariates

Here I introduce a new method to narrow the bounds by leveraging continuous covariates. For a known measurable function $g : \mathcal{X}_1 \to \mathbb{R}$, define $\varepsilon \equiv Y(1) - g(X_1)$. Intuitively, $g(X_1)$ can be thought of as OLS fitted values, and ε can be thought of as OLS residuals after regressing Y on X_1 among women whose first ACP succeeded. The idea behind the new approach is that the component of $\mathbb{E}[Y(1)|R = 1]$ explained by $g(.)$ can be identified. As a result, only the residual component needs to be bounded, and the distribution of ε can be tighter than the distribution of $Y(1)$, which results in

Table A2: Moment Functions for Covariate-Conditional Monotonicity

	Moment functions
$\psi_L^-(W,\zeta_0)$	$\frac{Z_1}{e_1(X_1)}(1-D^+)Y - \Pi_{j=1}^{\overline{w}}\frac{1-Z_j}{1-e_j(X_j)}(1-D^+)Y1_{\{Y>q^0(1-1/p(X_1),X_1)\}}$ $-q^0(1-1/p(X_1),X_1)\Big[\frac{Z_1}{e_1(X_1)}(1-D^+-r^+(X_1))$ $-\Pi_{j=1}^{\overline{w}}\frac{1-Z_j}{1-e_j(X_j)}\frac{1}{p(X_1)}(1-D^+-r_1(X_1))$ $-\Pi_{j=1}^{\overline{w}}\frac{1-Z_j}{1-e_j(X_j)}(1-D^+)(1_{\{Y>q^0(1-1/p(X_1),X_1)\}}-1/p(X_1))\Big]$ $-\frac{Z_1-e_1(X_1)}{e_1(X_1)}\beta^+(1,X_1)r^+(X_1)$ $+\sum_{k=1}^{\overline{w}}1_{\{A\geq k\}}\Pi_{j=1}^{k-1}\frac{1-D_j}{1-e_j(X_j)}\frac{e_k(X_k)-D_k}{1-e_k(X_k)}$ $\times\Big[\Big(r_k(X_1)r_k^L(X_k)z_k^{L-}(X_k)+\frac{q^0(1-1/p(X_1),X_1)}{p(X_1)}(r_1(X_1)-r_k(X_1))\Big)$ $+q^0(1-1/p(X_1),X_1)r_k(X_1)(1/p(X_1)-r_k^L(X_k))\Big]$
$\psi_U^-(W,\zeta_0)$	$\frac{Z_1}{e_1(X_1)}(1-D^+)Y - \Pi_{j=1}^{\overline{w}}\frac{1-Z_j}{1-e_j(X_j)}(1-D^+)Y1_{\{Y<q^0(1/p(X_1),X_1)\}}$ $-q^0(1/p(X_1),X_1)\Big[\frac{Z_1}{e_1(X_1)}(1-D^+-r^+(X_1))$ $-\Pi_{j=1}^{\overline{w}}\frac{1-Z_j}{1-e_j(X_j)}\frac{1}{p(X_1)}(1-D^+-r_1(X_1))$ $-\Pi_{j=1}^{\overline{w}}\frac{1-Z_j}{1-e_j(X_j)}(1-D^+)(1_{\{Y<q^0(1/p(X_1),X_1)\}}-1/p(X_1))\Big]$ $-\frac{Z_1-e_1(X_1)}{e_1(X_1)}\beta^+(1,X_1)r^+(X_1)$ $+\sum_{k=1}^{\overline{w}}1_{\{A\geq k\}}\Pi_{j=1}^{k-1}\frac{1-D_j}{1-e_j(X_j)}\frac{e_k(X_k)-D_k}{1-e_k(X_k)}$ $\times\Big[\Big(r_k(X_1)r_k^U(X_k)z_k^{U-}(X_k)+\frac{q^0(1/p(X_1),X_1)}{p(X_1)}(r_1(X_1)-r_k(X_1))\Big)$ $+q^0(1/p(X_1),X_1)r_k(X_1)(1/p(X_1)-r_k^U(X_k))\Big]$
$\psi^{R+}(G,\zeta^0)$	$r^+(X_1)+(1-D^+-r^+(X_1))\frac{Z_1}{e_1(X_1)}$
	Nuisance functions
$\zeta^0(x_1,\ldots,x_{\overline{w}})$	$\{e_1(x_1),\ldots,e_{\overline{w}}(x_{\overline{w}}),r_1(x_1),\ldots,r^{\overline{w}}(x_{\overline{w}}),r^+(x_1),q(p(x_1),x_1),q(1-p(x_1),x_1),$ $\beta^1(x_1),\ldots,\beta^{\overline{w}}(x_{\overline{w}}),\beta^+(x_1),z^{U+}(x_1),z^{L+}(x_1),z_1^{U-}(x_1),\ldots,z_{\overline{w}}^{U-}(x_{\overline{w}}),q^0(1/p(x_1),x_1),$ $q^0(1-1/p(x_1),x_1),z_1^{L-}(x_1),\ldots,z_{\overline{w}}^{L-}(x_{\overline{w}}),r_1^L(x_1),\ldots,r_{\overline{w}}^L(x_{\overline{w}}),r_1^U(x_1),\ldots,r_{\overline{w}}^U(x_{\overline{w}})\}$
$q^0(u,x)$	$inf\{q:u\leq\mathbb{E}[1_{\{Y\leq q\}}/\Pi_{j=2}^{\overline{w}}(1-e_j(X_j))\mid X_1=x,D=0]/$ $\mathbb{E}[\Pi_{j=2}^{\overline{w}}(1-e_j(X_j))\mid X_1=x,D=0]\}$
$z_k^{L-}(x)$	$\mathbb{E}[Y/\Pi_{j=k+1}^{\overline{w}}(1-e_j(X_j)\vert Y\geq q^0\left(1-1/p(X_1),X_1\right),D=0,X_k=x]$ $\mathbb{E}[\Pi_{j=k+1}^{\overline{w}}(1-e_j(X_j)\vert Y\geq q^0\left(1-1/p(X_1),X_1\right),D=0,X_k=x]$
$z_k^{U-}(x)$	$\mathbb{E}[Y/\Pi_{j=k+1}^{\overline{w}}(1-e_j(X_j)\vert Y\leq q^0\left(1/p(X_1),X_1\right),D=0,X_k=x]$ $\mathbb{E}[\Pi_{j=k+1}^{\overline{w}}(1-e_j(X_j)\vert Y\leq q^0\left(1/p(X_1),X_1\right),D=0,X_k=x]$
$r_k^L(x)$	$\mathbb{E}[1_{Y>q^0(1-1/p(X_1),X_1)}/\Pi_{j=k+1}^{\overline{w}}(1-e_j(X_j)\vert D=0,X_k=x]$ $\mathbb{E}[\Pi_{j=k+1}^{\overline{w}}(1-e_j(X_j)\vert Y\leq q^0\left(1/p(X_1),X_1\right),D=0,X_k=x]$
$r_k^U(x)$	$\mathbb{E}[1_{Y<q^0(1/p(X_1),X_1)}/\Pi_{j=k+1}^{\overline{w}}(1-e_j(X_j)\vert D=0,X_k=x]$ $\mathbb{E}[\Pi_{j=k+1}^{\overline{w}}(1-e_j(X_j)\vert Y\leq q^0\left(1/p(X_1),X_1\right),D=0,X_k=x]$

narrower bounds. Formally, first, by definition, $\mathbb{E}[Y(1)|R=1]=\mathbb{E}[g(X_1)+\varepsilon|R=1]$. Second, since X_1 is observed, $\mathbb{E}[g(X_1)|R=1]$ can be identified using women who

remain childless similar to $\mathbb{E}[Y(0)|R=1]$, specifically:

$$\mathbb{E}\left[g(X_1)\frac{(1-D)}{\Pi_{j=1}^{\overline{w}}(1-e_j(X_j))}\right]\Big/\mathbb{E}\left[\frac{(1-D)}{\Pi_{j=1}^{\overline{w}}(1-e_j(X_j))}\right]=\mathbb{E}\left[g(X_1)\mid R=1\right].$$

Since among women whose first ACP succeeds, $Y(1)$ and $g(X_1)$ are observed, ε is observed, meaning that $\mathbb{E}[\varepsilon|R=1]$ can be bounded similar to how $\mathbb{E}[Y(1)|R=1]$ is bounded using the baseline method without covariates. Then, it can be combined with the point-identified $\mathbb{E}[g(X_1)|R=1]$ to obtain bounds on $\mathbb{E}[Y(1)|R=0]$.

The bounds obtained using this approach need not be narrower and could even be wider than the baseline bounds that ignore covariates. To see this, consider a case where Y is constant. In this case, the baseline bounds collapse to a point, whereas the new bounds may not, since $g(X_1)$ need not be constant, meaning that ε is not constant either. In practice, however, the bounds can be substantially narrower than those that do not leverage covariates, and in some cases, they can match the sharp bounds. Bounds based on different $g(.)$'s can also be compared empirically. This approach can also be used to leverage continuous covariates in the Lee (2009) setting and can be combined with the method proposed by Lee (2009) to use discrete covariates by estimating bounds for each discrete covariate cell before aggregating.

I implement the above approach in the following steps: (1) estimate $e_j(x_j)$ for all j to obtain estimates of weights $w_i^w = Z_{1i}/e_1(X_{1i}) + (1-Z_{Ai})/\Pi_{j=1}^{A_i}(1-e_j(X_{ji}))$, (2) estimate $g(x_1)$ by regressing Y on X_1 using women with $Z_1=1$ and estimates of weights w_i^w, (3) separately regress D on X_1 using women whose first ACP succeeded and women whose ACPs failed, with estimates of weights w_i^w, (4) split the sample into quintiles based on differences in fitted values for the two regressions in step (3), (5) estimate bounds on the effect in each quintile using $Y-g(X_1)$ as the outcome with weights w_i^w, (6) aggregate across bins with weights proportional to the estimated relier share in each bin with estimates of weights w_i^w. Confidence intervals are based on Stoye (2020) with the covariance matrix obtained via Bayesian bootstrap with 150 draws. Weights $w_i \sim \exp(1)$ are used for step (1), and weights w_i^w are replaced with $w_i w_i^w$ for other steps. Figure A12 presents the results for women's outcomes, showing that the estimates remain largely consistent with the baseline approach.

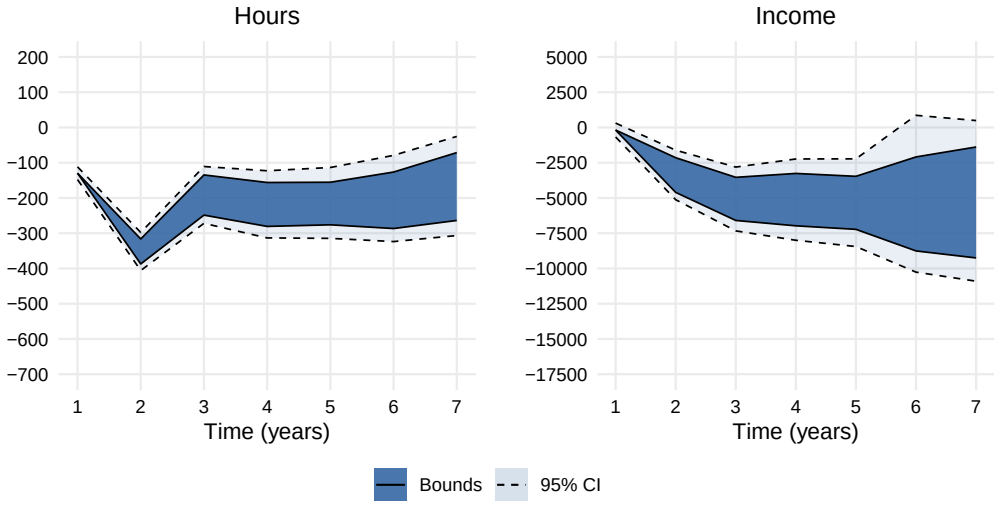

Figure A12: Effects on Women Using Residualization Approach (Leave-Adjusted Hours)

A3.3 Accounting for Age Difference Between Partners

My estimates of the share of gender inequality caused by parenthood focus on the within-couple gender gap in each year after becoming parents. This gap also captures differences related to the within-couple age gap, which may distort the picture of aggregate gender inequality in the economy because men's outcomes are measured at systematically older ages. A particular concern is that if work hours and income increase with age, my estimates might understate the share of aggregate gender inequality caused by parenthood.

Ideally, using cumulative lifetime outcomes would directly address this issue, but since such data is unavailable, a different approach is required. One way would be to correct for age differences parametrically, but this relies on strong and potentially opaque assumptions. Instead, I opt for a simple approach that fits into my framework: I adjust the timing of when men's outcomes are measured based on the women's age. For example, if a woman is two years younger than her male partner, I lag the male's outcome in each period by two years. This adjustment ensures that gender gaps in outcomes are assessed at comparable life-cycle stages within couples. The adjustment reduces my sample by 22%, as it excludes couples where the male partner is much older or younger, leaving me with 12,146 observations.

Figure A13 presents the results, showing that the adjustment has minimal impact

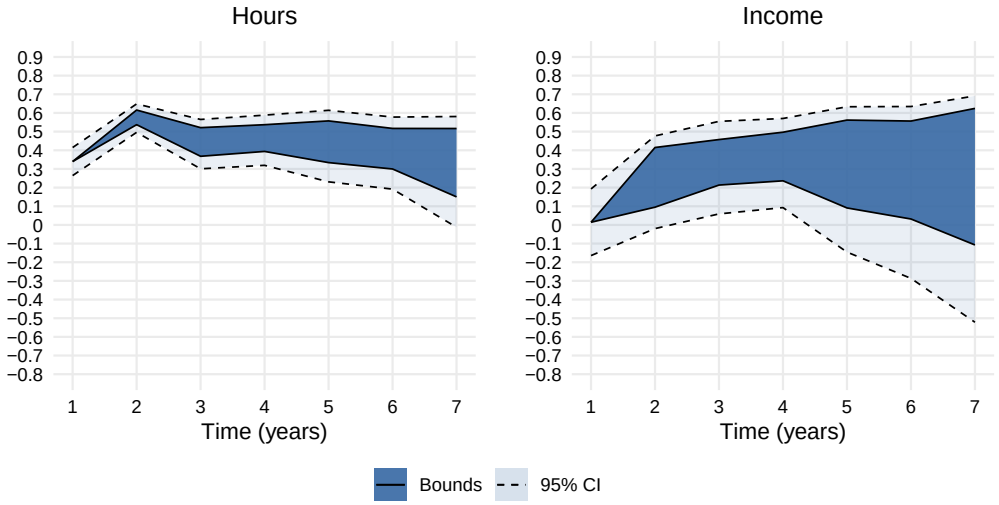

Figure A13: Share of Gender Inequality Caused by Parenthood Using Partner's Income Lagged to Match Woman's Age (Leave-Adjusted Hours)

on estimates. The upper bound for the share of gender inequality in hours decreases slightly, while that for income increases by no more than 10 percentage points.

A3.4 Effects Over Time and Stable Relier Group

The changes in my main estimates over time reflect a combination of two factors. First, how the effect of parenthood evolves with time spent in parenthood. Second, since the group of reliers shrinks over time, how effects differ between women who remain reliers for a different duration. This means that my main results provide limited insight into how the effects evolve with time spent in parenthood. A similar concern regarding changing compliers applies to both the IV and the ES estimates.

To address this, I adapt my approach to bound the effects for women who remain reliers until the last period, allowing evaluation of how effects change with time spent in parenthood. This is enabled by the irreversibility of fertility, which ensures that past control outcomes are observable for any childless group at a given time, and the full trajectory of treated outcomes is observable for all women whose first ACP succeeded. For identification and estimation, this involves replacing outcomes Y_t

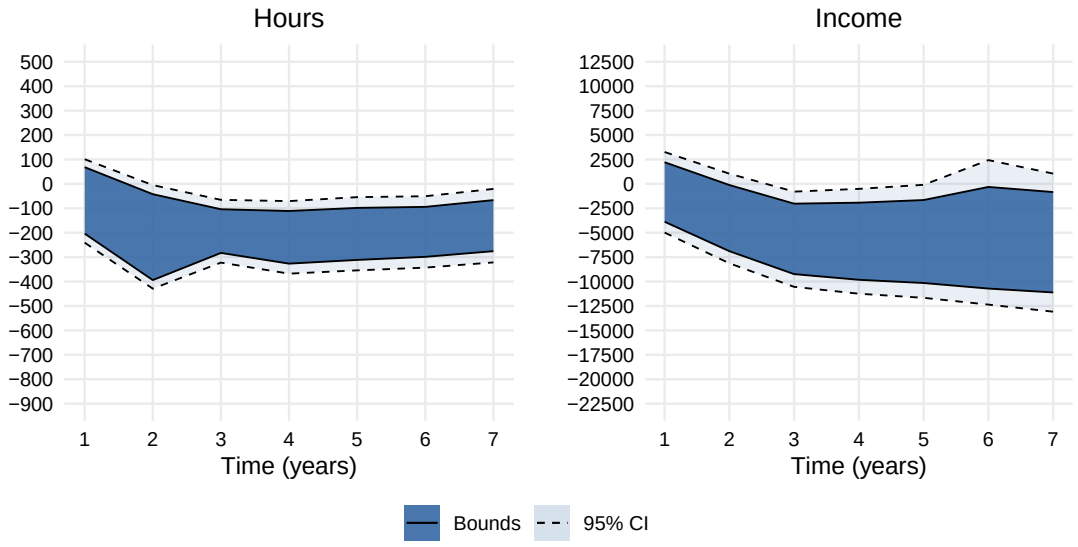

Figure A14: Effects for a Stable Group of Women

with past outcomes Y_k for $0 < k < t$.[35] Figure A14 presents estimates for women who remain reliers seven years after their first ACP, which align with the baseline results.

A3.5 Bounds without Monotonicity

Figure A15 presents effects on women's outcomes. In the first three years, the results are similar to baseline. By the fourth year, the bounds widen, including zero effects. Nonetheless, these results imply that parenthood causes at most 64% of gender inequality in hours and 70% in income over the sample period.

[35]The sequential unconfoundedness assumption must be adapted accordingly. Without monotonicity, these bounds are at least as wide as baseline since the trimmed non-relier share cannot decrease over time. With monotonicity, they may narrow if the decline in subsequent reliers exceeds the increase in reliers.

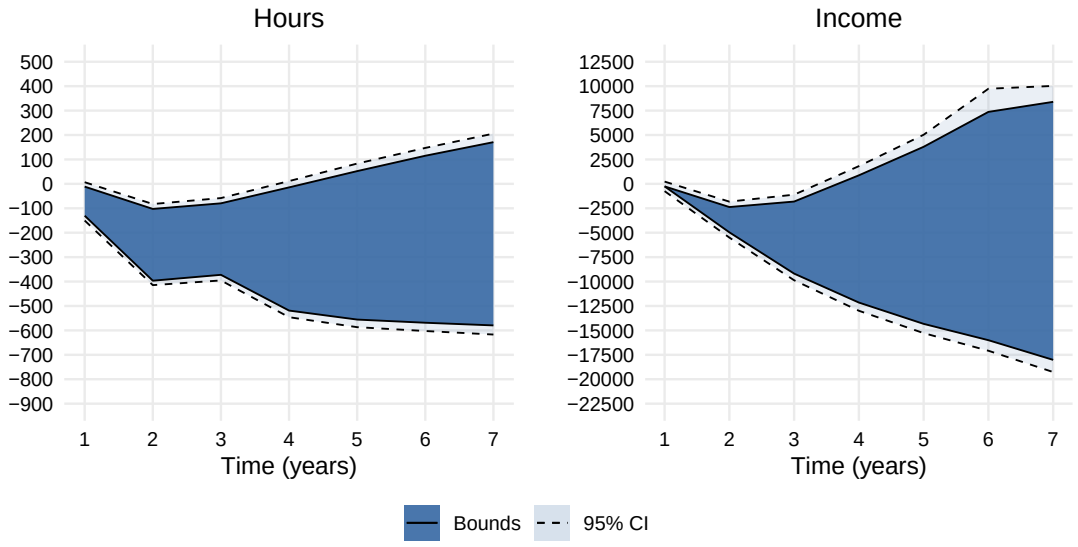

Figure A15: Effects on Women without Monotonicity

A4 Procedure Descriptives and Balance

Table A3 presents balance results for subsequent ACPs up to the tenth. Since these ACPs also include IVF, I additionally control for each partner's age interacted with treatment type. This ensures that ACP success only needs to be as good as random among women who undergo the same procedure (and are of similar age), allowing for selection into IUI or IVF based on women's types and potential outcomes. Overall, the results suggest no systematic differences in pre-ACP outcomes between those with successful and unsuccessful subsequent ACPs, supporting the conditional sequential unconfoundedness assumption.

Figure A16 presents the realized distribution of the number of ACPs women undergo, along with the estimated distributions of willingness to undergo ACP and relier shares, all measured seven years after the first ACP. It also presents the relationship between R and W, suggesting the two are uncorrelated.

Table A3: Balance in Later ACPs

	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}
Work (W)	0.009	-0.004	0.022	0.014	0.039	-0.003	-0.011	0.022	0.030
	(0.010)	(0.011)	(0.011)	(0.012)	(0.012)	(0.017)	(0.018)	(0.019)	(0.024)
Work (P)	0.006	0.016	0.012	0.020	-0.004	-0.004	-0.019	0.017	0.030
	(0.010)	(0.010)	(0.012)	(0.012)	(0.015)	(0.015)	(0.019)	(0.020)	(0.027)
Hours (W)	32.885	-4.482	52.999	41.332	81.957	11.894	-18.836	72.659	24.819
	(18.721)	(20.032)	(21.045)	(22.686)	(25.131)	(31.187)	(32.937)	(38.210)	(48.490)
Hours (P)	21.655	24.730	23.756	38.965	9.666	-6.580	-28.458	30.525	43.722
	(21.018)	(21.089)	(23.574)	(25.255)	(30.585)	(31.513)	(37.976)	(44.856)	(52.821)
Income 1000s € (W)	1.481	-0.015	1.685	1.802	2.086	0.150	-0.043	0.866	-0.444
	(0.615)	(0.624)	(0.767)	(0.830)	(0.913)	(1.000)	(1.092)	(1.234)	(1.629)
Income 1000s € (P)	-0.749	1.002	2.040	0.800	0.774	0.025	0.259	-0.324	0.149
	(0.835)	(0.912)	(1.066)	(1.115)	(1.424)	(1.424)	(1.563)	(1.737)	(2.203)
Observations	12,974	10,774	8,726	6,977	5,411	3,944	2,723	1,850	1,174
Joint p-val.	0.175	0.976	0.234	0.303	0.140	1.000	0.956	0.704	0.917

Note: Each column reports the difference in average characteristics between women whose respective ACP succeeded and those for whom it failed, among those who underwent the procedure, using inverse probability weights for each ACP following the main specification. Labor market outcomes measured in the year before first ACP. (W) - woman, (P) - partner. Standard errors in parentheses.

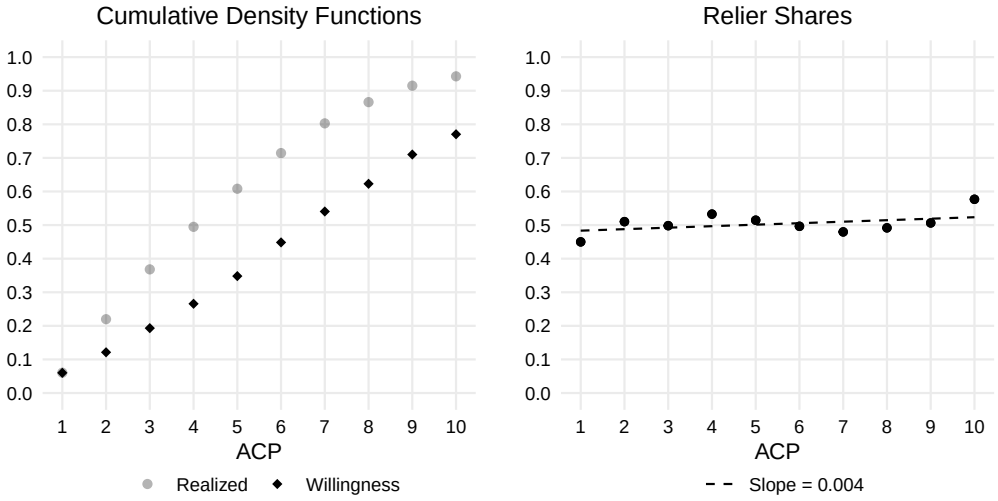

Figure A16: ACP Histories and Reliance

A5 Instrumental Variable and Event Study

I implement the IV following Lundborg et al. (2017), where the first stage specification is:

$$D_{it} = Z_{i1}\beta_t^{FS} + X_{i1}\chi_t^{FS} + \varepsilon_{it}^{FS},$$

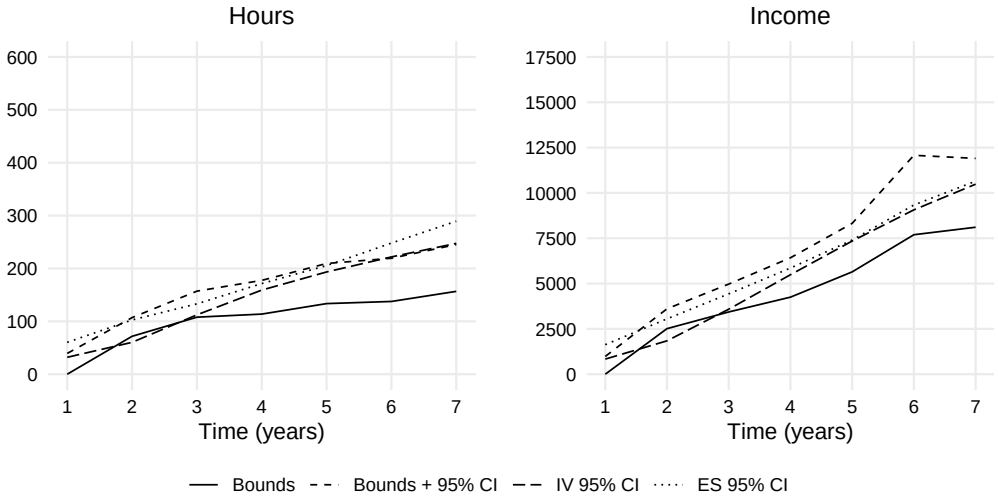

Figure A17: 95% Confidence Interval Width for Different Methods

and the second stage specification is:

$$Y_{it} = \widehat{D}_{it}\beta_t^{IV} + X_{i1}\chi_t^{IV} + \varepsilon_{it}^{IV},$$

where the parameters for the effect of parenthood in period t is β_t^{IV}.

I implement the ES following the fixed effect specification of Kleven et al. (2019):

$$Y_{it} = \beta_0^{ES} + \sum_{j\neq 0} \beta_j^{ES} 1_{\{t=j\}} + \sum_a \alpha_a 1_{\{age_{it}=a\}} + \sum_y \gamma_y 1_{\{year_{it}=y\}} + \upsilon_{it}, \tag{19}$$

where the parameters for the effect of parenthood in period t is β_t^{ES}.

A6 Confidence Interval Comparison

While my method only partially identifies the effects, my estimates are substantially more precise. Figure A17 compares the width of 95% confidence intervals for my bounds, IV estimates, and ES estimates. The ES is implemented using women whose first ACP succeeded. The confidence intervals for the three methods are almost identical. The primary source of uncertainty in my estimates arises from identifying bounds rather than sampling variability in the estimation process. Similarly, the

method introduced in Section A7, used to estimate τ_{ATR} under the assumption of static effects, provides a more precise alternative to the IV method. Intuitively, this improvement occurs because much of the uncertainty around IV estimates stems from scaling the reduced form by a low first stage. Leveraging women's complete ACP histories improves the first stage by expanding it from compliers to reliers, thereby reducing the amplification of noise.

A7 Testing Assumptions for Bias Correction

In this section, I introduce an estimator of $\tau_{ATR}(t)$ that parallels the Wald estimator of $\tau_{LATE}(t)$, as it identifies a linear combination of a relier average treatment effect and a relier average effect of delaying parenthood. I then demonstrate how each of the two estimators can be used to identify $\tau_{ATE}(t)$ under the parametric assumption that justifies the methods by Bensnes et al. (2023) and Gallen et al. (2023). Since the assumption implies that both methods should yield identical results, differing results suggest the assumption is violated.

I first introduce functions for estimating τ_{LATE} and τ_{ATR} in a doubly-robust manner to maximize precision:

$$
\begin{aligned}
g_a^{0+}(G) &= \gamma_a^{0,1+}(X_1) + (a - \gamma^{0,1+}(X_1))\Pi_{j=1}^{\overline{w}} \frac{(1-Z_j)}{(1-e_j(X_j))} \\
&\quad + \Sigma_{k=1}^{\overline{w}} \Big[1_{\{A \geq k\}} \Pi_{j=1}^{k-1} \frac{(1-Z_j)}{1-e_j(X_j)} \frac{(e_k(X_k) - Z_k)}{1-e_k(X_k)} [\gamma_a^{0,1+}(X_1) - \gamma_a^{0,k+}(X_k)] \Big] \\
g_a^0(G) &= \gamma_a^0(X_1) + (a - \gamma_a^0(X_1)) \frac{Z_1}{e_1(X_1)} \\
g_a^1(G) &= \gamma_a^1(X_1) + (a - \gamma_a^1(X_1)) \frac{1-Z_1}{1-e_1(X_1)},
\end{aligned}
$$

where $\gamma_a^1(X_1)$ is the OLS prediction of a given X_1 among observations with $Z_1 = 1$ with weights $1/e_1(X_1)$, $\gamma_a^0(X_1)$ is the OLS prediction of a given X_1 among observations with $Z_1 = 0$ with weights $1/(1-e_1(X_1))$, $\gamma_a^{0,k+}(X_1)$ is the OLS prediction of a at X_k given X_k among observations with $Z_1 = 0, A \geq k$ with weights $1/(\Pi_{j=1}^{A}(1-e_j(X_j)))$.

$\mathbb{E}[g_{Y_t}^1(G) - g_{Y_t}^0(G)] / \mathbb{E}[g_{D_t}^1(G) - g_{D_t}^0(G)]$ corresponds to a Wald estimator of $\tau_{LATE}(t)$ where the reduced form and the first stage are both implemented in a doubly-robust

manner. $\mathbb{E}[g^1_{Y_t}(G) - g^{0+}_{Y_t}(G)]/\mathbb{E}[g^1_{D_t}(G) - g^{0+}_{D_t}(G)]$ corresponds to the Wald-like estimator of $\tau_{ATR}(t)$, where the reduced form and the first stage are also implemented in a doubly-robust manner.

Following standard argument gives:

$$\frac{\mathbb{E}[g^1_{Y_1}(G) - g^0_{Y_1}(G)]}{\mathbb{E}[g^1_{D_1}(G) - g^0_{D_1}(G)]} = \tau_{LATE}(1),$$

and similarly, using the Lemma and the standard argument gives:

$$\frac{\mathbb{E}[g^1_{Y_1}(G) - g^{0+}_{Y_1}(G)]}{\mathbb{E}[g^1_{D_1}(G) - g^{0+}_{D_1}(G)]} = \tau_{ATR}(1).$$

After the first period, both estimators may be biased. In the second period:

$$\frac{\mathbb{E}[g^1_{Y_2}(G) - g^0_{Y_2}(G)]}{\mathbb{E}[g^1_{D_2}(G) - g^0_{D_2}(G)]} = \tau_{LATE}(2) + \frac{\Pr(C_2 = 0, C_1 = 1)}{\Pr(C_2 = 1)}\mathbb{E}[Y_2(1) - Y_2(2)|C_2 = 0, C_1 = 1]$$
$$\frac{\mathbb{E}[g^1_{Y_2}(G) - g^{0+}_{Y_2}(G)]}{\mathbb{E}[g^1_{D_2}(G) - g^{0+}_{D_2}(G)]} = \tau_{ATR}(2) + \frac{\Pr(R_2 = 0, R_1 = 1)}{\Pr(R_1 = 1)}\mathbb{E}[Y_2(1) - Y_2(2)|R_2 = 0, R_1 = 1].$$

The correction methods by Bensnes et al. (2023); Gallen et al. (2023) are valid when:

Assumption 7 (Parametric effects). *$Y_t(k) - Y_t(0) = \tau_{ATE}(1 + t - k)$ for all t and $k \leq t$.*

The assumption has two implications: first, the effects are homogeneous across individuals; second, the effects depend only on time spent in parenthood and not on the moment of becoming a parent. Gallen et al. (2023) discuss how the assumption can be relaxed; the relaxed version can be tested following similar steps.

Under parametric effects, the parameter identified by the Wald estimator in the second period simplifies to:

$$\frac{\mathbb{E}[g^1_{Y_2}(G) - g^0_{Y_2}(G)]}{\mathbb{E}[g^1_{D_2}(G) - g^0_{D_2}(G)]} = \tau_{ATE}(2) + \frac{\Pr(C_2 = 0, C_1 = 1)}{\Pr(C_1 = 1)}(\tau_{ATE}(2) - \tau_{ATE}(1))$$

Since under Assumption 7 $\tau_{ATE}(1) = \tau_{LATE}(1)$, and since $\tau_{LATE}(1)$, $\Pr(C_2 = 0, C_1 = 1)$, and $\Pr(C_2 = 1)$ are identified, $\tau_{ATE}(2)$ can be backed out. Following similar

reasoning for subsequent periods allows to back out $\tau_{ATE}(t)$ for all t.

My test for Assumption 7 uses the fact that $\tau_{ATE}(t)$ can also be backed out using the Wald-like estimates of $\tau_{ATR}(t)$, and that when the Assumption 7 holds, the two approaches should give similar results. To ease exposition, define the pseudo-outcome:

$$\widehat{Y_t^l} = \begin{cases} Y_t, \text{ if } D_1 = 1 \text{ or } D_t = 0, \\ Y_t - \tau^l(k), \text{ otherwise, where } k = 1 + t - (min\{j : D_j = 1\}) \ , \end{cases}$$

for $l \in \{C, R\}$, where:

$$\tau^C(t) = \frac{\mathbb{E}[g^1_{\widehat{Y_t}}(G) - g^0_{\widehat{Y_t}}(G)]}{\mathbb{E}[g^1_{D_1}(G) - g^0_{D_1}(G)]},$$

and

$$\tau^R(t) = \frac{\mathbb{E}[g^1_{\widehat{Y_t}}(G) - g^{0+}_{\widehat{Y_t}}(G)]}{\mathbb{E}[g^1_{D_1}(G) - g^{0+}_{D_1}(G)]}.$$

For women who become mothers in later periods, the pseudo-outcome is the realized outcome adjusted by subtracting the effect of being a mother for their motherhood duration, which is identified in previous periods. Under Assumption 7, the pseudo-outcome equals their control outcome. $\tau^C(t)$ corresponds to how $\tau_{ATE}(t)$ is identified using the Gallen et al. (2023) method based on $\tau^{LATE}(t)$. $\tau^R(t)$ corresponds to how it can be identified using $\tau_{ATR}(t)$. Under Assumption 7, $\tau^R(t) = \tau^C(t)$ for all t; if the two are not equal, at least one of the assumptions must be violated. Note that the only additional assumption that I require relative to Bensnes et al. (2023) and Gallen et al. (2023) is that the outcomes of subsequent ACPs are as good as random, conditional on observables.

Figure A18 presents the results for women's outcomes. Confidence intervals for the difference between the estimates in each period are estimated using Bayes bootstrap with weights $w_i \sim exp(1)$ and 150 draws, where all parameters are estimated sequentially in each draw. Estimates of $\tau^C(t)$ suggest a substantially smaller career cost of motherhood than $\tau^R(t)$, which indicates that the parametric effects assumption is violated.

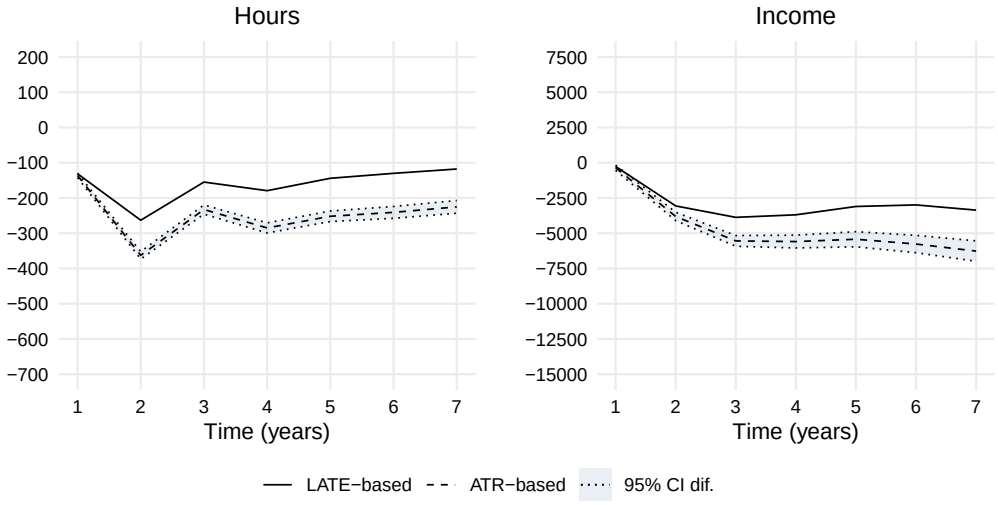

Figure A18: Estimates Using Parametric Bias Correction

A8 Selection Event Study

Section A8.1 discusses identification for the selection event study, and Section A8.2 discusses its implementation.

A8.1 Identification

To formally illustrate the ES approach in my model, I focus on women who conceive through their first ACP. The parallel trends assumption states that conditional on age and calendar time, control outcomes t periods after becoming a mother are the same as control outcomes in the period before becoming a mother, on average: $\mathbb{E}[Y_t(0)|age_t = a, year_t = y, Z_1 = 1] = \mathbb{E}[Y_0(0)|age_0 = a, year_0 = y, Z_1 = 1]$, for all t, a, y, where age_t and $year_t$ are the woman's age and calendar year in period t, respectively. This assumption allows for unbiased predictions of childless outcomes in period t based on women's age and calendar year in period t, using outcomes of women who were of the same age and in the same calendar year in period 0—just before becoming mothers. Comparing the realized treated outcomes in period t with these predictions gives the

average treatment effect:[36]

$$\tau_{ATE}(t) = \mathbb{E}[Y_t|Z_1 = 1] - \mathbb{E}[\mathbb{E}[Y_0|age_0 = age_t, year_0 = year_t, Z_1 = 1]|Z_1 = 1].$$

When the parallel trends assumption does not hold, the bias term for $\tau_{ATE}(t)$ is:

$$\mathbb{E}[Y_t(0)|Z_1 = 1] - \mathbb{E}[\mathbb{E}[Y_0(0)|age_0 = age_t, year_0 = year_t, Z_1 = 1]|Z_1 = 1].$$

It measures the difference in age- and year-conditional average childless outcomes between women who have their first child earlier versus later, reflecting selective fertility timing.

I quantify the extent of selective timing specifically for reliers. Formally, this procedure involves two steps. First, I identify the relier average control outcomes in the period before their first ACP, conditional on age and calendar year: $\mathbb{E}[Y_0(0) \mid age_0 = a, year_0 = y, R_t = 1]$. The identification procedure follows the same steps as for the relier average control outcome $\mathbb{E}[Y_t(0) \mid R_t = 1]$, except that all expectations are conditioned on pre-ACP age and calendar year, and the realized labor market outcome in period t is replaced by that in period 0. Then, as in the ES approach, I use these conditional average control outcomes in period 0 to construct age- and calendar-year-conditional predictions for control outcomes after t periods. The difference in the relier average childless career trajectories from the baseline approach and the constructed ones quantifies the extent of selective timing, specifically for reliers:

$$\mathbb{E}[Y_t(0)|R_t = 1] - \mathbb{E}[\mathbb{E}[Y_0(0)|age_0 = age_t, year_0 = year_t, R_t = 1]|R_t = 1].$$

This allows for a comparison with τ_{ATR}, making it possible to distinguish, for a consistent group, how much of the gender inequality associated with parenthood is driven by the effect of parenthood itself versus selective timing.

[36]The inner expectation is over Y_0, the outer expectation is over age_t and $year_t$. The inner expectation may not be well defined when the support of age_0 and $year_0$ may differ from that of age_t and $year_t$.

A8.2 Implementation

I implement the selection ES using (19) on a sample of women who remain childless seven years after their first ACP, with weights $w_i^w = 1/\Pi_{j=1}^{A_{i7}}(1 - e_j(X_{ji}))$ to ensure proportional representation of reliers with different willingness to undergo ACPs. Standard errors are estimated via Bayes bootstrap as follows: (1) draw weights $w_i \sim \exp(1)$, (2) estimate $e_j(x_j)$ for all j using weights w_i to obtain an estimate of w_i^w, (3) estimate the selection ES using weights $w_i w_i^w$, (4) repeat steps 1–3 150 times to obtain bootstrap estimates, (5) compute their variance.

Share of gender inequality due to parenthood and selective fertility in year t estimated in the following steps: (1) construct separate sample moments for the control mean and the upper and lower bounds for the treated mean (a_1, a_2, and a_3, respectively), where Y is the female labor market outcome subtracted from the male labor market outcome in period t, (2) implement the selection ES using the female labor market outcome and age, repeat it for the male labor market outcome and age, obtain the estimate for period t, a_4, by subtracting the female estimate for period t from the male estimate for period t, (3) construct the bounds $(a_2 - (a_1 + a_4))/a_2$ and $(a_3 - (a_1 + a_4))/a_3$. Confidence intervals are based on Stoye (2020), with covariance matrices estimated using Bayes bootstrap and delta method in the following steps: (1) estimate $\widehat{\xi}_i$ using cross-fitting where Y is the female labor market outcome subtracted from the male labor market outcome, (2) draw weights $w_i \sim exp(1)$, (3) implement the selection ES with weights w_i using the female labor market outcome and age, repeat it for the male labor market outcome and age, obtain the estimate a_4 for the difference between the male and the female estimates, (4) construct separate sample moments for the control mean and the upper and lower bounds for the treated mean evaluated at $\widehat{\xi}_i$ with weights w_i (a_1, a_2, and a_3, respectively), (5) repeat steps 2-4 150 times to obtain a collection of bootstrap estimates, (6) estimate the joint covariance matrix of a_1, a_2, a_3 and a_4, (7) obtain the joint covariance matrix for $(a_2 - (a_1 + a_4))/a_2$ and $(a_3 - (a_1 + a_4))/a_3$ using delta method.

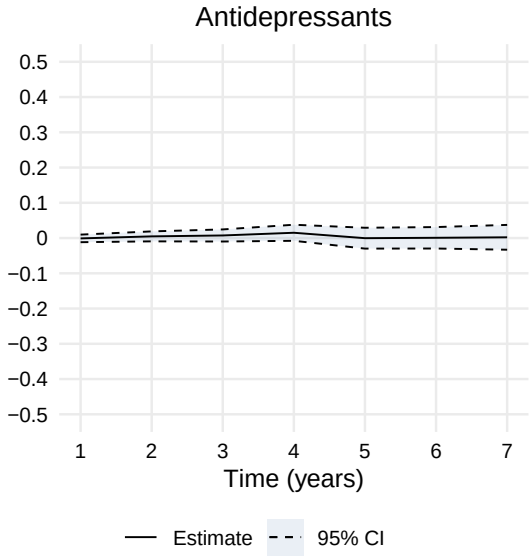

Figure A19: Effects on Antidepressant Uptake

A9 Antidepressants

To maximize precision in estimating the impact on antidepressant uptake, I use the method described in Section A7 with target parameter:

$$\frac{\mathbb{E}[g^1_{Y_t}(G) - g^{0+}_{Y_t}(G)]}{\mathbb{E}[g^1_{D_t}(G) - g^{0+}_{D_t}(G)]},$$

where the outcome is taking antidepressants in a given year. In the absence of dynamic effects, it identifies τ_{ATR}. Figure A19 presents the results, the effects are precisely estimated and indistinguishable from zero.

A10 Monotonicity

Monotonicity states that all reliers are subsequent reliers, implying that the relier share is at least as large as the subsequent relier share: $\Pr(R^+ = 1) \geq \Pr(R = 1)$. Figure A20 plots the estimated shares over time, showing that the subsequent relier share consistently exceeds the relier share, in line with monotonicity.

Monotonicity further implies that the relier share is at least as large as the sub-

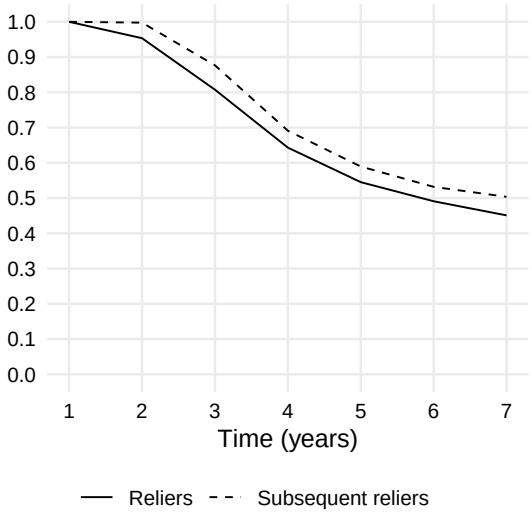

Figure A20: Estimated Relier and Subsequent Relier Shares

sequent relier share at each covariate value: $r^+(X_1) \geq r(X_1)$. Since the conditional shares are estimated nonparametrically, formally testing whether their differences allow rejecting monotonicity is not trivial, but comparing them offers some insight. The top left graph in Figure A21 plots the empirical distribution of the difference between estimated conditional subsequent relier and relier shares in year 7. For 25% of observations, the difference is below zero. While this would contradict monotonicity if observed in the true shares, such deviations can result from estimation error when the shares are close. Namely, when all subsequent reliers are reliers $\Pr(R^+ = R) = 1$, the difference should be below zero for 50% of observations. Consistent with this, the differences are generally small, with only 5% of observations below -0.1, suggesting no clear monotonicity violations.

The right graph in Figure A21 repeats the above for the partial monotonicity assumption, which permits violations among women who would separate from their partners or uptake antidepressants after ACP failure. The estimated difference between the two shares is below zero for only 5% of observations and below -0.1 for just 1%, providing stronger support for partial monotonicity. The second-row panels in Figure A21 repeat this analysis, allowing monotonicity to fail separately for those who would uptake antidepressants or separate, yielding similar results. Equivalent

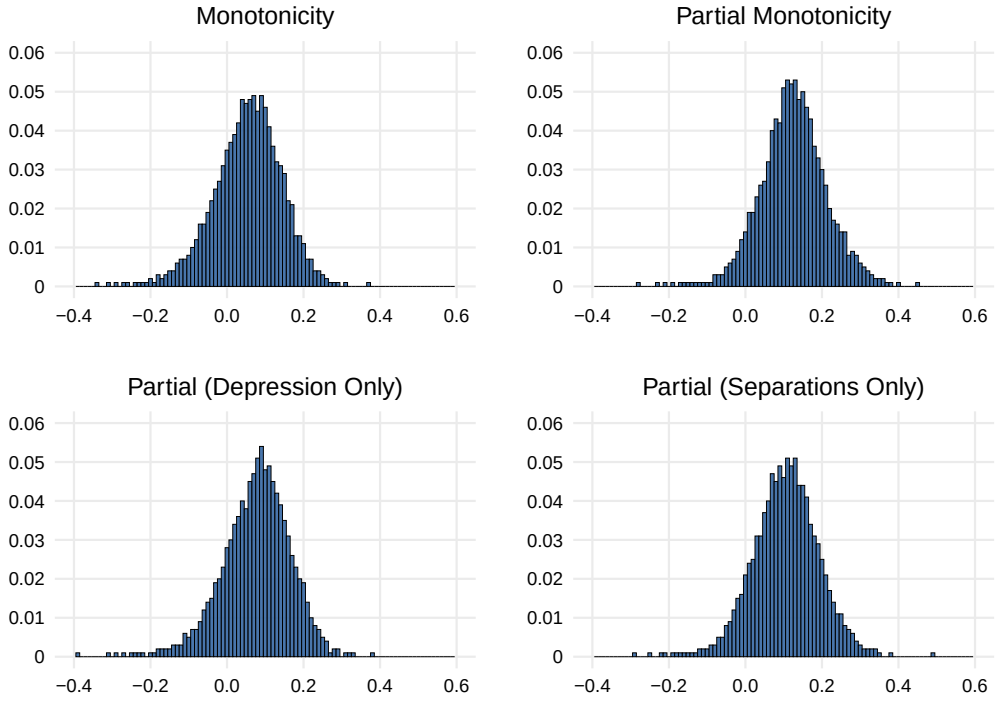

Figure A21: Histogram of Estimated $r^+(X_1) - r(X_1)$

results from earlier years are even more favorable for the assumption.

To formally test monotonicity using covariates, I adapt the approach of Semenova (2023). I partition the sample into $J = 25$ discrete cells C_j based on quintiles of women's work hours and age in the year prior to their first ACP. Since these two covariates are highly predictive of the remaining covariates used in the analysis, including additional ones results in small cells (e.g., almost no women have extremely high work hours while having extremely low income). Monotonicity implies that the subsequent relier share is at least as large as the relier share in each cell, meaning that each value in the vector $\mu = (\mathbb{E}[r^+(x) - r(x) \mid x \in C_j])_{j=1}^J$ must be non-negative. The null hypothesis is $-\mu \leq 0$, and the test statistic is $\max_{j \in J} \frac{-\widehat{\mu}_j}{\widehat{\sigma}_j}$. The critical value is the self-normalized critical value of Chernozhukov et al. (2019). Consistent with the results in Figure A21, in 24% of cells, $\widehat{\mu}_j$ in year 7 is negative. However, the p-value for the test statistic is close to 1, indicating that these differences are not statistically significant, providing support for monotonicity. Using women's income instead of hours yields similar results.

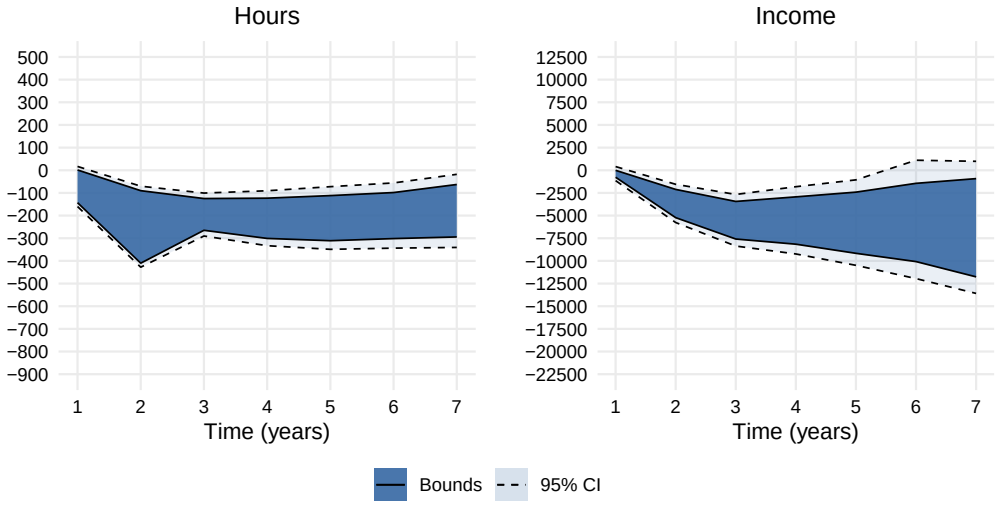

Figure A22: Effects on Resilient Women (Depression Only)

In Section A3.1, I discuss how my estimation method accounts for reversed relier and subsequent relier shares, propose an approach that relaxes monotonicity to allow the direction to vary with covariates, and present corresponding estimates, which align with my baseline results.

Finally, Figure A22 presents estimated labor market impacts of parenthood for reliers who would not uptake antidepressants after ACP failure (but might separate from their partner). The estimates are close to the baseline results.