MONASH University

MONASH BUSINESS SCHOOL

CENTRE FOR HEALTH ECONOMICS

WORKING PAPERS

Taking an Extra Moment to Consider Treatment Effects on Distributions

Discussion Paper no. 2024-18

Gawain Heckley and Dennis Petrie

Keywords: Causal Inference, Policy Evaluation, Distribution Impacts, Income Inequality, Education Inequality

JEL Classification: I24, I26, I28, C10

Gawain Heckley: Health Economics, Faculty of Medicine, Lund University (email: gawain.heckley@med.lu.se); Dennis Petrie: Centre for Health Economics, Monash Business School, Monash University (email: dennis.petrie@monash.edu).

Taking an Extra Moment to Consider Treatment Effects on Distributions

Gawain Heckley and Dennis Petrie *

Abstract

This paper presents a flexible method, Parameter Estimation by Raw Moments (PERM), to evaluate a policy's impact on parameters of the distribution of outcomes. Such parameters include the variance ($\mathbb{E}[Y^2] - \mathbb{E}[Y]^2$), skewness and covariance. While many studies estimate the mean (first moment), PERM extends this to estimate higher order moments, enabling calculation of distribution parameter treatment effects. Two implementations are discussed: regression with controls and DiD with staggered roll-out. Applying PERM DiD to a Swedish school reform shows it reduced education inequality but increased earnings variance resulting in a lower covariance between education and earnings.

Keywords: Causal Inference, Policy Evaluation, Distribution Impacts, Income Inequality, Education Inequality

JEL Classification: *I24, I26, I28, C10*

Acknowledgements: The authors would like to thank Maarten Lindeboom, Lars Kirkebøen, Christian Hansen, Joakim Westerlund, Johannes Kunz, Denzil Fiebig and participants at Applied Econometrics Conference (Oslo, 2023), Australian Health Economics Society (Brisbane, 2022), Workshop on Education Economics and Policy (Trondheim, 2023), Lund University for helpful comments and discussions. *Funding:* Financial support from Crafoord (Heckley), Swedish Research Council (Heckley, dnr: 2019-06292) and Swedish Research Council for Health, Working Life and Welfare (FORTE dnr:2023-01128) is gratefully acknowledged. *Heckley:* Health Economics, Faculty of Medicine, Lund University, 220 07 Lund, Sweden. E-mail: gawain.heckley@med.lu.se. *Petrie:* Centre for Health Economics, Monash Business School, Monash University, Level 5 Building H, Caulfield, Victoria 3145, Australia E-mail: dennis.petrie@monash.edu.

It is ironic that most of the research investigating compulsory schooling reforms, egalitarian policies designed to improve equality of opportunities and reduce inequality, has focused on mean outcomes rather than their wider distributional effects (see e.g. Meghir and Palme 2005, Oreopoulos 2006, Clark and Royer 2013, Fischer et al. 2021). The focus on the mean in the policy evaluation literature can likely be attributed to its desirable statistical properties that include unbiasedness, consistency, and additive separability. These properties are crucial to several of the most commonly used policy evaluation methods, such as those based on linear regression or differences-in-differences analysis. As such, it is empirically challenging to apply these same policy evaluation tools to estimate distribution parameters beyond the mean. This paper introduces a new method, Parameter Estimation by way of Raw Moments (PERM), a simple extension to common policy evaluation techniques that enables investigation of a broader range of univariate and bivariate distribution parameters.

PERM builds on the simplicity of mean analysis. Although the variance itself is not additively separable, it can be expressed exclusively as a nonlinear function of additively separable components $\mu_2 = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$.[1] $\mathbb{E}[Y]$ and $\mathbb{E}[Y^2]$ are the first and second raw moments of the distribution and are additively separable being the means of Y and Y^2 respectively. PERM first estimates a policy's impact on each raw moment, noting that most studies will have already estimated the impact on the mean $\mathbb{E}[Y]$. Second, these raw moments are then combined to estimate the observed and counterfactual distribution parameter of interest. For the variance, this requires only an additional estimate of the second raw moment, whose identification and estimation is a simple extension of the mean. The PERM method of combining estimates of raw moments provides a new flexible way to estimate treatment effects on distribution parameters of outcomes.

Raw moments (e.g.$\mathbb{E}[Y]$, $\mathbb{E}[Y^2]$,...) are just means of polynomials of the observed outcomes and thus share the same desirable statistical properties as the mean (unbiasedness, consistency and additive separability). The same empirical tools used to investigate the first raw moment can also be applied to higher-order raw moments. Such estimates in turn enable causal analysis of distribution parameters that can be expressed as functions exclusively of raw moments. This includes not only the variance but also the skewness. In fact, several univariate inequality measures, such as the Coefficient of Variation and the Thiel Index, can also be expressed as functions exclusively of raw moments.[2] In some settings, the impact of

[1] The Law of Total Variance states that the total variance can expressed as a function subgroup (G) variance and variance of subgroup means: $\mu_2 = \mathbb{E}[\mu_2(Y|G)] + \mu_2(\mathbb{E}[Y|G])$. The second term represents the between group component of the variance that prevents additive separability.

[2] Note that means of ranks are not raw moments as ranks are themselves functions. Rank-dependent inequality measures such as the Gini index therefore fall out of scope. Rank-based measures are particularly tricky because they make the contribution of observations dependent on other observations (i.e. not additively separable into the contribution of each observation or group). This is because a rank change for one observation will, by construction, change the rank for at least one other observation. Rank-dependent inequality measures therefore require estimation of the entire distribution of counterfactual outcomes.

a treatment on the joint distribution of outcomes may also be of interest. Treatment may not only affect these outcomes independently, but also change how these outcomes relate to each other. Distribution parameters that capture the bivariate relationship between multiple outcomes include the covariance and the slope coefficient, which are functions exclusively of both univariate raw moments as well as bivariate or joint raw moments (e.g.$\mathbb{E}[YW]$).

We focus on raw moments because they are informative about the distribution of outcomes and relatively simple to evaluate. Moments are as old as the field of statistics. Andrey Markov utilised the 'Method of Moments', building on the work of Chebyshev (1891), to provide a proof for the Central Limit Theorem. Karl Pearson utilised the 'method of moments' to define distributions based on empirical moments from a sample (Pearson 1894). Not only are raw moments helpful in parameterising distributions, certain distributions can also be uniquely characterised by their raw moments.[3] While knowing every raw moment might identify the exact distribution, even knowing only a small number of lower order raw moments can help understand the nature of a distribution. The distribution information contained in raw moments and their straightforward estimation are the two key insights that motivate PERM. PERM builds on the progress made in credibly estimating mean equations by extending these to higher-order raw moments that, in turn, can be used to estimate treatment effects on distribution parameters. It thus turns a complex analytical problem of estimating the treatment effect on the distribution of outcomes into a more tractable two step non-parametric procedure.

For the sake of clarity, we focus on the treatment effect on the distribution of outcomes, not on the distribution of treatment effects.[4] In particular, we estimate the difference between the distribution of outcomes and what it would have been without treatment - what we call the *Distribution Parameter Treatment Effect* (DPTE).[5] The DPTE will depend on both the heterogeneous treatment effects and how these are distributed in relation to the untreated outcomes.

Several methods already exist that allow estimation of distribution parameter treatment effects. One popular method is the marginal estimation technique of Recentered Influence Function (RIF) Regression (see e.g. Firpo, Fortin and Lemieux 2009, 2018, Essama-Nssah and Lambert 2012, Heckley, Gerdtham and Kjellsson 2016, for alternative use cases). This method provides a linear approximation of how the distribution parameter will change in

[3]The Hamburger moment problem solves whether a unique distribution is identified by a sequence of moments across the entire real line.

[4]The distribution of treatment effects can be considered as the distribution of the difference in potential outcomes (see e.g. Fan and Park 2010, Firpo and Ridder 2019, Melly and Wüthrich 2017, for recent contributions). The distribution of treatment effects ignores the relationship between individual treatment effects and the baseline distribution of outcomes.

[5]Firpo and Pinto (2016) call this the *inequality treatment effect*. We utilise more general terminology because our focus here is not specific measures of inequality.

response to a marginal change in treatment. As a consequence, RIF only estimates 'local effects' and is therefore only useful for 'small' policy changes. For 'larger' policies that substantially affect the distribution of outcomes, local linearisation techniques like the RIF approach, may provide a poor approximation (Rothe 2015).

Several non-local estimation techniques exist. These methods require the assumption of treatment unconfoundedness, an assumption requiring selection into treatment based only on observable characteristics, also known as strong ignorability. Examples include re-weighting based approaches (Firpo and Pinto 2016, DiNardo, Fortin and Lemieux 1996, Card et al. 2004), location shift estimators (Juhn, Murphy and Pierce 1993), and methods that parametrically or non-parametrically estimate the full conditional distribution (Chernozhukov, Fernández-Val and Melly 2013, Rothe 2010).

In this paper we first introduce a regression-based PERM approach (PERM regression) which assumes weak ignorability (independence of treatment across the required raw moments) and compare it to methods that require strong ignorability. Strong ignorability requires ALL raw moments to be conditionally independent of treatment given observables. In a Monte Carlo exercise alongside an empirical application of union coverage in the USA we show that PERM regression yields very similar results compared to the Inverse Probability re-Weighting approach (IPW). Our finding that PERM performs well when evaluated against IPW, combined with the results of Firpo and Pinto (2016) who find that IPW performs well against the methods of Juhn, Murphy and Pierce (1993), Chernozhukov, Fernández-Val and Melly (2013), together suggest that PERM regression performs well generally, while requiring weaker assumptions. These results are confirmed in our first empirical application assessing the impact of union coverage on US log hourly wages. The results show that union coverage not only improves the mean of log earnings, but also reduces variance and standardised skewness. Both PERM and IPW yield very similar conclusions.

PERM, however, can also be utilised under alternative identifying assumptions. In this paper, we also introduce a difference-in-difference based PERM approach (PERM DiD) that relaxes the raw moment independence assumption and instead utilises parallel trend assumptions. We show that PERM DiD identifies the second raw moment under the assumption of parallel trends in the mean and variance of groups. A parallel trend in the mean is, in general, incompatible with parallel trends in the mean of any non-linear transformations of the outcome variable. This is because any trend in the mean will, in general, lead to non-parallel trends in the mean of the transformation.[6] We show that under a parallel group variance assumption the effect of the trend in the mean has a mechanical effect on the second raw moment and this can be estimated and used to correct the DiD of the second raw moment yielding unbiased results. We extend this idea to higher-order

[6] Roth and Sant'Anna (2023) provide a proof of this.

and bivariate raw moments. Our empirical implementation of PERM DiD utilises the staggered event study DiD regression approach of De Chaisemartin and dHaultfoeuille (2020) and De Chaisemartin and d'Haultfoeuille (2024), for which we provide a Stata command *did_multiplegt_PERM*.

Our second empirical application illustrates PERM DiD by estimating the DPTE of a major Swedish educational reform in the 1940's and 1950's that increased the minimum years of schooling, delayed ability streaming (tracking) and thereby mixed ability peer groups for longer. This reform had the explicit aim of reducing inequalities in outcomes through improved equality of opportunity. Previous research has shown that this reform increased average years of education as well as income (Meghir and Palme 2005, Holmlund 2007, Fischer et al. 2021). We replicate these findings and then extend the previous analysis to consider distribution impacts. We find that the reform substantially reduced the variance in years of schooling, but the impacts on labour earnings indicate an increase in the variance. We also consider the impact on the relationship between education and earnings and find a clear reduction in the covariance and slope of earnings and education. This suggests that the comprehensive school reform reduced education inequality and weakened the education gradient in earnings in Sweden but increased labour market inequalities. The results suggest that even well-targeted education policies aimed at reducing education inequalities may not be effective policy tools in reducing labour market inequalities.

Several alternative distribution DiD estimators have been suggested (Athey and Imbens 2006, Bonhomme and Sauder 2011, Callaway and Li 2019, Roth and Sant'Anna 2023, Fernández-Val et al. 2024). Each impose alternative parallel trends assumptions to allow the identification of the distribution effects. If one's interest is only focused on identifying only a few distribution parameters, then the alternative parallel trend assumptions required by distribution DiD approaches are in general stronger than those required by PERM DiD. This is because these alternative methods require identification of the full counterfactual distribution in order to identify the counterfactual distribution parameter, whereas PERM only requires identification of the relevant counterfactual raw moments, which is less demanding. Furthermore, it is not obvious how the distribution DiD estimators of (Athey and Imbens 2006, Bonhomme and Sauder 2011, Callaway and Li 2019, Roth and Sant'Anna 2023) can be easily extended to consider the nature of a policy's impact on the bivariate outcome distribution, such as, the covariance of education and labour earnings. The analysis of parameters of the bivariate outcome distribution is relatively straightforward using PERM DiD.

PERM is a flexible framework that identifies the DPTE under alternative identifying assumptions. PERM regression and PERM DiD are just two PERM based approaches that could be used to identify the DPTE. More generally, PERM is valid in any empirical setting as long as credible counterfactual raw moments can be identified. PERM is relatively straight forward and quick to implement and provides a natural extension to causal inference methods that focus on the mean. Finally, there is an intuitiveness to the PERM approach which potentially opens up its usefulness to a wider audience compared to alternative DPTE estimators. PERM builds on what is already familiar, means and regressions. No new concepts are required, rather PERM provides a framework for the analysis of DPTE using concepts commonly understood.

1 Parameter Estimation using Raw Moments

1.1 Distribution Parameter Treatment Effects

We want to know how exposure to a treatment changes the distribution of outcomes. First, let us define a binary variable for treatment, $D_i = \{0,1\}$ where 0 is untreated and 1 is treated. $Y_{0,i}$ is the potential outcome for individual i if they were unexposed to treatment and $Y_{1,i}$ is the potential outcome if they were exposed. Outcomes can be expressed in terms of potential outcomes:

$$Y_i = Y_{0,i} + \tau_i D_i \quad \forall_i. \tag{1}$$

where $\tau_i = (Y_{1,i} - Y_{0,i})$ and is the treatment effect for each individual i.

The distribution of outcomes Y_i for a particular population can be defined in terms of the parameters, $v(Y)$, which describe it. Common parameters used to describe distributions include the mean, variance and standardised skewness. We need not restrict ourselves to parameters that describe a univariate outcome distribution. Suppose that a treatment affects the distribution of two outcomes. It may also be of interest to understand how the relationship between these two outcomes changes as a result of treatment. This relationship can be summarised using parameters that describe the joint distribution of two or more outcomes, $v(Y,W)$, such as the covariance of Y and W, for example.

A natural way to compare two potential outcome distributions is to consider the difference in their parameters. The difference between the parameters of the observed and potential outcome distribution for the treated population gives the *Distribution Parameter Treatment effect on the Treated* (DPTT):

$$\begin{aligned} \Delta_{DPTT} =& v(Y_1|D=1) - v(Y_0|D=1) \\ =& v_{11} - v_{01} \end{aligned} \tag{2}$$

where v_{11} and v_{01} are the parameters of the distribution of potential outcomes for the treated group, if they were treated and untreated, respectively. Firpo and Pinto (2016) refer to the DPTT as *inequality treatment effect on the treated* due to their inequality focus. DPTT examples include the Average Treatment effect on the Treated (ATT) and the Variance Treatment effect on the Treated (VTT).

In this paper, we focus on the treated population to simplify our exposition but there are other population comparisons that may also be of interest. For example, Firpo and Pinto (2016) define the *overall inequality treatment effect* as the difference in the distribution of outcomes for the whole population if everyone was treated compared to a counterfactual where no one was treated, and the *current inequality treatment effect* as how the observed outcome distribution for the overall population has been impacted by only part of the population being treated. These treatment effects can be estimated by the methods we introduce in this paper, but we leave this as an extension.

1.2 Identifying Distribution Parameter Treatment effects on the Treated

To identify the distribution parameter treatment effect on the treated we need to estimate a counterfactual. Two assumptions commonly used in the inequality policy evaluation literature (see e.g. Firpo and Pinto 2016, Firpo, Fortin and Lemieux 2009, Card et al. 2004, Card, Lemieux and Riddell 2020, Chernozhukov, Fernández-Val and Melly 2013, Juhn, Murphy and Pierce 1993) to identify a counterfactual distribution are:

Assumption 1 (Unconfoundedness). *Given a set of observed characteristics x in χ, then our potential outcomes (Y_1, Y_0) are jointly independent of treatment D given $X = x$.*

Assumption 2 (Common Support). *For all x in χ, there is a positive probability of being both treated and untreated:* $0 < P(D = 1|X = x) < 1$.

Assumption 1 is also known as selection on observables. Assumption 2 ensures that a control group exists that closely matches the treated group for all values of x. When both assumptions hold Rosenbaum and Rubin (1983) define treatment assignment as strongly ignorable.

Under strong ignorability, it is possible to identify not only the average treatment effect on the treated, $\Delta'_{\mu} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$, but also every raw moment treatment effect on the treated, $\Delta_{\mu'_q} = \mathbb{E}[Y_1^q] - \mathbb{E}[Y_0^q]$, where Y^q is the q^{th} order polynomial of Y (Imbens 2004). That is, we can identify not only the counterfactual first raw moment (the mean) but also all counterfactual raw moments.

Raw moments beyond the mean are of interest because they tell us more about the distribution of outcomes. Table 1 provides a non-exhaustive set of discrete distribution parameters

Table 1: Distribution Parameters Expressed as Functions of Raw Moments

PARAMETER	FORMULA
Univariate Distribution:	
Mean μ	$\mathbb{E}[Y]$
Variance μ_2	$\mathbb{E}[Y^2]-\mathbb{E}[Y]^2$
Coefficient of Variation $(\mu_2)^{1/2}/\mu$	$\frac{(\mathbb{E}[Y^2]-\mathbb{E}[Y]^2)^{1/2}}{\mathbb{E}[Y]}$
Skewness μ_3	$\mathbb{E}[Y^3]-3\,\mathbb{E}[Y^2]\,\mathbb{E}[Y]+2\,\mathbb{E}[Y]^3$
Standardised Skewness $\frac{\mu_3}{(\mu_2)^{3/2}}$	$\frac{(\mathbb{E}[Y^3]-3\,\mathbb{E}[Y^2]\,\mathbb{E}[Y]+2\,\mathbb{E}[Y]^3)}{(\mathbb{E}[Y^2]-\mathbb{E}[Y]^2)^{3/2}}$
Bivariate Distribution:	
Covariance μ_{YW}	$\mathbb{E}[YW]-\mathbb{E}[Y]\,\mathbb{E}[W]$
Slope $\frac{\mu_{YW}}{\mu_2(W)}$	$\frac{(\mathbb{E}[YW]-\mathbb{E}[Y]\,\mathbb{E}[W])}{(\mathbb{E}[W^2]-\mathbb{E}[W]^2)}$

Standarised Skewness is sometimes referred to as the Coefficient of Skewness, or just as Skewness.

that can be expressed as functions of raw or joint moments. These distribution parameters, familiar to many, are often used in the analysis of inequality, and can be used to help depict full distributions using fitted distribution models. For example, the normal distribution, the gamma distribution, and Poisson distribution can all be parameterised with only knowledge of the mean and variance. More flexible models such as the Pearson I-IV family of distributions can be parameterised by the mean, variance, standardised skewness and standardised kurtosis. Moments can also be used to calculate parameters of a joint distribution including the covariance and the slope coefficient.

Identification of a larger and larger set of raw moments allows a more and more detailed description of the counterfactual distribution. As long as the relevant raw moments can be consistently identified, then it follows from the Slutsky Theorem that any distribution parameter that can be expressed as functions exclusively of these raw moments can also be consistently identified. This leads us to the following proposition.

Proposition 1. ***Consistency of distribution parameter treatment effects estimated by way of raw moments*** *Consider a distribution parameter v that can be expressed as a function g*

exclusively of a vector of a subset of the population's raw moments[7] $\boldsymbol{\mu'_q} \in (\mu'_1, \mu'_2, \ldots, \mu'_q)$ *such that* $v = g(\boldsymbol{\mu'_q})$. $\boldsymbol{\mu'_q}$ *for the observed distribution of the treated is given by* $\boldsymbol{\mu'_{11,q}}$ *and for the counterfactual distribution of the treated is given by* $\boldsymbol{\mu'_{01,q}}$. *The population DPTT is therefore given by* $\Delta_{DPTT} = g(\boldsymbol{\mu'_{11,q}}) - g(\boldsymbol{\mu'_{01,q}})$. *If* $v_{11}^{\mu'_q}(Y)$ *and* $v_{01}^{\mu'_q}(Y)$ *are weakly consistent sample estimators for* $\mu'_{11,q}$ *and* $\mu'_{01,q}$, *respectively, for all* $\mu'_q \in \boldsymbol{\mu'_q}$, *such that* $\underset{n\to\infty}{plim}\ v_n^{\mu'_q}(Y) = \mu'_q$, *then the sample estimator of DPTT,* $v_n^{\Delta_{DPTT}} = g(\mathbf{v_{11,n}^{\mu'_q}}(\mathbf{Y})) - g(\mathbf{v_{01,n}^{\mu'_q}}(\mathbf{Y}))$ *will also be weakly consistent,* $\underset{n\to\infty}{plim}\ v_n^{\Delta_{DPTT}} = \Delta_{DPTT}$.

Proof. Given weakly consistent estimates of the relevant raw moments, then it follows from Slutsky's Theorem, that any function soley of these raw moments will also provide a weakly consistent estimator of that function:

$$\begin{aligned}
\underset{n\to\infty}{plim}\ v_n^{\Delta_{DPTT}} &= \underset{n\to\infty}{plim}(g(\mathbf{v_{11,n}^{\mu'_q}}(\mathbf{Y})) - g(\mathbf{v_{01,n}^{\mu'_q}}(\mathbf{Y}))) \\
&= g(\underset{n\to\infty}{plim}(\mathbf{v_{11,n}^{\mu'_q}}(\mathbf{Y}))) - g(\underset{n\to\infty}{plim}(\mathbf{v_{01,n}^{\mu'_q}}(\mathbf{Y}))) \\
&= g(\boldsymbol{\mu'_{11,q}}) - g(\boldsymbol{\mu'_{01,q}}) \\
&= \Delta_{DPTT}
\end{aligned}$$

assuming $g(\mathbf{v_n^{\mu'_q}}(\mathbf{Y}))$ exists. □

Consider the variance for example, which can be expressed as a function of raw moments $\mathbb{E}[Y^2] - \mathbb{E}[Y]^2$. It follows from Proposition 1 that the counterfactual variance can be consistently estimated given consistent estimates of the counterfactual raw moments $\mathbb{E}[Y_0]$ and $\mathbb{E}[Y_0^2]$ and likewise for the observed variance. Given consistent estimates of the counterfactual and observed variance we can thereby estimate the variance treatment effect on the treated parameter. This is what we term Parameter Estimation by way of Raw Moments (PERM), which is applicable to any distribution parameter or inequality measure expressible solely as a function of raw moments.

PERM turns a complex analytical problem of estimating distribution parameter treatment effects into a more tractable challenge of estimating raw moments. This approach not only makes estimation of discrete distribution parameters more manageable, but also requires weaker identifying assumptions. Unlike strong ignorability, which is necessary to identify the entire counterfactual distribution (e.g., Juhn, Murphy and Pierce 1993, Firpo 2007, Chernozhukov, Fernández-Val and Melly 2013), PERM only requires independence in the specific raw moments of interest:

[7] Note the set of raw moments can also include joint raw moments if the outcome distribution of interest is multivariate, such as the first joint raw moment of two random variables X and Y, $\mathbb{E}[XY]$.

Assumption 3. *[q$^{\text{th}}$ **Raw Moment Independence**] Given a set of observed characteristics x in χ, then our raw moment outcomes $(\mathbb{E}[Y_1^q], \mathbb{E}[Y_0^q])$ are jointly independent of treatment D given $X = x$.*

Mean independence for the first raw moment ($q = 1$) is known as weak ignorability and is unquestionably weaker than unconfoundedness (Imbens 2004). If one's interest lies in the variance, then we require raw moment independence in both the first and second raw moments (Assumption 3 holds for $q \in \{1,2\}$). Mean independence in the first two raw moments is a stronger assumption than just mean independence, but it is still a weaker assumption than unconfoundedness. This weaker assumption, however, comes at the cost of revealing less about the counterfactual outcome distribution. If raw moment independence is invoked for $q \in \{1,...,\infty\}$ only then does assumption 3 become as strong as the unconfoundedness assumption.

1.3 How Does PERM Work?

PERM utilises the fact that higher-order moments tell us more about the distribution. To illustrate what information we gain from considering the average treatment effect for higher-order raw moments let us assume the following simple Data Generating Process (DGP) for outcome Y:

$$Y_i = \alpha_1 + \tau_i D_i + \varepsilon_i \tag{3}$$

Under this DGP, α_1 is the mean untreated outcome, ε_i determines the distribution of untreated outcomes and τ_i is individual i's own treatment effect. Unfortunately, it is impossible to identify the full joint distribution of τ_i and ε_i without making strong un-testable assumptions. A common simplification to this identification problem is to consider the expectation of the outcome. This is sometimes referred to as the 'mean' equation, or a model of the first raw moment of the population distribution. Taking conditional expectations of the DGP in equation 3 and assuming mean independence between the errors and treatment yields:

$$\begin{aligned}
\mathbb{E}[Y_i|D=1] =& \mathbb{E}[(\alpha_1 + \tau_i D_i + \varepsilon_i)|D=1] \\
=& \mathbb{E}[\alpha_1|D=1] + \mathbb{E}[\tau_i|D=1]D_i + \mathbb{E}[\varepsilon_i|D=1] \\
=& \alpha_1 + \tau_1 D_i
\end{aligned} \tag{4}$$

The mean equation 4 assigns a common coefficient $\mathbb{E}[\tau_i|D=1] = \tau_1$ to all treated individuals. Given mean independence $\mathbb{E}[\varepsilon_i|D=1] = \mathbb{E}[\varepsilon_i]$, τ_1 identifies how the mean of the outcome distribution Y for the treated population changes in response to being treated (the ATT). However, we can reveal additional information about how treatment changes the unconditional distribution of Y by considering expectations of polynomial transformations of

Y_i. Let us consider what is revealed about the distribution of Y from an equation for the second raw moment:

$$\begin{aligned}
\mathbb{E}[Y_i^2|D=1] =& \mathbb{E}[(\alpha_1+\tau_i D_i+\varepsilon_i)^2|D=1] \\
=& \mathbb{E}[\alpha_1^2]+\mathbb{E}[\tau_i^2 D_i^2|D=1]+\mathbb{E}[\varepsilon_i^2|D=1] \\
& +2\mathbb{E}[\alpha_1]\mathbb{E}[\tau_i D_i|D=1]+2\mathbb{E}[\alpha_1]\mathbb{E}[\varepsilon_i|D=1]+2\mathbb{E}[\tau_i D_i \varepsilon_i|D=1] \\
=& \alpha_1^2+\mathbb{E}[\varepsilon_i^2|D=1]+\mathbb{E}\left[(\tau_i^2+2\alpha_1\tau_1+2\tau_i\varepsilon_i)|D=1\right]D_i+2\alpha_1\mathbb{E}[\varepsilon_i|D=1] \\
=& \alpha_2+\tau_2 D_i
\end{aligned} \tag{5}$$

where $\alpha_2 = \alpha_1^2+\mathbb{E}[\varepsilon_i^2|D=1]$ and $\tau_2 = \mathbb{E}\left[(\tau_i^2+2\alpha_1\tau_1+2\tau_i\varepsilon_i)|D=1\right]$. τ_2 provides an unbiased estimate of the ATT on Y^2, conditional on first raw moment independence $\mathbb{E}[\varepsilon_i|D=1]=\mathbb{E}[\varepsilon_i]$ and second raw moment independence $\mathbb{E}[\varepsilon_i^2|D=1]=\mathbb{E}[\varepsilon_i^2]$.

Equation 5 tells us more about how treatment affects the distribution of outcomes. To illustrate the information gained from analysing higher-order raw moments, let us consider the PERM estimate of the variance. PERM utilises consistent estimates of observed and counterfactual first and second raw moments to provide the VTT:

$$\begin{aligned}
\Delta_{VTT}[Y|D=1] =& Var[Y_1|D=1]-Var[Y_0|D=1] \\
=& (\mathbb{E}[Y_1^2|D=1]-\mathbb{E}[Y_1|D=1]^2)-(\mathbb{E}[Y_0^2|D=1]-\mathbb{E}[Y_0|D=1]^2) \\
=& (\alpha_2+\tau_2-(\alpha_1+\tau_1)^2)-(\alpha_2-\alpha_1^2) \\
=& (\alpha_2+\tau_2-(\alpha_1^2+\tau_1^2+2\alpha_1\tau_1))-(\alpha_2-\alpha_1^2) \\
=& (\tau_2-(\tau_1^2+2\alpha_1\tau_1))
\end{aligned} \tag{6}$$

Equation 6 illustrates that for treatment to impact the variance, then $\tau_2 \neq (\tau_1^2+2\alpha_1\tau_1)$. Substituting $\tau_2 = \mathbb{E}[(\tau_i^2+2\alpha_1\tau_1+2\tau_i\varepsilon_i)|D=1]$ into equation 6 and simplifying gives:

$$\Delta_{VTT}[Y|D=1] = (\mathbb{E}[(\tau_i^2|D=1]-\tau_1^2)+\mathbb{E}[2\tau_i\varepsilon_i|D=1] \tag{7}$$

So while the two components of τ_2, namely $\mathbb{E}[\tau_i^2|D=1]$ and $\mathbb{E}[2\tau_i\varepsilon_i|D=1]$ are themselves not individually identifiable, they do shed light on how treatment can impact the variance and thereby highlights the information contained in τ_2. Specifically, the variance changes due to both the variability in treatment effects $(\mathbb{E}[(\tau_i^2|D=1]D_i-\tau_1^2)$ and how the distribution of individual treatment effects are related to the distribution of their untreated outcomes $(\mathbb{E}[2\tau_i\varepsilon_i|D=1])$.

This example demonstrates how estimates of τ_1 and τ_2 from regressions of the first and second raw moments provide greater information regarding the impact of treatment on the (unconditional) distribution of outcomes, compared to just an estimate of τ_1 alone. It follows that additional estimates of ATTs from regressions of higher-order polynomial transformations of the outcome variable provide more and more information. PERM combines these informative estimates of raw moment treatment effects to provide consistent estimates of the impact of treatment on distributional parameters of the outcome.

Here we have simplified the exposition by not including covariates in the DGP. However, it follows from assumption 3 and proposition 1 that the same approach could be applied to a model with covariates.

1.4 Differences-in-Differences Identification of Raw Moments

In this section we consider a differences-in-differences (DiD) set-up to identify counterfactual raw moments and thereby identify counterfactual distribution parameters using PERM. Let us consider a reform whose implementation depends on the time period T and group G. Without any loss of generality, let $(g,t) \in \{0,1\} \times \{0,1\}$ so that there are only two time periods and two groups. Let us consider a simple DGP, where $\delta_{i,g}$ is an individual-specific group effect, $\rho_{i,t}$ is an individual-specific time effect, $\tau_{i,g,t}$ is an individual treatment effect, and $\alpha_{i,g,t}$ is the sum of the intercept and error term. We can express our outcome variable as the sum of these terms:

$$Y_{i,g,t} = \alpha_{i,g,t} + \delta_{i,g} + \rho_{i,t} + \tau_{i,g,t} \tag{8}$$

Equation 8 is not identifiable. We therefore cannot estimate the full distribution of $Y_{i,g,t}$. Instead, let us assume we are interested in the first raw moment, the average treatment effect on the treated (ATT). To estimate the ATT, the following DiD assumption is often made.

Assumption 4 (Parallel trends in the first raw moment). *Let* $\alpha_1 = \mathbb{E}[\alpha_i]$ *be an intercept,* $\mathbb{E}[\delta_{i,g}]$ *be a first raw moment time-invariant group effect and* $\mathbb{E}[\rho_{i,t}]$ *be a first raw moment time effect common to all groups, then* $\mathbb{E}[Y_{01,i,g,t}|g,t] = \alpha_1 + \mathbb{E}[\delta_{i,g}] + \mathbb{E}[\rho_{i,t}]$.

Assumption 4 states that the counterfactual first raw moment is determined by an additive time-invariant mean group effect and an additive mean common time effect across groups. If assumption 4 holds we are able to identify the ATT.

DiD performed on higher-order raw moments tells us more about the distributional impacts of a social policy intervention. Estimation of the variance treatment effect on the treated

(VTT) using PERM DiD requires consistent estimates of both the first raw moment and the second raw moment. A possible assumption one could make using PERM DiD is the assumption of parallel variances. Assuming parallel group variances states that the change in the group variances in the control group is informative as to how the group variances would have changed in the treatment group in the absence of treatment.

Assumption 5 (Parallel trends in each group's variance). *Let α_2 be an intercept term, ω be a variance time-invariant group effect, γ be a variance time effect common to both treatment and control groups, then the counterfactual variance is given by $\mu_2(Y_{01,i,g,t}|g,t) = \alpha_2 + \omega + \gamma$.*

PERM provides a consistent estimate of the population variance treatment effect if there are consistent estimators for the relevant raw moments. PERM DiD can provide a consistent estimate of the second raw moment under assumptions 4 and 5, but an adjustment is required.

Proposition 2. ***Consistent estimation of the counterfactual second raw moment:*** *Given assumptions 4 and 5, then the counterfactual second raw moment is given by $\mathbb{E}[Y^2_{01,i,g,t}|g,t] = \alpha_2 + \omega + \gamma + 2\,\mathbb{E}[\delta_{i,g}]\,\mathbb{E}[\rho_{i,t}]$*

The proof is found in Appendix A. Proposition 2 states that the counterfactual second raw moment is determined by an additive time-invariant second raw moment group effect and an additive second raw moment common time effect across groups plus an additional term $2\,\mathbb{E}[\delta_{i,g}]\,\mathbb{E}[\rho_{i,t}]$. If the first raw moment is changing over time but parallel in both groups, and the variance is parallel between the treatment and control group, then $\mathbb{E}[Y^2]$ has to change by $2\,\mathbb{E}[\delta_{i,g}]\,\mathbb{E}[\rho_{i,t}]$. A standard DiD of the second raw moment would not adjust for $2\,\mathbb{E}[\delta_{i,g}]\,\mathbb{E}[\rho_{i,t}]$ and would therefore be biased if a time trend exists in the first raw moment, $(\mathbb{E}[\rho_{i,t}] \neq 0)$, and there are differences across the group's first raw moments, $(\mathbb{E}[\delta_{i,g}] \neq 0)$. Assumptions 4 and 5 enable PERM DiD to adjust a standard DiD of the second raw moment for $2\,\mathbb{E}[\delta_{i,g}]\,\mathbb{E}[\rho_{i,t}]$ estimated from the DiD of the first raw moment. PERM DiD applies this adjustment and thereby provides the variance treatment effect on the treated (VTT).

Assumption 5 enables identification of treatment effects on *group* level variances. However, the analysis of *group* level variances by itself does not reveal the impact on the *population* level variance because such an analysis ignores the between group effects. This follows from the variance decomposition formula, which states the population variance can be decomposed as the average of within group variances plus the variance of group means.[8] Analysis of just within group variance ignores the variance of the means, and therefore does not provide the full information needed to understand the effect of treatment on the population

[8]The law of total variance provides how the variance can be decomposed: $V(Y) = \mathbb{E}[V(Y|G)] + V(\mathbb{E}[Y|G])$

variance. PERM DiD, by utilising proposition 2 and the assumptions therein, allows the analysis of the population's first and second raw moments and thereby the population level variance.

For higher-ordered central moments, further assumptions are required in order to identify the PTT parameter using PERM DiD. For the unstandardised skewness, a potential source of information about how the treatment group's unstandardised skewness would have changed in the absence of treatment is how the unstandardised skewness of the control group changed over time.

Assumption 6 (Parallel trends in each group's unstandardised skewness). *Let α_3 be an intercept term, θ be a unstandardised skewness time-invariant group effect and ϕ be a unstandardised skewness time effect common to both treatment and control groups, then the counterfactual unstandardised skewness of the treated is given by $\mu_3(Y_{01,igt}|g,t) = \alpha_3 + \theta + \phi$.*

Proposition 3. ***Consistent estimation of the counterfactual third raw moment:*** *Given assumptions 4, 5, and 6, then the counterfactual third raw moment is given by: $(\mathbb{E}[Y^3_{01,i,g,t}|g,t]) = \alpha_3 + \theta + \phi + 3\gamma\mathbb{E}[\delta_{i,g}] + 3\omega\,\mathbb{E}[\rho_{i,t}] - 6\alpha_1\,\mathbb{E}[\delta_{i,g}]\,\mathbb{E}[\rho_{i,t}]$*

The proof of proposition 3 is relegated to Appendix A.

Proposition 3 states that if the mean, group variance and group unstandardised skewness are changing over time but are parallel across groups, then the counterfactual third raw moment, $\mathbb{E}[Y^3]$, in the treatment group has to change by $3\gamma\mathbb{E}[\delta_{i,g}] + 3\omega\,\mathbb{E}[\rho_{i,t}] - 6\alpha_1\,\mathbb{E}[\delta_{i,g}]\,\mathbb{E}[\rho_{i,t}]$ in the absence of treatment. PERM DiD can then be used to calculate the *population* unstandardised skewness. Note, PERM DiD of the third raw moment that follows from proposition 3 allows PERM DiD estimation of not only the unstandardised skewness, but also the standardised skewness. This follows from the definition of the skewness as the ratio of the unstandardised skewness and the standard deviation cubed. PERM DiD can be used to estimate either definition of the skewness under these assumptions.

PERM DiD can also be used to consider multivariate outcomes such as the covariance or the slope of two outcomes, W and Y, if there are parallel trends in the group's covariance.

Assumption 7 (Parallel trends in each group's covariance). *Let α_{WY} be an intercept term, ω_{WY} be a covariance time-invariant group effect and γ_{WY} be a covariance time effect common to both treatment and control groups, then $\mu_{WY}[W_{01,i,g,t}Y_{01,i,g,t}|g,t] = \alpha_{WY} + \omega_{WY} + \gamma_{WY}$.*

Proposition 4. ***Consistent estimation of the joint (bivariate) raw moment:*** *Given assumption 4 holds for both outcomes Y and W and assumption 7 holds, then the counterfactual joint raw moment is given by: $\mathbb{E}[W_{01,i,g,t}Y_{01,i,g,t}|g,t] = \alpha_{WY} + \omega_{WY} + \gamma_{WY} +$*

$\mathbb{E}[\delta_{i,g,W}]\,\mathbb{E}[\rho_{i,t,Y}] + \mathbb{E}[\delta_{i,g,Y}]\,\mathbb{E}[\rho_{i,t,W}]$

The proof of proposition 4 is relegated to Appendix A.

2 Estimation of PERM

In this section we consider the empirical estimation of PERM Regression and PERM DiD, including its small sample properties as well as alternate ways of estimating standard errors.

2.1 Estimation of PERM Regression

PERM Regression is the combination of regression and the PERM method to estimate the DPTT. Assume the following set of raw moment equations for functions of the outcome variable Y^q:

$$\begin{aligned} y_i &= \alpha_1 + \tau_1 d_i + X_i'\beta_1 + (d_i \times X_i)'\zeta_1 + r_{1i} \\ y_i^2 &= \alpha_2 + \tau_2 d_i + X_i'\beta_2 + (d_i \times X_i)'\zeta_2 + r_{2i} \\ &\vdots \\ y_i^q &= \alpha + \tau_q d_i + X_i'\beta_q + (d_i \times X_i)'\zeta_q + r_{qi} \end{aligned} \tag{9}$$

where y_i is the outcome variable of individual i; X_i is a vector of observables; β is the return to observables; r_i is the residual, y_i^q is a functional form applied to y_i, and q is a positive integer polynomial. These are our raw moment regressions of Y^q which allow us to predict both the observed and counterfactual raw moments under assumptions 1 and 2, or under the weaker assumptions of 2 and 3.

In eq. 9 we fully interact treatment status with covariates. This is because the ATT estimated without fully interacting with covariates can yield a biased estimator if there are heterogenous treatment effects (Słoczyński 2022). Fully interacting treatment with covariates solves this (Wooldridge 2010). This holds for all raw moments.

To provide PERM regressions we use the sample analogues to these formulas. In our Monte Carlo simulation exercise and empirical example, we use OLS to estimate the raw moment regressions in eq. 9. Standard errors for the treatment effects can be obtained via estimating all raw moments in a seemingly unrelated regression framework in combination with the linearization technique of (Graubard and Korn 1999). This approach requires us to estimate unconditional standard errors for our predicted means that account for the uncertainty in the explanatory variables at the population level. Non-parametric bootstrap standard errors and bootstrap 95% percentile confidence intervals are also alternative approaches that can be used. We compare these alternatives in the following Monte Carlo simulation.

2.2 Estimation of PERM DiD

We utilise the method of De Chaisemartin and d'Haultfoeuille (2024) to illustrate PERM DiD utilising a staggered DiD set-up. Our aim is to provide event study figures and event-study regressions for a binary treatment whose implementation is staggered across a group variable. The population distribution parameters we consider are the mean, variance, un-standardised and standardised skewness, observed for a cross-section where time elapsed since the reform was first implemented, ℓ, varies depending the date of first treatment for each treatment group. The event-study regression estimator provides a weighted average of the effect of treatment for those that were exposed ℓ periods after it was first implemented.[9]

The following notation defines the DiD set-up as introduced in De Chaisemartin and d'Haultfoeuille (2024) utilising their notation and definitions, followed by the introduction of the PERM DiD estimator for any raw moment. Let us consider a "sharp" binary reform where the implementation is staggered over time periods T and groups G. We assume the use of individual level data, and the total number of observations, N, are divided across every $(g,t) \in \{1,\ldots,G\} \times \{1,\ldots,T\}$ so that there are $N_{g,t}$ observations in each group g and time period t. From hereon in the $N_{g,t}$ notation is suppressed to improve legibility, noting that it is just a mechanical exercise to extend the estimators to include weights. Let $D_{g,t}$ denote the treatment of group g at period t. Let $\boldsymbol{D}_g = (D_{g,1},\ldots,D_{g,T})$ be a vector stacking g's treatments from period 1 to T, and let $\boldsymbol{D} = (\boldsymbol{D}_1,\ldots,\boldsymbol{D}_G)$ be a vector stacking the treatments of all groups 1 to G at every period.

The grouping variable is defined as those who are first treated in the same period. For all g, let $F_g = min\{t : t \geq 2, D_{g,t} \neq D_{g,t-1}\}$ be the first treated period for group g. For a school reform rolled out across states and time, this would entail grouping by year of reform implementation. In the DiD design we consider all groups are initially untreated, so there is at least one untreated observation for each group ($F_g \geq 2$).

The last period we are able to consider must also have a relevant control group that has remained untreated throughout the period. For every g, let $T_g = \max_{g':D_{g',1}=D_{g,1}} F_{g'} - 1$ denote the last period where there is still a group who were untreated in period-one, as g, and and have remained untreated. Consequently, if all g are eventually treated then the last treated group

[9] Note that the methods of De Chaisemartin and dHaultfoeuille (2020), De Chaisemartin and d'Haultfoeuille (2024) allow application to a broader set of empirical designs than the design we consider. We leave the alternative designs and their implementation with PERM for future research.

is dropped due to lack of a relevant control group.

A treatment group g in period t can have a number of relevant control groups that have remained untreated between the first period and period t. For any finite set A, let $\#A$ be the number of elements of A. For all (g,t), let $N_t^g = \#\{g' : D_{g',1} = D_{g,1}, F_{g'} > t\}$ be the number of control groups g' that like g were untreated in period-one and if they were treated, this treatment happened after period t.

The DiD estimate comparing the period just before treatment implementation $F_g - 1$ to the outcome evolution of g ℓ periods later $F_g - 1 + \ell$ to that of groups that have remained untreated from period 1 to $F_g - 1 + \ell$ for any raw moment $\mathbb{E}[Y^q]$ for all $q \in \mathbb{Z}^+$ is given by:

$$DID_{g,\ell}(Y^q) = Y^q_{g,F_g-1+\ell} - Y^q_{g,F_g-1} - \frac{1}{(N^g_{F_g-1+\ell})} \sum_{g':D_{g',1}=D_{g,1},F_{g'}>F_g-1+\ell} (Y^q_{g',F_g-1+\ell} - Y^q_{g',F_g-1}) \tag{10}$$

Note that $DID_{g,\ell}(Y^q)$ from eq. 10 uses groups' $F_g - 1$ outcome as the baseline always, not an average of their period 1 to $F_g - 1$ outcomes. For $q = 1$, eq. 10 is the DiD estimator of De Chaisemartin and d'Haultfoeuille (2024) and is an unbiased and consistent estimator of the ATT.

For any $q > 1$, eq. 10 is a biased DiD estimator if there are trends and pre-treatment group raw moment differences in lower order moments. Propositions 2 - 4 state that this bias can be accounted for. To account for this bias, estimates of the differences in the relevant raw moments between group g and their counterfactual group(s) g' in period $F_g - 1$ are required. For each raw moment and g the mean group difference for raw moment Y^q is given by:

$$\Delta_{g,F_g-1}(Y^q) = Y^q_{g,F_g-1} - \frac{1}{(N^g_{F_g-1+\ell})} \sum_{g':D_{g',1}=D_{g,1},F_{g'}>F_g-1+\ell} (Y^q_{g',F_g-1})$$

For $q = 1$ this gives $\mathbb{E}[\delta_i]$, for $q = 2$ this gives ω , and for $q = 3$ this gives θ. Propositions 2 - 4 also state that estimates of the time trend between period $F_g - 1$ and period ℓ for all relevant raw moments for group(s) g' are required. For each raw moment and ℓ, the time trend in group(s) g' is given by:

$$\Delta_{g',\ell}(Y^q) = \frac{1}{(N^g_{F_g-1+\ell})} \sum_{g':D_{g',1}=D_{g,1},F_{g'}>F_g-1+\ell} (Y^q_{g',F_g-1+\ell} - Y^q_{g',F_g-1})$$

For $q = 1$ this gives $\mathbb{E}[\rho_i]$, for $q = 2$ this gives γ, and for $q = 3$ this gives ϕ.

PERM DiD adjusts $DID_{g,\ell}(Y^q)$ for parallel trends in lower order moments using a function of the parameters $\Delta_{g,F_g-1}(Y^q)$ and $\Delta_{g',\ell}(Y^q)$:

$$PERM\ DID_{g,\ell}(Y^q) = DID_{g,\ell}(Y^q) - f(\Delta_{g,F_g-1}(Y^q), \Delta_{g',\ell}(Y^q)) \tag{11}$$

where $f(\Delta_{g,F_g-1}(Y^q), \Delta_{g',\ell}(Y^q))$ is defined by propositions 2-4 for the second and third and joint first raw moments under parallel mean, variance, skewness and covariance assumptions respectively.

Let $L = max_g(T_g - F_g + 1)$ denote the largest ℓ such that $\delta_{g,\ell}$ can be estimated for at least one g. Under design restriction $L \geq 1$. For every $\ell \in \{1, \ldots, L\}$, let $N_\ell = \#\{g : F_g - 1 + \ell \leq T_g\}$ be the number of groups for which $PERM\ DID_{g,\ell}(Y^q)$ can be estimated. The PERM DID estimate of the average effect of being exposed for ℓ periods is therefore:

$$PERM\ DID_\ell(Y^q) = \frac{1}{N_\ell} \sum_{g:F_g-1+\ell \leq T_g} PERM\ DID_{g,\ell}(Y^q). \tag{12}$$

De Chaisemartin and d'Haultfoeuille (2024) show that $PERM\ DID_\ell(Y^{(q=1)})$ is a consistent estimator of the ATT. It then follows that, conditional on the additional assumptions and adjustment for trends in lower order raw moments, that PERM DiD will also consistently estimate the ATT of higher-order raw moments. As shown in Proposition 1, Slutsky's Theorem ensures that by combining these consistent raw moment estimates one can consistently estimate the DPTT.

We provide the Stata package *did_multiplegt_PERM* to estimate PERM DiD, which is an adaptation of De Chaisemartin and dHaultfoeuille (2020)'s Stata package *did_multiplegt*. *did_multiplegt_PERM* provides the extra terms set out in propositions 2, 3 and 4 that are required to estimate variance and skewness treatment effects as well as the treatment effect on the covariance of two outcomes on the treated using PERM DiD. Like De Chaisemartin and dHaultfoeuille (2020) we use a non parametric bootstrap to provide standard errors for the DiD estimate of the DPTT by bootstrapping the whole PERM DiD procedure.

The Stata package *did_multiplegt* provides the option of estimating the ATT for each value of ℓ, and these can be used to provide event study figures. Event study figures are useful to understand the dynamics of treatment over time and can also help explore the validity of the parallel trend assumptions by assessing differences in the pre-treatment period. Roth (2024) suggests using the *long differences* (default) option in *did_multiplegt* to compare pre-treatment trends between $F - (1 + b)$ and $F - 1$, where b is time before the reference period. This method of assessing parallel trends in the pre-treatment period is straight forward for the first raw moment.

The Stata package *did_multiplegt_PERM* allows extension of event study figures to the variance, skewness and covariance. PERM DiD based event study figures plot the DPTT for all relevant treatment and control groups in each period. The same event study figures are valid for testing for pre-trends. As shown in Appendix D, the DPTT for any period $F-(1+b)$ is informative of group parallel trends for variance, skewness and covariance. A zero DPTT estimate in the pre-treatment period for the variance, skewness and covariance is consistent with parallel trends at the group level in the respective distribution parameters. *did_multiplegt_PERM* based event study figures are therefore both informative going forward from the reference period but also backwards assessing the pre-treatment period.

2.3 Properties of the PERM Sample Estimator

Just like the standard sample variance and sample covariance estimators, PERM's sample variance and covariance estimates are biased in small samples. This is because the PERM estimates rely on the sample estimate of the mean which itself is measured with error. In the variance formula the squaring of the sample mean which includes its measurement error introduces bias, and in the covariance formula the cross multiplication of the two sample means and their included error terms has the potential to introduce bias.

In Appendix B we derive the small sample bias for the sample (co)variance and sample skewness and propose a correction that uses the standard errors of the raw moments. Because the DPTT is the difference between the treated and counterfactual distribution parameters, the bias of the DPTT will be the difference in the biases in each term, and because both biases will be in the same direction as each other, the bias of the DPTT will likely be of minor concern compared to the overall precision of the DPTT.

In our empirical estimation of the small sample bias corrected DPTT we utilise both the linearisation technique and bootstrapping with replacement to provide estimates of the measurement errors of the raw moments, that are in turn used to correct the PERM DPTT estimates.[10]

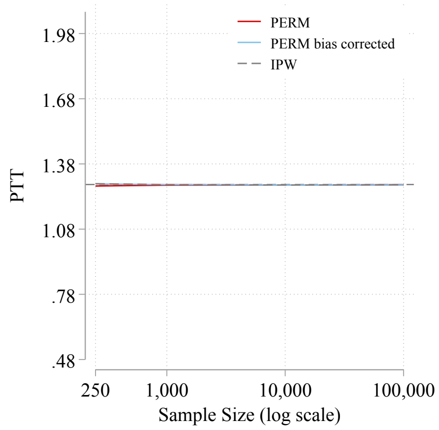
3 Monte Carlo Exercise

In this section, we aim to better understand the performance of our PERM regression estimator in small samples and to compare it with the Inverse Probability Weighting (IPW) estimator using a Monte Carlo Exercise. The experiment is designed around a data generating process (DGP) that models selection into treatment based on observables. The DGP is an exact replica of the DGP used in Firpo and Pinto (2016) and is chosen because it allows

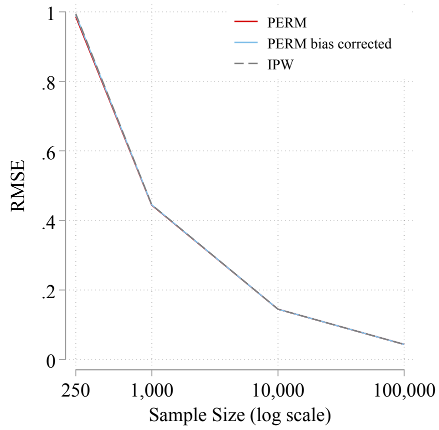
[10]Note that the standard errors for the bias corrected PERM sample (co)variance estimates do not account for the uncertainty in the bias correction term. Bootstrapping the whole procedure would address this, but it turns out that bias is of low order importance compared to the overall precision in all the examples considered (i.e. if one has enough data to accurately estimate the impact of treatment on the distribution of outcomes then the small sample bias is likely to be very small).

PERM regression to be compared with the wider simulation results from other distributional methods presented in Firpo and Pinto (2016).

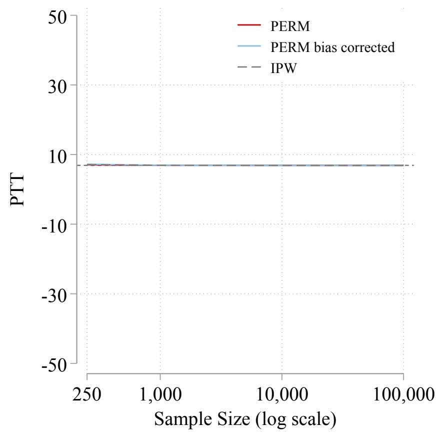
The DGP has a very simple set-up consisting of two explanatory variables, $X = [X_1, X_2]^T$, which determine both the treatment status D and the potential outcomes Y_0 and Y_1. The observed outcome is therefore $Y = Y_0(1-D) + D \cdot Y_1$. Full details of the Monte Carlo simulation setup are provided in Appendix E. The simulated data are non-normal and highly positively skewed, but has a much more symmetrical distribution in logs (see Figure E.1 in Appendix E).



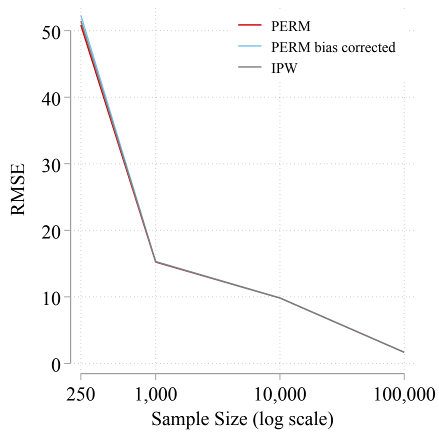
(a) Bias Variance



(b) RMSE Variance



(c) Bias Skewness



(d) RMSE Skewness

Figure 1: Monte Carlo Exercise - Bias and Root Mean Squared Error

Notes: Figures (a), (c) and (e) plot a line of the average sample estimator of the Parameter Treatment effect on the Treated from the Monte Carlo Exercise with varying sample size and 20,000 replications for each sample size versus the 'truth' (horizontal dotted line). Figures (b), (d), (f) plot Root Mean Squared Error. Estimates are provided by two methods, PERM and IPW. PERM Bias corrected utilises the sample bias correction from section 2.3 and Appendix B. The y-axis scale for the bias figures is the true value +- 1 RMSE.

We extend the Monte Carlo exercise of Firpo and Pinto (2016) to consider the variance (VTT), skewness (STT) and standardised skewness (SSTT) treatment effect on the treated.

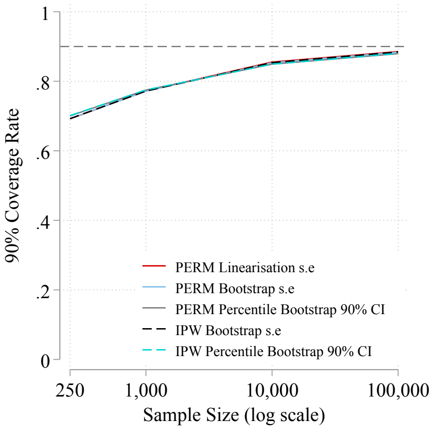
Two criteria are used to judge the estimates: bias (average difference between the small sample estimate and the true value); and precision as measured by root mean squared error (RMSE). An infeasible estimator is calculated using the 'unobserved' potential outcomes from the DGP for the given sample size. Because the DGP does not have an easily derived closed analytical form, the true, or target values, are calculated based on the average of an unfeasible estimator with 1000 replications of sample size 10,000,000. A naive estimator is calculated as the raw difference between the treated and untreated groups, assuming no selection into treatment. Comparison of the infeasible and naive estimators provides a sense of the selection bias.

A replication of Firpo and Pinto (2016) for the mean and the coefficient of variation is presented in Appendix F Table F.1 that considers 100,000 replications of sample sizes of 250 and 1,000. The Monte Carlo simulation results show that both IPW and PERM regression yield very similar results when assessed for bias and RMSE, even under miss-specification. This, combined with the conclusion of Firpo and Pinto (2016) that IPW performs well compared to the methods of Juhn, Murphy and Pierce (1993) and Chernozhukov, Fernández-Val and Melly (2013) in terms of bias and RMSE, suggests PERM regression performs well against these methods too. The replication is extended to the variance and shows that IPW and PERM regression yield similar results, both in terms of bias and RMSE. Finally, the small sample bias correction of the PERM variance estimator is shown to perform well, although the bias is very small compared to the statistical precision.

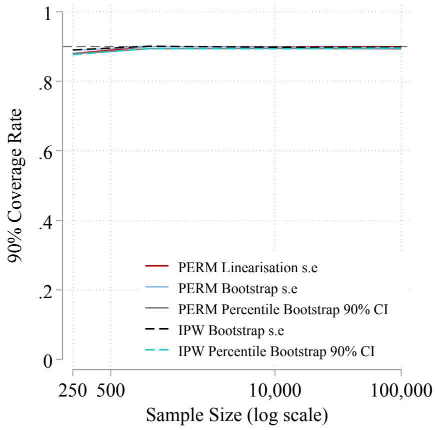
Figure 1 presents the Monte Carlo simulation estimates from 20,000 replications for sample sizes 250 through to 100,000 of the bias and the RMSE for both the VTT and STT.[11] According to the bias criteria, both PERM regression and IPW based estimates are similar and perform well in all sample sizes. The similarity of the two methods also extends to the RMSE criteria. Both methods show a clear downward trend in RMSE towards zero as the sample size increases to 100,000. A decreasing RMSE is consistent with a sample estimator that approaches the true value with reducing variance as sample size increases. In Appendix G Table G.1 we present results for 100,000 replications of sample sizes 250 and 1,000 to see in more detail the small sample properties of the estimators. Across the VTT and STT the sample bias for the PERM estimator is very similar to the infeasible estimator and very small relative to the level of precision. Again PERM and IPW show very similar results.

In figure 2 we consider the coverage rates for alternative 90% confidence intervals from PERM and IPW estimators for DPTTs from the Monte Carlo simulation of 20,000 replications for sample sizes of 250 through to 100,000. PERM and IPW 90% confidence intervals are estimated based on the standard errors estimated using non-parametric bootstrap with

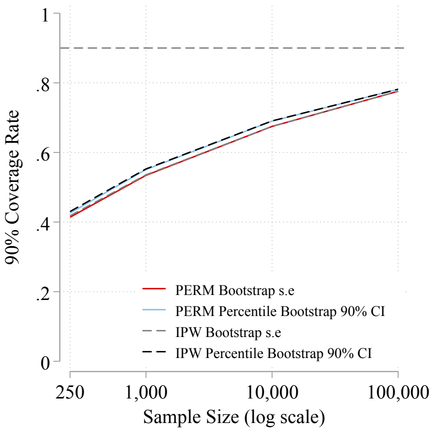
[11]To calculate the STT using IPW we multiplied the standardised statistics by the variance estimate raised to the relevant power, as there was no standard procedure available in Stata for these statistics.



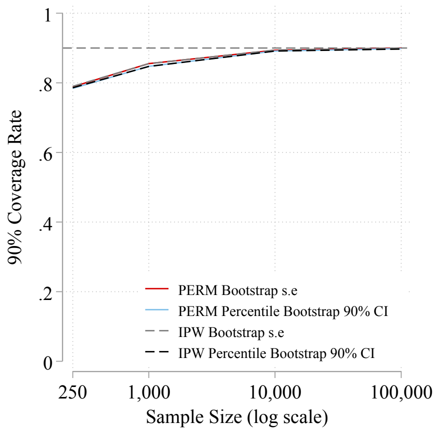
(a) Variance Y



(b) Variance ln(Y)



(c) Skewness Y



(d) Skewness ln(Y)

Figure 2: Monte Carlo Exercise - 90% Coverage Rates

Notes: These figures plot a line of the 90% confidence interval coverage rates of the Parameter Treatment effect on the Treated from the Monte Carlo Exercise with varying sample size and 20,000 replications for each sample size. Estimates are provided by two methods, PERM and IPW and alternative methods of standard error estimation. PERM linearisation standard errors are provided by the linear approximation method of Graubard and Korn (1999). Bootstrap standard errors and Percentile Bootstrap 90% Confidence Intervals are from a non-parametric bootstrap with 200 replications.

200 replications, as well as by using the percentile method. For VTT, PERM 90% confidence intervals are also provided by the standard errors estimated using the linearisation technique following simultaneous regression of Y and Y^2 to estimate the VTT.[12]

The coverage rates from the main Monte Carlo simulation are reported in figure 2 panels (a), (c) and (e). Panel (a) reports results for the VTT and shows that the coverage rates for all estimation methods are relatively low $(< .7)$ in small samples, but coverage does improve with increasing sample size in a very similar way for both PERM and IPW. The small sample

[12]We only use the linearisation technique to estimate standard errors for the VTT as Stata fails when simultaneously estimating higher-order raw moments for our DGP we use, suggesting this method may be of limited practical use beyond the variance.

coverage rates are even lower for STT (panel (c)) with both PERM and IPW again showing very similar results but again both improve in a similar way for larger sample sizes. The choice of bootstrap standard errors to calculate the confidence intervals or bootstrap 90% percentile confidence intervals yields very similar conclusions, and again this is true for both PERM and IPW.

The results in figure 2 panels (a) and (c) show that statistical inference for all standard error approaches is relatively poor in small samples when the distribution of the outcome variable is highly skewed (non-normal) for both IPW and PERM. Panels (b) and (d) show that standard errors of parameters of an outcome distribution that is more normally distributed (i.e. the logged outcome in this case) achieve better coverage in smaller samples, and this true for both PERM and IPW.[13] Further results for the log transformation of Y are presented in appendix G summarising the bias and RMSE results. Appendix G also presents results for the standardised skewness, showing similar performance of PERM regression to that of IPW. The results in Appendix G also illustrate the well-known small sample bias of the estimators for standardised skewness Joanes and Gill (1998).

4 Empirical Application I: Unionisation in the USA

Our first empirical application of PERM allows further comparison with IPW using non-simulated data to confirm the conclusions drawn from the Monte Carlo exercise. Following a long line of research on distributions that relies on a selection on observables assumption (see e.g. Card et al. (2004), Card, Lemieux and Riddell (2020), Firpo, Fortin and Lemieux (2009)), we also consider how unionisation coverage impacts the distribution of hourly log wages for those covered by unionisation in the US. We maintain the assumption that unionisation status is exogenous, conditional on observables, to allow comparison with the IPW estimation procedure, acknowledging that this assumption may introduce a bias.

Data is from a sample of 266,956 U.S. males from the 1983 - 1985 Outgoing Rotation Group (ORG) supplement of the Current Population Survey as used in Firpo, Fortin and Lemieux (2009).[14] Hourly log wages in 1979 dollar terms are provided alongside information on union coverage status and other factors that may be associated with both union coverage and wages; ethnicity, marital status, level of education and years of experience. PERM regression follows the methodology described in section 2.1. IPW estimation follows the

[13] An alternative method that could potentially improve inference performance suggested by Wilcox (2012) is to estimate bootstrap-t confidence intervals. The cost of this approach is that it requires a procedure that provides bootstrap standard errors for each bootstrap replication of the t-statistic. That is, it requires a bootstrap of a bootstrap, which extends processing time substantially.

[14] We use the replication files for Firpo, Fortin and Lemieux (2009) provided here `https://sites.google.com/view/nicole-m-fortin/data-and-programs`.

methodology described in Firpo and Pinto (2016). More detailed description of the data and methods is found in Appendix H.

Table 2: Impact of Unionisation on log Hourly Wages Inequality

	Observed Statistic	PERM	IPW
Mean:			
Union	1.9625	0.1770***	0.1792***
		(0.0019)	(0.0018)
Variance:			
Union	0.1655	-0.1655***	-0.1629***
		(0.0016)	(0.0015)
Standardised Skewness:			
Union	-0.1787	-0.4109***	-0.3945***
		(0.0183)	(0.0194)

Notes: This table presents the estimated impact of unionisation on the mean, variance and standardised skewness of the natural log of hourly wages in the USA. Each cell represents results from separate estimates with corresponding standard errors in parenthesis. PERM estimates utilise fully factorised controls for ethnicity, marriage status, education and experience and also considers that the treatment effect of unionisation may vary by these controls and thus includes these controls all fully interacted with unionisation status. IPW estimates utilise a probit regression of union status using the same factorised controls variables as used in PERM. PERM standard errors are provided by the linearization method and IPW standard errors are provided by bootstrapping (200 replications). *** p<0.01, ** p<0.05, * p<0.1.

Table 2 presents the PERM regression and IPW based DPTT estimates of the impact of union coverage on the mean, variance and standardised skewness. The PERM regression estimates show that for the population covered by unionisation, the mean of log hourly wages is higher but the variance and standardised skewness in log hourly wages are lower for those covered by union membership. These results are consistent with previous findings, that union coverage in the U.S. increased mean log wages, and also reduced log wage inequalities Card et al. (2004), Firpo, Fortin and Lemieux (2009). The results presented in table 2 show this reduction in inequality is due to a compression of the distribution which also makes it less skewed. The IPW based estimates are very similar to those estimated using PERM, confirming the finding from the Monte Carlo exercise, that PERM regression performs as well as IPW when selection on observables holds. The precision of both PERM and IPW are also very similar.

Firpo and Pinto (2016) show that IPW performs well against other distributional methods under selection on observables, and the results presented in this section show that PERM produces very similar results to IPW under the same empirical conditions. This suggests

PERM also performs well against other distributional methods under selection on observables. The advantage of PERM over IPW, as we shall illustrate, is that it can be extended to situations where selection on observables does not hold, which is not possible for IPW.

Instead of PERM or IPW, we could have also used alternative methods to understand the impact of union coverage. One popular alternative is Recentered Influence Function (RIF) regression, first introduced by Firpo, Fortin and Lemieux (2009), that like PERM is computationally simple.[15] However, a key difference and drawback of the RIF approach is that RIF regression only provides the partial policy effect (PPE), not the DPTT. While PERM can estimate the DPTT it can also be used to provide the PPE and thereby clearly illustrate the difference between the DPTT and PPE.

The PPE is the average impact of treatment on the first derivative of the distribution parameter. The PPE is therefore a linear approximation of the DPTT and will have approximation error. The size of the approximation error will depend on the degree of non-linearity of the distribution parameter function and the size of the treatment effect. PERM can be used to provide the PPE by calculating the differential of the distributional parameter expressed as a function of raw moments that is then estimated using PERM. In appendix J we illustrate this using the variance, describe the PERM procedure for the PPE and show that PERM PPE provides the same PPE as RIF. This result illustrates nicely that the PPE provided by RIF regression is only a linear approximation of the DPTT which could include substantial error for policies with wide-reaching effects.

Finally, for many, the results of Table 2 in terms of their magnitude may not be straightforward to conceptualise. In appendix I we present a distribution fitting exercise depicting the results of Table 2 visually illustrating how PERM based PTT results can be explained to a wider audience. As well as depicting the results graphically, one may be interested in whether particular groups are driving the results. The variance decomposition formula allows contributions of subgroups to be estimated, and this is straightforward with PERM regression and is illustrated in appendix K.

5 Empirical Application II: Inequality Impacts of a Compulsory School Reform

Our second empirical example illustrates PERM DiD applied in a staggered DiD setting. We consider a major school reform in Sweden that had the specific aim of reducing inequality. This reform attempted to reduce inequality by raising the minimum years of schooling

[15] For example Bossler and Schank (2023) apply RIF regression to investigate the impact of minimum wages on the variance of earnings in Germany.

from seven (or eight) to nine years and also by postponing entry into selective schooling (tracking), replacing it with comprehensive schooling for all abilities up to the ninth grade. The reform was rolled out slowly over time across municipalities (see Appendix L for more details regarding the reform).

A school system reform is an interesting application because it has the potential to substantially change the formation of education and earnings inequalities (Blanden, Doepke and Stuhler 2023). Many similar reforms were also introduced in other Western countries making the results of wider interest (Holmlund 2016). Sweden is especially interesting because it has achieved a high degree of earnings equality, relative to the US and its European counterparts. An improved understanding of how Sweden achieved this relatively low level of inequality would be useful. The Swedish school reform we consider has been extensively evaluated, with a focus on the mean, the first raw moment, of education and income (Meghir and Palme 2005, Fischer et al. 2022), intergenerational transmission of human capital (Lundborg, Nilsson and Rooth 2014, Lundborg and Majlesi 2018), health (Lager and Torssander 2012, Palme and Simeonova 2015, Meghir, Palme and Simeonova 2018, Fischer et al. 2021) and crime (Hjalmarsson, Holmlund and Lindquist 2015). There is no evidence of its impact on the unconditional distribution of education or earnings or on their joint distribution.

5.1 Data and Empirical Strategy

We use Swedish administrative data provided by Statistics Sweden covering the entire population born in Sweden during the years 1932-1952 and their parents.[16] Using personal identification numbers we are able to match individuals to various administrative data including the income and tax records for years 1968-2016, the national census for 1960 and 1970, population registers, and education records for years 1990-2016. For further details see Appendix M.

To identify the impact of the school reform, we utilise PERM DiD, as outlined in section 1.4, estimated following the empirical approach outlined in section 2.2. We are interested in the school reform's impact on the mean, variance, and standardised skewness of years of education, and earnings, as well as the covariance and slope coefficient (association) of education on income for those exposed to the reform. This requires identification of the impact of the comprehensive school reform on the first three raw moments, $\mathbb{E}[Y]$, $\mathbb{E}[Y^2]$ and $\mathbb{E}[Y^3]$ of the outcome distributions of education and income as well as their first joint raw moment, $\mathbb{E}[WY]$. Under assumptions 4, 5, 6 and 7, PERM DID provides credible estimates of the impact of the school reform on the first three observed and counterfactual

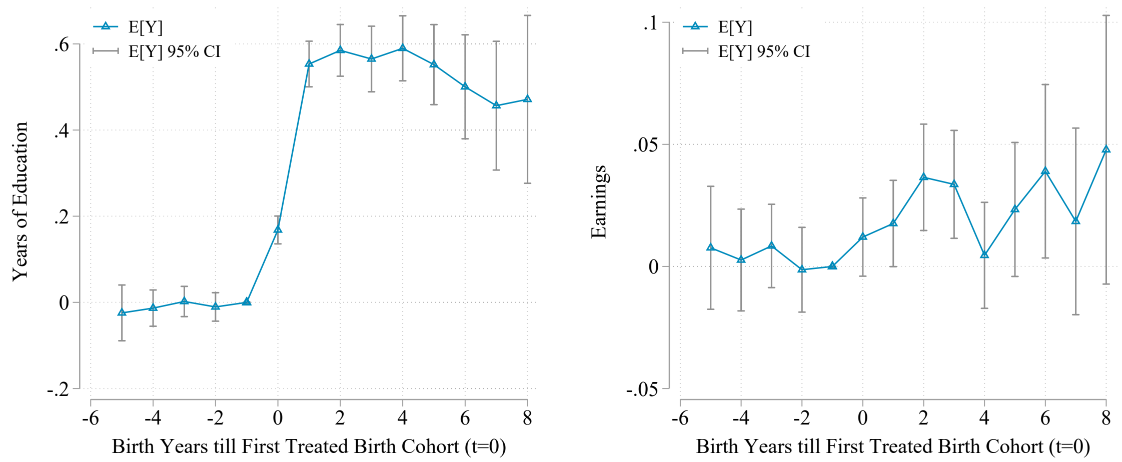
[16]We exclude immigrants to ensure individuals were in fact exposed to the reform

raw moments as well as their first joint raw moment. PERM utilises these estimates to provide the distribution parameter treatment effect on the treated.

As a visual presentation of our results we provide PERM DiD event study results that illustrate the impact since first implementation for the mean, variance, and skewness following the empirical approach set out in 2.2 that builds upon the DiD estimator of De Chaisemartin and dHaultfoeuille (2020). As noted in 2.2, event study figures for the mean, variance and skewness allow assessment of the DPTT going forward, and assessment of the parallel trends assumptions from 1.4 in the pre-period (Roth 2024). We provide event study figures with four periods pre-treatment and nine periods post-treatment, the longest our sample allows.[17]

The PERM DiD regression results are utilised to summarise the overall DPTT for the whole combined treated population presented in the event study figures. Standard errors and 95% confidence intervals are provided by 200 replications that bootstrap the whole PERM procedure, combining the Stata command *did_multiplegt_PERM* with PERM estimation of the DPTT parameters.

5.2 PERM Event Studies



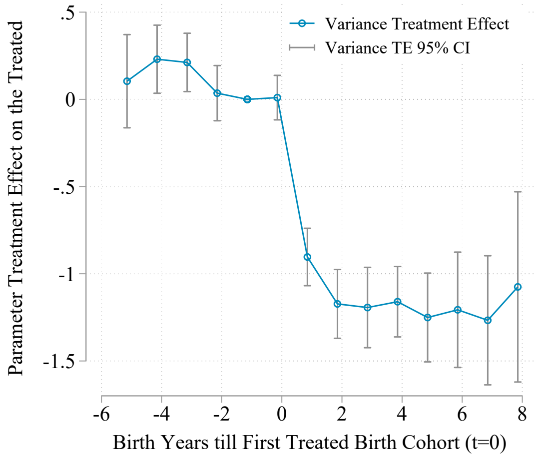
(a) Years of Education **(b)** Earnings (100,000 SEK)

Figure 3: Education and Earnings First Raw Moment Event Studies

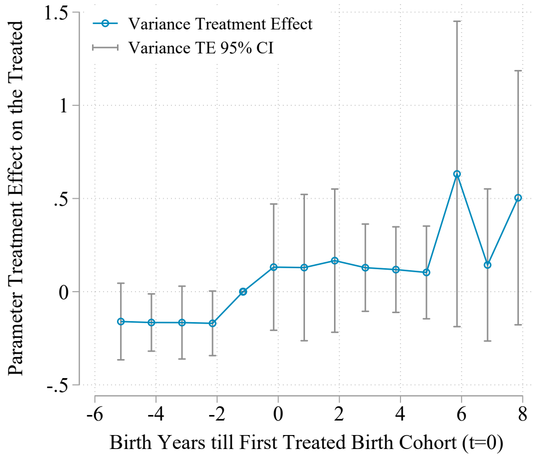
Notes: These figures plot the ATT by time since reform was first implemented in their municipality using the method described in section 2.2. 95%CI clustered by municipality of residence and provided by bootstrapping the whole PERM DiD procedure with 200 replications are shown as capped vertical lines.

In this section, we present PERM DiD event study evidence of the impact of the school reform on our parameters of interest for our outcomes years of education, and earnings.

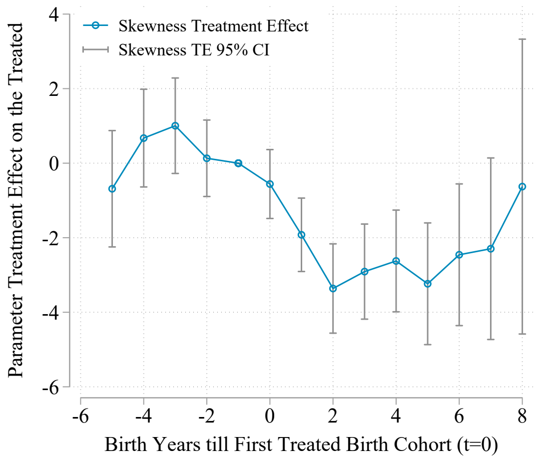
[17] Note we only need parallel group variance and parallel group (unstandardised) skewness assumptions to provide estimates of the PERM standardised skewness.



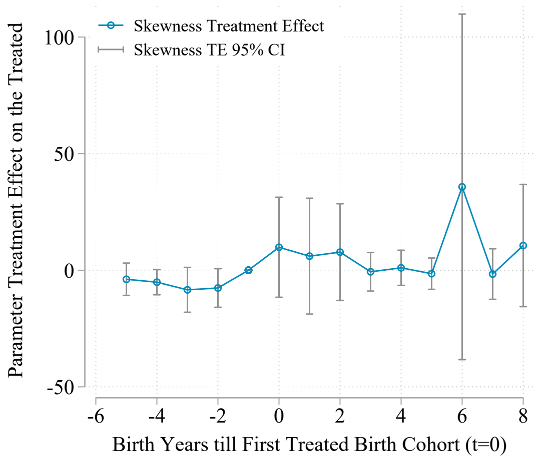
(a) Variance (Education)



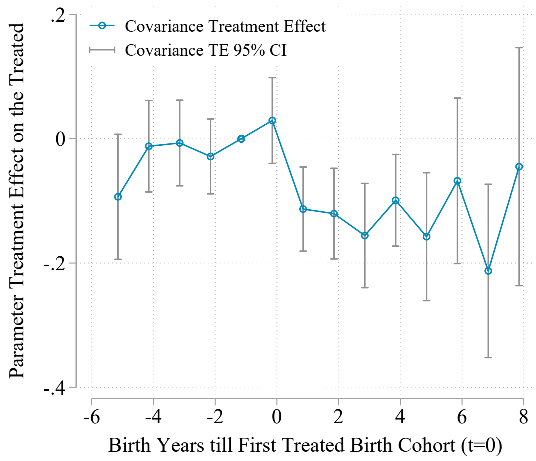
(b) Variance (Earnings)



(c) Skewness (Education)



(d) Skewness (Earnings)



(e) Covariance education and earnings)

Figure 4: PERM Event Study Figures

Notes: These figures plot the DPTT by time since reform was first implemented in their municipality using the methods set out in section 2.2. 95%CI clustered by municipality of residence are shown as capped vertical lines, provided by bootstrapping the whole PERM DiD procedure with 200 replications.

Figure 3 presents the event study figures for the first raw moment of education and earnings. Both education and earnings observe parallel trends in the pre-treatment period. A test of joint significance of the placebo pre-treatment effects fails to reject the null for both outcomes. Focusing first on years of education (panel (a)), the first treated birth cohort ($t = 0$) shows a meaningful impact of the reform, which then increases substantially in the $t = 1$ period that is then largely sustained. The partial jump at $t = 0$ is due to partial implementation also observed in prior studies examining the impact of the comprehensive school reform (Holmlund, Lindahl and Plug 2011, Fischer et al. 2021). The overall impact of the reform indicated by the event study figures is an increase in years of education of over 0.5 years, large relative to other school reforms (Galama, Lleras-Muney and van Kippersluis 2018) and an increase in average annual earnings of around 20,000 SEK (Panel (b)).

Figure 4 presents PERM event study figures for the population variance and population (unstandardised) skewness of education as well as earnings. The event study results for education indicate clear negative impacts of the reform on the variance and skewness of education. For earnings, the event study figures indicate no precise impacts, although there was perhaps a slight increase in variance. We consider the joint significance in the next section.

Event study figures of the population variance and population skewness as presented in Figure 4 also allow assessment of pre-trends in group variance and group skewness respectively, as discussed in section 2.2. The figures suggest no clear trends in the pre-treatment period, supporting the PERM DiD assumptions for these parameters. Finally, panel (e) of Figure 4 presents the event study figure for the covariance of years of education and earnings. Again, like for years of education, the covariance observes clear impact post reform and no clear pre-trends. The impact of the reform appears to have reduced the covariance between earnings and education.

In Appendix N we present some additional material that illustrates how PERM DiD works. As set out in section 1.4, PERM DiD corrects for the bias in standard DiD of higher-order raw moments. This is illustrated in Appendix N, where standard event study estimates are compared to PERM DiD event study estimates.

The PERM distribution parameter event study estimates shown in Figure 4 are the difference between a function of PERM based estimates of the lower order distribution parameters and the highest raw moment relevant for the distribution parameter of interest. For example, an impact on the variance requires the impact on $\mathbb{E}[Y^2]$ to be different compared to the impact on $\mathbb{E}[Y]^2$. PERM DiD is therefore a set of building blocks where distribution parameters utilising higher-order raw moments require estimates of lower order raw moments. Appendix N illustrates this. Figures N.2 panel (b) and N.3 present the raw moment components as event study figures. The difference between the two curves is the DPTT for each birth cohort shown in Figure 4.

5.3 PERM DiD Main Results

Table 3: PERM DiD Regression Results

	Observed Parameter (1)	Treatment Effect (2)
Years of Education		
Mean	11.05	0.47 ***
	(0.04)	(0.03)
Variance[a]	5.68	-0.83 ***
	(0.11)	(0.08)
Skewness[a]	14.36	-2.50 ***
	(0.47)	(0.50)
Standardised Skewness[a,b]	1.06	0.05 ***
	(0.01)	(0.02)
Earnings - 100,000 SEK pa		
Mean	2.37	0.02 ***
	(0.01)	(0.01)
Variance[a]	2.17	0.18 *
	(0.09)	(0.11)
Skewness[a]	20.15	6.78
	(4.46)	(5.41)
Standardised Skewness[a,b]	6.29	1.55
	(1.06)	(1.38)
Covariance and Slope		
$Covariance(WY)^a$	1.26	-0.09 ***
	(0.04)	(0.03)
$Beta(WY)^{a,b}$	0.22	-0.02 ***
	(0.00)	(0.01)
N	304,313	304,313

Notes: This table presents distribution parameter estimates and their standard errors in column (1) and PERM DiD based estimates of the comprehensive school reform's effect on these distribution parameters in column (2). The population is the population exposed to the reform. Cluster along municipality robust standard errors are provided by non-parametric bootstrap with 200 replications and shown in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. [a] standard errors may be downward biased in small samples, especially for very non-normal distributions. [b] standardised skewness can be biased in small samples, especially for very non-normal distributions.

The main PERM DID results are presented in Table 3 and show that the comprehensive school reform increased mean years of schooling by about 0.47 years and reduced the population variance by 0.83, both substantial effects compared to the observed baseline. The school reform also reduced the population skewness by about 2.5. These are also relatively large effects, but when standardising the population skewness by the variance, the relative impact size becomes smaller and positive, suggesting the negative effect was driven by the reduction in variance.

The results for earnings show an economically meaningful positive mean effect of 20,000 SEK per annum due to the reform. A positive variance impact of 0.18 is also observed, a substantial increase compared to baseline, although imprecisely estimated. Imprecise positive effects are found for the skewness (unstandardised and standardised). Whilst raising the floor of education clearly reduces inequalities in education, this reduction in inequality has not translated into a reduction in labour market earnings inequality, rather the results suggest labour market inequalities increased. In Appendix M figure M.1 plots a histogram of earnings of those with less than comprehensive schooling and not treated by the reform. This shows considerable variation in earnings with a large right tail, even for this low education group. If low-educated but high-potential earners benefited from this school reform, then this would be consistent with an increase in earnings variance due to the reform. A school reform that targets the low educated does not necessarily target low earners and thereby is not necessarily an effective inequality policy. Note, we consider earnings and not log earnings. These conclusions are therefore in terms of absolute differences, rather than relative differences.

The impacts on our bivariate distribution measures, the covariance and beta slope coefficient of education and earnings, suggest the relationship between education and earnings has been weakened due to the school reform. The impact of the school reform on the slope suggests that an additional year of schooling is now associated with 20,000 SEK less compared to prior the reform. This suggests a reduction in education related income inequalities.

Overall, the Swedish comprehensive school reform helped reduce education inequalities, increase mean earnings and reduce education related earnings differentials. However, earnings differences, as measured by the variance, appear to have increased as a consequence of the reform. These conclusions are supported by analysis of pre-trends. The PERM event study figures show parallel trends in the parameters prior to introduction of the reform supporting our identification assumptions. To further support our conclusions we provide evidence of covariate balance between treated and counterfactual groups, and also evidence of robustness to outliers by removing the single highest earner in each reform year - birth cohort cell (293 cells in total), details found in Appendix O.

6 Discussion

This paper has developed a new empirical framework for the analysis of treatment effects on the distribution of outcomes, an approach we call Parameter Estimation by Raw Moments. The advantages of the PERM approach include its familiarity, its relative ease of application, and the opportunity to invoke weaker identifying assumptions than most alternative approaches. Ease of empirical application of the PERM approach follows from its familiarity as merely an extension of commonly used causal inference methods that focus on the mean. The ability to make weaker assumptions follows from PERM's focus on estimating counterfactual distribution parameters by first identifying a low dimensional set of counterfactual raw moments, then calculating the parameter of interest. For the variance, this requires one additional raw moment, as generally the mean has already been identified. Other available methods generally require identifying the full counterfactual distribution to then calculate the variance, which is a much more demanding approach. Compared to alternative approaches PERM tells us less, but also requires less demanding assumptions as a consequence.

The results of our Monte Carlo simulation illustrate that under the assumption of selection of observables PERM implemented with regression (PERM regression) produces almost identical results to those produced using Inverse Probability Weighting (IPW), a method that performs well in terms of bias and root mean squared error compared to the leading alternative distribution methods (Firpo and Pinto 2016). This result is confirmed in our first empirical example analysing the impact of union coverage on US log wages, where IPW and PERM regression yield almost identical results. We found that not only are mean wages higher due to unionisation, but also more equal in terms of reduced variance, and (standardised) skewness for those who were covered by union membership.

We have also shown that the PERM framework is not only applicable in a selection on observables setting, but can also be extended to cases where selection may depend on unobservables. We suggest some reasonable identifying assumptions and a PERM DiD estimator that yields unbiased estimators of the distribution parameter treatment effect on the treated under these assumptions. Employing PERM DiD we assessed a Swedish comprehensive school reform that introduced a higher minimum years of schooling, from seven or eight years to nine years, and kept mixed ability groups together for longer. The results show that the reform led not only to an increase in the average years of education but also to a meaningful reduction in the variance of years of education, compressing the education distribution. However, this compression in years of education did not translate to earnings: earnings were found to have increased on average (significant at the 1% level), yet suggestive evidence is found for an increase in earnings variability (significant at the 10% level).

Measures of skewness also suggest increases, although these were not significant. Finally, we find that the relationship between years of education and earnings is weakened by the reform. Both the covariance and the slope were reduced by the reform. This last finding illustrates a further advantage of PERM, it's ability to consider the distribution of multivariate outcomes in a DiD set-up.

PERM DiD is not the only DiD based method available that allows analysis of distribution impacts of social policies. Other methods have been proposed that utilise alternative assumptions to identify distribution impacts by way of DiD (see Roth et al. (2023) for a recent overview). The Changes-in-Changes model of Athey and Imbens (2006), allows one to identify the full counterfactual distribution under the assumption that the mapping between each and every quantile of interest across treated and untreated groups remains stable over time. The distribution DiD model of Bonhomme and Sauder (2011) proposes a method utilising a parallel trends assumption for the characteristic function. Callaway and Li (2019) propose a distribution DiD model based on a copula stability assumption and Fernández-Val et al. (2024) propose a distribution regression DiD approach utilising a trends in the transformed distribution assumption. Roth and Sant'Anna (2023) show that if a form of parallel trends assumption for the cumulative distribution of Y means parallel trends holds for all functional forms of Y then one can also estimate the full counterfactual distribution of the treated group. This implies that for any distributional DiD method to provide a counterfactual distributional parameter, some form of a 'parallel distribution' assumption is required.

PERM DiD offers potential advantages compared to the alternative distribution DiD methods noted above when the aim is to estimate DPTTs. This is due to the general properties of PERM when estimating distribution parameters – the ability to invoke weaker assumptions, relative computational simplicity in estimating distribution parameters, and the ability to easily consider multivariate outcomes – which also apply to PERM DiD. If the end goal is to summarise the impacts of a social policy using distribution parameters, then PERM DiD offers an approach that is computationally less burdensome than the distribution DiD methods outlined above. Distributional DiD methods require identification of the full counterfactual distribution to then calculate the counterfactual distributional parameter. This is computationally burdensome. Furthermore, compared to these alternative methods, the identifying assumption(s) of PERM DiD, parallel trends in group-level distribution parameters of interest, are collectively a weaker assumption than a parallel distribution type assumption required by distributional DiD methods. The larger the set of parameters considered, the less clear this distinction will be. Finally, to date, only PERM DiD and the method of Fernández-Val et al. (2024) has provided a way to estimate how treatment affects not only the distribution of outcomes but also the multivariate distribution of two outcomes.

Potential extensions to the work presented here include consideration of a doubly robust estimator (Robins, Rotnitzky and Zhao 1994), which combines regression and IPW to estimate the relevant raw moments. Alternative assumptions could also be considered. For example, PERM regression can permit one to relax the common support assumption to consider out-of-sample predictions and also flexibility when faced with missing data, cases we have not considered here but are potentially useful in empirical practice. Alternative treatment effects could also be considered. In this paper, we have focused on the distribution parameter treatment effect on the treated. Alternative treatment effects have been suggested by Firpo and Pinto (2016). These treatment effects consider other policy relevant scenarios, such as the comparison of where no-one is treated to where everyone is treated, what Firpo and Pinto (2016) call the *overall inequality treatment effect*. Another policy relevant treatment effect is the impact of an intervention on current observed inequality if only a subset were exposed to treatment, the *current inequality treatment effect* (Firpo and Pinto 2016). PERM can be easily extended to consider these treatment effects.

Whilst the PERM approach has several positive attributes that suggest its empirical feasibility and applicability in the estimation of distribution parameter treatment effects, there are some important limitations worth noting. The first is that PERM is limited to distribution parameters and inequality measures that can be expressed exclusively as functions of raw moments. This disqualifies rank-based measures of inequality such as the Gini Index for example. This is because the rank itself is a function not expressible in terms of raw moments. IPW based approaches, for example, are not limited in this regard and allow analysis of any distribution measure. A second drawback, related to the first, PERM's statistical features depend on the parameter of interest. As our Monte Carlo exercise has shown, bootstrap-based standard errors of higher-order distribution parameters achieve lower rates of coverage in small samples, although this limitation is also common with IPW approaches.

To conclude, this paper has introduced a new framework to examine the impact on distributions, PERM, which is built around the analysis of raw moments. PERM is familiar because it can utilised our established toolkit for means. As a result, PERM allows flexible identifying assumptions to yield unbiased distribution parameter treatment effects compared to several leading alternative methods. This paper has provided two use cases for PERM, PERM regression and PERM DiD, but PERM can also be extended to cases beyond those considered in this paper.

References

Athey, Susan, and Guido W Imbens. 2006. "Identification and inference in nonlinear difference-in-differences models." *Econometrica*, 74(2): 431–497.

Blanden, Jo, Matthias Doepke, and Jan Stuhler. 2023. "Educational inequality." In *Handbook of the Economics of Education*. Vol. 6, 405–497. Elsevier.

Bonhomme, Stéphane, and Ulrich Sauder. 2011. "Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling." *Review of Economics and Statistics*, 93(2): 479–494.

Bossler, Mario, and Thorsten Schank. 2023. "Wage inequality in Germany after the minimum wage introduction." *Journal of Labor Economics*, 41(3): 813–857.

Callaway, Brantly, and Tong Li. 2019. "Quantile treatment effects in difference in differences models with panel data." *Quantitative Economics*, 10(4): 1579–1618.

Card, David, Thomas Lemieux, and W Craig Riddell. 2020. "Unions and wage inequality: The roles of gender, skill and public sector employment." *Canadian Journal of Economics/Revue canadienne d'économique*, 53(1): 140–173.

Card, David, Thomas Lemieux, D Card, and W Craig Riddell. 2004. "Unions and wage inequality." *Journal of Labor Research*, 25(4): 519–62.

Chebyshev, P.L. 1891. "Sur deux theoremes relatifs aux probabilities." *Acta Math*, 14.

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. 2013. "Inference on counterfactual distributions." *Econometrica*, 81(6): 2205–2268.

Clark, Damon, and Heather Royer. 2013. "The effect of education on adult mortality and health: Evidence from Britain." *The American Economic Review*, 103(6): 2087–2120.

De Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2024. "Difference-in-differences estimators of intertemporal treatment effects." *Review of Economics and Statistics*, 1–45.

De Chaisemartin, Clément, and Xavier dHaultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*, 110(9): 2964–2996.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." *Econometrica*, 64(5): 1001–1044.

Essama-Nssah, Boniface, and Peter J Lambert. 2012. "Chapter 6 Influence Functions for Policy Impact Analysis." In *Inequality, mobility and segregation: Essays in honor of Jacques Silber*. 135–159. Emerald Group Publishing Limited.

Fan, Yanqin, and Sang Soo Park. 2010. "Sharp bounds on the distribution of treatment effects and their statistical inference." *Econometric Theory*, 26(3): 931–951.

Fernández-Val, Iván, Jonas Meier, Aico van Vuuren, and Francis Vella. 2024. "Distribution Regression Difference-in-Differences." *arXiv preprint arXiv:2409.02311*.

Firpo, Sergio. 2007. "Efficient semiparametric estimation of quantile treatment effects." *Econometrica*, 75(1): 259–276.

Firpo, Sergio, and Cristine Pinto. 2016. “Identification and estimation of distributional impacts of interventions using changes in inequality measures.” *Journal of Applied Econometrics*, 31(3): 457–486.

Firpo, Sergio, and Geert Ridder. 2019. “Partial identification of the treatment effect distribution and its functionals.” *Journal of Econometrics*, 213(1): 210–234.

Firpo, Sergio, Nicole M Fortin, and Thomas Lemieux. 2009. “Unconditional quantile regressions.” *Econometrica*, 77(3): 953–973.

Firpo, Sergio P, Nicole M Fortin, and Thomas Lemieux. 2018. “Decomposing wage distributions using recentered influence function regressions.” *Econometrics*, 6(2): 28.

Fischer, Martin, Gawain Heckley, Martin Karlsson, and Therese Nilsson. 2022. “Revisiting Sweden’s comprehensive school reform: Effects on education and earnings.” *Journal of Applied Econometrics*, 37(4): 811–819.

Fischer, Martin, Ulf-G Gerdtham, Gawain Heckley, Martin Karlsson, Gustav Kjellsson, and Therese Nilsson. 2021. “Education and health: long-run effects of peers, tracking and years.” *Economic Policy*, 36(105): 3–49.

Galama, Titus J, Adriana Lleras-Muney, and Hans van Kippersluis. 2018. “The Effect of Education on Health and Mortality: A Review of Experimental and Quasi-Experimental Evidence.” National Bureau of Economic Research.

Graubard, Barry I, and Edward L Korn. 1999. “Predictive margins with survey data.” *Biometrics*, 55(2): 652–659.

Heckley, Gawain, Ulf-G Gerdtham, and Gustav Kjellsson. 2016. “A general method for decomposing the causes of socioeconomic inequality in health.” *Journal of Health Economics*, 48: 89–106.

Hjalmarsson, Randi, Helena Holmlund, and Matthew J Lindquist. 2015. “The Effect of Education on Criminal Convictions and Incarceration: Causal Evidence from Micro-data.” *The Economic Journal*.

Holmlund, Helena. 2007. “A Researcher’s Guide to the Swedish Compulsory School Reform.” Working paper 9/2007, Swedish Institute for Social research, Stockholm University.

Holmlund, Helena. 2016. “Education and equality of opportunity: what have we learned from educational reforms?” Working paper.

Holmlund, Helena, Mikael Lindahl, and Erik Plug. 2011. “The causal effect of parents’ schooling on children’s schooling: A comparison of estimation methods.” *Journal of Economic Literature*, 49(3): 615–651.

Imbens, Guido W. 2004. “Nonparametric estimation of average treatment effects under exogeneity: A review.” *Review of Economics and statistics*, 86(1): 4–29.

Joanes, Derrick N, and Christine A Gill. 1998. "Comparing measures of sample skewness and kurtosis." *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1): 183–189.

Juhn, Chinhui, Kevin M Murphy, and Brooks Pierce. 1993. "Wage inequality and the rise in returns to skill." *Journal of political Economy*, 101(3): 410–442.

Lager, Anton Carl Jonas, and Jenny Torssander. 2012. "Causal effect of education on mortality in a quasi-experiment on 1.2 million Swedes." *Proceedings of the National Academy of Sciences*, 109(22): 8461–8466.

Lundborg, Petter, and Kaveh Majlesi. 2018. "Intergenerational transmission of human capital: Is it a one-way street?" *Journal of Health Economics*, 57: 206–220.

Lundborg, Petter, Anton Nilsson, and Dan-Olof Rooth. 2014. "Parental education and offspring outcomes: evidence from the Swedish compulsory school reform." *American Economic Journal: Applied Economics*, 6(1): 253–278.

Meghir, Costas, and Mårten Palme. 2005. "Educational reform, ability, and family background." *American Economic Review*, 414–424.

Meghir, Costas, Mårten Palme, and Emilia Simeonova. 2018. "Education and mortality: Evidence from a social experiment." *American Economic Journal: Applied Economics*, 10(2): 234–56.

Melly, Blaise, and Kaspar Wüthrich. 2017. "Local quantile treatment effects." *Handbook of quantile regression*, 145–164.

Oreopoulos, Philip. 2006. "Estimating average and local average treatment effects of education when compulsory schooling laws really matter." *The American Economic Review*, 96(1): 152–175.

Palme, Mårten, and Emilia Simeonova. 2015. "Does women's education affect breast cancer risk and survival? Evidence from a population based social experiment in education." *Journal of Health Economics*, 42: 115–124.

Pearson, Karl. 1894. "Contributions to the mathematical theory of evolution." *Philosophical Transactions of the Royal Society of London. A*, 185: 71–110.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of regression coefficients when some regressors are not always observed." *Journal of the American statistical Association*, 89(427): 846–866.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70(1): 41–55.

Rothe, Christoph. 2010. "Nonparametric estimation of distributional policy effects." *Journal of Econometrics*, 155(1): 56–70.

Rothe, Christoph. 2015. "Decomposing the composition effect: the role of covariates in determining between-group differences in economic outcomes." *Journal of Business & Economic Statistics*, 33(3): 323–337.

Roth, Jonathan. 2024. "Interpreting event-studies from recent difference-in-differences methods." *arXiv preprint arXiv:2401.12309*.

Roth, Jonathan, and Pedro HC Sant'Anna. 2023. "When is parallel trends sensitive to functional form?" *Econometrica*, 91(2): 737–747.

Roth, Jonathan, Pedro HC SantAnna, Alyssa Bilinski, and John Poe. 2023. "Whats trending in difference-in-differences? A synthesis of the recent econometrics literature." *Journal of Econometrics*.

Słoczyński, Tymon. 2022. "Interpreting OLS estimands when treatment effects are heterogeneous: Smaller groups get larger weights." *Review of Economics and Statistics*, 104(3): 501–509.

Van der Vaart, Aad W. 2000. *Asymptotic statistics.* Vol. 3, Cambridge university press.

Wilcox, Rand R. 2012. *Introduction to robust estimation and hypothesis testing.* Academic press.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data.* MIT press.

Online Appendix

Appendix A Proofs

Proof. (*Proposition 2*)

In the case of no treatment, the PERM DiD estimate of the variance treatment effect on the treated is given by:

$$\begin{aligned}
\Delta_{VarTT}(Y_{i,g,t}) =& 0 \\
=& (v^{Var}(Y_{i,1,1}) - v^{Var}(Y_{i,1,0})) - (v^{Var}(Y_{i,0,1}) - v^{Var}(Y_{i,0,0})) \\
=& (\mathbb{E}[Y_{i,1,1}^2] - \mathbb{E}[Y_{i,1,1}]^2) - (\mathbb{E}[Y_{i,1,0}^2] - \mathbb{E}[Y_{i,1,0}]^2) \\
&- (\mathbb{E}[Y_{i,0,1}^2] - \mathbb{E}[Y_{i,0,1}]^2) - (\mathbb{E}[Y_{i,0,0}^2] - \mathbb{E}[Y_{i,0,0}]^2)
\end{aligned}$$

Rearranging for the counterfactual second raw moment, $\mathbb{E}[Y_{i,1,1}^2]$, and subsequently substituting in the DGP for Y_i we get:

$$\begin{aligned}
\mathbb{E}[Y_{i,1,1}^2] =& \mathbb{E}[Y_{i,1,1}]^2 + (\mathbb{E}[Y_{i,1,0}^2] - \mathbb{E}[Y_{i,1,0}]^2) \\
&+ (\mathbb{E}[Y_{i,0,1}^2] - \mathbb{E}[Y_{i,0,1}]^2) - (\mathbb{E}[Y_{i,0,0}^2] - \mathbb{E}[Y_{i,0,0}]^2) \\
=& \mathbb{E}[\alpha_i + \delta_{i,g} + \rho_{i,t}]^2 + (\mathbb{E}[(\alpha_i + \delta_{i,g})^2] - \mathbb{E}[\alpha_i + \delta_{i,g}]^2) \\
&+ (\mathbb{E}[(\alpha_i + \rho_{i,t})^2] - \mathbb{E}[\alpha_i + \rho_{i,t}]^2) - (\mathbb{E}[(\alpha_i)^2] - \mathbb{E}[\alpha_i]^2) \\
=& \mathbb{E}[\alpha_i^2] + 2\,\mathbb{E}[\delta_{i,g}]\,\mathbb{E}[\rho_{i,t}] \\
&+ \mathbb{E}[\delta_{i,g}^2] + 2\,\mathbb{E}[\delta_{i,g}\alpha_i] + \mathbb{E}[\rho_{i,t}^2] + 2\,\mathbb{E}[\rho_{i,t}\alpha_i]
\end{aligned}$$

Let the intercept term be given by $\alpha_2 = \mathbb{E}[\alpha_i^2]$, the mean group difference that is constant over time be given by $\omega = \mathbb{E}[\delta_{i,g}^2] + 2\,\mathbb{E}[\delta_{i,g}\alpha_i]$, the mean time effect common to both groups be given by $\gamma = \mathbb{E}[\rho_{i,t}^2] + 2\,\mathbb{E}[\rho_{i,t}\alpha_i]$, and how $\mathbb{E}[Y^2]$ changes if both the mean and variance are parallel across treated and control groups be given by $2\,\mathbb{E}[\delta_{i,g}]\,\mathbb{E}[\rho_{i,t}]$, then the second raw moment for the treated group in the treated period in the absence of treatment, is given by:

$$\mathbb{E}[Y_{i,1,1}^2] = \alpha_2 + \omega + \gamma + 2\,\mathbb{E}[\delta_{i,g}]\,\mathbb{E}[\rho_{i,t}]$$

where $\mathbb{E}[Y_{01i,g,t}^2|g,t]$ is the more general case of $\mathbb{E}[Y_{i,1,1}^2]$. This completes the proof. □

Proof. (*Proposition 3*) In the case of no treatment, the PERM DiD estimate of the skewness treatment effect on the treated is given by:

$$
\begin{aligned}
0 =& (v^{skew}(Y_{i,1,1}) - v^{skew}(Y_{i,1,0}) - (v^{skew}(Y_{i,0,1}) - v^{skew}(Y_{i,0,0})) \\
=& \Big((\mathbb{E}[Y_{i,1,1}^3] - 3\,\mathbb{E}[Y_{i,1,1}^2]\,\mathbb{E}[Y_{i,1,1}] + 2\,\mathbb{E}[Y_{i,1,1}]^2) \\
& - (\mathbb{E}[Y_{i,1,0}^3] - 3\,\mathbb{E}[Y_{i,1,0}^2]\,\mathbb{E}[Y_{i,1,0}] + 2\,\mathbb{E}[Y_{i,1,0}]^2) \Big) \\
& - \Big((\mathbb{E}[Y_{i,0,1}^3] - 3\,\mathbb{E}[Y_{i,0,1}^2]\,\mathbb{E}[Y_{i,0,1}] + 2\,\mathbb{E}[Y_{i,0,1}]^2) \\
& - (\mathbb{E}[Y_{i,0,0}^3] - 3\,\mathbb{E}[Y_{i,0,0}^2]\,\mathbb{E}[Y_{i,0,0}] + 2\,\mathbb{E}[Y_{i,0,0}]^2) \Big)
\end{aligned}
$$

Rearranging for the counterfactual third raw moment, $\mathbb{E}[Y_{i,1,1}^3]$, and subsequently substituting in the DGP for Y_i we get:

$$
\begin{aligned}
\mathbb{E}[Y_{i,1,1}^3] =& \Big((3\,\mathbb{E}[Y_{i,1,1}^2]\,\mathbb{E}[Y_{i,1,1}] - 2\,\mathbb{E}[Y_{i,1,1}]^2) \\
& + (\mathbb{E}[Y_{i,1,0}^3] - 3\,\mathbb{E}[Y_{i,1,0}^2]\,\mathbb{E}[Y_{i,1,0}] + 2\,\mathbb{E}[Y_{i,1,0}]^2) \Big) \\
& + \Big((\mathbb{E}[Y_{i,0,1}^3] - 3\,\mathbb{E}[Y_{i,0,1}^2]\,\mathbb{E}[Y_{i,0,1}] + 2\,\mathbb{E}[Y_{i,0,1}]^2) \\
& - (\mathbb{E}[Y_{i,0,0}^3] - 3\,\mathbb{E}[Y_{i,0,0}^2]\,\mathbb{E}[Y_{i,0,0}] + 2\,\mathbb{E}[Y_{i,0,0}]^2) \Big) \\
=& \Big((3(\mathbb{E}[(\alpha_i + \delta_{i,g} + \rho_{i,t})^2])\,\mathbb{E}[\alpha_i + \delta_{i,g} + \rho_{i,t}] - 2\,\mathbb{E}[\alpha_i + \delta_{i,g} + \rho_{i,t}]^2) \\
& + (\mathbb{E}[(\alpha_i + \delta_{i,g})^3] - 3\,\mathbb{E}[(\alpha_i + \delta_{i,g})^2]\,\mathbb{E}[\alpha_i + \delta_{i,g}] + 2\,\mathbb{E}[\alpha_i + \delta_{i,g}]^2) \Big) \\
& + \Big((\mathbb{E}[(\alpha_i + \rho_{i,t})^3] - 3\,\mathbb{E}[(\alpha_i + \rho_{i,t})^2]\,\mathbb{E}[\alpha_i + \rho_{i,t}] + 2\,\mathbb{E}[\alpha_i + \rho_{i,t}]^2) \\
& - (\mathbb{E}[\alpha_i^3] - 3\,\mathbb{E}[\alpha_i^2]\,\mathbb{E}[\alpha_i] + 2\,\mathbb{E}[\alpha_i]^2) \Big) \\
=& \Big((3(\alpha_2 + \gamma + \omega + 2\,\mathbb{E}[\delta_{i,g}]\,\mathbb{E}[\rho_{i,t}])(\alpha_1 + \mathbb{E}[\delta_{i,g}] + \mathbb{E}[\rho_{i,t}]) \\
& - 2(\alpha_1 + \mathbb{E}[\delta_{i,g}] + \mathbb{E}[\rho_{i,t}])^2) \\
& + (\mathbb{E}[(\alpha_i + \delta_{i,g})^3] - 3(\alpha_2 + \omega)(\alpha_1 + \mathbb{E}[\delta_{i,g}]) + 2(\alpha_1 + \mathbb{E}[\delta_{i,g}])^2) \Big) \\
& + \Big((\mathbb{E}[(\alpha_i + \rho_{i,t})^3] - 3(\alpha_2 + \gamma)(\alpha_1 + \mathbb{E}[\rho_{i,t}]) + 2(\alpha_1 + \mathbb{E}[\rho_{i,t}])^2) \\
& - (\mathbb{E}[\alpha_i^3] - 3\alpha_2\alpha_1 + 2(\alpha_1)^2) \Big) \\
=& \alpha_3 + \theta + \phi + 3\gamma\mathbb{E}[\delta_{i,g}] + 3\omega\,\mathbb{E}[\rho_{i,t}] - 6\alpha_1\,\mathbb{E}[\delta_{i,g}]\,\mathbb{E}[\rho_{i,t}]
\end{aligned}
$$

where $\mathbb{E}[Y^3_{01i,g,t}|g,t]$ is the more general case of $\mathbb{E}[Y^3_{i,1,1}]$.This completes the proof. $\square$

Proof. (*Proposition 4*) In the case of no treatment, the PERM DiD estimate of the covariance treatment effect on the treated is given by:

$$\begin{aligned}
\Delta_{CovTT}((W_{i,g,t},Y_{i,g,t}) =& 0\\
=&(\nu^{Cov}(W_{i,1,1}Y_{i,1,1}) - \nu^{Cov}(W_{i,1,0},Y_{i,1,0}))\\
&-(\nu^{Cov}(W_{i,0,1},Y_{i,0,1}) - \nu^{Cov}(W_{i,0,0},Y_{i,0,0}))\\
=&(\mathbb{E}[W_{i,1,1}Y_{i,1,1}] - \mathbb{E}[W_{i,1,1}]\,\mathbb{E}[Y_{i,1,1}])\\
&-(\mathbb{E}[W_{i,1,0}Y_{i,1,0}] - \mathbb{E}[W_{i,1,0}]\,\mathbb{E}[Y_{i,1,0}])\\
&-(\mathbb{E}[W_{i,0,1}Y_{i,0,1}] - \mathbb{E}[W_{i,0,1}]\,\mathbb{E}[Y_{i,0,1}])\\
&-(\mathbb{E}[W_{i,0,0}Y_{i,0,0}] - \mathbb{E}[W_{i,0,0}]\,\mathbb{E}[Y_{i,0,0}])
\end{aligned}$$

Rearranging for the counterfactual joint raw moment, $\mathbb{E}[W_{i,1,1}Y_{i,1,1}]$, and subsequently substituting in the DGP for W_i and Y_i we get:

$$\begin{aligned}
\mathbb{E}[W_{i,1,1}Y_{i,1,1}] =& \mathbb{E}[W_{i,1,1}]\,\mathbb{E}[Y_{i,1,1}] + (\mathbb{E}[W_{i,1,0}Y_{i,1,0}] - \mathbb{E}[W_{i,1,0}]\,\mathbb{E}[Y_{i,1,0}])\\
&+(\mathbb{E}[W_{i,0,1}Y_{i,0,1}] - \mathbb{E}[W_{i,0,1}]\,\mathbb{E}[Y_{i,0,1}]) - (\mathbb{E}[W_{i,0,0}Y_{i,0,0}] - \mathbb{E}[W_{i,0,0}]\,\mathbb{E}[Y_{i,0,0}])\\
=&\mathbb{E}[\alpha_{i,W} + \delta_{i,g,W} + \rho_{i,t,W}]\,\mathbb{E}[\alpha_{i,Y} + \delta_{i,g,Y} + \rho_{i,t,Y}]\\
&+(\mathbb{E}[(\alpha_{i,W} + \delta_{i,g,W})(\alpha_{i,Y} + \delta_{i,g,Y})] - \mathbb{E}[\alpha_{i,W} + \delta_{i,g,W}]\,\mathbb{E}[\alpha_{i,Y} + \delta_{i,g,Y}])\\
&+(\mathbb{E}[(\alpha_{i,W} + \rho_{i,t,W})(\alpha_{i,Y} + \rho_{i,t,Y})] - \mathbb{E}[\alpha_{i,W} + \rho_{i,t,W}]\,\mathbb{E}[\alpha_{i,Y} + \rho_{i,t,Y}])\\
&-(\mathbb{E}[(\alpha_{i,W})(\alpha_{i,Y})] - \mathbb{E}[\alpha_{i,W}]\,\mathbb{E}[\alpha_{i,Y}])\\
=&\mathbb{E}[(\alpha_{i,W})(\alpha_{i,Y})]\\
&+\mathbb{E}[\delta_{i,g,W}]\,\mathbb{E}[\rho_{i,t,Y}] + \mathbb{E}[\delta_{i,g,Y}]\,\mathbb{E}[\rho_{i,t,W}]\\
&+\mathbb{E}[\delta_{i,g,W}\delta_{i,g,Y}] + \mathbb{E}[\delta_{i,g,W}\alpha_{i,Y}] + \mathbb{E}[\delta_{i,g,Y}\alpha_{i,W}]\\
&+\mathbb{E}[\rho_{i,g,W}\rho_{i,g,Y}] + \mathbb{E}[\rho_{i,g,W}\alpha_{i,Y}] + \mathbb{E}[\rho_{i,g,Y}\alpha_{i,W}]
\end{aligned}$$

Let the intercept term be given by $\alpha_{WY} = \mathbb{E}[(\alpha_{i,W})(\alpha_{i,Y})]$, the joint raw moment group difference that is constant over time be given by $\omega_{WY} = \mathbb{E}[\delta_{i,g,W}\delta_{i,g,Y}] + \mathbb{E}[\delta_{i,g,W}\alpha_{i,Y}] + \mathbb{E}[\delta_{i,g,Y}\alpha_{i,W}]$, the joint raw moment time effect common to both groups be given by $\gamma_{WY} = \mathbb{E}[\rho_{i,g,W}\rho_{i,g,Y}] + \mathbb{E}[\rho_{i,g,W}\alpha_{i,Y}] + \mathbb{E}[\rho_{i,g,Y}\alpha_{i,W}]$, and how $\mathbb{E}[WY]$ changes if both the mean and variance are parallel across treated and control groups be given by $\mathbb{E}[\delta_{i,g,W}]\,\mathbb{E}[\rho_{i,t,Y}] + \mathbb{E}[\delta_{i,g,Y}]\,\mathbb{E}[\rho_{i,t,W}]$, then the joint raw moment for the treated group in the treated period in the absence of treatment, is given by:

$$\mathbb{E}[W_{i,1,1}Y_{i,1,1}] = \alpha_{WY} + \omega_{WY} + \gamma_{WY} + \mathbb{E}[\delta_{i,g,W}]\,\mathbb{E}[\rho_{i,t,Y}] + \mathbb{E}[\delta_{i,g,Y}]\,\mathbb{E}[\rho_{i,t,W}]$$

where $\mathbb{E}[W_{01i,g,t}Y_{01i,g,t}|g,t]$ is the more general case of $\mathbb{E}[W_{i,1,1}Y_{i,1,1}]$. This completes the proof. □

Appendix B Properties of the sample PERM estimator

We first consider the case of the covariance and extend the argument to the (unstandardised) skewness.

Proposition B.1. ***Bias of the PERM sample covariance estimator:*** *Let $W_1, \ldots, W_n$ be i.i.d random variables with first raw moment μ_W and variance $\mu_{2,W}$, $Y_1, \ldots, Y_n$ be .i.d random variables with first raw moment μ'_Y and variance $\mu_{2,Y}$, that together have a joint first raw moment μ'_{WY} and covariance $\mu_{WY} = \mu'_{WY} - \mu_W \mu_Y$. Let the corresponding sample raw moments be defined as:*

$$\bar{W} = \frac{1}{n}\sum_{i=1}^{n} W_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \overline{WY} = \frac{1}{n}\sum_{i=1}^{n} W_i Y_i$$

The sample raw moments are measured with error, but are unbiased:

$$E[u] = E[\bar{W} - \mu_W] = 0, \quad E[e] = E[\bar{Y} - \mu_Y] = 0, \quad E[m] = E[\overline{WY} - \mu'_{WY}] = 0$$

The PERM sample covariance is biased due to measurement error in the sample raw moments:

$$\mathbb{E}[S_{WY}] = \mu_{WY} - \mathbb{E}[ue]$$

Proof. (Proposition B.1) The PERM sample covariance is:

$$S_{WY} = \frac{1}{n}\sum_{i=1}^{n} W_i Y_i - \frac{1}{n}\sum_{i=1}^{n} W_i \frac{1}{n}\sum_{i=1}^{n} Y_i$$

The PERM sample covariance is biased due to measurement error in the sample raw moments:

$$
\begin{aligned}
\mathbb{E}[S_{WY}] =& \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} W_i Y_i - \frac{1}{n}\sum_{i=1}^{n} W_i \frac{1}{n}\sum_{i=1}^{n} Y_i] \\
=& \mathbb{E}[\overline{WY} - \bar{W}\bar{Y}] \\
=& \mathbb{E}[\overline{WY}] - \mathbb{E}[\bar{W}\bar{Y}] \\
=& \mathbb{E}[\mu'_{WY} + m] - \mathbb{E}[(\mu_W + u)(\mu_Y + e)] \\
=& \mathbb{E}[\mu'_{WY}] + \mathbb{E}[m] - \mathbb{E}[\mu_W \mu_Y] - \mathbb{E}[\mu_W e] - \mathbb{E}[\mu_Y u] - \mathbb{E}[ue] \\
=& \mu'_{WY} - \mu_W \mu_Y - \mu_W \mathbb{E}[e] - \mu_Y \mathbb{E}[u] - \mathbb{E}[ue] \\
=& \mu_{WY} - \mathbb{E}[ue]
\end{aligned}
$$

□

The direction of the bias for the covariance is negative. The variance is a special case of the covariance, where $W = Y$. The PERM sample variance for Y is:

$$
S_{2,Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - (\frac{1}{n}\sum_{i=1}^{n} Y_i)^2
$$

By similar argument for the PERM sample covariance, the PERM sample variance is biased:

$$
\mathbb{E}[S_{2,Y}] = \mu_{2,Y} - \mathbb{E}[e^2]
$$

The bias of the PERM sample variance, $-\mathbb{E}[e^2]$, can also be expressed as $-\mathbb{E}[(\bar{Y} - \mu_Y)^2] = -\frac{1}{n}\mu_{2,Y}$, which follows from the Bienaymé formula. The standard sample variance bias correction formula therefore follows, $\frac{n}{n-1}\mathbb{E}[S_{2,Y}] = \mu_{2,Y}$. A similar line of argument tells us that the bias of the PERM sample covariance $-\mathbb{E}[eu]$, can also be expressed as $-\mathbb{E}[(\bar{W} - \mu_W)(\bar{Y} - \mu_Y))] = -\frac{1}{n}\mu_{WY}$,

An alternative to the standard $\frac{n}{n-1}$ sample covariance correction is to employ an empirical estimator of $\mathbb{E}[eu]$. The i.i.d bootstrap provides consistent estimates of the standard error of the mean (see chapter 23 of Van der Vaart 2000, for a relevant proof). Given consistent estimates of the covariance of the sample mean measurement errors for the observed outcomes, $\mathbb{E}[\widehat{u_1 e_1}|D = 1]$ and counterfactual outcomes $\mathbb{E}[\widehat{u_0 e_0}|D = 1]$ for the treated group, then the bias corrected PERM sample covariance treatment effect on the treated is given by:

$$
\begin{aligned}
\delta_{PTT}^{\mu_{WY}} &= (S_{W_1 Y_1}|D = 1] + \mathbb{E}[\widehat{u_1 \varepsilon_1}|D = 1]) \\
&- (S_{W_0 Y_0}|D = 1] + \mathbb{E}[\widehat{u_0 \varepsilon_0}|D = 1])
\end{aligned}
$$

The PERM sample skewness is also biased and for the same reason as the (co)variance:

the measurement errors of the raw moments bias the sample distribution parameter estimates.

Proposition B.2. ***Bias of the PERM sample skewness estimator:*** *Let Y have first raw moment μ, second raw moment μ_2', third raw moment μ_3', variance $\mu_2 = \mu_2' - (\mu)^2$, and skewness $\mu_3 = \mu_3' - 3\mu_2'\mu + 2(\mu)^3$. Let the sample raw moments, and sample skewness, S_3 be defined as:*

$$\bar{Y^2} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2, \quad \bar{Y^3} = \frac{1}{n}\sum_{i=1}^{n} Y_i^3, \quad \bar{Y^4} = \frac{1}{n}\sum_{i=1}^{n} Y_i^4$$

$$S_3 = \frac{1}{n}\sum_{i=1}^{n} Y_i^3 - 3(\frac{1}{n}\sum_{i=1}^{n} Y_i^2)(\frac{1}{n}\sum_{i=1}^{n} Y_i) + 2(\frac{1}{n}\sum_{i=1}^{n} Y_i)^3$$

The sample raw moments are measured with error, but are unbiased:

$$E[e_1] = E[\bar{Y} - \mu] = 0, \quad [E[e_2] = E[\bar{Y^2} - \mu_2'] = 0$$

$$E[e_3] = E[\bar{Y^3} - \mu_3'] = 0$$

The PERM sample skewness is biased due to measurement error in the sample raw moments:

$$\mathbb{E}[S_3] = \mu_3 - 3\,\mathbb{E}[e_1 e_2] + 6\mu\,\mathbb{E}[e_1^2] + 2\,\mathbb{E}[e_1^3]$$

Proof. (Proposition B.2) The PERM sample skewness is biased due to measurement error in the sample raw moments:

$$\begin{aligned}
\mathbb{E}[S_3] =& \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} Y_i^3 - 3(\frac{1}{n}\sum_{i=1}^{n} Y_i^2)(\frac{1}{n}\sum_{i=1}^{n} Y_i) + 2(\frac{1}{n}\sum_{i=1}^{n} Y_i)^3] \\
=& \mathbb{E}[\bar{Y^3} - 3\bar{Y^2}\bar{Y} + 2(\bar{Y})^3] \\
=& \mathbb{E}[\bar{Y^3}] - 3\,\mathbb{E}[\bar{Y^2}\bar{Y}] + 2\,\mathbb{E}[(\bar{Y})^3] \\
=& \mathbb{E}[\mu_3' + e_3] - 3\,\mathbb{E}[(\mu_2' + e_2)(\mu + e_1)] + 2\,\mathbb{E}[(\mu + e_1)^3] \\
=& \mathbb{E}[\mu_3'] - 3\,\mathbb{E}[(\mu_2'\mu] + 2\,\mathbb{E}[(\mu)^3 - 3\,\mathbb{E}[e_1 e_2] + 6\,\mathbb{E}[\mu e_1^2] + 2\,\mathbb{E}[e_1^3] \\
=& \mu_3 - 3\,\mathbb{E}[e_1 e_2] + 6\mu\,\mathbb{E}[e_1^2] + 2\,\mathbb{E}[e_1^3]
\end{aligned}$$

□

Given consistent estimates of the sample mean measurement errors for the observed outcomes and counterfactual outcomes for the treated group, then the bias corrected PERM sample skewness treatment effect on the treated is given by:

$$\delta_{PTT}^{\mu_3} = ((S_{3,Y_{11}}) - (-3\,\mathbb{E}[\widehat{e_{11,1}e_{11,2}}] + 6\widehat{\mu_{Y_{11,1}}}\,\mathbb{E}[\widehat{e_{11,1}^2}] + 2\,\mathbb{E}[\widehat{e_{11,1}^3}]))$$
$$- ((S_{3,Y_{01}}) - (-3\,\mathbb{E}[\widehat{e_{01,1}e_{01,2}}] + 6\widehat{\mu_{Y_{01,1}}}\,\mathbb{E}[\widehat{e_{01,1}^2}] + 2\,\mathbb{E}[\widehat{e_{01,1}^3}]))$$

Appendix C Law of total third central moment (skewness)

In this section we derive the law of total third central moment. This expressions allow the third central moment to be decomposed into contributions from sub-groups, akin to the law of total variance.

Proposition C.1. ***Law of total third central moment***: *Let the third central moment of Y is defined as:*

$$\mu_3(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^3]$$

and the conditional expectation of Y given X be $\mathbb{E}[Y \mid X]$, then the law of total total third central moment is given by:

$$\mu_3(Y) = \mathbb{E}[\mu_3(Y \mid X)] + 3\mathbb{E}[\mu_2(Y \mid X)(\mathbb{E}[Y \mid X] - \mathbb{E}[Y])] + \mu_3(\mathbb{E}[Y \mid X])$$

Proof. of proposition C.1:

We can introduce the conditional expectation $\mathbb{E}[Y \mid X]$ into the third central moment. By adding and subtracting $\mathbb{E}[Y \mid X]$, we have:

$$(Y - \mathbb{E}[Y]) = (Y - \mathbb{E}[Y \mid X]) + (\mathbb{E}[Y \mid X] - \mathbb{E}[Y])$$

Thus, we can write:

$$(Y - \mathbb{E}[Y])^3 = ((Y - \mathbb{E}[Y \mid X]) + (\mathbb{E}[Y \mid X] - \mathbb{E}[Y]))^3$$

Now, expand the cubic term:

$$\begin{aligned}(Y - \mathbb{E}[Y])^3 =& (Y - \mathbb{E}[Y \mid X])^3 + 3(Y - \mathbb{E}[Y \mid X])^2(\mathbb{E}[Y \mid X] - \mathbb{E}[Y]) \\ &+ 3(Y - \mathbb{E}[Y \mid X])(\mathbb{E}[Y \mid X] - \mathbb{E}[Y])^2 + (\mathbb{E}[Y \mid X] - \mathbb{E}[Y])^3\end{aligned}$$

Next, we take the expectation of both sides to compute the overall third central moment $\mu_3(Y)$:

$$\mathbb{E}[(Y-\mathbb{E}[Y])^3] = \mathbb{E}[(Y-\mathbb{E}[Y \mid X])^3] + 3\mathbb{E}[(Y-\mathbb{E}[Y \mid X])^2(\mathbb{E}[Y \mid X]-\mathbb{E}[Y])]$$

$$+3\mathbb{E}[(Y-\mathbb{E}[Y \mid X])(\mathbb{E}[Y \mid X]-\mathbb{E}[Y])^2] + \mathbb{E}[(\mathbb{E}[Y \mid X]-\mathbb{E}[Y])^3]$$

Each of these terms corresponds to different moments or interactions.

The first term, $\mathbb{E}[(Y-\mathbb{E}[Y \mid X])^3]$, is the expected third central moment of Y within each group defined by X. It is the third central moment of Y conditional on X, denoted as $\mu_3(Y \mid X)$. So the first term becomes:

$$\mathbb{E}[\mu_3(Y \mid X)]$$

The second term, $3\mathbb{E}[(Y-\mathbb{E}[Y \mid X])^2(\mathbb{E}[Y \mid X]-\mathbb{E}[Y])]$ is the conditional variance of Y given X, $(Y-\mathbb{E}[Y \mid X])^2$, denoted as $_2(Y \mid X)$ multiplied by the difference between the group mean and population mean. So this term becomes:

$$3\mathbb{E}[\mu_2(Y \mid X)(\mathbb{E}[Y \mid X]-\mathbb{E}[Y])]$$

The third term, $3\mathbb{E}[(Y-\mathbb{E}[Y \mid X])(\mathbb{E}[Y \mid X]-\mathbb{E}[Y])^2]$, cancels out to zero. Here, $Y-\mathbb{E}[Y \mid X]$ has zero expectation given X, because $\mathbb{E}[Y-\mathbb{E}[Y \mid X] \mid X]=0$. Therefore, the expectation of this term is zero:

$$3\mathbb{E}[(Y-\mathbb{E}[Y \mid X])(\mathbb{E}[Y \mid X]-\mathbb{E}[Y])^2] = 0$$

The fourth term, $\mathbb{E}[(\mathbb{E}[Y \mid X]-\mathbb{E}[Y])^3]$ is the third central moment of the conditional expectation $\mathbb{E}[Y \mid X]$, measuring the skewness of the conditional means. We can denote this as $\mu_3(\mathbb{E}[Y \mid X])$.

Combining all the terms, the **law of total third central moment** for Y becomes:

$$\mu_3(Y) = \mathbb{E}[\mu_3(Y \mid X)] + 3\mathbb{E}[\mu_2(Y \mid X)(\mathbb{E}[Y \mid X]-\mathbb{E}[Y])] + \mu_3(\mathbb{E}[Y \mid X])$$

This provides a full decomposition of the third central moment of Y in terms of conditional expectations and variances with respect to X. □

Appendix D Parallel trends in centered moments

Proposition D.1. ***Parallel group variance is parallel population variance****: Consider two time periods, t-1 and t. Between t-1 and t let the trend in group variance be Δ_B and the* *trend in group first raw moment be Δ_A, then by the law of the total variance, the trend in the population variance will be:*

$$\Delta\mu_2(Y) = \Delta_B$$

Proof. of proposition D.1

The law of total variance states:

$$\mu_2(Y) = \mathbb{E}[\mu_2(Y|X)] + \mu_2(\mathbb{E}[Y|X])$$

The change in total variance is therefore:

$$\mu_2(Y)_t - \mu_2(Y)_{t-1} = \mathbb{E}[\mu_2(Y|X)]_t - \mathbb{E}[\mu_2(Y|X)]_{t-1} + \mu_2(\mathbb{E}[Y|X])_t - \mu_2(\mathbb{E}[Y|X])_{t-1}$$

Let $\Delta\mu_2(Y)$ be the change between t-1 and t, then:

$$\Delta\mu_2(Y) = \Delta\mathbb{E}[\mu_2(Y|X)] + \Delta\mu_2(\mathbb{E}[Y|X])$$

If the trend in group variance is Δ_B and the trend in group first raw moment is Δ_A, then by the law of the total variance, the trend in the population variance will be:

$$\begin{aligned}\Delta\mu_2(Y) =& \Delta_B + \mu_2(\mathbb{E}[Y + \Delta_A|X]) - \mu_2(\mathbb{E}[Y|X])\\ =& \Delta_B + \mu_2(\mathbb{E}[Y|X]) + \mu_2(\Delta_A) - \mu_2(\mathbb{E}[Y|X])\\ =& \Delta_B\end{aligned}$$

Noting that because the variance of a constant is zero, $\mu_2(\Delta_A)$ is equal to zero.

□

Proposition D.2. ***Parallel group skewness is parallel population skewness****: Consider two time periods, t-1 and t. Between t-1 and t let the trend in group skewness be Δ_C, the trend in group variance be Δ_B and the trend in group first raw moment be Δ_A, then by the law of the third central moment, the trend in the population skewness will be:*

$$\Delta\mu_3(Y) = \Delta_C$$

Proof. of proposition D.2

The law of total third central moment provided by proposition C.1 states:

$$\mu_3(Y) = \mathbb{E}[\mu_3(Y \mid X)] + \mu_3(\mathbb{E}[Y \mid X]) + 3\mathbb{E}[\mu_2(Y \mid X)(\mathbb{E}[Y \mid X] - \mathbb{E}[Y])]$$

A change in total third central moment is therefore composed of :

$$\Delta\mu_3(Y) = \Delta\mathbb{E}[\mu_3(Y \mid X)] + \Delta\mu_3(\mathbb{E}[Y \mid X]) + 3\Delta\mathbb{E}[\mu_2(Y \mid X)(\mathbb{E}[Y \mid X] - \mathbb{E}[Y])]$$

If the trend in group skewness is Δ_C, the trend in group variance is Δ_B and the trend in group first raw moment is Δ_A, then by the law of the third central moment, the trend in the population skewness will be:

$$\begin{aligned}\Delta\mu_3(Y) = {} & \Delta_C + \mu_3(\mathbb{E}[Y + \Delta_A|X]) - \mu_3(\mathbb{E}[Y|X]) \\ & + 3\mathbb{E}[(\mu_2(Y \mid X) + \Delta_B)(\mathbb{E}[Y + \Delta_A \mid X] - \mathbb{E}[Y + \Delta A]) - (\mu_2(Y \mid X))(\mathbb{E}[Y \mid X] - \mathbb{E}[Y])]\end{aligned}$$

The skewness of the conditional means simplify, and the changes in the first raw moment inside the last term cancel out:

$$\begin{aligned}\Delta\mu_3(Y) = {} & \Delta_C + \mu_3(\Delta_A) \\ & + 3\mathbb{E}[(\mu_2(Y \mid X) + \Delta_B)(\mathbb{E}[Y \mid X] - \mathbb{E}[Y]) - (\mu_2(Y \mid X))(\mathbb{E}[Y \mid X] - \mathbb{E}[Y])]\end{aligned}$$

The skewness of Δ_A is the skewness of a constant, which equals zero, and the conditional variances multiplied by the difference in the conditional mean and population mean cancel out:

$$\Delta\mu_3(Y) = \Delta_C + 3\mathbb{E}[\Delta_B(\mathbb{E}[Y \mid X] - \mathbb{E}[Y])]$$

By law of iterated expectations, $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$, $\Delta\mu_3(Y)$ is therefore:

$$\Delta\mu_3(Y) = \Delta_C$$

□

Appendix E Monte Carlo Exercise - The Setup

Two explanatory variables, X_1 and X_2, predict both the outcome variable Y and the treatment exposure variable D and are given by:

$$X_1 \sim Uniform\left[1 - \frac{(12)^{1/2}}{2}, 1 + \frac{(12)^{1/2}}{2}\right] \tag{E.1}$$

$$X_2 \sim Uniform\left[5-\frac{(12)^{1/2}}{2}, 5+\frac{(12)^{1/2}}{2}\right] \tag{E.2}$$

Treatment is set as:

$$D = \mathbb{1}\{-0.5+1.35X_1-0.2X_2+0.15X1^2-0.1X2^2+0.5X_1\cdot X_2+\eta>0\} \tag{E.3}$$

where $\eta \sim Normal(0,100)$. The potential outcomes are:

$$Y_0 = exp(0.01-0.01X_1+0.01X_2+0.01X_1^2-0.01X_2^2-0.02X_1\cdot X_2+e_0) \tag{E.4}$$

$$Y_1 = exp(0.1+0.01X_1+0.01X_2+0.01X_1^2+0.01X_2^2+0.01X_1\cdot X_2+e_1) \tag{E.5}$$

where

$$e_0 = (0.01-0.01X_1+0.01X_2+0.01X_1^2-0.01X_2^2-0.02X_1\cdot X_2)\cdot k_0 \tag{E.6}$$

$$e_1 = (0.01+0.01X_1+0.01X_2+0.01X_1^2+0.01X_2^2+0.01X_1\cdot X_2)\cdot k_1 \tag{E.7}$$

where $k_0 \sim normal(0,1)$ and $k_1 \sim normal(0,1)$.

The parameters of interest are the mean (MTT), variance (VTT), skewness (STT) and standardised skewness (SSTT) treatment effect of the treated. An unfeasible estimator is calculated using the potential outcomes from the DGP for the given sample size. A naive estimator is calculated using the observed values of Y and treatment status assuming no selection into treatment.

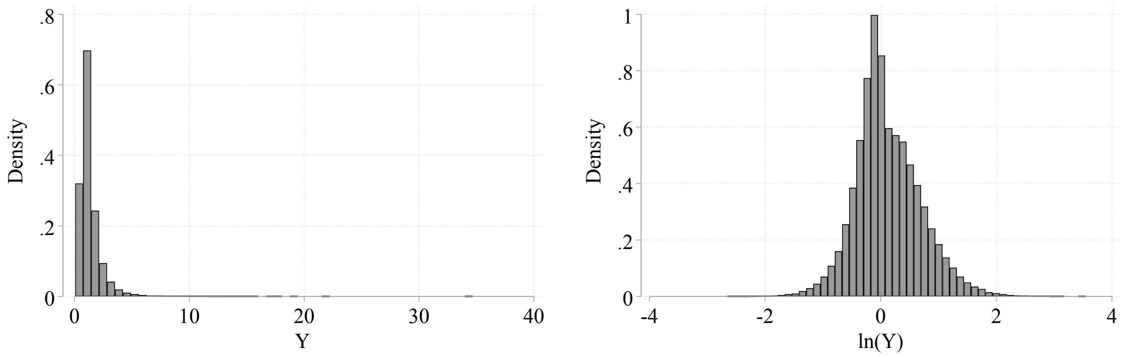
To estimate the PTT using PERM regression the correct specification of X_1 and X_2 is interacted with treatment status in regressions of the first four raw moments, providing observed and counterfactual raw moments following the method set out in section 2.1. The PTTs are then calculated and both standard errors following linearisation technique and standard errors utilising a bootstrap procedure with replacement for the whole procedure are provided.

Estimation of the PTTs using IPW requires two weighting functions that convert the observed distribution function into weighted distribution functions of the observed and counterfactual distribution functions, conditional on being treated. The propensity score, the conditional probability of being treated, is defined as $p(x) \equiv Pr[D=1|X=x]$, where $X \in \chi$. The unconditional probability of being treated is defined as $p \equiv Pr[D=1]$, which we assume exists and is positive. The two weighting functions of interest are:

$$\omega_{11}(d,p(x)) = d/p$$

$$\omega_{01}(d,p(x)) = ((1-d)/(1-p(x)))\times(p(x)/p)$$

p is observed. $p(x)$ is estimated using logit regression, with treatment status as the dependent variable and X_1 and X_2 correctly specified as explanatory variables. PTTs are then estimated using IPW re-weighting to provide the treated and counterfactual estimates. This whole procedure is then bootstrapped with replacement to provide standard errors.



(a) **(b)**

Figure E.1: Distribution of observed Y

Notes: Probability density functions of Y and $ln(Y)$ for a sample of N=100,000 from Monte Carlo Exercise.

Appendix F Monte Carlo Exercise - A Replication

The Monte Carlo experiment of 100,000 replications with sample size of 250 is shown in table F.1 columns 1-4 and a sample size of 1000 shown in columns 5-8 of the same table. The results for the truth, naive, unfeasible and IPW based estimators of the ATT and the Coefficient of variation treatment effect on the treated (CVTT) are precise replications of the results of Firpo and Pinto (2016). Firpo and Pinto (2016) conclude that IPW yields favourable results in terms of bias and RMSE compared to the alternative inequality treatment effect estimators by Chernozhukov, Fernández-Val and Melly (2013) and Juhn, Murphy and Pierce (1993). PERM regression based results are presented alongside these replication results and show very similar results to IPW based estimates. PERM regression performs in terms of bias and RMSE compared to IPW, and by extension also compared to the alternative inequality treatment effect estimators by Chernozhukov, Fernández-Val and Melly (2013) and Juhn, Murphy and Pierce (1993).

Extending the Monte Carlo simulation to consider the variance, both IPW and PERM regression based estimators of the variance are similar in terms of bias and RMSE. The naive estimates show there is meaningful selection into treatment and both IPW and PERM regression are able to account for this when correctly specified (quadratic specification). The standard deviation and RMSE show that the variance of both IPW and PERM regression point estimates decrease as the sample size increases, as expected. The sample bias correction of the PERM regression VTT estimator reduces the bias, but the SSB is small compared to the precision of the estimates.

Table F.1: Monte Carlo Exercise (replications 100,000)

	Sample Size 250				Sample Size 1000			
	Average	Standard Deviation	Bias	Root Mean Squared Error	Average	Standard Deviation	Bias	Root Mean Squared Error
PANEL A: REPLICATION OF FIRPO AND PINTO (2016)								
Mean Treatment effect on the Treated:								
True : 1.102								
Infeasible	1.102	0.109	-0.000	0.109	1.102	0.054	-0.000	0.054
IPW quadratic	1.100	0.111	-0.002	0.111	1.100	0.055	-0.002	0.055
PERM quadratic	1.102	0.111	-0.000	0.111	1.102	0.055	-0.000	0.055
Coefficient of Variation Treatment effect on the Treated:								
True : 0.269								
Infeasible	0.250	0.126	-0.019	0.128	0.263	0.077	-0.006	0.077
IPW quadratic	0.259	0.130	-0.010	0.131	0.270	0.079	0.001	0.079
PERM quadratic	0.253	0.130	-0.016	0.131	0.264	0.080	-0.005	0.080
PANEL B: EXTENSION								
Variance Treatment effect on the Treated:								
True : 1.285								
Infeasible	1.285	0.872	-0.000	0.872	1.284	0.431	-0.001	0.431
IPW quadratic	1.288	0.872	0.003	0.872	1.286	0.431	0.002	0.431
PERM quadratic	1.275	0.865	-0.010	0.865	1.282	0.430	-0.003	0.430
PERM quadratic sample bias corrected	1.285	0.871	0.000	0.871	1.284	0.431	-0.001	0.431

Appendix G Monte Carlo Exercise - Additional Material

In table G.1 we present Monte Carlo Simulation results for 100,000 replications of sample sizes 250 and 1,000 to see in more detail the small sample properties of the estimators. The difference between the unfeasible and naive estimates is relatively minor suggesting the degree of selection is into treatment is small for the variance, and skewness. Across both the VTT and STT the sample bias for the PERM estimator is very similar to the unfeasible estimator and small relative to the precision. Again PERM and IPW show very similar results. Finally there is indicative evidence that the bias correction of the PERM sample estimators improves the estimate. However, the statistical precision is not high enough, even with 100,000 replications, to say anything with meaningful certainty.

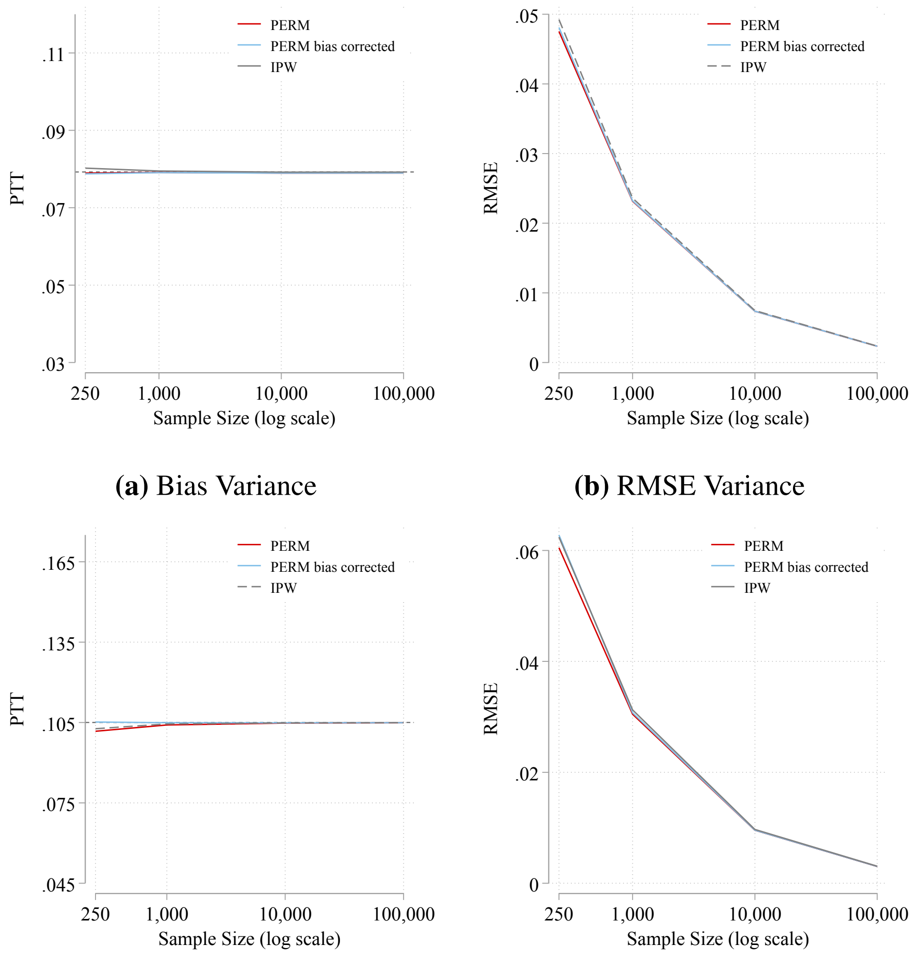
Figure G.1 presents MCS results for the bias and RMSE by sample size when considering the log of outcomes, a distribution that more closely follows a normal distribution compared to the non-logged outcomes distribution. For both the VTT and STT there is no clear evidence of meaningful bias and the RMSE figures suggest the estimates approach the true value with increasing precision as the sample size increases.

Figure G.2 represents MCS results for the bias of the standardised skewness (SSTT) for both unlogged and logged outcomes. The results show that the SSTT is substantially biased in small samples, relative to the precision. This is because they are ratios, which perform poorly in small samples (Joanes and Gill 1998). The RMSE error results in Figure G.3 suggest that the SSTT approaches the true value with increasing precision as sample size increases, however. Coverage of the SSTT is poor in small samples as shown in Figure G.4 but this improves with sample size, or with data that more closely approximates a normal distribution.

The results are very similar across PERM and IPW.

Table G.1: Monte Carlo Exercise (replications 100,000)

	Sample Size 250				Sample Size 1000			
	Average	Standard Deviation	Bias	Root Mean Squared Error	Average	Standard Deviation	Bias	Root Mean Squared Error
Variance Treatment effect on the Treated:								
True : 1.285								
Unfeasible	1.287	0.884	0.001	0.884	1.283	0.436	-0.002	0.436
Naive	1.302	0.885	0.017	0.885	1.299	0.436	0.013	0.436
IPW	1.288	0.884	0.002	0.884	1.284	0.436	-0.002	0.436
PERM	1.277	0.877	-0.008	0.877	1.281	0.435	-0.004	0.435
PERM bias corrected	1.287	0.883	0.001	0.883	1.283	0.436	-0.002	0.436
Skewness Treatment effect on the Treated:								
True : 6.893								
Unfeasible	6.852	31.020	-0.041	31.020	6.835	15.506	-0.058	15.506
Naive	6.864	31.020	-0.029	31.020	6.846	15.506	-0.046	15.506
IPW	6.853	31.021	-0.040	31.020	6.835	15.506	-0.058	15.507
PERM	6.771	30.643	-0.122	30.643	6.815	15.460	-0.078	15.460
PERM bias corrected	6.935	31.461	0.042	31.461	6.855	15.552	-0.037	15.552



(a) Bias Variance

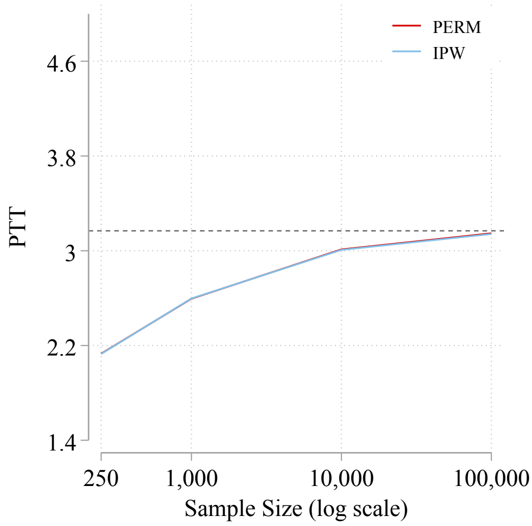
(b) RMSE Variance

(c) Bias Skewness

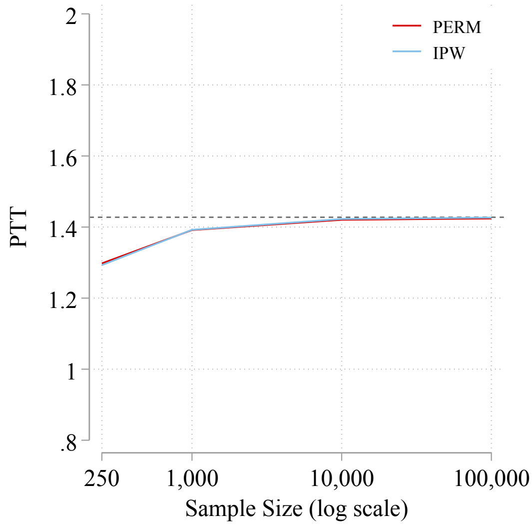
(d) RMSE Skewness

Figure G.1: Monte Carlo Exercise - Bias, Root Mean Square Error for ln(Y)

Notes: This figure is a replication of Figure 1 but for log transformation of outcome. These figures plot lines of sample estimator vs the 'truth' (to illustrate bias) and lines of the RMSE of the Parameter Treatment effect on the Treated by sample size from the Monte Carlo Exercise of 20,000 replications for each sample size. Estimates are provided by two methods, PERM and IPW. PERM Bias corrected utilises the sample bias correction from Appendix B. The y scale for the bias figures is the true value +- 1 RMSE.



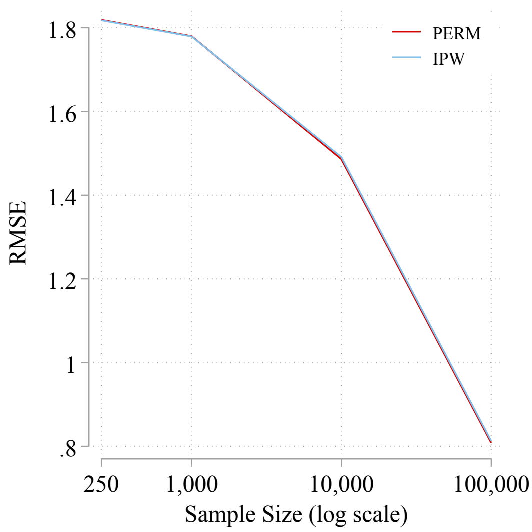
(a) Std. Skewness Y



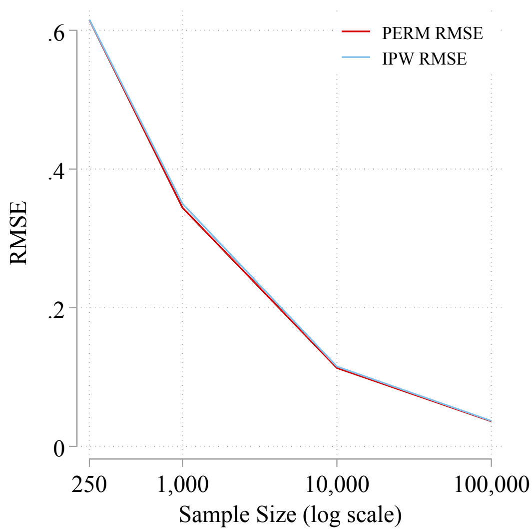
(b) Std. Skewness ln(Y)

Figure G.2: Monte Carlo Exercise - Bias

Notes: These figures plot a line of the average sample estimator of the Parameter Treatment effect on the Treated from the Monte Carlo Exercise with varying sample size and 20,000 replications for each sample size versus the 'truth' (horizontal dotted line). Estimates are provided by two methods, PERM and IPW and alternative methods of standard error estimation. The y scale is the true value +- 1 RMSE.



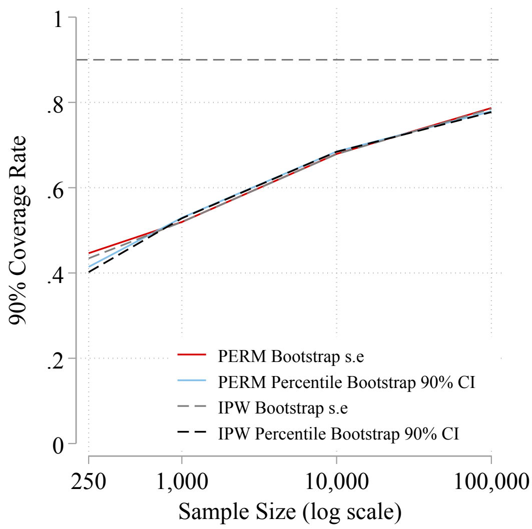
(a) Std. Skewness Y



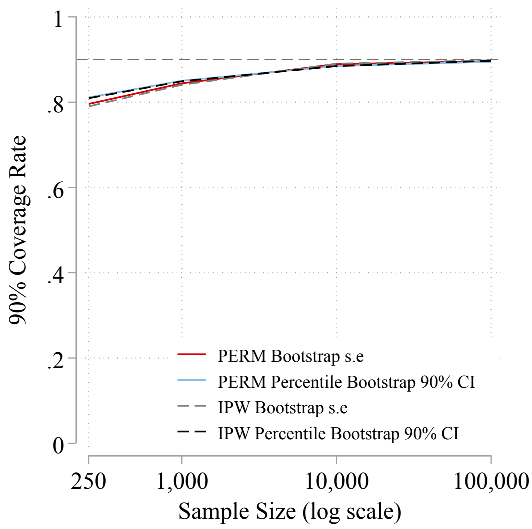
(b) Std. Skewness ln(Y)

Figure G.3: Monte Carlo Exercise - Root Mean Square Error

Notes: These figures plot a line of the RMSE of the Parameter Treatment effect on the Treated from the Monte Carlo Exercise with varying sample size and 20,000 replications for each sample size.



(a) Std. Skewness Y



(b) Std. Skewness ln(Y)

Figure G.4: Monte Carlo Exercise - 90% Coverage Rates

Notes: These figures plot a line of the 90% coverage rates of the Parameter Treatment effect on the Treated from the Monte Carlo Exercise with varying sample size and 20,000 replications for each sample size. Estimates are provided by two methods, PERM and IPW and alternative methods of standard error estimation. Bootstrap standard errors and Percentile Bootstrap 90% Confidence Intervals are from a non-parametric bootstrap with 200 replications.

Appendix H Unions - Data and Methods

We begin our analysis from 1983 because it marks the initial year when the ORG supplement inquired about union membership status. The dependent variable is the real logarithm of hourly wages for all wage and salary workers. Our set of explanatory variables encompasses six education categories, marital status, non-white ethnicity, and nine experience classifications.

The dependent variable refers to the real logarithm of hourly wages for all wage and salary earners. Hourly wages are directly assessed for hourly employees and are determined by dividing typical earnings by usual hours worked for other employees. Explanatory variables include six levels of education, nine levels of labour market experience, marital status and ethnicity. Education is grouped into six categories: 0 – 8 years of schooling, High school dropout, High school, some college, college graduate, college post-graduate which are in turn defined as dummy variables (ed0 – ed5) Labour market experience is grouped into eight categories: 0 – 4 years, 5 – 9 years, 10 – 14 years, 15 – 19 years, 25 – 29 years, 30 – 34 years, 35 – 39 years, and 40+ plus years, which are in turn defined as dummy variables (ex1 – ex9). Ethnicity is defined as a dummy variable equal to one if non-white, marital status is defined as a dummy variable equal to one if married. The reference category is high school education, 20-24 years experience, non-married and white.

The parameters of interest are the mean (MTT), variance (VTT) and standardised skewness (SSTT) treatment effect of the treated. To estimate these parameters using PERM regression dummy variables for education (ed0 – ed5), labour market experience (ex1 - ex9), ethnicity (non-white), marital status (married) are interacted with union coverage status in regressions of the first four raw moments, providing observed and counterfactual raw moments following the method set out in section 2.1. The PTTs are then calculated and this whole procedure is bootstrapped with replacement to provide standard errors.

Application of IPW to estimate the PTTs of interest requires two weighting functions that convert the observed distribution functions into weighted distribution functions of the observed and counterfactual distribution functions, conditional on being treated. The propensity score, the conditional probability of being treated, is defined as $p(x) \equiv Pr[D=1|X=x]$, where $X \in \chi$. The unconditional probability of being treated is defined as $p \equiv Pr[D=1]$, which we assume exists and is positive. The two weighting functions of interest are:

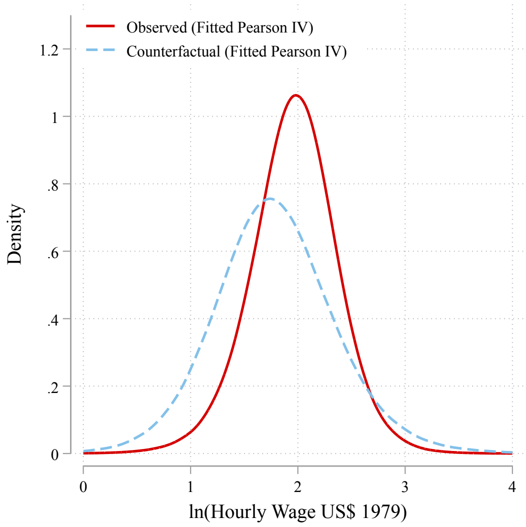
$$\omega_{11}(d, p(x)) = d/p$$

$$\omega_{01}(d, p(x)) = ((1-d)/(1-p(x))) \times (p(x)/p)$$

p is observed. $p(x)$ is estimated using probit regression, with union coverage status as the dependent variable and dummy variables for education (ed0 – ed5), labour market experience (ex1 - ex9), ethnicity (non-white), marital status (married) as explanatory variables. PTTs are then estimated using IPW re-weighting to provide the treated and counterfactual estimates. This whole procedure is then bootstrapped with replacement to provide standard errors.

Appendix I Unions - Additional Material

For some it may be easier to understand the implications of differences in the mean, variance, and skewness by illustrating these on representative probability density functions. For this purpose we fit the observed and counterfactual parameters to Pearson distributions using the 'pearsonFitM' package in 'R'. The Pearson distribution is a four-parameter family of continuous probability distributions. Figure I.1 shows the log hourly wage distribution of those covered by unionisation. The solid red curve depicts the probability density function of a Pearson IV distribution which is fitted to the sample estimates of the mean, variance, and skewness for log wages of those unionised. The dashed blue curve depicts their estimated counterfactual probability density function if they were not unionised as represented by a Pearson I distribution fitted based on the PERM estimates of the counterfactual mean, variance, and skewness. The figure illustrates how those covered by union membership have a wage distribution whose center of mass is higher and has a higher peak around the mean compared to the counterfactual. The more 'peaky' distribution means more individuals are nearer the mean compared to the counterfactual, and thereby a less variation in wages.



(a)

Figure I.1: Hourly Wage distribution; Observed and Counterfactual

Notes: This figure compares the fitted distribution of log hourly wages for those that were unionized to a fitted counterfactual distribution if these same people were not unionised. The density function of a Pearson IV distribution estimated using the mean, variance, skewness and kurtosis of the observed log hourly wages (see table 2 for the parameter values, observed kurtosis is 4.09) is compared with a Pearson IV distribution fitted to the mean, variance, skewness and kurtosis of the counterfactual distribution estimated by PERM regressions. Note we keep the kurtosis unchanged for the counterfactual distribution.

Appendix J Unions - PERM versus RIF

In this section we show how the Partial Policy Effect (PPE) is a first order approximation of the DPTT. PERM can estimate both these parameters and thereby illustrates that the Recentered Influence Functions (RIF), a linearisation method, provides an approximation for the true treatment effect.

We compare the estimated Partial Policy Effects of union coverage on the variance in log hourly wages using the PERM based approach and the Recentered Influence Function based approach. RIF regression introduced by Firpo, Fortin and Lemieux (2009) estimates the Partial Policy Effect (PPE) of a covariate on any parameter of interest as long as the parameter is differentiable, and applications have included the unconditional quantiles (Firpo, Fortin and Lemieux 2009), the variance, the Gini (Fortin, Lemieux and Firpo 2011, Firpo, Fortin and Lemieux 2018) and bivariate measures of income related health inequality (Heckley, Gerdtham and Kjellsson 2016).

While weighting approaches do not allow estimation of the PPE on the parameter of interest the PERM regression approach does. For example with the variance we have $\frac{dVar[Y]}{dX} = \frac{d(\mathbb{E}[Y^2]-\mathbb{E}[Y]^2)}{dX} = \frac{d\,\mathbb{E}[Y^2]}{dX} - 2\,\mathbb{E}[Y]\frac{\mathbb{E}[Y]}{dX}$ where the estimated regressions for $\mathbb{E}[Y^2]$ and

$\mathbb{E}[Y]$ are used in their place. We illustrate this by estimating the PPE of union coverage on the variance of log hourly wages, results shown in table J.1). PERM regression and RIF regression find the same PPE estimate of -0.1328 on the variance of log hourly wages and with the same standard errors. However, RIF regression cannot estimate the true PTT and PPTE impacts because it first linearises the distribution parameter before estimating the regression and thus can only be used to approximate the impact of small changes. So although RIF regression can be extended to account for selection into treatment when selection is not observed such as the use of Fixed Effects or Differences in Differences, it can only estimate the PPE and not the PTT. Yet the PTT is the parameter of policy interest.

Table J.1: Partial Policy Effect of Unionisation on Log Hourly Wages Inequality

	PERM	RIF
Union	-0.1328***	-0.1328***
	(0.0016)	(0.0016)

Notes: This table presents the Partial Policy Effect of unionisation on the variance of hourly wages in the USA. Each cell represents results from separate estimates with corresponding standard errors in parenthesis. PERM linearisation standard errors are provided by the linear approximation method of Graubard and Korn (1999) and RIF robust standard errors are analytical. PERM and RIF estimates utilise fully factorised controls for ethnicity, marriage status, education and experience all fully interacted with unionisation status. PERM estimates are of the derivative over the whole population.

Appendix K Unions - PERM Based Sub-group Analysis of the Variance

In this section we illustrate decomposing the variance PTT of unionisation by ethnic group. The variance decomposition formula states that the variance is the sum of conditional group variances plus a between groups variance effect, and is given by:

$$Var(Y) = \mathbb{E}[Var(Y|G)] + Var(\mathbb{E}[Y|G]) \tag{K.1}$$

The results in table K.1 show that the population variance effect of unionisation is fairly uniform across ethnicity groups and not driven by changes in the means between the groups due to unionisation.

Table K.1: PTT of Unionisation on Log Hourly Wages Inequality, sub-group analysis

	PERM
Total union effect	-0.1655***
	(0.0016)
Within ethnic groups effect	-0.1644***
	(0.0019)
Variance within white ethnic group	-0.1658***
	(0.0020)
Variance within non-white ethnic group	-0.1530***
	(0.0053)
Between ethnic groups effect	0.0011

Notes: Each cell represents results from separate estimates with corresponding standard errors in parenthesis. PERM linearisation standard errors are provided by the linear approximation method of Graubard and Korn (1999). PERM utilises fully factorised controls for ethnicity, marriage status, education and experience all fully interacted with unionisation status.

Appendix L Institutional Background to Sweden's School System and the Compulsory School Reform

The Swedish comprehensive school system was first introduced in the late 1940s and rolled out over time through the 1950s and 1960s across the municipalities of Sweden. As noted in Husén (1986) "The main motive for a structural [educational] reform was in their view to provide increased equality by providing equal access to further education irrespective of social class and place of residence. The inequalities existed not only between social classes but also, and to an even greater extent, between rural and urban areas". The reform is described in detail in Holmlund (2007). Here, we provide details of its most important elements.

Prior to the comprehensive school reform children went to Swedish primary school (*folkskola*), for seven or eight years depending on which municipality one lived in. To go on to higher education children had to apply to a selective junior secondary school (*realskola*) which was subject to ability testing. Selective schools were from either fourth grade or sixth grade, depending on the school and continued up to ninth or tenth grade. Attendance at a selective junior secondary school was necessary in order to attend upper secondary school, which in turn was necessary in order to be accepted to university. Students who stayed on at primary school for the full seven or eight years could not pursue further education beyond the minimum years of schooling. They could attend vocational training or find work. The old school system in Sweden therefore meant that children not tracked into a selective school had fewer years of schooling than those who went to a selective school, and the peer groups children were exposed to were more homogeneous in terms of ability.

The two most important components of the comprehensive school reform were the delay of the selective school system, bringing together higher ability and lower ability children for longer, and raising the minimum years of schooling to nine years for the lower ability children. All schools would now share a unified curriculum, however in reality this was a small change (Fischer et al. 2021). The reform roll-out was gradual and was rolled out municipality by municipality rather than by separate schools or classes. Municipalities could choose the timing of the roll-out of the reform, but eventually it was implemented across the whole of Sweden, where nearly 100 per cent coverage was achieved for cohorts born 1955 onwards. The reform was finally implemented for birth cohorts born in 1962.

Appendix M Compulsory School Reform: Data and Descriptives

We use Swedish administrative data provided by Statistics Sweden covering the universe born in Sweden during the years 1932-1950 and their parents. We exclude immigrants to ensure individuals were in fact impacted by the reform. Using each individual's personal identification number we are able to match individuals to various administrative data including the income and tax records for years 1968-2016, the national census for 1960 and 1970, population registers, and education records for years 1990-2016.

To measure exposure to the comprehensive school reform we utilise information the roll-out of the new comprehensive school system from the Swedish National Archives as used in Hjalmarsson, Holmlund and Lindquist (2015). For each municipality we have the first year in which the reform was implemented, which we minus 14 years to give the first birth cohort exposed to the reform. Individuals are assigned as exposed or unexposed to the reform using their year of birth and municipality of residence obtained from the 1960 and 1965 censuses. We follow Holmlund (2008) and Fischer et al. (2021) and assume that place of residence in 1960 is where individuals born between 1943 and 1948 went to school, and place of residence in 1965 for individuals born on or after 1949. For individuals born before 1943 we use place of residence of the mother in 1960 (father if information is missing for the mother). We do not use municipality of birth to assign treatment status as this refers to the location of the hospital in which they were born, which is not necessarily the same as their location of residence, inducing substantial measurement error in reform assignment (cf., Fischer et al. (2021)). Municipalities merged into larger units over time and to make municipalities consistent we map pre-1952 municipalities to municipalities as defined by the 1960 census.

In the few cases when a treated municipality merged with an untreated area, we have defined the whole new municipality as treated. Because there is a well documented partial roll-out of the reform in many municipalities one year before the reform was officially enacted (see e.g Holmlund (2007), Hjalmarsson, Holmlund and Lindquist (2015)), we recode for all municipalities the first exposed birth cohort in each municipality to be one birth cohort earlier.

Our earnings measure is annual total pre-tax labour earnings and includes earnings from work as an employee and as self-employed. We express earnings in 100,000 SEK and in 2016 prices (100,000 SEK translates to about $12,000 in 2016). We are interested in long-run earnings and follow the literature by using an average of earnings aged 36-55 years (Bhuller, Mogstad and Salvanes 2017).

We measure years of education by combining highest achieved level of schooling (primary and upper secondary school) as measured in the 1970 Census with highest achieved level of post-mandatory schooling (vocational training and tertiary education) as recorded in the education records as documented for the years 1990-2016. We then assign the years of education typically associated with the achieved level of schooling and post-mandatory schooling to provide an approximation of the total years of education (see Fischer et al. (2022) for the exact algorithm used). For mothers and fathers education we use highest level of schooling achieved (primary and upper secondary school) as measured in the 1970 Census.

Our sample starts with 2,136,250 born in Sweden between and including the years 1932 through to 1950. We drop individuals whose data is not available on their place of residence, years of education, earnings and reform status. We then restrict the sample to those who were not living in the cities of Stockholm, Gothenburg and Malmö. This restriction is made because the reform was rolled out in different parts of these cities in different years making reform assignment difficult. These individuals made up about 19% of the total population in 1960 (Statistics Sweden, 1961). Finally we restrict the sample to individuals born ten years before and nine years after the pivotal reform cohort to ensure control cohorts are not too dissimilar to treated cohorts (there are no relevant controls for the treated group 10 years after reform implementation with a restriction of born 10 years before first affected cohorts, which is why it becomes an asymmetric sample restriction). Our final sample size is 1,096,170 individuals.

Table M.1 provides the descriptive statistics of the variables used in the analysis for the whole sample and split by those treated and not treated by the 9-year reform. Individuals in our sample have about 10 years worth of education, a variance of 7 years and the distribution is positively standardised skewed. The untreated have lower mean level of education, higher variance and skewness than the treated population. Average earnings between 36-55 years

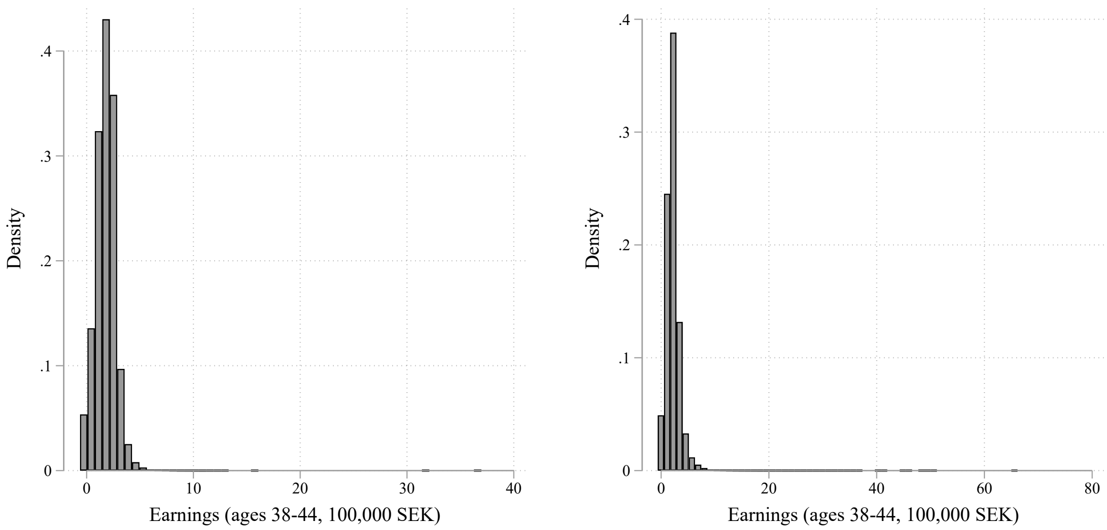
Table M.1: Descriptive Statistics

	WHOLE SAMPLE	UNTREATED	TREATED
YEARS OF EDUCATION			
Mean	10.28	9.98	11.05
Variance	6.94	7.10	5.68
Skewness	1.06	1.19	1.06
EARNINGS (100,000 SEK PER ANNUM)			
Mean	2.24	2.19	2.37
Variance	1.77	1.60	2.17
Skewness	4.33	3.08	6.29
EARNINGS AND EDUCATION			
Covariance	1.39	1.38	1.26
Beta	0.20	0.19	0.22
MEAN BACKGROUND CHARACTERISTICS			
Birth Cohort	1946	1945	1949
	[4.21]	[4.05]	[3.12]
9 Year Reform Exposure	0.28	0.00	1.00
	[0.45]	[0.00]	[0.00]
Male	0.51	0.51	0.51
	[0.50]	[0.50]	[0.50]
N	1,096,170	791,857	304,313

Notes: This table presents descriptive statistics of our two main outcomes (mean, variance and standardised skewness), years of education and earnings earnt between 36-55 years of age, and important background variables, and where relevant their standard deviations (shown in brackets) for the whole sample and for those treated and untreated by the 9 year comprehensive school reform.

old (expressed as 100,000 SEK per annum in 2016 prices) is 224,000 SEK for the whole population, with a variance of 1.77 and a positive standardised skew of 4.33. The untreated have lower mean earnings, lower variance and lower standardised skewness compared to the treated group. The treated group are born later but gender remains balanced across the groups.

We characterise the joint distribution of education and income using the covariance and also the slope (the covariance of education and earnings divided by the variance of education), or more commonly thought of as the raw association between years of education and earnings. The covariance for the whole sample and the untreated is circa 1.4, but the treated have a lower covariance of 1.26. This lower covariance does not clearly translate across to the slope coefficient of education on earnings, which is almost constant across all samples at 0.2, suggesting earnings is on average 20,000 SEK higher with each additional year of education.



(a) Less Than 9 Years of Education

(b) All

Figure M.1: Earnings Histograms of Untreated

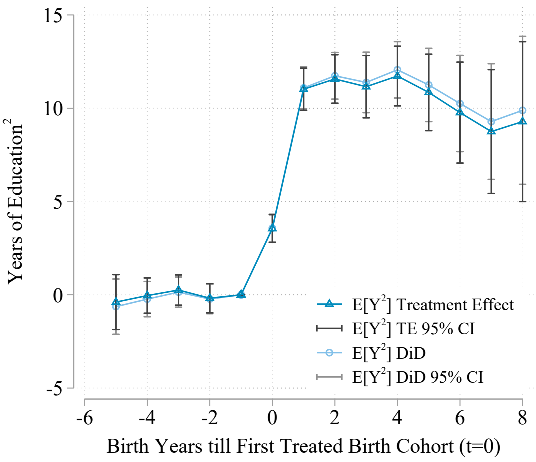
Figure M.1 presents histograms of the earnings distribution of the untreated for those with less than 9 years of education and for all. This illustrates that even within the low educated group there were still many high earners even before they were exposed to the comprehensive school reform.

Appendix N Compulsory School Reform: Additional Event Study Results

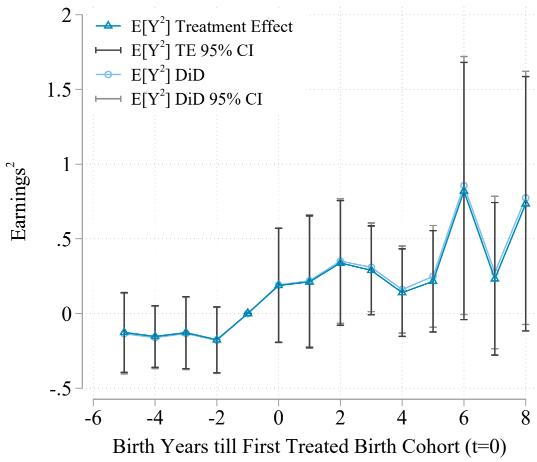
The PERM event studies in Figure 4 build upon PERM DiD estimates of raw moment treatment effects that are subsequently plugged into the raw moment based formulas for the parameters of interest. We illustrate these two steps. In Figures N.1 and N.2 panel (a) standard event study figures and PERM event study figures are compared for the second, third and fourth raw moments of education and income. The differences between the two is due to standard event study figures not accounting for the impact of trends and group differences in lower order raw moments mechanically affecting the estimates and biasing the raw moment treatment effects of higher-order raw moments. There are greater differences between the standard event studies and PERM event studies for years of education, suggesting education had a more substantial time trend during this time period.

Utilising the PERM event study raw moment estimates shown in Figures N.1 and N.2 panel (a) we can compare these estimates to a no parameter effect event study curve. The no parameter event study curve is the impact on the parameter driven by lower order raw moments. For there to be a parameter effect, the highest order raw moment effect must be different to the remaining part of the parameter formula formed of the lower order moments. For the variance, given by the formula $\mu_2 = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$, the no variance effect is given by $\mathbb{E}[Y]^2$. For the skewness, given by the formula $\mu_3 = \mathbb{E}[Y^3] - 3\,\mathbb{E}[Y^2]\,\mathbb{E}[Y] + 2\,\mathbb{E}[Y]^3$, the no skewness effect is given by $-(-3\,\mathbb{E}[Y^2]\,\mathbb{E}[Y] + 2\,\mathbb{E}[Y]^3)$.

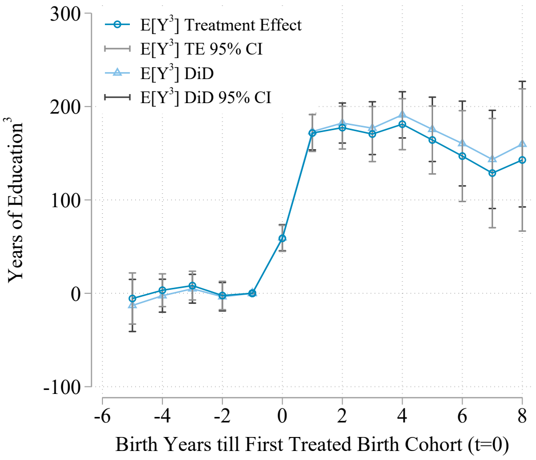
The difference between the raw moment curves and the no parameter effect curves shown in Figures N.2 panel (b) and N.3 is the parameter treatment effect for the treated, shown in Figure 4.



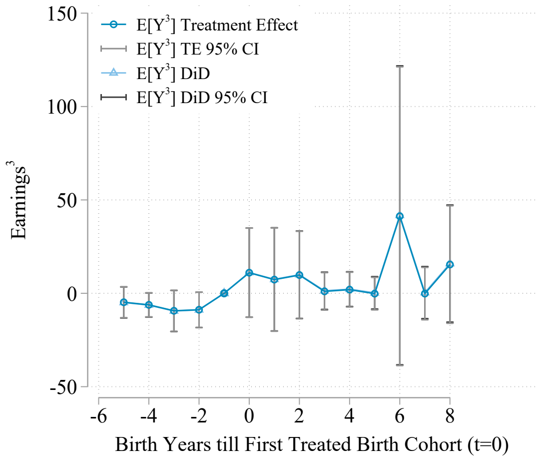
(a) $\mathbb{E}[Education^2]$



(b) $\mathbb{E}[Earnings^2]$



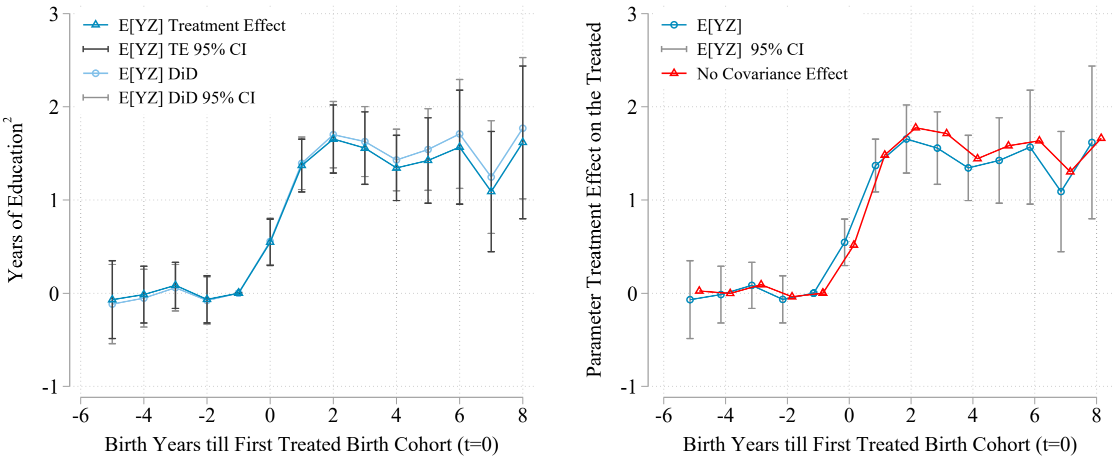
(c) $\mathbb{E}[Education^3]$



(d) $\mathbb{E}[Earnings^3]$

Figure N.1: Raw Moment Event study figures

Notes: Event study figures for raw moments of education and earnings, comparing standard DiD with PERM DiD (Treatment Effect) estimates. DiD is implemented using the method of De Chaisemartin and dHaultfoeuille (2020), PERM DiD utilises the method of 2.2. 95% clustered along municipality of residence confidence intervals are shown as capped vertical lines, provided by bootstrapping with 200 replications.

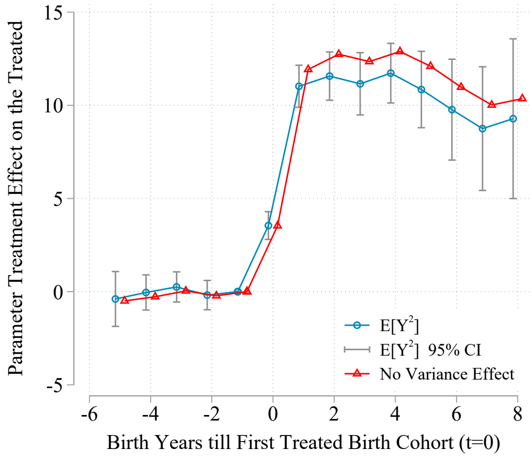


(a) PERM DiD vs DiD

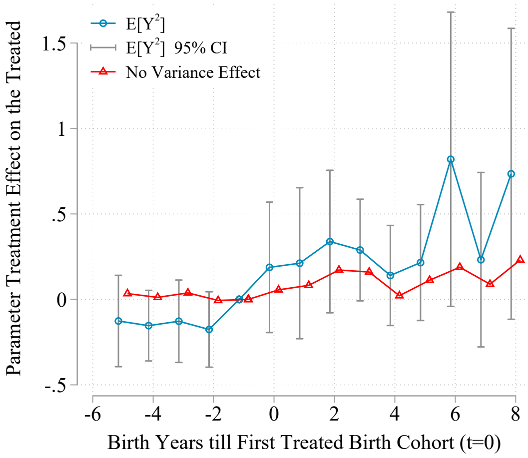
(b) $\mathbb{E}[WY]$ vs No Covariance Effect

Figure N.2: Bivariate Event Study Figures

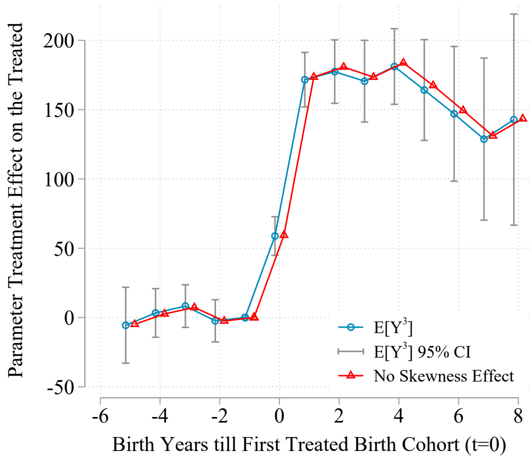
Notes: Event study figures for joint raw moment of education and earnings. DiD is implemented using the method of De Chaisemartin and dHaultfoeuille (2020), PERM DiD utilises the method of 2.2. 95% clustered along municipality of residence confidence intervals are shown as capped vertical lines, provided by bootstrapping with 200 replications.



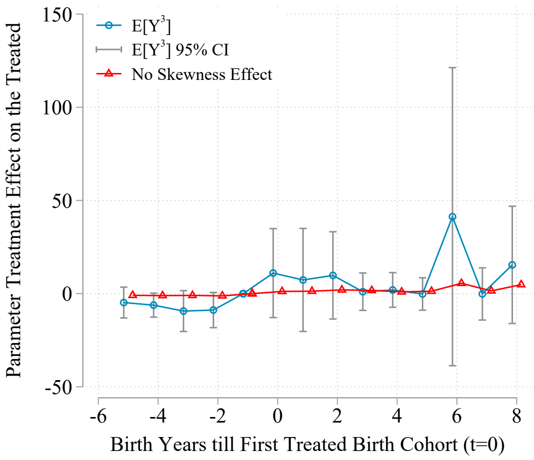
(a) $\mathbb{E}[Y^2]$ vs No Variance Effect



(b) $\mathbb{E}[Y^2]$ vs No Variance Effect



(c) $\mathbb{E}[Y^3]$ vs No Skewness Effect



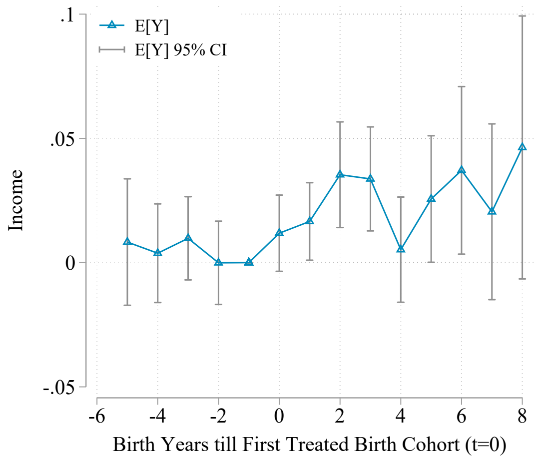
(d) $\mathbb{E}[Y^3]$ vs No Skewness Effect

Figure N.3: Raw Moment PERM Event Studies

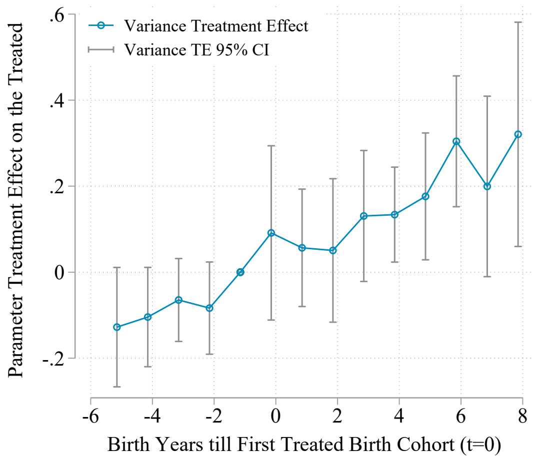
Notes: PERM event studies are estimated using the method of 2.2. 95% clustered along municipality of residence confidence intervals are shown as capped vertical lines, provided by bootstrapping with 200 replications.

Appendix O Compulsory School Reform: Robustness of Empirical Results

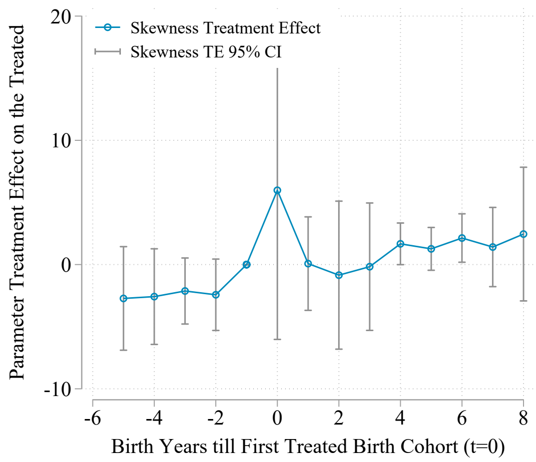
Figure O.1 presents PERM event studies for the mean, variance and skewness PTTs of earnings for a sample removing the highest earner from the 293 reform year - birth cohort cells. The precision of the estimates improves, especially for the higher-order distribution parameters. The large jump in t=6 is no longer present. However, the substantial conclusions remain the same; there was a positive mean earnings effect of the compulsory school reform, a possible positive earnings variance effect but no impact on the skewness of earnings.



(a) $\mathbb{E}[Y]$



(b) $Variance(Y)$



(c) $Skewness(Y)$

Figure O.1: Income Distribution Parameter Event study figures, excluding outliers

Notes: These figures are as Figure 4 but instead estimated on a sample that excludes the single highest earner from each g,t cell. See notes for Figure 4 for details.

Table O.1 presents the results of covariate balancing regression estimates. The outcomes are parent's occupational status as recorded in the 1960 status, and whether either parent is recorded as having the outcome. Column one provides estimates from an OLS regression of reform status on parent occupation group controlling for child birth cohort fixed effects. Column two utilises the DiD method of De Chaisemartin and dHaultfoeuille (2020) to estimate the difference between the treated and control groups. The results in column one show economically meaningful and statistically significant associations between reform status and parent occupation groupings suggesting non-randomness of reform roll-out even when controlling for child birth cohort fixed effects. The results in column two suggest that applying a DiD approach substantially improves covariate balance.

Table O.1: Balancing Tests

	OLS	DiD
	(1)	(2)
PARENT'S CHARACTERISTICS		
Blue Collar Worker	0.021 ***	-0.008 ***
	(0.006)	(0.003)
White Collar Worker	0.066 ***	0.004
	(0.013)	(0.003)
Farmer	-0.079 ***	-0.002
	(0.012)	(0.002)

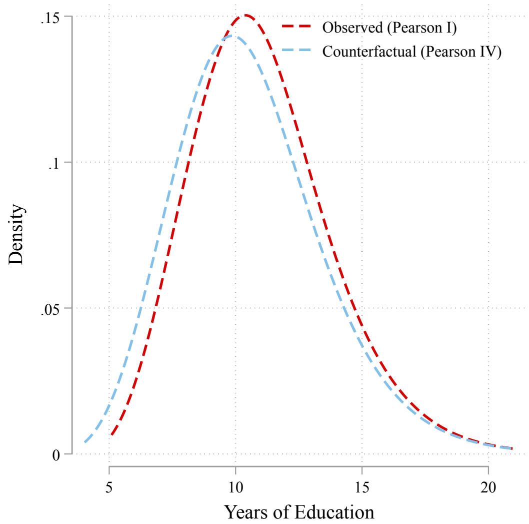
Notes: This table presents estimates of the impact of reform exposure on background characteristics of the parents of children exposed to the reform. Each variable is dummy variable equal to one if either parent has the characteristic. Column (1) presents results of an OLS regression of parent characteristic on reform status and cohort fixed effects, column (2) presents results of a DiD event study regression using the approach of De Chaisemartin and dHaultfoeuille (2020). Clustered along municipality level robust standard errors are presented in parenthesis in column (1) and bootstrap standard errors accounting for clustering at municipality level from 200 replications are presented in column (2). *** p<0.01, ** p<0.05, * p<0.1.

Appendix P Compulsory School Reform: Fitted PERM results

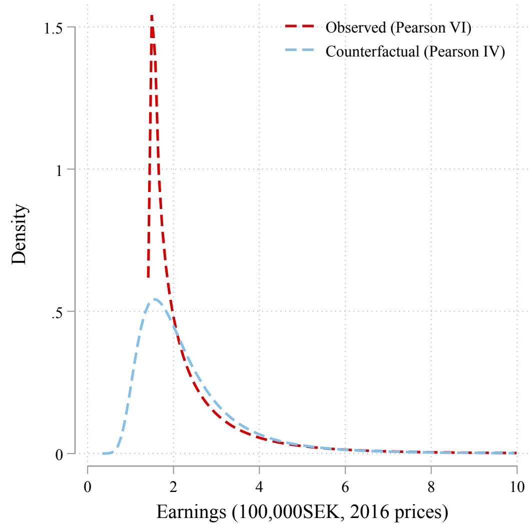
The parameter estimates in Table 3 provide important information as to how the distributions of earnings and education independently and jointly have changed as a result of the school reform. To illustrate graphically what these estimates imply about the observed and counterfactual distribution of outcomes we supplement these results with probability density functions and cumulative density functions predicted using Pearson distributions with the same mean, variance, skewness and kurtosis. Pearson distributions are a four-parameter family of unimodal and continuous distributions. The skewness and kurtosis help determine the type of Pearson distribution.

Figure P.1 illustrates the observed and counterfactual distributions provided by fitting a Pearson I/IV distribution for education and Pearson IV/VI distribution for earnings parameterised using the observed and counterfactual distribution parameters obtained from Table 3.[18] The fitted Pearson distribution function illustrates that for years of education (panel (a)), individuals at the lower tail increased their years of education while the upper end of the distribution was largely unaffected. The corresponding cumulative density curve in panel (c) suggests that the observed distribution dominates the counterfactual, no-one was made worse off, whilst many were made better off. Therefore, these illustrative figures suggest that the reform resulted in a Pareto welfare improvement in educational outcomes.

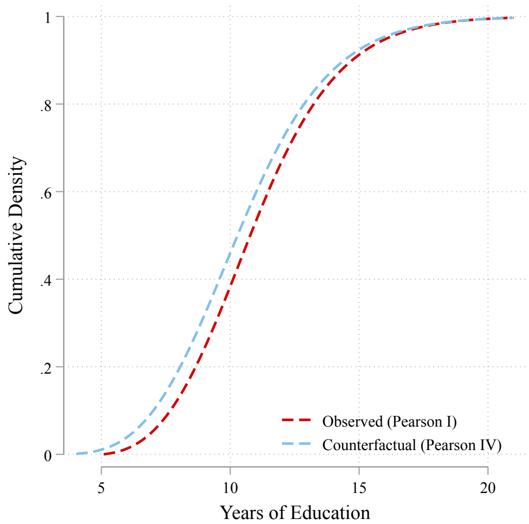
[18]To estimate the Pearson distributions we use the R program PearsonDS for 100,000 observations and then estimate a kernel function of the probability densities.



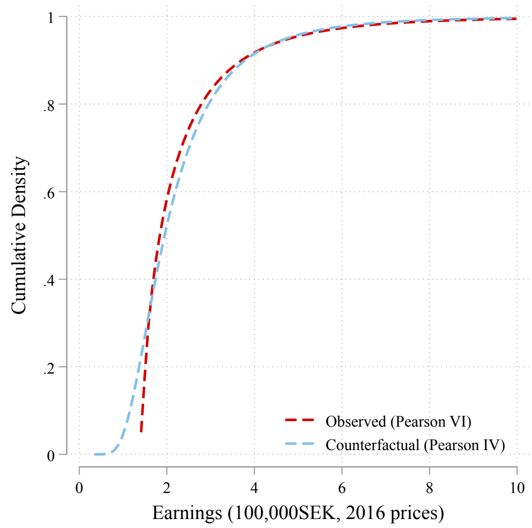
(a) Probability Density (education)



(b) Probability Density (Earnings)



(c) Cumulative Density (education)



(d) Cumulative Density (Earnings)

Figure P.1: Modelled Education and Earnings Distributions

Notes: These figures plot the distribution of years of education (panel (a)) and earnings (average earnings aged 36-55, 100,000 SEK per annum (panel (b)) for those that were exposed to the school reform. The density function for the observed treated population from a fitted Pearson distribution for education and earnings is presented as the red line and superimposed over the counterfactual fitted Pearson curves in blue. Both curves are calculated using the observed and counterfactual mean, variance, and skewness provided by PERM in table 3. Note that we keep standardised kurtosis unchanged from the observed distribution.

Turning to earnings, the fitted probability distribution in panel (b) illustrates how the bottom tail of the distribution have seen an improvement from the reform. The Cumulative density curve shown in panel (d) shows that the observed distribution of earnings crosses the counterfactual and therefore there is no dominance, suggesting an unclear welfare outcome in terms of earnings.

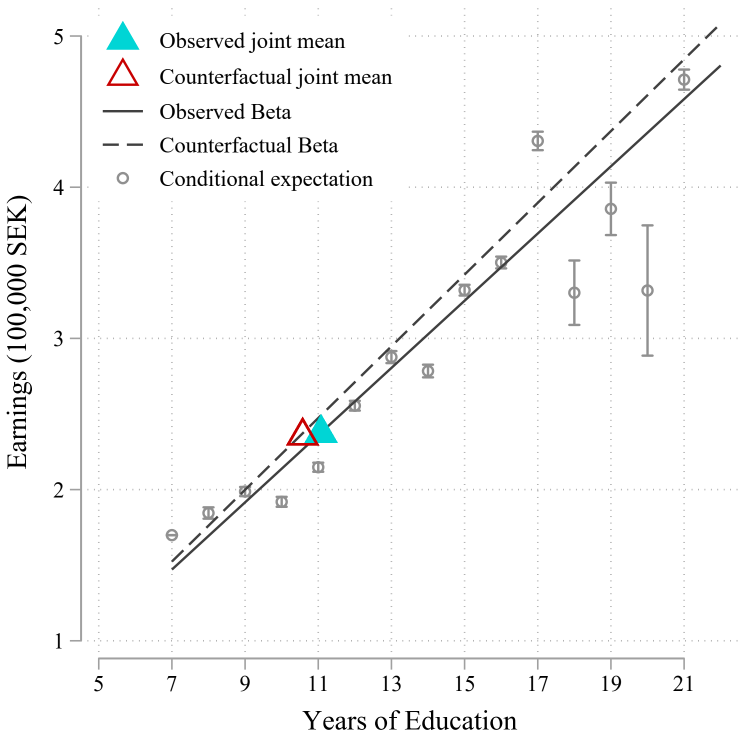


Figure P.2: Joint Education and Earnings Inequality

Notes: This figure plots the joint relationship between years of education and earnings. The observed joint mean is represented by the solid green triangle, it's counterfactual as a hollow red triangle. The linear fit of observed income and years of education (beta in Table M.1) is represented by a solid line graph and its counterfactual as a dashed line. The observed conditional expectation of income for each year of education are presented as hollow circles with corresponding 95% confidence intervals as spikes.

In Figure P.2 we illustrate the bivariate relationship between education and log income as a slope coefficient, Beta (solid line). This line represents the linear approximation of the expected income for a given years of schooling. The true conditional expectation function is plotted as grey circles and suggests this linear line is a good approximation of the education and log income relationship. The observed joint mean of education and log income is represented as a turquoise triangle. The counterfactual joint mean is to the left, reflecting that earnings changes were small for the relatively large changes in years of education. The change in the slope reflects this, with a flatter slope for the observed treated group.

Appendix References

Bhuller, Manudeep, Magne Mogstad, and Kjell G Salvanes. 2017. "Life-cycle earnings, education premiums, and internal rates of return." *Journal of Labor Economics*, 35(4): 993–1030.

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. 2013. "Inference on counterfactual distributions." *Econometrica*, 81(6): 2205–2268.

De Chaisemartin, Clément, and Xavier dHaultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*, 110(9): 2964–2996.

Firpo, Sergio, and Cristine Pinto. 2016. "Identification and estimation of distributional impacts of interventions using changes in inequality measures." *Journal of Applied Econometrics*, 31(3): 457–486.

Firpo, Sergio, Nicole M Fortin, and Thomas Lemieux. 2009. "Unconditional quantile regressions." *Econometrica*, 77(3): 953–973.

Firpo, Sergio P, Nicole M Fortin, and Thomas Lemieux. 2018. "Decomposing wage distributions using recentered influence function regressions." *Econometrics*, 6(2): 28.

Fischer, Martin, Gawain Heckley, Martin Karlsson, and Therese Nilsson. 2022. "Revisiting Sweden's comprehensive school reform: Effects on education and earnings." *Journal of Applied Econometrics*, 37(4): 811–819.

Fischer, Martin, Ulf-G Gerdtham, Gawain Heckley, Martin Karlsson, Gustav Kjellsson, and Therese Nilsson. 2021. "Education and health: long-run effects of peers, tracking and years." *Economic Policy*, 36(105): 3–49.

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. "Decomposition methods in economics." In *Handbook of labor economics*. Vol. 4, 1–102. Elsevier.

Graubard, Barry I, and Edward L Korn. 1999. "Predictive margins with survey data." *Biometrics*, 55(2): 652–659.

Heckley, Gawain, Ulf-G Gerdtham, and Gustav Kjellsson. 2016. "A general method for decomposing the causes of socioeconomic inequality in health." *Journal of Health Economics*, 48: 89–106.

Hjalmarsson, Randi, Helena Holmlund, and Matthew J Lindquist. 2015. "The Effect of Education on Criminal Convictions and Incarceration: Causal Evidence from Micro-data." *The Economic Journal*.

Holmlund, Helena. 2007. "A Researcher's Guide to the Swedish Compulsory School Reform." Working paper 9/2007, Swedish Institute for Social research, Stockholm University.

Holmlund, Helena. 2008. "A researcher's guide to the Swedish compulsory school reform." Centre for the Economics of Education, London School of Economics and Political Science.

Husén, Torsten. 1986. "Why did Sweden go comprehensive?" *Oxford Review of Education*, 12(2): 153–163.

Joanes, Derrick N, and Christine A Gill. 1998. "Comparing measures of sample skewness and kurtosis." *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1): 183–189.

Juhn, Chinhui, Kevin M Murphy, and Brooks Pierce. 1993. "Wage inequality and the rise in returns to skill." *Journal of political Economy*, 101(3): 410–442.