Evaluating the Impact of Large Scale Social Programs

with an application to labor supply and child benefits

Filip Premik*
Monash University and FAME|GRAPE

November 27, 2024

Abstract

I propose a method of evaluating universal government programs in which the lack of appropriate control groups limits the usability of conventional frameworks. The method accentuates the researcher's role in making an identification argument and supports it with a flexible choice model estimated nonparametrically using recent advances in machine learning. I apply my framework to estimate the effects of a large-scale child benefit program on the female labor supply. Aggregating impacts on labor market flows, I show that the program led to a 2–3 percentage points decrease in labor supply among eligible women, driven by discouragement among job seekers.

Keywords: large-scale government programs, program evaluation, child benefit programs, female labor supply, generalized random forest

*filip.premik@monash.edu. The author thanks Joanna Tyrowicz without whose support and effort this paper would have never been written, as well as multiple discussants whose comments improved its quality. The author gratefully acknowledges the support of the Polish National Science Centre (grant #2016/23/N/HS4/03637).

1 Introduction

Large-scale government programs have become increasingly prevalent worldwide, particularly since the COVID-19 pandemic and subsequent energy price crisis. Given that these untargeted or broadly targeted transfers and subsidies often constitute a significant cost to the taxpayers, understanding the effects of such policies is of paramount interest. However, broad eligibility implies that an appropriate control group may be difficult to define, hindering the utilization of the standard program evaluation methods. Even if an ineligible group exists, its behavior is rarely a good counterfactual for the behavior of the eligible group in the absence of the program due to significant differences in the decision processes among individuals related to various stages of life-cycle, systematic heterogeneity in unobservable characteristics or incentives motivating their decisions.

In this paper, I propose an alternative approach to evaluate large-scale government programs. It follows the same logic as traditional methods in first identifying the variation in the evolution of the outcome of interest attributable to the program and second choosing suitable counterfactual realizations of the outcome variable in the absence of treatment. However, instead of imposing rigid one-fit-all assumptions, it relies on the informed choices of a researcher in fulfilling the two tasks and producing a convincing identification argument.

A flexible choice model supports the researcher in identifying variation caused by the program by providing a counterfactual decomposition of the evolution of the outcome variable. It separates dynamics resulting from changes in an individual's decision rule (driven by relative payoffs and future beliefs) from the intertemporal shift in the composition of individual characteristics (summarizing self-selection), which can then be compared across groups of individuals and over time to determine the variation attributable to the program. Next, the researcher shuts it down by choosing a suitable counterfactual that mimics the counterfactual decomposition obtained for another group of individuals. Alternatively, it may also be inspired by economic theory or other application-specific knowledge. Lastly, an implied path of the outcome of interest is simulated using the proposed counterfactual variation that excludes the program's impacts. Comparing the simulated and realized paths of the outcome of interest yields the desired effect.

I use my framework to evaluate the effects of a universal large-scale child benefit program in Poland on the female labor supply. The program – *Family 500 Plus* (henceforth: *P500*, or the *intervention*) – costs approximately 2% GDP yearly and provides transfers to families raising kids. From 2016 on, families raising children receive a monthly nonequivalent transfer of approximately 20% of the median wage per second and any further child, and additional criteria define the eligibility of the first child. Given their monetary amount, the benefits are unlikely to provide a substitute for labor income at a large scale. Instead, the program is expected to affect the marginal individuals predominantly. To better capture this environment feature, I focus on labor market flows, studying women's choices to enter or leave the labor force.

The estimates of the decision model suggest that the child benefits decreased female labor force participation mainly by discouraging the activation of women outside of the labor force. The discouragement was driven by changes in perceived trade-offs and future beliefs. These direct effects have been propagating and accumulating over time, leading up to a 1.74 percentage point drop in the labor force participation rate among eligible women in two years and up to a 3.37 percentage point drop in four years after the program's introduction.

The decrease in labor force participation may have also affected employers who improved the working conditions facing increased difficulty maintaining staff. If this is true, one would expect an implied decrease in outflow rates driven by changes in the economic environment, which is confirmed by the model decomposition. Removing this effect in a counterfactual labor force participation path simulation counteracts the further propagation of initial shocks resulting in the estimate of up to a 2.7 percentage point drop in the labor force participation among the eligible women after four years since the program's introduction.

Even though I focus primarily on changes in labor force participation among Polish women in response to the program *P500*, the methodology proposed in this paper generalizes easily to study many other universal or near-universal government programs as long as the researcher is able to specify an underlying individual decision problem. The decision model shares the flexibility of reduced form approaches (by conditioning on potentially a large set of state variables and not requiring functional form assumptions) and benefits from the appealing interpretation of the underlying choice model (which is a typical aspect of structural models) without the necessity to impose restrictive assumptions on expectations or future evolution of state variables. It is estimated non-parametrically and provides a convenient framework to inform better researcher's choice of which variation is likely attributable to the program. It can effortlessly exploit large amounts of information regarding individuals' socioeconomic backgrounds to approximate individuals' decisions as closely as possible. Datasets including such information are increasingly available for researchers evaluating the impacts of large-scale government policies.

My paper speaks to a few strands of the literature. The first two focus on evaluating large-scale government programs using reduced-form and structural approaches, respectively. I also relate to papers using machine learning methods in studying labor market outcomes and add to the prior research on *P500* in Poland.

In related studies concerning reduced-form large-scale program evaluations, Schirle (2015); Koebel and Schirle (2016) show that the Canadian Universal Child Care Benefit decreases the labor supply of married women. Baker, Messacar, and Stabile (2021) provide an overview of a few Canadian child benefit system reforms showing a reduction in child poverty and no evidence of labor supply response on both extensive and intensive margins. González (2013) investigates a universal child benefit program in Spain and finds a decrease in the maternal labor force after childbirth. A common denominator of these studies (and many more, for a review of literature evaluating the labor supply effects of child benefits and other family-related welfare programs see Moffitt, 2002; Immervoll, Kleven, Kreiner, and Saez, 2007) is the reduced form approach taken as a tool to describe changes in labor supply as a result of a benefit program. In my paper, I explicitly model women's decision rules, which allows me to avoid the restrictive assumptions regarding the data-generating process associated with the potential outcomes model. In addition, my results have an appealing interpretation of a micro-founded model.

Another strand of the literature uses structural modeling as a tool to evaluate the impacts of large child support programs. Blundell, Duncan, McCrae, and Meghir (2000) study Working Families' Tax Credit program in the UK. Using a structural labor supply model with childcare costs, they showed increased labor force participation as a response to the program. Stephens Jr and Unayama (2015) investigate the effects of the Japanese child benefit system on household wealth accumulation. A fully specified structural model requires several assumptions regarding agent expectations and the law of motion of state variables. There are also computational constraints limiting the number of state variables. In turn, my simple choice model framework does not impose strong assumptions on the structure of the decision problem. Moreover, it uses machine learning techniques that allow to condition women's decisions on many observed state variables tractably.

My paper applies machine learning methods to study labor force participation. In a related setting, Cengiz, Dube, Lindner, and Zentler-Munro (2021) uses similar tools to predict which individuals are likely to be affected by the minimum wage reforms. Sigurdsson (2019) applies forest-based estimators in studying labor supply responses to temporary variation in wages, exploiting exogenous variation in a tax cut. Angrist and Frandsen (2019) study performance of machine learning algorithms in causal studies, illustrating it with an example concerning effects of college characteristics on wages. A common denominator of these studies is that they focus on recovering causal parameters in the potential outcomes framework. My study differs from these papers by using the machine learning algorithm to estimate a flexible structural choice model and then simulate counterfactual decisions.

My study adds to the discussion concerning the effects of the program *p500*. Myck (2016) and Myck and Trzciński (2019) utilize a microsimulation model to evaluate ex-ante potential effects of *P500*. Their model relies on a discrete choice model of labor supply in which a household with two adults chooses labor supply for both of them. Results indicate a drop in the labor force supply of roughly 150 thousand, or approximately 2%, of economically active women, which is similar to my findings concerning early-stage program evaluation. The simulations are obtained in the short-run and partial equilibrium, ignoring potential changes in the wage structure and working conditions. My approach allows for implicit consideration of these effects. Magda, Kiełczewska, and Brandt (2018) use a difference-in-difference approach to provide an early evaluation of the effects on the female labor supply. Their identifying assumption relies on the short term between the introduction of the program and the measurement of their effects. They find treatment effects implying a 2-3 percentage points drop in the female labor force supply as a result of introducing the program, which is consistent with my estimates of the initial effects of the program.

The remainder of the paper is structured as follows. Section 2 describes the policy design. Section 3 reviews the data. Section 4 presents the method and explains its use in the program evaluation exercise. Section 5 introduces details of the estimation routine. Empirical results are discussed in section 6. Section 7 concludes.

2 The Program Family 500Plus

The program Family 500Plus provides a universal child benefit for each second and further child aged 0-17 in a household. Until 2019, the first child in the household was also eligible if the household's income was below a threshold. The program has been extended to all children in the household in 2019. The benefit comes as a monthly non-equivalent payment of roughly 20-25% of the net average wage (PLN500, or approximately US$130) per eligible child. The program's main goals are to improve the financial well-being of families raising children and stimulate fertility in the long run. The program constitutes a significant financial effort to the government budget, at the cost of approximately 1.5-2% of GDP yearly.

An eligible household should apply for the benefit at the local administration to obtain it. Table 1 indicates high participation rates among the eligible. Approximately 95% of households with two or more children below 18 are obtaining the benefit. Extending eligibility for the first child in 2019 shifted the participation rate among single-child households to a comparable level.

Table 1: Program Participation

2 or more children	1 child	
2017Q1 - 2019Q4	2017Q1-2019Q1	2019Q4
.948	.234	.899

The program was announced in the first quarter of 2016, and the first payments arrived in the next quarter. However, some regions experienced delays in the distribution of the initial payments, which were eventually distributed in the second half of 2016. I divide the timeline into four general periods. The pre-intervention period includes all years up to

2015. At that time, any influence of the program can be ruled out. I refer to 2016 as a transition year because that was a period in which the program was announced and gradually introduced. Analogously, the second transition period occurred in 2018Q4-2019Q3, covering the official announcement and implementation of the program expansion to cover all children. The remaining periods belong to the post-intervention period.

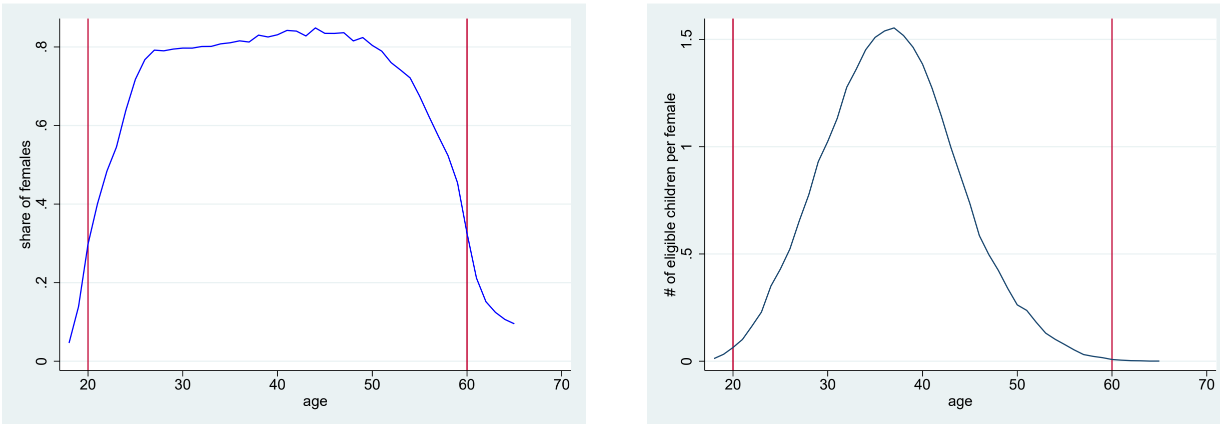
3 Data

Data comes from the Labor Force Survey conducted by the Central Statistical Office in Poland. Approximately 30,000 households are interviewed each quarter using a detailed questionnaire concerning their labor market outcomes. The sample is representative of the population and constitutes a rotating panel. Each household is interviewed 4 times. The first two waves are collected in two consecutive quarters. The third wave is collected after a year after the first, and the fourth follows in the quarter right after the third. For example, if a household enters the sample in 2016Q1, then it is re-interviewed in 2016Q2, 2017Q1, and 2017Q2. In each wave of the survey, the responses of all adult members of the household are recorded.

I restrict my attention to the subsample of females of age between 20 and 60. Typically, individuals in Poland leave the schooling system in the 19th year after birth. The lower threshold allows me to abstract from schooling and birth date effects. In turn, Polish women are eligible for retirement at the age of 60, which motivates the choice of the upper threshold. Figure 1 illustrates this reasoning. Most of the economic activity and child upbringing are performed by women not younger than 20 and not older than 60.

I distinguish three groups of women. First, females with two or more children below 18 (shortly: $\geq$2 children) are eligible to receive the benefit for at least one child. Second, women with one child below 18 (shortly: 1 child) are all eligible after the program expan-

Figure 1: Labor force participation and children bearing in the life cycle.



(a) female labor force participation

(b) number of children below 18 per female

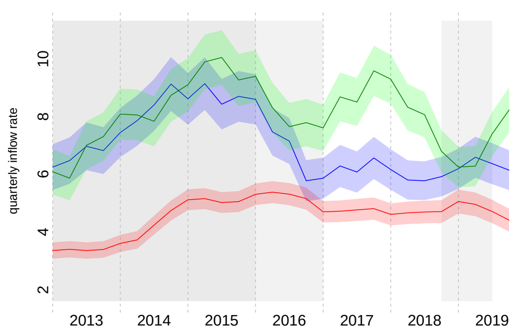
sion. Third, females who do not have children below 18 (shortly: childless) and therefore are not eligible for the benefit. This allows me to keep track of the differential impacts of the program on the participating and non-participating individuals and better inform the identification arguments.

Since nearly all eligible women participate, I focus on the distinction between eligible and ineligible females. This is motivated by the fact that, given the short panel dimension in my data, it is impossible to determine which of the eligible individuals observed before the program's introduction would participate.

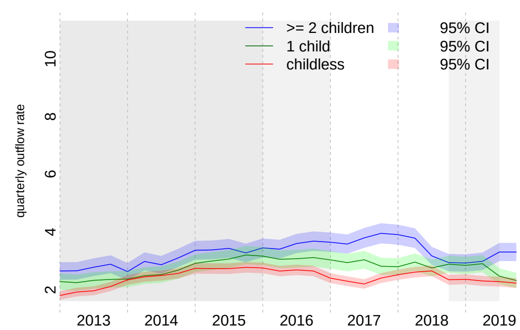
3.1 Labor Market flows

In this paper, I identify the effects of child benefits on labor force participation through changes in the labor market flows. A woman is a member of the labor force in a given period if she works or is actively searching for a job. Labor force participation is determined by labor market flows. Given the structure of my data, I measure inflows in a given period as the share of females in the labor force who were not there in the previous period. Analogously, I define outflows in a given period as the share of women who are not currently in the labor force but were there in the previous period. Figure 2 depicts the quarterly time series of the flows and stock of labor force participation. Differences in their dynamics suggest that other economic processes drive the inflows than the outflows. In addition, the child benefit program is likely to affect predominantly individuals at the margin. The variation in labor market flows (as opposed to the levels) is also less subject to trends and impacts from sources other than the program. All of these factors motivate the focus on the flows in analyzing the program's impacts.

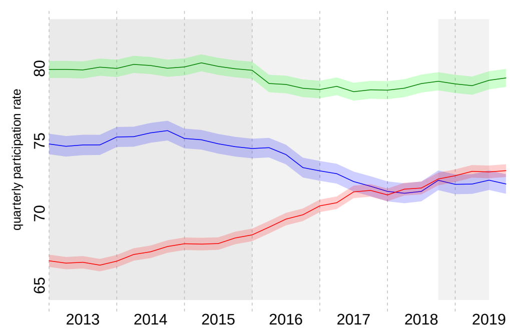
Figure 2: Quarterly labor force indicators by program eligibility.



(a) y_q^0: rate of inflows to labor force.



(b) y_q^1: rate of outflows from labor force.



(c) y_q^1: labor force participation.

Note: Note: All values expressed as percentage points. Dark gray background indicates pre-intervention periods, light gray indicates the transition periods.

Before the *P500* was introduced, the inflows for all the groups followed roughly the same trend, which mostly stabilized after 2016, as shown in figure 2. The introduction of *P500* has coincided with a significant drop in the inflows among women with two or more children below 18 and a moderate decline among women with one child below 18. This probably reflects the fact that not all women with one child below 18 were eligible in 2016. In turn, 2019 brought an additional drop in the inflow rate among women with one child following the extension of the eligibility.

The average inflows among the eligible females decreased by 2.7 percentage points after the program's introduction, as shown in table 2. The change in inflows among the ineligible is also negative but smaller in order of magnitude. Changes in the outflows among eligible and ineligible females are of opposite signs but low magnitude. However, the measure of participating females is much larger than inactive. Therefore, relatively smaller changes in the outflows may translate into larger shocks to the aggregate labor supply.

Table 2: Labor force participation - inflows and outflows.

	>=2 children		1 child		childless	
	inflows	outflows	inflows	outflows	inflows	outflows
post-int. (2017-2019)	6.202	4.648	2.672	8.258	3.393	2.672
pre-int. (2014-2015)	8.901	4.893	2.940	8.997	3.188	2.940
difference	-2.699	-.245	-.268	-.738	.205	-.268

Note: Note: Inflows and outflows expressed in percentage points.

Figure 2 also presents trends in labor force participation levels. The participation rates of women with two or more children remained roughly constant in the pre-intervention period, dropped by approximately 3 percentage points in 2016-2017, and stabilized at the end of the observation window. The share of active women with one child below 18 remained relatively constant throughout the sample. Participation rates among females without children below 18 steadily increase across the sample window. Capturing how much of this variation can be attributed to the program *p500* constitutes a crucial part of the empirical exercise described in the following sections.

3.2 Predictors of Female's Labor Market Decisions

The data includes a detailed description of the household's socio-economic background and labor market activities. This information is crucial for predicting female labor market decisions and evaluating the program's effects *P500*. I classify available variables into several groups.

First, I consider a set of household-level variables describing the household composition, number of earners, basic demographics, and month in which the interview has been taken - a particularly important covariate that controls for seasonal variation in labor force participation. Second, I consider a range of demographic characteristics of the woman, including age, marital status, dummies for their spouse, parents, and children's presence in the household. This group also contains the number of children below 18, which is a fundamental variable driving the program eligibility. The third group describes female's employment status. It provides a comprehensive description of the job (for the employed), reasons for not having a job (for the jobless), and past working experience. The fourth group summarizes a woman's job search effort (including intensity, duration, and type of searched job), and the fifth describes her educational background.

In this paper, I focus on women's decisions, which are likely to also depend on the outcomes of other household members. For example, they typically share responsibility for financial well-being with their spouses. To account for that, I consider another group of variables that describe spouse's outcomes, conditional on their presence in the household. Specifically, I include the spouse's employment situation, job search, and educational background. Female's decisions may also depend on parental support. Guided by this fact, I consider a subset of variables describing the mother's and father's sources of income and subjective evaluation of their labor market status.

Table 3: Choice of the Observed State Variables – Summary

	female	husband	mother	father
household level covariates	✓			
female demographics	✓			
employment status	✓	✓	✓*	✓*
job search	✓	✓		
education	✓	✓		

Note: Note: * only a selection of variables in the group is chosen.

Table 3 summarizes available predictors of women's labor market decisions. I take a set of 379 observed state variables to the estimation. Section B in the online supplementary materials provides a detailed description of these variables.

4 The method

In this section, I present the analytical tool designed to study the effects of a large-scale government policy. I start with defining a flexible decision model to explain the flows in the labor market. Predictions of the model are then used to derive a counterfactual decomposition of the evolution of the flows that allows me to identify the variation in the data likely attributable to the program. In the next step, shutting down this variation, I simulate a counterfactual path of aggregate labor supply in the absence of the program. Comparing the realized path with the simulated one quantifies the program's effects.

4.1 Decision model

In period t, a woman chooses $y \in \{0, 1\}$ conditionally a finite set of state variables that are known by her at the time the decision is taken. I model inflows into and outflows from the labor force separately. In analyzing inflows, $y = 1$ describes a woman's decision to enter the labor force. In analyzing outflows, $y = 1$ denotes her decision to leave the labor force.

There are two distinct types of state variables: observed by both woman and econometrician (s, henceforth the *observed* state variables) and observed only by a woman (ε, henceforth the *unobserved* state variables). I assume that the latter are drawn from a joint distribution $F_t(\varepsilon|s)$ with finite first moments.

A woman's payoff function in period t is $v_t(y, s, \varepsilon)$. It depends on her choice, values of the state variables, and period t itself—to emphasize that the payoffs may vary across periods as influenced by macroeconomic conditions and changing beliefs about the future. This specification implicitly allows for an arbitrary scheme of discounting future outcomes and beliefs. The value of the decision problem at time t can be written as:

$$V_t(s, \varepsilon) = \max_{y \in \{0,1\}} \Big\{ v_t(y, s, \varepsilon) \Big\} \tag{1}$$

and gives rise to the optimal policy function:

$$y_t(s, \varepsilon) = \mathbf{1}\Big[v_t(1, s, \varepsilon) \geq v_t(0, s, \varepsilon) \Big] \tag{2}$$

Using assumptions of the model, I derive the probability that the woman chooses $y = 1$ conditionally on the set of observed state variables by integrating out the unobserved state variables:

$$\mathrm{P}[y = 1|s, t] = \int\limits_{\varepsilon : v_t(1,s,\varepsilon) - v_t(0,s,\varepsilon) \geq 0} \mathrm{d}F_t(\varepsilon|s) \quad \equiv \quad \varrho_t(s) \tag{3}$$

$\varrho_t(s)$ is a conditional choice probability and describes woman's decision rule given s.

4.2 Decomposing differences in conditional choice probabilities

For any two periods $t \in \{0,1\}$, the expected change in woman's labor force participation decisions can be decomposed into two counterfactual elements:

$$\mathbb{E}_{\varepsilon}\big[y_1(s_1,\varepsilon_1) - y_0(s_0,\varepsilon_0)\big] = \varrho_1(s_1) - \varrho_0(s_0) = \\ = \underbrace{\varrho_1(s_0) - \varrho_0(s_0)}_{\beta(s_0)} + \underbrace{\varrho_1(s_1) - \varrho_1(s_0)}_{\gamma_1(s_1,s_0)} \tag{4}$$

First, $\beta(s_0)$ describes changes in woman's conditional choice probabilities between pre- and post-intervention periods holding fixed the pre-intervention vector of observed state variables. It summarizes the inter-temporal changes in the functional form of the within-period payoff functions, including the effects of changes in individual beliefs regarding the future. For example, consider a woman who is unemployed but actively searches for a job in the pre-intervention period and has two or more children below 18 (that means she is eligible for receiving the benefit). In the post-intervention period, the additional income from the program *p500* may magnify the significance of disutility from a potentially costly job search process in the per-period payoff because the salary income becomes less necessary to sustain the household. If the woman expects the transfers to arrive regularly in the future, her probability of continuing job search would drop significantly without a change in any of the state variables s. I refer to the parameter $\beta(s_0)$ as to *decision rule* effect to accentuate its dependence on the evolution of conditional choice probabilities.

Second, $\gamma_1(s_1, s_0)$ describes changes in woman's conditional choice probabilities between pre- and post-intervention periods holding fixed the decision rule. This parameter describes changes resulting from the inter-temporal shift in individual characteristics. By adjusting elements of s_1, a woman may increase her probability of receiving a benefit, affecting her labor force participation. For example, consider an ineligible woman with one child. Suppose she derives high utility from staying out of the labor force and raising her child. If she believes that the benefit program will be sustained in the long term, she may decide to give birth to another child, self-selecting to the program. The benefit would provide additional financial means that would lower her probability of being in the labor force without changing the functional form of her payoff function. I label $\gamma_1(s_1, s_0)$ the *composition* effect to emphasize its dependence on the shift in observed characteristics.

The decomposition is similar to the Oaxaca-Blinder decomposition to some extent. However, I focus on the changes in the decision rules and individual characteristics. In addition, my decomposition is free of functional form assumptions.

4.2.1 Sample decomposition

The decomposition given by equation (4) is complete if the researcher can observe the true optimal policy functions $\varrho_t(\cdot)$. This object has to be estimated from the data in a real-world setting. In finite samples, there may not be enough variation to average out the impact of unobserved state variables fully. Define the resulting error as:

$$\hat{\xi}(s_1, s_0) \equiv \big(\overline{y}_1(s_1) - \overline{y}_0(s_0)\big) - \big(\hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0)\big) \tag{5}$$

where $\overline{y}_t(s_t)$ is the sample average of the outcome variable at time t among individuals with realization of state variables s_t. This error (henceforth the *residual* parameter) refers to the variation in the unobservables and enables writing down the exact decomposition of sample averages:

$$\begin{aligned} \overline{y}_1(s_1) - \overline{y}_0(s_0) &= \hat{\varrho}_1(s_1) - \hat{\varrho}_0(s_0) + \hat{\xi}(s_1, s_0) = \\ &= \underbrace{\hat{\beta}(s_0)}_{\text{decision rule}} + \underbrace{\hat{\gamma}_1(s_1, s_0)}_{\text{composition}} + \underbrace{\hat{\xi}(s_1, s_0)}_{\text{residual}} \end{aligned} \tag{6}$$

If the model specification reflects the true data-generating process, the variation in observed choices resulting from $\hat{\xi}(s_1, s_0)$ should be negligible. That suggests a simple specification test with the null hypothesis $H_0 : \hat{\xi}(s_1, s_0) = 0$.

The parameters of 5 provide the basis for identifying the effects of the program. In the empirical part, I estimate quarterly time series of the decomposition and average obtained results over nearly all variables in s, distinguishing solely between the broad categories defined by program eligibility.

4.3 Identification arguments

Any evaluation of a large government program aims first to identify the variation in the outcome of interest attributable to the program and then compare it to a counterfactual outcome that would be realized in the absence of the program. This is often a challenging task.

The traditional approaches typically assume that all changes in the outcome of interest past the introduction of the program among units receiving the treatment are attributable to the intervention and require defining a control group whose behavior would be assumed to be also the behavior of the treated units if the intervention does not occur. My method relaxes the traditional rigid assumptions in a few dimensions.

First, the decomposition 6 offers insights into qualitatively different sources of variation in the evolution of the outcome of interest in terms of counterfactual changes in the decision rule and composition of state variables characterizing individuals. The researcher's role is to point out which part of the variation is attributable to the program. For example, a result of this stage would be a claim that the non-zero value of the decision rule parameter in the model of inflows among women with two or more children after introducing the benefits results from the program. Allowing the researcher to decide which part of the variation is likely driven by the program gives them an advantage of using the knowledge of the economic environment, specifics of the program, and economic theory to deliver a convincing story motivating the identification of the program's effects. In particular, the researchers are not limited solely to analyzing the behavior of the treated group but also may easily accommodate general equilibrium effects affecting units that were not directly subjected to the intervention.

Second, having identified the variation reflecting the effects of the program, the researcher has to choose an appropriate counterfactual for each of its instances. Depending on the application, average behavior among units of group not subjected to the intervention may be a good counterfactual. However, this is not the only choice. For example, a suitable counterfactual may come from assuming a steady state value for the outcome of interest (for example, zero values for the parameters of the decomposition of labor market flows) or the average variation in periods preceding the intervention.

My approach's construction of the identification argument is application-specific and gives the researcher a lot of flexibility. It does not require but also does not preclude the existence of suitable control groups. Moreover, for different parts of variation, the behavior of various groups may be used to form a counterfactual. The counterfactual variation can also be derived from economic restrictions on the underlying model or other knowledge available to the researcher. All of these may seem somewhat arbitrary, but in many cases, they are actually less arbitrary than rigidly assuming the entirety of the one-fits-all standard assumptions. To decrease the arbitrariness, the researcher should always explain in detail the reasoning for choosing a specific variation and a suitable counterfactual. In addition, the researcher may consider a few different scenarios, ranking them according to the strength or credibility of the underlying assumptions.

The result of this stage is a set of new time series of $\tilde{\beta}(s_0), \tilde{\gamma}_1(s_1, s_0)$ that describe what should be the values of counterfactual parameters of the decomposition 6 in the absence of the program.

4.4 Simulation

The law of motion of labor force participation y_t is given by:

$$\mathrm{P}[y_t = 1] = \underbrace{\mathrm{P}[y_t = 1 | y_{t-1} = 0]}_{\text{inflow rate}} \cdot \mathrm{P}[y_{t-1} = 0] + \underbrace{\mathrm{P}[y_t = 1 | y_{t-1} = 1]}_{\text{(negative) outflow rate}} \cdot \mathrm{P}[y_{t-1} = 1] \tag{7}$$

The flow rates can be further decomposed into:

$$P[y_t = 1|y_{t-1} = y] = \underbrace{P[y_{t-1} = 1|y_{t-2} = y]}_{\text{lagged flow rate}} + \underbrace{P[y_t = 1|y_{t-1} = y] - P[y_{t-1} = 1|y_{t-2} = y]}_{\text{change in flow rate}} \quad (8)$$

for $y \in \{0, 1\}$. Given the decomposition (6), the change in flows can be written as:

$$\text{change in flow rate}_{t,t-1} = \beta^{flow}(s_{t-1}) + \gamma_1^{flow}(s_t, s_{t-1}) + \xi^{flow}(s_t, s_{t-1}) \quad (9)$$

Substituting for the counterfactual time-series of $\tilde{\beta}(s_0)$ and $\tilde{\gamma}_1(s_1, s_0)$, I simulate forward the dynamics of the aggregate. The result is an implied path of aggregate labor force participation in each group had the program not been introduced. The simulated path is then compared to the original one to quantify the aggregate effects of the program.

5 Estimating the Conditional Choice Probabilities

The conditional choice probabilities $\varrho_t(s)$ reflecting women's decision rules at different values of state variables and time are key primitives of the model that need to be estimated from the data. For each t and s, they are point-identified through a conditional moment restriction:

$$\mathbb{E}\Big[y - \varrho|t, s\Big] = 0 \quad (10)$$

Estimation based on conditional moment restriction is often subject to the curse of dimensionality, which effectively limits the analysis to very few state variables. I estimate conditional choice probability function $\varrho_t(\cdot)$ using the Generalized Random Forest estimator developed by Athey, Tibshirani, and Wager (2019, GRF) that allows me to condition women's decisions on a large set of state variables without facing the curse of dimensionality. Although GRF relies on a machine learning algorithm, it is shown to produce consistent and asymptotically normal estimates of the conditional choice probabilities. This appealing feature enables statistical inference, making GRF particularly suitable for applications in applied economics. Section A in the online supplementary materials contains an intuitive description of the mechanics behind the GRF estimator.

5.1 Estimation Details

Motivated by the differences in flow dynamics presented in the descriptive analysis, I estimate separate decision models of inflows and outflows. In the model of inflows, I estimate the probability of a woman being in the labor force in questionnaire waves 2 or 4, conditionally on being out of the labor force in questionnaire waves 1 or 3, respectively. Analogously, in the model of outflows, I estimate the probability of a woman being out of the labor force in questionnaire waves 2 or 4, conditionally on being in the labor force in questionnaire waves 1 or 3, respectively.

In all model specifications, a woman conditions her choice on a set of observed state variables s, which cannot result from the decision. To account for this, I exploit the rotating panel structure of the survey. I focus on inter-quarter changes in the labor force participation decisions. Specifically, I condition women's choices regarding labor force participation observed in questionnaire waves 2 and 4 on the responses given in waves 1 and 3, respectively. Quarter-lagged state variables cannot result from the decision and provide a good source of variation relevant to women's choices available in my data.

The empirical strategy relies on uncovering the underlying women's decision rule regarding labor force participation. In the real world, a decision regarding labor force participation usually takes into account a series of various factors describing a woman's current life situation. I consider a high-dimensional set of observed state variables to approximate the optimal policy as closely as possible. The main idea is to avoid making arbitrary choices regarding which variables available in the questionnaire to include in the model.

I do not model explicitly the joint decisions in the household, but my approach does not preclude joint decisions in the data-generating process. In the empirical part, I condition women's decisions also on the characteristics of other household members in a previous period.

In theory, the GRF framework can incorporate any non-linearity pattern in how the observed state variables affect the outcome variable at the cost of quickly increasing forest size (in terms of the size of a single tree or number of trees holding the tree size constant) and resulting computational complexity. With a forest large enough, it is sufficient to estimate one model of female labor force participation that pools together data from all periods, previous quarter employment, and program eligibility statuses. In practice, this is not a convenient approach due to the high computational complexity of the algorithm. Therefore, I place a-priori restrictions to help the algorithm perform well without the necessity of growing a large-sized forest. I estimate a separate forest for each combination of period, treatment eligibility (females without children below 18, females with one child below 18, and females with at least two children below 18), and labor force status (in or out) in the preceding quarter.

The GRF routine produces estimates of conditional choice probabilities[1]. Given program eligibility and initial labor force status, I obtain an estimate of $\varrho_t(s)$ for any period t and vector of observed state characteristics s. I aggregate the estimated conditional choice probabilities by averaging over all dimensions in s using survey population weights. I obtain counterfactual conditional choice probabilities by using the estimated model in period t to predict the outcomes using observations from period $s \neq t$.

The parameters β, γ, and ξ are functions of the counterfactual conditional choice probabilities obtained for the same individuals. That makes it difficult to derive appropriate standard errors. However, as conditional choice probabilities obtained from the GRF procedure are asymptotically normal, bootstrap techniques are expected to perform well. For each period t, I repeatedly draw a sample of N_t individuals with replacement and estimate

[1] I use `R` package `grf` developed by Tibshirani, Athey, and Wager (2020).

the decision model. A distribution of parameters obtained by repeating this procedure is expected to converge to the true sampling distribution of the effects of interests as the number of repetitions goes large. All statistical inference performed in the empirical part of this paper is based on 200 bootstrap replications per decision model.

I investigate quarterly dynamics in parameters of decomposition 6. I estimate a series of models in a quarterly rolling observation window to control for seasonal variation in labor force flows and increase estimation precision through increased sample size. That means a decision model for a quarter q is estimated using data on quarters $q-3$ to q. I refer to this approach as *quarterly* models. The quarterly models give more precise insights into the timing of adjustments in labor force flows and help identify the variation resulting from the introduction of the program.

6 Results

In this section, I present the results of my analysis. I begin with discussing the estimates of the parameters in the decomposition 6, suggesting which part of the variation in the evolution in labor market flows comes from the introduction of the child benefits, and choosing counterfactual paths for these parameters in the absence of the program. Then, I move to simulating paths of aggregate labor supply have the benefits not been introduced and comparing them with the realized path to infer the effects of the program on female labor supply.

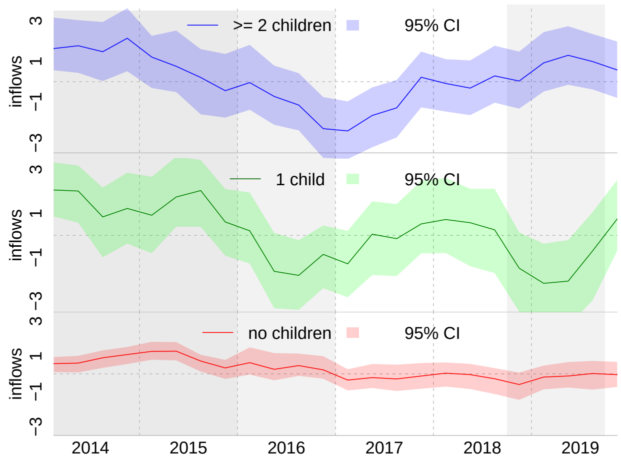
6.1 Decomposition

Figure 3 presents the quarterly time series of the estimates of the decomposition 6 parameters. In the remainder of this subsection, I investigate which part of the variation in these estimates can be attributed to the intervention. I am looking specifically for variation that causes the estimates to be different than zero (a departure from the steady state value) or to be different than each other across the program eligibility groups (motivated by the possibility that the program has a differential effect on the women without children below 18, with one child below 18, and with two or more children below 18).

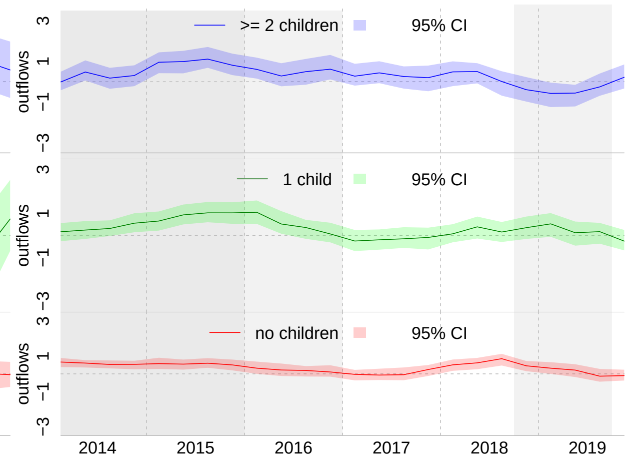
6.1.1 The direct effects of the program

The estimates of the decision rule parameters concerning the inflows into the labor market are presented in figure 3a. Initially, the estimates for women with no children below 18, women with one child below 18, and women with two or more children below 18 exhibit a common trend of positive values converging towards a steady-state value of zero at the end of 2015. The program's introduction does not seem to affect the decision rule parameters for the non-eligible females, which remain non-distinguishable from zero until the end of the sample. In turn, the decision rule parameters for women with two or more children below 18 become negative for a few quarters after the introduction. Therefore, I interpret these changes as a direct impact of the program on women's decision rules—with additional child benefit transfers, women's perceived payoffs changed, elevating the relative costs of labor market activation and making it less profitable to engage in an active search for employment. A similar result is also observed among women with one child below 18. However, the magnitude of the estimates is smaller, consistent with the fact that less than 25% of women with one child below 18 were eligible for the benefits in 2016 (table 1). In turn, when the program eligibility was extended in 2019 to all children, the decision rule channel decreased the inflow rates among women with one child again, stronger than in 2016. The estimates for the remaining groups remained unchanged in 2019, reflecting that 2019 policy changes did not directly impact their payoffs. The decision rule estimates for the inflows suggest that the program's economic impact has been promptly internalized—it took approximately a year for the inflows to return to a steady state value of zero after any

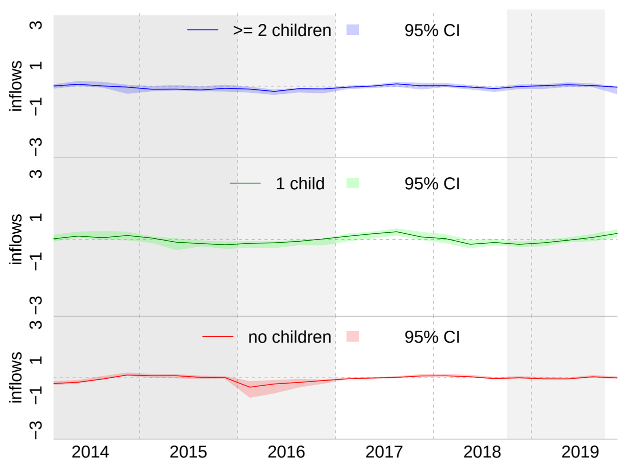
Figure 3: Estimated Parameters of Decomposition (6), Time Series Approach.



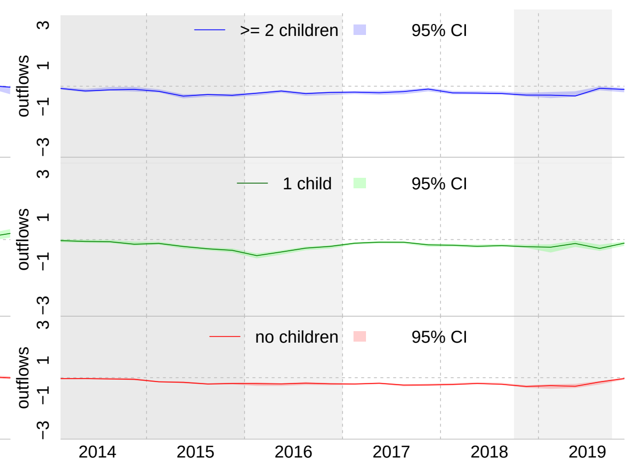
(a) inflows: *decision rule* parameters



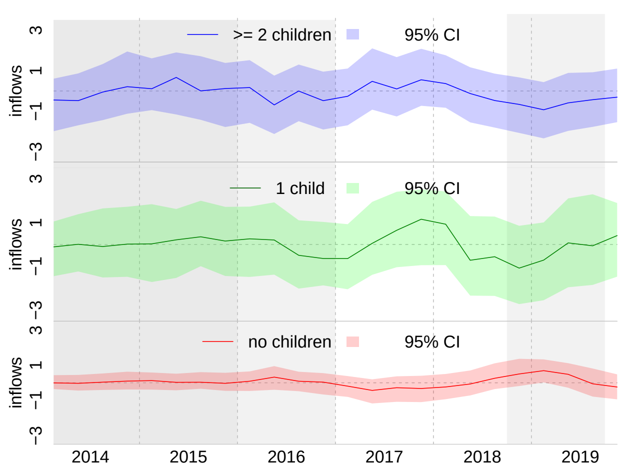
(b) outflows: *decision rule* parameters



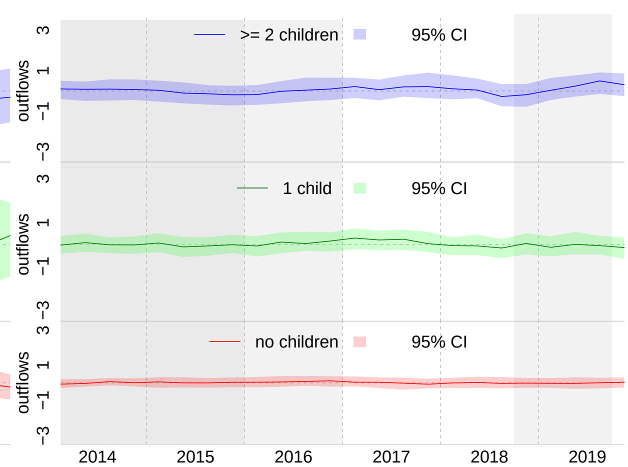
(c) inflows: *composition* parameters



(d) outflows: *composition* parameters



(e) inflows: *residual* parameters



(f) outflows: *residual* parameters

Note: Note: Dark gray background indicates pre-intervention periods, light gray indicates the transition periods.

shock related to the program.

The identification argument concerning the impacts of the program through the decision rule channel for the inflows is supported by both comparing the estimates across groups defined by eligibility and searching for deviations from steady state. It delivers the most compelling evidence of the effects of the program and hence constitutes a basis for the main result of the paper. For the sake of the simulation exercise, I label the variation discussed above as *scenario 1*, describing the *direct* effects of the program. The counterfactual values of the decision rule parameters are defined as follows. I eliminate the impacts of the program through the decision rule channel in predicting inflows among the eligible females by assuming that $\beta(s_{t-1}^{\geq 2\, children})$ and $\beta(s_{t-1}^{1child})$ take values of $\beta(s_{t-1}^{no\, children})$ in periods 2016Q2-2017Q1, and additionally that $\beta(s_{t-1}^{1child})$ is equal to $\beta(s_{t-1}^{no\, children})$ in periods 2018Q4-2019Q3. The remaining values of the time series of decomposition 6 remain unchanged. Note that this definition of counterfactual can be interpreted as both using the outcomes for a control group and shutting down the variation to its steady state value of zero.

6.1.2 The indirect effects of the program

The decision rule parameters in outflow decomposition followed similar patterns for every group, positive in 2014-2015 and zero afterward, as depicted in figure 3b. They diverge for the first time in mid-2018 after an increase in parameters for ineligible females. It is unlikely that this effect is related to the program *P500*. Economic theory does not predict that child benefits would change the economic environment or individual beliefs among ineligible females, which would support increased outflows from the labor force. First, the program does not directly affect women who do not obtain the transfers. Second, the indirect effects would rather be associated with decreased outflows. As discussed above, the benefits discouraged inflows to the labor market among women with 2 or more children in 2016. The implied shrinkage of the labor supply is likely to make it more difficult for employers to keep the current staff and hire new suitable employees. As a result, employers are likely to improve the job conditions for existing employees or new hires, which in turn would limit the outflows. This is precisely what happens afterward. The decision rule parameters in the outflow decomposition decreased in all three groups of females between mid-2018 and 2019 in a nearly parallel fashion. The mechanism described above is consistent with the decision rule channel, as the improved employment conditions affect the economic environment and future beliefs accommodated in women's decision rules. There should not be differences in the dynamics of this effect between the eligible and ineligible women, which is confirmed by my results.

Even though this evidence may not seem as conclusive as in the case of scenario 1 variation, the economic theory and empirical evidence indicate that the drop in decision rule parameters in the outflow rate changes decomposition at the end of the sample may be an indirect result of the *P500*. I refer to this variation as *scenario 2* in the simulation exercise, describing the *indirect* effects of the program. I turn off the *indirect* effects of the program by assuming that the drop in the decision rule parameters in the outflow decomposition does not occur. Specifically, I set all three βs in periods 2018Q4-2019Q3 to their respective averages over the preceding quarters 2018Q1-2018Q3. The remaining values of the time series of decomposition 6remain unchanged.

6.1.3 Additional results

Estimates of the composition parameters in the decomposition of both the inflows (figure 3c) and outflows (figure 3d) show similar flat trends for every group of women. Hence, the program is unlikely to manifest its effects through this channel. Any deviations are also small and do not significantly affect the flows or the aggregate.

Lastly, the residual parameters in both decompositions and across groups (figures 3e and 3f) are statistically insignificant throughout the whole sample, supporting my choice of specification. The model performs well at the quarterly level.

6.1.4 Further remarks

The generalized random forest allows a much more in-depth analysis of the decision model and the resulting decomposition 6. Section C in the online supplementary materials discusses the heterogeneity in program impacts on women in different demographic groups and provides evidence on the fact that after the introduction of the program, the variables related to child benefits have strengthened their power in predicting labor market flows.

6.2 Simulation

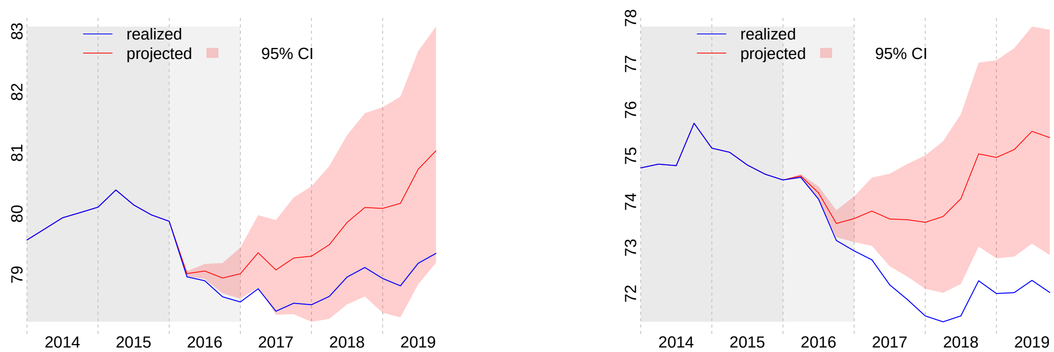
With the definition of counterfactual time series of the parameters of decomposition 6, reflecting their expected realizations had the program not been introduced, I proceed to simulate counterfactual paths of aggregate labor supply. First, I simulate the path under scenario 1 only, considering the program's direct effects. Second, I also allow for the indirect effects, simulating the path under both scenario 1 and scenario 2.

6.2.1 Main result – direct effects

Through its direct effect, the program *P500* decreased the aggregate labor force participation among eligible females by discouraging labor market activation among eligible women outside of the labor force. The direct channel does not predict changes in the behavior of ineligible women who were not directly affected by child benefits.

Figure 4 presents the realized and simulated paths of aggregate labor force participation among women with one child below 18 (panel 4a) and women with two or more children below 18 (panel 4b). The point estimates obtained for both groups indicate that the labor force participation would have been higher without introducing the child benefits.

Figure 4: Actual and Simulated Paths of Labor Supply (direct effects).



(a) one child (b) two or more children

Note: Note: Dark gray background indicates pre-intervention periods, light gray indicates the transition periods.

Table 4: Effects on the labor force participation – simulation results (direct effects).

	1 child		2+ children	
	2017Q4	2019Q4	2017Q4	2019Q4
labor force participation	-0.743 (-1.456)	-1.687 (-1.596)	-1.74 $(-2.762)^{***}$	-3.373 $(-2.702)^{***}$

Note: Note: The table presents the results of the program evaluation exercise, that is, the differences between the realized and simulated paths of female labor force participation at a given point in time. t-Statistics based on 200 bootstrap replications are presented in parentheses. Stars denote *** p-val.<0.01, **, p-val.<.05, * p-val.<0.1.

To further quantify the direct effects of the program *P500*, table 4 presents the differences between realized and simulated paths at the end of 2017 (measuring immediate effects of the intervention) and at the end of 2019 (measuring more prolonged effects of the intervention).

The results indicate that the child benefits significantly decreased the labor force participation rate among women with two or more children by 1.74 percentage points at the end of 2017. The changes in flows interacted and accumulated in each consecutive quarter, leading to further evolution of intervention impacts on labor force participation. At the end of 2019, the estimates suggest a 3.37 percentage point drop in labor force participation among women with two or more children below 18 solely due to the propagation of the direct (immediate) shocks.

The effects were lower among women with one child below 18, with the estimated effects at -0.74 pp. at the end of 2017 and 1.69 pp. at the end of 2019. The relatively lower magnitude of these effects is associated with the fact that only a fraction of women with one child below 18 were eligible until 2019. Moreover, having a single child below 18 implies the lowest benefit rate, which may be a weaker disincentive to becoming active in the labor market.

The statistical precision of the estimated counterfactual paths of labor force participation is not very high, as evidenced by relatively wide 95% confidence intervals, systematically widening over time. This comes from the fact that based on assumptions on the economic environment that are not explicitly accounted for in the model, the estimated path comprises accumulated counterfactual predictions from multiple periods. The uncertainty at each data point interacts with others and propagates into subsequent periods. The low precision affects, in particular, the interpretation of the effects on women with one child below 18, which fails the statistical significance test at a 95% confidence level from 2017 on.

6.2.2 Additional Result – The Indirect Effect

In this subsection, I simulate the counterfactual paths of labor market participation under scenario 1 and scenario 2 simultaneously. Table 5 presents the effects after two and four years since the introduction of child benefits.

The total effect is a sum of the direct effects that limited the inflows and indirect effects that limited the outflows from the labor force. The overall effect is positive since only the

Table 5: Effects on the labor force participation – simulation results (direct and indirect effects).

	no children		1 child		2+ children	
	2017Q4	2019Q4	2017Q4	2019Q4	2017Q4	2019Q4
labor force participation	—	0.975 (3.006)***	−0.743 (−1.456)	−1.811 (−1.545)	−1.74 (−2.762)***	−2.701 (−1.956)*

Note: Note: The table presents the results of the program evaluation exercise, that is, the differences between the realized and simulated paths of female labor force participation at a given point in time. t-Statistics based on 200 bootstrap replications are presented in parentheses. Stars denote *** p-val.<0.01, **, p-val.<.05, * p-val.<0.1.

latter affected women with no children below 18. Interestingly, the estimated effects for women with one child below 18 did not change much. The total decrease in labor force participation among women with two or more children below 18 at the end of 2019 amounts to 2.7 pp. – a more moderate decrease compared to the scenario only considering the direct effects. However, this effect is not precisely estimated. With the increase in the number of assumptions about counterfactual outcomes attributable to the program, the precision of the simulated paths decreases.

6.2.3 Remarks

My result shed further light on the effects of *P500* obtained in other studies, relying on standard difference-in-difference assumptions. Magda et al. (2018) focused mainly on the immediate impacts of the program. Their comparisons between women with two or more children and women without children in the initial quarters after the introduction of the program reveal similar differences to my estimates. However, they miss the effects on women with one child and cannot take into account the potential indirect effects. Gromadzki (2024) fails to find significant effects of the program on labor supply while comparing women with two or more children below 18 to women with one child below 18. As indicated above, both groups are affected by the program similarly; hence, differencing the outcome between these groups annihilates the effect of child benefits on labor supply.

7 Conclusion

In this paper, I propose a simple but comprehensive approach to analyze the effects of large-scale government programs and apply the framework to study how a child benefit program impacts women's labor supply. I develop and estimate a general discrete choice model of women's decisions that explains labor force participation flows. Changes in flows are decomposed into components related to changes in a woman's decision rule, her observed characteristics, and residual factors. By shutting down variation triggered by the program, I simulate counterfactual flow rate paths. Using the law of motion for the aggregate labor supply, I recover the impacts of the program on labor force participation.

My estimates reveal a moderate but statistically significant drop in labor force participation among eligible women as a result of the program, driven mainly by changes in women's decision rules accommodating perceived trade-offs and beliefs that discouraged labor force inflows immediately after the program's introduction. Changes in flow rates induced by the program may not seem large but accumulate over time, leading to significant changes in labor force participation.

The flexibility of my approach may enable researchers to study many other large-scale programs in which it is difficult to define appropriate control groups, and hence, the identification strategies based on traditional frameworks may seem less appealing. The only trade-off for the researchers is that due to the relative simplicity of my model, an analysis of more elaborative counterfactual scenarios at the aggregate level requires more ad-hoc assumptions about the market and may lead to an increase in estimation uncertainty.

References

Angrist, J. and B. Frandsen (2019). Machine labor. Technical report, National Bureau of Economic Research.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Baker, M., D. Messacar, and M. Stabile (2021). The effects of child tax benefits on poverty and labor supply: Evidence from the canada child benefit and universal child care benefit. Technical report, National Bureau of Economic Research.

Blundell, R., A. Duncan, J. McCrae, and C. Meghir (2000). The labour market impact of the working families' tax credit. *Fiscal studies 21*(1), 75–104.

Cengiz, D., A. Dube, A. S. Lindner, and D. Zentler-Munro (2021). Seeing beyond the trees: Using machine learning to estimate the impact of minimum wages on labor market outcomes. Technical report, National Bureau of Economic Research.

González, L. (2013). The effect of a universal child benefit on conceptions, abortions, and early maternal labor supply. *American Economic Journal: Economic Policy 5*(3), 160–88.

Gromadzki, J. (2024). Labor supply effects of a universal cash transfer. *Journal of Public Economics 239*, 105248.

Immervoll, H., H. J. Kleven, C. T. Kreiner, and E. Saez (2007). Welfare reform in european countries: a microsimulation analysis. *The Economic Journal 117*(516), 1–44.

Koebel, K. and T. Schirle (2016). The differential impact of universal child benefits on the labour supply of married and single mothers. *Canadian Public Policy 42*(1), 49–64.

Magda, I., A. Kiełczewska, and N. Brandt (2018). The effects of large universal child benefits on female labour supply. *IZA Discussion Paper No. 11652*.

Moffitt, R. A. (2002). Welfare programs and labor supply. *Handbook of public economics 4*, 2393–2430.

Myck, M. (2016). Estimating labour supply response to the introduction of the family 500+ programme. *Centre For Economic Analysis (CenEA) Working Paper Series WP01/16*.

Myck, M. and K. Trzciński (2019). From partial to full universality: The family 500+ programme in poland and its labor supply implications. *ifo DICE Report 17*(03), 36–44.

Schirle, T. (2015). The effect of universal child benefits on labour supply. *Canadian Journal of Economics/Revue canadienne d'économique 48*(2), 437–463.

Sigurdsson, J. (2019). Labor supply responses and adjustment frictions: A tax-free year in iceland. *Available at SSRN 3278308*.

Stephens Jr, M. and T. Unayama (2015). Child benefit payments and household wealth accumulation. *The Japanese Economic Review 66*(4), 447–465.

Tibshirani, J., S. Athey, and S. Wager (2020). *grf: Generalized Random Forests*. R package version 1.2.0.