Efficient Instrumental Variable Estimation under Sequential Missingness *

Jiqing Zhao

This version: October 2026

(Click here for the latest version)

Abstract

Missing data poses a challenge for instrumental variable estimation when endogenous variables and instruments are incompletely observed across survey modules, linked records, or multiple waves. I introduce a sequential missing-at-random condition that allows instrument availability to depend on observed endogenous variables and permits dependence among instrument observation indicators. This relaxes the common restriction that instrument availability depends only on the always observed variables. Under this condition, I establish identification and develop a two-step inverse probability weighted GMM estimator that combines available moments across observation patterns using sequential inverse probability weights. I establish consistency and asymptotic normality and derive feasible weighting matrices that minimize asymptotic variance within the proposed estimator class. Monte Carlo simulations illustrate the consequences of misspecified observation models and the benefits of flexible estimation. An application to family size and schooling using an extended sample from the Indonesia Family Life Survey finds statistically significant evidence that larger sibling counts reduce children's schooling under the maintained identifying assumptions. The resulting framework generalizes IPW-GMM estimation under weaker restrictions on data availability.

*Jiqing Zhao, University of Texas at Austin. Email: jiqingzhao@utexas.edu. I would like to thank my advisor, Jason Abrevaya; my committee members, Stephen Donald, Brendan Kline, and Chris Muris; other econometrics faculty members at UT Austin, Daniel Ackerberg, Xinyue Bei, and Haiqing Xu; and seminar participants for their support and helpful advice. All errors are my own.

1 Introduction

The problem of missing data is prevalent in economic studies across nearly all fields. According to Abrevaya and Donald (2017), at least 40% of empirical publications in top economics journals encounter the challenge of missing data, with 70% of these studies opting to address this by discarding all observations with missing values (complete-case analysis). However, this approach can lead to both efficiency loss and potential bias. This problem becomes even more complex with multidimensional missing data, where covariates, instruments, or other variables are missing simultaneously.

This paper shows how instrumental-variable (IV) models can be identified and estimated when instrument availability depends on endogenous variables that are themselves missing for part of the sample. These models relate an observed outcome to explanatory variables that may be correlated with unobserved effect of the outcome, using instruments to identify their causal effects. I allow both endogenous variables and multiple instruments to be missing, while the outcome and baseline controls are observed. The proposed estimator combines the moment conditions available under different instrument patterns. I establish asymptotic normality and derive optimal weights for combining these moments within the proposed class of estimators.

This setting is relevant when information is assembled from different survey modules, linked records, survey waves, or administrative sources. These sources can create distinct layers of missing information. The empirical application studies the effect of family size on completed schooling using the Indonesia Family Life Survey, following Chaudhuri and Guilkey (2016). Child records provide schooling and demographics, maternal links provide the number of siblings, and community records provide family planning instruments. Some children have no usable sibling count, while others have a sibling count but only a subset of the instruments. The application examines how accounting for both sources of missingness changes the estimated relationship between family size and schooling.

I use a sequential missing-at-random (SMAR) condition to model these two observation layers. The availability of the endogenous variables is assumed to be independent of their values and the instrument values after conditioning on the always observed information. Among observations with the endogenous variables available, instrument availability may additionally depend on those variables, but not further on the instrument values. In the application, this permits the instrument observation pattern to depend on the number of siblings whenever that count is observed. The ordering specifies the information used to explain each observation process and does not need to coincide with the time or waves of the underlying interviews.

The inverse-probability-weighted GMM estimator groups observations according to which instruments are observed and uses the moment conditions that can be evaluated in each group. It reweights the computable products of instruments and regression errors to recover the relationships that would be available with complete data. The inverse probability weights account jointly for the availability of the endogenous variables and the relevant instrument pattern. Under the identifying conditions, information from different patterns can identify the model even when no individual pattern does so alone. The optimal combination of these moments accounts for their precision and correlation. The large sample analysis incorporates estimation of the observation probabilities, and Monte Carlo simulations illustrate the consequences of ignoring dependence on the endogenous variables and the performance of flexible probability estimation.

Early research established the fundamentals of missing data and the limitations of ad hoc fixes. Dagenais (1973) discusses linear model performance in the presence of missing data, while Rubin (1976) formally defines standard missingness patterns. Subsequent studies formalized why simple corrections, such as complete-case analysis or the missing-indicator approach, often yield biased or inefficient results, particularly when missingness patterns are complex (Gourieroux and Monfort, 1981; Little and Rubin, 2002; Donders et al., 2006).

To address these shortcomings, the econometric literature increasingly relies on moment-based estimators under the MAR assumption. A rich line of work focuses on handling missing covariates and instruments in estimating equations using Generalized Method of Moments (GMM) frameworks (Mogstad and Wiswall, 2012; Abrevaya and Donald, 2017; Coe, 2018; Rai, 2023). Within this literature, recent papers have developed more sophisticated IPW-GMM estimators to handle arbitrary strata of missingness (Chaudhuri and Guilkey, 2016; Kennedy et al., 2019; Muris, 2020). Related work studies missing endogenous variables in conditional moment restriction models (Cosma, Kostyrka, and Tripathi, 2026). These frameworks impose MAR restrictions under which the observation probabilities depend on variables observed for everyone. This paper allows instrument-pattern probabilities to depend additionally on endogenous variables within the subsample where those variables are observed.

A related strand of literature models missingness through sequential observation mechanisms. Robins and Gill (1997) study randomized monotone missingness, while Zhou, Little, and Kalbfleisch (2010) develop block-conditional missing-at-random models. Yang (2026) introduces a sequential MAR assumption with asymmetric conditioning sets, and Chaudhuri (2020) quantifies efficiency gains under monotone missingness. Yang's augmented estimator uses additive separability of the moments across the two partially observed variables. The IV moments studied here generally contain products involving both

missing-data layers and cannot be evaluated directly when the endogenous variables are unavailable. The proposed IPW estimator addresses this joint dependence and does not require the additional restriction that Yang imposes on the second observation process when the first variable is missing.

Imputation provides an alternative path for missing data, ranging from traditional linear imputation (Gourieroux and Monfort, 1981) to modern machine learning techniques such as K-nearest neighbors (Batista and Monard, 2002) and Multivariate Imputation by Chained Equations (Azur et al., 2011). McDonough and Millimet (2017) examine imputation in models with endogenous regressors, illustrating the importance of accounting for endogeneity when handling missing data. Imputation approaches specify models for the missing variables. This paper focuses on recovering the identifying moments through models of observation probabilities, and does not pursue imputation-based estimators.

Ultimately, this paper connects the literature on missing data in moment condition models with the literature on sequential nonresponse. Building on the semiparametric two step literature, including Hahn (1998), Hirano, Imbens, and Ridder (2003), Cattaneo (2010), and Graham (2011), I provide a framework for multidimensional missing data in IV models. The paper makes five contributions.

First, I introduce a sequential missing at random condition for IV models in which both the endogenous variables and the instruments can be missing. The condition follows the order in which information becomes available and permits the observation probability of the instruments to depend on an endogenous variable once that variable has been observed. This dependence is ruled out by conventional MAR specifications that condition only on variables observed for everyone. I show that the resulting observed data have a nested information structure. Although the raw instrument patterns can be nonmonotone, no IV moment can be evaluated before the endogenous variable is observed. Under SMAR, inverse weighting recovers the complete data moments, and moments from different instrument patterns can jointly identify the structural parameter even when no single pattern contains all instruments and no individual stratum identifies the model on its own.

Second, I propose a two step IPW GMM estimator that uses every observation pattern contributing a feasible moment. The first step estimates the probability that the endogenous variable is observed and the conditional probabilities of the instrument patterns. The second step weights each feasible stratum moment by the inverse product of these probabilities and combines the moments into a unified system. The instrument pattern probabilities are estimated jointly by a multinomial model, so the procedure does not require the missingness indicators of the instruments to be conditionally independent. The

first step can also be estimated flexibly, which allows the observation probabilities to be nonlinear functions of the conditioning variables.

Third, I establish asymptotic normality and derive optimal weighting matrices for the proposed estimator. I derive its asymptotic linear representation and asymptotic variance when the observation probabilities are estimated in the first step. The sequential structure produces a variance decomposition with three mutually uncorrelated components: variation explained by the always observed variables, additional variation associated with observing the endogenous variable, and additional variation associated with the instrument pattern. This decomposition clarifies how each missingness layer enters inference. I also compare the feasible estimator with the unaugmented estimator that inserts the true observation probabilities and show how estimating sufficiently flexible observation models can remove part of the variation generated by selection. Using the joint covariance of the adjusted moments across instrument patterns, I derive weights that minimize asymptotic variance within the proposed class of IPW GMM estimators. These weights account for correlations across patterns as well as differences in their precision. I also provide a feasible implementation based on consistent estimators of the required covariance and derivative matrices.

Fourth, I study the method in a Monte Carlo experiment and an empirical application. The simulation considers nonlinear observation probabilities and a severe rate of incompleteness. The flexible SMAR estimator becomes centered near the true structural parameter and approaches the true probability benchmark as the sample size grows, while complete case analysis, conventional MAR weighting, and a restricted parametric first step become concentrated around different probability limits. The simulation also documents the finite sample cost of flexible inverse weighting when overlap is weak.

Fifth, in the empirical application, I revisit the effect of family size on completed schooling in Indonesia, following the IV design of Chaudhuri and Guilkey (2016). The application extends their setting by allowing the number of siblings to be missing for part of the child sample while the family planning instruments are observed in different overlapping patterns. It illustrates how the proposed procedure can be implemented without changing the underlying causal design.

The remainder of the paper proceeds as follows: Section 2 introduces the IV model and the pattern of missingness used in this paper; Section 3 establishes identification and states the necessary identifying assumptions; Section 4 presents the two-step IPW-GMM estimation strategy; Section 5 studies the asymptotic theory of the proposed estimator; Section 6 presents the Monte Carlo simulation design and results; Section 7 applies the method to an empirical example and studies the effect of the number of siblings on a

child's education using the Indonesia Family Life Survey; and Section 8 concludes.

2 Model

2.1 The IV Model Setup

Consider the following outcome equation in an IV model:

$$Y = D'\alpha_0 + X'\beta_0 + \epsilon, \qquad E[\epsilon|X, Z] = 0 \tag{1}$$

where Y is the dependent variable; $D \in \mathbb{R}^p$ is a vector of p endogenous variables, where $p > 0$; $X \in \mathbb{R}^q$ is a vector of q exogenous control variables, where $q \geqslant 0$; and ϵ is the corresponding error term. In addition, let $Z \in \mathbb{R}^k$ denote the set of k ($k \geqslant p$) instruments. α_0 is the vector of the parameters of interest.

Write $\theta = (\alpha', \beta')' \in \Theta \subset \mathbb{R}^{p+q}$, $V = (Y, X')'$ for the always-observed variables, $W = (Y, X', D', Z')'$ for the complete data, and

$$\tilde{\epsilon}(\theta) := Y - D'\alpha - X'\beta, \qquad \tau(\theta) := \begin{pmatrix} Z\tilde{\epsilon}(\theta) \\ X\tilde{\epsilon}(\theta) \end{pmatrix} \in \mathbb{R}^{k+q} \tag{2}$$

Under Equation (1), $\tilde{\epsilon}(\theta_0) = \epsilon$ and the complete-data moment condition is $E[\tau(\theta_0)] = 0$. Since the model is linear, we have:

$$\tau(\theta) = t - T\theta, \qquad t := \begin{pmatrix} ZY \\ XY \end{pmatrix}, \qquad T := \begin{pmatrix} Z \\ X \end{pmatrix} \begin{pmatrix} D' & X' \end{pmatrix} \tag{3}$$

Let $G := E[\partial\tau/\partial\theta']$, then from Equation (3), we have $G = -E[T]$ and $E[\tau(\theta)] = G(\theta - \theta_0)$ for every θ.

2.2 Missingness

In this paper, we consider the case where the endogenous variables D and the instrumental variables Z can be sometimes-missing. Following the notation defined in Section 2.1, $V = (Y, X')'$ denotes the always-observed variables. Let $g_D \in \{0, 1\}$ be the data availability

indicator for D, which equals 1 if every element in D is observed and 0 otherwise, i.e.,

$$g_D = \begin{cases} 1, & D \text{ is fully observed} \\ 0, & \text{some element of } D \text{ is missing} \end{cases}$$

Let $\lambda_Z \in \{s_1, \cdots, s_J\}$ denote the observed subset of instruments with $s_j \subseteq \{1, \cdots, k\}$. All the different strata form a partition, which means that, with

$$g_{Z,j} := \mathbb{1}\{\lambda_Z = s_j\}$$

we have $\sum_{j=1}^{J} g_{Z,j} = 1$, and the empty stratum $s_j = \emptyset$ is allowed.

Let $S_j \in \mathbb{R}^{k_j \times k}$ ($k_j := |s_j|$) be the selection matrix selecting the components of Z indexed by s_j and I_N denote the $N \times N$ identity matrix. With $S_j S_j' = I_{k_j}$, and

$$\tilde{S}_j := \begin{bmatrix} S_j & 0 \\ 0 & I_q \end{bmatrix} \in \mathbb{R}^{(k_j+q)\times(k+q)}$$

we have that $\tilde{S}_j \tau(\theta)$ is the subvector of moments that we can compute for an observation from stratum j when $g_D = 1$. When $s_j = \emptyset$, this still holds since the moment conditions reduce to the exogenous moments only, i.e., $\tilde{S}_j \tau(\theta) = X\tilde{\epsilon}(\theta)$.

In a 2-instrument example, the notations for the missingness can be simplified to the following:

Example 1. *Consider a 2-instrument ($k = 2$) case. In addition, let $q = 1$ for simplicity, the four possible strata s_j can be expressed as:*

$$s_1 = \{1, 2\}, \qquad s_2 = \{1\}, \qquad s_3 = \{2\}, \qquad s_4 = \emptyset$$

Here, s_1 denotes the stratum where both instruments Z_1 and Z_2 are observed; s_2 denotes the stratum where only instrument Z_1 is observed; s_3 denotes the stratum where only instrument Z_2 is observed; and s_4 denotes the stratum where neither Z_1 nor Z_2 is observed.

The selection matrices $S_j \in \mathbb{R}^{k_j \times 2}$ can be expressed as:

$$S_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad S_2 = \begin{bmatrix} 1 & 0 \end{bmatrix}, \qquad S_3 = \begin{bmatrix} 0 & 1 \end{bmatrix}, \qquad S_4 = (\ ) \in \mathbb{R}^{0\times 2}$$

with $k_j = 2, 1, 1, 0$, respectively.

The augmented selection matrices $\tilde{S}_j \in \mathbb{R}^{(k_j+1)\times 3}$ *can be written as:*

$$\tilde{S}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \tilde{S}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \tilde{S}_3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \tilde{S}_4 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

And hence the moment subvectors that we can compute when $g_D = 1$ *are*

$$\tilde{S}_1\tau(\theta) = \begin{bmatrix} Z_1\tilde{\epsilon}(\theta) \\ Z_2\tilde{\epsilon}(\theta) \\ X\tilde{\epsilon}(\theta) \end{bmatrix}, \qquad \tilde{S}_2\tau(\theta) = \begin{bmatrix} Z_1\tilde{\epsilon}(\theta) \\ X\tilde{\epsilon}(\theta) \end{bmatrix}, \qquad \tilde{S}_3\tau(\theta) = \begin{bmatrix} Z_2\tilde{\epsilon}(\theta) \\ X\tilde{\epsilon}(\theta) \end{bmatrix}, \qquad \tilde{S}_4\tau(\theta) = \begin{bmatrix} X\tilde{\epsilon}(\theta) \end{bmatrix}$$

We now present the sequential missing-at-random (SMAR) assumption that we use throughout this paper. The name for such a missingness pattern is borrowed from Yang (2026), but the version in this paper is weaker since we do not impose any restriction on λ_Z when D is not observed, i.e. $g_D = 0$.

Assumption 1 (SMAR)**.**

(i). $g_D \perp (D, Z)|V$

(ii). $\lambda_Z \perp Z|(D, V, g_D = 1)$

Assumption 1(i) says that missingness of the endogenous variables is ignorable given the outcome and the exogenous covariates; Assumption 1(ii) says that, among observations with D observed, which instruments are observed is ignorable given the outcome, exogenous covariates, and the endogenous variables. The distinctive feature is that Assumption 1(ii) conditions on D, which is a variable that is itself missing for part of the sample. This is what conventional MAR frameworks in the econometrics literature (e.g., Chaudhuri and Guilkey, 2016; Muris, 2020) rule out.[1]

The SMAR assumption is a selection process that follows the nested information structure $\sigma(V) \subseteq \sigma(D, V) \subseteq \sigma(W)$. The first part of SMAR uses the baseline information; the second part additionally conditions on D on where it is observed. And while the information is monotone, the observed pattern of instruments within themselves can be nonmonotone.

The analysis treats the outcome and baseline controls as observed. This provides the information used to model first-layer availability throughout the sample. Missing outcomes

[1]Here, conventional MAR refers to the common-conditioning-variable restrictions used in the comparison estimators, which set the missingness to condition on always observed variables. This should be distinguished from the general definition of MAR, which allows the conditioning information to be all of the observed information and to vary across observation patterns (See Rubin, 1976; Seaman et al., 2013; Sun and Tchetgen Tchetgen, 2018).

or controls would change both this conditioning information and the moment subvectors that can be computed. Such an extension requires a corresponding observation model and a new moment-recovery argument. Grouping a missing control with another data layer is possible only after those changes are specified.

Such a pattern is very common when different layers of information becomes available through different survey modules, linked records, waves, or administrative sources. To generalize, a data collecting process that follows some order may satisfy SMAR. And the order here does not need to be time or waves alone. Example 2 below is an illustrative example of what SMAR in dataset may look like:

Example 2 (Family size and schooling)**.** *Consider a study of number of siblings* (D) *and children's schooling* (Y)*, with schooling and baseline characteristics* (X) *observed for everyone. A fertility module provides sibling counts, while a separate residential-history module supplies information for linking family-planning instruments* (Z) *from community records. These modules create two distinct layers of missingness.*

Suppose the fertility module is administered to a random subsample within groups defined by the always observed information, with complete responses from selected households. Sibling-count availability is then independent of family size and instrument values within these groups, satisfying SMAR (i).

The residential-history module is offered to all households. Parents with more children may face greater reporting burdens, leaving different sections incomplete and different subsets of instruments available. Among households with observed sibling counts, suppose the resulting instrument pattern is independent of instrument values conditional on family size and the always observed information. This satisfies SMAR (ii), which allows instrument availability to depend on family size and to be correlated across instruments.

The ordering is plausible since fertility-module selection depends only on the always observed information, whereas residential-history completion may also depend on family size.

Define the two data availability probabilities and their product:

$$p_D(V) := P(g_D = 1|V),\ \ p_{Z,j}(D,V) := P(\lambda_Z = s_j|D,V,g_D = 1),\ \ p_j(D,V) = p_D(V)p_{Z,j}(D,V)$$

Under SMAR, the joint probability of data availability can be recovered from the observed

part of the sample:

$$\begin{aligned}
&P(\lambda_Z = s_j, g_D = 1|V, Z, D) \\
=&P(\lambda_Z = s_j|V, Z, D, g_D = 1)P(g_D = 1|V, Z, D) \\
=&p_{Z,j}(D, V)p_D(V) \\
=&p_j(D, V)
\end{aligned}$$

Note that the order of SMAR matters. It is a restriction of the conditioning information, not merely an algebraic choice of factors. Reversing the order of D and Z would require different conditional independence restrictions and different observed samples for the probability model. Yet, if by the nature of data collecting process, the order does switch, the calculation of the joint probability itself proceeds similarly. Again, echoing to our Example 2, the order does not need to coincide with the calendar dates at which the underlying variables were measured.[2]

Remark 1. *(Why SMAR): Since the estimator below uses only the product $p_j(D, V)$, a question would be why we do not simply assume that the joint probability $P(\lambda_Z = s_j, g_D = 1|V, Z, D)$ equals some identified function of (D, V), leaving the missingness unrestricted. We impose SMAR for two reasons. First, recoverability is a conclusion rather than a primitive: the probability above conditions on D, so it cannot in general be estimated from the observed data. SMAR derives recoverability from conditions stated in terms of the data-collection process, which can be assessed against the survey design. Second, SMAR is constructive: it splits the joint probability into pieces that each condition only on variables observed for the data points used to estimate them, whereas a high-level recoverability condition implies no estimation strategy.*

3 Identification

This section establishes identification under SMAR. We first construct the observed-data moment conditions and state the rank, coverage, and overlap assumptions. The IPW identities recover their complete-data expectations. A rank argument then shows when the stacked system identifies the parameter and when combining its rows preserves that identification.

[2]Suppose that we switch the order of the two missingness layers, then the first layer would become the pattern of instrument availability is independent of the values of the instruments themselves and the endogenous variables conditional on the always-observed variables (controls and the outcome). The second layer would become that the availability of endogenous variables all at once is independent of the values of themselves once conditioned on some observed pattern of instruments, along with the always-observed variables.

3.1 Moment Conditions

Let $w_j := \frac{g_D g_{Z,j}}{p_D(V) p_{Z,j}(D,V)}$. For stratum j, the IPW moment condition can be expressed as:

$$E[w_j \tilde{S}_j \tau(\theta_0)] = 0, \qquad j = 1, \cdots, J \tag{4}$$

Let $A_j \in \mathbb{R}^{(p+q)\times(k_j+q)}$, $j = 1, \cdots, J$, be a set of conformable weighting matrices so that the weighted sum of the IPW moments forms a single system of dimension $p+q$, i.e.,

$$m(\theta) := \sum_{j=1}^{J} A_j w_j \tilde{S}_j \tau(\theta) \tag{5}$$

In addition, let $\bar{A} := \sum_{j=1}^{J} A_j \tilde{S}_j \in \mathbb{R}^{(p+q)\times(k+q)}$.

3.2 Assumptions

Beyond Assumption 1, we impose additional assumptions needed for identification in this subsection. We assume the data satisfy the following conditions: $(W_i, g_{D,i}, \lambda_{Z,i})_{i=1}^n$ is an i.i.d. sample; and $(V_i, g_{D,i}, g_{D,i}D_i, g_{D,i}g_{Z,ij}, g_{D,i}S_{\lambda_{Z,i}}Z_i)_{i=1}^n$ is observed.

Assumption 2 (Identification of the complete-data model)**.** *rank*$(G) = p+q$*, equivalently* $E[(Z', X')'(D', X')]$ *has full column rank.*

Assumption 3 (Coverage)**.** *For every instrument* $\kappa \in \{1, \cdots, k\}$*, there exists a stratum* j *with* $\kappa \in s_j$ *and* $P(g_D = 1, \lambda_Z = s_j) > 0$*.*

Assumption 4 (Overlap)**.** *There exists* $\delta \in (0, 1)$ *such that* $p_D(V) \geqslant \delta$ *a.s. and* $p_{Z,j}(D, V) \geqslant \delta$ *a.s. for every stratum with* $P(\lambda_Z = s_j | g_D = 1) > 0$*.*

Assumption 5 (Moments)**.** $E\|t\|^2 < \infty$ *and* $E\|T\|^2 < \infty$*.*

Assumption 6 (Weighting Matrices)**.** *The collection* $\{A_j\}_{j=1}^J$ *satisfies the condition that* $\bar{A}G = \sum_{j=1}^{J} A_j \tilde{S}_j G$ *is nonsingular.*

Assumption 2 is the standard setup for IV models, which requires $k \geqslant p$. It separates identification of the complete-data model from the additional conditions needed to recover its moments under missingness. See Hansen (1982) for the GMM framework. Assumption 3 says that while we do not require that all variables be observed simultaneously with positive probability, the probability of each instrument being observed along with D should be positive so that each instrument plays a role in identification rather than completely dropping out. This is also a core assumption in Muris (2020). See Remark 3

for an extended discussion. Assumption 4 ensures that the inverse probability weights are bounded, so extreme weights cannot dominate the moment conditions. Similar assumption can be found in Wooldridge (2007). Assumption 5 imposes standard second-moment bounds, ensuring the expectations in Equation (3) and Equation (4) exist (Hansen, 1982). Assumption 6 is the usual nonsingularity condition for the matrix inverted below. Optimal combinations subsequently uses the GMM covariance criterion (Hansen, 1982). Under Assumptions 2 and 3, matrices $\{A_j\}$ satisfying Assumption 6 exist even when no single stratum identifies θ_0.

3.3 Main Identification Result

The inverse propensity weighting in the moment conditions works in the following way:

Lemma 1 (IPW identities)**.** *Let $h(W)$ be any integrable function of W. Under Assumptions 1 and 4:*

$$E[g_D g_{Z,j} \mid W] = p_D(V)p_{Z,j}(D,V) = p_j \quad a.s., \text{ for every } j \tag{6}$$

$$E[g_D h(W) \mid V] = p_D(V)E[h(W) \mid V] \quad a.s. \tag{7}$$

$$E[g_D g_{Z,j} h(W) \mid D,V] = p_j E[h(W) \mid D,V] \quad a.s., \text{ for every } j \tag{8}$$

Consequently,

$$E[(g_D/p_D)h(W)] = E[h(W)] \tag{9}$$

$$E[w_j h(W)] = E[h(W)] \tag{10}$$

Lemma 1, combined with the complete-data moment condition, establishes Equation (4) under Assumptions 1 and 4. The main takeaway is:

$$E[g_D g_{Z,j}|W] = p_D(V)p_{Z,j}(D,V) = p_j \quad a.s. \tag{11}$$

so that the inverse weighting by the product of the two data availability probabilities removes the sequential selection bias.

Lemma 2 (Rank)**.** *Suppose Assumptions 2 and 3 hold. Then*

$$rank \begin{pmatrix} \tilde{S}_1 G \\ \vdots \\ \tilde{S}_J G \end{pmatrix} = p + q \tag{12}$$

Consequently, there exists a sequence $\{A_j\}_{j=1}^J$ *that satisfies Assumption 6. One constructive choice of* A_j *is* $A_j = (\tilde{S}_j G)'$.

Lemma 2 gives a constructive way to preserve the rank of the stacked system. And arbitrary matrices A_j may not achieve this, since contribution from each stratum can cancel through the linear combination of A_j.

Now we establish identification as follows:

Theorem 1 (Identification)**.** *Under Assumptions 1 to 5:*

(a) θ_0 *is the unique solution to the stacked moment conditions*

$$E[w_j \tilde{S}_j \tau(\theta)] = 0$$

for every j *with* $P(g_D = 1, \lambda_Z = s_j) > 0$.

(b) For any fixed matrices $\{A_j\}_{j=1}^J$,

$$E[m(\theta)] = \bar{A}G(\theta - \theta_0) \qquad \text{for all } \theta.$$

Hence, $E[m(\theta)] = 0$ *has* θ_0 *as its unique solution if* $\bar{A}G$ *is nonsingular (Assumption 6).*

Theorem 1(a) says that θ_0 is the only parameter value that solves the fully stacked moment conditions from different strata at once. Theorem 1(b) says that once we map each stratum's moments to a common dimension and sum them, θ_0 is still the only parameter value that solves these combined moment conditions at once if $\bar{A}G$ is nonsingular. The proof is short once we have Equation (11): the weights restore each stratum moment to its complete-data expectation, and Assumption 3 guarantees that we can use every instrument and stack the strata to reproduce the complete-data moment conditions $E[\tau(\theta_0)] = 0$. The fact that this is a linear model turns the uniqueness of the solution into a rank condition.

We now revisit Example 1 of the 2-instrument case below and see how Theorem 1(a) and Theorem 1(b) are related (Remark 2). We also discuss an extreme case where we never observe all instruments at once and no single stratum identifies the model on its own. Combining the strata, however, identifies θ_0 (Remark 3).

Remark 2. *Let* $p = 1$, $q = 1$, *and suppose all four strata occur with positive probability. Stratum* $\{1, 2\}$ *contributes* $k_j + q = 3$ *moment conditions; strata* $\{1\}$ *and* $\{2\}$ *each contribute 2; stratum* $\emptyset$ *contributes 1. The stacked moment condition of Theorem 1(a) has* $\sum_j (k_j + q) = 3 + 2 + 2 + 1 = 8$ *in total and says that* θ_0 *satisfies all 8 restrictions*

simultaneously. In this example, it is worth noting that the total number of raw moments is $k+q=3$, e.g., stratum $\{1\}$'s Z_1 row and stratum $\{1,2\}$'s Z_1 row share the same restriction $E[Z_1\tilde{\epsilon}(\theta)]=0$. Theorem 1(b) considers the $p+q=2$ distinct moment conditions, so the model is overidentified.

Remark 3. *Let $p=2$, $q=0$, and suppose the only strata that we observe with some positive probability are $s_1=\{1\}$ and $s_2=\{2\}$. This means that we never observe both instruments at the same time. We can either observe instrument Z_1 with D, or instrument Z_2 with D. In either stratum alone, θ_0 is underidentified. Yet with $A_j \in \mathbb{R}^{2\times 1}$, $\bar{A}=[A_1,A_2]$ and $\bar{A}G=[A_1,A_2]G$ is nonsingular whenever $[A_1,A_2]$ and $E[ZD']$ are. Combining the two strata identifies θ_0 under the complete-data rank condition. This is the analogue of the extreme case (one observes either Z_1 or Z_2, but never both) discussed in Muris (2020), though my SMAR assumption is more relaxed than his MAR assumption as I allow $p_{Z,j}$ to depend on D.*

Remark 2 also demonstrates why we need $\bar{A}G$ to be nonsingular: Each A_j maps the (k_j+q)-dimensional stratum moments into a $(p+q)$-dimensional vector of moment conditions. Since the A_j generally have different dimensions, we need $\bar{A}G$ to convert the information from each stratum to the complete-data Jacobian.

As the number of instruments k increases, the number of possible missingness patterns grows exponentially. This growth does not by itself undermine identification. Under Theorem 1, only strata with positive probability enter the moment system, while the coverage requirement is imposed instrument by instrument. Patterns with zero probability therefore contribute no moment conditions and can be omitted.

4 Estimation

In this section, we propose a two-step IPW-GMM estimator for this model. The first step is to estimate the propensities $p_j(D,V),\ j=1,\cdots,J$. The second step uses the inverses of the estimated probabilities as weights and estimates θ_0 by GMM.

Step 1 (propensity score, nonparametric):
Estimate $p_D(\cdot)$ by a nonparametric (e.g., series or other sieve estimators) regression of g_D on V using all observations. Then estimate $p_{Z,j}(\cdot)$ by a nonparametric multinomial regression of the stratum indicator $(g_{Z,1},\cdots,g_{Z,J})$ on (D,V) using the subsample with $g_D=1$. Trim all estimated probabilities to $[\tilde{\delta},1]$ for a small $\tilde{\delta}\in(0,\delta]$, and set $\hat{p}_j:=\hat{p}_D\hat{p}_{Z,j}$, $\hat{w}_{ij}:=\frac{g_{D,i}g_{Z,ij}}{\hat{p}_j(D_i,V_i)}$.

Step 2 (GMM):
Solve

$$\hat{g}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} A_j \hat{w}_{ij} \tilde{S}_j \tau_i(\theta) = 0$$

and denote the solution by $\hat{\theta}$.

Lemma 3. *The solution of the GMM function in Step 2 above is*

$$\hat{\theta} = \hat{Q}_n^{-1} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} A_j \hat{w}_{ij} \tilde{S}_j t_i \tag{13}$$

where $\hat{Q}_n$ is assumed to be nonsingular and takes the form:

$$\hat{Q}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} A_j \hat{w}_{ij} \tilde{S}_j T_i$$

Moreover,

$$\hat{\theta} - \theta_0 = \hat{Q}_n^{-1} \hat{g}_n(\theta_0)$$

Under Assumptions 2 and 3, a preliminary choice of A_j, such as $A_j = (\tilde{S}_j \hat{G}_n)'$ for a consistent pilot $\hat{G}_n$, satisfies Assumption 6.

We estimate $p_{Z,j}(\cdot)$ using a multinomial model because, unlike the binary indicator g_D, λ_Z is a categorical variable and records membership in one of J instrument strata. An alternative is to estimate $\hat{p}_{Z,j}(\cdot)$ by separate binary sieve regressions, one per instrument and multiplying the resulting probabilities to obtain the stratum probabilities. This factorization requires the instrument observation indicators to be mutually independent conditional on (D, V) among observations with $g_D = 1$. Although this restriction holds in our Monte Carlo design, it may be restrictive in applications where instrument availability reflects common survey waves or reporting processes. Joint multinomial estimation accommodates dependence among these instrument missingness patterns without imposing such factorization. Cattaneo (2010, Appendix B) establishes uniform convergence results for multinomial logistic series estimators under suitable regularity conditions.

A large number of instruments can nevertheless make first-step estimation demanding. When many instrument patterns occur with positive probability, the sample is distributed across more strata, potentially leaving few observations for estimating some conditional pattern probabilities. Small conditional observation probabilities also increase the sensitivity of inverse probability weights to estimation error. Thus, although the number of possible patterns does not by itself undermine identification, the number and relative

frequency of the patterns that occur affect the precision and stability of estimation. The asymptotic analysis in this paper treats J as fixed; allowing it to increase with the sample size would require additional analysis.

Now we introduce the assumption that is sufficient to establish the consistency of this estimator:

Assumption 7 (First-step conditions for consistency)**.** *The estimated probabilities are trimmed to* $[\tilde{\delta}, 1]$ *for some fixed* $\tilde{\delta} \in (0, \delta]$*, and*

$$\|\hat{p}_D - p_D\|_\infty = o_p(1), \qquad \max_{1 \leqslant j \leqslant J} \|\hat{p}_{Z,j} - p_{Z,j}\|_\infty = o_p(1)$$

Theorem 2 (Consistency)**.** *Suppose Assumptions 1 to 7. Then*

$$\hat{\theta} \to_p \theta_0$$

5 Asymptotic Theory

5.1 Asymptotic Variance

Since the SMAR assumption makes the information monotone even though the observation patterns are not, we derive the asymptotic theory by sequential projections of the moment function (see Robins, Rotnitzky, and Zhao, 1995; Tsiatis, 2006).

Define

$$\mu_1(D, V; \theta) := E[\tau(\theta)|D, V] = \begin{pmatrix} m_Z(D, V)\tilde{\epsilon}(\theta) \\ X\tilde{\epsilon}(\theta) \end{pmatrix}, \quad m_Z(D, V) := E[Z|D, V] \tag{14}$$

$$\mu_0(V; \theta) := E[\mu_1(D, V; \theta)|V], \quad e_D := \mu_1 - \mu_0, \quad e_Z := \tau - \mu_1 = \begin{pmatrix} (Z - m_Z)\tilde{\epsilon}(\theta) \\ 0_q \end{pmatrix} \tag{15}$$

Notice that for every $\theta \in \Theta$, we have:

$$E[e_D|V] = 0, \qquad E[e_Z|D, V] = 0$$

An immediate result of SMAR is that all these conditional means are estimable by unweighted regressions on observed data. For example, $m_{Z,\kappa}$ can be obtained by regressing Z_κ on (D, V) over the observations that satisfy $\{g_D = 1, \kappa \in \lambda_Z\}$; and $E[\cdot|V]$ can be obtained by regression on V over the observations that satisfy $\{g_D = 1\}$. These conditional means are not necessary for estimating θ, but are needed for inference and efficiency.

Consider a simplified case where the data availability probabilities p_D and $p_{Z,j}$ are known. Theorem 3 gives the asymptotic distribution for this case:

Theorem 3 (Asymptotic distribution with known probabilities)**.** *Suppose p_D and $p_{Z,j}$ $(\forall j)$ are known and Assumptions 1 to 6 hold. Then*

$$\sqrt{n}(\hat{\theta}_{known} - \theta_0) = -(\bar{A}G)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i^{known} + o_p(1)$$

and

$$\sqrt{n}(\hat{\theta}_{known} - \theta_0) \to^d N(0, (\bar{A}G)^{-1}\Omega_{known}(\bar{A}G)^{-1\prime})$$

where

$$\psi^{known}(\cdot) = \sum_{j=1}^{J} A_j w_j \tilde{S}_j \tau(\theta_0)$$

$$\Omega_{known} = Var(\psi^{known}) = \sum_{j=1}^{J} A_j E[\frac{1}{p_j}\tilde{S}_j\tau(\theta_0)\tau(\theta_0)'\tilde{S}_j']A_j'$$

This theorem serves as a benchmark, and we compare its asymptotic variance with that of our estimator later in this paper.

Note that the estimator we proposed in Section 4 never computes μ_0 or μ_1. These terms are generated by estimating the two probabilities in the first step and serve as adjustment terms.

Lemma 4. *Under Assumption 4, for every j,*

$$w_j\tilde{S}_j\tau = \underbrace{\tilde{S}_j\mu_0 + \frac{g_D}{p_D}\tilde{S}_j e_D + w_j\tilde{S}_j e_Z}_{=:\ \psi_j^*} + \underbrace{\frac{g_D - p_D}{p_D}\tilde{S}_j\mu_0}_{=:\ r_{D,j}} + \underbrace{\frac{g_D}{p_D}\cdot\frac{g_{Z,j} - p_{Z,j}}{p_{Z,j}}\tilde{S}_j\mu_1}_{=:\ r_{Z,j}} \tag{16}$$

We need some additional assumptions on the first step estimation for our estimator to be well behaved.

Assumption 8 (First step estimation)**.**

(i). The estimated probabilities $\hat{p}_D$ and $\hat{p}_{Z,j}$'s $(\forall j)$ take values in $[\tilde{\delta}, 1]$, $\tilde{\delta} \in (0, \delta]$

(ii). $\|\hat{p}_D - p_D\|_{L^2(V)} = o_p(n^{-1/4})$ and $\max_j \|\hat{p}_{Z,j} - p_{Z,j}\|_{L^2(D,V|g_D=1)} = o_p(n^{-1/4})$

(iii). The estimated probabilities are in a fixed Donsker class with probability approaching 1, and are uniformly consistent

(iv). $E[\|\tau(\theta_0)\|^2] < \infty$.

(v). For every j.

$$\frac{1}{n}\sum_{i=1}^{n}\hat{w}_{ij}\tilde{S}_j\tau_i(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}^{*} + o_p(n^{-1/2}) \tag{17}$$

Assumption 8(i) is a standard setup for IPW-type estimators to keep the inverse weights bounded; Assumption 8(ii) imposes convergence rates on the estimated probabilities. The first-order effect of their estimation is specified separately by Assumption 8(v). Suitable series estimators can attain this rate under sufficient smoothness and appropriate choices of the series dimension (Newey, 1997), and a correctly specified parametric first step with the usual root-n convergence rate satisfies it; Assumption 8(iii) restricts the complexity of the first-step estimators and provides uniform consistency, which is standard for series and sieve estimators under smoothness (see Newey, 1994, 1997; Chen, Hong, and Tarozzi, 2008; Cattaneo, 2010); Assumption 8(iv) ensures square integrability of the complete-data moment function. Together with overlap, it guarantees finite second moments of the adjusted moments. This condition is also implied by Assumption 5. Furthermore, Assumption 8(v) imposes a high-level first-step expansion analogous to the condition in Cattaneo (2010, Theorem 4, condition (4.2)) and restricts the adjustment terms from the first-step estimators shown in Lemma 4. Such adjustment terms and restrictions also appear in Muris (2020).

Now, we present the asymptotic distribution of our proposed estimator:

Theorem 4 (Asymptotic distribution with estimated probabilities)**.** *Under Assumptions 1 to 6 and 8*

$$\sqrt{n}(\hat{\theta} - \theta_0) = -(\bar{A}G)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i^{*} + o_p(1) \to^{d} N(0, \Gamma^{*}) \tag{18}$$

where, with all objects evaluated at θ_0,

$$\psi^{*} = \bar{A}\mu_0 + \frac{g_D}{p_D}\bar{A}e_D + \sum_{j=1}^{J} w_j A_j \tilde{S}_j e_Z \tag{19}$$

$$\Omega^{*} = \bar{A}M_0\bar{A}' + \bar{A}M_D\bar{A}' + \sum_{j=1}^{J} A_j\tilde{S}_j M_{Z,j}\tilde{S}_j'A_j' \tag{20}$$

$$\Gamma^{*} = (\bar{A}G)^{-1}\Omega^{*}(\bar{A}G)^{-1\prime}$$

where $M_0 = E[\mu_0\mu_0']$, $M_D = E[p_D^{-1}e_De_D']$, and $M_{Z,j} = E[(p_j)^{-1}e_Ze_Z']$.

Notice that the three terms of Equation (20) are the variances of the three terms of Equation (19), which are mutually uncorrelated. Recall that they are based on the nested information structure $\sigma(V) \subseteq \sigma(D, V) \subseteq \sigma(W)$. Since the missingness is in the instru-

ments and the endogenous variables, and we can still recover partial information from a data point where instruments are not fully observed as long as we observe D in that point, Ω^* takes the form of the sum of three terms. The distinctions from Muris (2020) and Yang (2026) are: Muris (2020) allows multiple variables to be missing, but conditions the observation probabilities on a common set of fully observed variables. Here, instrument-pattern probabilities may additionally depend on D when $g_D = 1$, giving rise to the intermediate projection on (D, V). Yang (2026) also allows sequential conditioning, but her augmented estimator exploits additive separability of the moment function and an additional observation restriction when the first variable is missing.

From Lemma 4, we know that the weighted moment decomposes into ψ^* and two additional terms: $r_{D,j}$ and $r_{Z,j}$. By Newey (1994) and Cattaneo (2010), the influence function correction terms from the first step nonparametric estimation cancel out $r_{D,j}$ and $r_{Z,j}$ under Assumption 8. This is why using estimated weights in fact reduces the asymptotic variance.[3] We introduce the corresponding Proposition 1 below:

Proposition 1 (Efficiency gain from estimated weights)**.** *Under the conditions of Theorems 3 and 4,*

$$\begin{aligned}\Omega_{known} - \Omega^* = E[&\frac{1-p_D}{p_D}\bar{A}\mu_0\mu_0'\bar{A}'] \\ &+ \sum_{j=1}^{J} E[\frac{1-p_{Z,j}}{p_D p_{Z,j}} A_j \tilde{S}_j \mu_1 \mu_1' \tilde{S}_j' A_j'] \\ &- \sum_{j\neq l} E[\frac{1}{p_D} A_j \tilde{S}_j \mu_1 \mu_1' \tilde{S}_l' A_l'] \\ &\succeq 0\end{aligned}$$

Consequently,

$$(\bar{A}G)^{-1}\Omega_{known}(\bar{A}G)^{-1'} - \Gamma^* \succeq 0$$

Proposition 1 says that $\Omega_{known} - \Omega^*$ is positive semidefinite and estimating the selection probabilities weakly improves on knowing them. This is consistent with the theoretical findings in the literature (Robins, Rotnitzky, and Zhao, 1994; Hirano, Imbens, and Ridder, 2003; Wooldridge, 2007; Prokhorov and Schmidt, 2009). In short, the nonparametric estimation strategy removes the noise coming from the randomness of the missingness indicators, given the conditioning variables, and hence improves the overall efficiency.

To estimate the variance Γ^*, given $\hat{\theta}$, we first estimate the set of nuisance functions $\mu_{1,i}$

[3]For general methods of deriving influence functions for estimators with nonparametric first steps, see Newey (1994) and Ichimura and Newey (2022).

and $\mu_{0,i}$ directly from the sample and denote the fitted values $\hat{\mu}_{1,i}$ and $\hat{\mu}_{0,i}$, then define:

$$\hat{e}_{D,i} = \hat{\mu}_{1,i} - \hat{\mu}_{0,i}, \qquad \tilde{S}_j \hat{e}_{Z,i} = \tilde{S}_j \tau_i(\hat{\theta}) - \tilde{S}_j \hat{\mu}_{1,i} \qquad \text{when } g_{D,i} g_{Z,ij} = 1$$

Then, estimate:

$$\hat{M}_0 = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_{0,i} \hat{\mu}_{0,i}', \quad \hat{M}_D = \frac{1}{n} \sum_{i=1}^{n} \frac{g_{D,i} \hat{e}_{D,i} \hat{e}_{D,i}'}{\hat{p}_D(V_i)^2} \tag{21}$$

and estimate the observable instrument block by:

$$\widehat{\tilde{S}_j M_{Z,j} \tilde{S}_j'} := \frac{1}{n} \sum_{i=1}^{n} \frac{g_{D,i} g_{Z,ij}}{\hat{p}_j(D_i, V_i)^2} (\tilde{S}_j \hat{e}_{Z,i})(\tilde{S}_j \hat{e}_{Z,i})' \tag{22}$$

Notice that an estimate of the full $M_{Z,j}$, or an unobserved element of τ_i, is not required. Plug these into Equation (20), and then obtain:

$$\begin{aligned} \hat{\Omega}^* &= \bar{A}(\hat{M}_0 + \hat{M}_D)\bar{A}' + \sum_{j=1}^{J} A_j \widehat{\tilde{S}_j M_{Z,j} \tilde{S}_j'} A_j' \\ \hat{\Gamma}^* &= \hat{Q}_n^{-1} \hat{\Omega}^* \hat{Q}_n^{-1\prime} \end{aligned} \tag{23}$$

Note that we need to square the estimated probabilities in the denominators in Equation (21) and Equation (22). For $\hat{M}_D$, we have $E[g_D p_D^{-2} e_D e_D'] = E[p_D^{-1} e_D e_D'] = M_D$, and likewise for $\hat{M}_{Z,j}$. Hence, the estimators proposed above are consistent under Assumption 8. If we do not square these probabilities in the denominators, the probabilities would cancel entirely and generate a complete-data-style variance that ignores missingness, understating every standard error.

Now we establish the consistency of the variance estimator.

Proposition 2 (Consistency of the variance estimator)**.** *Under Theorem 4, if*

$$\frac{1}{n} \sum_{i=1}^{n} \|\hat{\mu}_{0,i} - \mu_0(V_i)\|^2 = o_p(1), \qquad \frac{1}{n} \sum_{i=1}^{n} g_{D,i} \|\hat{\mu}_{1,i} - \mu_1(D_i, V_i)\|^2 = o_p(1) \tag{24}$$

then,

$$\hat{\Omega}^* \to_p \Omega^*, \qquad \hat{\Gamma}^* \to_p \Gamma^*$$

The proof is in the Appendix.

5.2 Optimal Weighting

Recall the influence function ψ_j^* from Lemma 4. Proposition 1 shows that under Assumption 8(v), such influence function can cancel out the additional noise terms $r_{D,j}$ and $r_{Z,j}$ that originate from missingness. Consider the covariance relevant for optimal weighting

$$\Sigma^* := Var \begin{pmatrix} \psi_1^* \\ \vdots \\ \psi_J^* \end{pmatrix} \tag{25}$$

The covariance decomposition in Section 5.1 gives that for the (j, l) block:

$$\Sigma_{jl}^* = \tilde{S}_j(M_0 + M_D)\tilde{S}_l' + \mathbb{1}\{j = l\}\tilde{S}_j M_{Z,j}\tilde{S}_j' \tag{26}$$

and consequently,

$$\Omega^*(A_1, \cdots, A_J) = \sum_{j=1}^{J}\sum_{l=1}^{J} A_j \Sigma_{jl}^* A_l' \tag{27}$$

Notice that although the indicators $g_{Z,j}$ and $g_{Z,l}$ cannot both equal 1 for $j \neq l$, ψ_j^* share the variation that exists in both μ_0 and e_D. Their covariance generally does not vanish.

We now introduce additional assumptions needed for the optimal weighting matrices:

Assumption 9 (Nondegeneracy of the moment functions)**.**

(i). θ_0 *is an interior point of* Θ

(ii). After removing redundant moment conditions, the covariance matrix of the remaining adjusted moments is positive definite, i.e.

$$\Sigma_{reduced} \succ 0$$

Assumption 9(i) is standard and rules out a boundary-driven limit distribution. Assumption 9(ii) states that every non-zero remaining linear combination has sampling variation. Similar assumptions can be found in Muris (2020), Graham (2011), and Ackerberg et al. (2014). Identification strategy still follows Section 3 and remains unchanged.

Theorem 5 (Information matrix for the observable moment conditions)**.** *Under Assumptions 1 to 6, 8, and 9, define the GMM information matrix*

$$\mathcal{I}_0 = G' \left[\sum_{j=1}^{J}\sum_{l=1}^{J} \tilde{S}_j' [(\Sigma^*)^+]_{jl} \tilde{S}_l \right] G \tag{28}$$

The matrix $\mathcal{I}_0$ is positive definite. Its inverse

$$\Gamma_{\mathit{eff}} = \mathcal{I}_0^{-1} \tag{29}$$

is the lower bound on the limiting covariance of $\sqrt{n}(\hat{\theta}-\theta_0)$ among the proposed IPW-GMM estimators satisfying Assumption 8, over all weighting matrices satisfying Assumption 6.

Note that here $[(\Sigma^*)^+]_{jl}$ is a block of Moore Penrose generalized inverse of the full covariance matrix instead of the inverse of Σ^*_{jl} alone.

Under the conditions of Theorem 5, define

$$A_j^* = G' \sum_{l=1}^{J} \tilde{S}_l' [(\Sigma^*)^+]_{lj}, \qquad j = 1, \cdots, J \tag{30}$$

and $\bar{A}^* = \sum_{j=1}^{J} A_j^* \tilde{S}_j$. Then

$$\bar{A}^* G = \mathcal{I}_0, \qquad \Omega^*(A_1^*, \cdots, A_J^*) = \mathcal{I}_0 \tag{31}$$

Hence these matrices satisfy Assumption 6. Under Assumption 8, the estimator using these matrices satisfies

$$\sqrt{n}(\hat{\theta}^* - \theta_0) = -\mathcal{I}_0^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} A_j^* \psi_{ij}^* + o_p(1) \to^d N(0, \mathcal{I}_0^{-1}) \tag{32}$$

Its influence function equals the one with lowest asymptotic variance for the model in Theorem 5. The feasible estimator therefore achieves that asymptotic variance for the model. This is the analogue of Muris's optimal-weighting remark. It uses the same principle as optimal GMM (Hansen, 1982). But I derive the covariance structure of the adjusted moments under sequential missingness and characterize feasible weighting matrices that minimize asymptotic covariance within the proposed IPW-GMM estimator class. The entries of the covariance accommodate the two-layer observation process. In general, a stratum's optimal matrix depends on the covariance with other strata, not only its own variance.

The optimal matrices contain unknown quantities, but we can use what we introduced previously in this paper to establish feasible weights and inference. Let $\hat{G}$ be a consistent estimate of G. Firstly, the covariance blocks can be expressed as

$$\hat{\Sigma}_{jl}^* = \tilde{S}_j(\hat{M}_0 + \hat{M}_D)\tilde{S}_l' + \mathbb{1}\{j = l\}\widehat{\tilde{S}_j M_{Z,j} \tilde{S}_j'} \tag{33}$$

And secondly, the feasible optimal weighting matrices are

$$\hat{A}_j^* = \hat{G}' \sum_{l=1}^{J} \tilde{S}_l' [(\hat{\Sigma}^*)^+]_{lj} \to_p A_j^* \tag{34}$$

Finally, the corresponding feasible optimal estimator of this model can be expressed as

$$\hat{\theta}^* = \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \hat{A}_j^* \hat{w}_{ij} \tilde{S}_j T_i \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \hat{A}_j^* \hat{w}_{ij} \tilde{S}_j t_i \right\} \tag{35}$$

6 Monte Carlo Experiment

6.1 Design

I compare the performance of my proposed estimator and its first-step variants with the complete-case estimator and an IPW-GMM estimator designed for MAR, under a DGP that satisfies SMAR. I consider different sample sizes and include a full-data estimator as an infeasible benchmark. The DGP is a linear IV model with one endogenous variable ($p = 1$), one exogenous control ($q = 1$), and two instruments ($k = 2$). The exogenous variables, instruments, endogenous variable, outcome, and error terms are generated as follows:

$$\begin{aligned}
&X, Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} N(0,1), \qquad (X, Z_1, Z_2) \perp (\eta, \epsilon), \\
&D = 0.1 + 0.3Z_1 + 0.1Z_2 + \eta, \\
&Y = 0.3D + 0.5X + \epsilon, \\
&(\eta, \epsilon)' \sim N\left((0,0)', \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right).
\end{aligned}$$

Thus, D is endogenous and the complete-data moment conditions are

$$E\left[(Z_1, Z_2, X)'(Y - \alpha D - \beta X)\right] = 0.$$

The parameter of interest is $\alpha_0 = 0.3$.

I then impose an SMAR missingness structure on the complete data. Let

$$\pi_c(u) := c + (1 - 2c)\Phi(u), \qquad c = 0.05,$$

where Φ denotes the standard normal cumulative distribution function. The following indices parameterize probabilities of being missing:

$$\begin{aligned} P(g_D = 0 \mid V, D, Z) &= \pi_c(-0.6 + 0.2Y + 0.5Y^2), \\ P(\lambda_{Z_1} = 0 \mid Z, D, V, g_D = 1) &= \pi_c(-0.4 - 0.1Y + 0.8D), \\ P(\lambda_{Z_2} = 0 \mid Z, D, V, g_D = 1) &= \pi_c(-0.5 - 0.3Y - 0.5D + 0.8D^2). \end{aligned}$$

where, for $\kappa \in \{1, 2\}$, let $\lambda_{Z_\kappa} = \mathbb{1}\{\kappa \in \lambda_Z\}$ denote the observation indicator for instrument Z_κ. Conditional on $(D, V, g_D = 1)$, the two instrument-observation indicators are generated independently. Defining

$$q_l(D, V) := P(\lambda_{Z_l} = 1 \mid D, V, g_D = 1), \qquad l \in \{1, 2\},$$

the four instrument-pattern probabilities are therefore

$$\begin{aligned} p_{Z,1}(D, V) &= q_1 q_2, & p_{Z,2}(D, V) &= q_1(1 - q_2), \\ p_{Z,3}(D, V) &= (1 - q_1)q_2, & p_{Z,4}(D, V) &= (1 - q_1)(1 - q_2). \end{aligned}$$

Hence, the number of instrument-observation strata is $J = 4$.

Three things are worth noting here. Firstly, the bounded missingness link implies that every binary observation probability and its complement lie in $[0.05, 0.95]$. Consequently, every instrument-pattern probability is bounded below by $0.05^2 = 0.0025$, and every joint probability of observing D and a particular instrument pattern is bounded below by $0.05^3 = 0.000125$. Assumption 4 is therefore satisfied by construction. Under the coefficients above, the complete-case probability is approximately 0.14, so about 86 percent of observations are incomplete in at least one of (D, Z_1, Z_2).

Secondly, the missingness probabilities for D and Z_2 are deliberately nonlinear and U-shaped. A bounded-probit model with an index linear in (Y, X) for D, or linear in (Y, D, X) for Z_2, is therefore misspecified. This is the situation that the flexible first step developed in Section 4 and analyzed in Theorem 4 is designed for. Unlike an ordinary probit link, which allows probabilities arbitrarily close to zero and one, the bounded link used here keeps the probabilities away from the boundary and so preserves overlap in the tails. Without such a lower bound, inverse-weighted moments can be dominated by observations with probabilities arbitrarily close to zero, leading to the irregular behavior discussed by Khan and Tamer (2010).

Thirdly, conditional independence between the two instrument-observation indicators is imposed only to simplify this simulation design. It allows the two probabilities q_1 and q_2

to be estimated separately and multiplied into the four pattern probabilities. The general estimator developed in the paper does not require this restriction and instead permits an unrestricted multinomial model for $p_{Z,j}(D, V)$.

Six main estimators are compared. The first estimator is full-data 2SLS, computed before missingness is imposed. It is infeasible in the observed data but provides a benchmark for coding, instrument strength, and the dispersion attainable when all variables are observed.

The second estimator is conventional complete-case 2SLS and uses only observations for which (D, Z_1, Z_2) are all observed. Since complete-case selection depends on both Y and D in this design, this estimator is generally inconsistent even though its sampling distribution can be tightly concentrated around a pseudo-true value.

The third estimator is an all-pattern IPW-MAR estimator. It estimates $p_D(V)$ flexibly using the always-observed variables and estimates the four instrument-pattern probabilities by a multinomial series model in $V = (Y, X)'$ on the subsample with $g_D = 1$. Unlike a weighted complete-case estimator, it uses all four instrument-observation strata. It is nevertheless misspecified in this DGP because the true instrument-pattern probabilities depend on D, which standard MAR excludes from the conditioning set.

The remaining three main estimators are different versions of the IPW-SMAR estimator proposed in this paper. The fourth one is the parametric first step uses the bounded-probit link that matches the support of the DGP but restricts each index to be linear: p_D is modeled in (Y, X) and the two instrument-missing probabilities are modeled in (Y, D, X). It is misspecified because the true models for p_D and q_2 contain quadratic terms. The fifth one is the sieve first step uses additive cubic B-splines in the same conditioning variables, with spline dimensions equal to 3, 4, and 5 at $n = 100$, 1,000, and 10,000, respectively. The last one is the true-probability estimator uses the data-generating probabilities directly and skips first-step estimation.

The MAR estimator and all three SMAR estimators use the same stacked stratum-specific moment conditions and pilot weightings $A_j = (\tilde{S}_j \hat{G})'$. These pilot weightings are a preliminary, rather than fully optimal, cross-stratum combination.

The first-step implementation uses probability regularization that matches the overlap imposed in the DGP. Estimated binary observation probabilities are clipped symmetrically to $[0.05, 0.95]$. The vector of four estimated pattern probabilities is regularized on the probability simplex so that every pattern probability is at least 0.0025 and the four probabilities continue to sum to one.

The simulation uses $r = 1{,}000$ Monte Carlo replications at $n \in \{100, 1{,}000, 10{,}000\}$. With

two instruments and one endogenous variable, the degree of overidentification is one. Finite-sample results for conventional k-class estimators show that just-overidentified IV estimators can have very heavy tails and that the existence of higher-order moments is delicate (see Sawa, 1969; Mariano, 1972; Kinal, 1980; and Phillips, 1983). These results do not mechanically establish the nonexistence of the variance for the selected-sample and estimated-weight GMM estimators used here. I therefore avoid making that stronger claim and instead emphasize robust finite-sample summaries. Following practice in the weak-IV and related simulation literature (Hahn, Hausman, and Kuersteiner, 2004; Andrews, Stock, and Sun, 2019), I report mean bias, median bias, the normal-consistent median absolute deviation,

$$\text{MAD} = 1.4826 \times \text{median}\,|\hat{\alpha} - \text{median}(\hat{\alpha})|\,,$$

the median absolute error, the 10th and 90th percentiles and their difference, the minimum, and maximum estimates. Mean bias is retained as a familiar statistic but should be read together with the median-based measures because it is sensitive to a small number of extreme draws. The baseline table omits the standard deviation and root mean squared error (RMSE) because these quantities can be unstable in heavy-tailed designs.

6.2 Results

Table 1 reports the Monte Carlo results for the six estimators. Figures 1 to 4 in Appendix B present the corresponding sampling distributions and absolute-error comparisons. Because several estimators exhibit substantial tail behavior in small samples, I emphasize median bias, MAD, median absolute error, and the central 80 percent range rather than mean bias alone.

The full-data estimator provides a useful benchmark for the underlying IV design. Its mean bias is essentially zero at every sample size, while its median bias falls from 0.0551 at $n = 100$ to 0.0009 at $n = 1{,}000$ and -0.0003 at $n = 10{,}000$. Its MAD declines from 0.3020 to 0.0965 and then to 0.0314, and its central 80 percent range declines from 0.8005 to 0.2500 and then to 0.0800. The nearly proportional reduction in dispersion as the sample size increases confirms that the complete-data IV model is well behaved and that the differences among the remaining estimators are generated by the missing-data procedures rather than by a failure of the underlying design. It also provides the relevant efficiency benchmark. At $n = 10{,}000$, for example, the sieve SMAR estimator has an MAD of 0.1887, compared with 0.0314 under full data, so the implemented incomplete-data estimator remains substantially less precise than the full-data benchmark under this

Table 1: Monte Carlo results for $\hat{\alpha}$ under the nonlinear SMAR design (1,000 replications)

n	Estimator	Mean bias	Med. bias	MAD	MedAE	Q_{10}	Q_{90}	$Q_{90}-Q_{10}$	Min	Max
100	Full data 2SLS	−0.0003	0.0551	0.3020	0.2184	−0.1214	0.6791	0.8005	−2.880	1.590
100	Complete case 2SLS	−0.0647	−0.0305	0.9042	0.6133	−1.294	1.571	2.866	−27.31	73.63
100	IPW MAR (all patterns)	0.3795	0.0286	0.7274	0.4803	−0.9141	1.565	2.479	−118.8	292.0
100	IPW SMAR (parametric)	−0.9206	−0.0621	0.6005	0.4077	−0.8036	1.236	2.040	−1,087	345.3
100	IPW SMAR (sieve)	−1.782	0.1315	0.8188	0.5922	−1.045	1.793	2.837	−1,446	59.88
100	IPW SMAR (true.p)	1.899	0.2280	0.8540	0.6081	−0.9061	1.893	2.799	−41.49	1,748
1,000	Full data 2SLS	−0.0017	0.0009	0.0965	0.0656	0.1752	0.4252	0.2500	−0.0800	0.5897
1,000	Complete case 2SLS	−0.4649	−0.3708	0.4428	0.4098	−0.8714	0.4327	1.304	−7.041	8.971
1,000	IPW MAR (all patterns)	−0.2917	−0.0808	0.4331	0.2921	−0.5771	0.6846	1.262	−65.88	23.04
1,000	IPW SMAR (parametric)	−0.2053	−0.1198	0.2083	0.1612	−0.1684	0.4248	0.5932	−38.42	1.062
1,000	IPW SMAR (sieve)	−0.0186	0.0512	0.4075	0.2820	−0.3741	0.8628	1.237	−66.20	69.43
1,000	IPW SMAR (true.p)	−0.0671	0.1062	0.4249	0.3127	−0.3780	0.9433	1.321	−104.0	17.42
10,000	Full data 2SLS	0.0000	−0.0003	0.0314	0.0213	0.2607	0.3407	0.0800	0.1593	0.3906
10,000	Complete case 2SLS	−0.5217	−0.5063	0.1498	0.5063	−0.4313	−0.0383	0.3930	−0.9891	0.2252
10,000	IPW MAR (all patterns)	−0.1987	−0.1599	0.2288	0.1918	−0.2558	0.4189	0.6748	−5.109	1.619
10,000	IPW SMAR (parametric)	−0.1260	−0.1221	0.0778	0.1225	0.0668	0.2820	0.2153	−0.3592	0.4500
10,000	IPW SMAR (sieve)	0.0367	0.0058	0.1887	0.1292	0.0151	0.5531	0.5379	−2.251	47.95
10,000	IPW SMAR (true.p)	0.0020	0.0139	0.1933	0.1327	0.0045	0.5724	0.5679	−0.8780	1.319

Notes: Mean bias and median bias are computed relative to $\alpha_0 = 0.3$. MedAE denotes median absolute error. Q_{10} and Q_{90} are quantiles of $\hat{\alpha}$, and $Q_{90}-Q_{10}$ is the central 80 percent range. MAD is the normal-consistent median absolute deviation defined in the design subsection. The statistics are computed over finite estimates.

high rate of missingness.

The small-sample results also show why the median-based statistics are important in this design. At $n = 100$, the weighted estimators have very large extreme draws, and their mean biases can be dominated by a small number of replications. For example, the sieve estimator has a mean bias of -1.782 but a median bias of 0.1315, while the true-probability estimator has a mean bias of 1.899 but a median bias of 0.2280. The same distinction remains relevant at larger sample sizes. At $n = 10{,}000$, the sieve estimator has a median bias of only 0.0058, but a very large positive draw raises its mean bias to 0.0367 and its maximum to 47.95. The minimum and maximum columns should therefore be viewed as tail diagnostics, while MAD, median absolute error, and the central quantile range provide a more stable description of the typical replication.

The complete-case and MAR estimators illustrate the consequences of imposing the wrong missingness structure. At $n = 100$, both appear nearly centered at the truth: their median biases are -0.0305 and 0.0286, respectively. This apparent performance is a small-sample feature, however, because sampling variation is large relative to the underlying misspecification bias. As the sample size increases and sampling variation falls, that misspecification becomes visible. The median complete-case estimate changes from 0.2695 at $n = 100$ to -0.0708 at $n = 1{,}000$ and -0.2063 at $n = 10{,}000$. The median MAR estimate similarly changes from 0.3286 to 0.2192 and then to 0.1401. At $n = 10{,}000$, even the complete-case central 80 percent interval, $[-0.4313, -0.0383]$, lies entirely below zero. The complete-

case estimator is therefore becoming concentrated near a severely biased selected-sample limit, while the MAR estimator retains a smaller but economically meaningful downward bias because it excludes D from the instrument-missingness probabilities.

Overall, the results show that both parts of the proposed procedure are needed. Conditioning the instrument-observation probabilities on D removes the bias left by the conventional MAR estimator, while the flexible first step is needed to accommodate the nonlinear missingness probabilities. The sieve estimator tracks the true-probability benchmark increasingly closely as the sample size grows and is centered near α_0 in the largest design. In contrast, complete case, MAR, and the misspecified parametric SMAR estimator become increasingly concentrated around different pseudo-true values, regardless of the degree of their dispersion.

7 Empirical Application

7.1 Research Question and Identification Strategy

Our empirical application studies the classical tradeoff between family size and child quality and asks whether a larger number of siblings reduces a child's completed schooling. In the model of Becker and Lewis (1973), parents choose the number of children and the investment per child jointly, so the shadow price of child quality rises with family size. Rosenzweig and Wolpin (1980) first took this implication to the data, and a large subsequent literature, including Angrist, Lavy, and Schlosser (2010), has studied the magnitude of this relationship.

The empirical difficulty comes from the joint choice itself. Family size is not assigned, but is chosen by the same parents who make schooling and other investments in their children. Unobserved parental preferences, resources, and expectations about a child's ability can therefore enter both the fertility decision and the schooling equation. A simple OLS regression of schooling on the number of siblings confounds the causal effect of family size with this selection. Resource dilution tends to produce a negative association, while parental heterogeneity can move the association in either direction. An IV model is therefore needed, with instruments that shift the cost or feasibility of fertility without directly entering the child's schooling equation.

Chaudhuri and Guilkey (2016, henceforth CG) study this question in Indonesia. Their instrument set uses the expansion of the national family planning program that began in 1970, when the national family planning board, BKKBN, developed a large village infras-

tructure for contraceptive supply and maternal and child services. Over the same period, the total fertility rate fell from well above five births per woman in the late 1960s to below three by the middle of the 1990s (Gertler and Molyneaux, 1994). The instruments in CG measure the local presence of institutions created by this expansion. The family planning post, or PPKBD, is a village distribution point whose local cadres supply contraceptives and refer users to clinics. The posyandu provides monthly community services that include growth monitoring, immunization, maternal care, and family planning. The village midwife indicator reflects the bidan di desa program, while the polindes indicator records the village maternity post associated with that program. CG also use a separate delivery post indicator and a village medicine post indicator. The number of family planning posts and the polindes indicator are the two sometimes-missing instruments. The number of posyandu and the indicators for a village midwife, a delivery post, and a village medicine post are always observed in the estimation data.

The relevance argument is direct. Each of these facilities can lower the money, time, or information cost of contraception and safe birth spacing, so villages with more extensive family planning infrastructure should have lower completed fertility among exposed mothers. The exclusion restriction requires that, conditional on the controls, these facilities affect a child's schooling only through family size. There are two important concerns. First, program placement was not random, and Pitt, Rosenzweig, and Gibbons (1993) document systematic placement of public programs in Indonesia. Following CG, we condition on the community characteristics that plausibly affected placement and also matter for schooling, including urban status and the local numbers of elementary, junior, and senior schools. Second, health facilities can affect schooling directly through child health. This channel cannot be ruled out a priori and remains a substantive limitation of the design. Our contribution does not alter this identification strategy. We hold the causal design fixed and study how the estimates change when the endogenous variable and part of the instrument set are missing.

This setting fits the theory of the paper in a direct way. The two sometimes-missing instruments are jointly observed for only about one third of the CG sample, which is why their paper develops an IPW GMM estimator that uses observations with incomplete instrument information. Their constructed data set, however, includes only children whose maternal fertility information is available. In the underlying IFLS, a sibling count can be constructed only when the child can be linked to a mother in the household roster and the relevant maternal information is observed. For a substantial share of age eligible children, this link or maternal record is unavailable because the mother is deceased, has migrated, the child is fostered, or the relevant survey book was not completed. Extending the analysis to the full child sample therefore produces the sequential observation pattern

studied in this paper. The endogenous variable D, the number of siblings, is missing when $g_D = 0$. Conditional on $g_D = 1$, the two sometimes-missing instruments are observed in four possible strata. The child's schooling, demographics, and community variables form the always-observed variables V.

The CG extract is used first as a validation sample, since the first missing layer is absent there. The extended sample is then used to estimate the full two layer model. Under the SMAR condition, the instrument observation probabilities may depend on the number of siblings once D is observed. This is the main difference from a conventional MAR specification that conditions only on the always observed instruments, controls, and the child's schooling.

7.2 Data

The data come from the 1993, 1997, 2000, and 2007 waves of the Indonesia Family Life Survey, a longitudinal household and community survey (Frankenberg and Karoly, 1995; Frankenberg and Thomas, 2000; Strauss et al., 2004, 2009). The 1993 baseline covered 321 enumeration areas in 13 provinces, representing about 83 percent of the Indonesian population. Each wave combines a household survey with a community survey that records local schools, health facilities, and family planning institutions.

We use two data sources. The first is the CG replication archive, which contains their constructed child by wave data with the outcome, fertility measure, controls, and community variables. Collapsing this file by community and survey wave produces a panel of 1,094 community and wave cells covering 313 communities. We use this panel for the six instruments, urban status, and the three school counts. This ensures that the community variables in our analysis follow the same construction as in CG. The second source is the raw IFLS household files from all four waves. These files provide the complete roster of age eligible children, including children for whom maternal fertility information cannot be constructed, and therefore allow us to define the first missing layer.

The outcome Y is the child's completed years of schooling. It is constructed from the roster questions on the highest education level attended and the highest grade completed at that level. Because the education codes change across survey waves, we use wave specific mappings to convert each level and grade pair into years of schooling. Primary, junior secondary, and senior secondary schooling are assigned six, three, and three years, respectively, while diploma, university, religious, and equivalency programs are mapped to the corresponding number of completed years.

The endogenous variable D is the number of siblings. Following CG, it is defined as the mother's number of children minus one and is censored above at four in the main specification. The construction of the mother's number of children requires some care. In the CG archive, the fertility measure is constant within mother across waves and has a correlation of 0.77 with the count of her children enumerated in the survey. We therefore construct a mother and child registry from the raw household rosters. Whenever a child's recorded biological mother can be linked to a household member, the pair is added to the registry. We then pool these links across waves using the person identifier and count the distinct children associated with each mother. The registry contains 36,427 mother and child pairs for 16,010 mothers. The resulting sibling count is constant within mother, as in the CG data. Since the censoring point binds for a nontrivial share of children, we also report results that censor the sibling count above at five.

The six instruments, with their definitions discussed in Section 7.1, are taken from the community panel at the child's community and survey wave. The controls follow CG as closely as the extended sample permits. They include a full set of age indicators for ages 12 through 20, with age 11 omitted, the child's sex, an indicator for Muslim, urban status, and the community numbers of elementary, junior, and senior schools. Mother's education and literacy are sometimes missing controls. For estimation, the first-layer indicator records joint availability of the sibling count, maternal education, and maternal literacy. Its probability is estimated using always-observed variables, and the instrument-pattern model is fitted on this jointly observed cell. They then enter the schooling equation on the cell where the maternal record is observed.

We construct a pooled child sample from the four IFLS waves. In each wave, we begin with roster members from ages 11 through 20 for whom schooling can be constructed. A child contributes one observation, taken from the last interview at which the child is in this age range. Stacking the four waves before applying this rule gives 27,000 child and wave observations and 18,487 distinct children. Each child is then matched to the community panel by community and wave. After imposing the availability of the always observed variables, the main analysis sample contains 16,238 children, of whom 8,028 are in urban communities and 8,210 are in rural communities.

7.3 The Missingness Pattern

The first missing layer concerns the number of siblings. A sibling count is available only if the child can be linked to a mother in the household roster and the maternal information needed to construct fertility is observed. Table 2 reports the main sources

Table 2: The first missing layer by wave of last interview

	1993	1997	2000	2007
Age eligible children	7,613	7,420	9,288	8,325
-No mother link in the roster	0.20	0.17	0.25	0.21
-Linked mother, no fertility book	0.19	0.14	0.16	0.16
$g_D = 1$ among children last seen in the wave	0.49	0.61	0.70	0.74
Children in the last interview sample	4,100	2,829	5,790	5,768

Notes: The first three rows refer to the full age eligible children in each wave. The last two rows refer to the pooled sample after assigning each child to the last interview at which the child is age eligible.

Table 3: Instrument missingness by survey wave

Wave	Family planning posts	Polindes	Posyandu	Delivery post	Medicine post	Midwife
1993	0.036	0.000	0.023	0.000	0.000	0.000
1997	0.027	0.000	0.016	0.000	0.000	0.000
2000	0.020	0.440	0.000	0.000	0.000	0.000
2007	1.000	0.287	0.010	0.000	0.000	0.000

Notes: Entries are the shares of community and wave cells with the corresponding instrument missing. The calculations use the panel of 1,094 cells covering 313 communities.

of this missingness by survey wave. Depending on the wave, between 17 and 25 percent of age eligible children do not have a usable mother link in the roster. A further 14 to 19 percent of the children have a linked mother but no usable fertility book. The share with $g_D = 1$ among children whose last eligible interview occurs in each wave is 0.49 in 1993, 0.61 in 1997, 0.70 in 2000, and 0.74 in 2007. In the pooled child sample, roughly one third of children therefore have no observed sibling count.

The second missing layer concerns the number of family planning posts and the polindes indicator. Conditional on observing D, the four instrument patterns contain 895 children with neither instrument, 3,355 with only the polindes indicator, 1,430 with only the family planning post count, and 5,137 with both instruments. The four cells are therefore large enough to contribute separately to estimation.

Table 3 shows that most of the instrument missingness is associated with the survey wave. The family planning post count is nearly complete from 1993 through 2000 and is not available in 2007. The polindes indicator is complete in the first two waves but is missing for 44.0 percent of community cells in 2000 and 28.7 percent in 2007. The other four instruments are almost completely observed.

7.4 Econometric Implementation

Estimation follows the two step procedure in Section 4. The first observation probability, $\hat{p}_D(V)$, is estimated by a series model for the joint availability of number of siblings and the sometimes missing control variables (mother's education and literacy) using all children. Its regressors are the always observed variables: completed schooling, age indicators, sex, Muslim, urban status, the three school counts, and the four always observed instruments. The four probabilities $\hat{p}_{Z,j}(D, V)$ are estimated by a multinomial logit model on the jointly observed cell. The specification includes the variables in the conventional MAR conditioning set and quadratic terms for the continuous covariates. The main SMAR specification also includes the number of siblings.

The estimated probabilities are trimmed at 0.02 in the main specification. Results using 0.01 and 0.05 are reported as robustness checks. The second step uses the stacked moments from all four instrument patterns and the optimal weighting matrices A_j^* from Section 5.2. Standard errors are calculated from the variance estimator in Section 5, including the first step adjustment and the squared estimated probabilities in the missingness components.

The empirical comparison contains five estimators. OLS and GMM with the four always observed instruments are computed on the jointly observed cell. The complete-case GMM estimator with all six instruments uses only observations for which the sibling count, maternal education, maternal literacy, and both sometimes-missing instruments are jointly observed. A complete case IPW estimator reweights this same cell using a single MAR probability that excludes D. The final estimator is the proposed SMAR IPW GMM estimator, which uses all four instrument patterns and allows the pattern probabilities to depend on the number of siblings.

7.5 Identifying Assumptions in the Application

The justification for SMAR rests on the sources of the information combined in the analysis and the conditioning information available at each stage. The IFLS collects household and individual information through separate survey books and supplements these records with a community–facility survey drawing on local informants and community records (Frankenberg and Thomas, 2000). In this application, these sources give rise to two distinct requirements for constructing the moments: the availability of the maternal information needed to construct the sibling count, and the availability of the community variables used as instruments. We discuss this structure first and then the two conditional independence restrictions in Assumption 1.

The two-layer structure. The child's schooling and demographic information establish the baseline record. Constructing the number of siblings requires an additional link to the mother's information, using the family relationships and person identifiers described in Section 7.2. The family planning instruments are obtained from a different source: the community records matched to the child's community and survey wave. A usable maternal record therefore supplies the endogenous variable, while the community record determines which instrument moments can be formed. Among children with observed sibling counts, the family-planning-post count and the polindes indicator may be observed separately or together, giving the four patterns described in Section 7.3.

This arrangement gives the layers a natural econometric order. The baseline information is sufficient for the maternal-record selection process, while allowing family size to provide additional information for the community-record selection process. The order thus follows the information needed for estimation and permits the household and community records to have been collected separately. It also accommodates children whose community information is available even when their maternal link is unavailable. Children without observed sibling counts contribute to estimating the first observation probability; children with observed sibling counts contribute the moments supported by their instrument pattern.

The first SMAR restriction. The first layer concerns whether the maternal information needed to construct the sibling count can be recovered. The use of recorded family relationships and identifiers, together with information pooled across waves, motivates an observation model based on the child's observed circumstances and the environment in which the household is surveyed. The first probability model allows availability to vary with completed schooling, age, sex, religion, urban status, the three school counts, and the four always observed instruments. These variables are available for children both with and without a usable sibling count.

The conditioning variables serve complementary roles. Age and schooling describe the child's stage of life and educational circumstances, while sex and religion allow availability to differ across demographic groups. Urban status, school counts, and the observed family planning instruments describe the local environment in which maternal information is obtained. Including schooling also permits maternal-record availability to be associated with the outcome under study. The comparison underlying the first restriction is therefore between children with similar observed educational, demographic, and community characteristics. Within these characteristics, we maintain that the availability of a usable maternal record is independent of the sibling count and the instrument values. In particular, the community variables provide information about local infrastructure that may be associated with both fertility and the availability of household information.

The longitudinal construction is relevant to maternal absence. Pooling mother–child links across survey waves allows a link established at an earlier interview to contribute to the registry when the mother is absent at the child's final eligible interview. This use of earlier information makes the construction less dependent on maternal presence at a single interview.[4]

The second SMAR restriction. The second layer concerns the community records supplying the family-planning-post count and the polindes indicator among children with observed sibling counts. Their availability is determined at the level of the matched community and survey wave. Table 3 shows the differences across waves. The observed community characteristics and the four always observed instruments supply information about the same local environment as the two sometimes-missing instruments.

The number of siblings supplies additional conditioning information because family composition can be associated with the community records to which children are matched. For example, if larger families are more heavily represented in communities for which only one of the two instruments is recorded, children from those families will more frequently appear in that instrument pattern. Location therefore provides plausible channels through which the sibling count can predict instrument availability. Allowing the second observation probability to depend on the observed sibling count accommodates this association alongside the observed child and community characteristics.

Conditional on the observed sibling count and the baseline characteristics, the second part of SMAR treats the remaining variation in the instrument pattern as independent of the underlying instrument values. Its substantive interpretation is that the sibling count and baseline characteristics account for the relationship between the composition of the matched child sample and the availability of community information. The four pattern probabilities also allow the availability of the two instruments to be dependent, as is appropriate when both variables originate from the same community records. The exclusion restrictions for the instruments continue to follow the identification strategy of CG discussed in Section 7.1.

[4]Maternal death and nonresidence can prevent recovery of the maternal information needed for some children. The first SMAR restriction applies to the availability of a usable record conditional on the child's observed schooling, demographics, and community characteristics. It permits maternal-record availability to vary with these characteristics and requires that, after conditioning on them, availability have no further association with the sibling count or instrument values. Thus, maternal death itself need not be a random event; the maintained restriction concerns the resulting selection into the sample with usable maternal information.

7.6 Results

Table 4 reports the main estimates for the 16,238 children in the extended sample. OLS on the jointly observed cell gives an estimate of -0.076 with a standard error of 0.016. The estimate becomes substantially more negative on the same jointly observed sample when instruments are used, at -1.113 with a standard error of 0.297. The GMM estimate with observed D is substantially larger in absolute magnitude than the OLS estimate; its first-stage F statistic is 19.2.

Among the estimators that use all six instruments, complete case GMM gives -0.324 with a standard error of 0.634, while the complete case IPW estimator gives -0.469 with a standard error of 0.641. The proposed SMAR estimator gives -0.920 with a standard error of 0.384. It has the same negative sign as the two complete case estimates and a reported standard error that is about 40 percent smaller than the unweighted complete case standard error. This comparison is not an efficiency statement, since the complete case and SMAR estimators need not have the same probability limit when the observation process depends on D. It does show that the proposed estimator can use the incomplete instrument strata without increasing the reported sampling uncertainty in this application.

Relative to the published CG benchmarks, the SMAR point estimate is larger in absolute magnitude: -0.920, compared with -0.665 for their AIPW-GMM estimator and -0.709 for their IPW-GMM estimator. Similar to these benchmarks, the SMAR estimate is also statistically significant, which means that we found evidence of a negative effect of the number of siblings on a child's schooling. The comparison is descriptive, since the panels differ in sample construction, controls, and the construction of the sibling count; the difference cannot therefore be attributed solely to the treatment of missingness. The SMAR specification allows instrument availability to depend on the observed sibling count, relaxing the restriction that observation probabilities depend only on always-observed variables.

Results of some additional specifications can be found in Appendix B.

Table 5 reports two robustness checks. First, changing the trimming constant from 0.02 to 0.01 makes trivial changes to both point estimate and the standard error, while increasing it to 0.05 moves the estimate from -0.920 to -1.022. The corresponding standard errors remain close. Across the reported trimming thresholds, the SMAR point estimate remains negative and is statistically significant.

Second, we censor the number of siblings above at five rather than four. The GMM

Table 4: Years of schooling on the number of siblings

Estimator	Estimate	Standard error
Panel A: CG (2016), original sample		
GMM with observed D	−0.731	0.325
Complete-case GMM with observed D and Z	−1.007	0.515
IPW-GMM with six IVs	−0.709	0.338
AIPW-GMM with six IVs	−0.665	0.294
Full CG sample	7,370	
Panel B: This paper, extended sample		
OLS with observed D	−0.076	0.016
GMM with observed D	−1.113	0.297
Complete-case GMM with observed D and Z	−0.324	0.634
Complete-case IPW-GMM (propensity excludes D)	−0.469	0.641
SMAR IPW-GMM	−0.920	0.384
Full analysis sample	16,238	
Instrument-pattern counts: 00/01/10/11	895 / 3,355 / 1,430 / 5,137	

Notes: The outcome is completed years of schooling and D is the number of siblings, censored above at four. Panel A reports the original CG benchmarks with their reported bootstrap standard errors based on 1,000 replications. The panels differ in sample construction, controls, and the construction of the sibling count.

In Panel B, controls include age indicators, sex, Muslim, mother's education and literacy for children with observed sibling counts, urban status, and the three community school counts. OLS and GMM with observed D require an observed sibling count, maternal education and literacy; in addition, GMM with observed D uses the four always-observed instruments; complete-case estimators additionally require both sometimes-missing instruments. "Excludes D" refers only to the indicated propensity model; D and the two sometimes missing controls: mother's education and literacy, remain in the schooling equation.

Pattern counts refer to children with observed D. The first and second digits indicate availability of the family-planning-post count and the polindes indicator, respectively. The full sample counts are not estimator-specific sample sizes. The main probability trim in Panel B is 0.02. Standard errors and optimal weightings for the SMAR estimators use the variance estimator and the optimal weighting estimator in Section 5. 10,783 children contribute directly computable structural moments, with 5,109 complete cases.

with observed D estimate becomes −0.951 with a standard error of 0.259. The complete case estimate becomes −0.186 with a standard error of 0.477, and the proposed SMAR estimate becomes −0.756 with a standard error of 0.329. Relaxing the censoring point changes the magnitude of the estimates but leaves the main conclusions unchanged: the SMAR estimate remains negative and statistically significant.

Substantively, the preferred SMAR specification provides statistically significant evidence that a larger sibling count reduces children's schooling under the maintained identifying assumptions. The estimate remains negative and statistically significant across the reported trimming thresholds and the alternative censoring point. These findings are consistent with a quantity–quality tradeoff in this application.

Table 5: Robustness to trimming and the sibling censoring point

	Estimate	Standard error
Trimming, sibling count censored above at 4		
Trim 0.05	−1.022	0.392
Trim 0.02, main specification	−0.920	0.384
Trim 0.01	−0.918	0.384
Sibling count censored above at 5		
GMM with observed D	−0.951	0.259
Complete case GMM with observed D and Z	−0.186	0.477
SMAR IPW GMM	−0.756	0.329

Notes: The first panel changes the trimming constant while keeping the sibling count censored above at four. The second panel censors the sibling count above at five with a trim level at 0.02.

Taken together, the empirical results support the main econometric motivation of the paper. The benchmark estimates using the CG replication extract are similar to their published estimates, as reported in Appendix B. In the extended sample, the proposed estimator accounts for missing sibling information and uses all four instrument patterns, producing a smaller reported standard error than complete case analysis. The application illustrates how the proposed framework can be implemented when the endogenous variable and instruments are incompletely observed.

8 Conclusion

This paper studies linear IV models in which the endogenous variables and the instruments are missing in different and overlapping parts of the sample. The main difficulty is not only that several variables are incomplete. It is that the observation probability of an instrument can depend on an endogenous variable that is itself observed only for part of the sample. Conventional MAR methods exclude this dependence by conditioning only on variables observed for everyone, while complete case analysis discards all observations outside the fully observed data. Both approaches can therefore fail even when each layer of the data collection process is generating missing values given the information available at that layer.

I address this problem through a sequential missing-at-random condition. Under SMAR, the observation probability of the endogenous variable depends on the always observed variables, and the instrument pattern is conditionally independent of the missing instrument values after conditioning on the endogenous variable and the always observed variables. This structure turns an apparently nonmonotone observation problem into a

nested information problem. The endogenous variable must be observed before any IV moment can be evaluated, while different subsets of instruments contribute different parts of the complete data moment vector. The resulting weighted moments recover their complete data counterparts, and combining them can identify the structural parameter even when no single observation pattern is sufficient for identification.

The proposed estimator implements this argument in two steps. It first estimates the probability that the endogenous variable is observed and the conditional probabilities of the instrument patterns. It then uses the product of these probabilities to construct inverse weighted moments from every usable stratum. The asymptotic analysis shows that the influence function reflects the same order in which information is revealed. Its variance contains separate components associated with the always observed information, the observation of the endogenous variable, and the observation of the instruments. This decomposition provides a direct variance estimator and also explains why a flexible estimated probability model can improve on an unaugmented estimator that simply inserts the true probabilities.

The Monte Carlo experiment supports the main identification and estimation results. Under nonlinear SMAR, complete case analysis and conventional MAR weighting become increasingly concentrated away from the structural parameter as the sample size grows. A restricted parametric SMAR specification also retains a nonvanishing misspecification bias. In contrast, the flexible SMAR estimator becomes centered near the truth and tracks the estimator using the data generating probabilities. The experiment also makes clear that this correction is not costless in small samples. Severe missingness and small observation probabilities produce heavy tails, and the flexible first step requires enough information to estimate the observation models reliably. The practical performance of the method therefore depends on overlap, first step specification, and the strength of the complete data moments.

The empirical application extends the family size design of Chaudhuri and Guilkey (2016) to a setting with two missingness layers. The number of siblings cannot be constructed for every child, and two family planning instruments are observed only in selected community and survey cells. The application shows how the complete case, conventional MAR, and SMAR procedures use different portions of the same data and how the sequential observation models can be estimated in practice. The preferred specification finds a negative and statistically significant effect of sibling counts on schooling under the maintained identifying assumptions.

The main takeaway is that the conditioning set for missingness must reflect the order in which the data become available. When the observation of an instrument can depend on

a sometimes missing endogenous variable (or the other way around), treating the entire problem as conventional MAR changes the maintained restriction and can change the probability limit of the estimator. Furthermore, observations with incomplete instrument information do not need to be discarded when they contain valid parts of the moment vector. A sequential formulation allows these observations to contribute while keeping the missingness assumptions transparent. The same logic can be extended to other moment condition models, additional layers of sequential nonresponse, and panel settings in which different variables disappear at different points in the data collection process.

References

Abrevaya, J., and S. G. Donald (2017): "A GMM Approach for Dealing with Missing Data on Regressors," *The Review of Economics and Statistics*, 99(4), 657–662.

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014): "Asymptotic Efficiency of Semiparametric Two-step GMM," *The Review of Economic Studies*, 81(3), 919–943.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

Angrist, J. D., V. Lavy, and A. Schlosser (2010): "Multiple Experiments for the Causal Link between the Quantity and Quality of Children," *Journal of Labor Economics*, 28(4), 773–824.

Azur, M. J., E. A. Stuart, C. Frangakis, and P. J. Leaf (2011): "Multiple Imputation by Chained Equations: What Is It and How Does It Work?" *International Journal of Methods in Psychiatric Research*, 20(1), 40–49.

Batista, G., and M. C. Monard (2002): "A Study of K-Nearest Neighbour As An Imputation Method," in *Soft Computing Systems: Design, Management and Applications*, ed. by A. Abraham, J. Ruiz-del-Solar, and M. Köppen, Frontiers in Artificial Intelligence and Applications, Vol. 87, pp. 251–260. IOS Press, Amsterdam.

Becker, G. S., and H. G. Lewis (1973): "On the Interaction between the Quantity and Quality of Children," *Journal of Political Economy*, 81(2), S279–S288.

Cattaneo, M. D. (2010): "Efficient Semiparametric Estimation of Multi-Valued Treatment Effects under Ignorability," *Journal of Econometrics*, 155(2), 138–154.

Chaudhuri, S. (2020): "On Efficiency Gains from Multiple Incomplete Subsamples," *Econometric Theory*, 36(3), 488–525.

Chaudhuri, S., and D. K. Guilkey (2016): "GMM with Multiple Missing Variables," *Journal of Applied Econometrics*, 31(4), 678–706.

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric Efficiency in GMM Models with Auxiliary Data," *The Annals of Statistics*, 36(2), 808–843.

Coe, J. E. (2018): "Estimation of Fixed Effects Models with Missing Covariate Data, with an Application to Valuing Local Water Quality," Job Market Paper, University of Texas at Austin.

Cosma, A., A. V. Kostyrka, and G. Tripathi (2026): "Missing Endogenous Variables in Conditional Moment Restriction Models," *Journal of Business & Economic Statistics*, published online May 4, 2026, doi:10.1080/07350015.2026.2619543.

Dagenais, M. G. (1973): "The Use of Incomplete Observations in Multiple Regression Analysis: A Generalized Least Squares Approach," *Journal of Econometrics*, 1(4), 317–328.

Donders, A. R. T., G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons (2006): "Review: A Gentle Introduction to Imputation of Missing Values," *Journal of Clinical Epidemiology*, 59(10), 1087–1091.

Frankenberg, E., and L. Karoly (1995): *The 1993 Indonesian Family Life Survey: Overview and Field Report.* RAND, DRU-1195/1-NICHD/AID.

Frankenberg, E., and D. Thomas (2000): *The Indonesia Family Life Survey (IFLS): Study Design and Results from Waves 1 and 2.* RAND, DRU-2238/1-NIA/NICHD.

Gertler, P. J., and J. W. Molyneaux (1994): "How Economic Development and Family Planning Programs Combined to Reduce Indonesian Fertility," *Demography*, 31(1), 33–63.

Gouriéroux, C., and A. Monfort (1981): "On the Problem of Missing Data in Linear Models," *The Review of Economic Studies*, 48(4), 579–586.

Graham, B. S. (2011): "Efficiency Bounds for Missing Data Models with Semiparametric Restrictions," *Econometrica*, 79(2), 437–452.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–331.

Hahn, J., J. Hausman, and G. Kuersteiner (2004): "Estimation with Weak Instruments: Accuracy of Higher-Order Bias and MSE Approximations," *The Econometrics Journal*, 7(1), 272–306.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

Ichimura, H., and W. K. Newey (2022): "The Influence Function of Semiparametric Estimators," *Quantitative Economics*, 13(1), 29–61.

Kennedy, E. H., J. A. Mauro, M. J. Daniels, N. Burns, and D. S. Small (2019): "Handling Missing Data in Instrumental Variable Methods for Causal Inference," *Annual Review of Statistics and Its Application*, 6, 125–148.

Khan, S., and E. Tamer (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78(6), 2021–2042.

Kinal, T. W. (1980): "The Existence of Moments of k-Class Estimators," *Econometrica*, 48(1), 241–249.

Little, R. J. A. (1988): "A Test of Missing Completely at Random for Multivariate Data with Missing Values," *Journal of the American Statistical Association*, 83(404), 1198–1202.

Little, R. J. A., and D. B. Rubin (2002): *Statistical Analysis with Missing Data*, 2nd ed.

Wiley, Hoboken, NJ.

Mariano, R. S. (1972): "The Existence of Moments of the Ordinary Least Squares and Two-Stage Least Squares Estimators," *Econometrica*, 40(4), 643–652.

McDonough, I. K., and D. L. Millimet (2017): "Missing Data, Imputation, and Endogeneity," *Journal of Econometrics*, 199(2), 141–155.

Mogstad, M., and M. Wiswall (2012): "Instrumental Variables Estimation with Partially Missing Instruments," *Economics Letters*, 114(2), 186–189.

Molenberghs, G., C. Beunckens, C. Sotto, and M. G. Kenward (2008): "Every Missingness Not at Random Model Has a Missingness at Random Counterpart with Equal Fit," *Journal of the Royal Statistical Society, Series B*, 70(2), 371–388.

Muris, C. (2020): "Efficient GMM Estimation with Incomplete Data," *The Review of Economics and Statistics*, 102(3), 518–530.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 1349–1382.

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79(1), 147–168.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, ed. by R. F. Engle and D. L. McFadden, pp. 2111–2245. Elsevier, Amsterdam.

Phillips, P. C. B. (1983): "Exact Small Sample Theory in the Simultaneous Equations Model," in *Handbook of Econometrics*, Vol. 1, ed. by Z. Griliches and M. D. Intriligator, pp. 449–516. North-Holland, Amsterdam.

Pitt, M. M., M. R. Rosenzweig, and D. M. Gibbons (1993): "The Determinants and Consequences of the Placement of Government Programs in Indonesia," *The World Bank Economic Review*, 7(3), 319–348.

Prokhorov, A., and P. Schmidt (2009): "GMM Redundancy Results for General Missing Data Problems," *Journal of Econometrics*, 151(1), 47–55.

Rai, B. (2023): "Efficient Estimation with Missing Data and Endogeneity," *Econometric Reviews*, 42(2), 220–239.

Robins, J. M., and R. D. Gill (1997): "Nonresponse Models for the Analysis of Nonmonotone Ignorable Missing Data," *Statistics in Medicine*, 16(1), 39–56.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89(427), 846–866.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90(429), 106–121.

Rosenzweig, M. R., and K. I. Wolpin (1980): "Testing the Quantity–Quality Fertility Model: The Use of Twins as a Natural Experiment," *Econometrica*, 48(1), 227–240.

Rubin, D. B. (1976): "Inference and Missing Data," *Biometrika*, 63(3), 581–592.

Sawa, T. (1969): "The Exact Sampling Distribution of Ordinary Least Squares and Two-Stage Least Squares Estimators," *Journal of the American Statistical Association*, 64(327), 923–937.

Seaman, S., J. Galati, D. Jackson, and J. Carlin (2013): "What Is Meant by 'Missing at Random'?" *Statistical Science*, 28(2), 257–268.

Strauss, J., K. Beegle, B. Sikoki, A. Dwiyanto, Y. Herawati, and F. Witoelar (2004): *The Third Wave of the Indonesia Family Life Survey (IFLS3): Overview and Field Report.* RAND, WR-144/1-NIA/NICHD.

Strauss, J., F. Witoelar, B. Sikoki, and A. M. Wattie (2009): *The Fourth Wave of the Indonesia Family Life Survey (IFLS4): Overview and Field Report.* RAND, WR-675/1-NIA/NICHD.

Sun, B., and E. J. Tchetgen Tchetgen (2018): "On Inverse Probability Weighting for Non-monotone Missing at Random Data," *Journal of the American Statistical Association*, 113(521), 369–379.

Tsiatis, A. A. (2006): *Semiparametric Theory and Missing Data.* Springer, New York.

Wooldridge, J. M. (2007): "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics*, 141(2), 1281–1301.

Xiao, Z. (2020): "Efficient GMM Estimation with Singular System of Moment Conditions," *Statistical Theory and Related Fields*, 4(2), 172–178.

Yang, S. (2026): "A Doubly Robust GMM Estimator for Sequential Non-monotone Missingness," arXiv:2201.01010v3.

Zhou, Y., R. J. A. Little, and J. D. Kalbfleisch (2010): "Block-Conditional Missing at Random Models for Missing Data," *Statistical Science*, 25(4), 517–532.

A Proofs

Proof of Lemma 1.

We first obtain the joint observation probability and then use iterated expectations to recover the complete-data moments. Fix a pattern j. By the conditional probability chain rule,

$$P(g_D = 1, \lambda_Z = s_j|W) = P(\lambda_Z = s_j|W, g_D = 1)P(g_D = 1|W)$$

Assumption 1(i) gives

$$P(g_D = 1|W) = P(g_D = 1|V) = p_D(V)$$

and Assumption 1(ii) gives

$$P(\lambda_Z = s_j|W, g_D = 1) = P(\lambda_Z = s_j|D, V, g_D = 1) = p_{Z,j}(D, V)$$

Hence, by multiplication, Equation (6) holds. The law of iterated expectations gives Equation (7):

$$\begin{aligned} E[g_D h(W)|V] &= E\{E[g_D h(W)|W]|V\} \\ &= E\{h(W)E[g_D|W]|V\} \\ &= p_D(V)E[h(W)|V] \end{aligned}$$

Dividing by $p_D(V)$ and integrating proves Equation (9). Applying the same argument conditional on (D, V) gives Equation (8):

$$E[g_D g_{Z,j} h(W)|D, V] = E\{h(W)E[g_D g_{Z,j}|W]|D, V\} = p_D(V)p_{Z,j}(D, V)E[h(W)|D, V]$$

Dividing by $p_D(V)p_{Z,j}(D, V)$ and integrating proves Equation (10). ■

Proof of Lemma 2.

We show that the selection matrices jointly retain all components of the moment vector, establish the rank of the stacked Jacobian, and construct weights satisfying Assumption 6.

Step 1: establish the joint null space of the selection matrices.

Consider a vector $a = (a_Z', a_X')' \in \mathbb{R}^{k+q}$ that satisfies

$$\tilde{S}_j a = 0 \qquad \text{for every } j$$

Every $\tilde{S}_j$ contains the full X block, so $a_X = 0$. By Assumption 3, for each instrument κ there is at least one active pattern j with $\kappa \in s_j$. The corresponding row of $S_j a_Z = 0$ implies $a_{Z,\kappa} = 0$. Thus $a = 0$.

Step 2: establish full column rank of the stacked Jacobian.

Suppose $\nu \in \mathbb{R}^{p+q}$ satisfies

$$\tilde{S}_j G\nu = 0 \qquad \text{for every } j$$

The preceding argument, applied to $a = G\nu$, gives $G\nu = 0$. Assumption 2 then implies $\nu = 0$. Hence $\begin{pmatrix} \tilde{S}_1 G \\ \vdots \\ \tilde{S}_J G \end{pmatrix}$ achieves full column rank $p + q$.

Step 3: construct weights with a nonsingular weighted Jacobian.

Define

$$[A_1, \ldots, A_J] = \begin{bmatrix} \tilde{S}_1 G \\ \vdots \\ \tilde{S}_J G \end{bmatrix}'$$

Then

$$\bar{A}G = \sum_{j=1}^{J} A_j \tilde{S}_j G = \begin{bmatrix} \tilde{S}_1 G \\ \vdots \\ \tilde{S}_J G \end{bmatrix}' \begin{bmatrix} \tilde{S}_1 G \\ \vdots \\ \tilde{S}_J G \end{bmatrix}$$

which is positive definite since the matrix has full column rank. Hence, $\bar{A}G$ is nonsingular. ■

Proof of Theorem 1.

The inverse-probability identities (Lemma 1) reduce the observed moments to the complete-data moments. The rank conditions (Lemma 2) then establish uniqueness.

For part (a), apply Lemma 1 with $h(W) = \tilde{S}_j \tau(\theta)$ for each j:

$$E[w_j \tilde{S}_j \tau(\theta)] = \tilde{S}_j E[\tau(\theta)] = \tilde{S}_j G(\theta - \theta_0)$$

where the second equality holds by Equation (3). Stacking over j proves that

$$\begin{pmatrix} E[w_1 \tilde{S}_1 \tau(\theta)] \\ \vdots \\ E[w_J \tilde{S}_J \tau(\theta)] \end{pmatrix} = \begin{pmatrix} \tilde{S}_1 G \\ \vdots \\ \tilde{S}_J G \end{pmatrix} (\theta - \theta_0)$$

By Lemma 2, the stacked matrix has full column rank. Its null space is therefore trivial, which proves uniqueness in part (a).

For part (b), multiply each pattern equation by A_j and sum over j to obtain

$$E[m(\theta)] = \sum_{j=1}^{J} A_j \tilde{S}_j G(\theta - \theta_0) = \bar{A} G(\theta - \theta_0)$$

If $\bar{A}G$ is nonsingular, $\bar{A}G(\theta - \theta_0) = 0$ implies $\theta = \theta_0$. ■

Proof of Lemma 3.

The affine form of the moment function gives the estimator and its estimation error directly. By Equation (3),

$$\begin{aligned} \hat{g}_n(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} A_j \hat{w}_{ij} \tilde{S}_j (t_i - T_i \theta) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} A_j \hat{w}_{ij} \tilde{S}_j t_i - \hat{Q}_n \theta \end{aligned}$$

Thus $\hat{g}_n(\theta) = 0$ is equivalent to $\hat{Q}_n \theta = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} A_j \hat{w}_{ij} \tilde{S}_j t_i$, proving Equation (13).

Since

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} A_j \hat{w}_{ij} \tilde{S}_j t_i - \hat{Q}_n \theta_0 = \hat{g}_n(\theta_0)$$

subtracting θ_0 from Equation (13) proves

$$\hat{\theta} - \theta_0 = \hat{Q}_n^{-1} \hat{g}_n(\theta_0)$$ ■

Proof of Theorem 2.

We first show that replacing the true probabilities by their estimates changes the sample moments and Jacobian by $o_p(1)$. The law of large numbers and the closed-form solution then yield consistency.

Step 1: control the error from estimated probabilities.

Since both true and estimated probabilities are bounded away from zero under Assumptions 4 and 7, the map $(a, b) \to 1/(ab)$ is Lipschitz on the relevant compact set. Therefore, when $g_{D,i}g_{Z,ij} = 1$,

$$\begin{aligned} |\hat{w}_{ij} - w_{ij}| &\leqslant C(|\hat{p}_D(V_i) - p_D(V_i)| + |\hat{p}_{Z,j}(D_i, V_i) - p_{Z,j}(D_i, V_i)|) \\ &\leqslant C(\|\hat{p}_D - p_D\|_\infty + \|\hat{p}_{Z,j} - p_{Z,j}\|_\infty) = o_p(1) \end{aligned} \tag{36}$$

where C is deterministic.

Since J is finite, the matrices A_j and $\tilde{S}_j$ are bounded, and Assumption 5 implies the necessary first moments, Equation (36) and the law of large numbers give

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} A_j(\hat{w}_{ij} - w_{ij})\tilde{S}_j t_i = o_p(1)$$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} A_j(\hat{w}_{ij} - w_{ij})\tilde{S}_j T_i = o_p(1)$$

Step 2: obtain the probability limits of the sample moments and Jacobian.

For the true-weight parts, Lemma 1 and the law of large numbers give

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} A_j w_{ij}\tilde{S}_j t_i \to_p \bar{A}E[t]$$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} A_j w_{ij}\tilde{S}_j T_i \to_p \bar{A}E[T] = -\bar{A}G$$

Consequently,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} A_j \hat{w}_{ij}\tilde{S}_j t_i \to_p \bar{A}E[t], \qquad \hat{Q}_n \to_p -\bar{A}G$$

Step 3: apply the closed-form solution.

Assumption 6 makes the second limit nonsingular, so $\hat{Q}_n$ is nonsingular with probability approaching one. At the truth, since $E[\tau(\theta_0)] = 0$, we have

$$E[t] = E[T]\theta_0$$

Hence,

$$\begin{aligned}\hat{\theta} &= \hat{Q}_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} A_j \hat{w}_{ij}\tilde{S}_j t_i \\ &\to_p (-\bar{A}G)^{-1}\bar{A}E[t] \\ &= (-\bar{A}G)^{-1}\bar{A}E[T]\theta_0 \\ &= \theta_0\end{aligned}$$

■

Proof of Theorem 3.

We use the linear representation of the estimator, compute the covariance of its weighted moments, and apply the central limit theorem.

Step 1: obtain the linear representation and Jacobian limit.

Consistency follows from Theorem 2 after setting the estimated probabilities equal to their true values. By Lemma 3,

$$\sqrt{n}(\hat{\theta}_{known} - \theta_0) = (\hat{Q}_n^{known})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i^{known}$$

By Lemma 1 and the law of large numbers,

$$\hat{Q}_n^{known} \to_p \sum_{j=1}^{J} A_j \tilde{S}_j E[T] = \bar{A}E[T] = -\bar{A}G$$

Step 2: compute the covariance of the weighted moments.

The inverse-probability identities and $E[\tau(\theta_0)] = 0$ give $E[\psi^{known}] = 0$. Hence,

$$\Omega_{known} = E[\psi^{known}(\psi^{known})']$$

Expanding the right hand side gives

$$\Omega_{known} = \sum_{j=1}^{J}\sum_{l=1}^{J} A_j E[\frac{g_D^2 g_{Z,j} g_{Z,l}}{p_D(V)^2 p_{Z,j}(D,V) p_{Z,l}(D,V)}\tilde{S}_j \tau(\theta_0)\tau(\theta_0)'\tilde{S}_l']A_l'$$

The missingness patterns of the instruments are partitioned in a mutually exclusive way,

so $g_{Z,j}g_{Z,l} = 0$ when $j \neq l$. Also $g_D^2 = g_D$ and $g_{Z,j}^2 = g_{Z,j}$. Hence

$$\Omega_{known} = \sum_{j=1}^{J} A_j E[\frac{g_D g_{Z,j}}{p_D(V)^2 p_{Z,j}(D,V)^2} \tilde{S}_j \tau(\theta_0)\tau(\theta_0)' \tilde{S}_j'] A_j'$$

Conditioning on W and applying Lemma 1 gives

$$\begin{aligned} E[\frac{g_D g_{Z,j}}{p_D(V)^2 p_{Z,j}(D,V)^2} \tilde{S}_j \tau(\theta_0)\tau(\theta_0)' \tilde{S}_j'] &= E[\frac{E[g_D g_{Z,j}|W]}{p_D(V)^2 p_{Z,j}(D,V)^2} \tilde{S}_j \tau(\theta_0)\tau(\theta_0)' \tilde{S}_j'] \\ &= E[\frac{1}{p_D(V) p_{Z,j}(D,V)} \tilde{S}_j \tau(\theta_0)\tau(\theta_0)' \tilde{S}_j'] \end{aligned}$$

Therefore,

$$\Omega_{known} = \sum_{j=1}^{J} A_j E[\frac{1}{p_j} \tilde{S}_j \tau(\theta_0)\tau(\theta_0)' \tilde{S}_j'] A_j'$$

Overlap and Assumption 5 give that this covariance matrix is finite.

Step 3: obtain the limiting distribution.

The central limit theorem gives

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^{known} \to^d N(0, \Omega_{known})$$

Since $(-\bar{A}G)^{-1} = -(\bar{A}G)^{-1}$, Slutsky's theorem gives

$$\sqrt{n}(\hat{\theta}_{known} - \theta_0) = -(\bar{A}G)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^{known} + o_p(1)$$

The covariance of the limiting distribution is

$$[-(\bar{A}G)^{-1}]\Omega_{known}[-(\bar{A}G)^{-1}]' = (\bar{A}G)^{-1}\Omega_{known}[(\bar{A}G)^{-1}]'$$

Hence,

$$\sqrt{n}(\hat{\theta}_{known} - \theta_0) \to^d N(0, (\bar{A}G)^{-1}\Omega_{known}(\bar{A}G)^{-1\prime}) \qquad \blacksquare$$

Throughout the following four proofs, $E[\tau(\theta_0)] = 0$, $G = -E[T]$, and the probability estimates are trimmed as in Section 4. The number of patterns J and the matrices A_j are fixed. Summation with a zero observation indicator are defined to be zero. Constants C in inequalities may change from line to line.

Proof of Lemma 4.

We verify the decomposition by substituting the conditional-mean residuals and collecting terms. The definitions imply $e_D = \mu_1 - \mu_0$ and $e_Z = \tau(\theta_0) - \mu_1$. Therefore,

$$\begin{aligned}\psi_j^* + r_{D,j} + r_{Z,j} &= \tilde{S}_j\mu_0 + \frac{g_D}{p_D}\tilde{S}_j(\mu_1 - \mu_0) + w_j\tilde{S}_j\{\tau(\theta_0) - \mu_1\} \\ &\quad + \frac{g_D - p_D}{p_D}\tilde{S}_j\mu_0 + \frac{g_D}{p_D}\frac{g_{Z,j} - p_{Z,j}}{p_{Z,j}}\tilde{S}_j\mu_1 \\ &= \frac{g_D}{p_D}\tilde{S}_j\mu_1 + w_j\tilde{S}_j\tau(\theta_0) - w_j\tilde{S}_j\mu_1 + \left(w_j - \frac{g_D}{p_D}\right)\tilde{S}_j\mu_1 \\ &= w_j\tilde{S}_j\tau(\theta_0).\end{aligned}$$

The denominators are positive by Assumption 4. Premultiplying by A_j and summing over strata gives the corresponding decomposition of the stacked moment:

$$m(\theta_0) - \sum_j A_j r_{D,j} - \sum_j A_j r_{Z,j} = \psi^*. \tag{37}$$

■

Proof of Theorem 4.

The proof uses an adjusted-moment expansion and the corresponding GMM linearization, following the approach in Cattaneo (2010, Theorem 4) and Muris (2020, Appendix A.3). We first establish consistency and convergence of the sample Jacobian. We then obtain an asymptotic linear representation of the feasible moment, compute its covariance, and apply the central limit theorem.

Step 1: convergence of the sample Jacobian and consistency.

We show that replacing the true probabilities by their estimates changes the sample Jacobian and the sample moment at θ_0 by $o_p(1)$. By Assumption 4 and the trimming rule in Section 4, both the true and fitted probability products are bounded away from zero. By some manipulation of the algebra, we have

$$\frac{1}{\hat{p}_D\hat{p}_{Z,j}} - \frac{1}{p_D p_{Z,j}} = \left(\frac{1}{\hat{p}_D} - \frac{1}{p_D}\right)\frac{1}{\hat{p}_{Z,j}} + \frac{1}{p_D}\left(\frac{1}{\hat{p}_{Z,j}} - \frac{1}{p_{Z,j}}\right).$$

Thus, on an observation contributing to pattern j,

$$|\hat{w}_{ij} - w_{ij}| \leq C\left\{\|\hat{p}_D - p_D\|_\infty + \max_l \|\hat{p}_{Z,l} - p_{Z,l}\|_\infty\right\} = o_p(1), \tag{38}$$

uniformly over the sample and the finite set of patterns. Both weights are zero when the corresponding observation indicator is zero.

Since $\tau_i(\theta_0) = t_i - T_i\theta_0$, Assumption 5 gives

$$E\|T\| < \infty, \qquad E\|\tau(\theta_0)\| < \infty, \qquad E\|\tau(\theta_0)\|^2 < \infty.$$

The bound Equation (38) and the law of large numbers imply

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i,j} A_j(\hat{w}_{ij} - w_{ij})\tilde{S}_j T_i\right\| &\leq C\left\{\|\hat{p}_D - p_D\|_\infty + \max_j \|\hat{p}_{Z,j} - p_{Z,j}\|_\infty\right\}\frac{1}{n}\sum_i \|T_i\| \\ &= o_p(1)O_p(1) = o_p(1).\end{aligned} \tag{39}$$

The same argument, with T_i replaced by $\tau_i(\theta_0)$, gives

$$\frac{1}{n}\sum_{i,j} A_j(\hat{w}_{ij} - w_{ij})\tilde{S}_j\tau_i(\theta_0) = o_p(1). \tag{40}$$

Lemma 1 gives the true-weight identities

$$E[w_j\tilde{S}_j T] = \tilde{S}_j E[T], \qquad E[w_j\tilde{S}_j\tau(\theta_0)] = \tilde{S}_j E[\tau(\theta_0)] = 0.$$

Consequently, the true-weight sample averages obey

$$\frac{1}{n}\sum_{i,j} A_j w_{ij}\tilde{S}_j T_i \to_p \bar{A}E[T] = -\bar{A}G, \qquad \frac{1}{n}\sum_{i,j} A_j w_{ij}\tilde{S}_j\tau_i(\theta_0) \to_p 0.$$

Adding the approximation errors to the true-weight averages yields the fitted-weight limits:

$$\begin{aligned}\hat{Q}_n &= \frac{1}{n}\sum_{i,j} A_j w_{ij}\tilde{S}_j T_i + \frac{1}{n}\sum_{i,j} A_j(\hat{w}_{ij} - w_{ij})\tilde{S}_j T_i \\ &= -\bar{A}G + o_p(1),\end{aligned} \tag{41}$$

and

$$\frac{1}{n}\sum_{i,j} A_j\hat{w}_{ij}\tilde{S}_j\tau_i(\theta_0) = \frac{1}{n}\sum_{i,j} A_j w_{ij}\tilde{S}_j\tau_i(\theta_0) + o_p(1) = o_p(1).$$

Nonsingularity of $\bar{A}G$ implies that $\hat{Q}_n$ is nonsingular with probability approaching one and

$$\hat{Q}_n^{-1} \to_p -(\bar{A}G)^{-1}.$$

Using the identity in Lemma 3,

$$\hat{\theta} - \theta_0 = \hat{Q}_n^{-1}\frac{1}{n}\sum_{i,j} A_j\hat{w}_{ij}\tilde{S}_j\tau_i(\theta_0), \tag{42}$$

we obtain $\hat{\theta} \to_p \theta_0$. Thus the feasible Jacobian has a nonsingular probability limit, and

the estimator is consistent.

Step 2: account for the first-step estimation effect.

We express the feasible sample moment as an average of adjusted moments. For each stratum, Assumption 8(v) and Lemma 4 imply

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_i \hat{w}_{ij}\tilde{S}_j\tau_i(\theta_0) &= \frac{1}{\sqrt{n}}\sum_i\{w_{ij}\tilde{S}_j\tau_i(\theta_0) - r_{D,ij} - r_{Z,ij}\} + o_p(1)\\ &= \frac{1}{\sqrt{n}}\sum_i \psi^*_{ij} + o_p(1).\end{aligned}$$

Premultiplying by A_j and summing over finitely many j gives

$$\frac{1}{\sqrt{n}}\sum_{i,j} A_j\hat{w}_{ij}\tilde{S}_j\tau_i(\theta_0) = \frac{1}{\sqrt{n}}\sum_i\sum_j A_j\psi^*_{ij} + o_p(1) = \frac{1}{\sqrt{n}}\sum_i \psi^*_i + o_p(1). \tag{43}$$

Hence ψ^*_i is the first-order contribution of observation i to the feasible sample moment.

Step 3: establish the mean and covariance of ψ^*.

We use the conditional-mean restrictions to show that the three components of ψ^* have mean zero and are pairwise uncorrelated. Their variances then just add to the covariance matrix in the theorem. The definitions of the conditional projections and the maintained moment restriction give

$$E[\mu_0] = 0, \qquad E[e_D|V] = 0, \qquad E[e_Z|D,V] = 0.$$

By SMAR, conditioning on the realized instrument pattern among observations with D observed does not change the conditional mean of e_Z:

$$E[e_Z|D,V,g_D = 1, g_{Z,j} = 1] = 0.$$

Then, we obtain:

$$E\left[\frac{g_D}{p_D}\bar{A}e_D\middle|V\right] = 0, \qquad E\left[\sum_j w_j A_j\tilde{S}_j e_Z\middle|D,V,g_D\right] = 0. \tag{44}$$

The first equation uses $E[g_D|D,V] = p_D(V)$ and $E[e_D|V] = 0$. The second is immediate when $g_D = 0$; when $g_D = 1$, it follows by averaging over $g_{Z,j}$ and the selected instrument distribution. These facts show $E[\psi^*] = 0$.

The three components in Equation (19) are pairwise uncorrelated. The first is measurable

with respect to V, so its covariance with the second is zero by the first one in Equation (44). Both the first and second are measurable with respect to (D, V, g_D), so their covariances with the third are zero by the second one. Integrability of these products follows from square integrability, verified below.

Their three covariance matrices can now be computed separately. First,

$$\mathrm{Var}(\bar{A}\mu_0) = \bar{A}E[\mu_0\mu_0']\bar{A}' = \bar{A}M_0\bar{A}'.$$

Second, because $g_D^2 = g_D$,

$$\begin{aligned}\mathrm{Var}\left(\frac{g_D}{p_D}\bar{A}e_D\right) &= \bar{A}E\left[\frac{g_D}{p_D^2}e_De_D'\right]\bar{A}' \\ &= \bar{A}E\left[\frac{E[g_D|D,V]}{p_D^2}e_De_D'\right]\bar{A}' \\ &= \bar{A}E[p_D^{-1}e_De_D']\bar{A}' = \bar{A}M_D\bar{A}'.\end{aligned}$$

Third,

$$\begin{aligned}\mathrm{Var}\left(\sum_j w_jA_j\tilde{S}_je_Z\right) &= \sum_j\sum_l A_j\tilde{S}_jE[w_jw_le_Ze_Z']\tilde{S}_l'A_l' \\ &= \sum_j A_j\tilde{S}_jE\left[\frac{g_Dg_{Z,j}}{p_j^2}e_Ze_Z'\right]\tilde{S}_j'A_j' \\ &= \sum_j A_j\tilde{S}_jE\left[\frac{E[g_Dg_{Z,j}|W]}{p_j^2}e_Ze_Z'\right]\tilde{S}_j'A_j' \\ &= \sum_j A_j\tilde{S}_jE[p_j^{-1}e_Ze_Z']\tilde{S}_j'A_j' \\ &= \sum_j A_j\tilde{S}_jM_{Z,j}\tilde{S}_j'A_j'.\end{aligned}$$

The second line uses $g_{Z,j}g_{Z,l} = 0$ for $j \neq l$, and the fourth uses $E[g_Dg_{Z,j}|W] = p_j$.

Adding the three uncorrelated covariance components proves

$$\mathrm{Var}(\psi^*) = \bar{A}M_0\bar{A}' + \bar{A}M_D\bar{A}' + \sum_j A_j\tilde{S}_jM_{Z,j}\tilde{S}_j'A_j' = \Omega^*.$$

Finally, Jensen's inequality gives

$$E\|\mu_1\|^2 \leq E\|\tau(\theta_0)\|^2 < \infty, \qquad E\|\mu_0\|^2 \leq E\|\tau(\theta_0)\|^2 < \infty.$$

It follows that e_D and e_Z have finite second moments. Overlap bounds p_D^{-1} and p_j^{-1}, and

A_j and $\tilde{S}_j$ are fixed finite matrices. Hence $E\|\psi^*\|^2 < \infty$.

Step 4: conclude the asymptotic distribution.

Combining the feasible-moment expansion with the limit of the inverse Jacobian gives the influence function of $\hat{\theta}$. The vectors ψ_i^* depend only on the individual observations and true population functions. They are therefore independent and identically distributed, have mean zero, and have covariance Ω^*. The multivariate central limit theorem gives

$$\frac{1}{\sqrt{n}} \sum_i \psi_i^* \to^d N(0, \Omega^*).$$

Multiplying Equation (42) by $\sqrt{n}$, using Equation (43), and applying Equation (41),

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta_0) &= \hat{Q}_n^{-1} \left\{ \frac{1}{\sqrt{n}} \sum_i \psi_i^* + o_p(1) \right\} \\ &= -(\bar{A}G)^{-1} \frac{1}{\sqrt{n}} \sum_i \psi_i^* + o_p(1). \end{aligned}$$

The limiting covariance is

$$[-(\bar{A}G)^{-1}]\Omega^*[-(\bar{A}G)^{-1}]' = (\bar{A}G)^{-1}\Omega^*\left[(\bar{A}G)^{-1}\right]' = \Gamma^*.$$

This proves the theorem. ■

Proof of Proposition 1.

We establish the variance reduction by separating the adjusted moment from orthogonal selection residuals, extending the estimated-propensity comparisons in Hirano, Imbens, and Ridder (2003, Section 3) and the augmentation representation in Graham (2011, equation (3)) to the two-layer missingness structure. We decompose the known-probability moment into the adjusted moment and two selection residuals. Orthogonality makes the covariance difference equal to the sum of the residual covariance matrices.

Step 1: establish the orthogonal decomposition.

All objects are evaluated at θ_0. Recall the selection residuals defined in Lemma 4:

$$r_{D,j} = \frac{g_D - p_D}{p_D} \tilde{S}_j \mu_0, \qquad r_{Z,j} = \frac{g_D}{p_D} \frac{g_{Z,j} - p_{Z,j}}{p_{Z,j}} \tilde{S}_j \mu_1.$$

The algebra in Equation (37) is equivalently

$$m(\theta_0) = \psi^* + \sum_j A_j r_{D,j} + \sum_j A_j r_{Z,j}. \tag{45}$$

The last two terms are

$$\sum_j A_j r_{D,j} = \frac{g_D - p_D}{p_D} \bar{A}\mu_0, \qquad \sum_j A_j r_{Z,j} = \frac{g_D}{p_D} \sum_j \frac{g_{Z,j} - p_{Z,j}}{p_{Z,j}} A_j \tilde{S}_j \mu_1.$$

We first show that they have mean zero, are uncorrelated with each other, and are each uncorrelated with ψ^*.

By the definitions of the probabilities,

$$E\left[\sum_j A_j r_{D,j} \middle| V\right] = 0, \qquad E\left[\sum_j A_j r_{Z,j} \middle| D, V, g_D\right] = 0.$$

For the second statement, the expression is zero when $g_D = 0$; when $g_D = 1$, use $E[g_{Z,j}|D, V, g_D = 1] = p_{Z,j}$. Both sums therefore have unconditional mean zero. The first sum is measurable with respect to (D, V, g_D), so the two sums are uncorrelated.

The first sum is uncorrelated with $\bar{A}\mu_0$ by its conditional mean zero given V. For its covariance with the e_D component of ψ^*,

$$\begin{aligned} &E\left[\frac{g_D}{p_D}\bar{A}e_D\left(\frac{g_D - p_D}{p_D}\bar{A}\mu_0\right)'\right] \\ &\quad = E\left[\frac{1 - p_D}{p_D}\bar{A}e_D\mu_0'\bar{A}'\right] \\ &\quad = E\left[\frac{1 - p_D}{p_D}\bar{A}E[e_D|V]\mu_0'\bar{A}'\right] = 0. \end{aligned}$$

The first equality conditions on (D, V) and uses $E[g_D(g_D - p_D)|D, V] = p_D(1 - p_D)$.

The second sum is uncorrelated with both $\bar{A}\mu_0$ and $(g_D/p_D)\bar{A}e_D$, because those vectors are measurable with respect to (D, V, g_D).

To deal with the last component of ψ^*, condition on (D, V, g_D, λ_Z). On $g_D = 1$, SMAR gives

$$E[e_Z|D, V, g_D = 1, \lambda_Z] = 0.$$

All remaining factors in the products with $r_{D,j}$ or $r_{Z,j}$ are measurable with respect to these conditioning variables. On $g_D = 0$, the last component of ψ^* is zero. Therefore it is uncorrelated with both selection-residual sums. This establishes all required cross-

covariance cancellations.

Since $\text{Var}\{m(\theta_0)\} = \Omega_{known}$ by Theorem 3 and $\text{Var}(\psi^*) = \Omega^*$ by Theorem 4, Equation (45) implies

$$\Omega_{known} - \Omega^* = \text{Var}\left(\sum_j A_j r_{D,j}\right) + \text{Var}\left(\sum_j A_j r_{Z,j}\right) \succeq 0. \tag{46}$$

It remains to express these two covariance matrices in the notation of the proposition.

Step 2: compute the contribution from observing D.

The first residual is proportional to the centered indicator $g_D - p_D$. Its covariance follows from the conditional Bernoulli variance. Condition on V. Since $E[(g_D - p_D)^2|V] = p_D(1-p_D)$,

$$\begin{aligned}\text{Var}\left(\sum_j A_j r_{D,j}\right) &= E\left[\frac{(g_D - p_D)^2}{p_D^2}\bar{A}\mu_0\mu_0'\bar{A}'\right] \\ &= E\left[\frac{1-p_D}{p_D}\bar{A}\mu_0\mu_0'\bar{A}'\right].\end{aligned} \tag{47}$$

Step 3: compute the contribution from the joint instrument pattern.

The multinomial covariance formula gives the diagonal and cross-pattern terms in the second residual covariance. Conditional on $(D, V, g_D = 1)$, $(g_{Z,1}, \ldots, g_{Z,J})$ is one multinomial draw. Thus

$$E[(g_{Z,j} - p_{Z,j})(g_{Z,l} - p_{Z,l})|D, V, g_D = 1] = \mathbf{1}\{j = l\}p_{Z,j} - p_{Z,j}p_{Z,l}.$$

Expanding the covariance and using $g_D^2 = g_D$ gives

$$\begin{aligned}&\text{Var}\left(\sum_j A_j r_{Z,j}\right) \\ &= \sum_j \sum_l E\left[\frac{g_D}{p_D^2}\frac{(g_{Z,j} - p_{Z,j})(g_{Z,l} - p_{Z,l})}{p_{Z,j}p_{Z,l}}A_j\tilde{S}_j\mu_1\mu_1'\tilde{S}_l'A_l'\right] \\ &= \sum_j \sum_l E\left[\frac{1}{p_D}\frac{\mathbf{1}\{j = l\}p_{Z,j} - p_{Z,j}p_{Z,l}}{p_{Z,j}p_{Z,l}}A_j\tilde{S}_j\mu_1\mu_1'\tilde{S}_l'A_l'\right].\end{aligned}$$

The last equality first averages over the multinomial pattern among observations with D observed, and then uses $E[g_D|D, V] = p_D(V)$. Separating the diagonal and off-diagonal

terms yields

$$\begin{aligned}\operatorname{Var}\left(\sum_j A_j r_{Z,j}\right) = &\sum_j E\left[\frac{1-p_{Z,j}}{p_D p_{Z,j}} A_j \tilde{S}_j \mu_1 \mu_1' \tilde{S}_j' A_j'\right] \\ &- \sum_{j\neq l} E\left[\frac{1}{p_D} A_j \tilde{S}_j \mu_1 \mu_1' \tilde{S}_l' A_l'\right].\end{aligned} \tag{48}$$

Substituting Equation (47) and Equation (48) into Equation (46) gives the asserted covariance difference and establishes its positive semidefiniteness.

Finally, premultiplication by $(\bar{A}G)^{-1}$ and postmultiplication by its transpose preserve positive semidefiniteness. This proves the comparison of the asymptotic covariance matrices of the parameter estimators. ■

Proof of Proposition 2.

We establish componentwise covariance consistency and then apply the sandwich mapping, following the variance-estimation argument in Muris (2020, Appendix A.6). The fitted-mean conditions Equation (24) control residual estimation errors, while the conditions of Theorem 4 control the estimated probabilities and Jacobian.

Step 1: fitted residuals approximate their true observed counterparts.

We first show that the fitted residuals converge in sample mean square on the observations contributing to each covariance component. On $g_{D,i} = 1$,

$$\hat{e}_{D,i} - e_D(D_i, V_i) = (\hat{\mu}_{1,i} - \mu_1(D_i, V_i)) - (\hat{\mu}_{0,i} - \mu_0(V_i)).$$

The inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and Equation (24) give

$$\frac{1}{n}\sum_i g_{D,i}\|\hat{e}_{D,i} - e_{D,i}\|^2 = o_p(1). \tag{49}$$

For a contributing observation in pattern j, linearity gives

$$\tilde{S}_j\hat{e}_{Z,i} - \tilde{S}_j e_{Z,i} = -\tilde{S}_j T_i(\hat{\theta} - \theta_0) - \tilde{S}_j(\hat{\mu}_{1,i} - \mu_1(D_i, V_i)).$$

Since the Euclidean operator norm of the selection matrix $\tilde{S}_j$ is at most one,

$$\begin{aligned}\frac{1}{n}\sum_i g_{D,i} g_{Z,ij}\|\tilde{S}_j\hat{e}_{Z,i} - \tilde{S}_j e_{Z,i}\|^2 \leq{}& 2\|\hat{\theta} - \theta_0\|^2 \frac{1}{n}\sum_i \|T_i\|^2 \\ &+ \frac{2}{n}\sum_i g_{D,i}\|\hat{\mu}_{1,i} - \mu_1(D_i, V_i)\|^2 = o_p(1).\end{aligned}$$

Here $\hat{\theta} \to_p \theta_0$ and $n^{-1}\sum_i \|T_i\|^2 = O_p(1)$ follow from Theorem 4 and Assumption 5.

Jensen's inequality, the moment assumptions, and the law of large numbers give

$$\frac{1}{n}\sum_i \|\mu_{0,i}\|^2 = O_p(1), \qquad \frac{1}{n}\sum_i g_{D,i}\|e_{D,i}\|^2 = O_p(1), \qquad \frac{1}{n}\sum_i g_{D,i}g_{Z,ij}\|\tilde{S}_j e_{Z,i}\|^2 = O_p(1).$$

The corresponding averages of fitted squared norms are also $O_p(1)$, by the squared-error bounds just established.

Step 2: convergence of $\hat{M}_0$.

We separate the fitted-mean error from the sample average of the true outer products. For conformable vectors, the identity $aa' - bb' = (a-b)a' + b(a-b)'$ implies

$$\|aa' - bb'\| \le \|a-b\|(\|a\| + \|b\|).$$

Cauchy–Schwarz therefore gives

$$\begin{aligned}&\left\|\hat{M}_0 - \frac{1}{n}\sum_i \mu_{0,i}\mu_{0,i}'\right\| \\ &\quad\le \left\{\frac{1}{n}\sum_i \|\hat{\mu}_{0,i} - \mu_{0,i}\|^2\right\}^{1/2}\left\{\frac{1}{n}\sum_i (\|\hat{\mu}_{0,i}\| + \|\mu_{0,i}\|)^2\right\}^{1/2} = o_p(1).\end{aligned}$$

Since $n^{-1}\sum_i \mu_{0,i}\mu_{0,i}' \to_p M_0$, we obtain $\hat{M}_0 \to_p M_0$.

Step 3: convergence of $\hat{M}_D$.

We separate the residual estimation error from the probability estimation error. Adding and subtracting the true-residual, fitted-probability average gives

$$\begin{aligned}&\hat{M}_D - \frac{1}{n}\sum_i \frac{g_{D,i}}{p_D(V_i)^2} e_{D,i}e_{D,i}' \\ &\quad= \frac{1}{n}\sum_i \frac{g_{D,i}}{\hat{p}_D(V_i)^2}\{\hat{e}_{D,i}\hat{e}_{D,i}' - e_{D,i}e_{D,i}'\} \\ &\qquad+ \frac{1}{n}\sum_i g_{D,i}\left\{\frac{1}{\hat{p}_D(V_i)^2} - \frac{1}{p_D(V_i)^2}\right\} e_{D,i}e_{D,i}'.\end{aligned}$$

The first term is $o_p(1)$ by bounded inverse probabilities, Equation (49), and the same outer-product/Cauchy–Schwarz bound as in Step 2. The norm of the second is bounded by

$$\sup_v \left|\frac{1}{\hat{p}_D(v)^2} - \frac{1}{p_D(v)^2}\right| \frac{1}{n}\sum_i g_{D,i}\|e_{D,i}\|^2 = o_p(1)O_p(1) = o_p(1).$$

The supremum tends to zero because the reciprocal-square map is Lipschitz away from zero and the probability estimate is uniformly consistent. Finally,

$$E\left[\frac{g_D}{p_D^2}e_D e_D'\right] = E\left[\frac{E[g_D|D,V]}{p_D^2}e_D e_D'\right] = E[p_D^{-1}e_D e_D'] = M_D.$$

The law of large numbers proves $\hat{M}_D \to_p M_D$.

Step 4: convergence of the selected instrument covariance blocks.

The same decomposition applies to each selected instrument covariance block, with observations restricted to pattern j:

$$\begin{aligned}
&\widehat{\tilde{S}_j M_{Z,j} \tilde{S}_j'} - \frac{1}{n}\sum_i \frac{g_{D,i}g_{Z,ij}}{p_j(D_i,V_i)^2}(\tilde{S}_j e_{Z,i})(\tilde{S}_j e_{Z,i})' \\
&= \frac{1}{n}\sum_i \frac{g_{D,i}g_{Z,ij}}{\hat{p}_j(D_i,V_i)^2}\left\{(\tilde{S}_j \hat{e}_{Z,i})(\tilde{S}_j \hat{e}_{Z,i})' - (\tilde{S}_j e_{Z,i})(\tilde{S}_j e_{Z,i})'\right\} \\
&\quad + \frac{1}{n}\sum_i g_{D,i}g_{Z,ij}\left\{\frac{1}{\hat{p}_j(D_i,V_i)^2} - \frac{1}{p_j(D_i,V_i)^2}\right\}(\tilde{S}_j e_{Z,i})(\tilde{S}_j e_{Z,i})'.
\end{aligned}$$

The first term is $o_p(1)$ by the selected-residual bound in Step 1 and Cauchy–Schwarz. The second is $o_p(1)$ because $\hat{p}_j = \hat{p}_D \hat{p}_{Z,j}$ converges uniformly to p_j and the products are bounded away from zero. Conditioning on W gives

$$\begin{aligned}
E\left[\frac{g_D g_{Z,j}}{p_j^2}\tilde{S}_j e_Z e_Z' \tilde{S}_j'\right] &= E\left[\frac{E[g_D g_{Z,j}|W]}{p_j^2}\tilde{S}_j e_Z e_Z' \tilde{S}_j'\right] \\
&= \tilde{S}_j E[p_j^{-1} e_Z e_Z'] \tilde{S}_j' = \tilde{S}_j M_{Z,j} \tilde{S}_j'.
\end{aligned}$$

The law of large numbers consequently proves

$$\widehat{\tilde{S}_j M_{Z,j} \tilde{S}_j'} \to_p \tilde{S}_j M_{Z,j} \tilde{S}_j' \qquad \text{for every } j.$$

Step 5: the sandwich covariance and standard errors.

The componentwise limits imply consistency of $\hat{\Omega}^*$; convergence of the sample Jacobian then gives consistency of the sandwich covariance. Since the matrices A_j are fixed and the number of patterns is finite, substitution into Equation (23) yields

$$\hat{\Omega}^* \to_p \bar{A}(M_0 + M_D)\bar{A}' + \sum_j A_j \tilde{S}_j M_{Z,j} \tilde{S}_j' A_j' = \Omega^*.$$

By Equation (41) and continuous mapping,

$$\hat{\Gamma}^* = \hat{Q}_n^{-1}\hat{\Omega}^*\left[\hat{Q}_n^{-1}\right]' \to_p (\bar{A}G)^{-1}\Omega^*\left[(\bar{A}G)^{-1}\right]' = \Gamma^*.$$

For any parameter θ_a with $\Gamma^*_{aa} > 0$, Theorem 4 and Slutsky's theorem give

$$\frac{\hat{\theta}_a - \theta_{0,a}}{\sqrt{\hat{\Gamma}^*_{aa}/n}} \to^d N(0,1).$$

This proves the variance and standard-error statements. ■

Proof of Theorem 5.

The proof applies the optimal-GMM covariance comparison of Hansen (1982) to the sequentially adjusted moments, with stratum-specific combining matrices as in Muris (2020, Remark 3). The reduction of fixed linear redundancies uses the generalized-inverse identities discussed in Xiao (2020, Proposition 2.1(ii) and Theorem 2.1). Step 1 computes the covariance of the adjusted moments. Step 2 removes fixed linear redundancies and obtains the information matrix. Step 3 derives the minimum asymptotic covariance within the proposed IPW-GMM class.

Step 1. Compute the covariance of the adjusted moments.

We obtain each block of Σ^* by evaluating the covariances of the three layers in the moment decomposition. From Lemma 4,

$$\psi_j^* = \tilde{S}_j\mu_0 + \frac{g_D}{p_D}\tilde{S}_j e_D + w_j\tilde{S}_j e_Z.$$

Each component has mean zero. In particular, $E\mu_0 = 0$, $E[e_D \mid V] = 0$, and $E[e_Z \mid D, V] = 0$. Sequential recoverability gives

$$E\left[\frac{g_D}{p_D}e_D \,\middle|\, V\right] = 0, \qquad E[w_j\tilde{S}_j e_Z \mid D, V, g_D] = 0.$$

The same conditioning arguments used in Section 5.1 therefore make different layers uncorrelated for any two strata. For example,

$$E\left[\tilde{S}_j\mu_0\left(\frac{g_D}{p_D}\tilde{S}_l e_D\right)'\right] = \tilde{S}_j E\left[\mu_0 E\left[\frac{g_D}{p_D}e_D' \,\middle|\, V\right]\right]\tilde{S}_l' = 0.$$

Conditioning on (D, V, g_D) makes every cross-product involving an instrument-residual

layer and a preceding layer zero. The remaining same-layer terms are

$$E[(\tilde{S}_j\mu_0)(\tilde{S}_l\mu_0)'] = \tilde{S}_j M_0 \tilde{S}_l',$$
$$E\left[\frac{g_D}{p_D^2}\tilde{S}_j e_D e_D' \tilde{S}_l'\right] = \tilde{S}_j E[p_D^{-1} e_D e_D'] \tilde{S}_l' = \tilde{S}_j M_D \tilde{S}_l'.$$

For the final layer, $g_{Z,j}g_{Z,l} = 0$ for $j \neq l$. When $j = l$, Lemma 1 gives

$$\begin{aligned} E[w_j^2 \tilde{S}_j e_Z e_Z' \tilde{S}_j'] &= \tilde{S}_j E\left[\frac{E[g_D g_{Z,j} \mid W]}{p_D^2 p_{Z,j}^2} e_Z e_Z'\right] \tilde{S}_j' \\ &= \tilde{S}_j M_{Z,j} \tilde{S}_j'. \end{aligned}$$

Adding these terms gives Equation (26), including the covariances across strata. Bilinearity of covariance then gives Equation (27).

Step 2. Reduce the moment vector and obtain the information matrix.

We express the information matrix using the linearly independent moments and then recover it in terms of the original moment vector stacked across strata, define

$$F_j(\theta) = E[w_j \tilde{S}_j \tau(\theta)] = \tilde{S}_j G(\theta - \theta_0), \tag{50}$$

where the second equality follows from Lemma 1 and linearity of the moment function. Let $F(\theta)$ be the column stack of the $F_j(\theta)$ and put $r = \sum_j (k_j + q)$. Let the fixed deletion and reconstruction matrices be

$$R_0 \in \mathbb{R}^{r_0 \times r}, \qquad L_0 \in \mathbb{R}^{r \times r_0}, \qquad R_0 L_0 = I_{r_0},$$

so that the retained vector $F_r = R_0 F$ satisfies

$$F(\theta) = L_0 F_r(\theta) \tag{51}$$

for all nearby parameter values, with the same fixed linear identities holding for the adjusted moments. These identities define the redundancy reduction used in Assumption 9(ii).

Write $\psi_r = R_0 \begin{pmatrix} \psi_1^* \\ \vdots \\ \psi_J^* \end{pmatrix}$ and $\Sigma_r = \mathrm{Var}(\psi_r)$. Let $\partial F_r / \partial \theta'$ denote the retained population Jacobian at θ_0. The fixed identities for the adjusted moments and differentiation of

Equation (51) in θ give

$$\begin{pmatrix} \psi_1^* \\ \vdots \\ \psi_J^* \end{pmatrix} = L_0 \psi_r, \qquad \begin{pmatrix} \tilde{S}_1 G \\ \vdots \\ \tilde{S}_J G \end{pmatrix} = L_0 \frac{\partial F_r}{\partial \theta'}, \qquad \Sigma^* = L_0 \Sigma_r L_0'. \tag{52}$$

Assumption 9 gives $\Sigma_r \succ 0$. Assumptions 2 and 3 give full column rank of the unreduced Jacobian. Equation (52) then gives full column rank of the retained Jacobian.

Because L_0 has full column rank, $L_0^+ = (L_0' L_0)^{-1} L_0'$ and

$$(\Sigma^*)^+ = (L_0^+)' \Sigma_r^{-1} L_0^+, \qquad L_0' (\Sigma^*)^+ L_0 = \Sigma_r^{-1}. \tag{53}$$

For verification, multiplication of the first right-hand side by $L_0 \Sigma_r L_0'$ in either order gives the symmetric projection $L_0 L_0^+$. The two remaining Moore–Penrose identities follow from $L_0^+ L_0 = I_{r_0}$. Thus

$$\mathcal{I}_0 = \left(\frac{\partial F_r}{\partial \theta'} \right)' \Sigma_r^{-1} \frac{\partial F_r}{\partial \theta'} \succ 0. \tag{54}$$

This proves positive definiteness and the equality with Equation (28). It also proves that

$$\Sigma^* (\Sigma^*)^+ \begin{pmatrix} \tilde{S}_1 G \\ \vdots \\ \tilde{S}_J G \end{pmatrix} = \begin{pmatrix} \tilde{S}_1 G \\ \vdots \\ \tilde{S}_J G \end{pmatrix}, \tag{55}$$

which is used in the optimal-weighting proof.

For the repeated X equations, the reduction is particularly simple: retain their q rows in one stratum and delete their other copies. For every j,

$$F_{j,X}(\theta) = E[X \tilde{\epsilon}(\theta)], \qquad \psi_{j,X}^* = \mu_{0,X} + \frac{g_D}{p_D} e_{D,X},$$

where the second equality uses the fact that the components of e_Z corresponding to the X moments are zero. Both expressions are independent of j, giving the same fixed redundancy in the population moments and their adjusted counterparts.

Step 3. Obtain the minimum asymptotic covariance.

We compare the influence functions obtained from different combining matrices satisfying Assumption 6. Consider the centered vector

$$\varphi_{\text{eff}} = -\mathcal{I}_0^{-1} \left(\frac{\partial F_r}{\partial \theta'} \right)' \Sigma_r^{-1} \psi_r. \tag{56}$$

It has finite second moment by Assumption 9 and Step 1. Direct multiplication using Equation (54) gives

$$\mathrm{Var}(\varphi_{\text{eff}}) = \mathcal{I}_0^{-1}.$$

Using Equation (53), we express it in terms of the original adjusted moment vector stacked across strata as

$$\varphi_{\text{eff}} = -\mathcal{I}_0^{-1} G' \sum_{l=1}^{J} \sum_{j=1}^{J} \tilde{S}_l' [(\Sigma^*)^+]_{lj} \psi_j^*. \tag{57}$$

By Equations (28) and (30), the matrices A_j^* satisfy $\bar{A}^* G = \mathcal{I}_0$. They therefore satisfy Assumption 6, and Theorem 4 gives Equation (57) as their estimator's influence function.

For any combining matrices satisfying Assumption 6, Theorem 4 gives the influence function

$$\tilde{\varphi} = -(\bar{A}G)^{-1} \sum_j A_j \psi_j^*.$$

By Equation (55),

$$E[\psi_j^* \varphi_{\text{eff}}'] = -\tilde{S}_j G \mathcal{I}_0^{-1}.$$

Consequently,

$$E[\tilde{\varphi} \varphi_{\text{eff}}'] = (\bar{A}G)^{-1} \sum_j A_j \tilde{S}_j G \mathcal{I}_0^{-1} = \mathcal{I}_0^{-1}.$$

The difference $\tilde{\varphi} - \varphi_{\text{eff}}$ is therefore uncorrelated with φ_{eff}, and

$$\mathrm{Var}(\tilde{\varphi}) - \mathrm{Var}(\varphi_{\text{eff}}) = E[(\tilde{\varphi} - \varphi_{\text{eff}})(\tilde{\varphi} - \varphi_{\text{eff}})'] \succeq 0. \tag{58}$$

Thus $\mathcal{I}_0^{-1}$ is the minimum asymptotic covariance among the proposed IPW-GMM estimators satisfying Assumption 8, over combining matrices satisfying Assumption 6. Equations (53) and (54) prove Equation (28)–Equation (29). ■

Proof of optimal weighting and feasible attainment (Section 5.2).

The proof uses the optimal-GMM covariance comparison of Hansen (1982). Steps 1 and 2 establish that the proposed combining matrices satisfy Assumption 6 and minimize the asymptotic covariance. Step 3 obtains consistent estimates of these matrices. Steps 4 and 5 establish the limiting distribution and consistent covariance estimation for the feasible estimator.

Step 1. Verify nonsingularity of the weighted Jacobian

We first calculate the Jacobian and covariance under the proposed matrices. By Equa-

tion (30),

$$\begin{aligned}\bar{A}^* G &= \sum_j A_j^* \tilde{S}_j G \\ &= G' \left[\sum_{l,j} \tilde{S}_l' [(\Sigma^*)^+]_{lj} \tilde{S}_j \right] G = \mathcal{I}_0.\end{aligned}$$

Since Theorem 5 establishes $\mathcal{I}_0 \succ 0$, Assumption 6 holds for this weighting choice.

Then, the Moore–Penrose identity $(\Sigma^*)^+\Sigma^*(\Sigma^*)^+ = (\Sigma^*)^+$ gives

$$\begin{aligned}\Omega^*(A_1^*, \ldots, A_J^*) &= \sum_{j,l} A_j^* \Sigma_{jl}^* (A_l^*)' \\ &= G' \left[\sum_{a,b} \tilde{S}_a' \left[(\Sigma^*)^+ \Sigma^* (\Sigma^*)^+ \right]_{ab} \tilde{S}_b \right] G \\ &= \mathcal{I}_0.\end{aligned}$$

This proves both identities in Equation (31). Substitution in the sandwich from Theorem 4 yields

$$(\bar{A}^* G)^{-1} \Omega^*(A^*) [(\bar{A}^* G)^{-1}]' = \mathcal{I}_0^{-1}.$$

Step 2. Establish optimality of the population matrices

We compare the covariance of any estimator in the proposed class whose combining matrices satisfy Assumption 6 with $\mathcal{I}_0^{-1}$ by decomposing its influence function into the proposed optimal influence function and an uncorrelated remainder. For any collection $\{A_j\}_{j=1}^J$ satisfying Assumption 6, normalization gives

$$\sum_j \left\{ (\bar{A}G)^{-1} A_j - \mathcal{I}_0^{-1} A_j^* \right\} \tilde{S}_j G = 0.$$

Using symmetry of the generalized inverse and Equation (55),

$$\begin{aligned}\sum_l \Sigma_{jl}^* (\mathcal{I}_0^{-1} A_l^*)' &= \sum_{l,a} \Sigma_{jl}^* [(\Sigma^*)^+]_{la} \tilde{S}_a G \mathcal{I}_0^{-1} \\ &= \tilde{S}_j G \mathcal{I}_0^{-1}.\end{aligned}$$

Then it gives that the covariance between

$$\sum_j \left\{ (\bar{A}G)^{-1} A_j - \mathcal{I}_0^{-1} A_j^* \right\} \psi_j^* \qquad \text{and} \qquad \sum_l \mathcal{I}_0^{-1} A_l^* \psi_l^*$$

is zero. Expanding the variance therefore gives

$$\Gamma^*(A_1,\ldots,A_J) - \mathcal{I}_0^{-1} = \operatorname{Var}\left[\sum_j \left\{(\bar{A}G)^{-1}A_j - \mathcal{I}_0^{-1}A_j^*\right\}\psi_j^*\right] \succeq 0. \tag{59}$$

Thus the matrices A_j^* attain the smallest asymptotic covariance among all matrices satisfying Assumption 6. If a common nonsingular matrix C premultiplies each A_j, then

$$(C\bar{A}G)^{-1}CA_j = (\bar{A}G)^{-1}A_j,$$

and the sample equation is also premultiplied by C. Both the estimator and its influence function are unchanged.

Step 3. Obtain consistent estimates of the population weights

We estimate G and the adjusted covariance, using the fixed reduction from the proof of Theorem 5 to obtain a consistent generalized inverse. Instrument coverage implies that $\sum_j \tilde{S}_j'\tilde{S}_j$ is a nonsingular diagonal matrix: its instrument entries count the strata revealing each instrument and its X entries equal J. Define

$$\hat{G} = -\left(\sum_j \tilde{S}_j'\tilde{S}_j\right)^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\sum_j \hat{w}_{ij}\tilde{S}_j'\tilde{S}_jT_i\right\}. \tag{60}$$

The term $\tilde{S}_jT_i$ is evaluated when $g_{D,i}g_{Z,ij} = 1$, and terms with a zero observation indicator are set to zero. The true-weight law of large numbers and the estimated-minus-true-weight bound from the proof of Theorem 2 yield

$$\hat{G} \to_p -\left(\sum_j \tilde{S}_j'\tilde{S}_j\right)^{-1}\sum_j \tilde{S}_j'\tilde{S}_jE[T] = G.$$

The variance-estimation conditions in Section 5.1 give

$$\hat{M}_0 \to_p M_0, \qquad \hat{M}_D \to_p M_D, \qquad \widehat{\tilde{S}_jM_{Z,j}\tilde{S}_j'} \to_p \tilde{S}_jM_{Z,j}\tilde{S}_j'.$$

Since J is fixed, Equation (33) gives joint convergence of all covariance blocks. Let R_0 and L_0 be the fixed deletion and reconstruction matrices from the proof of Theorem 5, and let $\hat{\Sigma}_r$ be the estimated covariance of the retained equations, formed by applying R_0 to Equation (33). Then

$$\hat{\Sigma}_r \to_p \Sigma_r \succ 0.$$

Reconstruct the unreduced covariance estimate as $\hat{\Sigma}^* = L_0\hat{\Sigma}_rL_0'$ so that it satisfies the

fixed known identities. For the repeated X equations, these identities also hold directly when the components of the fitted e_Z residuals corresponding to the X moments are set to zero. With probability approaching one, the retained covariance is nonsingular, and Equation (53) implies

$$(\hat{\Sigma}^*)^+ = (L_0^+)'\hat{\Sigma}_r^{-1}L_0^+ \to_p (L_0^+)'\Sigma_r^{-1}L_0^+ = (\Sigma^*)^+.$$

Consequently,

$$\hat{A}_j^* \to_p A_j^*, \qquad \hat{\mathcal{I}}_0 \to_p \mathcal{I}_0. \tag{61}$$

Computation with the retained equations gives the same normalized influence function and hence the same first-order result.

Step 4. Obtain the expansion of the feasible estimator

We show that estimating the combining matrices leaves the first-order moment expansion unchanged. Assumption 8(v) states, for every stratum,

$$\frac{1}{n}\sum_i \hat{w}_{ij}\tilde{S}_j\tau_i(\theta_0) = \frac{1}{n}\sum_i \psi_{ij}^* + o_p(n^{-1/2}) = O_p(n^{-1/2}).$$

Therefore,

$$\begin{aligned}\sqrt{n}\left\|\sum_j(\hat{A}_j^* - A_j^*)\frac{1}{n}\sum_i \hat{w}_{ij}\tilde{S}_j\tau_i(\theta_0)\right\| \\ \leqslant \sum_j \|\hat{A}_j^* - A_j^*\|\left\|\frac{1}{\sqrt{n}}\sum_i \hat{w}_{ij}\tilde{S}_j\tau_i(\theta_0)\right\| = o_p(1).\end{aligned}$$

The final sample matrix converges to

$$\frac{1}{n}\sum_{i,j}\hat{A}_j^*\hat{w}_{ij}\tilde{S}_jT_i \to_p -\sum_j A_j^*\tilde{S}_jG = -\mathcal{I}_0.$$

Applying the linear error identity from Lemma 3 now gives

$$\sqrt{n}(\hat{\theta}^* - \theta_0) = -\mathcal{I}_0^{-1}\frac{1}{\sqrt{n}}\sum_i\sum_j A_j^*\psi_{ij}^* + o_p(1).$$

Step 1 supplies the term's covariance, and the multivariate central limit theorem proves Equation (32). By Equation (30), this estimator's influence function equals Equation (57).

Step 5. Establish consistency of the covariance estimator

Finally, Equation (61) and continuity of inversion at $\mathcal{I}_0 \succ 0$ give

$$\hat{\mathcal{I}}_0^{-1} \to_p \mathcal{I}_0^{-1}.$$

The final sample Jacobian converges to $\mathcal{I}_0$, and the fitted combined covariance converges to $\mathcal{I}_0$. Hence the conventional sandwich has the same limit. Dividing either estimate by n gives the first-order covariance estimator for $\hat{\theta}^*$. The final sandwich consistency argument follows the standard GMM plug-in approach; see Newey and McFadden (1994). ■

B More Results

B.1 More on Monte Carlo Results

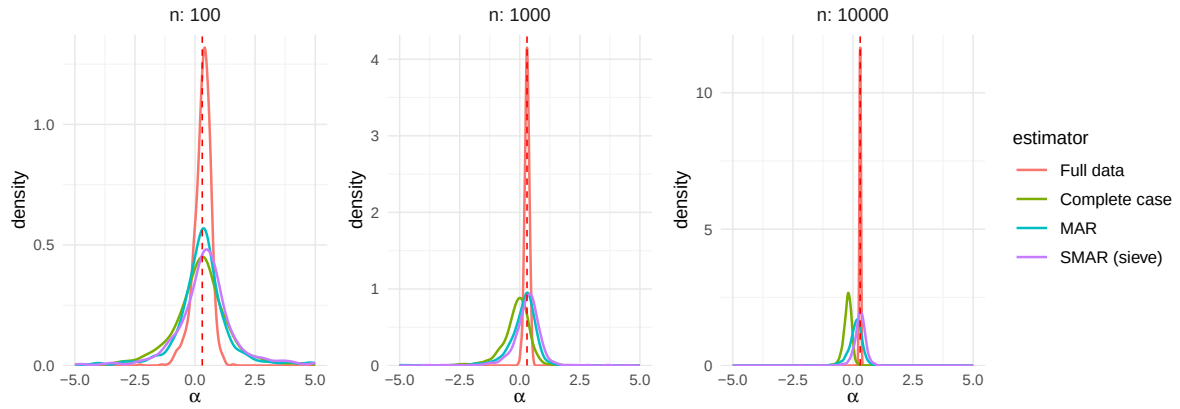


Figure 1: Sampling distributions of $\hat{\alpha}$ for full-data 2SLS, complete-case 2SLS, all-pattern MAR-IPW, and the proposed SMAR-IPW estimator with a sieve first step, by sample size. The vertical line marks $\alpha_0 = 0.3$. The densities are computed using all finite draws; the horizontal display window is $[-5, 5]$.

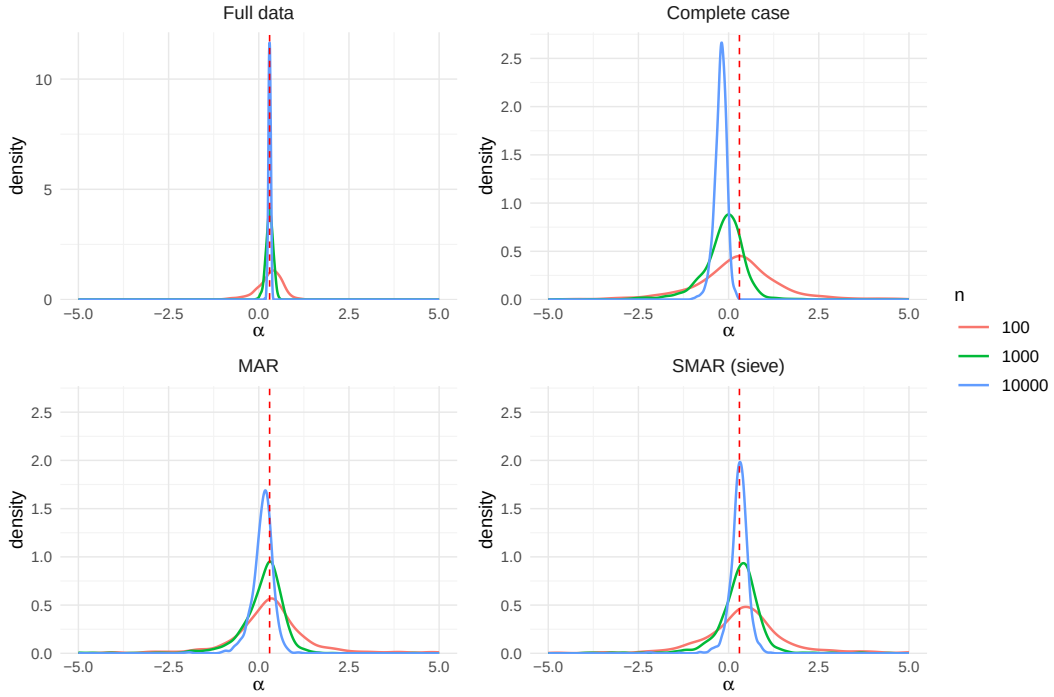


Figure 2: Sampling distributions of full-data 2SLS, complete-case 2SLS, all-pattern MAR-IPW, and the proposed SMAR-IPW estimator with a sieve first step across sample sizes. The vertical line marks $\alpha_0 = 0.3$. The densities are computed using all finite draws; the horizontal display window is $[-5, 5]$.

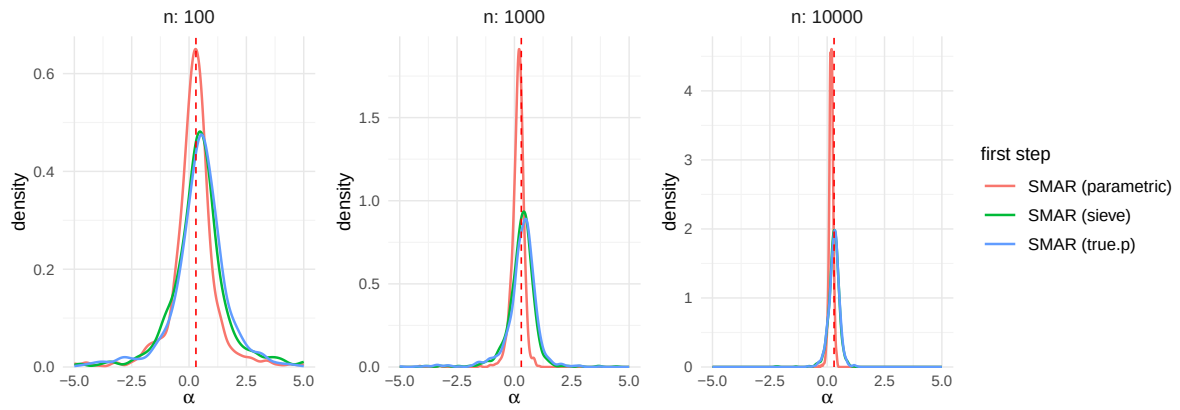


Figure 3: First-step comparison for the SMAR estimator under the nonlinear design: a misspecified linear-index bounded-probit first step, a flexible spline-sieve first step, and the true observation probabilities. The vertical line marks $\alpha_0 = 0.3$. The densities are computed using all finite draws; the horizontal display window is $[-5, 5]$.

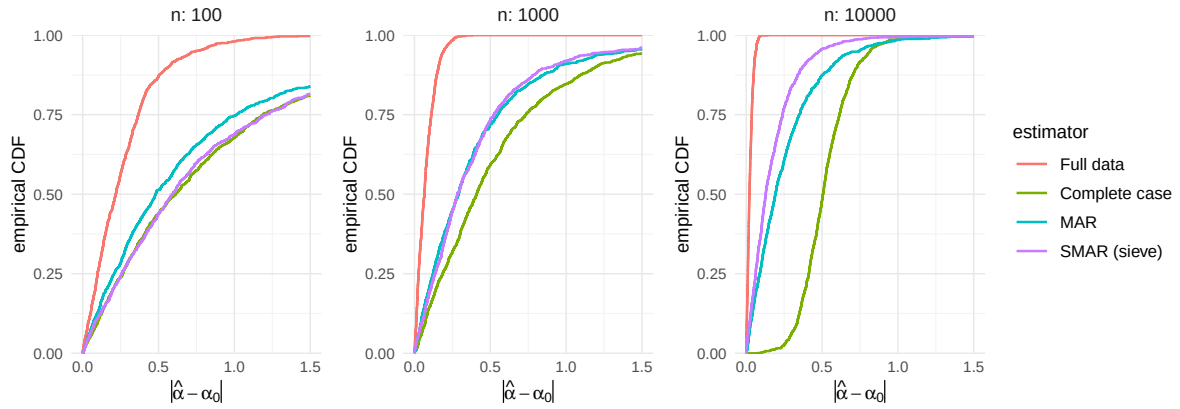


Figure 4: Empirical cumulative distribution functions of the absolute estimation error $|\hat{\alpha} - \alpha_0|$ for full-data 2SLS, complete-case 2SLS, all-pattern MAR-IPW, and the proposed SMAR-IPW estimator with a sieve first step, by sample size. At a given error threshold, a higher curve indicates a larger fraction of estimates lying within that threshold.

Figures 1 and 2 make this distinction between dispersion and accuracy especially clear. Figure 1 shows that the four densities overlap substantially at $n = 100$, but separate as the sample size increases. By $n = 10{,}000$, the full-data and sieve-SMAR distributions are centered near the vertical line at 0.3, whereas the complete-case and MAR distributions are centered well to its left. Figure 2 follows each estimator across sample sizes. The complete-case distribution becomes much narrower as n increases, but it becomes concentrated around approximately -0.21 rather than around 0.3. Its MAD at $n = 10{,}000$ is 0.1498, which is smaller than the sieve estimator's MAD of 0.1887, but its median absolute error is 0.5063, compared with 0.1292 for the sieve estimator. The lower complete-case dispersion is therefore not an efficiency gain for estimating α_0; it reflects increasingly

precise estimation of the wrong probability limit. The MAR estimator shows the same issue in a less severe form, with median bias -0.1599 and median absolute error 0.1918 at $n = 10{,}000$.

Figure 3 isolates the role of the first-step specification within the proposed SMAR procedure. The parametric estimator is comparatively concentrated, but its distribution moves toward a value below the truth as the sample size increases. At $n = 10{,}000$, its median estimate is 0.1779, its median bias is -0.1221, and its central 80 percent range is only 0.2153. The narrow distribution is again not evidence of greater efficiency, since the linear-index first step omits the quadratic terms in the true missingness probabilities and the estimator becomes concentrated around a pseudo-true value. The parametric restriction gives it a favorable median absolute error at each of the three reported sample sizes, including 0.1225 compared with 0.1292 for the sieve estimator at $n = 10{,}000$. This is a finite-sample tradeoff rather than evidence of consistency: the parametric estimator retains a nonvanishing misspecification error, whereas the sieve error continues to decline as the first step becomes more flexible and the sample size grows.

In contrast, the sieve and true-probability SMAR estimators become centered near α_0. Their median estimates at $n = 10{,}000$ are 0.3058 and 0.3139, corresponding to median biases of 0.0058 and 0.0139. Their median absolute errors are also similar, at 0.1292 and 0.1327. The central dispersion of the sieve estimator is slightly smaller than that of the true-probability estimator: its MAD is 0.1887 rather than 0.1933, and its central 80 percent range is 0.5379 rather than 0.5679. This pattern is consistent with the first-step adjustment described in Proposition 1, under which estimating the observation probabilities can remove part of the variation generated by selection. The numerical difference is modest, and the pilot cross-stratum weightings used here are not fully optimal, so this comparison should be read as supporting evidence rather than as a direct efficiency-bound calculation.

Figure 4 summarizes the economically relevant comparison by plotting the empirical distribution of $|\hat{\alpha} - \alpha_0|$. At $n = 100$, the flexible first step is costly relative to the amount of information in the sample: among the incomplete-data estimators shown in the figure, MAR has the smallest median absolute error, 0.4803, compared with 0.5922 for sieve SMAR and 0.6133 for complete case. At $n = 1{,}000$, the sieve estimator begins to improve on both alternatives, with median absolute error 0.2820, compared with 0.2921 for MAR and 0.4098 for complete case. At $n = 10{,}000$, the ordering becomes clear. The sieve-SMAR curve lies above the MAR and complete-case curves over most of the relevant error range, and its median absolute error falls to 0.1292, compared with 0.1918 for MAR and 0.5063 for complete case. The full-data estimator remains substantially more precise, with median absolute error 0.0213, as expected for a benchmark that observes

every moment without inverse-propensity weighting.

B.2 More on Empirical Results

I use the replication extract provided by CG to benchmark the empirical implementation. The sibling count is observed throughout this extract, so the first missingness layer is absent. This setting is therefore a special case of the proposed sequential framework in which only the instruments are incompletely observed and the first observation probability equals one. Table 6 compares sample characteristics and standard IV estimates computed using their prepared data with their published benchmarks. The estimates are similar in sign and magnitude. This exercise provides a reference point for the main application of the full sequential procedure to the extended sample, where both the sibling count and the instruments can be missing.

Table 6: Benchmark estimates using the Chaudhuri and Guilkey replication extract

	This paper	CG (2016)
Mother and child pairs in the $g_D = 1$ extract	7,391	7,370
Mean years of schooling	7.73	7.74
Mean number of siblings, censored at 4	2.32	2.34
Share with both sometimes-missing IVs observed	0.333	0.333
GMM with 4 always observed IVs	−0.795 (0.318)	−0.73 (0.325)
Complete case GMM with 6 IVs	−0.947 (0.438)	−1.01 (0.515)

Notes: The first four rows compare sample size and descriptive moments. The last two rows compare point estimates, with standard errors in parentheses. The validation specification includes age indicators, sex, birth order, parental education categories, literacy, Muslim, urban status, and the three community school counts.

Table 7: Years of schooling on the number of siblings: alternative specifications

Estimator	Estimate	Standard error
Specification 2: Categorical sibling count in the instrument observation model		
OLS	−0.076	0.016
GMM with observed D	−1.113	0.297
Complete-case GMM with observed D and Z	−0.324	0.634
Complete-case IPW-GMM (propensity excludes D)	−0.469	0.641
SMAR IPW-GMM	−0.940	0.377
Specification 3: Always-observed controls only		
OLS	−0.183	0.017
GMM with observed D	−1.594	0.274
Complete-case GMM with observed D and Z	0.024	0.766
Complete-case IPW-GMM (propensity excludes D)	−1.217	0.746
SMAR IPW-GMM	−1.399	0.334

Notes: The outcome is completed years of schooling. The number of siblings is censored above at four, and the probability trimming threshold is 0.02. Both specifications use the same baseline analysis sample of 16,238 children. The SMAR rows report the feasible optimal-weighting estimates; their standard errors include the first-step adjustment.

Specification 2 retains the main specification's schooling controls and joint first-layer availability requirement for the sibling count, maternal education, and maternal literacy. Relative to the main specification, it uses indicators for one, two, three, and four-or-more siblings, with zero siblings as the reference category, in the conditioning set for the second-layer instrument observation model. The remaining probability-model specification is unchanged.

Specification 3 also uses categorical sibling-count terms in the second-layer instrument observation model, but omits maternal education and literacy from the schooling equation and that observation model. Its first-layer availability indicator therefore requires only an observed sibling count. Age indicators, sex, Muslim status, urban status, and the three community school counts remain included. Removing the maternal availability requirement expands the observations eligible to contribute structural moments relative to the main specification. The complete-case IPW estimator excludes the sibling count from its propensity model in both panels.