

On Causal Inference with Model-Based Outcomes*

Dmitry Arkhangelsky[†]

Kazuharu Yanagimoto[‡]

Tom Zohar [§]

June 26, 2025

Abstract

We study a causal inference problem with group-level outcomes, which are themselves parameters identified from microdata. We formalize these outcomes using population moment conditions and demonstrate that one-step Generalized Method of Moments (GMM) estimators are generally inconsistent due to an endogenous weighting bias, where policy affects the implicit GMM weights. In contrast, two-stage Minimum Distance (MD) estimators perform well when group sizes are sufficiently large. While MD estimators can still be inconsistent in small groups due to a policy-induced sample selection, we demonstrate that this can be addressed by incorporating auxiliary population information. An empirical application illustrates the practical importance of these findings.

Keywords: Causal Inference, Model-Based Outcomes, GMM, Minimum Distance, Child Penalty, Firm Wage-Premium, Credit Supply, TFP.

*We thank Manuel Arellano, Stephane Bonhomme, Kirill Borusyak, Nezih Guner, Lance Lochner, Victor Sancibrian, Pedro Sant'Anna, Davide Viviano, and Dean Yang for their feedback, which has greatly improved the paper. We also thank seminar participants at Harvard, UCLA, Carlos III, CEMFI, IFAU, Tel-Aviv U., Hebrew U., and Warwick, and participants in the 15th Research Workshop Banco de España - CEMFI, ELSE 2024, SAE 2024, Econometric Society Winter meeting 2024, Barcelona GSE Summer Forum 2024, ASSA 2025, and Causal Inference with Panel Data workshop 2025 at Emory for their helpful comments. This paper incorporates and extends the ideas developed in the previously circulated draft “Using Event Studies as an Outcome in Causal Analysis” by the same authors. We gratefully acknowledge financial support from the Maria de Maeztu Unit of Excellence CEMFI MDM-2016-0684 Grants CEX2020-001104-M, PID2021-128992NB-I00, PID2022-143184NA-I00, and RYC2022-037963-I; Funded by MICIU/AEI/10.13039/501100011033, ESF+, and by “ERDF/EU”, which covered data access to the Dutch Central Bureau of Statistics and travel expenses to present the project. All views and any errors are our own.

[†]CEMFI, CEPR, darkhangel@cemfi.es, corresponding author.

[‡]CEMFI, kazuharu.yanagimoto@cemfi.edu.es.

[§]CEMFI, CESifo, tom.zohar@cemfi.es.

1 Introduction

A significant body of empirical work assesses policies implemented at a group level (e.g., firm, school, local labor market, or county) and their impact on aggregate group-level outcomes. Rather than simple averages of individual data, research interest often centers on specific parameters estimated from micro-level models; examples include firm-specific wage premia ([Abowd et al., 1999](#)), bank-specific credit supply ([Khwaja and Mian, 2008](#)), firm-level productivity ([Amiti and Konings, 2007](#)), or the “child penalty” reflecting labor market consequences following childbirth ([Kleven et al., 2019](#)). These model-based parameters represent key group-level features potentially shaped by the policy of interest. Common estimation strategies involve either one-step approaches, such as micro-level regressions with policy interactions, or two-step methods, where group-specific parameters are estimated first before being related to the policy in a second stage.

Estimating causal effects on such model-based outcomes faces a core complication: the policy typically shifts the entire distribution of the underlying micro-level data. For instance, labor policies can alter labor market composition or mobility decisions relevant for wage premia, while childcare expansion can change the timing of fertility decisions in a given municipality. This distributional response fundamentally challenges causal analysis, raising key questions: How should the causal effect itself be defined when the underlying population distribution is policy-dependent? And under what conditions do standard one-step or two-step estimation strategies identify such effects?

This paper addresses these questions by providing a systematic approach to causal inference for policy effects on group-level parameters defined via micro-level models, motivated by applications like those mentioned above. We make three main contributions.

First, we develop a formal econometric framework that explicitly defines group-specific parameters through population moment conditions. These moments are computed using the potential distribution of micro-level data, inducing causal structure on the group-level outcomes of interest. This framework clarifies the target causal estimand—the average effect of the policy on these well-defined model-based outcomes—and helps emphasize different channels through which the policy can affect these outcomes.

Second, we demonstrate a fundamental problem with standard one-step estimation methods, such as OLS with policy interactions. By analyzing these approaches within the Generalized Method of Moments (GMM) framework, we identify a source of inconsistency arising from an endogenous weighting problem: by altering the underlying distribution of the micro-level data, the policy generically affects the implicit GMM weights. This correlation between the policy and the effective weights renders the estimator inconsistent for the target causal effect. This mechanism—distinct from recent concerns established in the difference-in-differences

(DiD) literature (discussed below)—presents a potentially widespread issue for procedures involving interactions between policy and micro-level variation.

Third, we analyze intuitive two-stage procedures, which we view as Minimum Distance (MD) estimators, that relate first-stage group-level estimates to policy. We demonstrate that standard MD estimators are generically inconsistent when group sizes remain fixed as the number of groups grows large, even if first-stage estimates are conditionally unbiased. This inconsistency stems from policy-induced sample selection (Heckman, 1979): if policy affects the probability that a group’s sample permits first-stage estimation (e.g., by influencing required sample variation), conditioning the second stage on estimability creates a selection bias, a potentially underappreciated limitation of the MD estimator with “small” groups. This inconsistency can, however, be overcome under relatively mild conditions. We show that the MD estimator becomes consistent and permits standard second-stage inference in the asymptotic regime where group sizes grow with the number of groups. Alternatively, even with fixed group sizes, consistency can be restored by incorporating auxiliary information, thereby linking this approach to design-based identification arguments.

Our framework and results provide three practical takeaways for applied researchers. First, the explicit definition of parameters via moment conditions encourages careful consideration of what object is truly being targeted and how policy might affect these model-based outcomes through direct and indirect channels. Second, our findings cast doubt on the general suitability of one-step GMM procedures, including OLS. Finally, while two-stage procedures avoid the specific GMM weighting bias, their reliability hinges on having sufficiently large within-group samples. In settings with many small groups, researchers should be aware of the potential for selection bias in standard MD estimators and consider whether auxiliary data can be leveraged to mitigate this issue.

Many empirical studies that focus on problems of a similar nature to the one we discuss often stop short of claiming causality, arguably in part because they lack a unifying theory for such analysis (see, e.g., discussions in Daruich et al., 2023; Schmieder et al., 2023; Bertheau et al., 2023). Our paper addresses this gap by developing a framework to define and identify causal effects on model-based outcomes.

Our analysis of one-step estimators relates to the recent literature on DiD methods dealing with heterogeneous treatment effects (e.g., Borusyak et al., 2024b, Callaway and Sant’Anna, 2021, de Chaisemartin and D’Haultfœuille, 2020, Goodman-Bacon, 2021, Sun and Abraham, 2021; see Arkhangelsky and Imbens (2024) for a survey) and related literature on heterogeneity in linear regressions (Goldsmith-Pinkham et al., 2024). While that work focuses on biases arising from problematic comparisons when effects vary across units and time, the endogenous weighting bias we identify in GMM-type procedures stems from a different source: the policy’s direct impact on the underlying population moments determining the estimator’s implicit

weights. Consequently, this bias can arise even in randomized experiments with homogeneous policy effects, representing a distinct challenge not addressed by fixing the comparison issues identified in the current literature.

Our findings connect to the literature on GMM estimation challenges, particularly concerning weighting matrices (e.g., Altonji and Segal, 1996; Newey and Smith, 2004). Unlike work focusing on finite-sample biases that arise from using estimated optimal weights, the GMM inconsistency we document can occur even with very simple, pre-specified weights. It is driven by policy affecting the population moments within the GMM objective function. Similarly, the related asymptotic inconsistency we find for MD estimators in small groups stems from policy-induced sample selection affecting the first stage, rather than challenges specific to estimating optimal second-stage weights.

Finally, our proposed use of known population moments to restore consistency for the MD estimator connects conceptually to design-based identification strategies. It parallels IPW estimation strategies, where known assignment probabilities are used to address composition issues, and resonates with recent design-based approaches that incorporate known population quantities directly into estimation (e.g., Arkhangelsky et al., 2024; Borusyak and Hull, 2023; Borusyak et al., 2024a). More broadly, as demonstrated by Imbens and Lancaster (1994), aggregate information generically improves micro-based estimation; in our model, the availability of such information mitigates inconsistency arising from finite group sizes.

To demonstrate the practical relevance of our theoretical findings, we present an empirical application analyzing the impact of the 2005 Dutch childcare expansion on "child penalty" measures of gender inequality. This policy setting provides a suitable case study in which the reform likely influenced both the outcome of interest (parental labor supply) and the distribution of the underlying individual-level data (fertility decisions). We demonstrate that a conventional one-step GMM/OLS estimation strategy suggests that the policy had a large impact on mothers' earnings and labor force participation. By contrast, our preferred two-step Minimum Distance (MD) approach, which is robust to the endogenous weighting bias we identify, finds substantially smaller effects. This notable difference in the results highlights that the methodological issues we raise are of practical consequence for applied work and that the choice of estimator can materially alter the conclusions of a policy evaluation exercise.

The remainder of the paper is organized as follows: we begin by developing our framework and illustrate it with three motivating examples (Section 2). We then analyze the properties of one-step GMM estimators (Section 3), and two-step Minimum Distance estimators (Section 4), highlighting pitfalls related to endogenous weighting and sample selection. Section 5 provides an empirical illustration using Dutch administrative data. Finally, Section 6 concludes.

2 Framework

This section develops an econometric framework for analyzing causal effects of group-level policies on parameters derived from micro-level data—such as ‘child penalty’ measures of gender inequality, local wage premia, firm-level productivity, or credit supply. We first formally define the setup, the parameters of interest, and estimators (Section 2.1). We then illustrate the scope using several examples (Section 2.2), and conclude by discussing the policy-induced composition effects (Section 2.3).

2.1 Econometric Setup

Consider a collection \mathcal{G} of G groups, indexed by g . Within each group g , we observe data $D_{g,i} \in \mathcal{D}$ for $i = 1, \dots, n_g$ units, drawn i.i.d. from a group-specific distribution F_g . A group-level policy $W_g \in \mathcal{W} \subseteq \mathbb{R}^p$ is assumed to causally affect the data-generating process F_g , inducing potential outcome distributions $F_g(w)$ for $w \in \mathcal{W}$. The observed distribution satisfies $F_g = F_g(W_g)$, thus effectively imposing a group-level Stable Unit Treatment Value Assumption (SUTVA).¹

We focus on a k -dimensional parameter vector $\theta_g \in \mathbb{R}^k$. We define this parameter implicitly via k population moment conditions:

$$\mathbb{E}_{F_g}[h(D_{g,i}, \theta_g)] = \mathbf{0}_k, \quad (1)$$

where $h : \mathcal{D} \times \Theta \rightarrow \mathbb{R}^k$ is a known function. This definition links θ_g directly to the underlying data distribution.² Since F_g depends causally on W_g , this naturally generates potential outcomes for the parameter:

$$\theta_g(w) \text{ solves } \mathbb{E}_{F_g(w)}[h(D_{g,i}, \theta_g(w))] = \mathbf{0}_k.$$

Defining the outcome $\theta_g(w)$ via moment conditions facilitates its economic interpretation, which, as we demonstrate using examples (Section 2.3), is not always straightforward and has critical implications for causal analysis. We focus on moment functions linear in the parameter:

$$h(D_{g,i}, \theta) = h_1(D_{g,i}) - h_2(D_{g,i})\theta.$$

Letting $H_{j,g}(w) := \mathbb{E}_{F_g(w)}[h_j(D_{g,i})]$ and assuming $H_{2,g}(w)$ is invertible, the potential outcome

¹See [Imbens and Rubin \(2015\)](#) for a textbook discussion of the potential outcome model of causality and related assumptions.

²Throughout the text, we index within-group expectations using F_g . Formally, any such expectation is a random variable due to its dependence on W_g . We use expectations without the F_g subscript to average over group-level uncertainty, which always includes W_g .

is $\boldsymbol{\theta}_g(w) = H_{2,g}(w)^{-1}H_{1,g}(w)$. Note that the dependence of $F_g(w)$ on w implies that $H_{1,g}(w)$ and $H_{2,g}(w)$ also depend on w .

To anchor our analysis, we consider an idealized benchmark estimator representing what applied researchers might compute if the true group-specific parameters $\boldsymbol{\theta}_g = \boldsymbol{\theta}_g(W_g)$ were directly observable. This benchmark, which we term the “oracle” estimator, takes the form of a linear regression motivated by common practice in policy evaluation:

$$(B^*, \boldsymbol{\alpha}^*, \{\boldsymbol{\lambda}_g^*\}) := \underset{\substack{B, \boldsymbol{\alpha}, \{\boldsymbol{\lambda}_g\} \\ \text{s.t. } \Gamma^\top B = 0, \Gamma^\top \boldsymbol{\alpha} = 0}}{\operatorname{argmin}} \sum_{g=1}^G \|\boldsymbol{\theta}_g - (\boldsymbol{\alpha} + \Gamma \boldsymbol{\lambda}_g + BW_g)\|_2^2. \quad (2)$$

Here, B ($k \times p$) captures the policy effects of primary interest, $\boldsymbol{\alpha}$ ($k \times 1$) is an intercept term, Γ ($k \times q$) defines dimensions of unobserved group heterogeneity, and $\boldsymbol{\lambda}_g$ ($q \times 1$) represents group-specific coefficients (fixed effects) along these dimensions. The constraints $\Gamma^\top B = 0$ and $\Gamma^\top \boldsymbol{\alpha} = 0$ are imposed to guarantee identifiability. This specification encompasses various standard approaches used for policy evaluation, including estimators based on two-way fixed effects models, such as the DiD approach. While the oracle regression (2) is infeasible in practice because $\boldsymbol{\theta}_g$ is unknown, it defines the target quantity of interest for our analysis. Our central question is whether, and under what conditions, feasible estimation strategies that rely (explicitly or implicitly) on estimates $\hat{\boldsymbol{\theta}}_g$ derived from micro-level data can mimic this oracle procedure. We do not analyze the statistical properties of the oracle estimator $(B^*, \boldsymbol{\alpha}^*, \{\boldsymbol{\lambda}_g^*\})$ itself, in particular its efficiency properties (see, however, Remark 3.1).

Motivated by empirical practice, we consider two estimation strategies: a two-step Minimum Distance (MD) and a one-step Generalized Method of Moments (GMM). The MD estimator first obtains group-specific estimates $\hat{\boldsymbol{\theta}}_g$, then performs the second-stage regression:

$$(\hat{B}^{MD}, \hat{\boldsymbol{\alpha}}^{MD}, \{\hat{\boldsymbol{\lambda}}_g^{MD}\}) := \underset{\substack{B, \boldsymbol{\alpha}, \{\boldsymbol{\lambda}_g\} \\ \text{s.t. } \Gamma^\top B = 0, \Gamma^\top \boldsymbol{\alpha} = 0}}{\operatorname{argmin}} \sum_{g=1}^G \left\| \hat{\boldsymbol{\theta}}_g - (\boldsymbol{\alpha} + \Gamma \boldsymbol{\lambda}_g + BW_g) \right\|_2^2. \quad (3)$$

The GMM estimator instead directly imposes the model structure $\boldsymbol{\theta}_g = \boldsymbol{\alpha} + \Gamma \boldsymbol{\lambda}_g + BW_g$ onto the sample moments $h_{g,n}(\boldsymbol{\theta}) := n_g^{-1} \sum_{i=1}^{n_g} h(D_{g,i}, \boldsymbol{\theta})$:

$$\begin{aligned} (\hat{B}^{GMM}, \hat{\boldsymbol{\alpha}}^{GMM}, \{\hat{\boldsymbol{\lambda}}_g^{GMM}\}) := & \\ & \underset{\substack{B, \boldsymbol{\alpha}, \{\boldsymbol{\lambda}_g\} \\ \text{s.t. } \Gamma^\top B = 0, \Gamma^\top \boldsymbol{\alpha} = 0}}{\operatorname{argmin}} \sum_{g=1}^G h_{g,n}(\boldsymbol{\alpha} + \Gamma \boldsymbol{\lambda}_g + BW_g)^\top \times A_g \times h_{g,n}(\boldsymbol{\alpha} + \Gamma \boldsymbol{\lambda}_g + BW_g), \end{aligned} \quad (4)$$

for chosen weighting matrices A_g . The subsequent sections analyze the statistical properties of \hat{B}^{MD} and \hat{B}^{GMM} , focusing on the conditions for consistency and highlighting potential biases.

Remark 2.1 (Data types). Our i.i.d. setup includes standard panel data as a special case, where each $D_{g,i}$ represents the entire path of outcomes for unit i . Alternatively, the group index g can denote time periods, accommodating analyses of independent cross-sections (e.g., policies varying by cohorts). This latter case can be extended from repeated cross-sections to incorporate richer panel structures, such as overlapping panels, at the cost of additional technical assumptions. When g represents time, SUTVA warrants careful consideration, as policies in one period may causally affect outcomes in subsequent periods.

Remark 2.2 (Overidentification). Economic applications often feature overidentified moment conditions, which allow for specification testing (e.g., parallel-trends tests). We focus on an exactly identified system (1) which could arise from combining a larger set of original moment restrictions. Overidentification raises a question of model misspecification and its interplay with the downstream policy analysis, which we do not engage with.

Remark 2.3 (IV instead of OLS). Our framework naturally extends to settings with endogenous policies instrumented at the group level. Appendix A.2.1 briefly illustrates this extension. Through a stylized example, we demonstrate how the challenges we identify in subsequent sections persist in applications with instrumental variables, potentially complicating recentering procedures emphasized in the recent work on formula instruments ([Borusyak and Hull, 2023](#); [Borusyak et al., 2024a](#)).

2.2 Motivating Examples

We now illustrate the scope of our framework using canonical examples from applied economics. These examples will show parameters of interest defined by moment conditions and illuminate contexts where the estimation challenges we analyze are particularly relevant.

2.2.1 DiD: The Child Penalty Example

A canonical application fitting our framework involves estimating group-level treatment effects using DiD, such as the “child penalty” – the effect of motherhood on women’s labor market outcomes (e.g., [Kleven et al., 2019](#)). Consider a simplified two-period setup where $Y_{g,i,t}$ is the labor market outcome (e.g., earnings) for individual i in group g (e.g., a region or cohort) at time $t \in \{1, 2\}$. Let $E_{g,i}$ be an indicator variable equal to one if individual i experiences a birth of the first child between periods 1 and 2, and zero otherwise.

A standard model in this literature relates the outcomes of an individual i to individual fixed effects $\gamma_{g,i}$, group-time effects $\delta_{g,t}$, and the event:

$$Y_{g,i,t} = \gamma_{g,i} + \delta_{g,t} + \tau_g E_{g,i} \mathbf{1}\{t = 2\} + \epsilon_{g,i,t}, \quad \mathbb{E}_{F_g}[\epsilon_{g,i,t}|E_{g,i}] = 0.$$

Here, τ_g represents the average effect of having the first child on the outcome in group g – the group-specific child penalty. We can then write the equation in differences:

$$\Delta Y_{g,i} = \tilde{\delta}_g + \tau_g E_{g,i} + \tilde{\epsilon}_{g,i}, \quad \mathbb{E}_{F_g}[\tilde{\epsilon}_{g,i}|E_{g,i}] = 0,$$

where $\Delta Y_{g,i} := Y_{g,i,2} - Y_{g,i,1}$, $\tilde{\delta}_g := \delta_{g,2} - \delta_{g,1}$, and $\tilde{\epsilon}_{g,i} := \epsilon_{g,i,2} - \epsilon_{g,i,1}$. The parameter vector of interest within group g is $\boldsymbol{\theta}_g = (\tilde{\delta}_g, \tau_g)^\top$. These parameters are identified via the standard OLS moments for the differenced regression:

$$\mathbb{E}_{F_g} \left[\begin{pmatrix} 1 \\ E_{g,i} \end{pmatrix} (\Delta Y_{g,i} - \tilde{\delta}_g - \tau_g E_{g,i}) \right] = \mathbf{0}_2.$$

This clearly fits the general structure $\mathbb{E}_{F_g}[h(D_{g,i}, \boldsymbol{\theta}_g)] = \mathbf{0}_k$ with linear moment functions, matching (1).

Suppose researchers are interested in the causal effect of a scalar group-level policy W_g (e.g., related to childcare provision or parental leave regulations) on the magnitude of the child penalty, τ_g . A common empirical strategy is to parameterize the child penalty as $\tau_g = \alpha + \beta W_g$ and estimate the parameters by pooling the data and running an OLS regression including individual and group-time fixed effects:

$$\min_{\{\gamma_{g,i}\}, \{\delta_{g,t}\}, \alpha, \beta} \sum_{g,i,t} (Y_{g,i,t} - \gamma_{g,i} - \delta_{g,t} - (\alpha + \beta W_g) E_{g,i} \mathbf{1}\{t=2\})^2.$$

The coefficient $\hat{\beta}$ from this regression is often interpreted as the causal effect of the policy W_g on the child penalty. This OLS estimator is numerically equivalent to a particular instance of the one-step GMM estimator (4). Specifically, it corresponds to applying GMM to the moment conditions derived from the differenced equation, imposing the structure $\boldsymbol{\theta}_g = (0, \alpha)^\top + (1, 0)^\top \lambda_g + (0, \beta)^\top W_g$, and using a GMM weighting matrix A_g proportional to the inverse of the sample design matrix within each group g :

$$A_g = n_g \left(\frac{1}{n_g} \sum_{i=1}^{n_g} \begin{pmatrix} 1 \\ E_{g,i} \end{pmatrix} (1, E_{g,i}) \right)^{-1}.$$

As we demonstrate in Section 3, using such data-dependent weights A_g can lead to inconsistency if the policy W_g affects the distribution of the event indicator $E_{g,i}$. An alternative is the two-stage MD approach (3), which first estimates $\hat{\tau}_g$ for each group and then regresses these estimates on W_g in a second stage.

While presented here in a simple two-period setting with a single event timing, this example naturally extends to dynamic contexts with multiple time periods and variation in event

timing across units, aligning with modern event-study specifications. Our framework accommodates such richer settings, with θ_g representing an entire vector of child penalties at different horizons identified through appropriate moment conditions.

Remark 2.4 (Choice between GMM and MD). [Pekkarinen et al. \(2009\)](#), studying an education reform's effect on intergenerational income mobility, exemplifies the researcher's dilemma in choosing between estimators. The authors initially motivate their analysis with a two-step logic akin to our MD estimator discussion (their equations (1)-(2)), but ultimately implement it using GMM/OLS (their equation (3)).

Remark 2.5 (The role of heterogeneity). While the exposition above assumes a constant τ_g identified by OLS moments, if true effects vary across units, these moments capture an average effect. A policy W_g might then influence this average not only through changes in individual effects $\tau_{g,i}$ but also by shifting the population composition relevant for the average. Section 2.3 discusses these direct and indirect channels within our moment-based framework.

2.2.2 DiD Meets AKM: Job Displacement, Average Wage Premia, and Policy Interactions

Our second example applies the framework to analyze the effects of group-level policies on local labor market structures, drawing inspiration from models in the style of [Abowd et al. \(1999\)](#). We consider a setting where policies implemented at the level of a local labor market (LLM), indexed by g , might influence not only average wage parameters but also patterns of worker sorting.

Consider log-wages $Y_{g,i,t}$ for worker i in LLM g at time $t \in \{1, 2\}$. Workers are employed by firms classified into $J = 2$ types (e.g., small and large), denoted by $j(g, i, t) \in \{1, 2\}$. A standard specification decomposes wages as:

$$Y_{g,i,t} = \gamma_{g,i} + \psi_{g,j(g,i,t)} + \epsilon_{g,i,t}, \quad \mathbb{E}_{F_g}[\epsilon_{g,i,t} | \{j(g, i, t')\}_{t'=1,2}] = 0.$$

where $\gamma_{g,i}$ is a worker-specific effect, and $\psi_{g,j}$ represents the wage premium associated with firm type j in market g . For identification purposes we normalize $\psi_{g,1} = 0$ for each g .

Consider a researcher interested in how a job displacement event alters the firm premia effectively experienced by workers, and whether this alteration varies with policy W_g . Let $T_{g,i}$ be an indicator for worker i in group g being affected by a job displacement occurring between periods $t = 1$ and $t = 2$. The research question centers on how this layoff event, interacts with the firm premium structure in the post-layoff period ($t = 2$), and how policy W_g changes this interaction. We can formalize this analysis using the following DiD-type specification:

$$\psi_{g,j(i,t)} = \mu_{g,i} + \delta_{g,t} + \tau_g T_{g,i} \mathbf{1}\{t = 2\} + \nu_{g,i,t}, \quad \mathbb{E}_{F_g}[\nu_{g,i,t} | T_{g,i}] = 0.$$

Using the same notation as in the previous section we define $\theta_g := (\tilde{\delta}_g, \tau_g)^\top$. The goal is to relate θ_g , in particular the effect of the job displacement event τ_g , to a policy of interest W_g . For instance, in [Daruich et al. \(2023\)](#), the authors are interested in how workers' firm wage premia after a job-displacement event (τ_g) depends on lifting constraints on the employment of temporary contract workers.³

In practice, such an investigation often proceeds in two stages. First, the firm-type premium $\hat{\psi}_{g,2}$ is estimated for each group g from individuals moving between firm types. Subsequently, these group-specific premium estimates inform an individual-level outcome, $\hat{\psi}_{g,j(i,t)}$. This constructed outcome is then used in a panel regression to estimate the policy effect. Specifically, parameters (α, β) along with nuisance parameters $(\{\gamma_{g,i}\}, \{\delta_{g,t}\})$ are estimated via OLS:

$$(\hat{\alpha}, \hat{\beta}, \{\hat{\gamma}_{g,i}\}, \{\hat{\delta}_{g,t}\}) = \underset{\alpha, \beta, \{\mu_{g,i}\}, \{\delta_{g,t}\}}{\operatorname{argmin}} \sum_{g,i,t} (\hat{\psi}_{g,j(i,t)} - \mu_{g,i} - \delta_{g,t} - (\alpha + \beta W_g) T_{g,i} \mathbf{1}\{t = 2\})^2.$$

The coefficient on the interaction term $T_{g,i} \mathbf{1}\{t = 2\}$ effectively models the group-specific layoff impact as $\tau_g = \alpha + \beta W_g$, thereby capturing the policy-dependent differential effect.

This empirical strategy fits within our moment-based framework. The first moment condition, identifying $\psi_{g,2}$ from the wage changes of movers ($\Delta Y_{g,i}$), is standard:

$$\mathbb{E}_{F_g}[(\Delta Y_{g,i} - \psi_{g,2}) \mathbf{1}\{h(i) = (1, 2)\} + (\Delta Y_{g,i} + \psi_{g,2}) \mathbf{1}\{h(i) = (2, 1)\}] = 0.$$

The second set of moment conditions identifies the change in the group-time effect, $\Delta \delta_g := \delta_{g,2} - \delta_{g,1}$, and the parameter τ_g . Differencing the population relationship for $\psi_{g,j(i,t)}$ with respect to time implies that the experienced change in an individual's firm premium, $\Delta \psi_{g,i}^{\text{exp}} := \psi_{g,j(i,2)} - \psi_{g,j(i,1)}$, satisfies the following moment restriction:

$$\mathbb{E}_{F_g} \left[\begin{pmatrix} 1 \\ T_{g,i} \end{pmatrix} (\Delta \psi_{g,i}^{\text{exp}} - \tilde{\delta}_g - \tau_g T_{g,i}) \right] = \mathbf{0}_2.$$

This two-stage procedure—initial estimation of $\psi_{g,2}$ followed by its use in a panel OLS regression—can be represented as a specific one-step GMM estimator (4). The corresponding GMM weighting matrix A_g reflects this sequential estimation logic, featuring a block-diagonal structure. The block corresponding to the $(\tilde{\delta}_g, \tau_g)$ moments would be based on the sample covariance of the regressors, structurally analogous to the GMM weighting matrix derived for the example in Section 2.2.1.

³In other studies, the authors relate the same parameter to unemployment rates ([Schmieder et al., 2023](#)) or variation in labor-market structures across countries ([Bertheau et al., 2023](#))

2.2.3 Supply-Side Effects in Financial Markets

Our framework also informs analyses of policies reshaping local financial markets, such as U.S. interstate banking deregulation (e.g., [Jayaratne and Strahan, 1996](#)). This state-level policy, intended to alter market competition, allows studying effects on local credit supply (e.g., bank interest rates). A key challenge is that observed market outcomes (loan rates, quantities) are equilibrium objects reflecting both supply and demand. Controlling for firm-specific demand shocks is crucial to isolate supply-side policy effects, and our framework allows for this.

Specifically, consider banks in state g classified into three types relevant to deregulation: type 1 (e.g., small, incumbent in-state banks, often a benchmark), type 2 (e.g., large, incumbent in-state banks), and type 3 (e.g., newly entering or expanding out-of-state banks, whose presence is directly influenced by W_g). Let $Y_{g,i,b}$ be the outcome (e.g., interest rate) for firm i from bank type b in state g . Following seminal work by [Khwaja and Mian \(2008\)](#) we model these outcomes as:

$$Y_{g,i,b} = \gamma_{g,i} + \psi_{g,b} + \epsilon_{g,i,b}, \quad \mathbb{E}_{F_g}[\epsilon_{g,i,b}|K_{g,i}] = 0,$$

where $\gamma_{g,i}$ is a firm-specific effect capturing its idiosyncratic credit demand and overall credit-worthiness (constant across bank types for firm i), and $\psi_{g,b}$ represents the average supply-side terms offered by bank type b in state g . $K_{g,i}$ denotes the set of bank types with which firm i interacts. Normalizing $\psi_{g,1} \equiv 0$, the parameters $\psi_{g,2}$ and $\psi_{g,3}$ represent the supply conditions of Type 2 and Type 3 banks relative to Type 1 banks. The state-level supply-side outcome vector is $\theta_g \equiv (\psi_{g,2}, \psi_{g,3})'$. These parameters are identified from firms interacting with multiple bank types, as $\gamma_{g,i}$ is differenced out.⁴

Suppose the goal is to quantify a causal effect of deregulation W_g on the relative bank-type effects, postulating, for example, a linear relationship

$$\psi_{g,b} = \alpha_b + \lambda_g + \beta W_{g,b},$$

where α_b are bank-type fixed effects, λ_g are state-level fixed effects, and $W_{g,b}$ measures the exposure of bank type b to the state-level policy W_g (e.g., $W_g \mathbf{1}\{b = 3\}$). A common empirical strategy then involves substituting this model for $\psi_{g,b}$ directly into the outcome equation for $Y_{g,i,b}$.⁵ The parameters, including β , are estimated by solving the following minimization problem:

$$\min_{\{\gamma_{g,i}\}, \{\alpha_b\}, \{\lambda_g\}, \beta} \sum_{g,i,b} (Y_{g,i,b} - \gamma_{g,i} - \mathbf{1}\{b \neq 1\}(\alpha_b + \lambda_g + \beta W_{g,b}))^2.$$

⁴We assume that in each state there are firms that interact with multiple banks, in particular types 1 and 2, and types 1 and 3. This assumption is analogous to the existence of movers in standard AKM-type models.

⁵Historically, this strategy has not been used to assess the effect of the deregulation in the US due to the lack of access to credit registry data, but similar approaches have been used to investigate the effects of macro-prudential policies in Europe, e.g., [Jiménez et al. \(2017\)](#).

This least squares estimator is an instance of a one-step GMM estimator (4).⁶

2.2.4 Production Functions, TFP, and Trade Policy

Estimating firm productivity and its response to trade policy (e.g., as explored by [Amiti and Konings \(2007\)](#)) provides another application of our framework. Consider an industry g , with firms i observed over time $t \in \{1, 2, 3\}$. The analysis typically unfolds in several stages, beginning with characterizing industry-specific production functions. Let $Y_{g,i,t}$ denote log output, and $L_{g,i,t}, K_{g,i,t}, M_{g,i,t}$ represent log inputs (labor, capital, materials). The production function in industry g is:

$$Y_{g,i,t} = \theta_g^l L_{g,i,t} + \theta_g^k K_{g,i,t} + \theta_g^m M_{g,i,t} + \omega_{g,i,t} + \eta_{g,i,t},$$

where $\omega_{g,i,t}$ is unobserved firm-specific productivity and $\eta_{g,i,t}$ is an idiosyncratic error term. To address the endogeneity arising from the correlation between input choices and $\omega_{g,i,t}$, a possible approach is to model the latter parametrically, for instance, as an AR(1) process. This allows for the identification of input elasticities $(\theta_g^l, \theta_g^k, \theta_g^m)$ by solving a system of g -specific linear GMM moment conditions (see, e.g., [Ackerberg et al. \(2015\)](#) for a relevant discussion). Using the industry-specific elasticities, we define the firm-level Total Factor Productivity (TFP):

$$TFP_{g,i,t} := Y_{g,i,t} - (\theta_g^l L_{g,i,t} + \theta_g^k K_{g,i,t} + \theta_g^m M_{g,i,t}).$$

The TFP is then used to specify the key outcomes of interest that characterize the features of the conditional TFP distribution and become the main objects for policy evaluation:

$$\Delta TFP_{g,i,t} = \delta_{g,t} + \tau_{g,t} \Delta FM_{g,i,t} + \epsilon_{g,i,t}, \quad \mathbb{E}_{F_{g,t}}[\epsilon_{g,i,t} | \Delta FM_{g,i,t}] = 0.$$

This specification in first differences implicitly accounts for any unobserved firm-specific fixed effects in the levels of TFP. $FM_{g,i,t}$ is an indicator for firm i importing intermediate inputs at time t , so $\Delta FM_{g,i,t} := FM_{g,i,t} - FM_{g,i,t-1}$ captures changes in import status. Thus, $\delta_{g,t}$ represents the baseline TFP growth in industry g at time t for firms that do not change their import status ($\Delta FM_{g,i,t} = 0$), while $\tau_{g,t}$ measures the additional TFP growth associated with starting (if $\Delta FM_{g,i,t} = 1$) or stopping (if $\Delta FM_{g,i,t} = -1$) the import of intermediate inputs.

⁶Our abstract model relates θ_g to W_g via $\theta_g = \alpha + \Gamma \lambda_g + BW_g$ with $\Gamma = (1, 1)^\top$ and $B = \text{diag}(\beta, \beta)$. Although $\Gamma^\top B \neq 0$ if $\beta \neq 0$, we can accommodate it by defining $B = \beta P_{\Gamma^\perp}$, where $P_{\Gamma^\perp} = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ projects orthogonal to Γ . Effectively, β is identified from the relationship between the policy difference $(W_{g,3} - W_{g,2})$ and the premium difference $(\psi_{g,3} - \psi_{g,2})$.

The full vector of model-based outcomes for industry g is thus

$$\boldsymbol{\theta}_g := (\theta_g^l, \theta_g^k, \theta_g^m, \delta_{g,2}, \delta_{g,3}, \tau_{g,2}, \tau_{g,3})^\top.$$

This parameter is identified via g -specific linear moment conditions (detailed in Appendix A.2.3). The policy analysis then relates the TFP growth parameters $(\delta_{g,t}, \tau_{g,t})$ to industry-level trade policies $W_{g,t} = (T_{g,t}^{out}, T_{g,t}^{in})^\top$, comprising output and input tariffs:

$$\begin{aligned}\delta_{g,t} &= \lambda_g^\delta + \alpha_t^\delta + \beta_{out}^\delta T_{g,t}^{out} + \beta_{in}^\delta T_{g,t}^{in}, \\ \tau_{g,t} &= \lambda_g^\tau + \alpha_t^\tau + \beta_{out}^\tau T_{g,t}^{out} + \beta_{in}^\tau T_{g,t}^{in}.\end{aligned}$$

This system is a particular representation of our general framework, $\boldsymbol{\theta}_g = \boldsymbol{\alpha} + \Gamma \boldsymbol{\lambda}_g + B W_g$, where W_g here would be a vector collecting $(T_{g,t}^{out}, T_{g,t}^{in})$ across the relevant periods $t \in \{2, 3\}$.

In practice (e.g., [Amiti and Konings, 2007](#)), researchers often estimate input elasticities in a first step, then use the constructed TFP measures in a second-stage OLS regression of $\Delta TFP_{g,i,t}$ on policy variables interacted with $\Delta FM_{g,i,t}$. This entire procedure can be cast as a single one-step GMM estimator (analogous to the example in Section 2.2.2). As established in Section 3, such one-step GMM estimators are susceptible to an endogenous weighting bias if the trade policy $W_{g,t}$ affects firms' decisions regarding the import of intermediate inputs, $FM_{g,i,t}$.⁷

Remark 2.6 (Markups as Model-Based Outcomes). The input elasticities estimated for TFP analysis are also central to firm markup estimation. Following insights from [Loecker and Warzynski \(2012\)](#) we can link markups to a variable input's output elasticity and its observed expenditure share. This relationship defines average group-level markups as solutions to linear population moment conditions, thereby allowing researchers to use our framework to investigate the effects of policy on market competition.⁸

2.3 Composition Effects

As foreshadowed by Remark 2.5, aggregate parameters may reflect not only direct policy effects but also policy-induced shifts in the composition of heterogeneous micro-units. This section uses the child penalty example to elaborate on this compositional challenge, showing how it complicates causal inference.

Consider again the DiD setup for estimating the child penalty (Section 2.2.1), but now

⁷For other empirical examples where differences in TFP are regressed on policy or idiosyncratic variation see [Greenstone et al. \(2010, 2012\)](#).

⁸For instance, [Edmond et al. \(2015\)](#) use a quantitative structural model with endogenously variable markups, calibrated with Taiwanese producer-level data, to investigate how international trade affects market competition.

allow for explicit unit-level heterogeneity in the penalty itself. Suppose the outcome $Y_{g,i,t}$ follows a linear model:

$$Y_{g,i,t} = \gamma_{g,i} + \delta_{g,t} + \tau_{g,i} E_{g,i} \mathbf{1}\{t = 2\} + \epsilon_{g,i,t}, \quad \mathbb{E}_{F_g}[\epsilon_{g,i,t}|E_{g,i}] = 0,$$

where $E_{g,i}$ indicates the first birth event between $t = 1$ and $t = 2$, and $\tau_{g,i}$ is the individual-specific child penalty for unit i in group g . A common aggregate parameter of interest in group g is the average penalty among those who experience the event (i.e., the Average Treatment Effect on the Treated, ATT):

$$\tau_g = \mathbb{E}_{F_g}[\tau_{g,i}|E_{g,i} = 1].$$

This parameter represents the average child penalty for mothers in group g (e.g., municipality) and is identified by population moments discussed in Section 2.2.1.

Consider a two-dimensional group-level policy $W_g = (W_{g,1}, W_{g,2})^\top$, with randomly assigned, yet potentially correlated, components. Suppose $W_{g,1}$ (e.g., local labor market conditions or social norms) influences selection into parenthood but not the individual child penalty $\tau_{g,i}$. Conversely, suppose $W_{g,2}$ (e.g., generosity of family policies like paid leave) directly affects $\tau_{g,i}$ but not selection into parenthood.

Under these conditions, the potential outcome for the aggregate ATT parameter $\tau_g(w)$ for a given policy vector $w = (w_1, w_2)$ is derived from its definition:

$$\tau_g(w) = \mathbb{E}_{F_g(w)}[\tau_{g,i}(w_2)|E_{g,i}(w_1) = 1].$$

This expectation averages the micro-level penalties $\tau_{g,i}(w_2)$ over the subpopulation of parents, defined by the condition $E_{g,i}(w_1) = 1$. Crucially, because $W_{g,1}$ affects selection into motherhood via its influence on $F_g(w)$, the aggregate parameter $\tau_g(w)$ generally depends on $W_{g,1}$ solely through this compositional channel, even though the underlying micro-level penalties $\tau_{g,i}$ only respond to $W_{g,2}$. This implies that the structural relationship for the aggregate parameter is more appropriately modeled as $\tau_g = \alpha + \beta_1 W_{g,1} + \beta_2 W_{g,2}$, where β_1 captures the composition effect and β_2 captures the direct effect of family policy generosity on the average penalty.

This dependence structure has critical implications for estimation strategies. First, consider approaches that misspecify the model by omitting $W_{g,1}$. A researcher, correctly believing that the individual penalty $\tau_{g,i}$ only depends causally on the family policy $W_{g,2}$, might estimate a

single equation using OLS:

$$\min_{\{\gamma_{g,i}\}, \{\delta_{g,t}\}, \alpha, \beta_2} \sum_{g,i,t} (Y_{g,i,t} - \gamma_{g,i} + \delta_{g,t} - (\alpha + \beta_2 W_{g,2}) E_{g,i} \mathbf{1}\{t = 2\})^2.$$

This approach generically produces an inconsistent estimate $\hat{\beta}_2$, suffering from standard omitted variable bias (since $W_{g,1}$ influences τ_g through composition and is likely correlated with $W_{g,2}$). Similarly, implementing a two-stage MD approach by first estimating $\hat{\tau}_g$ within each group and then running the second-stage regression,

$$\min_{\alpha, \beta_2} \sum_g (\hat{\tau}_g - \alpha - \beta_2 W_{g,2})^2,$$

also yields an inconsistent estimate $\hat{\beta}_2$ due to the same omitted variable issue. This occurs despite the fact that the individual-level penalty $\tau_{g,i}$ causally depends only on $W_{g,2}$.

Now consider approaches that correctly specify the model for the aggregate τ_g by including both policy components. One might be tempted to estimate the appropriately specified equation using OLS:

$$\min_{\{\gamma_{g,i}\}, \{\delta_{g,t}\}, \alpha, \beta_1, \beta_2} \sum_{g,i,t} (Y_{g,i,t} - \gamma_{g,i} - \delta_{g,t} - (\alpha + \beta_1 W_{g,1} + \beta_2 W_{g,2}) E_{g,i} \mathbf{1}\{t = 2\})^2.$$

However, as we formally demonstrate in Section 3, this one-step GMM/OLS approach still generally yields inconsistent estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. The reason is the endogenous weighting bias: the implicit GMM weights depend on $W_{g,1}$ (via its effect on the distribution of $E_{g,i}$, which determines the sample moments used in the equivalent GMM formulation), violating the conditions required for consistency.

Finally, consider the two-stage MD approach using the correctly specified second stage: first estimate $\hat{\tau}_g$ (the average penalty for mothers in group g) in the first stage, then run the second-stage regression:

$$\min_{\alpha, \beta_1, \beta_2} \sum_g (\hat{\tau}_g - \alpha - \beta_1 W_{g,1} - \beta_2 W_{g,2})^2.$$

Under certain regularity conditions (including sufficiently large n_g , as discussed in Section 4), this approach yields consistent estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. The first stage consistently estimates the parameter $\tau_g(W_{g,1}, W_{g,2})$ as defined by the population moments. The second stage then correctly relates this parameter to the exogenously assigned policy variables $W_{g,1}$ and $W_{g,2}$. This avoids both the omitted variable bias (present in models omitting $W_{g,1}$) and the endogenous weighting bias (present in the OLS/GMM).

This example underscores that grounding aggregate parameters $\boldsymbol{\theta}_g(w)$ in micro-moments from $F_g(w)$ is critical for their precise interpretation and downstream policy analysis. Our approach clarifies how policy shapes $\boldsymbol{\theta}_g(w)$, revealing composition effects (e.g., the average child penalty τ_g varying with $W_{g,1}$ through selection) crucial for model specification. The example also illuminates distinct pitfalls of common estimators: one-step GMM (or OLS) may obscure these compositional effects and suffer from endogenous weighting biases if policy influences underlying moments. Conversely, while the two-step MD estimation mitigates the weighting bias, its consistency depends on the properties of the first-stage estimates, especially with small groups (Section 4). Subsequent sections analyze GMM and MD, formalizing these concerns.

3 GMM

This section analyzes the consistency properties of the GMM estimator (4) as the number of groups $G \rightarrow \infty$. We consider an asymptotic regime where the number of units per group is infinitely large ($n_g = \infty$), which simplifies the analysis by allowing us to work directly with population moments within each group g .

In this setting, the population moments $H_{1,g} = \mathbb{E}_{F_g}[h_1(D_{g,i})]$ and $H_{2,g} = \mathbb{E}_{F_g}[h_2(D_{g,i})]$ are directly observed, yielding the true group parameter $\boldsymbol{\theta}_g = H_{2,g}^{-1}H_{1,g}$. Therefore, the two-stage MD estimator (3) using the true $\boldsymbol{\theta}_g$ coincides with the infeasible oracle OLS estimator (2), ensuring $\hat{B}^{MD} = B^*$, as noted earlier. However, the one-step GMM estimator, introduces potential complications even when $n_g = \infty$. Substituting population moments into the GMM objective (4), the estimator minimizes a weighted sum of squared deviations from the target model:

$$\min_{\substack{B, \boldsymbol{\alpha}, \{\boldsymbol{\lambda}_g\} \\ \text{s.t. } \Gamma^\top B = 0, \Gamma^\top \boldsymbol{\alpha} = 0}} \sum_{g=1}^G (\boldsymbol{\theta}_g - (\boldsymbol{\alpha} + \Gamma \boldsymbol{\lambda}_g + BW_g))^\top \tilde{A}_g (\boldsymbol{\theta}_g - (\boldsymbol{\alpha} + \Gamma \boldsymbol{\lambda}_g + BW_g)),$$

where the effective weighting matrix is $\tilde{A}_g := H_{2,g}^\top A_g H_{2,g}$.⁹ This objective differs from the oracle OLS target (2), which implicitly uses identity weights ($\tilde{A}_g = I_k$). Crucially, if the underlying moments $H_{2,g}$ depend on the policy W_g (because W_g changes the group-specific distribution F_g), then the effective weights \tilde{A}_g may also depend on W_g , potentially leading to inconsistency.

To formalize the conditions for GMM consistency, we adopt a simple data-generating process characterized by linear potential outcomes with common slopes:

⁹Recall that A_g is the original weighting matrix in (4), which in practice is often defined implicitly by using a particular estimator, e.g., OLS.

Assumption 3.1. (a) Policy assignment W_g is i.i.d. across groups g , independent of the potential outcome generating process $F_g(\cdot)$, with $\mathbb{V}[W_g]$ finite and positive definite. (b) The potential distribution function $F_g(w)$ is i.i.d. across groups g for any given w . (c) The potential outcome function $\theta_g(w)$ is linear in w with a common slope B_0 : $\theta_g(w) = \alpha_g + B_0 w$, where B_0 satisfies the constraint $\Gamma^\top B_0 = 0$.

We now use this assumption to investigate when $\text{plim}_{G \rightarrow \infty} \hat{B}^{GMM} = B_0$. Define the optimal population intercept α^0 and the associated residual ϵ_g^0 relative to the GMM weighting \tilde{A}_g :

$$\begin{aligned}\alpha^0 &:= \underset{\alpha: \Gamma^\top \alpha = 0}{\operatorname{argmin}} \mathbb{E} \left[\min_{\lambda_g} (\alpha_g - \alpha - \Gamma \lambda_g)^\top \tilde{A}_g (\alpha_g - \alpha - \Gamma \lambda_g) \right] \\ \epsilon_g^0 &:= (\alpha_g - \alpha^0) - \Gamma \hat{\lambda}_g^0(\alpha^0),\end{aligned}$$

where $\hat{\lambda}_g^0(\alpha)$ minimizes the inner quadratic form for a given α . The expectation $\mathbb{E}[\cdot]$ is over the joint distribution of $(W_g, F_g(\cdot))$. The residual ϵ_g^0 represents the component of baseline heterogeneity $\alpha_g - \alpha^0$ orthogonal to Γ in the \tilde{A}_g -metric.

GMM consistency hinges on the weighted correlation between this residual heterogeneity and the policy.

Proposition 1. Suppose Assumption 3.1 holds and relevant moments exist. The probability limit $B_{lim} = \text{plim}_{G \rightarrow \infty} \hat{B}^{GMM}$ satisfies $B_{lim} = B_0$ if and only if

$$M_\Gamma \mathbb{Cov}[\tilde{A}_g \epsilon_g^0, W_g] = \mathbf{0}_k, \quad (5)$$

where $M_\Gamma = I_k - \Gamma(\Gamma^\top \Gamma)^{-1} \Gamma^\top$ projects orthogonal to the columns of Γ , and the covariance is over the population distribution of groups.

Proposition 1 formalizes the key issue: GMM is consistent if and only if the policy W_g is uncorrelated with the heterogeneity component ϵ_g^0 in the GMM metric (i.e., weighted by $\tilde{A}_g = H_{2,g}^\top A_g H_{2,g}$). Even under random assignment of W_g (implying $\mathbb{Cov}[\alpha_g, W_g] = 0$), this condition can fail if the weights \tilde{A}_g depend on W_g . This dependence implies that ϵ_g^0 (directly related to \tilde{A}_g because of the fixed effects λ_g) can be correlated with W_g . It is further exacerbated by the explicit presence of \tilde{A}_g in the covariance.

Remark 3.1 (MD vs. GMM Consistency). The oracle MD estimator, which uses θ_g directly, implicitly relies on identity weights ($\tilde{A}_g = I_k$). Under Assumption 3.1(a), the consistency condition (5) holds because $\mathbb{Cov}[\epsilon_g^0, W_g] = 0$. Thus, the MD estimator is consistent for B_0 under Assumption 3.1. More generally, two-stage MD estimators using fixed second-stage weights (independent of W_g) are also consistent under these conditions. For instance, this allows for MD estimators that assign different deterministic weights to the groups, e.g., reflecting

their size, or more broadly for MD estimators that reweight the objective using deterministic weighting matrices to improve efficiency or protect from misspecification.

Remark 3.2 (Theory vs. Practice). While MD and GMM estimators are theoretically interchangeable via appropriate weighting matrices, the distinction is crucial in practice. Applied researchers routinely use default weights (e.g., implied by OLS, as in preceding examples) rather than specifying them explicitly. Moreover, selecting an appropriate GMM weighting matrix to guarantee consistency is non-trivial; e.g., standard optimal weights (derived from the inverse moment variance) depend on the policy W_g , exacerbating the very biases we focus on.

3.1 Revisting examples

The GMM inconsistency identified in Proposition 1 arises when policy W_g influences implicit GMM weights \tilde{A}_g through policy-induced changes in the population moments $H_{2,g}$. We now revisit the examples from Section 2.2 to illustrate why this mechanism can appear in common empirical settings where one-step GMM (or equivalent OLS) is applied.

3.1.1 Child Penalty Analysis

In the child penalty DiD setup, the model-based outcome of interest is τ_g , representing the effect of motherhood ($E_{g,i} = 1$) on labor market outcomes. Following Assumption 3.1 we model the relationship between τ_g and policy of interest W_g (e.g., childcare provision) using a linear model:

$$\tau_g = \alpha_g + \beta_0 W_g,$$

where β_0 is the true policy effect and α_g is unobserved group-specific heterogeneity in the baseline penalty (e.g., due to unrelated differences in labor-market and parenthood-related policies across municipalities). The one-step GMM estimator (equivalent to pooled OLS on the differenced equation $\Delta Y_{g,i} = \tilde{\delta}_g + \tau_g E_{g,i}$, with group fixed effects for $\tilde{\delta}_g$) involves effective weights equal to

$$\tilde{A}_g = \mathbb{E}_{F_g} \left[\begin{pmatrix} 1 \\ E_{g,i} \end{pmatrix} \begin{pmatrix} 1 & E_{g,i} \end{pmatrix} \right]$$

If the policy W_g affects the propensity to have a first child ($\mathbb{P}_{F_g}[E_{g,i} = 1]$), then \tilde{A}_g depends on W_g .

Specifically, the implicit weight in estimating β_0 from the τ_g component is proportional to $\sigma_{E,g}^2 := \mathbb{V}_{F_g}[E_{g,i}]$. As long as $\sigma_{E,g}^2$ depends on the policy, i.e., $\sigma_{E,g}^2 = \sigma_{E,g}^2(W_g)$, it induces a correlation between W_g and the heterogeneity term α_g (even if W_g is randomly assigned with

respect to α_g). The resulting asymptotic bias for $\hat{\beta}^{GMM}$ is:

$$\text{plim}_{G \rightarrow \infty} (\hat{\beta}^{GMM} - \beta_0) = \frac{\text{Cov}^{\sigma_E^2(W_g)}[W_g, \alpha_g]}{\mathbb{V}^{\sigma_E^2(W_g)}[W_g]},$$

where the covariance and variance use $\sigma_{E,g}^2(W_g)$ as weights.

3.1.2 Job Displacement and Wage Premia Analysis

In the wage premia example, the outcome of interest is τ_g , representing the effect of a job displacement event ($T_{g,i} = 1$) on experienced firm-type wage premia. Consistent with Assumption 3.1, we model this as:

$$\tau_g = \alpha_g + \beta_0 W_g,$$

where β_0 is the policy effect and α_g is unobserved group-specific heterogeneity in the baseline displacement impact (e.g., due to differences in local-labor-market conditions across municipalities and sectors). The one-step GMM estimator implicitly estimates τ_g from moments involving $\Delta\psi_{g,i}^{\text{exp}}$ and regressors $(1, T_{g,i})^\top$. As in the previous case, the effective weights are thus

$$\tilde{A}_g = \mathbb{E}_{F_g} \left[\begin{pmatrix} 1 \\ T_{g,i} \end{pmatrix} \begin{pmatrix} 1 & T_{g,i} \end{pmatrix} \right].$$

If the policy W_g (e.g., unemployment benefits) influences the composition of displaced workers (i.e., affects the distribution F_g such that $\mathbb{P}_{F_g}[T_{g,i} = 1]$ depends on W_g), then \tilde{A}_g becomes $\tilde{A}_g(W_g)$.

As in the previous example, the weight in estimating β_0 from the τ_g component is proportional to $\sigma_{T,g}^2 := \mathbb{V}_{F_g}[T_{g,i}]$. Under this weighting W_g and α_g are generically correlated leading to bias:

$$\text{plim}_{G \rightarrow \infty} (\hat{\beta}^{GMM} - \beta_0) = \frac{\text{Cov}^{\sigma_T^2(W_g)}[W_g, \alpha_g]}{\mathbb{V}^{\sigma_T^2(W_g)}[W_g]},$$

where the covariance and variance are using $\sigma_{T,g}^2 = \sigma_{T,g}^2(W_g)$ as weights.

Remark 3.3 (Distinct bias channels). The GMM weighting bias for $\hat{\beta}_0^{GMM}$ is driven only by W_g 's impact on the distribution of job displacement $T_{g,i}$. Notably, the bias does not feature any effects the policy might have on worker mobility patterns that identify the firm-type premia $\psi_{g,2}$ —a common empirical possibility. This apparent discrepancy is because the two-step procedure effectively uses the MD approach to construct $\psi_{g,2}$, and that particular step does not interact with the estimation of β_0 . A different GMM specification would feature the additional bias arising from the mobility decisions.

3.1.3 Supply-Side Analysis

In the banking deregulation example, the policy W_g influences bank-specific supply conditions $\psi_{g,b}$. Consistent with Assumption 3.1, we model this relationship linearly:

$$\psi_{g,b} = u_{g,b} + \beta_0 W_{g,b} \quad \text{for } b \in \{2, 3\},$$

where $u_{g,b}$ denotes unobserved group-specific heterogeneity in baseline supply conditions independent of the policy $W_{g,b}$, and β_0 is the effect of interest.

The one-step GMM/OLS estimator for β_0 outlined in Section 2.2.3, implicitly uses the micro-data structure of firm-bank relationships to determine $\psi_{g,b}$. In particular, it relies on firms with specific multi-bank connections. For $b \in \{2, 3\}$ we define the population shares of such firms:

$$p_{g,b} := \mathbb{P}_{F_g} [\text{firm } i \text{ is a client of type 1 and type } b \text{ banks}].$$

These shares determine the effective weighting matrix:¹⁰

$$\tilde{A}_g = \text{diag}\{p_{g,2}, p_{g,3}\}.$$

The resulting estimation problem for β_0 (in population) then corresponds to a $p_{g,b}$ -weighted regression with group and bank fixed effects:

$$\min_{\{\alpha_b\}, \beta} \mathbb{E} \left[\min_{\lambda_g} \sum_{b=2,3} p_{g,b} (\psi_{g,b} - \lambda_g - \alpha_b - \beta W_{g,b})^2 \right].$$

To estimate β , this two-way models combines the shares $p_{g,2}$ and $p_{g,3}$ together in a single weight,

$$\nu_g := \frac{p_{g,2}p_{g,3}}{(p_{g,2} + p_{g,3})^2}.$$

If the deregulation affects the matching process between firms and banks, then shares $p_{g,b}$ and thus derived weights ν_g , depend on W_g , and we can write $\nu_g = \nu_g(W_g)$. This relationship in turn induces bias in $\hat{\beta}^{GMM}$, which has the following form:

$$\text{plim}_{G \rightarrow \infty} (\hat{\beta}^{GMM} - \beta_0) = \frac{\mathbb{Cov}^{\nu_g(W_g)}[\Delta W_g, \Delta u_g]}{\mathbb{V}^{\nu_g(W_g)}[\Delta W_g]},$$

where $\Delta u_g := u_{g,3} - u_{g,2}$ and $\Delta W_g := W_{g,3} - W_{g,2}$, and we use $\nu_g(W_g)$ as weights.

¹⁰In deriving this matrix, we assume that no firms interact with banks of all three types, or only with banks of types 2 and 3. We impose this assumption only to simplify the calculations of the bias.

3.2 Discussion

Proposition 1 reveals a distinct source of inconsistency in one-step GMM estimators: the endogeneity of the implicit weighting matrix $\tilde{A}_g = H_{2,g}^\top A_g H_{2,g}$. This subsection puts this result in context, discussing its implications and relationship to existing literature.

First, the linear potential outcome model in Assumption 3.1(c), $\theta_g(w) = \alpha_g + B_0 w$, is intentionally simple. Its purpose is not realism but transparency: it allows us to isolate the bias stemming purely from the GMM weighting mechanism, separate from any other complexities. Introducing non-linearities in the response to W_g or allowing the policy effect B_0 to be heterogeneous across groups ($B_{0,g}$) would only exacerbate estimation challenges and potentially introduce additional sources of bias for GMM estimators. The weighting bias we focus on persists even in this basic linear, constant-effects setting.

Second, the GMM weighting bias we identify is distinct from issues arising in the recent DiD literature with heterogeneous treatment effects and staggered adoption (e.g., [Borusyak et al., 2024b](#); [Callaway and Sant'Anna, 2021](#); [de Chaisemartin and D'Haultfœuille, 2020](#); [Goodman-Bacon, 2021](#), [Sun and Abraham, 2021](#)), and from related concerns in linear regressions with multiple regressors ([Goldsmith-Pinkham et al., 2024](#)). The latter biases arise when moment conditions predicated on coefficient homogeneity are applied in contexts of effect heterogeneity, potentially leading to “forbidden comparisons” and “contamination biases”. In contrast, the inconsistency we identify in Proposition 1 arises from the correlation between policy-dependent weights (\tilde{A}_g) and the policy itself. Critically, unlike the aforementioned DiD-related issues, this weighting bias can manifest even in models with constant treatment effects and a single policy adoption time.

4 Minimum Distance Estimation

We now turn to the analysis of the Minimum Distance (MD) estimator (3). This two-stage procedure first obtains group-specific parameter estimates, $\hat{\theta}_g$, and then uses them in a second-stage regression, mimicking the oracle objective (2). We focus on practical settings where the number of units per group, n_g , is finite, making the first-stage estimates $\hat{\theta}_g$ inherently noisy. We demonstrate that even under the favorable assumption that $\hat{\theta}_g$ are conditionally unbiased, the reliability of the resulting MD policy estimator, \hat{B}^{MD} , critically depends on n_g .

A key concern is that for “small” group sizes (e.g., $n_g = 2$) the standard MD estimator is generically inconsistent due to a policy-induced sample selection problem (Corollary 1). However, this inconsistency can be avoided. Proposition 2 quantifies the impact of first-stage estimation noise, clarifying what we mean by a “sufficiently large” n_g , and then Proposition 3 demonstrates that in such regimes the MD estimator is statistically equivalent to the oracle one,

ensuring standard second-stage inference is valid. Furthermore, for settings with very small group sizes, we show that consistency can be restored by incorporating auxiliary information, analogous to approaches in the design-based identification literature.

4.1 Selection Problem

Suppose the group-specific parameters θ_g are identified via conditional moment restrictions:

$$\mathbb{E}_{F_g}[h_1(D_{g,i}) - h_2(\tilde{D}_{g,i})\theta_g | \tilde{D}_{g,i}] = \mathbf{0}_k.$$

where $\tilde{D}_{g,i}$ is a part of $D_{g,i}$. Several examples in Section (2.2) have this property, and it is, in general, common in regression models. A natural first-stage estimator for θ_g uses the sample analog within group g :

$$\hat{\theta}_g := \left(\sum_{i=1}^{n_g} h_2(\tilde{D}_{g,i}) \right)^{-1} \sum_{i=1}^{n_g} h_1(D_{g,i}) =: (\hat{H}_{2,g})^{-1}(\hat{H}_{1,g}).$$

This estimator requires the sample matrix $\hat{H}_{2,g}$ to be invertible. Let $\omega_g := \mathbf{1}\{\hat{H}_{2,g}\text{ is invertible}\}$ indicate whether $\hat{\theta}_g$ is well-defined for group g .

The conditional moment restriction implies

$$\mathbb{E}_{F_g}[\hat{H}_{1,g} | \{\tilde{D}_{g,i}\}_{i=1}^{n_g}] = \mathbb{E}_{F_g}[\hat{H}_{2,g} | \{\tilde{D}_{g,i}\}_{i=1}^{n_g}] \theta_g.$$

Therefore, conditional on the covariates $\{\tilde{D}_{g,i}\}$ and on the estimator being well-defined, $\hat{\theta}_g$ is unbiased for the true group parameter θ_g :

$$\mathbb{E}_{F_g}[\hat{\theta}_g | \omega_g = 1, \{\tilde{D}_{g,i}\}_{i=1}^{n_g}] = \theta_g.$$

When it exists, we can decompose the estimator into the true parameter and a conditionally mean-zero estimation error:

$$\hat{\theta}_g = \theta_g + \varepsilon_g, \quad \text{where } \mathbb{E}_{F_g}[\varepsilon_g | \omega_g = 1, \{\tilde{D}_{g,i}\}_{i=1}^{n_g}] = \mathbf{0}_k.$$

The feasible MD estimator (3) uses these first-stage estimates $\hat{\theta}_g$. Since $\hat{\theta}_g$ is only computed for groups with $\omega_g = 1$, the estimator effectively solves:

$$(\hat{B}^{MD}, \hat{\alpha}^{MD}, \{\hat{\lambda}_g^{MD}\}) := \underset{\substack{B, \alpha, \{\lambda_g\} \\ \text{s.t. } \Gamma^\top B = 0, \Gamma^\top \alpha = 0}}{\operatorname{argmin}} \sum_{g=1}^G \left\| \hat{\theta}_g - (\alpha + \Gamma \lambda_g + BW_g) \right\|_2^2 \omega_g. \quad (6)$$

Although the estimation error ε_g has conditional mean zero, the consistency of \hat{B}^{MD} is not guaranteed when n_g is fixed. The core issue is that the selection indicator ω_g depends on the specific sample $\{\tilde{D}_{g,i}\}_{i=1}^{n_g}$ drawn from F_g . Since the policy W_g affects the distribution F_g , it can influence $\mathbb{P}_{F_g}[\omega_g = 1]$. If this probability varies with W_g , restricting the second-stage regression to groups where $\omega_g = 1$ induces a sample selection bias, analogous to the GMM weighting bias discussed in Section 3.

Substituting $\hat{\theta}_g = \theta_g + \varepsilon_g$ into the objective (6), the resulting \hat{B}^{MD} decomposes additively due to the linearity of least squares: $\hat{B}^{MD} = \hat{B}_0^{MD} + \hat{B}_1^{MD}$. The first component, \hat{B}_0^{MD} , arises from using the estimation error ε_g as the outcome. By construction this error is orthogonal to W_g conditional on $\omega_g = 1$ and thus under standard technical conditions this component converges to zero as $G \rightarrow \infty$: $\text{plim}_{G \rightarrow \infty} \hat{B}_0^{MD} = \mathbf{0}$.

The potential inconsistency stems entirely from the second component, \hat{B}_1^{MD} , which arises from using the true parameter θ_g , but only for the selected sample determined by ω_g :

$$(\hat{B}_1^{MD}, \hat{\alpha}_1^{MD}, \{\hat{\lambda}_{1,g}^{MD}\}) := \underset{\substack{B, \alpha, \{\lambda_g\} \\ \text{s.t. } \Gamma^\top B = 0, \Gamma^\top \alpha = 0}}{\operatorname{argmin}} \sum_{g=1}^G \|\theta_g - (\alpha + \Gamma \lambda_g + BW_g)\|_2^2 \omega_g.$$

This objective function is precisely the population GMM objective analyzed in Section 3 with an effective weighting matrix $\tilde{A}_g = \omega_g I_k$. Crucially, because n_g is fixed, the selection indicator ω_g remains random even as $G \rightarrow \infty$, as it depends on the finite sample draw within group g . Consequently, the consistency condition derived for GMM in Proposition 1 applies directly to \hat{B}_1^{MD} and thus determines the consistency of the feasible MD estimator \hat{B}^{MD} .

Corollary 1 (Consistency of MD with Fixed n_g). *Suppose Assumption 3.1 holds and the number of units per group n_g is fixed. Define the population intercept adjusted for selection probability:*

$$\alpha^0 := \underset{\alpha: \Gamma^\top \alpha = 0}{\operatorname{argmin}} \mathbb{E} \left[\omega_g \min_{\lambda_g} \|\alpha_g - \alpha - \Gamma \lambda_g\|_2^2 \right]$$

and the associated residual heterogeneity $\epsilon_g^0 := (\alpha_g - \alpha^0) - \Gamma \hat{\lambda}_g^0(\alpha^0)$. The feasible MD estimator \hat{B}^{MD} defined in (6) converges in probability to B_0 as $G \rightarrow \infty$ if and only if

$$M_\Gamma \mathbb{C}\text{ov}[\omega_g \epsilon_g^0, W_g] = \mathbf{0}_k. \quad (7)$$

Corollary 1 reveals that MD consistency with fixed n_g requires a potentially strong condition. Even if policy W_g is randomly assigned such that it is independent of baseline heterogeneity α_g (implying $\mathbb{C}\text{ov}[\alpha_g, W_g] = 0$), consistency can fail if W_g influences the probability of selection $\mathbb{P}_{F_g}[\omega_g = 1]$. For example, if policy affects the distribution of covariates $\tilde{D}_{g,i}$ such that $\hat{H}_{2,g}$ is more likely to be invertible for certain values of W_g , the condition (7) fails. This

contrasts with the large- n_g case where $\omega_g \rightarrow 1$, $\tilde{A}_g \rightarrow I_k$, and consistency holds under random assignment.

This inconsistency of the fixed- n_g MD estimator due to sample selection bears resemblance to biases identified in the GMM literature, though the mechanism differs. [Altonji and Segal \(1996\)](#) highlighted finite-sample bias in optimally weighted GMM arising from spurious correlation between estimated optimal weights and sample moments. Subsequent work (e.g., [Newey and Smith, 2004](#)) analyzed such biases, particularly in settings with many or weak moments, and proposed corrections. In our MD context, the selection indicator ω_g acts as a crude, sample-dependent weight (1 for included groups, 0 for excluded). The bias arises because this "estimated" inclusion probability can be systematically correlated with the policy regressor W_g through the underlying data distribution $F_g(W_g)$. Unlike the higher-order bias analyzed in the optimal GMM literature, this MD bias persists asymptotically as $G \rightarrow \infty$ because ω_g remains random due to the fixed group size n_g .

4.2 Inference

Section 4.1 and Corollary 1 established that the Minimum Distance (MD) estimator \hat{B}^{MD} is generally inconsistent when group sizes n_g are fixed, due to a sample selection bias. This section analyzes the behavior of \hat{B}^{MD} when group sizes n_g grow with the number of groups G . We first quantify the selection bias (Section 4.2.1) and derive conditions under which it becomes asymptotically negligible. We show that under these same conditions, the first-stage estimation error also vanishes asymptotically, leading to the asymptotic equivalence of the feasible MD and infeasible oracle estimators and validating standard inference procedures (Section 4.2.2).

4.2.1 Bias Bound

While Corollary 1 establishes the potential inconsistency of the MD estimator with fixed group sizes n_g , its practical relevance hinges on the bias magnitude. We now quantify the deviation $\Delta B = \hat{B}_1^{MD} - B^*$ from the infeasible oracle estimator B^* .

Proposition 2 (Bias Bound for MD Estimator Component). *Let $(\hat{B}_1^{MD}, \hat{\alpha}_1^{MD}, \{\hat{\lambda}_{1,g}^{MD}\})$ be the solution using θ_g as the outcome in the weighted regression (6). Define $\Delta B = \hat{B}_1^{MD} - B^*$ and $\Delta \alpha = \hat{\alpha}_1^{MD} - \alpha^*$. Define the matrix $M =: \frac{1}{G} \sum_{g=1}^G \omega_g (1, W_g)^\top (1, W_g)$. Then,*

$$\|(\Delta \alpha, \Delta B)\|_F \leq \frac{\sqrt{1 + \max_g \|W_g\|_2^2}}{\lambda_{\min}(M)} \left(\max_g \|\mathbf{r}_g^*\|_2 \right) \frac{\|\omega - 1\|_1}{G}$$

where $\mathbf{r}_g^* := \theta_g - \alpha^* - \Gamma \lambda_g^* - B^* W_g$ is the oracle residual for group g , $\|\cdot\|_F$ denotes the Frobenius

norm, $\|\cdot\|_2$ the Euclidean norm, and $\|\omega - 1\|_1 = \sum_{g=1}^G (1 - \omega_g)$ is the number of groups excluded due to $\omega_g = 0$.

Proposition 2 shows that the magnitude of the bias effectively depends on the fraction of excluded groups, $\|\omega - 1\|_1/G$. In particular, the bias term is asymptotically negligible relative to the standard $G^{-1/2}$ estimation error if $\|\omega - 1\|_1 \ll \sqrt{G}$ (potentially up to additional factors depending on maximal residuals $\max_g \|\mathbf{r}_g^*\|_2$). This condition connects bias to the rate at which group sizes n_g grow relative to G . Assuming the probability of exclusion decays sufficiently fast, e.g., exponentially $\mathbb{P}_{F_g}[\omega_g = 0] \lesssim e^{-cn_g}$ (as often implied by standard concentration results), then n_g growing slightly faster than $\log G$ is sufficient to ensure $\|\omega - 1\|_1 \ll \sqrt{G}$.

Importantly, the condition requiring n_g to grow sufficiently fast to eliminate selection bias also ensures that the noise from the first-stage estimation of $\hat{\theta}_g$ becomes asymptotically negligible under \sqrt{G} scaling. Recall that the feasible estimator decomposes as $\hat{B}^{MD} = \hat{B}_1^{MD} + \hat{B}_0^{MD}$, where \hat{B}_0^{MD} captures the influence of the first-stage estimation errors $\varepsilon_g = \hat{\theta}_g - \theta_g$. The variance of this first-stage error is typically $O(n_g^{-1})$. Consequently, the contribution of these errors to the asymptotic variance of $\sqrt{G}(\hat{B}^{MD} - B_0)$ depends on the average precision, scaling with $\lim_{G \rightarrow \infty} \frac{1}{G} \sum_{g=1}^G O(n_g^{-1})$. Any growth rate of n_g sufficient to make the selection bias negligible ensures this average inverse group size vanishes, implying $\sqrt{G}\hat{B}_0^{MD} \xrightarrow{p} 0$.

Remark 4.1 (Bias in practice). Applied researchers can use the finite sample bound in Proposition 2 to quantify the importance of the bias. Its only unobservable component is the maximum oracle residual, $\max_g \|\mathbf{r}_g^*\|_2$, which might be approximated using the empirical analog from the feasible MD estimation. Researchers can then compare the resulting estimated bound to the standard error of the feasible MD estimator to gauge the relative importance of the bias.

4.2.2 Asymptotic Equivalence

The combined implication of the previous results is that when group sizes n_g grow sufficiently fast with G , both the sample selection bias and the first-stage estimation noise become irrelevant for the asymptotic distribution of \hat{B}^{MD} under standard \sqrt{G} scaling. The feasible MD estimator behaves asymptotically like the infeasible oracle estimator B^* that uses the true parameters θ_g . This leads to the following result:

Proposition 3 (Asymptotic Distribution of MD with Growing n_g). *Let $B_0 = \text{plim}_{G \rightarrow \infty} B^*$. Assume the oracle estimator is asymptotically normal: $\sqrt{G}(B^* - B_0) \xrightarrow{d} N(0, V_{B^*})$ for some asymptotic variance matrix V_{B^*} . Assume $\mathbb{P}_{F_g}[\omega_g = 0] \lesssim \exp(-cn_g)$ and $\sum_{g=1}^G \exp(-cn_g) \ll \sqrt{G}$, potentially up to log factors. Assume that conditional on $\{\omega_h, \{\tilde{D}_{h,i}\}, W_h\}_{h=1}^G$, the error terms ε_g are independent across g . Assume the variance $\|n_g \mathbb{E}[\varepsilon_g \varepsilon_g^\top | \omega_g = 1, \{\tilde{D}_{g,i}\}, W_g]\|_F$ is uniformly*

bounded. Finally, suppose that $\frac{1}{G} \sum_{g=1}^G W_g$ and $\frac{1}{G} \sum_{g=1}^G W_g W_g^\top$ converge in probability to well-behaved deterministic limits. Then the feasible MD estimator \hat{B}^{MD} defined in (6) satisfies:

$$\sqrt{G}(\hat{B}^{MD} - B_0) \xrightarrow{d} N(0, V_{B^*}).$$

Proposition 3 establishes the asymptotic equivalence of the feasible MD estimator \hat{B}^{MD} and the infeasible oracle estimator B^* when group sizes n_g grow sufficiently fast. They share the same probability limit B_0 and the same limiting distribution, characterized by the asymptotic variance V_{B^*} .

This equivalence has crucial implications for estimating the asymptotic variance V_{B^*} required for constructing confidence intervals and conducting hypothesis tests using \hat{B}^{MD} . The result suggests that standard approaches to variance estimation used in the second stage remain valid, provided they would be appropriate for the oracle estimator.

Consider, for instance, the commonly used Eicker-Huber-White (EHW) robust variance estimator. Its validity for the oracle regression depends on underlying assumptions about the data-generating process (e.g., independence or specific dependence structures across groups g). If these assumptions hold, and the EHW estimator applied to the infeasible oracle regression consistently estimates V_{B^*} , then Proposition 3 implies that the feasible EHW estimator, computed using the feasible second-stage residuals $\hat{r}_g = \hat{\theta}_g - (\hat{\alpha}^{MD} + \Gamma \hat{\lambda}_g^{MD} + \hat{B}^{MD} W_g)$, also consistently estimates V_{B^*} .

The first-stage estimation noise embedded within the feasible residuals does not distort the asymptotic properties of such standard second-stage variance estimators, provided n_g grows sufficiently fast. The contribution of this noise is asymptotically negligible for variance estimation, just as it is for the point estimate. Therefore, researchers can employ robust variance estimation techniques developed for linear models (like EHW or potentially cluster-robust variants) on the second-stage MD regression, contingent only on the appropriateness of the chosen technique for the underlying structure of the oracle problem itself. No explicit correction for the first-stage estimation is needed asymptotically under these conditions.

Remark 4.2 (No unbiased estimation). In models which do not have the conditional structure we discuss in this section (e.g., those we discuss in Section 2.2.4), $\hat{\theta}_g$ is generically biased for any finite n_g . This first-stage estimation bias, rather than sample selection, then becomes the primary obstacle to the consistency of the MD estimation. Mitigating this generally requires constructing bias-corrected estimators (e.g., using jackknife) and stronger requirements on the growth of n_g than needed in Proposition 3.

4.3 Restoring Consistency

The potential inconsistency of the MD estimator with fixed n_g , identified in Corollary 1, stems from the sample-based selection criterion $\omega_g = \mathbf{1}\{\hat{H}_{2,g}\text{ is invertible}\}$. This issue can be circumvented if external information about the population moment structure is available.

Suppose the population matrix $H_{2,g} := \mathbb{E}_{F_g}[h_2(\tilde{D}_{g,i})]$ is known for each group g . Such knowledge can originate from various sources. First, in settings with explicit research designs, such as experiments or quasi-experiments, parts of $H_{2,g}$ (like treatment assignment probabilities) might be known directly or can be computed. Second, $H_{2,g}$ could be accurately estimated from richer auxiliary datasets (e.g., administrative records or large-scale surveys) that detail the population distribution of covariates $\tilde{D}_{g,i}$, even if these datasets lack the primary outcome variables. Third, one might specify and potentially estimate a structural model for the conditional distribution of $\tilde{D}_{g,i}$ given the policy W_g , from which $H_{2,g}$ can then be derived as a model output. Crucially, regardless of its origin, knowing $H_{2,g}$ allows constructing an alternative first-stage estimator that bypasses the sample invertibility problem:

$$\hat{\boldsymbol{\theta}}_g^{alt} := H_{2,g}^{-1} \left(\frac{1}{n_g} \sum_{i=1}^{n_g} h_1(D_{g,i}) \right) = H_{2,g}^{-1} \hat{H}_{1,g}.$$

This estimator uses the population matrix $H_{2,g}$ in place of its potentially singular sample counterpart $\hat{H}_{2,g}$. Since $H_{2,g}$ is assumed known and invertible for all g , $\hat{\boldsymbol{\theta}}_g^{alt}$ is always well-defined, regardless of the specific sample drawn within group g . Furthermore, this estimator is unconditionally unbiased for $\boldsymbol{\theta}_g$:

$$\mathbb{E}_{F_g}[\hat{\boldsymbol{\theta}}_g^{alt}] = H_{2,g}^{-1} \mathbb{E}_{F_g}[\hat{H}_{1,g}] = H_{2,g}^{-1} H_{1,g} = \boldsymbol{\theta}_g.$$

We can write $\hat{\boldsymbol{\theta}}_g^{alt} = \boldsymbol{\theta}_g + \boldsymbol{\varepsilon}_g^{alt}$, where the estimation error $\boldsymbol{\varepsilon}_g^{alt} = H_{2,g}^{-1}(\hat{H}_{1,g} - H_{1,g})$ has unconditional mean zero.

Consider the MD estimator constructed using these design-based first-stage estimates:

$$(\hat{B}^{MD,alt}, \hat{\boldsymbol{\alpha}}^{MD,alt}, \{\hat{\boldsymbol{\lambda}}_g^{MD,alt}\}) := \underset{\substack{B, \boldsymbol{\alpha}, \{\boldsymbol{\lambda}_g\} \\ \text{s.t. } \Gamma^\top B = 0, \Gamma^\top \boldsymbol{\alpha} = 0}}{\operatorname{argmin}} \sum_{g=1}^G \left\| \hat{\boldsymbol{\theta}}_g^{alt} - (\boldsymbol{\alpha} + \Gamma \boldsymbol{\lambda}_g + BW_g) \right\|_2^2.$$

Since $\hat{\boldsymbol{\theta}}_g^{alt}$ is defined for all g and is unconditionally unbiased, the selection problem related to ω_g is eliminated. Following the logic used previously, the noise component of the estimator converges to zero ($\text{plim}_{G \rightarrow \infty} \hat{B}_0^{MD,alt} = 0$), ensuring that $\hat{B}^{MD,alt}$ converges in probability to B^* (or B_0 under Assumption 3.1) as $G \rightarrow \infty$, even with fixed n_g . The asymptotic variance of $\hat{B}^{MD,alt}$ is typically larger than that of B^* , with the estimation error contributing to the

asymptotic distribution, unlike in Proposition 3.

The principle of using known population moments $H_{2,g}$ to restore MD consistency by substituting population structure for sample-based information is analogous to Inverse Probability Weighting (IPW) and broader design-based methods for causal inference (e.g., [Arkhangelsky et al. \(2024\)](#); [Aronow and Samii \(2017\)](#); [Borusyak and Hull \(2023\)](#); [Borusyak et al. \(2024a\)](#)). These strategies leverage known population quantities—e.g., a known propensity score—instead of the in-sample analogs. The use of auxiliary aggregate information ($H_{2,g}$) also echoes the work by [Imbens and Lancaster \(1994\)](#), who showed that incorporating known population moments can enhance the efficiency of maximum likelihood estimation (MLE) based on micro survey data, and sometimes aid identification. While the context differs (MLE efficiency vs. MD consistency), the underlying principle is similar: leveraging external information on the population distribution can overcome limitations inherent in finite micro-samples. Here, the benefit is starker, shifting the estimator from being generally inconsistent (with fixed n_g) to consistent.

Remark 4.3 (Sampling Design). When auxiliary information is unavailable, specialized within-group sampling might ensure $\hat{H}_{2,g}$ invertibility, avoiding selection. However, if the sampling protocol depends on the data, then n_g itself becomes endogenous. Using such endogenous n_g for standard second-stage precision weighting could thus induce bias.

Illustrative Examples: The Role of Auxiliary Information

We now illustrate how the use of auxiliary information for $H_{2,g}$ can address estimation challenges in our motivating examples, particularly when group sizes are small.

DiD example Consider the DiD model from Section 2.2.1, where group-specific parameters $\theta_g = (\tilde{\delta}_g, \tau_g)^\top$ are estimated from moments involving a binary event indicator $E_{g,i}$ (e.g., first childbirth). The standard first-stage estimator fails if all units within a group g have the same value of $E_{g,i}$ (no within-group variation), rendering the corresponding $\hat{H}_{2,g}$ singular. If, however, the population matrix $H_{2,g}$ is known and invertible—specifically, if the population probability $\pi_g = \mathbb{E}_{F_g}[E_{g,i}]$ is known and $0 < \pi_g < 1$ —one can construct $\hat{\theta}_g^{alt} = H_{2,g}^{-1}\hat{H}_{1,g}$ even for groups with no variation in $\tilde{D}_{g,i}$. This estimator takes a familiar IPW form:

$$\hat{\tau}_g^{alt} = \frac{1}{n_g} \sum_{i=1}^{n_g} \frac{(E_{g,i} - \pi_g)}{\pi_g(1 - \pi_g)} \Delta Y_{g,i}.$$

Why does this approach work even if a specific group g lacks in-sample variation in $E_{g,i}$? For any single group $\hat{\theta}_g^{alt}$ is a noisy estimator of θ_g but it is unconditional unbiased: $\mathbb{E}_{F_g}[\hat{\theta}_g^{alt}] =$

θ_g .¹¹ The second-stage MD estimation averages these noisy but unbiased estimates across many groups g . As the number of groups G grows, the law of large numbers ensures that the estimation noise averages out, and the unconditional unbiasedness of the first-stage estimates (for each g) prevents systematic error in the second stage. This allows the MD estimator $\hat{B}^{MD,alt}$ to consistently estimate the policy effect B_0 , bypassing the sample selection bias.

AKM and Banking Examples The same principle extends to more complex settings. In the AKM example from Section 2.2.2, estimating firm wage premia $\psi_{g,2}$ and the displacement effect τ_g can face two distinct challenges if group sizes are small. First, estimating $\psi_{g,2}$ requires sufficient movers between firm types; lacking these, the relevant block of $\hat{H}_{2,g}$ would be singular. Second, estimating τ_g requires variation in the job displacement event $T_{g,i}$. Auxiliary information could address both of these problems: population moments for firm-to-firm mobility (from census data or a structural model of worker-firm matching) could inform components of $H_{2,g}$ relevant for $\psi_{g,2}$, while population displacement rates (or models thereof) could inform those for τ_g . Similarly, in the banking example (Section 2.2.3), identifying relative bank supply conditions relies on observing firms interacting with multiple bank types within group g . If a sample in a small group lacks such firms, standard estimation fails. Here, the information for $H_{2,g}$ might come from broader data on firm-bank relationships or a structural model of bank-firm matching and market structure.

5 Empirical application

This section empirically demonstrates the practical consequences of our theoretical findings. We show that using a conventional one-step GMM/OLS estimator to evaluate a major policy reform yields dramatically different—and likely misleading—conclusions compared to our preferred two-step MD approach. We analyze the impact of the 2005 Dutch childcare expansion on “child penalty” measures (Kleven et al., 2019), a setting where the policy plausibly affects both labor market outcomes and fertility decisions, creating the conditions for the endogenous weighting bias. The conflicting evidence on the effects of such policies in the literature (Andresen and Nix, 2022; Kleven et al., 2024; Rabaté and Rellstab, 2021) highlights the need for a careful choice of estimation strategy. Our analysis proceeds by first defining the model-based outcomes (Section 5.1) and the policy intervention (Section 5.2). We then present a stylized comparison of GMM and MD estimators, followed by a richer MD specification that showcases its advantages (Section 5.3).

¹¹Note that $\hat{\theta}_g^{alt}$ is generically biased for θ_g conditionally on $\{E_{g,i}\}_{i=1}^{n_g}$.

5.1 Measuring the Child Penalty as a Model-Based Outcome

To implement our empirical analysis, we use administrative data from the Central Bureau of Statistics Netherlands (CBS) on the universe of Dutch residents. Different data sources, such as municipal registers or tax records, are matched through unique, anonymized identifiers for individuals or households. Appendix C presents the main variables used and sample construction.

Our analysis employs a dynamic event-study specification building on the DiD examples in Sections 2.2.1 and 2.3. We use data for individuals i residing in municipalities g over time t .¹² For each individual, we observe covariates $X_{g,i}$ (gender $G_{g,i}$ and birth cohort $B_{g,i}$) and the timing of their first childbirth relative to their birth year, $E_{g,i}$ (relative event time). We model $Y_{g,i,t}$ (earnings or employment) within relevant age ranges as:

$$Y_{g,i,t} = \gamma_{g,i} + \delta_{g,t}(X_{g,i}) + \sum_{h \geq h_0} \tau_{g,i,h} \mathbf{1}\{B_{g,i} + E_{g,i} = t - h\} + \varepsilon_{g,i,t}, \quad (8)$$

$$\mathbb{E}_{F_g}[\varepsilon_{g,i,t}|X_{g,i}, E_{g,i}] = 0.$$

Here $\gamma_{g,i}$ is an individual fixed effect, $\delta_{g,t}(X_{g,i})$ represents covariate-specific group-time effects, and $\tau_{g,i,h}$ is the individual-specific effect at horizon h relative to childbirth (the “child penalty”). Our primary parameter of interest, the group-level average child penalty, is the quintessential model-based outcome that is the focus of our paper. It is defined as:

$$\tau_{g,h}(x, e) := \mathbb{E}_{F_g}[\tau_{g,i,h}|X_{g,i} = x, E_{g,i} = e].$$

This parameter corresponds to θ_g in our general framework. As detailed in Section 2.3, a group-level policy can affect this outcome through two distinct channels: a direct effect on the individual-level penalties $\tau_{g,i,h}$, and an indirect composition effect, by altering the distribution of fertility timing $E_{g,i}$ (i.e., changing the population over which the average is taken).

For our subsequent policy analysis, we assume that the childcare expansion policy (detailed below) does not affect the distribution of baseline covariates $X_{g,i}$ (gender and birth cohort). Figure B.1 plots the histogram of $n_g(x)$ —the total number of individuals in group g with given gender and birth cohort. In our analysis we focus on $\tau_{g,h}(x, e)$ for $h \in \{-2, \dots, 3\}$ and $e \in \{27, \dots, 33\}$.¹³ We estimate the group-level parameters using the imputation-based estimator of Borusyak et al. (2024b) to construct $\hat{\tau}_{g,i,h}$, which we then aggregate to the (g, x) level.

Before proceeding to the policy evaluation, we examine $\hat{\tau}_{g,h}(x, e)$, particularly for pre-event horizons, to assess model validity. A key concern for MD estimators with finite group sizes is

¹²We assign individual i to location g based on the municipality in which they resided in the year of the childbirth.

¹³The full group parameter vector θ_g also includes all fixed effects $\delta_{g,t}(x)$.

the selection problem discussed in Section 4. We find that we can construct estimates for 95% of the (g, x) cells in our data.¹⁴ The fact that the share of excluded groups is small suggests that the selection bias quantified in the bound from Proposition 2 is likely to be negligible relative to the estimation error. This gives us confidence in the validity of the MD approach here. Figure B.2, which displays the estimated average child penalties by gender and age at first birth, reveals significant heterogeneity but finds limited evidence of anticipation effects (effects for $h < 0$ are near zero), supporting the credibility of the event-study design.

5.2 Institutional background – Dutch childcare provision

The 2005 Dutch Childcare Reform The 2005 Dutch Childcare Act fundamentally reformed early childhood care provision, transforming its financing and structure.¹⁵ The reform replaced fragmented local subsidies with a national, demand-driven, tripartite funding model (parents, employers, government), substantially reducing out-of-pocket costs for many families. Concurrently, the Act liberalized the supply side, allowing for-profit entry under a streamlined national regulatory framework. This combination of stimulated demand and relaxed entry conditions triggered rapid growth in formal childcare capacity. The substantial reduction in costs and expansion of supply makes it plausible that the reform influenced not only parental labor supply but also the timing and incidence of childbirth ($E_{g,i}$).

Childcare index We measure local childcare availability using a municipality-level Childcare Capacity Index ($CCI_{g,t}$), defined as the ratio of childcare workers ($J_{g,t}$) to the mandated staffing level (Decree, 1996):

$$CCI_{g,t} := \frac{J_{g,t}}{\sum_{l=0}^5 N_{g,t,l}/R_l}, \quad (9)$$

where $N_{g,t,l}$ is the number of children aged $l \in \{0, \dots, 5\}$. A value $CCI_{g,t} = 1$ signifies that the childcare workforce exactly meets the regulated staffing requirement. Recognizing that this index reflects an equilibrium outcome, our empirical strategy controls for municipality-specific trends correlated with baseline demographic demand (S_g). Identification of treatment intensity then relies on the remaining spatial and temporal variation in $CCI_{g,t}$ following the nationwide policy implementation. Figure Ia shows the wide variation in CCI across space and time, confirming that the reform had a powerful and heterogeneous impact.

¹⁴We use an $n_g(x)$ -weighted share, because our subsequent analysis uses such weighting. Proposition 2 naturally extends to accommodate this.

¹⁵For a more detailed discussion of the reform's institutional background and implementation timeline, see Appendix D.

5.3 Policy evaluation

Our empirical analysis of the childcare expansion proceeds in two stages. Initially, we present a stylized comparison of GMM and MD approaches. This first exercise serves not only as a practical illustration of the pitfalls discussed in Section 3 but also reflects a common empirical strategy, thereby underscoring the relevance of the biases we identify. Subsequently, we undertake a more nuanced MD analysis that directly incorporates our framework’s insights regarding policy-induced compositional effects, demonstrating its advantages for applied research.

5.3.1 Comparison between GMM and MD

To empirically demonstrate the GMM/OLS challenges identified in Section 3, we first examine a common evaluation strategy for policies with continuous local intensity ($CCI_{g,t}$): binarizing the treatment variable (e.g., Kleven et al., 2024; Lim and Duletzki, 2023; Rabaté and Rellstab, 2021). For our illustration, we construct a binary treatment indicator, W_g , based on the median expansion in childcare capacity:

$$W_g := \mathbf{1}\{\overline{CCI}_g^{\text{post}} - \overline{CCI}_g^{\text{pre}} > \text{Med}(\overline{CCI}_g^{\text{post}} - \overline{CCI}_g^{\text{pre}})\},$$

where $\overline{CCI}_g^{\text{pre}}$ and $\overline{CCI}_g^{\text{post}}$ are average CCI levels in group g during 1999-2004 and 2011-2016, respectively.¹⁶

GMM estimator Figure Ib shows apparent parallel evolution of average childcare index for the two groups, but also reveals that “control” groups ($W_g = 0$) still responded to the nationwide policy. This highlights the information loss from binarization. Nonetheless, this simplification allows for a straightforward OLS specification that is common in applied work:

$$\begin{aligned} Y_{g,i,t} = & \gamma_{g,i} + \delta_t + \lambda_{t-B_{g,i}} + \sum_{h \geq h_0} \alpha_h \mathbf{1}\{t - B_{g,i} - E_{g,i} = h\} \\ & + \sum_{h \geq h_0} \rho_h \mathbf{1}\{t - B_{g,i} - E_{g,i} = h\} W_g \\ & + \sum_{h \geq h_0} \xi_h \mathbf{1}\{t - B_{g,i} - E_{g,i} = h\} \mathbf{1}\{E_{g,i} + B_{g,i} > 2005\} \\ & + \sum_{h \geq h_0} \beta_h \mathbf{1}\{t - B_{g,i} - E_{g,i} = h\} W_g \mathbf{1}\{E_{g,i} + B_{g,i} > 2005\} + \nu_{g,i,t}. \end{aligned} \tag{10}$$

¹⁶The childcare index is relatively stable in these periods, with most of the expansion happening in 2005–2010, making them appropriate for defining W_g .

Here, β_h aims to capture the policy's effect. This OLS specification is precisely the type of one-step estimator vulnerable to the endogenous weighting bias identified in Section 3. Mechanically, the OLS estimate of β_h is a weighted average of group-specific effects, where the implicit weights are determined by the within-group distribution of $E_{g,i}$. If the policy W_g affects the distribution of fertility timing it alters these implicit weights, creating a spurious correlation between the policy and the weighting matrix that renders the estimator inconsistent.

MD estimator An alternative, MD approach relies on the first-stage estimates $\hat{\tau}_{g,h}(x, e)$, relating them directly to the same binarized policy:

$$\hat{\tau}_{g,h}(x, e) = \alpha_h + \rho_h W_g + \xi_h \mathbf{1}\{e + b > 2005\} + \beta_h \mathbf{1}\{e + b > 2005\} W_g + v_{g,h}(x, e). \quad (11)$$

Here, x includes covariates (education and birth cohort b). We run this regression separately for men and women, weighting by $n_g(x)$. Unlike the implicit and potentially endogenous weights in the GMM specification, here we apply explicit weights, $n_g(x)$. This choice of exogenous weights is the key feature of the MD approach that mechanically purges the endogenous weighting bias.

Comparing empirical results Figure II reveals a big difference between the two approaches. The conventional GMM/OLS estimator (panel a) suggests the policy had large effects, increasing mothers' post-childbirth earnings by up to 12,000 euros and labor force participation by 13 percentage points. In contrast, the MD approach (panel b), which is robust to the endogenous weighting bias, finds that these effects are almost an order of magnitude smaller and, for earnings, statistically indistinct from the effects for men.

This divergence is a direct empirical manifestation of the theoretical problem identified in this paper. The GMM/OLS specification (10) implicitly uses the conditional distribution of fertility timing ($E_{g,i}$) to weigh observations, and this distribution may be affected by the policy. While this is a population-level concern, the problem is exacerbated in practice by finite-sample issues. For the many groups with a small number of individuals $n_g(x)$ (Figure B.1), the observed distribution of $E_{g,i}$ can be highly variable due to random sampling alone. Not all of this variation in the implicit weights can and should be attributed to a causal effect of the policy, but the GMM/OLS estimator uses it regardless, without raising any warnings.

This is precisely why the MD specification is more attractive empirically. It forces the researcher to confront this variation—whether it stems from policy-induced effects or idiosyncratic noise—and make an explicit choice about weighting. By using pre-determined weights like $n_g(x)$, the MD estimator (11) avoids relying on potentially noisy, endogenous, and data-driven implicit weights. The fact that this single methodological choice changes the results

so profoundly underscores the practical importance of our recommendation for transparent, two-step estimation.

Remark 5.1 (On Normalizing Child Penalties). The child penalty literature often normalizes estimates (e.g., by baseline earnings). While entirely feasible within our framework, such normalization would produce an additional finite-sample bias discussed in Remark 4.2. As a result, we opt to analyze unnormalized child penalties, isolating the core methodological issues of GMM endogenous weighting from important but separate statistical problems.

5.3.2 A Richer MD Specification for Policy Evaluation

Moving beyond the stylized comparison, we now leverage the flexibility of the MD approach to conduct a more nuanced evaluation. This allows us to use the continuous nature of the policy intensity, $CCI_{g,t}$, and investigate the compositional effects discussed in Section 2.3. We estimate the following specification:¹⁷

$$\hat{\tau}_{g,h}(x, e) = \alpha_{g,h}(x) + \lambda_h^{(0)}(x, e) + \lambda_h^{(1)}(x, e)S_g + \sum_{j=-2}^h \beta_{h,j} CCI_{g,b+e+j} + v_{g,h}(x, e). \quad (12)$$

The specification includes granular fixed effects ($\alpha_{g,h}(x)$, $\lambda_h^{(0)}(x, e)$, $\lambda_h^{(1)}(x, e)S_g$) to control flexibly for a wide array of potential confounders. The key term of interest is the summation, which captures the rich, dynamic effects of the policy. Crucially, incorporating CCI levels from periods prior to childbirth ($j < 0$) provides a powerful specification test. It allows us to probe for policy-induced compositional shifts that might manifest as pre-event “effects” on the average child penalty. If $\beta_{h,j}$ for $j < 0$ were significant, it may suggest that changes in childcare capacity are correlated with our outcome even before they can have a direct causal effect, likely via changes in who becomes a mother in a given municipality.

Table 1 presents the estimated effects for mothers. We find no evidence of significant effects for $j < 0$; the coefficients are small and statistically insignificant. This null result suggests that the compositional confounding is not a first-order issue in this richer specification. For the post-birth periods, we find that an expansion of childcare capacity in the year of childbirth has a persistent positive impact on mothers’ labor force participation. We also observe positive effects on mothers’ earnings from expansions one year after birth, when formal maternity leave ends. These patterns are not mirrored for fathers (Appendix Table B.1), whose labor market outcomes are largely unchanged, though their earnings respond positively to contemporaneous increases in childcare capacity, suggesting an intensive margin response. These plausible, nuanced results showcase the value of the MD approach.

¹⁷As with the previous analysis, we weight using $n_g(x)$.

6 Conclusion

This paper develops a framework for analyzing causal effects of group-level policies on model-based outcomes, defined via micro-level models. We highlight how policy-induced changes in the distribution of the micro-level data can bias common estimators. We find that one-step GMM estimators, including OLS with policy interactions, generally yield inconsistent estimates due to an endogenous weighting problem. Two-stage MD estimators, while avoiding this specific issue, can be inconsistent with small group sizes if policy influences first-stage sample estimability, creating a selection bias. Our analysis shows that MD estimators achieve consistency in the asymptotic regime where the group sizes are allowed to grow. In settings with small groups, consistency can be restored by using auxiliary data, such as known population moments, to bypass the sample-dependent selection.

Future research could extend this framework to nonlinear moment conditions, explore random effects models in settings with small groups where MD estimators face challenges, and investigate models with complex interdependencies across groups, allowing for the analysis of network data. Addressing these extensions would enhance the applicability of causal inference methods for a broader range of policy evaluation problems involving model-based outcomes.

References

- Abowd, John M., Francis Kramarz, and David N. Margolis.** 1999. “High Wage Workers and High Wage Firms.” *Econometrica* 67 (2): 251–333. [10.1111/1468-0262.00020](https://doi.org/10.1111/1468-0262.00020).
- Ackerberg, Daniel A, Kevin Caves, and Garth Frazer.** 2015. “Identification properties of recent production function estimators.” *Econometrica* 83 (6): 2411–2451.
- Adao, Rodrigo, Michal Kolesár, and Eduardo Morales.** 2019. “Shift-share designs: Theory and inference.” *The Quarterly Journal of Economics* 134 (4): 1949–2010.
- Altonji, Joseph G, and Lewis M Segal.** 1996. “Small-sample bias in GMM estimation of covariance structures.” *Journal of Business & Economic Statistics* 14 (3): 353–366.
- Amiti, Mary, and Jozef Konings.** 2007. “Trade liberalization, intermediate inputs, and productivity: Evidence from Indonesia.” *American economic review* 97 (5): 1611–1638.
- Andresen, Martin Eckhoff, and Emily Nix.** 2022. “What Causes the Child Penalty? Evidence from Adopting and Same-Sex Couples.” *Journal of Labor Economics* 40 (4): 971–1004. [10.1086/718565](https://doi.org/10.1086/718565).
- Arkhangelsky, Dmitry, and Guido Imbens.** 2024. “Causal Models for Longitudinal and Panel Data: A Survey.” *The Econometrics Journal* 27 (3): C1–C61. [10.1093/ectj/utae014](https://doi.org/10.1093/ectj/utae014).
- Arkhangelsky, Dmitry, Guido W. Imbens, Lihua Lei, and Xiaoman Luo.** 2024. “Design-Robust Two-Way-Fixed-Effects Regression for Panel Data.” *Quantitative Economics* 15 (4): 999–1034. [10.3982/QE1962](https://doi.org/10.3982/QE1962).
- Aronow, Peter M, and Cyrus Samii.** 2017. “Estimating average causal effects under general interference, with application to a social network experiment.” *The Annals of Applied Statistics* 11 (4): 1912.
- Bertheau, Antoine, Edoardo Maria Acabbi, Cristina Barceló, Andreas Gulyas, Stefano Lombardi, and Raffaele Saggio.** 2023. “The Unequal Consequences of Job Loss across Countries.” *American Economic Review: Insights* 5 (3): 393–408. [10.1257/aeri.20220006](https://doi.org/10.1257/aeri.20220006).
- Bettendorf, Leon J. H., Egbert L. W. Jongen, and Paul Muller.** 2015. “Childcare Subsidies and Labour Supply — Evidence from a Large Dutch Reform.” *Labour Economics* 36 112–123. [10.1016/j.labeco.2015.03.007](https://doi.org/10.1016/j.labeco.2015.03.007).
- Borusyak, Kirill, and Peter Hull.** 2023. “Nonrandom Exposure to Exogenous Shocks.” *Econometrica* 91 (6): 2155–2185. [10.3982/ECTA19367](https://doi.org/10.3982/ECTA19367).

- Borusyak, Kirill, Peter Hull, and Xavier Jaravel.** 2022. “Quasi-experimental shift-share research designs.” *The Review of economic studies* 89 (1): 181–213.
- Borusyak, Kirill, Peter Hull, and Xavier Jaravel.** 2024a. “Design-Based Identification with Formula Instruments: A Review.” *The Econometrics Journal* utae003. [10.1093/ectj/utae003](https://doi.org/10.1093/ectj/utae003).
- Borusyak, Kirill, Xavier Jaravel, and Jann Spiess.** 2024b. “Revisiting Event-Study Designs: Robust and Efficient Estimation.” *The Review of Economic Studies* rdae007. [10.1093/restud/rdae007](https://doi.org/10.1093/restud/rdae007).
- Callaway, Brantly, and Pedro H.C. Sant’Anna.** 2021. “Difference-in-Differences with Multiple Time Periods.” *Journal of Econometrics* 225 (2): 200–230. [10.1016/j.jeconom.2020.12.001](https://doi.org/10.1016/j.jeconom.2020.12.001).
- Daruich, Diego, Sabrina Di Addario, and Raffaele Saggio.** 2023. “The Effects of Partial Employment Protection Reforms: Evidence from Italy.” *The Review of Economic Studies* rdad012. [10.1093/restud/rdad012](https://doi.org/10.1093/restud/rdad012).
- de Chaisemartin, Clément, and Xavier D’Haultfœuille.** 2020. “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects.” *American Economic Review* 110 (9): 2964–2996. [10.1257/aer.20181169](https://doi.org/10.1257/aer.20181169).
- Decree.** 1996. “Childcare Quality Requirements (Temporary Measures) Decree, 1996.” [https://splash-db.eu/?cHash=3d3f1e98078d34205fd2c1a80af9ff46&tx_perfar_policydocument\[action\]=show&tx_perfar_policydocument\[controller\]=Policydocument&tx_perfar_policydocument\[policydocument\]=1352](https://splash-db.eu/?cHash=3d3f1e98078d34205fd2c1a80af9ff46&tx_perfar_policydocument[action]=show&tx_perfar_policydocument[controller]=Policydocument&tx_perfar_policydocument[policydocument]=1352).
- Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu.** 2015. “Competition, markups, and the gains from international trade.” *American Economic Review* 105 (10): 3183–3221.
- Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár.** 2024. “Contamination Bias in Linear Regressions.” *American Economic Review* 114 (12): 4015–4051. [10.1257/aer.20221116](https://doi.org/10.1257/aer.20221116).
- Goodman-Bacon, Andrew.** 2021. “Difference-in-differences with variation in treatment timing.” *Journal of econometrics* 225 (2): 254–277.
- Greenstone, Michael, Richard Hornbeck, and Enrico Moretti.** 2010. “Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings.” *Journal of political economy* 118 (3): 536–598.
- Greenstone, Michael, John A List, and Chad Syverson.** 2012. “The effects of environmental regulation on the competitiveness of US manufacturing.” Technical report, National Bureau of Economic Research.

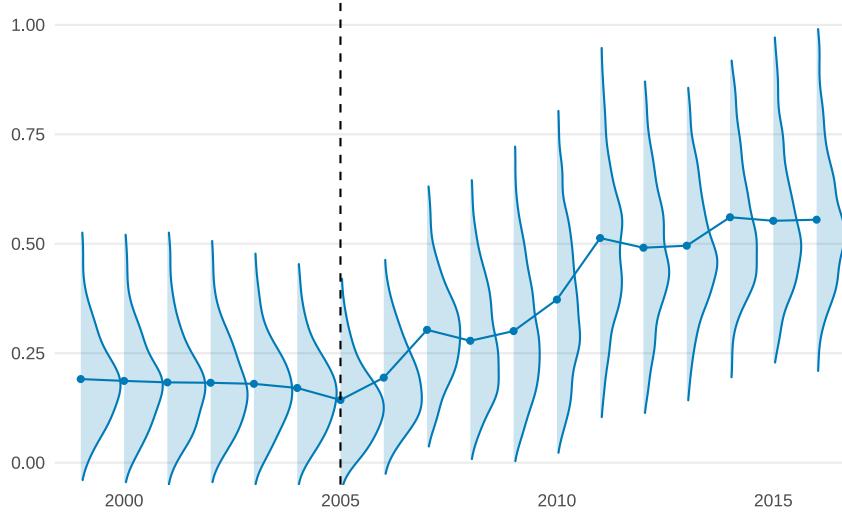
- Heckman, James J.** 1979. “Sample selection bias as a specification error.” *Econometrica: Journal of the econometric society* 153–161.
- Imbens, Guido W, and Tony Lancaster.** 1994. “Combining micro and macro data in microeconometric models.” *The Review of Economic Studies* 61 (4): 655–680.
- Imbens, Guido W., and Donald B. Rubin.** 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press, . [10.1017/CBO9781139025751](https://doi.org/10.1017/CBO9781139025751).
- Jayaratne, Jith, and Philip E Strahan.** 1996. “The finance-growth nexus: Evidence from bank branch deregulation.” *The Quarterly Journal of Economics* 111 (3): 639–670.
- Jiménez, Gabriel, Steven Ongena, José-Luis Peydró, and Jesús Saurina.** 2017. “Macro-prudential policy, countercyclical bank capital buffers, and credit supply: Evidence from the Spanish dynamic provisioning experiments.” *Journal of Political Economy* 125 (6): 2126–2177.
- Khwaja, Asim Ijaz, and Atif Mian.** 2008. “Tracing the impact of bank liquidity shocks: Evidence from an emerging market.” *American Economic Review* 98 (4): 1413–1442.
- Kleven, Henrik, Camille Landais, Johanna Posch, Andreas Steinhauer, and Josef Zweimüller.** 2024. “Do Family Policies Reduce Gender Inequality? Evidence from 60 Years of Policy Experimentation.” *American Economic Journal: Economic Policy* 16 (2): 110–149. [10.1257/pol.20210346](https://doi.org/10.1257/pol.20210346).
- Kleven, Henrik, Camille Landais, and Jakob Egholt Søgaard.** 2019. “Children and Gender Inequality: Evidence from Denmark.” *American Economic Journal: Applied Economics* 11 (4): 181–209. [10.1257/app.20180010](https://doi.org/10.1257/app.20180010).
- Lim, Nayeon, and Lisa-Marie Duletzki.** 2023. “The Effects of Public Childcare Expansion on Child Penalties - Evidence From West Germany.”
- Lloyd, Eva.** 2013. “Childminders in the Netherlands.” <https://childcarecanada.org/documents/research-policy-practice/13/08/childminders-netherlands>, Childcare Resource and Research Unit.
- Loecker, Jan De, and Frederic Warzynski.** 2012. “Markups and firm-level export status.” *American economic review* 102 (6): 2437–2471.
- Newey, Whitney K, and Richard J Smith.** 2004. “Higher order properties of GMM and generalized empirical likelihood estimators.” *Econometrica* 72 (1): 219–255.

- Noailly, Joëlle, and Sjors Visser.** 2009. “The impact of market forces on the provision of childcare: Insights from the 2005 Childcare Act in the Netherlands.” *CPB Memorandum* 176, <https://www.cpb.nl/sites/default/files/publicaties/download/memo176.pdf>.
- Pekkarinen, Tuomas, Roope Uusitalo, and Sari Pekkala Kerr.** 2009. “School Tracking and Intergenerational Income Mobility: Evidence from the Finnish Comprehensive School Reform.” *Journal of Public Economics* 93 (7-8): 965–973.
- Rabaté, Simon, and Sara Rellstab.** 2021. “The Child Penalty in the Netherlands and Its Determinants.” *CPB Discussion Paper*. [10.34932/TRKZ-QH66](https://doi.org/10.34932/TRKZ-QH66).
- Schmieder, Johannes F., Till Von Wachter, and Jörg Heining.** 2023. “The Costs of Job Displacement over the Business Cycle and Its Sources: Evidence from Germany.” *American Economic Review* 113 (5): 1208–1254. [10.1257/aer.20200252](https://doi.org/10.1257/aer.20200252).
- Sun, Liyang, and Sarah Abraham.** 2021. “Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects.” *Journal of Econometrics* 225 (2): 175–199. [10.1016/j.jeconom.2020.09.006](https://doi.org/10.1016/j.jeconom.2020.09.006).

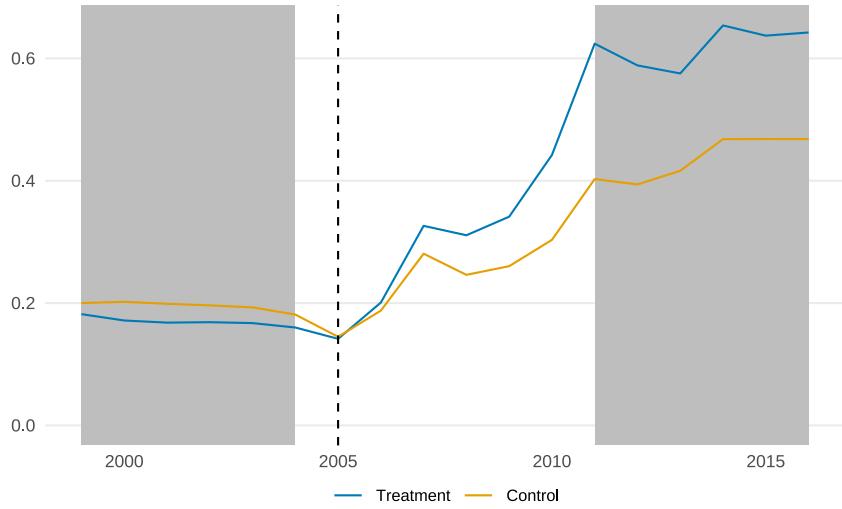
Figures and tables

Figure I: Childcare supply expansion

(a) Distribution of childcare index by municipality

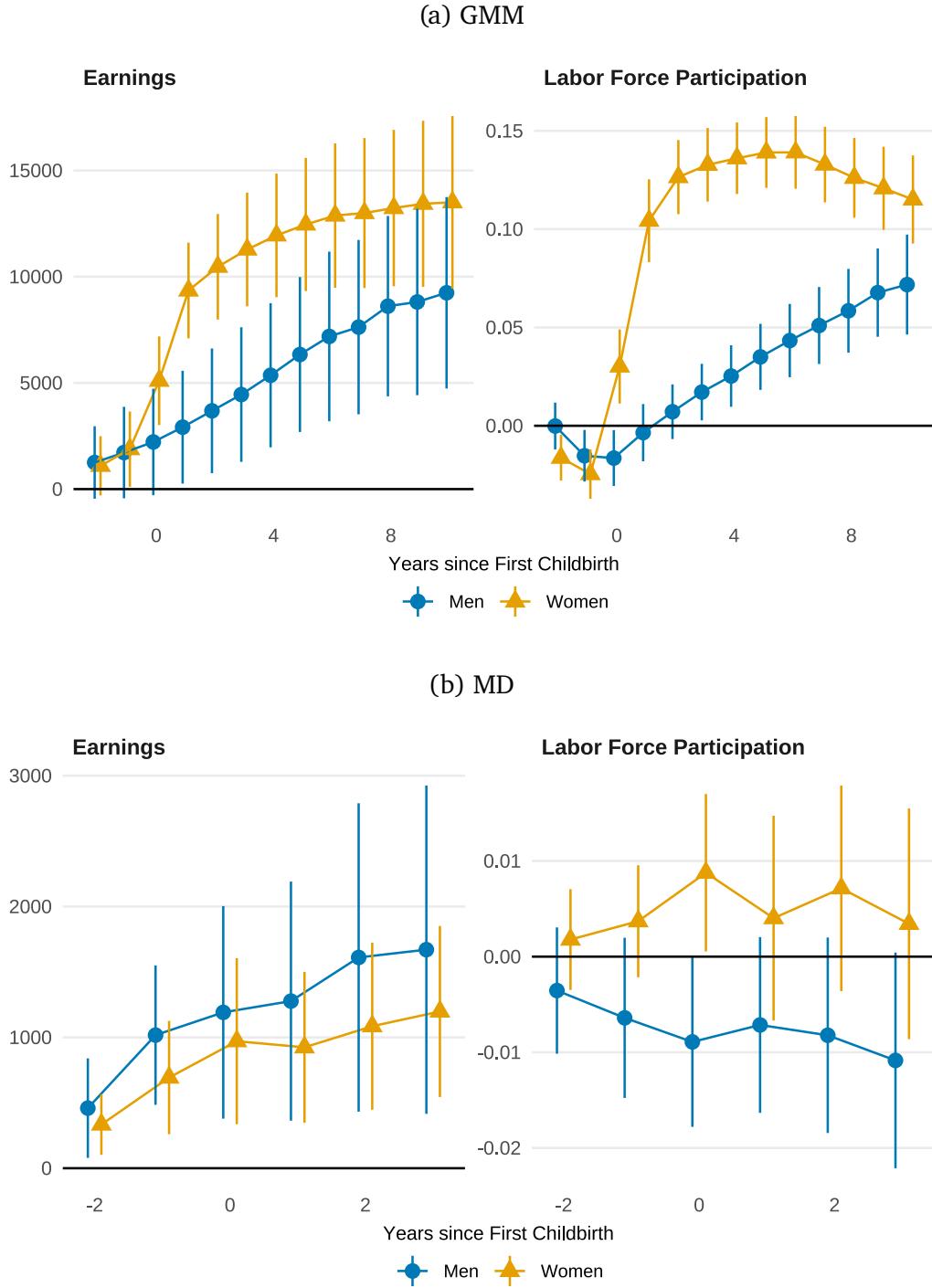


(b) Simplified 2×2 DiD design



Notes: These figures present the variation in childcare supply per preschool-aged children across municipalities from 1999 to 2016. Our childcare capacity index (CCI) for each municipality is calculated by dividing the number of childcare jobs in a given municipality g and year t ($N_{g,t}^{jobs}$) by the number of required number of childminders in the same locality (see Definition 9). The vertical line illustrates the timing of the 2005 Dutch childcare expansion reform. Panel (a) illustrates the substantial variation in childcare availability between different municipalities and the large increase due to the 2005 childcare expansion reform. Dots represent the mean CCI in a given year, whereas the shaded area represents the distribution of CCI across municipalities in that year. Panel (b) illustrates the equivalent simplified 2×2 DiD design, where the time variation is binary (gray area) and treatment is binary (see Section 5.3.1). Treatment is defined as municipalities with above median expansion of CCI between the baseline period (1999-2004) and the post-expansion period (2011-2016).

Figure II: GMM vs MD: Effect of the childcare provision expansion on CP



Notes: This figure presents the effect of the childcare provision expansion on child penalties (CP) in earnings, estimated separately using the GMM and the MD approaches (see Section 5.3.1 for more details). Figure IIa presents the results of the GMM approach commonly used in the literature (specification (10)). Figure IIb presents the results of the MD approach we suggest (specification (11)). We split the estimation between men (blue) and women (orange) and present the results for both the intensive margin (earnings, as shown in the panel above) and the extensive margin (participation, as shown in the panel below). Each dot presents the corresponding coefficient and its marginal 95% confidence interval based on standard errors clustered by municipality g .

Table 1: Rich MD specification – the childcare provision expansion on CP (mothers)

	(a) Earnings					
	$\hat{\tau}_{g,-2}$	$\hat{\tau}_{g,-1}$	$\hat{\tau}_{g,0}$	$\hat{\tau}_{g,1}$	$\hat{\tau}_{g,2}$	$\hat{\tau}_{g,3}$
$CCI_{g,b+e-2}$	283.5 (403.1)	384.8 (631.4)	798.3 (602.3)	355.2 (714.4)	398.6 (735.4)	-25.3 (895.9)
$CCI_{g,b+e-1}$		735.5 (544.3)	778.0 (541.7)	866.9 (753.6)	838.7 (778.7)	1091.3 (690.2)
$CCI_{g,b+e}$			394.9 (485.9)	191.7 (485.3)	109.2 (716.6)	287.8 (803.4)
$CCI_{g,b+e+1}$				1510.9* (618.4)	1522.9* (634.0)	1969.4* (892.2)
$CCI_{g,b+e+2}$					407.1 (630.2)	-59.1 (635.2)
$CCI_{g,b+e+3}$						1330.8 (889.7)
N	10,941	10,941	10,941	10,941	10,941	10,941
R^2	0.163	0.258	0.314	0.369	0.385	0.369
FE: Municipality g	X	X	X	X	X	X
FE: $B_{g,i} \times E_{g,i}$	X	X	X	X	X	X
FE: $(B_{g,i} \times E_{g,i})S_g$	X	X	X	X	X	X

+ p < 0.1, * p < 0.05, ** p < 0.01

	(b) Participation					
	$\hat{\tau}_{g,-2}$	$\hat{\tau}_{g,-1}$	$\hat{\tau}_{g,0}$	$\hat{\tau}_{g,1}$	$\hat{\tau}_{g,2}$	$\hat{\tau}_{g,3}$
$CCI_{g,b+e-2}$	-0.012 (0.010)	-0.026+ (0.014)	-0.007 (0.015)	-0.006 (0.017)	-0.021 (0.019)	-0.018 (0.020)
$CCI_{g,b+e-1}$		0.027* (0.011)	0.003 (0.011)	0.005 (0.013)	-0.002 (0.017)	-0.006 (0.016)
$CCI_{g,b+e}$			0.031** (0.011)	0.031* (0.012)	0.031* (0.014)	0.038* (0.016)
$CCI_{g,b+e+1}$				0.009 (0.012)	-0.000 (0.016)	-0.007 (0.017)
$CCI_{g,b+e+2}$					0.005 (0.014)	0.001 (0.012)
$CCI_{g,b+e+3}$						0.006 (0.015)
N	10,941	10,941	10,941	10,941	10,941	10,941
R^2	0.050	0.056	0.068	0.109	0.121	0.128
FE: Municipality g	X	X	X	X	X	X
FE: $B_{g,i} \times E_{g,i}$	X	X	X	X	X	X
FE: $(B_{g,i} \times E_{g,i})S_g$	X	X	X	X	X	X

+ p < 0.1, * p < 0.05, ** p < 0.01

Notes: These tables present the effect of the childcare provision expansion on child penalties (CP) of mothers in earnings (above) and labor force participation (below). See Section 5.3.2 for more details and Equation (12) for the specification. Standard errors clustered by municipality g are in parentheses.

Supplementary Appendix

A Theoretical details

A.1 Proofs

A.1.1 Proof of Proposition 1

Proof. The GMM estimator $(\hat{\alpha}, \hat{B})$ minimizes

$$L(\alpha, B) = \sum_{g=1}^G (\theta_g - \alpha - BW_g)^\top \tilde{A}_g (\theta_g - \alpha - BW_g)$$

subject to $\Gamma^\top \alpha = \mathbf{0}$ and $\Gamma^\top B = \mathbf{0}$. Since \tilde{A}_g is symmetric and positive definite by assumption, matrix $\check{A}_g := U^\top \tilde{A}_g U$ is also positive definite, where U is a $k \times k'$ matrix with orthonormal columns ($U^\top U = I_{k'}$) forming a basis for $\text{Im}(P_{\Gamma^\perp})$. We transform α and B to $\tilde{\alpha} := U^\top \alpha$ and $\tilde{B} := U^\top B$.

The estimators $\hat{\alpha}$ and \hat{B} satisfy the normal equations:

$$\left(\sum_{g=1}^G \check{A}_g \right) \hat{\alpha} + \left(\sum_{g=1}^G \check{A}_g \hat{B} W_g \right) = \sum_{g=1}^G U^\top \tilde{A}_g \theta_g \quad (\text{A.1})$$

$$\left(\sum_{g=1}^G W_g \otimes \check{A}_g \right) \hat{\alpha} + \left(\sum_{g=1}^G W_g W_g^\top \otimes \check{A}_g \right) \text{vec}(\hat{B}) = \text{vec} \left(\sum_{g=1}^G U^\top \tilde{A}_g \theta_g W_g^\top \right) \quad (\text{A.2})$$

Under Assumption 3.1(c), $\theta_g(W_g) = \alpha_g + B_0 W_g$. Since $\Gamma^\top B_0 = \mathbf{0}$, $B_0 = U \tilde{B}_0$ for some $\tilde{B}_0 \in \mathbb{R}^{k' \times p}$. The population intercept α^0 also satisfies $\Gamma^\top \alpha^0 = \mathbf{0}$, so $\alpha^0 = U \tilde{\alpha}^0$. We define $\epsilon_g^0 = \alpha_g - \alpha^0$. By the definition of α^0 as the GLS population intercept in the subspace, it satisfies $\mathbb{E}[\check{A}_g(U^\top \alpha_g - \tilde{\alpha}^0)] = \mathbf{0}$, which implies $\mathbb{E}[U^\top \tilde{A}_g \epsilon_g^0] = \mathbf{0}$. We substitute the model into the right-hand side (RHS) of the normal equations:

$$\text{RHS of (A.1)} = \sum U^\top \tilde{A}_g \epsilon_g^0 + \left(\sum \check{A}_g \right) \tilde{\alpha}^0 + \left(\sum \check{A}_g \tilde{B}_0 W_g \right),$$

$$\text{RHS of (A.2)} = \text{vec} \left(\sum U^\top \tilde{A}_g \epsilon_g^0 W_g^\top \right) + \text{vec} \left(\sum \check{A}_g \tilde{\alpha}^0 W_g^\top \right) + \text{vec} \left(\sum \check{A}_g \tilde{B}_0 W_g W_g^\top \right).$$

Note that $\text{vec}(\check{A}_g \tilde{\alpha}^0 W_g^\top) = (W_g \otimes \check{A}_g) \tilde{\alpha}^0$ and $\text{vec}(\check{A}_g \tilde{B}_0 W_g W_g^\top) = (W_g W_g^\top \otimes \check{A}_g) \text{vec}(\tilde{B}_0)$.

Let $\Delta \tilde{\alpha} = \hat{\alpha} - \tilde{\alpha}^0$ and $\Delta \tilde{B} = \hat{B} - \tilde{B}_0$. We substitute these into the left-hand side (LHS) of

the normal equations:

$$\begin{aligned}\text{LHS of (A.1)} &= \left(\sum \check{A}_g\right) \Delta \tilde{\alpha} + \sum \check{A}_g \Delta \tilde{B} W_g + \left(\sum \check{A}_g\right) \tilde{\alpha}^0 + \sum \check{A}_g \tilde{B}_0 W_g \\ \text{LHS of (A.2)} &= (\sum W_g \otimes \check{A}_g) \Delta \tilde{\alpha} + (\sum W_g W_g^\top \otimes \check{A}_g) \text{vec}(\Delta \tilde{B}) + (\sum W_g \otimes \check{A}_g) \tilde{\alpha}^0 + \\ &\quad (\sum W_g W_g^\top \otimes \check{A}_g) \text{vec}(\tilde{B}_0)\end{aligned}$$

Equating the LHS and RHS for both normal equations, the terms involving $\tilde{\alpha}^0$ and \tilde{B}_0 cancel out. This leaves a system for $(\Delta \tilde{\alpha}, \text{vec}(\Delta \tilde{B}))$:

$$\begin{aligned}\left(\sum_{g=1}^G \check{A}_g\right) \Delta \tilde{\alpha} + \left(\sum_{g=1}^G W_g^\top \otimes \check{A}_g\right) \text{vec}(\Delta \tilde{B}) &= \sum_{g=1}^G U^\top \tilde{A}_g \epsilon_g^0 \\ \left(\sum_{g=1}^G W_g \otimes \check{A}_g\right) \Delta \tilde{\alpha} + \left(\sum_{g=1}^G W_g W_g^\top \otimes \check{A}_g\right) \text{vec}(\Delta \tilde{B}) &= \text{vec}\left(\sum_{g=1}^G U^\top \tilde{A}_g \epsilon_g^0 W_g^\top\right).\end{aligned}$$

Let $\tilde{H}_{11} = \sum \check{A}_g$, $\tilde{H}_{12} = \sum (W_g^\top \otimes \check{A}_g)$, $\tilde{H}_{21} = \sum (W_g \otimes \check{A}_g)$, $\tilde{H}_{22} = \sum (W_g W_g^\top \otimes \check{A}_g)$. Let $C_1^\epsilon = \sum U^\top \tilde{A}_g \epsilon_g^0$ and $C_2^\epsilon = \text{vec}(\sum U^\top \tilde{A}_g \epsilon_g^0 W_g^\top)$. Solving for $\text{vec}(\Delta \tilde{B})$ using the Schur complement $\tilde{S} = \tilde{H}_{22} - \tilde{H}_{21} \tilde{H}_{11}^{-1} \tilde{H}_{12}$:

$$\text{vec}(\Delta \tilde{B}) = \tilde{S}^{-1} \left(C_2^\epsilon - \tilde{H}_{21} \tilde{H}_{11}^{-1} C_1^\epsilon \right).$$

Now, we consider the probability limits as $G \rightarrow \infty$. Under Assumption 3.1 and assuming necessary moments for the LLN exist: $\frac{1}{G} \tilde{H}_{ij} \xrightarrow{p} H_{ij,plim}$, $\frac{1}{G} C_1^\epsilon \xrightarrow{p} \mathbb{E}[U^\top \tilde{A}_g \epsilon_g^0] = \mathbf{0}$, $\frac{1}{G} C_2^\epsilon \xrightarrow{p} \mathbb{E}[\text{vec}(U^\top \tilde{A}_g \epsilon_g^0 W_g^\top)] = \text{vec}(\mathbb{E}[U^\top \tilde{A}_g \epsilon_g^0 W_g^\top])$.

Let $S_{plim} = \text{plim}_{G \rightarrow \infty} \frac{1}{G} \tilde{S} = H_{22,plim} - H_{21,plim} H_{11,plim}^{-1} H_{12,plim}$. For the estimator to be well-defined, S_{plim} must be invertible. S_{plim} is the Schur complement of $H_{11,plim} = \mathbb{E}[\check{A}_g(W_g)]$ in the overall expected Hessian matrix

$$H_{plim} = \mathbb{E} \begin{pmatrix} \check{A}_g & W_g^\top \otimes \check{A}_g \\ W_g \otimes \check{A}_g & W_g W_g^\top \otimes \check{A}_g \end{pmatrix}.$$

Thus, S_{plim} is positive definite if H_{plim} is positive definite. H_{plim} is positive definite if $\mathbb{E}[\text{Tr}((\mathbf{v}_1 + V_2 W_g)^\top \check{A}_g(W_g)(\mathbf{v}_1 + V_2 W_g))] > 0$ for any non-zero pair (\mathbf{v}_1, V_2) , where $\mathbf{v}_1 \in \mathbb{R}^{k'}$ and $V_2 \in \mathbb{R}^{k' \times p}$. Given that $\check{A}_g(W_g)$ is positive definite almost surely, this condition requires that $\mathbf{v}_1 + V_2 W_g = \mathbf{0}_{k'}$ almost surely only if $\mathbf{v}_1 = \mathbf{0}$ and $V_2 = \mathbf{0}_{k' \times p}$. Assumption 3.1(a), which states that $\mathbb{V}[W_g]$ is positive definite, ensures this linear independence.

Then,

$$\begin{aligned}\text{plim}_{G \rightarrow \infty} \text{vec}(\Delta \tilde{B}) &= S_{plim}^{-1} \left(\text{vec}(\mathbb{E}[U^\top \tilde{A}_g \epsilon_g^0 W_g^\top]) - H_{21,plim} H_{11,plim}^{-1} \cdot \mathbf{0} \right) \\ &= S_{plim}^{-1} \text{vec}(\mathbb{E}[U^\top \tilde{A}_g \epsilon_g^0 W_g^\top])\end{aligned}$$

So, $\text{plim}_{G \rightarrow \infty} (\hat{\tilde{B}} - \tilde{B}_0) = \text{unvec}(S_{plim}^{-1} \text{vec}(\mathbb{E}[U^\top \tilde{A}_g \epsilon_g^0 W_g^\top]))$. Consistency holds if and only if $U^\top \mathbb{E}[\tilde{A}_g \epsilon_g^0 W_g^\top] = \mathbf{0}_{k' \times p}$. The condition in Proposition 1 is $M_\Gamma \text{Cov}[\tilde{A}_g \epsilon_g^0, W_g] = \mathbf{0}_{k \times p}$. Since $U^\top \mathbb{E}[\tilde{A}_g \epsilon_g^0] = \mathbf{0}$ the condition $M_\Gamma \text{Cov}[\tilde{A}_g \epsilon_g^0, W_g] = \mathbf{0}$ simplifies to $M_\Gamma \mathbb{E}[\tilde{A}_g \epsilon_g^0 W_g^\top] = \mathbf{0}$, which itself is equivalent to $U^\top \mathbb{E}[\tilde{A}_g \epsilon_g^0 W_g^\top] = \mathbf{0}_{k' \times p}$. \square

A.1.2 Proof of Proposition 2

Proof. Let $\boldsymbol{\theta}_g \in \mathbb{R}^k$ be the group-specific parameter vector. Let $\boldsymbol{\alpha} \in \mathbb{R}^k$ be the intercept, $B \in \mathbb{R}^{k \times p}$ be the matrix of policy effects, $\Gamma \in \mathbb{R}^{k \times q}$ be the matrix defining dimensions of unobserved group heterogeneity, and $\boldsymbol{\lambda}_g \in \mathbb{R}^q$ be group-specific coefficients. The objective function for the weighted regression is

$$L(\boldsymbol{\alpha}, B, \{\boldsymbol{\lambda}_g\}; \{\omega_g\}) = \sum_{g=1}^G \|\boldsymbol{\theta}_g - \boldsymbol{\alpha} - \Gamma \boldsymbol{\lambda}_g - BW_g\|_2^2 \omega_g$$

For fixed $(\boldsymbol{\alpha}, B)$, minimizing with respect to $\boldsymbol{\lambda}_g$ yields $\Gamma \hat{\boldsymbol{\lambda}}_g(\boldsymbol{\alpha}, B) = P_\Gamma(\boldsymbol{\theta}_g - \boldsymbol{\alpha} - BW_g)$, where $P_\Gamma = \Gamma(\Gamma^\top \Gamma)^{-1} \Gamma^\top$ is the orthogonal projection matrix onto the column space of Γ . Substituting this into the objective function, it becomes $\sum_{g=1}^G \|P_{\Gamma^\perp}(\boldsymbol{\theta}_g - \boldsymbol{\alpha} - BW_g)\|_2^2 \omega_g$, where $P_{\Gamma^\perp} = I_k - P_\Gamma$.

Let $\boldsymbol{\theta}'_g := P_{\Gamma^\perp} \boldsymbol{\theta}_g$. The parameters of interest $(\boldsymbol{\alpha}, B)$ are subject to the orthogonality constraints $\Gamma^\top \boldsymbol{\alpha} = \mathbf{0}$ and $\Gamma^\top B = \mathbf{0}$. These constraints imply that $\boldsymbol{\alpha}$ and the columns of B lie in the orthogonal complement of the column space of Γ . Thus, $P_{\Gamma^\perp} \boldsymbol{\alpha} = \boldsymbol{\alpha}$ and $P_{\Gamma^\perp} B = B$ (where the projection acts on B column-wise). The objective function simplifies to

$$L'(\boldsymbol{\alpha}, B; \{\omega_g\}) = \sum_{g=1}^G \|\boldsymbol{\theta}'_g - \boldsymbol{\alpha} - BW_g\|_2^2 \omega_g$$

The parameters $(\boldsymbol{\alpha}, B)$ effectively operate in a subspace where their dimension, after projection, corresponds to $k' = k - q$ rows. Let $X = (\boldsymbol{\alpha}, B)$ represent this collection of parameters, which can be viewed as a $k' \times (1 + p)$ matrix.

Let $X_1 := (\hat{\boldsymbol{\alpha}}_1^{MD}, \hat{B}_1^{MD})$ be the minimizer of $L'(\boldsymbol{\alpha}, B; \{\omega_g\})$, and $X_0 := (\boldsymbol{\alpha}^*, B^*)$ be the minimizer of $L'(\boldsymbol{\alpha}, B; \{1\})$ (i.e., $\omega_g = 1$ for all g) and define $\Delta X := X_1 - X_0 = (\Delta \boldsymbol{\alpha}, \Delta B)$. The oracle residual is $r_g^* := \boldsymbol{\theta}_g - \boldsymbol{\alpha}^* - \Gamma \boldsymbol{\lambda}_g^* - B^* W_g$, which is by construction equal to $r'_g :=$

$\theta'_g - \alpha^* - B^*W_g$. The objective $L'(\alpha, B; \{\omega_g\})$ is quadratic in $X = (\alpha, B)$. The first-order conditions (FOCs) are $\nabla L'_1(X_1) = \mathbf{0}$ and $\nabla L'_0(X_0) = \mathbf{0}$. A Taylor expansion of $\nabla L'_1(X_1)$ around X_0 is exact:

$$\nabla L'_1(X_1) = \nabla L'_1(X_0) + H'_1 \Delta X,$$

where $H'_1 := \nabla^2 L'_1(X)$ is the Hessian matrix, which is constant. Since $\nabla L'_1(X_1) = \mathbf{0}$, we have $H'_1 \Delta X = -\nabla L'_1(X_0)$. The gradient $\nabla L'_1(X_0)$ can be related to the unweighted problem's residuals. Let $L'^{(g)}(X) := \|\theta'_g - \alpha - BW_g\|_2^2$. Then,

$$\nabla L'_1(X_0) = \sum_{g=1}^G \omega_g \nabla L'^{(g)}_0(X_0) = \sum_{g=1}^G (\omega_g - 1) \nabla L'^{(g)}_0(X_0) + \sum_{g=1}^G \nabla L'^{(g)}_0(X_0)$$

Since $\sum_{g=1}^G \nabla L'^{(g)}_0(X_0) = \nabla L'_0(X_0) = \mathbf{0}$,

$$\nabla L'_1(X_0) = \sum_{g=1}^G (\omega_g - 1) \nabla L'^{(g)}_0(X_0).$$

The matrix of partial derivatives $\nabla L'^{(g)}_0(X_0)$ with respect to $X = (\alpha, B)$ is $-2(r'_g, r'_g W_g^\top)$. Let $G'_0 := -\nabla L'_1(X_0)$. Then $G'_0 = 2 \sum_{g=1}^G (1 - \omega_g)(r'_g, r'_g W_g^\top)$.

The Hessian H'_1 corresponds to the second derivatives of $L'(\alpha, B; \{\omega_g\})$. This objective can be viewed as a sum of k' independent weighted least squares problems. For each of these k' components, say j , the parameters could be written as $(\alpha_j, B_{j.})$, where α_j is the j -th component of the projected α and $B_{j.}$ is the j -th row of the projected B . The "regressors" for $(\alpha_j, B_{j.}^\top)$ are $Z_g = (1, W_g^\top)^\top$. The Hessian for each such scalar component problem is $2 \sum_{g=1}^G \omega_g Z_g Z_g^\top = 2GM$, where $M = \frac{1}{G} \sum_{g=1}^G \omega_g Z_g Z_g^\top$. For the matrix equation $H'_1 \Delta X = G'_0$, the operation $(H'_1)^{-1}$ effectively scales by the inverse of $2GM$ for each of the k' "rows" of parameters in ΔX . Thus, the smallest singular value of H'_1 relevant for $\|(H'_1)^{-1}\|_{\text{op}}$ is $\sigma_{\min}(H'_1) = 2G\lambda_{\min}(M)$, assuming M is positive definite. Therefore,

$$\|\Delta X\|_F = \|(H'_1)^{-1} G'_0\|_F \leq \|(H'_1)^{-1}\|_{\text{op}} \|G'_0\|_F = \frac{1}{2G\lambda_{\min}(M)} \|G'_0\|_F.$$

Next, we bound $\|G'_0\|_F$:

$$\|G'_0\|_F = \left\| 2 \sum_{g=1}^G (1 - \omega_g)(r'_g, r'_g W_g^\top) \right\|_F \leq 2 \sum_{g=1}^G |1 - \omega_g| \|(r'_g, r'_g W_g^\top)\|_F =$$

$$\begin{aligned}
&= 2 \sum_{g=1}^G |1 - \omega_g| \sqrt{\|r'_g\|_F^2 + \|r'_g W_g^\top\|_F^2} = 2 \sum_{g=1}^G |1 - \omega_g| \sqrt{\|r'_g\|_2^2 + \|r'_g W_g^\top\|_F^2} \\
&\leq 2 \sum_{g=1}^G |1 - \omega_g| \sqrt{\|r'_g\|_2^2 + \|r'_g\|_2^2 \|W_g\|_2^2} = 2 \sum_{g=1}^G |1 - \omega_g| \|r'_g\|_2 \sqrt{1 + \|W_g\|_2^2} \\
&\quad \left(\max_g \|r_g^*\|_2 \right) \sqrt{1 + M_W^2} \|\omega - 1\|_1.
\end{aligned}$$

where $M_W = \max_g \|W_g\|_2$, $\|\omega - 1\|_1 = \sum_{g=1}^G (1 - \omega_g)$. Combining the pieces we get the stated bound:

$$\|\Delta X\|_F \leq \frac{\sqrt{1 + M_W^2}}{\lambda_{\min}(M)} \left(\max_g \|r_g^*\|_2 \right) \frac{\|\omega - 1\|_1}{G}.$$

□

A.1.3 Proof of Proposition 3

Proof. The feasible MD estimator \hat{B}^{MD} can be decomposed based on its relationship with the oracle estimator B^* . We can write $\hat{B}^{MD} - B_0 = (\hat{B}_1^{MD} - B^*) + \hat{B}_0^{MD} + (B^* - B_0)$. We need to show that the first two components, when scaled by \sqrt{G} , vanish in probability:

$$(A) \quad \sqrt{G}(\hat{B}_1^{MD} - B^*) \xrightarrow{p} \mathbf{0}$$

$$(B) \quad \sqrt{G}\hat{B}_0^{MD} \xrightarrow{p} \mathbf{0}$$

Part (A): Proposition 2 provides a bound for the Frobenius norm of the bias component $\Delta B = \hat{B}_1^{MD} - B^*$:

$$\|\hat{B}_1^{MD} - B^*\|_F \leq K \frac{\sum_{g=1}^G (1 - \omega_g)}{G}$$

where $K = \frac{\sqrt{1+M_W^2}}{\lambda_{\min}(M)} (\max_g \|r_g^*\|_2)$. We assume K is $O_p(1)$ under standard regularity conditions (e.g., $M_W = \max_g \|W_g\|_2$ and $\max_g \|r_g^*\|_2$ are $O_p(1)$, and $\lambda_{\min}(M)$ is bounded away from zero in probability). Let $N_{excl} = \sum_{g=1}^G (1 - \omega_g)$ be the number of excluded groups. The expectation of N_{excl} is $\mathbb{E}[N_{excl}] = \sum_{g=1}^G \mathbb{P}[\omega_g = 0]$. By assumption we have $\sum_{g=1}^G \mathbb{P}[\omega_g = 0] \ll \sqrt{G}$. This implies $\mathbb{E}[N_{excl}] = o(\sqrt{G})$. By Markov's inequality, for any $\epsilon > 0$, $P(N_{excl} \geq \epsilon\sqrt{G}) \leq \mathbb{E}[N_{excl}]/(\epsilon\sqrt{G})$. Since $\mathbb{E}[N_{excl}]/(\epsilon\sqrt{G}) \rightarrow 0$, it follows that $N_{excl} = o_p(\sqrt{G})$. Therefore,

$$\|\hat{B}_1^{MD} - B^*\|_F = O_p(1) \cdot \frac{o_p(\sqrt{G})}{G} = o_p\left(\frac{1}{\sqrt{G}}\right)$$

This implies $\sqrt{G} \left\| \hat{B}_1^{MD} - B^* \right\|_F = o_p(1)$, and thus $\sqrt{G}(\hat{B}_1^{MD} - B^*) \xrightarrow{p} 0$.

Part (B): Assume $\frac{1}{G} \sum_{g=1}^G W_g \xrightarrow{p} \mu_W$ and $\frac{1}{G} \sum_{g=1}^G W_g W_g^\top \xrightarrow{p} \Sigma_{WW}$ for deterministic and bounded μ_W, Σ_{WW} . Also assume the asymptotic variance $\Sigma_{WW} - \mu_W \mu_W^\top > 0$. Finally, suppose W_g are uniformly bounded, i.e., $\|W_g\|_F \leq C_W$ for some constant C_W .

The component \hat{B}_0^{MD} and its associated intercept $\hat{\alpha}_0^{MD}$ solve:

$$(\hat{\alpha}_0^{MD}, \hat{B}_0^{MD}) = \underset{\alpha, B \text{ s.t. } \Gamma^\top \alpha = 0, \Gamma^\top B = 0}{\operatorname{argmin}} \sum_{g=1}^G \omega_g \|\varepsilon_g - \alpha - BW_g\|_2^2,$$

where $\varepsilon_g = \hat{\theta}_g - \theta_g$ is the first-stage estimation error satisfying $\mathbb{E}_{F_g}[\varepsilon_g | \omega_g = 1, \{\tilde{D}_{g,i}\}_{i=1}^{n_g}] = \mathbf{0}_{k'}$. This structure mirrors Proposition 1's GMM estimator, with ε_g replacing θ_g and $\omega_g I_k$ as the effective weighting matrix \tilde{A}_g . Let $\hat{\alpha}_0^{MD} = U^\top \hat{\alpha}_0^{MD}$ and $\hat{B}_0^{MD} = U^\top \hat{B}_0^{MD}$ be the transformed parameters. The "true" parameters for ε_g in this regression are $\mathbf{0}_{k'}$ and $\mathbf{0}_{k' \times p}$. The analogous normal equations are:

$$\begin{aligned} \left(\sum_{g=1}^G \omega_g I_{k'} \right) \hat{\alpha}_0^{MD} + \left(\sum_{g=1}^G \omega_g I_{k'} \hat{B}_0^{MD} W_g \right) &= \sum_{g=1}^G \omega_g U^\top \varepsilon_g \\ \left(\sum_{g=1}^G \omega_g (W_g \otimes I_{k'}) \right) \hat{\alpha}_0^{MD} + \left(\sum_{g=1}^G \omega_g (W_g W_g^\top \otimes I_{k'}) \right) \operatorname{vec}(\hat{B}_0^{MD}) &= \operatorname{vec} \left(\sum_{g=1}^G \omega_g U^\top \varepsilon_g W_g^\top \right). \end{aligned}$$

Let $\tilde{H}_{11}^\omega = \sum_{g=1}^G \omega_g I_{k'}$, $\tilde{H}_{12}^\omega = \sum_{g=1}^G \omega_g (W_g^\top \otimes I_{k'})$, $\tilde{H}_{21}^\omega = \sum_{g=1}^G \omega_g (W_g \otimes I_{k'})$, and $\tilde{H}_{22}^\omega = \sum_{g=1}^G \omega_g (W_g W_g^\top \otimes I_{k'})$. Let $C_1^{\varepsilon, \omega} = \sum_{g=1}^G \omega_g U^\top \varepsilon_g$ and $C_2^{\varepsilon, \omega} = \operatorname{vec} \left(\sum_{g=1}^G \omega_g U^\top \varepsilon_g W_g^\top \right)$.

First, establishing the probability limits of the Hessian components: Since $\frac{1}{G} \sum_{g=1}^G (1 - \omega_g) = o_p(1)$, then $\frac{1}{G} \sum_{g=1}^G \omega_g \xrightarrow{p} 1$. Thus, $\frac{1}{G} \tilde{H}_{11}^\omega = \left(\frac{1}{G} \sum_{g=1}^G \omega_g \right) I_{k'} \xrightarrow{p} I_{k'} =: H_{11,plim}^\omega$. For $\frac{1}{G} \tilde{H}_{21}^\omega = \frac{1}{G} \sum_{g=1}^G \omega_g (W_g \otimes I_{k'})$:

$$\begin{aligned} \left\| \frac{1}{G} \sum_{g=1}^G \omega_g (W_g \otimes I_{k'}) - \frac{1}{G} \sum_{g=1}^G (W_g \otimes I_{k'}) \right\|_F &= \left\| \frac{1}{G} \sum_{g=1}^G (1 - \omega_g) (-W_g \otimes I_{k'}) \right\|_F \\ &\leq \frac{1}{G} \sum_{g=1}^G (1 - \omega_g) \|W_g \otimes I_{k'}\|_F \\ &\leq \sqrt{k'} C_W \left(\frac{1}{G} \sum_{g=1}^G (1 - \omega_g) \right) \xrightarrow{p} 0. \end{aligned}$$

So, $\frac{1}{G} \tilde{H}_{21}^\omega \xrightarrow{p} \mu_W \otimes I_{k'} =: H_{21,plim}^\omega$. Similarly, $\frac{1}{G} \tilde{H}_{12}^\omega \xrightarrow{p} \mu_W^\top \otimes I_{k'} =: H_{12,plim}^\omega$, and $\frac{1}{G} \tilde{H}_{22}^\omega \xrightarrow{p}$

$\Sigma_{WW} \otimes I_{k'} =: H_{22,plim}^\omega$. The Schur complement $\tilde{S}^\omega := \tilde{H}_{22}^\omega - \tilde{H}_{21}^\omega (\tilde{H}_{11}^\omega)^{-1} \tilde{H}_{12}^\omega$ thus satisfies:

$$\frac{1}{G} \tilde{S}^\omega \xrightarrow{p} H_{22,plim}^\omega - H_{21,plim}^\omega (H_{11,plim}^\omega)^{-1} H_{12,plim}^\omega = (\Sigma_{WW} - \mu_W \mu_W^\top) \otimes I_{k'} =: S_{plim}^\omega,$$

where S_{plim}^ω is invertible by assumption.

Next, examining $\frac{1}{\sqrt{G}} C_1^{\varepsilon,\omega}$ and $\frac{1}{\sqrt{G}} C_2^{\varepsilon,\omega}$: Define $\xi_{1g} := \omega_g U^\top \varepsilon_g$ and $\xi_{2g} := \text{vec}(\omega_g U^\top \varepsilon_g W_g^\top)$. These terms have zero mean by definition of ε_g . Focusing on variances:

$$\begin{aligned} \mathbb{V} \left[\frac{1}{\sqrt{G}} C_1^{\varepsilon,\omega} \right] &= \frac{1}{G} \sum_{g=1}^G \mathbb{E}[\omega_g (U^\top \varepsilon_g \varepsilon_g^\top U)] = \frac{1}{G} \sum_{g=1}^G \mathbb{E} \left[\omega_g \frac{1}{n_g} \check{\Sigma}_{0,g} \right] \Rightarrow \\ \left\| \mathbb{V} \left[\frac{1}{\sqrt{G}} C_1^{\varepsilon,\omega} \right] \right\|_F &\leq \frac{C_{\Sigma'}}{G} \sum_{g=1}^G \frac{\mathbb{E}[\omega_g]}{n_g} = \frac{C_{\Sigma'}}{G} \left(\sum_{g=1}^G \frac{1}{n_g} - \sum_{g=1}^G \frac{\mathbb{P}[\omega_g = 0]}{n_g} \right) \end{aligned}$$

From Part (A) $\frac{1}{G} \sum \mathbb{P}[\omega_g = 0] = o(G^{-1/2})$. Thus, $\mathbb{V} \left[\frac{1}{\sqrt{G}} C_1^{\varepsilon,\omega} \right]$ has norm bounded by $O \left(\frac{1}{G} \sum \frac{1}{n_g} \right) + o(G^{-1/2})$. Since $\frac{1}{G} \sum \frac{1}{n_g} \rightarrow 0$ by Lemma 1, this variance converges to 0. Similarly, $\mathbb{V} \left[\frac{1}{\sqrt{G}} C_2^{\varepsilon,\omega} \right] \rightarrow 0$, as $\mathbb{E}[\xi_{2g} \xi_{2g}^\top] = \mathbb{E} \left[\omega_g (W_g \otimes I_{k'}) (\frac{1}{n_g} \check{\Sigma}_{0,g}) (W_g^\top \otimes I_{k'}) \right]$, which with bounded W_g implies the variance is $O \left(\frac{1}{G} \sum \frac{1}{n_g} \right)$.

Since $\frac{1}{\sqrt{G}} C_1^{\varepsilon,\omega}$ and $\frac{1}{\sqrt{G}} C_2^{\varepsilon,\omega}$ have zero mean and vanishing variances, they converge in probability to zero. Consequently:

$$\sqrt{G} \text{vec}(\hat{B}_0^{MD}) = \left(\frac{1}{G} \tilde{S}^\omega \right)^{-1} \left(\frac{1}{\sqrt{G}} C_2^{\varepsilon,\omega} - \left(\frac{1}{G} \tilde{H}_{21}^\omega \right) \left(\frac{1}{G} \tilde{H}_{11}^\omega \right)^{-1} \frac{1}{\sqrt{G}} C_1^{\varepsilon,\omega} \right) = o_p(1).$$

This implies $\sqrt{G} \hat{B}_0^{MD} \xrightarrow{p} \mathbf{0}$, and thus $\sqrt{G} \hat{B}_0^{MD} = \sqrt{G} U \hat{B}_0^{MD} \xrightarrow{p} \mathbf{0}$. □

A.1.4 Technical lemmas

Lemma 1. Let $\{n_g\}_{g=1}^G$ be a sequence of positive integers such that $n_g \geq 1$ for all g . Let c be a strictly positive constant ($c > 0$). If the condition

$$\lim_{G \rightarrow \infty} \frac{\sum_{g=1}^G e^{-cn_g}}{\sqrt{G}} = 0$$

holds, then it implies that

$$\lim_{G \rightarrow \infty} \frac{1}{G} \sum_{g=1}^G \frac{1}{n_g} = 0.$$

Proof. Assume that the implication does not hold. This means that there exists a $\delta > 0$ and a subsequence $G_k \rightarrow \infty$ such that for all k :

$$\frac{1}{G_k} \sum_{g=1}^{G_k} \frac{1}{n_g} \geq \delta \quad (\text{A.3})$$

Since $n_g \geq 1$, we have $1/n_g \leq 1$. Thus, $0 < \delta \leq 1$. Let M_0 be a constant chosen such that $M_0 > 1/\delta$.

For each G_k in the subsequence, partition the indices $\{1, \dots, G_k\}$ into two sets: $S_k = \{g \in \{1, \dots, G_k\} \mid n_g \leq M_0\}$ $L_k = \{g \in \{1, \dots, G_k\} \mid n_g > M_0\}$. Let $|S_k|$ denote the number of elements in S_k . So, $|L_k| = G_k - |S_k|$. From (A.3), we have:

$$\delta \leq \frac{1}{G_k} \left(\sum_{g \in S_k} \frac{1}{n_g} + \sum_{g \in L_k} \frac{1}{n_g} \right) \leq \frac{|S_k|}{G_k} \left(1 - \frac{1}{M_0} \right) + \frac{1}{M_0}$$

Let $f_k = |S_k|/G_k$ be the fraction of groups for which $n_g \leq M_0$. Rearranging the inequality:

$$f_k \left(\frac{M_0 - 1}{M_0} \right) \geq \delta - \frac{1}{M_0} = \frac{\delta M_0 - 1}{M_0}$$

Since we chose $M_0 > 1/\delta$, $\delta M_0 - 1 > 0$. Also, since $M_0 > 1$, $M_0 - 1 > 0$. Let $\eta = \frac{\delta M_0 - 1}{M_0 - 1} > 0$. For all k , $f_k = |S_k|/G_k \geq \eta$, which means $|S_k| \geq \eta G_k$, and thus a fixed positive fraction η of the groups have their n_g bounded by M_0 .

Now consider the sum $\sum_{g=1}^{G_k} e^{-cn_g}$:

$$\sum_{g=1}^{G_k} e^{-cn_g} \geq \sum_{g \in S_k} e^{-cn_g} \geq |S_k| e^{-cM_0} \geq \eta G_k e^{-cM_0} \Rightarrow \frac{\sum_{g=1}^{G_k} e^{-cn_g}}{\sqrt{G_k}} \geq \eta \sqrt{G_k} e^{-cM_0}$$

As $k \rightarrow \infty$, $G_k \rightarrow \infty$, so $\sqrt{G_k} \rightarrow \infty$. Therefore,

$$\lim_{k \rightarrow \infty} \eta \sqrt{G_k} e^{-cM_0} = \infty$$

This implies that $\liminf_{k \rightarrow \infty} \frac{\sum_{g=1}^{G_k} e^{-cn_g}}{\sqrt{G_k}} = \infty$ leading to a contradiction. \square

A.2 Examples

A.2.1 IV analysis

This section discusses an extension of our analysis to models where a policy variable of interest is endogenous, necessitating an instrumental variable (IV) approach. To illustrate concretely the endogenous weighting issues analogous to those discussed for GMM estimators in Section 3, we focus our discussion on the estimation of how local policies affect returns to schooling.

We start by augmenting the framework from Section 2. For each group g (e.g., a local labor market), assume the data-generating process for individual outcomes, F_g , can depend on the group-level instrument Z_g , i.e., $F_g = F_g(Z_g)$. A local policy of interest, W_g , is a group-specific function of Z_g , i.e., $W_g = W_g(Z_g)$. We assume that the triplet $(Z_g, W_g(\cdot), F_g(\cdot))$, is i.i.d. across groups and Z_g is independent of $W_g(\cdot)$ and $F_g(\cdot)$.

Consider individuals i in market g . We observe log wages $Y_{g,i}$ and years of schooling $S_{g,i}$. The local returns to schooling, τ_g , are determined by a Mincer-style equation:

$$Y_{g,i} = \delta_g + \tau_g S_{g,i} + u_{g,i}, \quad \mathbb{E}_{F_g}[u_{g,i}|Q_{g,i}] = 0, \quad (\text{A.4})$$

where potentially endogenous schooling $S_{g,i}$ (e.g., due to ability bias) is instrumented via $Q_{g,i}$ (e.g., quarter of birth). The data for group g are $\{(Y_{g,i}, S_{g,i}, Q_{g,i})\}_{i=1}^{n_g}$, drawn i.i.d. from F_g . For simplicity, we assume $n_g = n$. Model (A.4) fits within our general setup (1), with $\boldsymbol{\theta}_g := (\delta_g, \tau_g)^\top$.

The local returns to schooling τ_g are modeled as a function of an endogenous local policy W_g :

$$\tau_g = \alpha_g + \beta W_g,$$

where β is the structural effect of policy W_g on returns, and α_g captures unobserved market-level heterogeneity in returns, potentially correlated with W_g . We employ a group-level instrument Z_g for the policy W_g , which by assumption satisfies

$$\mathbb{E} \left[(Z_g - \mu_Z) \begin{pmatrix} 1 \\ \alpha_g \end{pmatrix} \right] = \mathbf{0}_2,$$

where $\mu_Z := \mathbb{E}[Z_g]$. A key premise is that the policy instrument Z_g (via W_g) may affect not only τ_g but also the distribution of schooling $S_{g,i}$ or its relationship with the instrument $Q_{g,i}$ within market g . In contrast, we assume the conditional distribution of $Q_{g,i}$ (given group g and individual characteristics other than $S_{g,i}$) does not directly depend on Z_g .

A common empirical strategy is then to estimate the equation:

$$Y_{g,i} = \delta_g + (\theta_0 + \beta W_g) S_{g,i} + \epsilon_{g,i},$$

via TSLS. The terms $S_{g,i}$ and $W_g S_{g,i}$ are treated as endogenous and instrumented using $Q_{g,i}$ and $Z_g Q_{g,i}$, respectively. Let β^{TSLS} denote the probability limit of this TSLS estimator for β as $G \rightarrow \infty$ and $n \rightarrow \infty$. One can show that it has the following representation:

$$\beta^{TSLS} = \beta + \frac{\mathbb{C}\text{ov}^{C_g(Z_g)}[\alpha_g, Z_g]}{\mathbb{C}\text{ov}^{C_g(Z_g)}[W_g, Z_g]},$$

where the weighted covariance relies on weights

$$C_g := \mathbb{C}\text{ov}_{F_g}[S_{g,i}, Q_{g,i}],$$

which generically depend on Z_g , and thus we write $C_g = C_g(Z_g)$.

The TSLS estimator β^{TSLS} is thus generically inconsistent for β . The bias arises if the C_g -weighted covariance between α_g (unobserved heterogeneity in returns) and Z_g is non-zero. This can occur even though the unweighted covariance, $\mathbb{C}\text{ov}[\alpha_g, Z_g]$, is zero by assumption, provided C_g is itself a function of Z_g . Such dependence of C_g on Z_g emerges if the policy instrument Z_g (via W_g) affects the joint distribution of $S_{g,i}$ and $Q_{g,i}$ (e.g., by altering schooling levels or the composition of individuals for whom $Q_{g,i}$ strongly predicts $S_{g,i}$). This introduces an endogenous weighting bias, analogous to the one discussed in Section 3.

These findings may seem at odds with design-based identification results for shift-share IV estimators (Adao et al., 2019; Borusyak et al., 2022) or, more generally, for "formula IV" estimators (Borusyak and Hull, 2023; Borusyak et al., 2024a), which show consistency of appropriately recentered IV estimators. The apparent contradiction is resolved by recognizing that our probability model allows the distribution of $S_{g,i}$ to vary with Z_g , whereas conventional design-based analyses often treat such characteristics or their distributions as fixed with respect to instrument assignment. If $S_{g,i}$'s distribution indeed covaries with Z_g , comprehensive recentering of Z_g requires conditioning on $S_{g,i}$. Standard correction by μ_Z is insufficient to eliminate the bias, as our analysis demonstrates.

Remark A.1. The discussion above focuses on GMM estimation. The MD estimation is standard in this case with one caveat: moments induced by (A.4) do not satisfy the restrictions in Section 4 and thus unbiased estimation of τ_g is generally impossible with finite n_g . This puts additional restrictions on the magnitude of n_g and may require using bias correction methods, analogous to those discussed in Remark 4.2.

A.2.2 Identifying Sorting and Heterogeneity Parameters

In this section, we discuss an extension of the example from Section 2.2.2 focusing on sorting and heterogeneity. Analyzing the full impact of the policy often requires looking beyond its effect on average premia $\psi_{g,k}$. For instance, does policy $W_{g,k}$ targeted at type k firms disproportionately attract high-ability workers ($\gamma_{g,i}$) to that type, or does it alter the dispersion of worker abilities within type k ?

Simplifying to two firm types ($j = 1, 2$), define a worker's mobility history

$$h(i) := (j(g, i, 1), j(g, i, 2)),$$

with four possible histories $h \in H = \{(1, 1), (2, 2), (1, 2), (2, 1)\}$. Using previous moment conditions we can identify the mean worker effect conditional on history, $\mu_{g,\gamma,h} = \mathbb{E}_{F_g}[\gamma_{g,i}|h(i) = h]$, which reveals patterns of worker sorting across different mobility paths. To identify the dispersion patterns we need stronger restrictions:

$$\mathbb{E}_{F_g}[\epsilon_{g,i,t}|\gamma_{g,i}, h(i)] = 0, \quad \mathbb{E}_{F_g}[\epsilon_{g,i,1}\epsilon_{g,i,2}|h(i)] = 0. \quad (\text{A.5})$$

The first condition strengthens the conditional moment restriction necessary for identifying the wage premium by conditioning on the worker effect $\gamma_{g,i}$, while the second assumes idiosyncratic errors are serially uncorrelated conditional on the mobility path. Armed with these restrictions we can identify the variance of worker effects conditional on history, $\sigma_{g,\gamma,h}^2 = \mathbb{V}_{F_g}[\gamma_{g,i}|h(i) = h]$, which captures the degree of heterogeneity within specific worker-path groups. Estimating the causal effect of a policy W_g on this expanded set of parameters allows for a more comprehensive analysis of the policy's impact on market structure. As detailed in Remark A.2 below, this full vector θ_g (including $\psi_{g,2}$, the four means $\{\mu_{g,\gamma,h}\}_h$, and the four variances $\{\sigma_{g,\gamma,h}^2\}_h$) is identified via a system of linear moment conditions. However, the reliance on worker mobility for identification persists, and the potential for the policy W_g to influence these mobility patterns remains a central challenge for estimation via either GMM or MD methods.

The extended analysis under assumptions (A.5) fits within our general moment-based framework $\mathbb{E}_{F_g}[h(D_{g,i}, \theta_g)] = 0_k$. This example highlights how identifying richer features of the economic structure often necessitates stronger assumptions on the underlying microdata process. Furthermore, it illustrates how the policy of interest W_g might affect the very variation (worker mobility) needed for identification, motivating the subsequent detailed analysis of estimation procedures in Sections 3 and 4.

Remark A.2 (Linear Moment Conditions for Full Vector). The full 9-dimensional parameter vector $\theta_g = (\psi_{g,2}, \{\mu_{g,\gamma,h}\}_h, \{\sigma_{g,\gamma,h}^2\}_h)$ for the $J = 2$ case is identified via a system of 9 linear

moment conditions consistent with the structure $\mathbb{E}_{F_g}[h_1(D_{g,i}) - h_2(D_{g,i})\boldsymbol{\theta}_g] = \mathbf{0}$. These arise from: (1) wage changes of movers 1 → 2 (identifying $\psi_{g,2}$); (2) average period-1 wages conditional on history h (identifying the four $\mu_{g,\gamma,h}$, given $\psi_{g,2}$); and (3) conditional covariance information $\sigma_{g,\gamma,h}^2 = \mathbb{E}_{F_g}[(Y_{g,i,1} - \mathbb{E}[Y_{g,i,1}|h(i) = h])Y_{g,i,2}|h(i) = h]$ for each history h (identifying the four $\sigma_{g,\gamma,h}^2$, linearly given the mean parameters).

A.2.3 Details about the TFP calculations

We now provide additional details about the TFP example discussed in Section 2.2.4. To simplify exposition, we focus on a single dynamic input, $L_{g,i,t}$ (the logic extends to multiple inputs with appropriate timing assumptions). The econometric specification for the production function is:

$$\begin{aligned} Y_{g,i,t} &= \theta_g^l L_{g,i,t} + \omega_{g,i,t} + \eta_{g,i,t}, \\ \omega_{g,i,t} &= \theta_g^0 + \theta_g^\omega \omega_{g,i,t-1} + \xi_{g,i,t}. \end{aligned}$$

The unobserved productivity component $\omega_{g,i,t}$ follows an AR(1) process. Key identifying assumptions for the $t \in \{1, 2, 3\}$ panel, particularly for estimation using data from $t = 3$, include:

- (a) The innovation to productivity $\xi_{g,i,3}$ is unpredictable by past inputs and output:

$$\mathbb{E}_{F_g}[\xi_{g,i,3}|L_{g,i,2}, L_{g,i,1}, Y_{g,i,1}] = 0.$$

- (b) The idiosyncratic error $\eta_{g,i,t}$ (e.g., measurement error or unanticipated shock) is contemporaneously uncorrelated with inputs and unpredictable by past variables:

$$\mathbb{E}_{F_g}[\eta_{g,i,3}|L_{g,i,2}, L_{g,i,1}, Y_{g,i,1}] = 0, \quad \mathbb{E}_{F_g}[\eta_{g,i,2}|L_{g,i,2}, L_{g,i,1}, Y_{g,i,1}] = 0.$$

Substituting $\omega_{g,i,t}$ from the AR(1) process into the production function and then expressing $\omega_{g,i,t-1}$ in terms of observables (i.e., $\omega_{g,i,2} = Y_{g,i,2} - \theta_g^l L_{g,i,2} - \eta_{g,i,2}$), we obtain an equation for $Y_{g,i,3}$:

$$Y_{g,i,3} = \theta_g^0(1 - \theta_g^\omega) + \theta_g^\omega Y_{g,i,2} + \theta_g^l L_{g,i,3} - \theta_g^l \theta_g^\omega L_{g,i,2} + \nu_{g,i},$$

where the composite error term is $\nu_{g,i} := \xi_{g,i,3} + \eta_{g,i,3} - \theta_g^\omega \eta_{g,i,2}$. Under assumptions (a) and (b), it follows that $\mathbb{E}_{F_g}[\nu_{g,i}|L_{g,i,2}, L_{g,i,1}, Y_{g,i,1}] = 0$. Letting $\tilde{\theta}_g^0 = \theta_g^0(1 - \theta_g^\omega)$ and $\tilde{\theta}_g^l := -\theta_g^l \theta_g^\omega$, the equation becomes:

$$Y_{g,i,3} = \tilde{\theta}_g^0 + \theta_g^\omega Y_{g,i,2} + \theta_g^l L_{g,i,3} + \tilde{\theta}_g^l L_{g,i,2} + \nu_{g,i}, \quad \mathbb{E}_{F_g}[\nu_{g,i}|L_{g,i,2}, L_{g,i,1}, Y_{g,i,1}] = 0.$$

Parameters $(\tilde{\theta}_g^0, \theta_g^\omega, \theta_g^l, \tilde{\theta}_g^l)$ are then identified using linear moment conditions with $L_{g,i,2}$, $L_{g,i,1}$, $Y_{g,i,1}$ (and a constant) as instruments.

This identification of $(\theta_g^l, \theta_g^\omega, \theta_g^0)$ implies that TFP, defined as

$$TFP_{g,i,t} = Y_{g,i,t} - \theta_g^l L_{g,i,t},$$

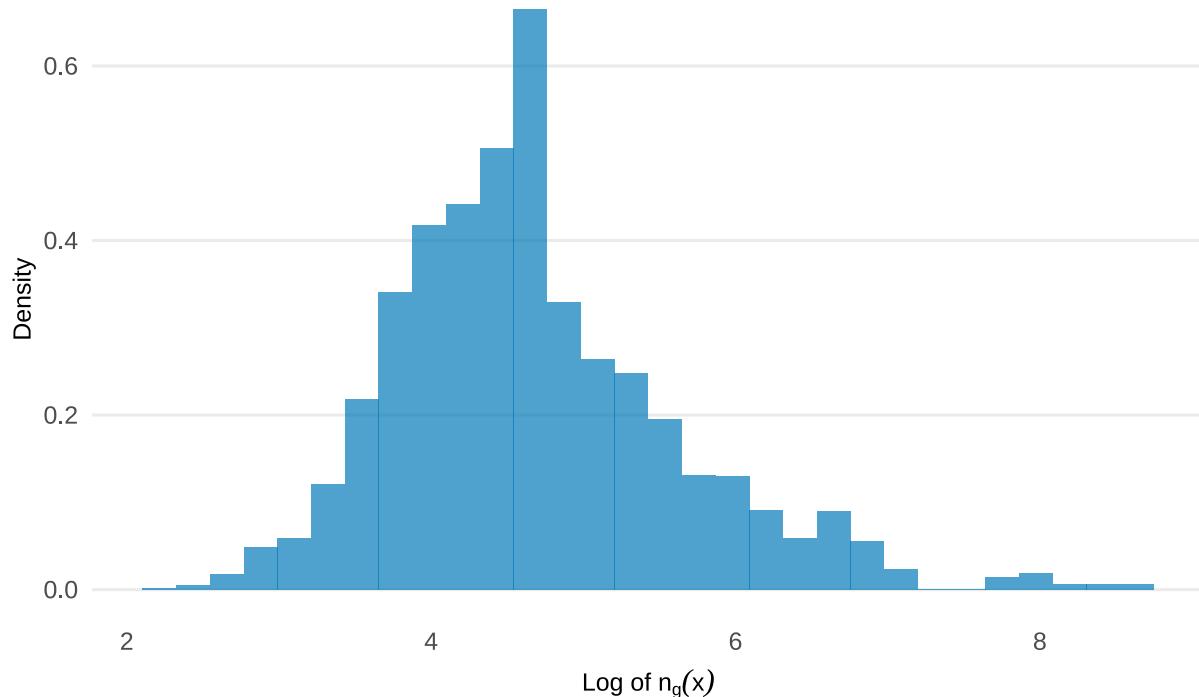
is identified. Key model-based parameters, $\delta_{g,t}$ and $\tau_{g,t}$, related to TFP growth, are then identified using moment conditions arising from the linear regression:

$$\Delta TFP_{g,i,t} = \delta_{g,t} + \tau_{g,t} \Delta FM_{g,i,t} + \epsilon_{g,i,t}, \quad \mathbb{E}_{F_{g,t}}[\epsilon_{g,i,t} | \Delta FM_{g,i,t}] = 0.$$

The structural parameters of the production function $(\theta_g^l, \theta_g^\omega, \theta_g^0)$ are separately identified from the coefficients in the GMM regression (e.g., θ_g^l is directly estimated, and θ_g^ω allows recovery of θ_g^0 from $\tilde{\theta}_g^0$).

B Figures and tables

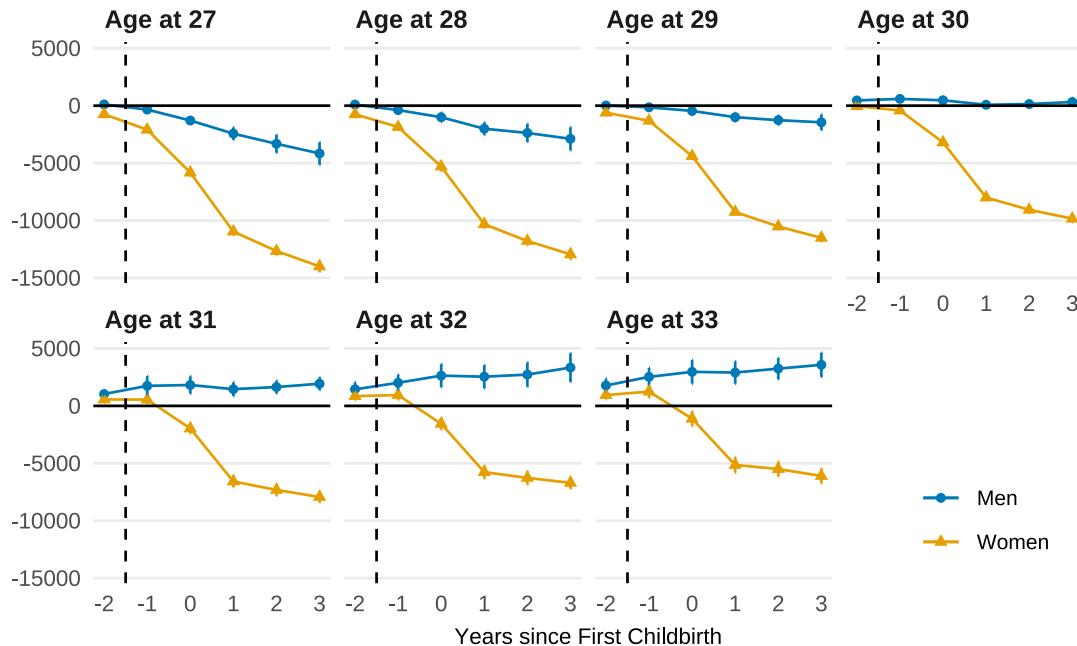
Figure B.1: Distribution of number of people with x at g : $n_g(x)$



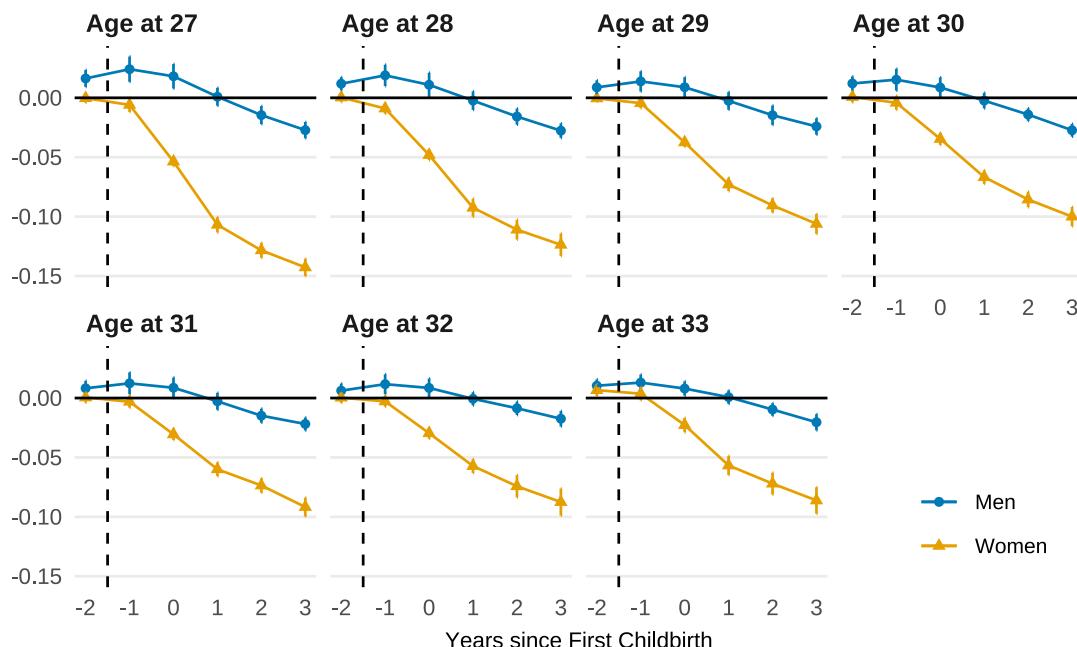
Notes: This figure presents the histogram of the population with covariates x (gender, birth year) who gave birth at municipality g . The x-axis shows the log of the population with x and g and y-axis shows its density. Due to the CBS confidentiality guidelines, $n_g(x) < 10$ are not included in the figure. The trimmed cells in the figure is 39 out of 4068 cells (0.96%).

Figure B.2: Child penalties by age at first childbirth

(a) Earnings



(b) Participation



Notes: This figure presents CP estimates for yearly earnings, aggregated by birth cohort, gender, municipality of residence at time of pregnancy, and the age of first childbirth, as described in Section 5.1. Figure B.2a reports the average CP for total yearly earnings. Similarly, Figure B.2b reports the CP for the labor-market participation margin.

Table B.1: Rich MD specification – the childcare provision expansion on CP (fathers)

(a) Earnings

	$\hat{\tau}_{g,-2}$	$\hat{\tau}_{g,-1}$	$\hat{\tau}_{g,0}$	$\hat{\tau}_{g,1}$	$\hat{\tau}_{g,2}$	$\hat{\tau}_{g,3}$
$CCI_{g,b+e-2}$	590.5 (397.8)	-148.5 (804.4)	189.5 (1113.7)	674.9 (1116.5)	1026.0 (1111.9)	1494.5 (1349.7)
$CCI_{g,b+e-1}$		1704.8** (592.7)	1532.7+ (804.7)	1550.1 (989.2)	940.2 (923.6)	638.2 (985.9)
$CCI_{g,b+e}$			1153.8* (560.5)	575.2 (666.6)	1394.8+ (709.5)	780.2 (840.9)
$CCI_{g,b+e+1}$				1209.2+ (720.0)	859.0 (718.8)	1976.2* (880.4)
$CCI_{g,b+e+2}$					1356.1* (675.0)	986.1 (981.8)
$CCI_{g,b+e+3}$						208.0 (850.9)
N	11,739	11,739	11,739	11,739	11,739	11,739
R^2	0.126	0.189	0.205	0.219	0.227	0.251
FE: Municipality g	X	X	X	X	X	X
FE: $B_{g,i} \times E_{g,i}$	X	X	X	X	X	X
FE: $(B_{g,i} \times E_{g,i})S_g$	X	X	X	X	X	X

+ p < 0.1, * p < 0.05, ** p < 0.01

(b) Participation

	$\hat{\tau}_{g,-2}$	$\hat{\tau}_{g,-1}$	$\hat{\tau}_{g,0}$	$\hat{\tau}_{g,1}$	$\hat{\tau}_{g,2}$	$\hat{\tau}_{g,3}$
$CCI_{g,b+e-2}$	-0.010 (0.010)	-0.008 (0.014)	-0.018 (0.014)	-0.010 (0.015)	-0.022 (0.015)	-0.007 (0.016)
$CCI_{g,b+e-1}$		0.000 (0.014)	0.012 (0.014)	0.008 (0.017)	0.013 (0.015)	-0.009 (0.015)
$CCI_{g,b+e}$			-0.025 (0.018)	-0.004 (0.019)	0.002 (0.018)	0.007 (0.014)
$CCI_{g,b+e+1}$				-0.015 (0.009)	-0.030* (0.012)	-0.014 (0.011)
$CCI_{g,b+e+2}$					0.010 (0.012)	0.003 (0.012)
$CCI_{g,b+e+3}$						-0.003 (0.011)
N	11,739	11,739	11,739	11,739	11,739	11,739
R^2	0.070	0.107	0.095	0.076	0.076	0.083
FE: Municipality g	X	X	X	X	X	X
FE: $B_{g,i} \times E_{g,i}$	X	X	X	X	X	X
FE: $(B_{g,i} \times E_{g,i})S_g$	X	X	X	X	X	X

+ p < 0.1, * p < 0.05, ** p < 0.01

Notes: These tables present the effect of the childcare provision expansion on child penalties (CP) of fathers in earnings (Panel (a)) and labor force participation (Panel (b)). See Section 5.3.2 for more details and Equation (12) for the specification. Parenthesis shows the clustered standard errors by municipality g .

C Data

Data Sources

We use administrative data from the Central Bureau of Statistics Netherlands (CBS) on the universe of Dutch residents. Different data sources, such as municipality registers or tax records, are matched through unique individual or household anonymized identifiers. The following section presents the main variables used and sample construction.

Tax and Employment Records Our primary data source is an extensive annual-level employer-employee data set derived from tax records (*baansommentab*) covering 1999 to 2016. We analyze two labor market outcomes: unconditional earnings and employment. Employment is specified as having a job based on an employment contract between a firm and a person, excluding self-employment. Second, earnings data consist of yearly gross earnings after social security contributions but before taxes and health insurance contributions from official tax data.

Demographic and Education Information To enrich our understanding of the workforce, we incorporate demographic data into our analysis (*gbapersoontab*). This includes birth year, date of death, sex, and annual information on the municipality of residence, household composition, marital status, and migration spells (*gbaadresobjectbus* and *vslgwbtab*). A unique aspect of our demographic data is the inclusion of a parent-child key (*kindoudertab*). We use information on birthdates and the linkage between parents and their children to determine the first child for all legal parents, which may include both adoptive and biological parents.

Childcare Provision Data An integral part of our study involves examining the role of childcare in labor market participation. To this end, we use records on childcare service providers using the firm's job classification (*betab*), and data on job location that we use to compute our index of childcare supply per municipality (given from *gemstplaatsbus/gemtplbus/ngemstplbus*). The job location data set contains each worker's municipality and firm ID, which we merge with the firm classification data.

Sample Definition

A key aspect of our study is the examination of labor market outcomes around the time of first childbirth. We restrict the sample to individuals born in 1993 or earlier to ensure we observe labor market outcomes at sufficiently mature ages. To capture transitions into parenthood, we include only those whose age at first birth was below 44, as observed from 2003 onward. To

ensure adequate labor market attachment before parenthood, we further restrict the sample to individuals who became parents at least 6 years after the typical graduation age for their highest educational attainment: 24 years for high school graduates, 26 years for vocational degree holders, and 27 years for those with a bachelor's degree. This approach provides a balanced panel of pre-parenthood labor market trajectories while minimizing censoring concerns.

D Dutch Childcare Act Reform

D.1 Key Features of the Reform

The *Dutch Childcare Act of 2005* (*Wet kinderopvang*) radically reformed childcare funding, provision, and regulation. It replaced a patchwork of local subsidies with a national, demand-driven system. The core expansion mechanisms included:

Demand-Side Funding. The Act shifted from supply-side grants to demand-led financing, giving parents direct subsidies or tax credits for childcare. Costs are now shared by parents, employers, and the state, with compulsory employer contributions (Lloyd, 2013). This tripartite funding greatly reduced out-of-pocket fees for families—on average cutting the effective parental fee by half over 2005–2009 (Bettendorf et al., 2015). All forms of formal care (daycare centers, childminders, after-school programs) became eligible for the same central subsidy, ending the old differentiation between “subsidized” and “unsubsidized” spots. Even informal care by relatives or licensed in-home providers (so-called *gastouder* or “guestparent” care) was brought under the subsidy scheme Lloyd (2013).

Market Liberalization. By decoupling funding from public provision, the Act introduced market forces into childcare. For-profit providers were explicitly allowed to enter and expand, competing on an equal footing with non-profits in a newly privatized market Noailly and Visser (2009). The expectation was that increased competition and profit incentives would spur new capacity. Indeed, following the reform, childcare supply grew fastest in high-demand areas and was led largely by for-profit firms, while non-profit providers saw their market share decline (especially in less affluent regions). This outcome reflected providers gravitating toward municipalities with greater demand and purchasing power, raising some concerns about equitable access in low-income or rural areas. Overall, however, the liberalization unlocked rapid growth in the number of childcare facilities nationwide Noailly and Visser (2009).

Regulatory Consistency. The Act sought to harmonize and simplify regulation of childcare. It introduced a single national quality framework (with light-touch regulation) applying to all providers Lloyd (2013). This replaced the prior mix of local rules and employer-based arrangements, ensuring consistent standards (e.g., safety, staff qualifications, and child-to-staff ratios) across the country. Notably, legally mandated child-to-staff ratios (set in 1996) remained in place to safeguard minimum quality Decree (1996), but administrative burdens were kept low to encourage new entrants. In essence, the reform fully privatized childcare provision but under a uniform set of basic rules. Subsequent legislation in 2010 further ce-

mented this consistency with a single statutory quality code for all early-childhood education and care services.

D.2 Implementation Timeline and Childcare Expansion

The Childcare Act took effect on **1 January 2005**, marking the start of a rapid expansion in Dutch childcare usage and supply. Key stages in its rollout include:

- **2005 – Transition Year:** The immediate impact was modest. The new funding system unified subsidies, slightly raising subsidies for previously “unsubsidized” parents and reducing them for the highest-income group, largely balancing out. Public spending on childcare actually dipped slightly in 2005, and the growth in childcare slots did not yet accelerate. Thus, no major change in labor supply was expected in this first year of reform implementation. The groundwork was laid, however, for broader participation: all formal childcare now qualified for support, and awareness of the new scheme grew among parents and providers.
- **2006–2007 – Surge in Generosity and Supply:** In the two years after 2005, the government dramatically increased the childcare subsidy rates. By 2007, the average parental co-payment share dropped from about 37% of the true cost to just 18%. Middle-income families saw subsidy rates rise by 20–40 percentage-points, and even higher-income families gained substantially. These changes effectively halved the cost of childcare for families within a short period. As a result, demand responded sharply: enrollment in formal childcare surged, and providers raced to expand capacity. The number of childcare places grew rapidly after 2006, with especially fast growth in daycare and out-of-school care participation. To support school-aged children’s care, a 2007 mandate required all primary schools to ensure out-of-school care was available (often by coordinating with childcare organizations).
- **Post-2007 – Consolidation and Further Growth:** Generous funding continued through the late 2000s. Public expenditure on childcare subsidies climbed from about €1.0 billion in 2004 to €3.0 billion by 2009, roughly 0.5% of GDP. The take-up of childcare subsidies expanded across all income groups (with low-income families eligible for nearly free care). The increased demand was met predominantly by private centers and child-minders entering the market under the light regulatory regime. By the end of 2009, the system had matured into a full-fledged childcare market with substantially higher coverage than pre-reform. The joint introduction of an expanded in-work tax credit for parents (the *combinatiekorting*) during 2005–2009 also provided additional incentive for parental employment [Bettendorf et al. \(2015\)](#).