GMM with Multiple Missing Variables*

Saraswata Chaudhuri† and David K. Guilkey‡

January 17, 2015

Summary

We consider efficient estimation in moment conditions models with non-monotonically missing-at-random (MAR) variables. A version of MAR point-identifies the parameters of interest and gives a closed-form efficient influence function that can be used directly to obtain efficient semi-parametric GMM estimators under standard regularity conditions. A small-scale Monte-Carlo experiment with MAR instrumental variables demonstrates that the asymptotic superiority of these estimators over the standard methods carries over to finite samples. An illustrative empirical study of the relationship between a child's years of schooling and number of siblings indicates that these GMM estimators can generate results with substantive differences from standard methods.

JEL Classification: C13; C14; C31; C36

Keywords: Non-monotonically missing data; Efficiency bounds; Inverse probability weighting; Double robustness; Generalized method of moments; Missing instruments.

*We thank the co-editor Edward Vytlacil and four anonymous referees for very useful comments that improved the content and the presentation of this paper substantially. We also thank D. Frazier, SJ. Lee, B. McManus, A. Prokhorov, B. Tsang and the participants at the NASM (2013), the Triangle Econometrics Conference (2012) and the Asian meeting of the Econometric Society (2012), and seminar participants at Concordia, U. Canterbury, McGill, U. New South Wales, U. Sydney (Business Analytics & Econ), U. West Virginia, U. Washington for helpful discussions.

†Corresponding author. Department of Economics, McGill University, 855 Sherbrooke Street West Montreal, Quebec H3A 2T7. Email: saraswata.chaudhuri@mcgill.ca. Tel: 514-398-4400, Extn: 09169. Fax: 514-398-4938.

‡Department of Economics, University of North Carolina, Chapel Hill.

1 Introduction

We study efficient estimation of parameters using a sample from which variables are missing jointly or individually for various units. Feasible efficient estimation is well studied when such missingness is monotone [see Tsiatis (2006)]. Our focus is on the more general non-monotone pattern of missingness. This arises in: (i) program evaluations where all but one counterfactual are, by definition, unobserved for any unit; (ii) panel studies with attrition where individuals may return to/enter the panel in a future period, (iii) general non-response in surveys, (iv) data combinations, etc.

The parameter value of interest, β^0, is defined by a set of unconditional moment restrictions:

$$E[g(Z,W;\beta)] = 0 \text{ for } \beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta} \text{ if and only if } \beta = \beta^0, \tag{1}$$

where $g(Z,W;\beta) : \text{support}(Z,W) \times \mathcal{B} \mapsto \mathbb{R}^{d_g}$ is a known function and $d_g \geq d_\beta$.

Let $Z = (Z_1, Z_2)$. We consider the observed sample to be $\{D_{1i}, D_{2i}, D_{1i}Z_{1i}, D_{2i}Z_{2i}, W_i\}_{i=1}^N$ where, for $j = 1, 2$: $D_{ji} = 1$ if Z_{ji} is observed and 0 otherwise. Non-monotonicity arises because we allow for both sub-samples whose units are characterized by $(D_1 = 1, D_2 = 0)$ and $(D_1 = 0, D_2 = 1)$.

A standard representation of this missing data pattern considers a missingness/coarsening indicator $C = 0, 1, 2, \infty$ and a transformation $G_C(Z,W)$ of Z such that $G_0(Z,W) = W, G_1(Z,W) = (Z_1, W), G_2(Z,W) = (Z_2, W)$ and $G_\infty(Z,W) = (Z,W)$. Point-identification of β^0 defined in (1) then follows from: (i) an overlap condition (to be duly defined in M(2)) on the complete sub-sample $(C = \infty)$[1], and (ii) the missing at random (MAR) assumption [see Heitjan and Rubin (1991)] that

$$P(C = r|Z) = P(C = r|G_r(Z)) \text{ almost surely in } G_r(Z) \text{ for all } r = 0, 1, 2, \infty. \tag{2}$$

Efficient estimation under similar setups was pioneered by Robins and his co-authors. See, among others, Robins et al. (1994), Robins et al. (1995), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995). In principle, the idea put forward by these papers also gives the semiparametric efficiency bounds for estimation of parameters when multiple variables are missing non-monotonically from the sample. However, under non-monotone missingness such bounds and the associated efficient influence functions generally do not have closed form expressions. Hence it is difficult to use these results and develop feasible but efficient estimators [see, for example, chapter 10 of Tsiatis (2006)].

The first theoretical result of this paper is to show that a different MAR assumption avoids the

[1] Existence of the complete sub-sample is in contrast to the literature on data combination that usually focuses on specifying minimal assumptions to ensure point-identification or non-trivial partial-identification of the parameters of interest. See Ichimura and Martinez-Sanchis (2005), Ridder and Moffitt (2007), etc.

above problem. In particular, we show that replacing (2) by

$$P(C = r|Z) = P(C = r|G_0(Z)) \text{ almost surely in } G_0(Z) \text{ for all } r = 0, 1, 2, \infty, \tag{3}$$

produces closed form expressions for the efficiency bound and the efficient influence function.[2] The latter has the augmented inverse probability weighted (AIPW) form [see Robins et al. (1994)]. We consider semiparametric GMM estimation of β^0 based on the efficient influence function after plugging in nonparametric series estimators for the associated nuisance parameters. We refer to this estimator of β^0 as the AIPW-GMM estimator. We show that under standard regularity conditions, similar to those in Cattaneo (2010), the AIPW-GMM estimator is semiparametrically efficient.

Since the original representation of the observed data as $\{D_1, D_2, D_1Z_1, D_2Z_2, W\}$ is more common in economics, we revert to it in the rest of the paper. For this purpose, it is useful to note the equivalent representations of missingness: $\mathbf{1}(C = 0) \equiv (1 - D_1)(1 - D_2)$, $\mathbf{1}(C = 1) \equiv D_1(1 - D_2)$, $\mathbf{1}(C = 2) \equiv (1 - D_1)D_2$, and $\mathbf{1}(C = \infty) \equiv D_1D_2$. Accordingly, we restate the MAR assumption in (3) as follows: There exists a set of observed random variables $\bar{W}$ such that

$$\text{MAR: } (D_1, D_2) \text{ independent of } (Z_1, Z_2) \text{ conditional on } \bar{W} \text{ almost surely in } \bar{W}. \tag{4}$$

While the term "exogenous" is confusing without a clear context/model or at least conditional moment restrictions, it is important to note that selection based on $\bar{W}$ need not be exogenous in the sense of its popular use in empirical research. See Moffitt et al. (1998). We abstract from the crucial empirical issue of determining the right $\bar{W}$, and take $\bar{W} = W$ in the rest of the paper unless otherwise specified. This is a difficult issue to resolve using statistical theory; we are unaware of any consistent (in all directions) tests for MAR assumptions without additional strong restrictions.[3]

The second set of results in our paper are related to an application of MAR to a specific scenario: multiple missing instrumental variables (IVs) in classical linear IV regressions.[4] We present

[2](3) is statistically neither stronger nor weaker than (2). These are just different assumptions, and depending on the application one may be preferred to the other. However, (2) is better suited for the most common application of MAR under more than two-level missingness: panel data with attrition in at least two periods that leads to monotone missingness. On the other hand, non-monotone missingness arises in this context only if individuals leaving the panel can subsequently return. Vansteelandt et al. (2007) argue that any form of MAR, i.e., selection on observables, is then unrealistic since the choice to return would depend on unobservables, i.e., on what happened when the individual was out of the panel. Also see Gill and Robins (1997), Gill et al. (1997) and Robins and Gill (1997). Therefore, given our focus on non-monotone missingness, we do not lose much *additional* applicability when we assume (3) instead of (2).

[3]However, the MAR assumption has been widely used in the recent econometrics literature. See, among many others, Moffitt et al. (1998) in the context of attrition, Imbens (2004) and Heckman and Vytlacil (2007) and the references therein in the context of treatment effect, Chen et al. (2008), Wooldridge (2002, 2007), Cattaneo (2010), Graham (2011), Graham et al. (2012), etc. in the context of general Z/M/GMM estimation with missing data.

[4]Consideration of missing outcomes or regressors has received a lot of attention. See, for example, the aforementioned papers by Robins and co-authors, Hahn (1998), Hirano et al. (2003), Graham et al. (2012), etc. However, the same is not true for missing IVs. Exceptions are Abrevaya and Donald (2011), Mogstad and Wiswall (2012), Chaudhuri and

a thorough treatment of MAR IV and thereby lay the groundwork for the empirical application in this paper where we need to use IV estimation with multiple missing IVs. We point out for the practitioner the sources of bias in estimation, and recommend a possible solution.

The case of multiple missing IVs falls under a general framework (see Section 3 for examples) that amounts to partitioning the rows of $g(Z, W; \beta)$ in (1) based on the variables that are missing from some sample units. We represent it as follows. Letting $\beta = (\beta_0', \beta_1', \beta_2')'$, consider the partition:

$$g(Z, W; \beta) \quad := \quad \left[g_1'(Z_1, W_1; \beta_0, \beta_1), g_2'(Z_2, W_2; \beta_0, \beta_2)\right]', \tag{5}$$

where $g_j(.) : \text{support}(Z_j, W_j) \times (\mathcal{B}_0 \times \mathcal{B}_j) \mapsto \mathbb{R}^{d_{g_j}}$ is a known function for $j = 1, 2$. $\mathcal{B} = \mathcal{B}_0 \times \mathcal{B}_1 \times \mathcal{B}_2$ is the parameter space of $\beta = (\beta_0', \beta_1', \beta_2')'$ and $\mathcal{B}_j \subseteq \mathbb{R}^{d_{\beta_j}}$ that of β_j for $j = 0, 1, 2$. $d_g := d_{g_1} + d_{g_2} \geq d_\beta := d_{\beta_0} + d_{\beta_1} + d_{\beta_2}$ and $d_{\beta_0} + d_{\beta_j} > 0$ for $j = 1, 2$. Let W denote the distinct collection of all elements of W_1 and W_2 allowing for the possibility that $W = W_1$ and/or $W = W_2$.

Cattaneo (2010)'s setup in the context of multi-valued treatments is closely related to it and can be mimicked, for example, by taking $\beta = (\beta_1', \beta_2')$, $W_1 = W_2 = W$ and ruling out the sample units with $D_1 = D_2$. The common complete sub-sample ($D_1 = D_2 = 1$) is no longer required for point-identification of β^0 under the partition in (5). We show that the formulation of the efficiency bound and the efficient influence function in Cattaneo (2010) remains valid under our extended setup.

As with the general case in (1), we also consider an AIPW-GMM estimator for β^0 under the partition in (5). Conditions given in Cattaneo (2010) prove sufficient for this estimator to be semi-parametrically efficient. The same conditions ensure efficiency of the IPW-GMM estimator discussed in Cattaneo (2010). However, we also show that the AIPW-GMM estimator is semiparametrically efficient under markedly weaker conditions than that required for the IPW-GMM estimator. This result is known in the context of parametric IPW and AIPW estimators. Our results are, however, for semiparametric estimators where the preliminary infinite dimensional nuisance parameters are estimated using series methods. Recently and prior to us, Rothe and Firpo (2012) provided similar results for cases where the nuisance parameters are estimated by local polynomial methods.

We show that *no such efficiency result* holds for similar IPW-GMM estimators in the general case in (1) (i.e., without the partition in (5)). Among others, Hirano et al. (2003), Cattaneo (2010), Chen et al. (2008) and Chen et al. (2012) considered various scenarios under which carefully chosen

Min (2012), Wang (2013), Muris (2014). Due to the recent trend that exploits the Mendelian randomization by using genetic markers as IVs, missing IVs is likely to become more common in empirical work. See, for example, Lawlor et al. (2008), Ding et al. (2009), Scholder et al. (2010), Scholder et al. (2011), Burgess et al. (2011), Palmer et al. (2012), Berry et al. (2012), etc. and the references therein. In most cases, genetic markers are collected in a two-phase sampling only for a sub-sample — for example, for worse outcome (dependent variable) or high exposure (explanatory variable) patients in medical studies, — perhaps due to the cost of data collection. This leads to MAR IVs.

regularity conditions ensure the convergence of the estimated nuisance parameters at a desired rate that gives efficiency of the IPW estimators. However, the general case in (1) is not such a scenario.

The rest of the paper is organized as follows. Section 2 contains all the theoretical results. The proofs of these results are collected in a Supplemental Appendix. Section 3 lists some examples where the partition in (5) is applicable. (Examples for the general case are standard.) Section 4 studies one such example, the case of missing IVs, in detail. It demonstrates the problems due to the common empirical practice of ignoring the sample units for which IVs are missing. It contrasts our proposed solution with that of Mogstad and Wiswall (2012). A small-scale Monte-Carlo experiment is provided to illustrate the adverse consequences of ignoring the missing IVs in a linear IV regression, and a possible solution under MAR through the use of the AIPW-GMM estimator discussed in Section 2.

Section 5 illustrates the estimators in an empirical study of the relationship between a child's education and number of siblings under a classical IV setup based on data from Indonesia. Access to community level health care and child-birth facilities are used as IVs. They could potentially affect the number of children of a mother but should not directly affect the child's education after controlling for relevant variables. Two of these IVs are missing jointly or individually for many sample units. The recommended methods show strong evidence of a significant negative relationship between a child's average educational attainment and the number of siblings. Section 6 concludes.

2 Efficiency bounds and efficient estimators

Notation used: Conditional probabilities are denoted by $p_{jk}(W) := P(D_1 = j, D_2 = k|W)$ for $j, k = 0, 1$. $p_1(W) := p_{11}(W)+p_{10}(W)$, $p_2(W) := p_{11}(W)+p_{01}(W)$. Unconditional probabilities are denoted accordingly by dropping W. Conditional expectations are denoted by $q(W;\beta) := E[g(Z, W;\beta)|W]$, $q(Z_j, W;\beta) := E[g(Z, W;\beta)|Z_j, W]$ for $j = 1, 2$. Under (5), $q(W;\beta) := [q_1(W;\beta_0, \beta_1)', q_2(W;\beta_0, \beta_2)']'$ where $q_j(W;\beta_0, \beta_j) := E[g_j(Z_j, W_j;\beta_0, \beta_j)|W]$ for $j = 1, 2$. Define the following quantities to be used to express the efficient influence functions under the general case (1) and under the partition (5):

$$\begin{aligned}\varphi(D_1, D_2, Z_1, Z_2, W;\beta) &:= \frac{D_1D_2}{p_{11}(W)}\left[g(Z, W;\beta) - q(W;\beta)\right] + q(W;\beta) \\ &\quad + \frac{p_{10}(W)}{p_1(W)}\left[\frac{D_1(1-D_2)}{p_{10}(W)} - \frac{D_1D_2}{p_{11}(W)}\right]\left[q(Z_1, W;\beta) - q(W;\beta)\right] \\ &\quad + \frac{p_{01}(W)}{p_2(W)}\left[\frac{(1-D_1)D_2}{p_{01}(W)} - \frac{D_1D_2}{p_{11}(W)}\right]\left[q(Z_2, W;\beta) - q(W;\beta)\right] \quad (6) \\ \text{and } \varphi_P(D_1, D_2, Z_1, Z_2, W;\beta) &:= \left[\varphi_{P1}(D_1, Z_1, W;\beta)', \varphi_{P2}(D_2, Z_2, W;\beta)'\right]' \quad (7)\end{aligned}$$

where $\varphi_{Pj}(D_j, Z_j, W_j;\beta) := \frac{D_j}{p_j(W)}\left[g_j(Z_j, W_j;\beta_0, \beta_j) - q_j(W;\beta_0, \beta_j)\right] + q_j(W;\beta_0, \beta_j)$ for $j = 1, 2$. The

subscript "P" stands for partition. Hereafter we use the term P-case to refer to this case. Unless confusing, we will suppress the dependence on β in all quantities evaluated at β^0.

Assumption M:

(1) $\{D_{1i}, D_{2i}, D_{1i}Z_{1i}, D_{2i}Z_{2i}, W_i\}_{i=1}^N$ are i.i.d. copies of $\{D_1, D_2, D_1Z_1, D_2Z_2, W\}$.

(2) $p_{11}(W), p_{10}(W), p_{01}(W) \in [\kappa, 1)$ almost surely in W where $\kappa > 0$.

(3) $E[g(Z, W; \beta)]$ is differentiable with respect to β in an open neighborhood $\mathcal{N} \subset \mathcal{B}$ of β^0, and $G(\beta) := \frac{\partial}{\partial \beta'} E[g(Z, W; \beta)]$ has full column rank d_β at $\beta = \beta^0$. $G(\beta)$ is $G_P(\beta)$ in the P-case.

(4) $V(\beta) := Var(\varphi(D_1, D_2, Z_1, Z_2, W; \beta))$ and $V_P(\beta) := Var(\varphi_P(D_1, D_2, Z_1, Z_2, W; \beta))$ are bounded and positive semidefinite for $\beta \in \mathcal{B}$. $V(\beta^0)$ and $V_P(\beta^0)$ are positive definite.

Remarks: (i) $\{D_{1i}Z_{1i}, D_{2i}Z_{2i}\}_{i=1}^N$ in M(1) are not i.i.d. copies of $\{Z_1, Z_2\}$ but of $\{D_1Z_1, D_2Z_2\}$. Selection bias is possible since the observed sample need not represent the target population. (ii) M(2) concerning the lower bound of $p_{11}(W)$ away from 0 is known as the strict overlap condition and suffices for a finite efficiency bound. See Khan and Tamer (2010) and Chaudhuri and Hill (2013) for more on this technical condition. M(2) concerning $p_{10}(W)$ and $p_{01}(W)$ is for convenience. (iii) The partition in the rows of the moment function in the P-case restricts the concerned elements of $G_P(\beta)$ to be zeros. (iv) Under (1), $V(\beta)$ defined in M(4) and evaluated at $\beta = \beta^0$ is:

$$V := V(\beta^0) = E\left[\frac{1}{p_{11}(W)} Var(g(Z, W)|W) + q(W)q(W)'\right] - \Delta \tag{8}$$

where the first term is the familiar Ω_β^2 in Theorem 1 of Chen et al. (2008) while the second term is

$$\begin{aligned}\Delta \;:=\; & Var\left(\frac{p_{10}(W)}{p_1(W)}\left[\frac{D_1(1-D_2)}{p_{10}(W)} - \frac{D_1D_2}{p_{11}(W)}\right]\left[q(Z_1, W; \beta^0) - q(W; \beta^0)\right]\right. \\ & \left. - \frac{p_{01}(W)}{p_2(W)}\left[\frac{(1-D_1)D_2}{p_{01}(W)} - \frac{D_1D_2}{p_{11}(W)}\right]\left[q(Z_2, W; \beta^0) - q(W; \beta^0)\right]\right).\end{aligned}$$

The benchmarks for estimation of β^0 under the general case (i.e., under (1)) and under the P-case (i.e., under (1) and (5)) are given below in Propositions 2.1 and 2.2 respectively.

Proposition 2.1 *Let* (1), (4) *and assumption M hold. Then for* β^0 *defined by* (1) *the asymptotic variance lower bound for* $\sqrt{N}(\widehat{\beta} - \beta^0)$ *of any regular estimator* $\widehat{\beta}$ *is given by* $\Omega := (G'V^{-1}G)^{-1}$. *An estimator whose asymptotic variance equals* Ω *has the asymptotically linear representation*

$$\begin{aligned}\sqrt{N}(\widehat{\beta} - \beta^0) &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(D_{1i}, D_{2i}, Z_{1i}, Z_{2i}, W_i) + o_p(1), \;\; \textit{where} \\ \psi(D_1, D_2, Z_1, Z_2, W) &:= -\Omega^{-1}G'V^{-1}\varphi(D_{1i}, D_{2i}, Z_{1i}, Z_{2i}, W_i; \beta^0).\end{aligned}$$

Proofs of all the stated results are collected in a Supplemental Appendix (Appendix C).

Remarks:

1. The key condition that produced closed forms for the efficiency bound and the efficient influence function is the version of MAR in (4) (or equivalently (3)). The technical role played by (4) is roughly as follows. Consider any row of the moment vector involving Z_1, Z_2 and W. (4) enables independence of D_1, D_2 (and hence D_1D_2, $D_1(1-D_2)$, $(1-D_1)D_2$) with Z_1, Z_2 by conditioning on the same set of variables W. This produces a closed form for the projection of an influence function to the tangent set of the model and thereby produces the closed form of the efficient influence function. Under the other version of MAR in (2) we need different conditioning sets to obtain each independence. Since those conditioning sets are not even monotone [see Chaudhuri (2014)], a closed form for the efficient influence function is generally ruled out unless additional restrictions are imposed.

2. While we report the result for two sets of missing variables Z_1 and Z_2, Proposition 2.1 can in principle be extended to, say, some general $R(> 2)$ sets similar to Chaudhuri (2014). Under non-monotone missingness such general exposition allowing for the 2^R possible types of sub-samples (we have 2^2 here) may not be practically relevant.

3. However, importantly, following the same method as in Chaudhuri (2014), it is also possible to extend the result to cases where β^0 is defined only in terms of sub-populations, for example, where $D_1 = 0$. The efficiency bound would then depend on if $\{p_{d_1d_2}(W) : d_1, d_2 = 0, 1\}$ is unknown, completely known or known up to finite dimensional parameters as are the cases in Propositions 1, 2 and 3 in Chaudhuri (2014). See Hahn (1998) and Chen et al. (2008) for the original work.

Proposition 2.2 *Let* (1), (4), (5) *and assumption M hold. Then for* β^0 *defined by* (1) *the asymptotic variance lower bound for* $\sqrt{N}(\widehat{\beta} - \beta^0)$ *of any regular estimator* $\widehat{\beta}$ *is given by* $\Omega_P := (G_P' V_P^{-1} G)_P^{-1}$. *An estimator whose asymptotic variance equals* Ω_P *has the asymptotically linear representation*

$$\begin{aligned} \sqrt{N}(\widehat{\beta} - \beta^0) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_P(D_{1i}, D_{2i}, Z_{1i}, Z_{2i}, W_i) + o_p(1), \text{ where} \\ \psi_P(D_1, D_2, Z_1, Z_2, W) &:= -\Omega_P^{-1} G_P' V_P^{-1} \varphi_P(D_{1i}, D_{2i}, Z_{1i}, Z_{2i}, W_i; \beta^0). \end{aligned}$$

Remarks:

1. Despite the additional structure in this extended setup, the efficient influence function and the efficiency bound have the same form as in Cattaneo (2010). Results for the sub-vectors β_0, β_1 and β_2 or their smooth functions (example 2 in Section 3) are straightforward to obtain from here.

2. As before, following the same method as in Chaudhuri (2014), it is also possible to extend the result to cases where β^0 is defined only in terms of sub-populations, for example, when $D_1 = 0$.

3. This proposition establishes the benchmark for the examples discussed in Sections 3 and 4 and the empirical application in Section 5.

Propositions 2.1 and 2.2 suggest the modified moment restrictions for the general case and the P-case:

$$E[\varphi(D_1, D_2, Z_1, Z_2, W; \beta)] = 0 \text{ for } \beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta} \text{ if and only if } \beta = \beta^0, \tag{9}$$

$$E[\varphi_P(D_1, D_2, Z_1, Z_2, W; \beta)] = 0 \text{ for } \beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta} \text{ if and only if } \beta = \beta^0, \tag{10}$$

to be used for efficient estimation of β^0. $\varphi(D_1, D_2, Z_1, Z_2, W; \beta)$ and $\varphi_P(D_1, D_2, Z_1, Z_2, W; \beta)$ involve elements of the infinite dimensional nuisance parameters $p(W) := (p_{d_1 d_2}(W))_{d_1, d_2 = 0,1}$ and $q(\beta) := (q(W; \beta)', q(Z_1, W; \beta)', q(Z_2, W; \beta)')'$. $q(Z_1, W; \beta)$ and $q(Z_2, W; \beta)$ appear only in the general case. (5) implies $q(W; \beta) = (q_1(W; \beta)', q_2(W; \beta)')'$ in the P-case. See the definitions in (6) and (7).

Preliminary estimators $\widehat{p}$ and $\widehat{q}(\beta)$ of these nuisance parameters can be plugged in $\varphi(D_1, D_2, Z_1, Z_2, W; \beta)$ and $\varphi_P(D_1, D_2, Z_1, Z_2, W; \beta)$, and then β^0 can be estimated in both cases by standard GMM as:

$$\widehat{\beta} = \arg\min_{\beta \in \mathcal{B}} \left(\bar{m}_N\left(\beta; \widehat{p}, \widehat{q}(\beta)\right)\right)' \Sigma_N \left(\bar{m}_N\left(\beta; \widehat{p}, \widehat{q}(\beta)\right)\right). \tag{11}$$

We refer to $\widehat{\beta}$ as the AIPW-GMM estimator. Σ_N is some positive semidefinite weighting matrix and

$$\bar{m}_N(\beta; p^*, q^*) = \frac{1}{N} \sum_{i=1}^{N} m_i(\beta; p^*, q^*)$$

for some generic vectors p^* and q^*. In the general case and the P-case respectively:

$$m_i(\beta; p^*, q^*) = \varphi(D_{1i}, D_{2i}, Z_{1i}, Z_{2i}, W_i; \beta) \text{ and } m_i(\beta; p^*, q^*) = \varphi_P(D_{1i}, D_{2i}, Z_{1i}, Z_{2i}, W_i; \beta)$$

with the relevant elements or functions of elements of $p(W)$ and $q(\beta)$ on both righthand sides replaced by the corresponding elements or functions of elements of some generic vectors p^* and q^* respectively.

We use the computationally convenient series estimators $\widehat{p}$ and $\widehat{q}(\beta)$ for the nuisance parameters $p(W)$ and $q(\beta)$ in (11). See the Supplemental Appendix (Appendix B) for details. If the number of terms K in the approximating series increases with sample size N, the series estimators for $p(W)$ and $q(\beta)$ are nonparametric. Then the estimator $\widehat{\beta}$ in (11) is the semiparametric AIPW-GMM estimator. On the other hand, a fixed K gives parametric series estimators for $p(W)$ and $q(\beta)$. Then the estimator $\widehat{\beta}$ in (11) is what is referred to as the parametric AIPW-GMM estimator in this paper.

Consistency, asymptotic normality and semiparametric efficiency of the semiparametric AIPW-GMM estimators $\widehat{\beta}$ under both cases follow under technical conditions similar to that in Cattaneo (2010). Some of these technical conditions are listed under Assumptions g, q and T in Appendix A.

Proposition 2.3 *Let* (1), (4) *and Assumptions M, g, q and T hold. Let the series estimators* $\widehat{p}$ *and* $\widehat{q}(\beta)$ *be based on power series or spline basis functions. Let* $p(w)$ *and* $q(u;\beta)$ *be* s *times differentiable with* $\frac{s}{d_u} > 5\frac{\eta}{2} + \frac{1}{2}$ *where* d_u *is the number of elements in* u*, and* $\eta = 1$ *for power series basis while* $\eta = \frac{1}{2}$ *for spline.* U *can be* $(Z_1', W')'$*,* $(Z_2', W')'$ *and* W *in the general case, while* U *is* W *in the P-case. Then results (i)-(ii) hold if the number of terms in the series is* $K = N^v$ *where* v *satisfies:*

$$4\eta + 2 < \frac{1}{v} < 4\frac{s}{d_u} - 6\eta. \tag{12}$$

(i) $\Sigma_N \xrightarrow{P} V^{-1}$ *implies that* $\sqrt{N}\left(\widehat{\beta} - \beta^0\right) = -\Omega^{-1}G'V^{-1}\sqrt{N}\bar{m}_N(\beta^0, p, \bar{q}(\beta^0)) + o_p(1)$.

(ii) $\widehat{V}_N := \frac{1}{N}\sum_{i=1}^N m_i(\widehat{\beta}, \widehat{p}, \widehat{q}(\widehat{\beta}))m_i(\widehat{\beta}, \widehat{p}, \widehat{q}(\widehat{\beta}))' \xrightarrow{P} V$ *and* $\widehat{G}_N := \frac{1}{N}\sum_{i=1}^N \frac{\partial}{\partial\beta'}\widehat{q}(U_i;\widehat{\beta}) \xrightarrow{P} G$.

(i) and (ii) hold in P-case under (5)*, and with* Ω*,* V *and* G *replaced by* Ω_P*,* V_P *and* G_P *respectively.*

Remarks: The results are not new except for minor differences from Theorems 2 - 8 in Cattaneo (2010). A sketch of the proof of (i) is provided under the general case since this requires additional verifications unique to the setup considered in our paper. Other proofs are omitted for brevity.

Proposition 2.3, however, does not fully exploit a special property of the moment restrictions in (9) and (10). The moment vectors in both have the special double-robustness property of Scharfstein et al. (1999) which, in terms of the notation in the general representation in (11), means

$$E\left[m_i(\beta; p(W), q^*)\right] = E\left[m_i(\beta; p^*, q(\beta))\right] = E\left[m_i(\beta; p(W), q(\beta))\right]. \tag{13}$$

Two implications of this property are well known for parametric AIPW-GMM estimators introduced by Robins and his co-authors, and their refinements studied in, for example, Robins et al. (1994), Scharfstein et al. (1999), Bang and Robins (2005), Tan (2010), Tan (2011), Cao et al. (2009), Graham et al. (2012), etc. The first implication of (13) is: the parametric AIPW-GMM estimators $\widehat{\beta}$ are consistent for β^0 as long as the parametric specification for either $p(W)$ or $q(\beta)$ is correct. The second implication is: when both parametric specifications are correct and $\Sigma_N \xrightarrow{P} V^{-1}$ in the general case and $\Sigma_N \xrightarrow{P} V_P^{-1}$ in the P-case, the parametric AIPW-GMM estimators have asymptotic variance equal to the lower bounds presented in Propositions 2.1 and 2.2. These results are extremely useful because in most empirical studies, including the one in Section 5, the dimension of W necessary to argue for (4) and the available sample size warrant the use of parametric AIPW-GMM estimators.

Recently Rothe and Firpo (2012) used local polynomial estimators $\widehat{p}$ and $\widehat{q}(\beta)$ for estimation of β^0 in moment conditions models such as (9) or (10), and studied the advantages of the double-robustness property in (13) in the context of semiparametric AIPW-GMM estimators. They showed

that (13) leads to asymptotic efficiency of such estimators of β^0 under weaker restrictions on the smoothness of $p(W)$ and $q(\beta)$, and the bandwidth for the kernels without requiring the standard faster than uniform $N^{1/4}$-consistency of the local polynomial estimators $\widehat{p}$ and $\widehat{q}(\beta)$.

Such results are also possible for the semiparametric AIPW-GMM estimators in (11) considered in our paper based on series estimators $\widehat{p}$ and $\widehat{q}(\beta)$. For example, in Proposition 2.4 we consider the P-case and exploit the particular structure of the moment function, that was not possible under the generality of exposition in Rothe and Firpo (2012), to show efficiency under even weaker restrictions on the convergence of $\widehat{p}$ and $\widehat{q}(\beta)$. A similar result can be obtained under the general case.

Proposition 2.4 *Let* (1), (4), (5) *and Assumptions M, g, q and T (defined in Appendix A) hold. Let* $p(w)$ *and* $q(w;\beta)$ *be respectively* s_p *and* s_q *times differentiable in* w*, and* d_w *be the number of elements in* w*. Let the series estimators* $\widehat{p}$ *and* $\widehat{q}(\beta)$ *be based on power series basis functions. Let* $K_p = N^{v_p}$ *and* $K_q = N^{v_q}$ *be the number of terms in the corresponding series. Let* $v_p, v_q > 0$ *and*

$$v_p < \frac{1}{3}, \quad v_p + v_q < \frac{1}{2}, \quad \frac{s_p}{d_w} > \max\left(\frac{3}{2}, 1 + \frac{v_q}{v_p}\right), \quad \frac{s_q}{d_w} > \frac{1}{2} + \frac{v_p}{v_q}, \quad v_p\frac{s_p}{d_w} + v_q\frac{s_q}{d_w} > \frac{1}{2} + v_p + \frac{v_q}{2}. \tag{14}$$

Then $\Sigma_N \xrightarrow{P} V_P^{-1}$ *implies that* $\sqrt{N}\left(\widehat{\beta} - \beta^0\right) = -\Omega_P^{-1} G_P' V_P^{-1} \sqrt{N}\bar{m}_N(\beta^0, p, \bar{q}(\beta^0)) + o_p(1)$.

Remark: The asymmetry in the restrictions required for $p(W)$ and $q(W;\beta)$ in (14) is due to the difference in how they enter the moment function and how we treat them to ensure that $\widehat{p}(W) \in (0,1)$.

The main message of Proposition 2.4 is condition (14). Recall that condition (12) in Proposition 2.3 ensures $\sup_w |\widehat{p}(w) - p(w)| = o_p(N^{-1/4})$ and $\sup_w |\widehat{q}(w;\beta^0) - q(w;\beta^0)| = o_p(N^{-1/4})$ similar to equation (4.13) in Andrews (1994) that are part of the set of conditions sufficient for the preliminary nonparametric estimation of the nuisance parameters to have no effect on the asymptotic variance of semiparametric estimators of β^0 (Assumption N(c) in Andrews (1994)). This is standard and well known. On the other hand, Proposition 2.4 is novel because now, by virtue of fully exploiting (13), taking $s_p = s_q = s$ and $v_p = v_q = v$ in (14) in Proposition 2.4 implies that

$$\frac{s}{d_w} > 2 \text{ and } 4 < \frac{1}{v} < 4\frac{s}{d_w} - 3$$

are sufficient for the same asymptotic result to hold for the AIPW estimator. Hence the required restriction on v (when $K = N^v$) with the less desirable power series basis ($\eta = 1$) in Proposition 2.4 is the same as that with the more desirable spline basis ($\eta = \frac{1}{2}$) in Proposition 2.3, while those for the smoothness s of $p(w)$ and $q(w;\beta)$ are closely comparable. (The non-binding restriction on s/d_w is related to the remark below Proposition 2.4.) This result is in the spirit of Rothe and Firpo (2012)

who also provide an intuitive discussion. They use local polynomial estimators for the nonparametric nuisance parameters and, based on a general framework, require $\sup_w |\widehat{p}(w) - p(w)| = o_p(N^{-1/6})$ and $\sup_w |\widehat{q}(w;\beta^0) - q(w;\beta^0)| = o_p(N^{-1/6})$. Condition (14) can accommodate for *such or slower rates*.

On the other hand, (14) in Proposition 2.4 is not sufficient for asymptotic efficiency in the P-case for the Horvitz and Thompson (1952)-type semiparametric IPW-GMM estimator defined as:

$$\widehat{\beta}_{P,\text{IPW}} := \arg\min_{\beta\in\mathcal{B}} \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} \frac{D_{1i}}{\widehat{p}_1(W_i)} g_1(Z_{1i}, W_i; \beta) \\ \frac{D_{2i}}{\widehat{p}_2(W_i)} g_2(Z_{2i}, W_i; \beta) \end{pmatrix}' \Sigma_N \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} \frac{D_{1i}}{\widehat{p}_1(W_i)} g_1(Z_{1i}, W_i; \beta) \\ \frac{D_{2i}}{\widehat{p}_2(W_i)} g_2(Z_{2i}, W_i; \beta) \end{pmatrix}$$

where the subscript P stands for P-case. A remarkable result due to Hirano et al. (2003), and later by Chen et al. (2008) and Cattaneo (2010), is that the IPW estimator $\widehat{\beta}_{P,\text{IPW}}$ is semiparametrically efficient under suitable convergence rates of the nonparametric estimators $\widehat{p}_1(W)$ and $\widehat{p}_2(W)$. In particular, Cattaneo (2010) showed that this holds under the conditions of Proposition 2.3.

We argue that this cannot hold in the general case in (1) (i.e., without the partition in (5)) which involves multi(more than two)-level missingness. (Recall that each element of the moment function in the P-case in (5) involves two levels of missingness: for the first d_{g_1} elements it is $D_1 = 0/1$ while for the next d_{g_2} it is $D_2 = 0/1$.) To see the deficiency of the semiparametric IPW-GMM estimator under multi-level missingness, let us first define the estimator precisely for the general case as

$$\widehat{\beta}_{\text{IPW}} := \arg\min_{\beta\in\mathcal{B}} \left(\frac{1}{N}\sum_{i=1}^{N} \frac{D_{1i}D_{2i}}{\widehat{p}_{11}(W_i)} g(Z_i, W_i; \beta)\right)' \Sigma_N \left(\frac{1}{N}\sum_{i=1}^{N} \frac{D_{1i}D_{2i}}{\widehat{p}_{11}(W_i)} g(Z_i, W_i; \beta)\right).$$

Cattaneo (2010)'s result shows that under the conditions of Proposition 2.3, $\widehat{\beta}_{\text{IPW}}$ is asymptotically normally distributed with mean 0 but asymptotic variance $\left(G'[V+\Delta]^{-1}G\right)^{-1}$ [see (8)], larger than the asymptotic variance bound $\Omega = (G'V^{-1}G)^{-1}$ in Proposition 2.1. This difference cannot be accounted for by merely maintaining a suitable rate of convergence of $\widehat{p}_{11}(W)$. Only the AIPW-approach of directly using the efficient influence function gives asymptotic efficiency in general.

An alternative but equivalent way to see the deficiency of the semiparametric IPW-approach in the general case (1), and appreciate the additional information that is effectively used by the AIPW-approach is to consider the extended set of unconditional and conditional moment restrictions:

$$E\left[\phi(D_1, D_2, Z, W; \beta)\right] = 0 \text{ for } \beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta} \text{ if and only if } \beta = \beta^0, \tag{15}$$

$$E[\phi_2(D_1, D_2, Z_2, W)|Z_2, W] = 0 \text{ almost surely in } Z_2 \text{ and } W, \tag{16}$$

$$E[\phi_1(D_1, D_2, Z_1, W)|Z_1, W] = 0 \text{ almost surely in } Z_1 \text{ and } W, \tag{17}$$

$$E[\phi_0(D_1, D_2, W)|W] = 0 \text{ almost surely in } W, \tag{18}$$

where the moment functions on the left hand side of (15)–(18) are respectively

$$\begin{aligned}
\phi(D_1, D_2, Z, W; \beta) &:= \frac{D_1 D_2}{p_{11}(W)} g(Z, W; \beta), \\
\phi_j(D_1, D_2, Z_j, W) &:= \frac{D_1 D_2}{p_{11}(W)} - \frac{D_j}{p_j(W)} \text{ for } j = 1, 2 \text{ and}, \\
\phi_0(D_1, D_2, W) &:= [D_1(1 - D_2) - p_{10}(W), (1 - D_1)D_2 - p_{01}(W), D_1 D_2 - p_{11}(W)]'.
\end{aligned}$$

For any random variable Y such that the concerned conditional expectations exist, define

$$\begin{aligned}
\overline{\text{Proj}}_W (Y|\phi_0) &:= Y - E[Y\phi_0'|W] \left(E[\phi_0\phi_0'|W]\right)^{-1} \phi_0, \\
\overline{\text{Proj}}_{Z_j,W} (Y|\phi_j) &:= Y - E[Y\phi_j|Z_j, W] \left(E[\phi_j^2|Z_j, W]\right)^{-1} \phi_j \text{ for } j = 1, 2 \text{ where}
\end{aligned}$$

ϕ_j and ϕ_0 are respectively the moment functions $\phi_j(D_1, D_2, Z_j, W)$ and $\phi_0(D_1, D_2, W)$.

The asymptotic variance of the semiparametric IPW-GMM estimator in the general case equals the efficiency bound for estimation of β by combining the moment restrictions in (15) and (18) and thus considering a modified restriction:

$$E\left[\overline{\text{Proj}}_W \left(\phi(D_1, D_2, Z, W; \beta)|\phi_0\right)\right] = 0 \text{ for } \beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta} \text{ if and only if } \beta = \beta^0.$$

See Lemma C in the Supplemental Appendix (Appendix C) for details. This approach of combining unconditional and conditional moment restrictions was originally proposed by Brown and Newey (1998), and subsequently used by Graham (2011) and Chaudhuri (2014).

On the other hand, the same approach also implies that the asymptotic variance of the semiparametric AIPW-GMM estimator in the general case equals the efficiency bound for estimation of β by combining the moment restrictions in (15)–(18) and thus considering a modified restriction:

$$E\left[\overline{\text{Proj}}_{Z_2,W} \left(\overline{\text{Proj}}_{Z_1,W} \left(\overline{\text{Proj}}_W \left(\phi(D_1, D_2, Z, W; \beta)|\phi_0\right)|\phi_1\right)|\phi_2\right)\right] = 0 \text{ for } \beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta} \text{ if and only if } \beta = \beta^0$$

if an additional conditional independence assumption, Z_1 independent of Z_2 conditional on W, is used only to obtain the projections but *not* the efficiency bound.[5] By the definitions of $\overline{\text{Proj}}_{(.)}(.|.)$ above, this bound cannot be larger than the last one based only on (15) and (18). Intuitively, this bound is smaller than the last one because it additionally uses the information in (16) and (17).

[5]The conditional independence assumption makes the results invariant to the ordering of the projections similar to the well known Frisch-Waugh-Lovell Theorem. Chaudhuri (2014) utilizes such invariance to ordering obtained through monotonicity of the conditioning sets in a similar context but under monotonically missing data. Note that we do not use the conditional independence assumption in Proposition 2.1 or in the rest of the paper. Such conditional independence has also been used by, e.g., Anderson and Perlman (1991) for maximum likelihood estimation with monotonically missing data, Abbring and Heckman (2007) to trivially identify the distribution of treatment effect (Section 2.5.1), Imai et al. (2010) to define a sequential ignorability assumption that identifies natural direct and indirect effects.

3 Three examples

Examples of the general case are standard — multiple missing regressors, missing regressor/outcome, etc. — but with multi-level non-monotone missingness. On the other hand, the structure in (5) for the P-case looks restrictive. However, the P-case is less restrictive than it looks. In this section we consider three simple examples of widely used estimation frameworks where the P-case is applicable.

Example 1: Seemingly unrelated regression (SUR) model

Consider a standard textbook-representation (for example, equation 12.21 in Davidson and MacKinnon (2004)) of a SUR model based on the moment restrictions:

$$\begin{aligned} E\left[X_1(y_1 - X_1'\beta_1)\right] &= 0 \text{ if and only if } \beta_1 = \beta_1^0 \\ E\left[X_2(y_2 - X_2'\beta_2)\right] &= 0 \text{ if and only if } \beta_2 = \beta_2^0. \end{aligned}$$

To relate this to the structure in (5) let us take $d_{\beta_0} = 0$. $W_j = X_j$ and $Z_j = y_j$ for $j = 1, 2$ lead to a missing outcome (y) model. $W_j = y_j$ and $Z_j = X_j$ for $j = 1, 2$ lead to a missing regressor (X) model. $W_1 = y_1$, $W_2 = X_2$, $Z_1 = X_1$ and $Z_2 = y_2$ lead to a missing regressor/outcome model. While the complete cases under the MAR outcome identify the true regression coefficients when the error is mean-independent of the regressor, this is not necessarily true under the weaker assumption of uncorrelatedness as above. On the other hand, the complete cases alone generally fail to identify the true regression coefficients in a MAR regressor model even under the mean-independence assumption.

Example 2: Two-sample instrumental variables (TSIV) estimation

The TSIV estimation discussed in Angrist and Krueger (1992) is an example of data combination.[6] Consider a linear instrumental variables regression model of y_1 on y_2 with X as a scalar instrument:

$$y_1 = y_2\theta^0 + \epsilon, \text{ and } y_2 = X\beta_2^0 + v.$$

TSIV considers a case with two samples where the first sample contains y_1 and X whereas the second one contains y_2 and X. y_1 and y_2 are never observed jointly. The above model gives

$$y_1 = X\beta_1^0 + u$$

where $\beta_1^0 = \theta^0\beta_2^0$ and $u = \epsilon + v\theta^0$. Taking $Z_1 = y_1$, $Z_2 = y_2$ and $W_1 = W_2 = X$, the TSIV setup is essentially a case of missing outcomes in two regressions: (i) y_2 on X and (ii) y_1 on X. Subsamples

[6] We thank a referee for suggesting this example. The technical issue of the randomness in the size of sub-samples (e.g. $\sum_{i=1}^N D_{1i}D_{2i}$) under our framework but not under standard data combination [Ichimura and Martinez-Sanchis (2005), Ridder and Moffitt (2007)] is overcome by conditioning. See Chen et al. (2008).

with $D_1 = D_2$ are also allowed in the P-case. The moment restrictions identifying β_1^0 and β_2^0:

$$\begin{aligned} E\left[X(y_1 - X\beta_1)\right] &= 0 \text{ if and only if } \beta_1 = \beta_1^0, \\ E\left[X(y_2 - X\beta_2)\right] &= 0 \text{ if and only if } \beta_2 = \beta_2^0 \end{aligned}$$

give the structure in (5). Since the parameter of interest is $\theta^0 = \frac{\beta_1^0}{\beta_2^0}$, we note that for any two estimators, say, $\widehat{\theta} = \widehat{\beta}_1/\widehat{\beta}_2$, $\widetilde{\theta} = \widetilde{\beta}_1/\widetilde{\beta}_2$, the delta method, when applicable, implies that $Avar(\widehat{\theta}) \geq Avar(\widetilde{\theta})$ if $\beta_2^0 \neq 0$ and $Avar(\widehat{\beta}_1, \widehat{\beta}_2) - Avar(\widetilde{\beta}_1, \widetilde{\beta}_2)$ is positive semidefinite. The classical minimum distance approach to the estimation of θ^0 with vector valued instruments is noted in Example 3.

Example 3: Multiple missing instrumental variables (IV)

Consider the following classical linear IV regression model of the scalar outcome y on the endogenous regressors X with IVs $Z = [Z_1', Z_2']'$:

$$y = X'\beta^0 + \epsilon. \tag{19}$$

The structure in (5) with $W_1 = W_2 = (y, X')'$ and $d_\beta = d_{\beta_0}$ follows from the moment restrictions:

$$\begin{aligned} E\left[Z_1(y - X'\beta)\right] &= 0 \text{ if and only if } \beta = \beta^0, \\ E\left[Z_2(y - X'\beta)\right] &= 0 \text{ if and only if } \beta = \beta^0. \end{aligned}$$

Viewed similarly as Example 2, the (non-monotone) missing IV problem under (19) and (20):

$$X = \Pi' Z + V \tag{20}$$

alternatively becomes a (non-monotone) missing regressors problem in two regressions (20) and (21):

$$y = Z'\Pi\beta^0 + (\epsilon + V'\beta^0) = Z'\theta + U. \tag{21}$$

Hence the missing IV setup under the P-case (5) with β^0 as the parameters of interest is equivalent to the missing regressors setup under the general case (1) with Π and θ as the parameters of interest.

Yet another representation is the classical minimum distance estimation problem under the general case with β^0 as the parameters of interest. Missing IVs affect the sample version of the problem:

$$\widehat{\beta} = \arg\min_{\beta \in \mathcal{B}} \left(\widehat{\theta} - \widehat{\Pi}\beta\right)' \Sigma_N \left(\widehat{\theta} - \widehat{\Pi}\beta\right) \text{ [for some positive semidefinite } \Sigma_N]$$

because $\widehat{\Pi}$ and $\widehat{\theta}$ need to estimated by regression with missing regressors in (20) and (21) respectively.

4 The case of MAR Instrumental Variables

To our knowledge only Mogstad and Wiswall (2012), Abrevaya and Donald (2011), Wang (2013), Chaudhuri and Min (2012) and Muris (2014) explicitly consider the case of missing IVs. Nevertheless, as documented in the former three papers, missing IVs are indeed not uncommon in empirical work.

Subsection 4.1 discusses the possible selection bias due to missing IV in a classical linear IV model and contrasts it with the bias/no-bias considerations in Mogstad and Wiswall (2012), and Abrevaya and Donald (2011). The exposition is based on a single IV that is missing for certain units in the sample. The case for multiple missing IVs is similar. Since the main theme of our paper is efficiency consideration, use of multiple IVs under the classical setup (no essential heterogeneity) is important.

Subsection 4.2 is a small-scale Monte-Carlo study showing the selection bias under MAR IV, the correction of this bias by the proposed estimators, and the gains in precision due to their use.

4.1 Missing IV: Selection bias when selection is endogenous but on observables

It is always the missingness mechanism of the IVs that characterizes the additional selection and determines if ignoring the sample units with missing IVs leads to inconsistency of the IV estimator. This is best understood by considering a simple linear model with one endogenous regressor X, one IV Z, and no covariates. Accordingly, consider the model in (19) with a scalar X. Let its coefficient β^0 be defined by the corresponding moment restriction (following (19)) with $Z \equiv Z_1 \equiv Z_2$. Under (19) and the assumption $E[ZX] \neq 0$, this is equivalent to assuming that $E[Z\epsilon] = 0$. The unconditional restriction nests as a special case the conditional moment restriction model that assumes $E[\epsilon|Z] = 0$. Abrevaya and Donald (2011) use the former while Mogstad and Wiswall (2012) use the latter. Both papers include exogenous covariates that can be accommodated easily in the following discussion.

Let the observed sample be $\{D_i, D_iZ_i, W_i = (y_i, X_i)\}_{i=1}^N$. $D = 1$ if Z is observed and 0 otherwise. Assuming that $E[DZX] \neq 0$, β^0 can be identified based on the complete cases ($D = 1$) if

$$E[DZ\epsilon] = 0. \tag{22}$$

This is the complete-case version of $E[Z\epsilon] = 0$. Abrevaya and Donald (2011) call it MAR which differs from our use of the term. Mogstad and Wiswall (2012) give sufficient conditions: (A) $E[\epsilon|Z, D = 1] = E[\epsilon|D = 1]$ and (B) $E[\epsilon|D = 1] = E[\epsilon]$ for (22). They call (B) MAR which again differs from our use of the term. (A) and (B) lead to $E[DZ\epsilon] = P(D = 1)E\left[ZE\left[\epsilon|Z, D = 1\right]|D = 1\right] = 0$ because

$$E\left[ZE\left[\epsilon|Z, D = 1\right]|D = 1\right] \overset{\text{(by A)}}{=} E\left[ZE\left[\epsilon|D = 1\right]|D = 1\right] \overset{\text{(by B)}}{=} E\left[ZE\left[\epsilon\right]|D = 1\right] = 0,$$

since $E[\epsilon] = 0$ (innocuous with an intercept). They argue that (B) is restrictive. They demonstrate the inconsistency of the IV estimator based on the complete cases by a simulation study that allows for the violation of (B) through a selection model $D = 1(a\epsilon - e \geq 0)$ where a is a non-random coefficient and e is the error independent of ϵ, X, Z (and hence y). They also propose two "robust IV"estimators that require, along with other standard assumptions, (A) but not (B) for consistency.

We note that (A) can also be restrictive especially if there is possible dependence of D on y or X that is not constrained by its dependence on ϵ, i.e., by the index $(y - \beta^0 X)$.[7] To illustrate with a simple example, append (19) with (20) (and scalar Z), and the following model for (non-)missingness:

$$D \quad = \quad \mathbf{1}(\gamma_0 y + \gamma_1 X - e \geq 0) = \mathbf{1}(\delta_0 \epsilon + \delta_1 V + \delta_2 Z - e \geq 0), \tag{23}$$

where $\delta_0 = \gamma_0$, $\delta_1 = \gamma_1 + \gamma_0\beta^0 = \gamma_1 + \delta_0\beta^0$ and $\delta_2 = [\gamma_1 + \gamma_0\beta^0]\Pi = \delta_1\Pi$ from the restrictions imposed by (19) and (20). For simplicity we maintain that e is independent of ϵ, X, Z (and hence y). While certainly not the most general specification, (23) is instructive in the context of (19) and (20). First, taking $\gamma_0 = \gamma_1 = 0$ trivially gives (A) and (B). Second, taking $\gamma_0 = a$, $\gamma_1 = -a\beta^0$ gives the violation of (B) similar to Mogstad and Wiswall (2012). Third, and importantly, note that (A) imposes

$$E[\epsilon \mid Z, D = 1] = E\left[\epsilon \mid Z, \epsilon \geq \frac{1}{\delta_0}(e - \delta_1 V - \delta_2 Z)\right] = E\left[\epsilon \mid \epsilon \geq \frac{1}{\delta_0}(e - \delta_1 V - \delta_2 Z)\right] = E[\epsilon \mid D = 1].$$

The equality in the middle may not hold in general unless $\delta_2 = 0$. This additionally imposes that $\delta_1 = 0$ unless $\Pi = 0$, i.e., unless the instrument Z is irrelevant for the endogenous regressor X to begin with. But $\delta_1 = 0$ implies that $\gamma_1 = -\gamma_0\beta^0$, i.e., the dependence of D on y and X is constrained by the index $(y - \beta^0 X)$. So the extent of selection on observables allowed by (A) is limited.

To accommodate for selection on observables, consider an alternative strategy for identification of β^0 based on the observed sample. Similar to MAR in (4), assume that almost surely in $W = (y, X)'$,

$$\text{MAR-IV: } D \text{ is independent of } Z \text{ conditional on } W. \tag{24}$$

Under MAR-IV, the identification condition (22) based on the complete cases may not hold since

$$E[DZ\epsilon] = E[DZ(y - \beta^0 X)] = E[P(D = 1|W)E[Z|W](y - \beta^0 X)] \neq 0$$

except by happenstance. This is true even under $E[\epsilon|Z] = 0$. On the other hand, the inverse

[7] This is likely if, for example, genetic markers to be used as instruments in a study are collected only for worse outcome or high exposure patients (i.e., selection based on outcome and/or endogenous regressor).

probability weighted complete case moment vector $DZ(y-\beta X)/P(D=1|W)$ identifies the original β^0 because MAR-IV in (24) gives $E\left[DZ(y-\beta X)/P(D=1|W)\right] = E\left[Z(y-\beta X)\right]$.

Now consider the limitations imposed by (24) on the selection equation defined in (23). δ_3 is not a free parameter. Therefore, although the two equivalent representations in (23) may suggest that (24) allows for selection based on unobservables from the models (19) and (20), the restriction $\delta_3 = \delta_1\Pi$ defines the extent to which it is possible. The restriction is less severe if $\delta_1 = 0$ and this mimics the selection in the simulation study of Mogstad and Wiswall (2012). Even so, the extent to which MAR-IV allows for selection based on unobservables is limited. On the other hand, MAR-IV does allow for full generality in terms of selection based on observables.

The MAR assumption in (24) is not (consistently) testable. Hence care should be taken to argue for its plausibility in any given application. On the other hand, if necessary, this opens up the possibility of cost cutting by post-survey selective collections of IVs relevant for a given study.

4.2 A Monte-Carlo study with MAR IV

The data generating process for the Monte-Carlo study is based on (19), (20) and (23) as follows:

First fix the specification of the classical linear IV model in (19) and (20). The parameter value of interest is taken as $\beta^0 = 0$ in (19). The errors ϵ and V are jointly normal with mean 0, variance 1 and three levels of correlation $\rho = 0, .5$ and $.9$. The instruments $Z = [Z_1', Z_2']'$ in (20) are mutually independent $U(1,2)$, and also independent of ϵ and V. $\Pi = [\Pi_1', \Pi_2']'$ in (20) is fixed such that the concentration parameter $\mu := N\Pi' E[ZZ']\Pi/2$ is 10 or 100. $\mu = 10$ signifies borderline strong/weak instruments while $\mu = 100$ is strong instruments. Draw i.i.d. copies of $\{\epsilon_i, V_i, Z_i\}_{i=1}^N$ and generate $\{y_i, X_i\}_{i=1}^N$ as above for the $3 \times 2 = 6$ standard IV specifications with sample size $N = 250$ and 1000.

Now fix the specification in (23) for the missingness mechanism of the instruments Z_1 and Z_2. For $j = 1, 2$, delete Z_{ji} for the i-th unit if $D_{ji} = 0$ where $D_{ji} = \mathbf{1}(\gamma_{y,j} y_i + \gamma_{X,j} X_i + e_{ji} \geq 0)$ and $e_{ji} \sim N(0,4)$ i.i.d., mutually independent, and independent of (ϵ_i, V_i, Z_i) for all $i = 1, \ldots, N$. We consider the following four specifications in Table 1 for the missingness mechanism: For $j = 1, 2$,

Specification 1	Specification 2	Specification 3	Specification 4
$\gamma_{y,j} = \gamma_{X,j} = 0$	$\gamma_{y,j} = 0, \gamma_{X,j} = 1$	$\gamma_{y,j} = 1, \gamma_{X,j} = 0$	$\gamma_{y,j} = \gamma_{X,j} = 1$

Table 1: Specifications for missingness of IVs Z_1 and Z_2. Simulation results reported in Tables 2–5.

We consider the following GMM estimators that use their respective efficient weighting matrices:

(a) "Full Obsn": The infeasible full observation two stage least squares estimator with data prior to deletion of instruments. Only Full Obsn utilizes $E[\epsilon^2|Z] = 1$ for the efficient weighting matrix.

(b) "Comp Case": The complete case estimator that ignores units with missing instruments, and hence is based on the possibly false moment restrictions in (22) corresponding to each instrument.

(c) "CC IPW" and "CC AIPW": The IPW and AIPW estimators based on the IPW-weighted, and thus bias-corrected, complete case moment restrictions: $E\left[D_1 D_2 g(Z_1, Z_2, W; \beta^0)/p_{11}(W)\right] = 0$.

(d) "IPW-GMM" and "AIPW-GMM": The estimators $\widehat{\beta}_{P,IPW}$ and $\widehat{\beta}$ ((11) for the P-case) defined in Section 2. They are semiparametrically efficient in the P-case of which this setup is a special case.

The mean bias, absolute bias, standard deviation, interquartile range of the estimators in (a)-(d), along with the coverage of the respective Wald-type confidence intervals with nominal 95% level are obtained based on 10000 Monte-Carlo (MC) trials. These are reported in Tables 2–5 for missingness mechanism Specifications 1-4 respectively. Each table contains the results for all 6 IV specifications with sample sizes $N = 250$ and 1000. For a given N, the level of endogeneity (ρ) increases from left to right and the strength of instruments (μ) from top to bottom in each table. The Full Obsn estimator always provides the infeasible benchmarks that would be attainable in the absence of missingness.

Let us first focus on the results for the Comp Case estimator to illustrate the problems with MAR IV. First, since in Table 2 the missingness is completely at random and, in this case exogenous, the Comp Case estimator should show no mean bias. This is seen when $\mu = 100$ (strong instruments). This is also seen when $\rho = 0$ and $\mu = 10$, i.e., no endogeneity and borderline strong instruments. However, note that μ is based on sample size N whereas the Comp Case only uses a fraction of these observations, and thus $\mu = 10$ based on full observations enters the domain of weak instruments under the complete case. This is reflected by the mean bias when $\rho > 0$, as would be expected in estimation with weak instruments. The above observation is important and it *counters the conventional wisdom that complete case estimation only results in efficiency loss but no bias under missingness that is completely at random.* The absolute bias, standard deviation and interquartile range follow predictable patterns with not much variation (with possible reduction in the last two) as ρ increases, but a significant decrease with the increase in μ that reduces the variability of the estimator. The second observation is from Table 3. Consider the case when $\rho = 0$. Since the missingness depends on X and hence on Z, it is not completely at random (but only at random conditional on X and y). However, since $\gamma_{y,j} = 0$ for $j = 1, 2$ (Specification 2), and since $\beta^0 = 0$ and $\rho = 0$, the missingness does not depend on ϵ (or y). Hence (22) actually holds in this case. As a result the estimated mean bias for the Comp Case estimator is very close to zero. Third observation: as ρ increases while moving right along the rows of this table, (22) does not hold any more and the mean bias increases severely. In fact, it increases to a point where the Wald-type confidence intervals with nominal confidence level 95% have actual coverage of 0. The same large mean bias and extremely

poor coverage, accompanied by large variability, are present to a greater extent in Tables 4 and 5.[8]

Now consider the CC IPW and CC AIPW estimators in (c) and the IPW-GMM and AIPW-GMM estimators in (d). The coverage of the Wald-type confidence intervals with nominal confidence level 95% are close to 95% and comparable for all these estimators. Both types of IPW estimators, in general, show worse finite-sample behavior than the corresponding AIPW estimators. In fact, the CC IPW estimator generally does not correct for the mean bias due to MAR IV in Tables 3–5. The IPW-GMM estimator that utilizes the partition in (5), and hence may be less susceptible to the overlap (common support) problem, is relatively more effective in correcting for this bias. While both AIPW estimators correct for mean bias, the AIPW-GMM estimator in (d) is especially more effective precisely for the above reason. This becomes important, for example, in Table 5 when $\rho > 0$. The AIPW estimators also display better behavior with respect to the absolute bias, standard deviation and interquartile range. In this respect they are often markedly more desirable than the IPW estimators under missingness specifications 2, 3 and 4.[9,10]

5 Empirical Application

The estimation methods with missing IVs are applied here in an empirical study of Becker (1960)'s quantity/quality model. With refinements by Becker and Lewis (1973), Willis (1973) and Becker and Tomes (1976), Becker's hypothesis was designed to explain the negative correlation between child quality related outcomes and family size. Rosenzweig and Wolpin (1980) provide the first rigorous test of this hypothesis using data from India where they use the exogenous occurrence of twins to construct an IV for fertility in a model that uses mother's fertility as an explanatory variable for child's education. Their findings suggest that even after controlling for the endogeneity of fertility, an increase in family size is related to a decrease in child quality. This supports Becker's hypothesis.

Since this early work, the quantity/quality model has been tested with a variety of data sets in several countries. Angrist et al. (2006) use data from Israel with twin births and mixed sibling-sex

[8]The coverage is only good in Table 4 for the case where $\rho = 0$ and $\mu = 10$ (i.e., borderline strong instruments) because the standard deviation is so large that these intervals are essentially uninformative.

[9]Interestingly, when $\mu = 10$ (i.e., when instruments were only borderline strong even prior to instrument deletion) the CC AIPW estimator is generally better in all these aspects that are related to lesser variability than the AIPW-GMM estimator. Hence, it should be noted that the first-order asymptotic theory, according to which the asymptotic variance of the AIPW-GMM estimator in (d) cannot be larger than that of the CC AIPW estimator in (c), may not provide a good approximation of the finite-sample behavior of the semiparametric estimators especially if the instrument strength is low. This is one reason why we reported the Monte-Carlo standard deviations of estimators. Exploring the second order asymptotics for better approximation is a promising possibility and is the topic of our future research.

[10]A referee pointed out that our proposed estimators will behave poorly under limited overlap that violates Assumption M(2). See Khan and Tamer (2010) for more on limited overlap. One could possibly deal with this serious problem by following Chaudhuri and Hill (2013) and negligibly trim the IPW-moment vector adaptively. Extending the results of Chaudhuri and Hill (2013) to the context of the present paper is a nontrivial task that is left for future research.

composition as IVs. While their ordinary least squares (OLS) results do show significant negative effects of family size on education, their IV results find no evidence of such a tradeoff. Black et al. (2005) use a very large data set from Norway and find that controlling for birth order causes the negative effect of family size on education to become negligible. This result is robust to the use of twins as an IV for family size. Juhn et al. (2013) point out that the negligible effect of family size in Norway may be related to the strong public education system in Norway which may reduce variation in educational outcomes. Using NLSY data from the United States, they find significant negative effect of family size on educational outcomes. The negative effect is stronger using IVs instead of OLS. Li et al. (1994) use data from the Chinese Population Census and also get a significant negative effect of family size which also becomes larger in their IV results (with twin births as IV) compared to OLS. The effect is more prominent in rural China where the public education system is poor. Malarani (1994) uses the 1993 and 1997 waves of the Indonesian Family Life Survey to analyze the quantity/quality tradeoff separately for urban and rural populations for several cohorts. She uses the occurrence of miscarriages as IV and finds mixed evidence in support of the hypothesis with major differences for the urban and rural samples.

The empirical study in this section takes advantage of community level information available in the Indonesian Family Life Survey (IFLS) to construct a very different set of IVs than have been used in the past to examine the quantity/quality tradeoff. The data used encompasses all four waves (1993, 1997, 2000, and 2007) of the longitudinal data from the IFLS.[11] An important feature of the IFLS is that it included community and facility surveys which provide current and retrospective information on community infrastructure and availability of family planning, health and schooling facilities that are relevant for the household survey respondents. This information was gathered through interviews with community leaders, visits to local health and family planning facilities, and data extracted from community records. This information is particularly relevant in Indonesia because Indonesia's family planning program is credited with being a major factor in the decline in fertility in Indonesia. This is important in our analysis for two reasons. First, because our data was gathered at least partly during the period of rapid program expansion there is a great deal of variation in family size in our data set which is useful in our empirical exercise. Second, since the program is considered to be a

[11]The data set gathered information at the individual and family level on fertility, health, education, migration, and employment. The sample was drawn from 13 of Indonesia's then 27 provinces (there are now 33). These 13 provinces (North, West, and South Sumatra; West, Central, and East Java; Lampung; DKI Jakarta; DI Yogyakarta; Bali; West Nusa Tenggara; South Kalimantan; and South Sulawesi) encompass roughly 83 percent of Indonesia's population. These provinces were selected in order to "maximize representation of the population, capture the cultural and socioeconomic diversity of Indonesia, and be cost effective given the size and terrain of the country". Enumeration areas were randomly selected from the 13 provinces using the 1993 SUSENAS (Survei Sosial Ekonomi Nasional) sampling frame, with over sampling of urban enumeration areas and those in smaller provinces (321 enumeration areas were selected in total). Households were randomly selected by field teams within each enumeration area.

major contributor to fertility decline, variables associated with the program are good candidates for use as IVs for the number of children born to a woman.

The expansion of the family planning program in Indonesia, however, was not random. It started on densely populated Java in the early 1970's and then expanded to other highly developed provinces [see Gertler and Molyneaux (1994) and Shiffman (2004)]. Because of this, we included a variety of measures of community infrastructure such as: if the community had transportation, paved roads, piped water, and good sanitation in addition to the program variables in preliminary runs. However, our results are robust to their omission and they are not included in the results reported below.

The sample used in the empirical analysis is cross-sectional in the sense that we use one observation for each mother/child pair from the last interview where the mother/child pair has information. All biological children are included except those who left home prior to the mother's first interview. If a child dies or leaves the household after the mother's first interview, they are included. Thus the full sample in our study contains 7,370 children between age 11 and 20. Simple descriptive statistics are presented in Table 6. The child quality measure that we use is the continuous years of education for each child which averages 7.74 years across the sample. The key explanatory variable is the number of siblings (censored above at 4) for the index child with 2.34 siblings being the average.[12,13]

The empirical model that we specify is essentially the same as that has been used in the literature:

$$\text{Years of Schooling} = \beta_{\text{sib}} \times (\text{Number of Siblings}) + \beta'_{\text{controls}} \times (\text{Control Variables}) + \epsilon$$

where the control variables include an intercept, dummies for the child's age, sex, dummies for education levels of mother, mother's husband, if the mother is Muslim, if she can read, and also community level variables such as number of elementary, junior and high schools. See Tables 7 and 8 for details. A more realistic but less parsimonious model would allow for interactions of the righthand side variables, or at least allow the sibling coefficient to vary with the number siblings.

The key argument for using IVs in our study (and in the quantity/quality model in general) to consistently estimate the coefficients is that the error term ϵ is correlated with the number of siblings which generally rules out consistency of the OLS estimates (in Table 7) of the coefficients on the righthand side. Under the classical IV assumptions, IVs uncorrelated with ϵ can be used. One does not need to worry about the "first stage" as long as the IVs are "strong". Moreover, with "reasonably" large sample size that rules out significant finite-sample bias, it is advisable to use more

[12]Number (percentage) of children with 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14 siblings are 578 (7.84%), 210 (2.85%), 211 (2.86%), 82 (1.11%), 26 (.35%), 3 (.04%), 1 (.01%) and 4 (.05%) respectively. About 15% observations were censored.

[13]The community variables are taken as the facilities in the child's community in the last year when the child was available for the survey. Ideally, one would like to backdate these variables to the beginning of the mother's child bearing years; unfortunately the date of establishment of the facilities was mostly missing.

such IVs to reduce the standard errors of the coefficient estimates. Of course, the coefficients thus estimated can be meaningless and even misleading (in sign) when the classical IV assumptions are wrong. See Heckman and Vytlacil (2007) and the references therein to the authors' earlier papers for a detailed exposition of this topic. However, in this section we abstract from such possibilities to focus instead on the application of the recommended methods in this simple empirical model.

The IFLS data contains six variables that could possibly work as IVs. The first four — the number of Posyandus (small health center providing family planning and other services), and three dummy variables for the presence of a delivery post, a medical post, and a midwife in the community — are always observed. The other IVs — the number of family planning posts and the dummy for the presence of a Polindes in the community — have many missing values. A family planning post is a government run facility specializing in family planning programs, contraceptive distribution, and child health. The presence of Polindes indicates the availability of health facilities in the community. Define $D_1 = 1/0$ if the IV related to Polindes is observed/missing, and $D_2 = 1/0$ if the IV related to family planning post is observed/missing. Then $D_1 = D_2 = 1$ for about 33% units only [see Table 6]. These units form the complete cases in our study. At the very least, this presents a tradeoff between using more IVs versus more sample units if IV estimation is based on the complete cases only.

The other issue with using the complete cases only is the selection problem: the topic of Section 4. Evidence of selection can be found from Table 6 that presents the means in the sub-samples broken down by the missingness of the two IVs. First, the key variable of our analysis, i.e., the years of schooling for the child has a much higher mean for the sub-sample where both IVs are missing relative to the sub-sample where both are observed. This is in spite of the fact that the mean age of children is larger in the latter. The same pattern holds for the mother's and her husband's education levels, and the mother's ability to read. The community variables tend to take on desirable values in the missing sub-sample as well which tends to be more urban than the sub-sample with both IVs not missing. Of course, we control for these variables in the regression model but it is not clear to what extent a simple linear model specification can account for this selection. In fact, the naive OLS estimates reported in Table 7 also indicate very different patterns in the relationship of the years of schooling with the number of siblings and the control variables in the sub-samples categorized by the missingness of these two IVs. Hence complete case analysis should proceed with caution.

A clean analysis, perhaps at the cost of precision, that rules out the possible selection problem (due to complete cases) under the classical IV setup would use the four IVs that are always observed. Results for this are reported in column 1 (GMM: no missing IV) of Table 8. The estimated coefficient for the number of siblings is -.73 with standard error .303. The magnitude is much larger than the

full sample OLS estimate in Table 7. This is reasonably consistent with the literature on this topic. (The other coefficient estimates are less different from the corresponding OLS estimates.) Our IV estimate of -.73 is indeed large compared to other IV estimates in this literature. For example, Juhn et al. (2013) obtained a point estimate -.36 for the effect of family size on years of schooling using the NLSY with twins as IV. Ignoring any heterogeneity in the "first stage", one could still attempt to explain this by noting that there is more variability in educational attainment in Indonesia than in the US — a standard deviation of a little over 3 in Indonesia versus a standard deviation of 2.3 in the NLSY data used by Juhn et al. (2013) — see Table 1 in their paper. Under our classical IV setup, we maintain the results from column 1 of Table 8 as the benchmark for our empirical illustration.

On the other hand, column 2 of Table 8 reports that the complete (COMP) case IV estimate of the coefficient for the number of siblings is -1.01 with standard error .524. This estimate is much larger in magnitude than the benchmark. Also, using more instruments did not reduce the standard error by offsetting the loss in sample size. The null hypothesis of underidentification (rank-deficient Jacobian) is strongly rejected, while that of overidentification cannot be rejected at the 44% level. However, the results of these specification tests should not be considered reassuring. The latter only supports the hypothesis that there exists a value, say, β^* at which the complete case moment restrictions hold [see (22)]. The former suggests that this β^* is locally identified and hence is precisely estimable. β^* need not be the same β at which the original IV moment restrictions hold. This is evident from the difference between the benchmark estimates and those based on the complete cases.

Now we turn to the results in columns 3 and 4 of Table 8 obtained by the IPW-GMM and AIPW-GMM estimators respectively. The estimates from both are very close to the benchmark. Both estimators assume MAR in (4) to hold with the conditioning set W specified as the four non-missing IVs, all the control variables, and the years of schooling of the child. As mentioned before, the specification of W is an extremely difficult task. In fact, absent complete knowledge of the missingness mechanism (which we do not have), it is impossible to be sure if a specification is correct.[14] Nevertheless, some incorrect specifications can be statistically ruled out in large samples (that allow for consistent nonparametric estimation of the nuisance parameters without specifying parametric models for them) under the conditions of Proposition 2.3/2.4 if the IPW-GMM and AIPW-GMM estimators are different (Hausman test). This, however, is not the case in our analysis.[15]

[14] Any ad-hoc forward/backward rule of adding or removing variables to determine the proper conditioning set (provided it exists) is misleading since, for example, given four random variables A, B, C, D, the condition $A \perp B|C, D$ neither implies nor is implied by the condition $A \perp B|C$ in general.

[15] It is important to remember that the different patterns of the descriptive statistics in Table 6 are not evidence against MAR, since MAR in (4) does not require the same marginal distributions of the conditioning variables W in the populations of the missing and non-missing samples. Common support of W is required for identification, but there is no evidence against it in our data. MAR does impose that the conditional distribution of the missing variables in Z

For all four estimators in Table 8 we report two sets of standard errors inside parentheses: the first one is obtained directly from STATA by using the asymptotic variance while the second one is based on 1000 bootstrap samples. Technically, the former standard errors for the IPW-GMM estimates are incorrect in this example because they do not take into account the preliminary estimation of the nuisance parameters. However, both sets of standard errors for each of the four estimators are more or less comparable and do not alter the significance levels of the corresponding coefficients. Finally, it should be noted that the standard errors of the AIPW-GMM estimates are uniformly smaller than the corresponding standard errors of the benchmark (no missing IV) and the IPW-GMM estimates.

In the context of this empirical study, both the IPW-GMM and AIPW-GMM estimates correct for the possible selection bias in the complete case estimates and corroborate the benchmark results in column 1 of Table 8. Under the classical IV setup and the other maintained assumptions, this empirical study provides strong evidence in support of Becker's quantity/quality model in Indonesia.

6 Conclusion

We showed that a version of the well known MAR assumption point-identifies and provides a closed-form efficient influence function for estimation of parameters in a general moment conditions model when relevant variables can be non-monotonically missing from the sample. We established asymptotic normality and efficiency for the recommended AIPW-GMM estimators under standard technical conditions. While many papers have recently established asymptotic equivalence of IPW and AIPW estimators under carefully chosen regularity conditions, we showed that such equivalence cannot hold generally: it does not hold for the general framework in our paper. The AIPW-GMM estimator provides a natural and convenient route for efficient estimation, and we showed that it is indeed based on markedly weaker conditions than what are required for general semiparametric estimators.

We also closely studied a particular case where the moment function partitions according to the observability of variables. Various commonly used estimation methods fall under this case. Special attention is paid to the less studied case of missing instrumental variables. We demonstrated the selection problem due to ignoring the sample units with missing instruments, and cautioned against such empirical practice. On the other hand, based on the theoretical results and the Monte-Carlo experiment we argued that the recommended estimation methods do not suffer from such problems asymptotically and possibly in finite samples under the framework of the present paper. Several promising extensions of these methods were pointed out in the paper and were left for future research.

given W are the same in all concerned populations but, as noted before, this is fundamentally untestable consistently.

	$\mu = 10,\ \rho = 0,\ N = 250$					$\mu = 10,\ \rho = .5,\ N = 250$					$\mu = 10,\ \rho = .9,\ N = 250$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.005	0.277	0.420	0.427	96.72	0.002	0.274	0.411	0.403	96.44	0.007	0.272	0.426	0.387	96.68
Comp Case	0.008	0.567	1.011	0.723	96.43	0.145	0.575	0.974	0.660	95.83	0.270	0.581	1.051	0.492	96.53
CC IPW	0.001	0.302	0.440	0.449	95.81	0.004	0.303	0.496	0.437	96.97	0.021	0.302	0.506	0.400	96.91
CC AIPW	0.005	0.283	0.394	0.431	95.21	0.015	0.289	0.484	0.417	97.3	0.030	0.286	0.449	0.391	96.7
IPW-GMM	0.009	0.305	0.450	0.457	96	-0.063	0.316	0.547	0.451	96.85	-0.121	0.338	0.681	0.456	97.09
AIPW-GMM	0.008	0.304	0.451	0.455	96	-0.087	0.333	1.691	0.452	99.71	-0.122	0.336	0.724	0.456	97.54
	$\mu = 100,\ \rho = 0,\ N = 250$					$\mu = 100,\ \rho = .5,\ N = 250$					$\mu = 100,\ \rho = .9,\ N = 250$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.001	0.081	0.102	0.134	94.97	0.001	0.081	0.101	0.136	95.08	0.002	0.081	0.103	0.134	95.37
Comp Case	-0.002	0.162	0.208	0.264	94.7	0.001	0.162	0.212	0.259	95.28	-0.001	0.163	0.220	0.253	95.61
CC IPW	0.001	0.088	0.111	0.148	94.9	0.001	0.087	0.111	0.146	94.82	0.002	0.087	0.111	0.144	95.48
CC AIPW	0.001	0.083	0.106	0.138	94.65	0.003	0.083	0.105	0.140	94.92	0.005	0.084	0.106	0.139	95.4
IPW-GMM	0.001	0.084	0.106	0.141	94.89	-0.004	0.084	0.106	0.142	95.12	-0.007	0.084	0.108	0.141	95.05
AIPW-GMM	0.001	0.084	0.106	0.139	94.81	-0.004	0.083	0.105	0.141	95.19	-0.007	0.084	0.107	0.141	95.08
	$\mu = 10,\ \rho = 0,\ N = 1000$					$\mu = 10,\ \rho = .5,\ N = 1000$					$\mu = 10,\ \rho = .9,\ N = 1000$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.004	0.269	0.368	0.416	95.12	0.003	0.274	0.430	0.406	96.9	0.004	0.272	0.479	0.374	97.17
Comp Case	-0.016	0.592	1.360	0.729	98.09	0.145	0.582	1.173	0.672	97.24	0.263	0.566	1.010	0.513	96.32
CC IPW	0.005	0.281	0.400	0.430	95.72	0.006	0.282	0.417	0.418	96.41	0.015	0.283	0.530	0.387	97.49
CC AIPW	0.005	0.276	0.389	0.425	95.48	0.008	0.279	0.412	0.414	96.45	0.014	0.279	0.477	0.378	97.21
IPW-GMM	0.005	0.295	0.429	0.443	95.96	-0.061	0.310	0.548	0.446	97.17	-0.117	0.335	0.759	0.443	97.6
AIPW-GMM	0.004	0.294	0.425	0.441	95.95	-0.060	0.309	0.530	0.448	96.98	-0.118	0.336	0.783	0.443	97.75
	$\mu = 100,\ \rho = 0,\ N = 1000$					$\mu = 100,\ \rho = .5,\ N = 1000$					$\mu = 100,\ \rho = .9,\ N = 1000$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.001	0.080	0.100	0.135	94.95	0.000	0.079	0.101	0.132	94.72	-0.002	0.081	0.103	0.135	95.34
Comp Case	-0.002	0.163	0.210	0.263	94.74	0.001	0.159	0.205	0.257	95.04	-0.002	0.162	0.219	0.253	96.03
CC IPW	0.001	0.083	0.105	0.139	94.97	-0.001	0.082	0.105	0.136	94.79	-0.002	0.083	0.106	0.138	95.24
CC AIPW	0.001	0.082	0.103	0.137	95.08	0.000	0.081	0.104	0.135	94.95	-0.001	0.082	0.105	0.136	95.09
IPW-GMM	0.001	0.082	0.103	0.136	94.83	-0.005	0.082	0.105	0.137	94.78	-0.011	0.083	0.106	0.138	94.86
AIPW-GMM	0.001	0.082	0.103	0.136	94.86	-0.005	0.082	0.104	0.136	94.84	-0.010	0.083	0.106	0.138	94.9

Table 2: Specification 1: $D_j = \mathbf{1}(e_j \geq 0)$ where $e_j \sim N(0,4)$ i.i.d. for $j = 1,2$. IV model: $y = \beta^0 X + \epsilon$ and $X = \Pi_1 Z_1 + \Pi_2 Z_2 + V$ where ϵ, V are $N(0,1)$ with $Cov(\epsilon, V) = \rho$ are independent of e and $Z_j \sim U(1,2)$ i.i.d for $j = 1,2$ are independent of e, ϵ, V. Parameter of interest is $\beta^0 = 0$. Mean bias (M. Bias), absolute bias (A. Bias), standard deviation (MC. Std) and interquartile range (MC. IQR) of estimators of β^0 and coverage of the respective nominal-95% Wald-type confidence intervals (95% CI) are reported based on 10000 trials with sample size $N = 250, 1000$ under various levels of $\rho = 0, .5, .9$ (i.e. X's endogeneity) and the concentration parameter $\mu := N\Pi' E[ZZ']\Pi/2 = 10, 100$ (i.e., instrument $Z = [Z_1', Z_2']'$'s strength).

	$\mu = 10,\ \rho = 0,\ N = 250$					$\mu = 10,\ \rho = .5,\ N = 250$					$\mu = 10,\ \rho = .9,\ N = 250$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.005	0.277	0.420	0.427	96.72	0.002	0.274	0.411	0.403	96.44	0.007	0.272	0.426	0.387	96.68
Comp Case	-0.002	0.114	0.144	0.190	94.83	0.366	0.366	0.129	0.167	18.58	0.667	0.667	0.071	0.093	0
CC IPW	-0.011	0.359	0.793	0.461	97.93	0.151	0.384	0.726	0.424	97.51	0.255	0.454	0.764	0.359	96.73
CC AIPW	0.000	0.281	0.388	0.426	95.14	0.050	0.289	0.418	0.407	96.45	0.110	0.314	0.485	0.384	97.7
IPW-GMM	0.001	0.308	0.453	0.449	95.56	-0.012	0.325	0.591	0.452	97.4	-0.034	0.371	0.922	0.456	97.68
AIPW-GMM	0.000	0.286	0.514	0.439	98.18	-0.030	0.294	0.509	0.423	97.15	-0.060	0.311	0.747	0.426	98.17
	$\mu = 100,\ \rho = 0,\ N = 250$					$\mu = 100,\ \rho = .5,\ N = 250$					$\mu = 100,\ \rho = .9,\ N = 250$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.001	0.081	0.102	0.134	94.97	0.001	0.081	0.101	0.136	95.08	0.002	0.081	0.103	0.134	95.37
Comp Case	-0.001	0.070	0.088	0.118	95.02	0.216	0.217	0.080	0.108	22.45	0.393	0.393	0.056	0.075	0
CC IPW	0.000	0.106	0.147	0.168	95.76	0.022	0.116	0.167	0.168	97.11	0.038	0.141	0.305	0.183	98.79
CC AIPW	0.001	0.083	0.105	0.137	94.72	0.009	0.083	0.105	0.139	94.95	0.016	0.087	0.108	0.139	95.22
IPW-GMM	0.001	0.088	0.116	0.146	95.44	-0.001	0.090	0.116	0.149	95.37	-0.005	0.095	0.125	0.153	95.71
AIPW-GMM	0.001	0.083	0.105	0.138	94.82	0.000	0.083	0.105	0.141	95.16	-0.001	0.084	0.107	0.140	95.33
	$\mu = 10,\ \rho = 0,\ N = 1000$					$\mu = 10,\ \rho = .5,\ N = 1000$					$\mu = 10,\ \rho = .9,\ N = 1000$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.004	0.269	0.368	0.416	95.12	0.003	0.274	0.430	0.406	96.9	0.004	0.272	0.479	0.374	97.17
Comp Case	0.001	0.065	0.082	0.109	94.88	0.426	0.426	0.071	0.096	0	0.771	0.771	0.038	0.051	0
CC IPW	0.012	0.320	0.585	0.420	96.85	0.174	0.372	0.755	0.396	98.22	0.322	0.476	0.861	0.330	97.59
CC AIPW	0.007	0.272	0.386	0.416	95.53	0.041	0.285	0.429	0.412	96.85	0.073	0.302	0.489	0.377	97.61
IPW-GMM	0.012	0.303	0.520	0.448	97.54	-0.019	0.325	0.716	0.447	98.3	-0.055	0.360	0.852	0.464	97.57
AIPW-GMM	0.007	0.278	0.394	0.425	95.76	-0.032	0.294	0.551	0.426	97.98	-0.058	0.303	0.587	0.414	97.06
	$\mu = 100,\ \rho = 0,\ N = 1000$					$\mu = 100,\ \rho = .5,\ N = 1000$					$\mu = 100,\ \rho = .9,\ N = 1000$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.001	0.080	0.100	0.135	94.95	0.000	0.079	0.101	0.132	94.72	-0.002	0.081	0.103	0.135	95.34
Comp Case	0.000	0.050	0.062	0.083	94.75	0.317	0.317	0.054	0.072	0.01	0.573	0.573	0.032	0.043	0
CC IPW	0.003	0.110	0.153	0.171	95.38	0.048	0.127	0.162	0.178	96.3	0.083	0.162	0.223	0.187	98.15
CC AIPW	0.001	0.082	0.103	0.137	94.88	0.005	0.082	0.105	0.137	94.79	0.008	0.085	0.107	0.139	95.24
IPW-GMM	0.002	0.087	0.111	0.144	94.86	-0.003	0.090	0.115	0.149	95.01	-0.007	0.097	0.126	0.158	95.32
AIPW-GMM	0.001	0.082	0.103	0.136	94.67	-0.002	0.082	0.104	0.136	95	-0.006	0.084	0.107	0.139	95.16

Table 3: Specification 2: $D_j = \mathbf{1}(X + e_j \geq 0)$ where $e_j \sim N(0,4)$ i.i.d. for $j = 1, 2$. IV model: $y = \beta^0 X + \epsilon$ and $X = \Pi_1 Z_1 + \Pi_2 Z_2 + V$ where ϵ, V are $N(0,1)$ with $Cov(\epsilon, V) = \rho$ are independent of e and $Z_j \sim U(1,2)$ i.i.d for $j = 1, 2$ are independent of e, ϵ, V. Parameter of interest is $\beta^0 = 0$. Mean bias (M. Bias), absolute bias (A. Bias), standard deviation (MC. Std) and interquartile range (MC. IQR) of estimators of β^0 and coverage of the respective nominal-95% Wald-type confidence intervals (95% CI) are reported based on 10000 trials with sample size $N = 250, 1000$ under various levels of $\rho = 0, .5, .9$ (i.e. X's endogeneity) and the concentration parameter $\mu := N\Pi' E[ZZ']\Pi/2 = 10, 100$ (i.e., instrument $Z = [Z_1', Z_2']'$'s strength).

	$\mu = 10,\ \rho = 0,\ N = 250$					$\mu = 10,\ \rho = .5,\ N = 250$					$\mu = 10,\ \rho = .9,\ N = 250$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.005	0.277	0.420	0.427	96.72	0.002	0.274	0.411	0.403	96.44	0.007	0.272	0.426	0.387	96.68
Comp Case	2.196	2.898	9.526	1.537	99.54	1.220	1.220	0.300	0.342	0.09	0.818	0.818	0.068	0.090	0
CC IPW	0.359	0.720	1.350	0.788	97.42	0.328	0.627	2.361	0.558	99.33	0.338	0.534	0.976	0.366	97.23
CC AIPW	0.025	0.321	0.469	0.480	95.76	0.079	0.335	0.910	0.449	99.49	0.121	0.339	0.569	0.407	98.19
IPW-GMM	0.048	0.382	0.589	0.555	96.46	-0.006	0.392	0.826	0.509	97.8	-0.022	0.396	0.910	0.475	97.32
AIPW-GMM	0.026	0.309	0.498	0.464	97.2	-0.029	0.307	0.529	0.441	97.14	-0.063	0.312	0.612	0.426	97.14
	$\mu = 100,\ \rho = 0,\ N = 250$					$\mu = 100,\ \rho = .5,\ N = 250$					$\mu = 100,\ \rho = .9,\ N = 250$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.001	0.081	0.102	0.134	94.97	0.001	0.081	0.101	0.136	95.08	0.002	0.081	0.103	0.134	95.37
Comp Case	0.988	0.988	0.275	0.335	1.89	0.660	0.660	0.108	0.143	0	0.521	0.521	0.054	0.071	0
CC IPW	0.114	0.215	0.328	0.279	98.25	0.099	0.206	0.430	0.239	99	0.081	0.205	0.431	0.222	98.7
CC AIPW	0.010	0.089	0.112	0.148	94.84	0.019	0.091	0.112	0.149	94.64	0.027	0.094	0.115	0.147	94.89
IPW-GMM	0.013	0.108	0.138	0.177	94.52	0.003	0.109	0.185	0.173	98.48	0.000	0.108	0.156	0.167	97.01
AIPW-GMM	0.006	0.085	0.107	0.142	94.9	0.002	0.085	0.107	0.141	95.3	0.000	0.085	0.109	0.138	95.4
	$\mu = 10,\ \rho = 0,\ N = 1000$					$\mu = 10,\ \rho = .5,\ N = 1000$					$\mu = 10,\ \rho = .9,\ N = 1000$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.004	0.269	0.368	0.416	95.12	0.003	0.274	0.430	0.406	96.9	0.004	0.272	0.479	0.374	97.17
Comp Case	4.624	5.650	10.477	2.799	98.16	1.512	1.512	0.201	0.258	0	0.943	0.943	0.040	0.054	0
CC IPW	0.564	0.869	1.552	0.888	97.32	0.443	0.655	1.003	0.549	97.65	0.380	0.544	1.058	0.340	97.96
CC AIPW	0.017	0.306	0.478	0.460	96.82	0.051	0.310	0.454	0.443	96.15	0.079	0.314	0.571	0.386	97.9
IPW-GMM	0.040	0.376	0.546	0.549	95.54	-0.012	0.371	0.635	0.529	97.1	-0.047	0.372	0.873	0.476	97.57
AIPW-GMM	0.016	0.299	0.449	0.446	96.33	-0.030	0.302	0.469	0.445	96.39	-0.064	0.311	0.722	0.423	97.98
	$\mu = 100,\ \rho = 0,\ N = 1000$					$\mu = 100,\ \rho = .5,\ N = 1000$					$\mu = 100,\ \rho = .9,\ N = 1000$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.001	0.080	0.100	0.135	94.95	0.000	0.079	0.101	0.132	94.72	-0.002	0.081	0.103	0.135	95.34
Comp Case	1.975	1.975	0.447	0.530	0	0.994	0.994	0.094	0.123	0	0.711	0.711	0.030	0.040	0
CC IPW	0.183	0.249	0.245	0.290	90.17	0.144	0.228	0.298	0.247	98.63	0.129	0.216	0.345	0.212	98.81
CC AIPW	0.005	0.087	0.109	0.145	94.77	0.008	0.087	0.110	0.145	95.13	0.011	0.088	0.111	0.141	95.25
IPW-GMM	0.016	0.105	0.131	0.173	94.81	0.006	0.104	0.133	0.167	95.07	0.001	0.105	0.136	0.170	95.54
AIPW-GMM	0.004	0.084	0.106	0.139	94.74	-0.001	0.084	0.107	0.139	95.04	-0.005	0.085	0.108	0.142	95.02

Table 4: Specification 3: $D_j = \mathbf{1}(y + e_j \geq 0)$ where $e_j \sim N(0, 4)$ i.i.d. for $j = 1, 2$. IV model: $y = \beta^0 X + \epsilon$ and $X = \Pi_1 Z_1 + \Pi_2 Z_2 + V$ where ϵ, V are $N(0, 1)$ with $Cov(\epsilon, V) = \rho$ are independent of e and $Z_j \sim U(1, 2)$ i.i.d for $j = 1, 2$ are independent of e, ϵ, V. Parameter of interest is $\beta^0 = 0$. Mean bias (M. Bias), absolute bias (A. Bias), standard deviation (MC. Std) and interquartile range (MC. IQR) of estimators of β^0 and coverage of the respective nominal-95% Wald-type confidence intervals (95% CI) are reported based on 10000 trials with sample size $N = 250, 1000$ under various levels of $\rho = 0, .5, .9$ (i.e. X's endogeneity) and the concentration parameter $\mu := N\Pi' E[ZZ']\Pi/2 = 10, 100$ (i.e., instrument $Z = [Z_1', Z_2']'$'s strength).

	$\mu = 10,\ \rho = 0,\ N = 250$					$\mu = 10,\ \rho = .5,\ N = 250$					$\mu = 10,\ \rho = .9,\ N = 250$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.005	0.277	0.420	0.427	96.72	0.002	0.274	0.411	0.403	96.44	0.007	0.272	0.426	0.387	96.68
Comp Case	0.704	0.704	0.201	0.262	3.72	0.760	0.760	0.115	0.152	0	0.792	0.792	0.048	0.062	0
CC IPW	0.281	0.491	0.869	0.517	97.72	0.414	0.543	0.638	0.383	96.73	0.513	0.583	0.594	0.233	97.32
CC AIPW	0.064	0.314	0.481	0.464	96.57	0.151	0.351	0.490	0.450	96.48	0.245	0.385	0.452	0.421	96.95
IPW-GMM	0.084	0.393	1.070	0.520	99.03	0.141	0.453	0.917	0.482	97.19	0.241	0.497	0.954	0.412	95.93
AIPW-GMM	0.037	0.296	0.450	0.450	96.47	-0.009	0.299	0.524	0.421	97.52	-0.023	0.305	0.742	0.410	98.45
	$\mu = 100,\ \rho = 0,\ N = 250$					$\mu = 100,\ \rho = .5,\ N = 250$					$\mu = 100,\ \rho = .9,\ N = 250$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.001	0.081	0.102	0.134	94.97	0.001	0.081	0.101	0.136	95.08	0.002	0.081	0.103	0.134	95.37
Comp Case	0.401	0.401	0.099	0.131	1.46	0.475	0.475	0.067	0.088	0	0.517	0.517	0.037	0.049	0
CC IPW	0.074	0.164	0.272	0.205	98.47	0.136	0.212	0.281	0.197	98.07	0.182	0.252	0.387	0.180	98.63
CC AIPW	0.013	0.085	0.107	0.142	94.66	0.027	0.092	0.112	0.147	94.32	0.044	0.101	0.118	0.149	94.01
IPW-GMM	0.009	0.108	0.165	0.169	97.66	0.006	0.135	0.239	0.186	97.67	0.014	0.169	0.427	0.197	98.53
AIPW-GMM	0.007	0.084	0.106	0.138	94.93	0.004	0.085	0.108	0.143	95.14	0.002	0.086	0.110	0.142	95.48
	$\mu = 10,\ \rho = 0,\ N = 1000$					$\mu = 10,\ \rho = .5,\ N = 1000$					$\mu = 10,\ \rho = .9,\ N = 1000$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.004	0.269	0.368	0.416	95.12	0.003	0.274	0.430	0.406	96.9	0.004	0.272	0.479	0.374	97.17
Comp Case	0.834	0.834	0.121	0.159	0	0.869	0.869	0.067	0.090	0	0.887	0.887	0.026	0.035	0
CC IPW	0.432	0.558	0.682	0.493	95.37	0.562	0.629	0.623	0.340	96.74	0.632	0.663	0.429	0.180	94.11
CC AIPW	0.067	0.306	0.471	0.446	96.61	0.134	0.335	0.456	0.443	96.01	0.204	0.371	0.499	0.402	97.71
IPW-GMM	0.091	0.394	0.682	0.525	97.32	0.190	0.464	0.882	0.490	97.38	0.257	0.502	1.405	0.393	98.78
AIPW-GMM	0.031	0.289	0.417	0.442	95.88	0.005	0.303	0.558	0.435	97.96	-0.006	0.311	0.854	0.398	98.76
	$\mu = 100,\ \rho = 0,\ N = 1000$					$\mu = 100,\ \rho = .5,\ N = 1000$					$\mu = 100,\ \rho = .9,\ N = 1000$				
Estimators	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI	M.Bias	A.Bias	MC.Std	MC.IQR	95% CI
Full Obsn	0.001	0.080	0.100	0.135	94.95	0.000	0.079	0.101	0.132	94.72	-0.002	0.081	0.103	0.135	95.34
Comp Case	0.600	0.600	0.078	0.105	0	0.666	0.666	0.047	0.064	0	0.701	0.701	0.021	0.029	0
CC IPW	0.158	0.214	0.223	0.222	93.66	0.248	0.293	0.293	0.204	98.04	0.312	0.350	0.324	0.163	98.91
CC AIPW	0.010	0.086	0.107	0.144	94.86	0.019	0.091	0.114	0.148	94.65	0.031	0.098	0.119	0.152	94.79
IPW-GMM	0.016	0.111	0.143	0.177	95.24	0.031	0.148	0.320	0.201	98.91	0.048	0.177	0.313	0.209	97.77
AIPW-GMM	0.006	0.084	0.105	0.139	94.8	0.002	0.085	0.108	0.142	94.87	-0.001	0.088	0.111	0.144	95.29

Table 5: Specification 4: $D_j = \mathbf{1}(y + X + e_j \geq 0)$ where $e_j \sim N(0, 4)$ i.i.d. for $j = 1, 2$. IV model: $y = \beta^0 X + \epsilon$ and $X = \Pi_1 Z_1 + \Pi_2 Z_2 + V$ where ϵ, V are $N(0, 1)$ with $Cov(\epsilon, V) = \rho$ are independent of e and $Z_j \sim U(1, 2)$ i.i.d for $j = 1, 2$ are independent of e, ϵ, V. Parameter of interest is $\beta^0 = 0$. Mean bias (M. Bias), absolute bias (A. Bias), standard deviation (MC. Std) and interquartile range (MC. IQR) of estimators of β^0 and coverage of the respective nominal-95% Wald-type confidence intervals (95% CI) are reported based on 10000 trials with sample size $N = 250, 1000$ under various levels of $\rho = 0, .5, .9$ (i.e. X's endogeneity) and the concentration parameter $\mu := N\Pi' E[ZZ']\Pi/2 = 10, 100$ (i.e., instrument $Z = [Z_1', Z_2']'$'s strength).

Variables	Full Sample	$D_1 = 1$	$D_2 = 1$	$D_1 = 0$	$D_2 = 0$	$D_1 = 1$ $D_2 = 1$	$D_1 = 1$ $D_2 = 0$	$D_1 = 0$ $D_2 = 1$	$D_1 = 0$ $D_2 = 0$
Child									
Years of schooling	7.740	7.463	7.827	8.626	7.667	7.382	7.523	8.922	8.260
	(3.012)	(2.928)	(3.232)	(3.106)	(2.812)	(3.125)	(2.771)	(3.233)	(2.902)
Number of siblings	2.339	2.347	2.286	2.312	2.383	2.298	2.383	2.255	2.383
	(1.307)	(1.319)	(1.416)	(1.268)	(1.206)	(1.438)	(1.221)	(1.360)	(1.140)
Age	15.503	15.436	15.846	15.719	15.215	15.714	15.228	16.172	15.158
	(2.848)	(2.832)	(2.820)	(2.887)	(2.839)	(2.823)	(2.821)	(2.787)	(2.910)
Male	0.508	0.508	0.509	0.509	0.508	0.510	0.506	0.506	0.513
Mother									
Middle school	0.702	0.711	0.682	0.675	0.719	0.679	0.735	0.689	0.657
High School	0.125	0.102	0.102	0.197	0.144	0.078	0.121	0.163	0.241
College	0.045	0.039	0.044	0.063	0.045	0.037	0.040	0.060	0.068
Can read	0.780	0.745	0.729	0.892	0.824	0.679	0.795	0.851	0.944
Muslim	0.873	0.871	0.849	0.878	0.893	0.854	0.884	0.838	0.926
Mom's husband:									
- High school	0.133	0.113	0.121	0.195	0.142	0.105	0.119	0.159	0.239
- College	0.050	0.045	0.056	0.064	0.045	0.053	0.039	0.062	0.068
Community									
Urban area	0.486	0.380	0.523	0.825	0.455	0.388	0.374	0.856	0.787
Number of Schools:									
(1) Elementary	5.001	4.508	5.164	6.575	4.863	4.460	4.543	6.900	6.173
	(3.214)	(2.723)	(3.617)	(4.049)	(2.824)	(2.849)	(2.625)	(4.595)	(3.206)
(2) Junior	3.645	3.487	3.615	4.151	3.671	3.420	3.537	4.095	4.222
	(1.802)	(1.723)	(1.780)	(1.949)	(1.821)	(1.763)	(1.692)	(1.729)	(2.192)
(3) Senior	3.297	3.202	3.276	3.603	3.316	3.120	3.263	3.659	3.534
	(2.307)	(2.302)	(2.603)	(2.299)	(2.026)	(2.742)	(1.907)	(2.177)	(2.443)
Health facilities:									
Medical post	0.071	0.074	0.092	0.064	0.054	0.094	0.058	0.086	0.036
Delivery post	0.290	0.380	0.273	0	0.304	0.384	0.378	0	0
Midwife	0.647	0.850	0.525	0	0.750	0.738	0.933	0	0
Polindes	0.388	0.388	0.403	x	0.377	0.403	0.377	x	x
Number of									
(1) Family plan- -ning post	4.270	3.525	4.270	6.106	x	3.525	x	6.106	x
	(9.335)	(7.296)	(9.335)	(12.888)	x	(7.296)	x	(12.888)	x
(2) Posyandu	7.586	6.547	7.654	10.904	7.528	6.537	6.554	10.406	11.520
	(6.686)	(5.783)	(7.005)	(8.140)	(6.406)	(6.376)	(5.300)	(7.695)	(8.624)
Sample size	7370	5613	3368	1757	4002	2396	3217	972	785

Table 6: Sample mean and standard deviation (in parentheses for non-binary variables) of the variables used in the analysis. The same are also reported for various sub-samples based on the observability ($D_j = 0/1$ for $j = 1, 2$) of the instruments: Community has a polindes and Number of family planning posts in the community. The two binary variables: Community has a delivery post and Community has a midwife are 0 when polindes is missing in the sample unit.

Regressors	Full Sample	$D_1 = 1$	$D_2 = 1$	$D_1 = 0$	$D_2 = 0$	$D_1 = 1$ $D_2 = 1$	$D_1 = 1$ $D_2 = 0$	$D_1 = 0$ $D_2 = 1$	$D_1 = 0$ $D_2 = 0$
Child									
Number of	-0.116	-0.107	-0.099	-0.138	-0.158	-0.090	-0.148	-0.116	-0.219
siblings	(0.019)	(0.021)	(0.028)	(0.037)	(0.025)	(0.033)	(0.029)	(0.050)	(0.055)
Age = 12	0.952	0.911	0.898	0.978	0.991	0.888	0.952	0.853	1.032
	(0.108)	(0.124)	(0.191)	(0.212)	(0.126)	(0.225)	(0.144)	(0.349)	(0.249)
Age = 13	1.855	1.833	1.667	1.874	2.005	1.581	2.024	1.830	1.943
	(0.108)	(0.125)	(0.187)	(0.210)	(0.127)	(0.217)	(0.147)	(0.360)	(0.240)
Age = 14	2.701	2.677	2.478	2.680	2.850	2.421	2.842	2.567	2.796
	(0.107)	(0.122)	(0.184)	(0.215)	(0.125)	(0.214)	(0.142)	(0.347)	(0.258)
Age = 15	3.588	3.486	3.352	3.816	3.776	3.143	3.735	3.822	3.805
	(0.108)	(0.125)	(0.181)	(0.204)	(0.129)	(0.216)	(0.148)	(0.324)	(0.254)
Age = 16	4.254	4.120	4.118	4.606	4.373	3.963	4.245	4.464	4.739
	(0.106)	(0.123)	(0.178)	(0.201)	(0.128)	(0.210)	(0.148)	(0.321)	(0.247)
Age = 17	4.742	4.518	4.725	5.362	4.756	4.420	4.624	5.417	5.294
	(0.105)	(0.122)	(0.173)	(0.197)	(0.129)	(0.206)	(0.148)	(0.307)	(0.261)
Age = 18	5.467	5.185	5.474	6.203	5.479	5.192	5.210	6.116	6.295
	(0.106)	(0.124)	(0.174)	(0.197)	(0.132)	(0.207)	(0.153)	(0.313)	(0.245)
Age = 19	5.724	5.359	5.565	6.665	5.885	4.997	5.672	6.702	6.586
	(0.108)	(0.126)	(0.176)	(0.198)	(0.134)	(0.212)	(0.154)	(0.308)	(0.260)
Age = 20	6.302	5.940	6.500	7.231	6.075	5.952	5.939	7.551	6.485
	(0.109)	(0.128)	(0.176)	(0.201)	(0.137)	(0.213)	(0.157)	(0.309)	(0.271)
Male	-0.277	-0.305	-0.286	-0.216	-0.263	-0.299	-0.301	-0.256	-0.169
	(0.048)	(0.056)	(0.078)	(0.092)	(0.060)	(0.093)	(0.068)	(0.133)	(0.122)
Mother									
Middle school	0.459	0.422	0.433	0.541	0.417	0.361	0.424	0.619	0.173
	(0.096)	(0.104)	(0.138)	(0.244)	(0.134)	(0.153)	(0.143)	(0.320)	(0.397)
Junior school	0.595	0.696	0.662	-0.106	0.473	0.821	0.493	-0.435	0.098
	(0.176)	(0.198)	(0.234)	(0.399)	(0.354)	(0.275)	(0.434)	(0.519)	(0.662)
High school	1.046	1.132	1.331	0.135	0.818	1.358	1.065	0.265	-0.501
	(0.280)	(0.304)	(0.355)	(0.793)	(0.644)	(0.399)	(0.725)	(1.005)	(1.374)
Can read	1.331	1.309	1.257	1.195	1.293	1.258	1.243	1.088	1.183
	(0.079)	(0.086)	(0.120)	(0.197)	(0.104)	(0.136)	(0.112)	(0.261)	(0.316)
Muslim	-0.452	-0.404	-0.516	-0.516	-0.420	-0.570	-0.322	-0.408	-0.882
	(0.073)	(0.084)	(0.109)	(0.142)	(0.099)	(0.134)	(0.111)	(0.184)	(0.243)
Mom's Husband:									
- High school	0.749	0.751	0.859	1.303	0.708	0.896	0.821	1.772	0.465
	(0.148)	(0.164)	(0.180)	(0.336)	(0.331)	(0.201)	(0.415)	(0.448)	(0.516)
- College	0.770	0.817	0.900	1.451	0.592	0.947	0.595	1.913	0.973
	(0.254)	(0.272)	(0.302)	(0.757)	(0.633)	(0.327)	(0.717)	(0.957)	(1.310)
Community									
Urban area	0.509	0.310	0.688	0.749	0.404	0.423	0.260	0.969	0.669
	(0.054)	(0.063)	(0.088)	(0.130)	(0.068)	(0.105)	(0.080)	(0.211)	(0.163)
Number of	0.017	0.003	0.036	0.013	-0.017	0.025	-0.026	0.023	-0.012
Elementary Sch.	(0.008)	(0.011)	(0.012)	(0.012)	(0.012)	(0.018)	(0.015)	(0.015)	(0.021)
Number of	0.039	0.033	0.049	0.076	0.052	0.056	0.032	0.025	0.086
Junior Sch.	(0.017)	(0.020)	(0.027)	(0.036)	(0.024)	(0.032)	(0.027)	(0.052)	(0.057)
Number of	-0.002	0.019	-0.003	-0.070	0.002	0.018	0.022	-0.087	-0.004
Senior Sch.	(0.013)	(0.014)	(0.018)	(0.031)	(0.020)	(0.020)	(0.023)	(0.041)	(0.050)
Intercept	2.894	3.036	2.699	2.853	3.227	2.941	3.272	2.675	3.850

Table 7: OLS regression of Years of schooling of the child on number of siblings and other covariates. Estimate and s.e. (in parentheses). Since (almost) all regressors are highly significant we omit the usual "*" notation for significance levels to avoid notational clutter. Sample sizes as in Table 6.

Dep Var: Years of Schooling Regressors	GMM: no missing IV	Comp Case GMM	IPW-GMM	AIPW-GMM
Child				
Number of siblings	-0.731**	-1.007*	-0.709**	-0.665**
	(0.303/0.325)	(0.524/0.515)	(0.294/0.338)	(0.284/0.294)
Age = 12	0.987***	0.869***	0.982***	0.982***
	(0.081/0.083)	(0.204/0.188)	(0.081/0.082)	(0.079/0.079)
Age = 13	1.876***	1.559***	1.874***	1.873***
	(0.081/0.082)	(0.206/0.185)	(0.081/0.081)	(0.079/0.079)
Age = 14	2.804***	2.446***	2.801***	2.792***
	(0.098/0.101)	(0.206/0.190)	(0.097/0.101)	(0.095/0.095)
Age = 15	3.703***	3.213***	3.697***	3.691***
	(0.101/0.106)	(0.213/0.193)	(0.101/0.105)	(0.098/0.101)
Age = 16	4.304***	3.890***	4.302***	4.299***
	(0.090/0.093)	(0.202/0.192)	(0.091/0.092)	(0.089/0.089)
Age = 17	4.823***	4.331***	4.820***	4.814***
	(0.105/0.104)	(0.217/0.203)	(0.105/0.106)	(0.103/0.101)
Age = 18	5.500***	4.979***	5.503***	5.497***
	(0.109/0.110)	(0.262/0.253)	(0.109/0.108)	(0.108/0.106)
Age = 19	5.784***	4.836***	5.777***	5.778***
	(0.128/0.129)	(0.272/0.251)	(0.128/0.128)	(0.127/0.127)
Age = 20	6.386***	5.905***	6.377***	6.376***
	(0.136/0.132)	(0.278/0.268)	(0.135/0.130)	(0.134/0.129)
Male	-0.301***	-0.429***	-0.304***	-0.297***
	(0.054/0.053)	(0.131/0.122)	(0.053/0.053)	(0.053/0.051)
Mother				
Middle school	0.462***	0.320*	0.468***	0.461***
	(0.120/0.122)	(0.193/0.183)	(0.120/0.122)	(0.119/0.121)
Junior school	0.396*	0.575*	0.399**	0.418**
	(0.204/0.216)	(0.325/0.314)	(0.203/0.217)	(0.200/0.207)
High school	0.760***	0.923*	0.723**	0.781***
	(0.285/0.314)	(0.476/0.471)	(0.285/0.315)	(0.280/0.301)
Can read	1.206***	1.196***	1.206***	1.218***
	(0.111/0.114)	(0.167/0.155)	(0.110/0.116)	(0.109/0.111)
Muslim	-0.437***	-0.364*	-0.439***	-0.440***
	(0.081/0.082)	(0.202/0.196)	(0.081/0.082)	(0.081/0.081)
Mom's Husband: High School	0.727***	1.020***	0.729***	0.728***
	(0.130/0.133)	(0.210/0.201)	(0.130/0.131)	(0.128/0.131)
Mom's Husband: College	0.834***	1.052***	0.880***	0.835***
	(0.203/0.214)	(0.298/0.293)	(0.203/0.211)	(0.200/0.206)
Community				
Urban area	0.610***	0.799***	0.609***	0.599***
	(0.078/0.082)	(0.245/0.245)	(0.077/0.084)	(0.075/0.078)
Number of. Elementary Schools	0.011	0.020	0.010	0.011
	(0.009/0.009)	(0.021/0.020)	(0.009/0.009)	(0.009/0.009)
Number Junior Schools	0.051***	0.054	0.052***	0.049***
	(0.018/0.018)	(0.034/0.035)	(0.017/0.018)	(0.017/0.018)
Number of Senior Schools	0.002	0.012	0.002	0.002
	(0.012/0.013)	(0.017/0.023)	(0.012/0.013)	(0.012/0.13)
Intercept	4.333***	4.989***	4.278***	4.181***
	(0.717/0.768)	(1.171/1.153)	(0.698/0.802)	(0.675/0.700)
Specification tests				
p-value of Underidentification test	0	0.05	0	0
p-value of Overidentification test	0.14	0.44	0.01	0.3

Table 8: Column (2): efficient GMM with IVs Medical post, Delivery post, Midwife and Number of Posyandu. Column (3): efficient (complete case) GMM with same IVs and also Polindes and Number of Family planning posts. Columns (4) & (5): IPW-GMM and AIPW-GMM with all IVs. Standard errors (s.e.) are reported inside parentheses based on asymptotic variance/1000 bootstrap samples. "*", "**" and "***" respectively signify that a 0 coefficient against a two-sided alternative is rejected at the 10%, 5% and 1% levels using the asymptotic s.e. that are also used for the p-values.

References

Abbring, J. and Heckman, J. (2007). Econometric evaluation of social programs, part iii: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume VIB, chapter 72, pages 5145–5303. Elsevier Science Publisher.

Abrevaya, J. and Donald, S. G. (2011). A GMM approach for dealing with missing data on regressors and instruments. Mimeo.

Anderson, S. A. and Perlman, M. D. (1991). Lattice-ordered conditional independence models for missing data. *Statistics and Probability Letters*, 12: 465–486.

Andrews, D. W. K. (1994). Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity. *Econometrica*, 62: 43–72.

Angrist, J. and Krueger, A. B. (1992). The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples. *Journal of American Statistical Association*, 87: 328–336.

Angrist, J., Lavy, V., and Schlosser, A. (2006). Multiple Experiments for the Causal Link Between the Quantity and Quality of Children. Technical report, M.I.T.

Bang, H. and Robins, J. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61: 962–972.

Becker, G. S. (1960). An Economic Analysis of Fertility. In *Demographic Change in Developed Countries*, 11. Princeton University Press. Universities-National Bureau Conference Series.

Becker, G. S. and Lewis, H. G. (1973). On the Interaction Between Quantity and Quality of Children. *Journal of Political Economy*, 82: 279–288.

Becker, G. S. and Tomes, N. (1976). Child Endowments and the Quantity and Quality of Children. *Journal of Political Economy*, 84: 143–162.

Berry, D. J., Vimaleswaran, K. S., Whittaker, J. C., Hingorani, A. D., and Hyppnen, E. (2012). Evaluation of Genetic Markers as Instruments for Mendelian Randomization Studies on Vitamin D. *PLOS ONE*, 7(5): e37465.

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2005). The More the Merrier? The Effects of Family Size and Birth Order on Childrens Education. *The Quarterly Journal of Economics*, 120: 669–700.

Brown, B. and Newey, W. (1998). Efficient Semiparametric Estimation of Expectations. *Econometrica*, 66: 453–464.

Burgess, S., Seaman, S., Lawlor, D. A., Casas, J. P., and Thompson, S. G. (2011). Missing Data Methods in Mendelian Randomization Studies With Multiple Instruments. *American Journal of Epidemiology*, 174: 1069–1076.

Cao, W., Tsiatis, A., and Davidian, M. (2009). Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data. *Biometrika*, 96: 723–734.

Cattaneo, M. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155: 138–154.

Chaudhuri, S. (2014). A Note on Efficiency Gains from Multiple Incomplete Subsamples. Mimeo.

Chaudhuri, S. and Hill, J. B. (2013). Heavy Tail Robust Estimation and Inference for Average Treatment Effect. Technical report, University of North Carolina, Chapel Hill.

Chaudhuri, S. and Min, H. (2012). Doubly-Robust Parametric Estimation in Moment Conditions Models with Missing Data. Mimeo.

Chen, X., Hahn, J., and Liao, Z. (2012). Asymptotic Efficiency of Semiparametric Two-step GMM. Working Paper, UCLA.

Chen, X., Hong, H., and Tamer, E. (2005). Measurement Error Models with Auxiliary Data. *Review of Economic Studies*, 72: 343–366.

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

Davidson, R. and MacKinnon, J. G. (2004). *Econometric Theory and Methods*. Oxford University Press.

Ding, W., Lehrer, S. F., Rosenquist, J. N., and Audrain-McGovern, J. J. (2009). The impact of poor health on academic performance: New evidence using genetic markers. *Journal of Health Economics*, 28: 578–597.

Gertler, P. and Molyneaux, J. (1994). How Economic Development and Family Planning Programs Combined to Reduced Fertility in Indonesia. *Demography*, 31: 33–63.

Gill, R. and Robins, J. (1997). Non-Response Models For The Analysis Of Non-Monotone Ignorable Missing Data. *Statistics in Medicine*, 16: 39–56.

Gill, R. D., van der Laan, M. J., and Robins, J. M. (1997). Coarsening at Random: Characterizations, Conjectures and Counterexamples. In Lin, D. Y. and Fleming, T. R., editors, *Proceedings of The First Seattle Symposium in Biostatistics: Survival Analysis*, Lecture Notes in Statitsics, pages 255–294. New York: Springer-Verlag.

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

Graham, B. S., Pinto, C., and Egel, D. (2012). Inverse Probability Tilting for Moment Condition Models with Missing Data. *Review of Economic Studies*, 79: 1053 – 1079.

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

Heckman, J. and Vytlacil, E. (2007). Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume VIB, chapter 71, pages 4875–5144. Elsevier Science Publisher.

Heitjan, D. F. and Rubin, D. B. (1991). Igonaribility and Coarse Data. *Annals of Statistics*, 19: 2244–2253.

Hirano, K. and Imbens, G. (2001). Estimation of Causal Effects using Propensity Score Weighting : An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology*, 2: 259–278.

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

Horvitz, D. and Thompson, D. (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of American Statistical Association*, 47: 663–685.

Ichimura, I. and Martinez-Sanchis, E. (2005). Identification and Estimation of GMM Models by Combining Two Data Sets. Working Paper.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25: 51–71.

Imbens, G., Newey, W., and Ridder, G. (2009). Mean-squared-error Calculations for Average Treatment Effects. Working Paper.

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics*, 86: 4–29.

Juhn, C., Rubinstein, Y., and Zuppann, A. (2013). The Quantity-Quality Tradeoff and the Formation of Cognitive and Non-cognitive Skills. University of Houston.

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

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Smith, G. D. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27: 1133–11163.

Li, H., Zhang, J., and Zhu, Y. (1994). The Quantity-Quality Trade-Off of Children in a Developing Country: Identification Using Chinese Twins. *Demography*, 45: 223–243.

Malarani, V. (1994). The Changing Relationship between Family Size and Educational Attainment over the Course of Socioeconomic Development: Evidence from Indonesia. *Demography*, 45: 693–717.

Moffitt, R., Fitzgerald, J., and Gottschalk, P. (1998). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. In *Working Paper Series*. NBER.

Mogstad, M. and Wiswall, M. (2012). Instrumental variables estimation with partially missing instruments. *Economics Letters*, 114: 186–189.

Muris, C. (2014). Efficient GMM Estimation with a General Missing Data Pattern. Technical report, Simon Frasier University.

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

Newey, W. (1997). Convergence rates and asymptotic normality of series estimators. *Journal of Econometrics*, 79: 147–168.

Pakes, A. and Pollard, D. (1989). Simulation and the Asymptotics of Optimization Estimators. *Econometrica*, 57: 1027–1057.

Palmer, T. M., Lawlor, D. A., Harbord, R. M., Sheehan, N. A., Tobias, J. H., Timpson, N. J., Smith, G. D., and Sterne, J. A. (2012). Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research*, 21: 223–242.

Ridder, G. and Moffitt, R. (2007). The Econometrics of Data Combination. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6B, chapter 75, pages 5470–5547. Elsevier Science Publisher.

Robins, J. and Gill, R. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine*, 16: 39–56.

Robins, J. and Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of American Statistical Association*, 90: 122–129.

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

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

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

Rothe, C. and Firpo, S. (2012). Semiparametric Estimation and Inference Using doubly-Robust Moment Conditions. Mimeo.

Rotnitzky, A. and Robins, J. (1995). Semiparametric Regression Estimation in the Presence of Dependent Censoring. *Biometrika*, 82: 805–820.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94: 1096–1146.

Scholder, S. V. H. K., Smith, G. D., Lawlor, D. A., Propper, C., and Windmeijer, F. (2010). Genetic Markers as Instrumental Variables: An Application to Child Fat Mass and Academic Achievement. Technical report, University of York.

Scholder, S. V. H. K., Smith, G. D., Lawlor, D. A., Propper, C., and Windmeijer, F. (2011). Genetic Markers as Instrumental Variables. Technical report, University of Bristol.

Shiffman, J. (2004). Political Management of the Indonesian Family Planning Program. *International Family Planning Perspectives*, 30: 27–33.

Tan, Z. (2010). Bounded, Efficient and Doubly Robust Estimation with Inverse Weighting. *Biometrika*, 97: 661–682.

Tan, Z. (2011). Efficient Restricted Estimators for Conditional Mean Models with Missing Data. *Biometrika*, 98: 663–684.

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

Vansteelandt, S., Rotnitzky, A., and Robins, J. (2007). Estimation of regression models for mean of repeated outcomes under nonignorable nonmotonone nonresponse. *Biometrika*, 94: 841–860.

Wang, L. (2013). Estimating Returns to Education when the IV sample is Selective. *Labour Economics*, 21: 74–85.

Willis, R. J. (1973). A New Approach to the Economic Theory of Fertility Behavior. *Journal of Political Economy*, 81: 14–64.

Wooldridge, J. (2002). Inverse Probability Weighted M-Estimation for Sample Selection, Attrition, and Stratification. *Portuguese Economic Journal*, 1: 117–139.

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

A Appendix: Technical assumptions on the nuisance parameters

Notation to be used in the rest of the Appendix: For any $a \times b$ matrix A (including $b = 1$ or $a = b = 1$), let $|A| := \sqrt{\text{Trace}(A'A)}$. For any $a \times b$ matrix $A(u, \beta)$ where the (i, j)-th element is a function $A_{ij}(u, \beta) : \mathcal{U} \times \Theta \to \mathbb{R}$, let $\|A(\beta)\|_\infty = \sup_{u \in \mathcal{U}} |A(u, \beta)|$ for any given $\beta \in \mathcal{B}$, and let $\|A\|_\infty = \sup_{\beta \in \mathcal{B}} \sup_{u \in \mathcal{U}} |A(u, \beta)|$.

The following assumptions from Cattaneo (2010) are maintained in Propositions 2.3 and 2.4. See Cattaneo (2010) for discussion on each assumption. Define a generic neighborhood of $\mathcal{N}_\delta := \{\beta \in \mathcal{B} : |\beta - \beta^0| < \delta\}$ of β^0 for some constant $\delta > 0$. $\mathcal{N}_\delta$ is referred to in Assumptions g(2)-g(4), q(2) and T(2).

Assumption g: The moment function in both the general case and the P-case satisfies:

(1) $\{g(.; \beta) : \beta \in \mathcal{B}\}$ is Glivenko-Cantelli.

(2) $\{g(.; \beta) : \beta \in \mathcal{N}_\delta\}$ is Donsker.

(3) $E[\sup_{\beta, \tilde{\beta} \in \mathcal{N}_\delta} |g(Z, W; \beta) - g(Z, W; \tilde{\beta})|^2] \leq C\delta^{2r}$ for some constants $C > 0$ and $r \in (0, 1)$.

(4) $E[\sup_{\beta \in \mathcal{N}_\delta} |g(Z, W; \beta)|^2] < \infty$.

Assumption q: The conditional expectation $q(U; \beta) = E[g(Z, W; \beta)|U]$ of the moment function in both the general case and the P-case satisfies the following. U can be $(Z_1', W)'$, $(Z_2', W)'$ and W in the general case while $U = W$ in P-case.

(1) $\{q(u; \beta) : \beta \in \mathcal{B}\}$ is Glivenko-Cantelli.

(2) $q(u; \beta)$ belongs to a class of functions $\mathcal{Q}(\beta) = \{\tilde{q}(w; \beta) : \beta \in \mathcal{B}\}$ that satisfies the following local (for $\beta \in \mathcal{N}_\delta$) restrictions for all $u \in \mathcal{U} := \text{Support}(U)$:

   (a) $\tilde{q}(u; \beta)$ is continuously differentiable in $\beta \in \mathcal{N}_\delta$, and $E\left[\sup_{\beta \in \mathcal{N}_\delta} \left|\frac{\partial}{\partial \beta'} \tilde{q}(U; \beta)\right|^2\right] < \infty$.

   (b) For all $\tilde{q}(u; \beta) \in \mathcal{Q}(\beta)$ satisfying $\sup_{\beta \in \mathcal{N}_\delta} \|q(u, \beta) - \tilde{q}(u; \beta)\|_\infty < \delta$ there exists $\epsilon > 0$ and a non-negative measurable function $b(U)$ with $E[b(U)] < \infty$ such that $\left|\frac{\partial}{\partial \beta'} q(u; \beta) - \frac{\partial}{\partial \beta'} \tilde{q}(u; \beta)\right| \leq b(u) \sup_{\beta \in \mathcal{N}_\delta} \|q(u, \beta) - \tilde{q}(u; \beta)\|_\infty^\epsilon$.

Assumption T: Other technical assumptions.

(1) U is continuous with density bounded and bounded away from 0 on its compact support $\mathcal{U}$.

(2) $Var(g(Z, W; \beta)|U = u)$ is uniformly bounded in $u \in \mathcal{U}$ for all $\beta \in \mathcal{N}_\delta$.

Remark: Since U can include Z_1 or Z_2 in the general case, assumption T(1) requiring compact $\mathcal{U}$ is uninteresting. The same is also true when $U = W$. Chen et al. (2008) relax this assumption.

Supplemental Appendix: GMM with Multiple Missing Variables

Appendix B gives further details on the series estimators used in the paper. Appendix C collects proofs of all the theoretical results and claims. The notations are the same as in the original paper, and the equations are also numbered accordingly in the same order.

B Appendix: Series estimation of the nuisance parameters

For a variable U, consider $P_K(U) = (P_K^1(U), \ldots, P_K^K(U))'$ that is a K-truncation of some approximating series such that the minimum eigenvalue of $E[P_K(U)P_K(U)']$ is bounded away from 0 uniformly in K. Let $R_K(U) := (E[P_K(U)P_K(U)'])^{-1/2}P_K(U)$ be a standardization for $P_K(U)$. See Newey (1997) for details. For estimation of $p(W)$ and $q(W;\beta)$ we take $U = W$, while for $q(Z_1, W;\beta)$ and $q(Z_2, W;\beta)$ we take $U = (Z_1', W')'$ and $U = (Z_2', W')'$ respectively.

$\widehat{p}(W) = (\widehat{p}_{00}(W), \widehat{p}_{10}(W), \widehat{p}_{01}(W), \widehat{p}_{11}(W))$ is estimated by multinomial series logit as:

$$\widehat{p}_{d_1d_2}(w) := L_{d_1d_2}(R_K(w), \widehat{\pi}_K) \text{ where } L_{d_1d_2}(R_K(w), \pi_K) = \frac{\exp[R_K(w)'\pi_{(d_1d_2)K}]}{\sum_{j,l=0,1}\exp[R_K(w)'\pi_{(jl)K}]}$$

for $d_1, d_2 = 0, 1$. $\widehat{p}_1(w) = \widehat{p}_{11}(w) + \widehat{p}_{10}(w)$ and $\widehat{p}_2(w) = \widehat{p}_{11}(w) + \widehat{p}_{01}(w)$. See Hirano et al. (2003) and Cattaneo (2010). For $\pi_K = [\pi'_{(00)K}, \pi'_{(10)K}, \pi'_{(01)K}, \pi'_{(11)K}]'$ the estimates are:

$$\widehat{\pi}_K := \arg\max_{\pi_K \in \mathbb{R}^{4d_K} | \pi_{(00)K}=0} \sum_{i=1}^{N} \sum_{d_1,d_2=0,1} 1(D_{1i} = d_1, D_{2i} = d_2) \log L_{d_1d_2}(R_K(W_i), \pi_K).$$

Series estimators for the elements of $q(\beta)$ are obtained following Newey (1997) and Cattaneo (2010):

$$\begin{aligned} \widehat{q}(w;\beta) &= (I_{d_g} \otimes R_K(w)')\widehat{\gamma}_K(\beta), \\ \widehat{q}(z_j, w;\beta) &= (I_{d_g} \otimes R_K(u_j)')\widehat{\gamma}_K^{(j)}(\beta) \text{ where } U_j = (Z_j', W')' \text{ for } j = 1, 2. \end{aligned}$$

For $\gamma_K(\beta) = [\gamma_{K1}(\beta)', \ldots, \gamma_{Kd_g}(\beta)']'$ and $\gamma_K^{(j)}(\beta) = [\gamma_{K1}^{(j)}(\beta)', \ldots, \gamma_{Kd_g}^{(j)}(\beta)']'$ the estimates are:

$$\widehat{\gamma}_K(\beta) := \arg\min_{\gamma \in \mathbb{R}^{Kd_g}} \frac{1}{N}\sum_{i=1}^{N} D_{1i}D_{2i}\left(g(Z_i, W_i;\beta) - (I_{d_g} \otimes R_K(W_i)')\gamma\right)'\left(g(Z_i, W_i;\beta) - (I_{d_g} \otimes R_K(W_i)'\gamma)\right),$$

$$\widehat{\gamma}_K^{(j)}(\beta) := \arg\min_{\gamma \in \mathbb{R}^{Kd_g}} \frac{1}{N}\sum_{i=1}^{N} D_{1i}D_{2i}\left(g(Z_i, W_i;\beta) - (I_{d_g} \otimes R_K(U_{ji})')\gamma\right)'\left(g(Z_i, W_i;\beta) - (I_{d_g} \otimes R_K(U_{ji})'\gamma)\right),$$

for $j = 1, 2$. Dividing $D_{1i}D_{2i}$ by $\widehat{p}_{11}(W_i)$ in the definitions for $\widehat{\gamma}_K(\beta)$ and $\widehat{\gamma}_K^{(j)}(\beta)$ often leads to improvement in the finite-sample properties of the AIPW estimators [see Hirano and Imbens (2001)]. Similarly, instead of using $\sum_{i=1}^N D_{1i}D_{2i}$ observations for all, improvement is possible in the P-case by taking advantage of the partition in (5) to estimate the elements of $\widehat{\gamma}_K(\beta)$ more precisely by using $\sum_{i=1}^N D_{1i}$ observations for the first Kd_{g_1} elements and $\sum_{i=1}^N D_{2i}$ observations for the last Kd_{g_2} elements. We follow this in Sections 4 and 5.

C Appendix: Proof of main results

Proof of Proposition 2.1:

Let f and F denote the density and distribution functions, with the concerned random variables specified inside parentheses. $L_0^2(F)$ denotes the space of mean-zero, square integrable functions with respect to F. The proof consists of three standard steps. (1) Obtain the tangent set for all regular parametric submodels satisfying the semiparametric assumptions on the observed data. (2) Conjecture the efficient influence function and then show pathwise differentiability of β^0 and verify that the efficient influence function lies in the tangent set. (3) Obtain the efficiency bound as the expectation of the outer product of the efficient influence function.

STEP - 1: Consider a regular parametric sub-model indexed by a finite-dimensional parameter θ for the joint distribution of the observed data $\mathcal{O} := (D_1, D_2, D_1Z_1, D_2Z_2, W)$. So the joint density $f_\theta(\mathcal{O})$ of the observed data can be expressed in terms of the full data (D_1, D_2, Z_1, Z_2, W) as

$$[p_{\theta,11}(W)f_\theta(Z_1, Z_2|W)]^{D_1D_2}\,[p_{\theta,10}(W)f_\theta(Z_1|W)]^{D_1(1-D_2)}\,[p_{\theta,01}(W)f_\theta(Z_2|W)]^{(1-D_1)D_2}\,[p_{\theta,00}(W)]^{(1-D_1)(1-D_2)}\,f_\theta(W)$$

where (4) gives the factorization of the first three terms. The score with respect to θ is

$$S_\theta(\mathcal{O}) = D_1D_2 s_\theta(Z_1, Z_2|W) + \sum_{j\neq k=1,2} D_j(1-D_k)s_\theta(Z_j|W) + s_\theta(W) + \sum_{d_1,d_2=0,1} \mathbf{1}(D_1 = d_1, D_2 = d_2)\frac{\dot{p}_{\theta,d_1d_2}(W)}{p_{\theta,d_1d_2}(W)}.$$

where $s_\theta(Z_1, Z_2|W) := \frac{\partial}{\partial\theta}\log f_\theta(Z_1, Z_2|W)$, $s_\theta(Z_1|W) := \frac{\partial}{\partial\theta}\log f_\theta(Z_1|W)$, $s_\theta(Z_2|W) := \frac{\partial}{\partial\theta}\log f_\theta(Z_2|W)$, $s_\theta(W) := \frac{\partial}{\partial\theta}\log f_\theta(W)$, and $\dot{p}_{\theta,d_1d_2}(W) := \frac{\partial}{\partial\theta}p_{\theta,d_1d_2}(W)$ for $d_1, d_2 = 0, 1$. Henceforth, we omit the subscript θ from quantities evaluated at $\theta = \theta_0$.

The tangent set for the model is characterized by functions of the form:

$$\mathcal{T} := D_1D_2 a(Z_1, Z_2, W) + \sum_{j\neq k=1,2} D_j(1-D_k)a_j(Z_j, W) + a_0(W) + \sum_{d_1,d_2=0,1} \mathbf{1}(D_1 = d_1, D_2 = d_2)\frac{b_{d_1d_2}(W)}{c_{d_1d_2}(W)} \quad (25)$$

where $a(Z_1, Z_2, W) \in L_0^2(F(Z_1, Z_2|W))$, $a_1(Z_1, W) \in L_0^2(F(Z_1|W))$, $a_2(Z_2, W) \in L_0^2(F(Z_2|W))$, $a_0(W) \in L_0^2(F(W))$ and $\sum_{d_1,d_2=0,1} \mathbf{1}(D_1 = d_1, D_2 = d_2)\frac{b_{d_1d_2}(W)}{c_{d_1d_2}(W)} \in L_0^2(F(D_1, D_2|W))$ with the additional restriction that $\sum_{d_1,d_2=0,1} b_{d_1d_2}(W) = 0$ and $\sum_{d_1,d_2=0,1} c_{d_1d_2}(W) = 1$ for all W.

STEP - 2: The moment conditions in (1) are equivalent to the requirement that for any $d_\beta \times d_g$ matrix A, the just-identified system of moment conditions $AE[g(Z, W; \beta^0)] = 0$ hold. Differentiating under the integral, and taking a full row rank A, we obtain by using (4) that

$$\frac{\partial\beta^0(\theta_0)}{\partial\theta'} = -(AG)^{-1}AE\left[g(Z, W; \beta^0)\frac{\partial\log f_{\theta_0}(Z, W)}{\partial\theta'}\right] = -(AG)^{-1}AE\left[g(Z, W; \beta^0)\left\{s(W)' + s(Z_1, Z_2|W)'\right\}\right].$$

For an arbitrary A, pathwise differentiability follows if we can find $\psi(A, D_1, D_2, Z_1, Z_2, W; \beta^0) \in \mathcal{T}$ such that

$$E[\psi(A, D_1, D_2, Z_1, Z_2, W; \beta^0)S(\mathcal{O})'] = \frac{\partial\beta^0(\theta_0)}{\partial\theta'}. \quad (26)$$

Conjecture: $\psi(A, D_1, D_2, Z_1, Z_2, W; \beta^0) = -(AG)^{-1} A\varphi(D_1, D_2, Z_1, Z_2, W; \beta^0)$. Then verify (26) by showing

$$E[\varphi(D_1, D_2, Z_1, Z_2, W) S(\mathcal{O})'] = E\left[g(Z, W; \beta^0) \{s(W)' + s(Z_1, Z_2|W)'\}\right]. \tag{27}$$

We proceed term-by-term for the four terms in $\varphi(D_1, D_2, Z_1, Z_2, W; \beta^0)$. Dependence on β^0 is suppressed. Consider the first term. Taking expectation conditional on W and then using (4) we obtain:

$$E\left[\frac{D_1 D_2}{p_{11}(W)} (g(Z, W) - q(W)) S(\mathcal{O})'\right] = E\left[(g(Z, W) - q(W)) s(Z_1, Z_2|W)'\right] = E\left[g(Z, W) s(Z_1, Z_2|W)'\right]$$

since $E[q(W)s(Z_1, Z_2|W)'] = 0$ by using $s(Z_1, Z_2|W) \in L_0^2(F(Z_1, Z_2|W))$. Now consider the second term. Taking expectation conditional on W and then using (4) we obtain: $E\left[q(W)S(\mathcal{O})'\right] = E\left[g(Z, W)s(W)'\right]$ since $s(Z_1, Z_2|W) \in L_0^2(F(Z_1, Z_2|W))$, $s(Z_1|W) \in L_0^2(F(Z_1|W))$, $s(Z_2|W) \in L_0^2(F(Z_2|W))$, and $\sum_{d_1, d_2 = 0,1} \dot{p}_{d_1 d_2}(W) = 0$. Now consider the third term and note that similar arguments give

$$\begin{aligned}
& E\left[\frac{p_{10}(W)}{p_1(W)} \left(\frac{D_1(1 - D_2)}{p_{10}(W)} - \frac{D_1 D_2}{p_{11}(W)}\right) (q(Z_1, W) - q(W)) S(\mathcal{O})'\right] \\
= \; & E\left[\frac{p_{10}(W)}{p_1(W)} \left(\frac{D_1(1 - D_2)}{p_{10}(W)} - \frac{D_1 D_2}{p_{11}(W)}\right) (q(Z_1, W) - q(W)) \{D_1 D_2 s(Z_1, Z_2|W) + D_1(1 - D_2) s(Z_1|W)\}'\right] \\
= \; & E\left[\frac{p_{10}(W)}{p_1(W)} (q(Z_1, W) - q(W)) \{s(Z_1|W) - s(Z_1, Z_2|W)\}'\right] \\
= \; & -E\left[\frac{p_{10}(W)}{p_1(W)} (q(Z_1, W) - q(W)) s(Z_2|Z_1, W)'\right] \; [\text{since } s(Z_1, Z_2|W) \stackrel{\text{def}}{=} s(Z_1|W) + s(Z_2|Z_1, W) \,] \\
= \; & 0
\end{aligned}$$

where the last line follows because, by definition of conditional score, $s(Z_2|Z_1, W) \in L_0^2(F(Z_2|Z_1, W))$. Similar arguments show that for the fourth term, $E\left[\frac{p_{01}(W)}{p_2(W)} \left(\frac{(1-D_1)D_2}{p_{01}(W)} - \frac{D_1 D_2}{p_{11}(W)}\right) (q(Z_2, W) - q(W)) S(\mathcal{O})'\right] = 0$.

Hence (27) (and thereby (26)) is verified. To show that $\varphi(D_1, D_2, Z_1, Z_2, W; \beta^0)$ belongs to the tangent set $\mathcal{T}$ in (25), rearrange its terms suitably as follows:

$$\begin{aligned}
\varphi(D_1, D_2, Z_1, Z_2, W; \beta^0) \; = \; & \frac{D_1 D_2}{p_{11}(W)} \left[(g(Z, W) - q(W)) - \frac{p_{10}(W)}{p_1(W)} (q(Z_1, W) - q(W)) - \frac{p_{01}(W)}{p_2(W)} (q(Z_2, W) - q(W))\right] \\
& + \frac{D_1(1 - D_2)}{p_1(W)} (q(Z_1, W) - q(W)) + \frac{(1 - D_1) D_2}{p_2(W)} (q(Z_2, W) - q(W)) + q(W).
\end{aligned}$$

The first term of the RHS involves $D_1 D_2, Z_1, Z_2, W$ and belongs to $L_0^2(F(Z_1, Z_2|W))$ by (4). It corresponds to $D_1 D_2 a(Z_1, Z_2, W)$ of $\mathcal{T}$ in (25). The second term of the RHS $D_1(1 - D_2), Z_1, W$ and belongs to $L_0^2(F(Z_1|W))$ by (4). It corresponds to $D_1(1 - D_2) a_1(Z_1, W)$ of $\mathcal{T}$ in (25). Similarly, the third term corresponds to $(1 - D_1) D_2 a_2(Z_2, W)$ of $\mathcal{T}$ in (25). The fourth term involves W and belongs to $L_0^2(F(W))$. It corresponds $a_0(W)$ of $\mathcal{T}$ in (25). Remaining terms of $\mathcal{T}$ are corresponded identically by 0s.

STEP - 3: We verified that any regular estimator for β^0 is asymptotically linear with influence function of the form $-(AG)^{-1} A g(Z, W; \beta^0)$. For a given A the projection of the above influence function on to the tangent set $\mathcal{T}$ is $\psi(A, D_1, D_2, Z_1, Z_2, W; \beta^0)$ which, therefore, is the efficient influence function given the A. The

variance of $\psi(A, D_1, D_2, Z_1, Z_2, W; \beta^0)$ is $(AG)^{-1}A\ V\ A'(AG)^{-1'}$ where $V := Var\left(\varphi(D_1, D_2, Z_1, Z_2, W; \beta^0)\right)$. The efficient influence function involves the A that minimizes the above variance. Standard arguments give that the minimizer is $A_* = G'V^{-1}$. Hence the efficiency bound is $\Omega := (G'V^{-1}G)^{-1}$ and the efficient influence function with variance equal to the efficiency bound is

$$\psi(D_1, D_2, Z_1, Z_2, W) := \psi(A_*, D_1, D_2, Z_1, Z_2, W) = -\Omega^{-1}G'V^{-1}\varphi(D_1, D_2, Z_1, Z_2, W; \beta^0). \blacksquare$$

Remark: From the verification of (27) in Step 2 involving the first two terms of $\varphi(D_1, D_2, Z_1, Z_2, W)$, it follows naturally that the conventional form [see Chen et al. (2008)] based on the common complete subsample ($D_1 = D_2 = 1$):

$$\frac{D_1 D_2}{p_{11}(W)}\left[g(Z, W; \beta^0) - q(W; \beta^0)\right] + q(W; \beta^0)$$

is an influence function. However, in general it does not belong in $\mathcal{T}$ defined in (25) because that requires the parametric submodel to satisfy $s(Z_1|W) \equiv s(Z_2|W) \equiv 0$ but $s(Z_1, Z_2|W) \neq 0$. This is not possible except in the special case $s(Z_1, Z_2|W) \equiv s(Z_1|Z_2, W) \equiv s(Z_2|Z_1, W)$ which imposes the additional restrictions on $\mathcal{T}$ that $a(Z_1, Z_2, W) \in L_0^2(F(Z_1|Z_2, W))$ and $a(Z_1, Z_2, W) \in L_0^2(F(Z_2|Z_1, W))$.

Proof of Proposition 2.2:

STEP - 1: Same as that in the proof of Proposition 2.1 with W denoting the distinct collection of all elements of W_1 and W_2, and allowing for the possibility that $W = W_1$ and/or $W = W_2$.

STEP - 2: The moment conditions in (1) under (5) are equivalent to the requirement that for any $d_\beta \times d_g$ matrix $A = [A_1, A_2]$ where A_j is $d_\beta \times d_{g_j}$ for $j = 1, 2$, the following just-identified system of moment conditions

$$AE[g(Z, W; \beta^0)] \equiv [A_1, A_2]E\begin{bmatrix} g_1(Z_1, W_1; \beta_0^0, \beta_1^0) \\ g_2(Z_2, W_2; \beta_0^0, \beta_2^0) \end{bmatrix} = 0$$

hold. Differentiating under the integral, and taking a full row rank A, we obtain by using (4) that

$$\frac{\partial \beta^0(\theta_0)}{\partial \theta'} = -(A\widetilde{G})^{-1}AE\begin{bmatrix} g_1(Z_1, W_1; \beta_0^0, \beta_1^0)\left\{s(W_1)' + s(Z_1|W_1)'\right\} \\ g_2(Z_2, W_2; \beta_0^0, \beta_2^0)\left\{s(W_2)' + s(Z_2|W_2)'\right\} \end{bmatrix}.$$

Now as in STEP-2 in the proof of Proposition 2.1 we show that, for $j = 1, 2$:

$$E\left[\varphi_{Pj}(D_j, Z_j, W; \beta_0^0, \beta_j^0)S(\mathcal{O})'\right] = E\left[g_j(Z_j, W_j; \beta_0^0, \beta_j^0)\left\{s(W)' + s(Z_j|W)'\right\}\right]. \tag{28}$$

For $j = 1$, taking expectation conditional on W and then noting that $D_1D_2s(Z_1, Z_2|W) + D_1(1-D_2)s(Z_1|W) = D_1s(Z_1|W) + D_1D_2s(Z_2|Z_1, W)$ in $S(D_1, D_2, D_1Z_1, D_2Z_2, W)$, it follows that

$$E\left[\frac{D_1}{p_1(W)}\left[g_1(Z_1, W_1) - q_1(W)\right]S(\mathcal{O})'\right] = E\left[\left[g_1(Z_1, W_1) - q_1(W)\right]s(Z_1|W)'\right] = E\left[g_1(Z_1, W_1)s(Z_1|W)'\right]$$

where the first equality follows from $s(Z_2|Z_1,W) \in L_0^2(F(Z_2|Z_1,W))$, and the second from $s(Z_1|W) \in L_0^2(F(Z_1|W))$. On the other hand, $E\left[q_j(W;\beta_0,\beta_j)S(\mathcal{O})'\right] = E\left[g_1(Z_1,W_1)s(W)'\right]$ by using $s(Z_1,Z_2|W) \in L_0^2(F(Z_1,Z_2|W))$, $s(Z_1|W) \in L_0^2(F(Z_1|W))$, $s(Z_2|W) \in L_0^2(F(Z_2|W))$ and $\sum_{d_1 d_2} \dot{p}_{d_1 d_2}(W) = 0$. Therefore, (28) is verified for $j = 1$. Similarly, it can be verified for $j = 2$.

Now rearranging the terms in $\varphi_P(D_1,D_2,Z_1,Z_2,W;\beta^0)$ as follows:

$$\varphi_P(D_1,D_2,Z_1,Z_2,W) = \begin{bmatrix} \varphi_{P1}(D_1,Z_1,W_1) \\ \varphi_{P2}(D_2,Z_2,W) \end{bmatrix} = D_1D_2 \begin{bmatrix} \frac{g_1(Z_1,W_1)-q_1(W)}{p_1(W)} \\ \frac{g_2(Z_2,W_2)-q_2(W)}{p_2(W)} \end{bmatrix} + D_1(1-D_2) \begin{bmatrix} \frac{g_1(Z_1,W_1)-q_1(W)}{p_1(W)} \\ 0 \end{bmatrix}$$
$$+(1-D_1)D_2 \begin{bmatrix} 0 \\ \frac{g_2(Z_2,W_2)-q_2(W)}{p_2(W)} \end{bmatrix} + \begin{bmatrix} q_1(W) \\ q_2(W) \end{bmatrix},$$

it is easy to see that when evaluated at $\beta = \beta^0$, the four terms on the RHS, by virtue of (4), are respectively in $L_0^2(F(Z_1,Z_2|W))$, $L_0^2(F(Z_1|W))$, $L_0^2(F(Z_2|W))$ and $L_0^2(F(W))$. The terms $D_1D_2a(Z_1,Z_2,W)$, $D_1(1-D_2)a_1(Z_1,W)$, $(1-D_1)D_2a_2(Z_2,W)$ and $a_0(W)$ of $\mathcal{T}$ in (25) are corresponded by these four terms, whereas the remaining terms in $\mathcal{T}$ are corresponded identically by 0s. This completes step 2.

STEP - 3: Follows similarly as in the proof of Proposition 2.1. ■

Notations to be used in the rest of the Appendix: For any $a \times b$ matrix A (including $b = 1$ or $a = b = 1$), let $|A| := \sqrt{\text{Trace}(A'A)}$. For any $a \times b$ matrix $A(u,\beta)$ where the (i,j)-th element is a function $A_{ij}(u,\beta) : \mathcal{U} \times \Theta \to \mathbb{R}$, let $\|A(\beta)\|_\infty = \sup_{u\in\mathcal{U}} |A(u,\beta)|$ for any given $\beta \in \mathcal{B}$, and let $\|A\|_\infty = \sup_{\beta\in\mathcal{B}} \sup_{u\in\mathcal{U}} |A(u,\beta)|$.

Proof of Proposition-2.3:

Define the following quantities that will be used throughout the proof for notational convenience:

$$\begin{aligned}
\omega_i &:= \frac{p_{10}(W_i)}{p_1(W_i)}, \text{ and } \widehat{\omega}_i := \frac{\widehat{p}_{10}(W_i)}{\widehat{p}_1(W_i)}, \\
\nu_i &:= \frac{D_{1i}(1-D_{2i})}{p_{10}(W_i)} - \frac{D_{1i}D_{2i}}{p_{11}(W_i)}, \text{ and } \widehat{\nu}_i := \frac{D_{1i}(1-D_{2i})}{\widehat{p}_{10}(W_i)} - \frac{D_{1i}D_{2i}}{\widehat{p}_{11}(W_i)}, \\
\tau_i(\beta) &:= q(Z_{1i},W_i;\beta) - q(W_i;\beta) \text{ and } \widehat{\tau}_i(\beta) := \widehat{q}(Z_{1i},W_i;\beta) - \widehat{q}(W_i;\beta).
\end{aligned}$$

Under the conditions of the proposition,

$$\|\widehat{p} - p\|_\infty = o_p(N^{-1/4}), \tag{29}$$
$$\sup_{|\beta-\beta^0|<\delta} \|\widehat{q}(\beta) - q(\beta)\|_\infty = o_p(1) \text{ for some constant } \delta > 0 \text{ and,} \tag{30}$$
$$\|\widehat{q}(U;\beta^0) - q(U;\beta^0)\|_\infty = o_p(N^{-1/4}). \tag{31}$$

(29) is Cattaneo's condition (5.1), and holds by Theorem B-1 of Cattaneo (2010). (30) is Cattaneo's condition (5.2), and holds from the first result of Proposition A1(i) of Chen et al. (2005). (31) is shown in the proof of Theorem 8 (page 152) in Cattaneo (2010) [Theorem 4 of Newey (1997)].

Therefore, all that are left to be verified are two conditions: a condition similar to (5.3) of Cattaneo (2010) and a stochastic equicontinuity condition that, by virtue of the previous condition, gives (iii) in Theorem 3.3 in Pakes and Pollard (1989). Thanks to (a) symmetry in the terms in the second and third lines of (6), and (b) the proofs of Theorem 5 and 8 in Cattaneo (2010) this boils down to verifying:

$$o_p(1) = \sqrt{N}\left(\bar{\xi}_N(\beta^0, \widehat{p}, \widehat{q}(\beta^0)) - \bar{\xi}_N(\beta^0, p, \bar{q}(\beta^0))\right), \tag{32}$$

$$o_p(1) = \sup_{|\beta-\beta^0|\leq\delta_N} \frac{\sqrt{N}\left|\bar{\xi}_N(\beta, \widehat{p}, \widehat{q}(\beta)) - E\left[\bar{\xi}_N(\beta, p, \bar{q}(\beta))\right] - \bar{\xi}_N(\beta^0, \widehat{p}, \widehat{q}(\beta^0))\right|}{1 + C\sqrt{N}|\beta - \beta^0|} \tag{33}$$

for all positive sequences $\delta_N = o(1)$ and a generic constant $C > 0$, where in terms of the notation above,

$$\bar{\xi}_N(\beta, \widehat{p}, \widehat{q}(\beta)) := \frac{1}{N}\sum_{i=1}^N \widehat{\omega}_i\widehat{\nu}_i\widehat{\tau}_i(\beta) \text{ and } \bar{\xi}_N(\beta, p, \bar{q}(\beta)) := \frac{1}{N}\sum_{i=1}^N \omega_i\nu_i\tau_i(\beta).$$

We start with the verification of (32). (Dependence on β is suppressed at $\beta = \beta^0$.) Note that (32)'s RHS is:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N(\widehat{\omega}_i - \omega_i)(\widehat{\nu}_i - \nu_i)(\widehat{\tau}_i - \tau_i) + \frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i(\widehat{\nu}_i - \nu_i)(\widehat{\tau}_i - \tau_i) + \frac{1}{\sqrt{N}}\sum_{i=1}^N(\widehat{\omega}_i - \omega_i)\nu_i(\widehat{\tau}_i - \tau_i)$$
$$+\frac{1}{\sqrt{N}}\sum_{i=1}^N(\widehat{\omega}_i - \omega_i)(\widehat{\nu}_i - \nu_i)\tau_i + \frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i\nu_i(\widehat{\tau}_i - \tau_i) + \frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i(\widehat{\nu}_i - \nu_i)\tau_i + \frac{1}{\sqrt{N}}\sum_{i=1}^N(\widehat{\omega}_i - \omega_i)\nu_i\tau_i.$$

Thanks to (29) and (31), and the fact that $\frac{1}{N}\sum_i|\omega_i|$, $\frac{1}{N}\sum_i|\nu_i|$ and $\frac{1}{N}\sum_i|\tau_i|$ are $O_p(1)$ under our assumptions, it is straightforward to show that the first four terms in the above expression are $o_p(1)$. We focus on the last three terms and show that each of them is $o_p(1)$. Since the convergence rates are faster for series estimators based on splines than on power series, it suffices to show the desired results for the power series case, i.e., with $\eta = 1$ in the statement of the proposition. We use Lemma A (below) with $s_p = s_q = s$ and $K_p = K_q = K$.

Let us start with the fifth term:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i\nu_i(\widehat{\tau}_i - \tau_i) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i\nu_i(\widehat{q}(Z_{1i}, W_i) - q(Z_{1i}, W_i)) - \frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i\nu_i(\widehat{q}(W_i) - q(W_i)).$$

We use Lemma A to show the second term on the RHS is $o_p(1)$. Modifying (B4)-(B6) in Lemma A accordingly, and then following the same steps as below will give the first term on the RHS is $o_p(1)$. Hence this is omitted for brevity. Note that $\frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i\nu_i(\widehat{q}(Z_{1i}, W_i) - q(Z_{1i}, W_i)) = T_{5aN} + T_{5bN}$ where

$$T_{5aN} := \frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i\nu_i(\widehat{q}(Z_{1i}, W_i) - (I_{d_g} \otimes R_K'(W_i))\gamma_K^*) \text{ and } T_{5bN} := \frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i\nu_i((I_{d_g} \otimes R_K'(W_i))\gamma_K^* - q(W_i)).$$

Consider T_{5aN} and note that $|T_{5aN}| \leq \left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i\nu_i(I_{d_g} \otimes R_K'(W_i))\right| |\widehat{\gamma}_K - \gamma_K^*| = O_p\left(KN^{-1/2} + K^{1/2-s/d_w}\right)$, because under our assumptions, $E[\omega_i\nu_i(I_{d_g} \otimes R_K'(W_i))] = 0$, $Var(\omega_i\nu_i(I_{d_g} \otimes R_K'(W_i))) = O(K)$ and hence $\frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i\nu_i(I_{d_g} \otimes R_K'(W_i)) = O_p(K^{1/2})$; while Lemma A (B4) gives $|\widehat{\gamma}_K - \gamma_K^*| = O_p\left(K^{1/2}N^{-1/2} + K^{-s/d_w}\right)$. Since taking $\eta = 1$ means $v < 1/6$ and $s/d_w > 3$ under our assumptions, noting $K = N^v$ gives $|T_{5aN}| = o_p(1)$.

Our assumptions give $E[\omega_i\nu_i((I_{d_g}\otimes R_K'(W_i))\gamma_K^* - q(Z_{1i},W_i))] = 0$ and $Var(\omega_i\nu_i((I_{d_g}\otimes R_K'(W_i))\gamma_K^* - q(Z_{1i},W_i)) = O\left(\sup_w \|((I_{d_g}\otimes R_K'(w))\gamma_K^* - q(w))\|^2\right)$. So $|T_{5bN}| = O_p(K^{-s/d_w}) = o_p(1)$ by Lemma A (B4).

Now consider the sixth term: $\frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i(\widehat{\nu}_i - \nu_i)\tau_i = T_{6aN} + T_{6bN}$ where

$$\begin{aligned} T_{6aN} &:= \frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i \frac{D_{1i}(1-D_{2i})}{\widehat{p}_{10}(W_i)p_{10}(W_i)}(p_{10}(W_i) - \widehat{p}_{10}(W_i))\tau_i, \\ T_{6bN} &:= \frac{1}{\sqrt{N}}\sum_{i=1}^N \omega_i \frac{D_{1i}D_{2i}}{\widehat{p}_{11}(W_i)p_{11}(W_i)}(\widehat{p}_{11}(W_i) - p_{11}(W_i))\tau_i. \end{aligned}$$

We will show $T_{6aN} = o_p(1)$. Define $\mathbf{1}_N^p := \mathbf{1}\left(\inf_{w\in\mathcal{W}}\widehat{p}_{10}(w) \geq \kappa\right)$. $\mathbf{1}_N^p \xrightarrow{P} 1$ under Assumption M(2) since $\|\widehat{p} - p\|_\infty = o_p(1)$. (4) gives $E[\mathbf{1}_N^p T_{6aN}|W_1,\ldots,W_N] = 0$ and hence $E[\mathbf{1}_N^p T_{6aN}] = 0$. Similarly, (4) also gives

$$E\left[\mathbf{1}_N^p\omega_i\omega_j \frac{D_{1i}(1-D_{2i})}{\widehat{p}_{10}(W_i)p_{10}(W_i)}(p_{10}(W_i) - \widehat{p}_{10}(W_i))\frac{D_{1j}(1-D_{2j})}{\widehat{p}_{10}(W_j)p_{10}(W_j)}(p_{10}(W_j) - \widehat{p}_{10}(W_j))\tau_i\tau_j|W_1,\ldots,W_N\right] = 0$$

for all $i \neq j$. Hence $Var(\mathbf{1}_N^p T_{6aN}) = E\left[Var(\mathbf{1}_N^p T_{6aN}|W_1,\ldots,W_N)\right] = E\left[O_p(\|\widehat{p} - p\|_\infty^2)\right]$ under our assumptions. Now (29) gives $Var(\mathbf{1}_N^p T_{6aN}) = o_p(1)$ and hence $|T_{6aN}| = o_p(1)$. Similar steps give $T_{6bN} = o_p(1)$.

Finally consider the seventh (last) term: $\frac{1}{\sqrt{N}}\sum_{i=1}^N(\widehat{\omega}_i - \omega_i)\nu_i\tau_i$ and note that steps similar to that for the sixth term also show that this last term is $o_p(1)$. Hence (32) is verified.

Now we verify (33), i.e., for all positive $\delta_N = o_p(1)$ and a generic positive constant C,

$$o_p(1) = \sup_{|\beta-\beta^0|\leq\delta_N} \frac{\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \widehat{\omega}_i\widehat{\nu}_i\left(\widehat{\tau}_i(\beta) - \widehat{\tau}_i(\beta^0)\right)\right|}{1 + C\sqrt{N}|\beta - \beta^0|}$$

using that $E\left[\bar{\xi}_N(\beta, p, \bar{q}(\beta))\right] = 0$ under (4). Define $\zeta_{1i}(\beta) = \tau_i(\beta) - E[\tau_i(\beta)]$ and $\zeta_{2i}(\beta) = \widehat{\tau}_i(\beta) - \tau_i(\beta)$. Therefore, the RHS of the above is

$$\begin{gathered} \sup_{|\beta-\beta^0|\leq\delta_N} \frac{\sqrt{N}\left|A_{1N}(\beta) + A_{2N}(\beta)\right|}{1 + C\sqrt{N}|\beta - \beta^0|}, \text{ where} \\ A_{1N}(\beta) = \frac{1}{N}\sum_{i=1}^N \widehat{\omega}_i\widehat{\nu}_i\left(\zeta_{1i}(\beta) - \zeta_{1i}(\beta^0)\right), \text{ and } A_{2N}(\beta) = \frac{1}{N}\sum_{i=1}^N \widehat{\omega}_i\widehat{\nu}_i\left(\zeta_{2i}(\beta) - \zeta_{2i}(\beta^0)\right) \end{gathered}$$

since $E[\tau_i(\beta)] = 0$ for all β by (4). Now, the verification of (33) follows directly by following *exactly the same steps* as that for the corresponding terms $R_{2N}(\beta)$ (for $A_{1N}(\beta)$) and $R_{3N}(\beta)$ (for $A_{2N}(\beta)$) in the proof of Proposition -2.4 below. (Details are available from the authors.) ■

The series estimators in Proposition-2.4 are based on power series. Lemma A summarizes some well known results for such estimators. The presentation omits an important intermediate step concerning the maximizer and minimizer of the limiting objective functions for the coefficients for the power series that are treated carefully in Hirano et al. (2003) and Imbens et al. (2009). Instead we directly consider the approximation error (B2) and (B5) for the intermediate target quantities defined below in (B1) and (B4) respectively. (B4)-(B6) can be modified to accommodate for the nuisance parameters $q(Z_1, W; \beta)$ and $q(Z_2, W; \beta)$.

Lemma A: The following results hold under the conditions of Proposition-2.4:

(B1) For a fixed K_p there exists a $\pi^*_{K_p} \in \mathbb{R}^{K_p}$ such that $\|p_{d_1 d_2} - L_{d_1 d_2}(R_{K_p}, \pi^*_{K_p})\|_\infty = O(K_p^{-s_p/d_w})$.

(B2) $|\widehat{\pi}_{K_p} - \pi^*_{K_p}| = O_p\left(K_p^{1/2} N^{-1/2} + K_p^{1/2} K_p^{-s_p/d_w}\right)$ as $N \to \infty$.

(B3) $\|\widehat{p} - p\|_\infty = O_p\left(K_p[K_p^{1/2} N^{-1/2} + K_p^{1/2} K_p^{-s_p/d_w}]\right)$ as $N \to \infty$.

(B4) For a fixed K_q there exists a $\gamma^*_{K_q}(\beta^0) \in \mathbb{R}^{K_q}$ such that $\|q(\beta^0) - (I_{d_g} \otimes R'_{K_q})\gamma^*_{K_q}(\beta^0)\|_\infty = O(K_q^{-s_q/d_w})$.

(B5) $|\widehat{\gamma}_{K_q}(\beta^0) - \gamma_{K_q}(\beta^0)| = O_p\left(K_q^{1/2} N^{-1/2} + K_q^{-s_q/d_w}\right)$ as $N \to \infty$.

(B6) For $\widehat{q}(\beta^0) = (I_{d_g} \otimes R'_{K_q})\widehat{\gamma}_{K_q}(\beta^0)$, $\|\widehat{q}(\beta^0) - q(\beta^0)\|_\infty = O_p\left(K_q[K_q^{1/2} N^{-1/2} + K_q^{-s_q/d_w}]\right)$ as $N \to \infty$.

Proof of Lemma A:

See Theorem B-1 of Cattaneo (2010) for (B1)-(B3). See Lemma 1 of Newey (1994) for (B4). See Theorem 1 (including the proof) and Theorem 4 of Newey (1997) for (B5) and (B6). ■

Proof of Proposition-2.4:

Since the idea is the same, for notational simplicity let us present the proof for the case with only two-level missingness. Accordingly, *only in this proof* let $D := D_1 D_2$, $p(W) := p_{11}(W)$, and let it be known that $D_1(1 - D_2) \equiv (1 - D_1)D_2 \equiv 0$ and $p_{10}(W) \equiv p_{01}(W) \equiv 0$. Without loss of generality, let $d_g = 1$. Define $L(u) := \exp(u)/[1 + \exp(u)]$ for some scalar u (to replace the general formula $L_{d_1 d_2}(.)$).

The proof is similar to that of Theorem 5 in Cattaneo (2010). The main difference is that we will not require his condition (5.1), i.e., $\|\widehat{p} - p\|_\infty = o_p(N^{-1/4})$. His condition (5.2), i.e.,

$$\sup_{|\beta - \beta^0| < \delta} \|\widehat{q}(\beta) - q(\beta)\|_\infty = o_p(1) \tag{34}$$

is still satisfied in the same way from the first result of Proposition A1(i) of Chen et al. (2005). His condition (5.3) also holds under our setup as is shown in Lemma B below (note the contrast with the proof of Theorem 8 in Cattaneo (2010)). Hence, similar to (33) in the proof of Proposition 2.3, we only need to verify that for any positive $\delta_N = o(1)$ and some generic positive constant C:

$$\sup_{|\beta - \beta^0| \le \delta_N} \frac{\sqrt{N}\ |R_N(\beta)|}{1 + C\sqrt{N}|\beta - \beta^0|} = o_p(1), \tag{35}$$

$$\text{where } R_N(\beta) \quad = \quad \bar{m}_N(\beta, \widehat{p}, \widehat{q}(\beta)) - E\left[\bar{m}_N(\beta, p, q(\beta))\right] - \bar{m}_N(\beta^0, \widehat{p}, \widehat{q}(\beta^0))$$

and $\bar{m}(\beta, p^*, q^*) := \frac{1}{N}\sum_{i=1}^N \left\{\frac{D_i}{p^*(W_i)}\left[g(Z_i, W_i; \beta) - q^*(W_i; \beta)\right] + q^*(W_i; \beta)\right\}$ for some generic p^* and q^*.

Since $E\left[\bar{m}_N(\beta, p, q(\beta))\right] = E[g(Z_i, W_i; \beta)] = E[q(W_i; \beta)]$ and $E[g(Z_i, W_i; \beta^0)] = E[q(W_i; \beta^0)] = 0$ by (1),

we obtain: $\sqrt{N}R_N(\beta) = R_{1N}(\beta) + R_{2N}(\beta) + R_{3N}(\beta)$ where

$$\begin{aligned}
R_{1N}(\beta) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{D_i}{\widehat{p}(W_i)}[v_1(Z_i,W_i;\beta) - v_1(Z_i,W_i;\beta^0)],\\
R_{2N}(\beta) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(1-\frac{D_i}{\widehat{p}(W_i)}\right)[v_2(W_i;\beta) - v_2(W_i;\beta^0)],\\
R_{3N}(\beta) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(1-\frac{D_i}{\widehat{p}(W_i)}\right)[v_3(W_i;\beta) - v_3(W_i;\beta^0)],
\end{aligned}$$

and the individual components are $v_1(Z_i,W_i;\beta) := g(Z_i,W_i;\beta) - E[g(Z_i,W_i;\beta)]$, $v_2(W_i;\beta) := \widehat{q}(W_i;\beta) - q(W_i;\beta)$ and $v_3(W_i;\beta) := q(W_i;\beta) - E[q(W_i;\beta)]$. Now we verify (35) by working through the terms $R_{1N}(\beta)$, $R_{2N}(\beta)$ and $R_{3N}(\beta)$ respectively. First choose δ_N converging to zero slowly enough to ensure that $\mathbf{1}_N^q := \mathbf{1}\left(\sup_{\beta\in\mathcal{N}_{\delta_N}}\|\widehat{q}(\beta) - q(\beta)\|_\infty \le \delta_N\right) \xrightarrow{P} 1$ by appealing to (34). Also define $\mathbf{1}_N^p := \mathbf{1}\left(\inf_{w\in\mathcal{W}}\widehat{p}(w) \ge \kappa\right)$. Lemma (B3) gives $\|\widehat{p} - p\|_\infty = o_p(1)$ because $v_p < \frac{1}{3}$ and $\frac{s_p}{d_w} > \frac{1}{2}$ by (14). So $\mathbf{1}_N^p \xrightarrow{P} 1$ by Assumption M(2).

Consider $R_{1N}(\beta)$. Using Assumption M(2), Assumptions g(2), g(3) and g(4), arguments along the line of Theorem 4 in Cattaneo (2010) imply that the class of functions $\left\{\mathbf{1}_N^p\frac{D}{\widehat{p}(.)}v_1(.;\beta) : \beta\in\mathcal{N}_{\delta_N}\right\}$ is Donsker with finite integrable envelope and L_2 continuous. Recalling that $\mathbf{1}_N^p \xrightarrow{P} 1$ (not depending on β), it follows that $\sup_{\beta\in\mathcal{N}_{\delta_N}}|R_{1N}(\beta)|/[1 + C\sqrt{N}|\beta - \beta^0|] = o_p(1)$.

Now consider $R_{2N}(\beta)$. First by a mean-value expansion (with the mean-value being subsumed by the $\sup_{\beta\in\mathcal{N}_{\delta_N}}$ clause) and then by appealing to (34) to use Assumption q(2b) we obtain

$$\sup_{\beta\in\mathcal{N}_{\delta_N}}\mathbf{1}_N^q\mathbf{1}_N^p|R_{2N}(\beta)| \le \sup_{\beta\in\mathcal{N}_{\delta_N},\|\tilde{q}-q\|_\infty\le\delta_N}\sqrt{N}|\beta-\beta^0|\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_N^q\mathbf{1}_N^p\left|1-\frac{D_i}{\widehat{p}(W_i)}\right|\left|\frac{\partial}{\partial\beta'}[\tilde{q}(W_i;\beta) - q(W_i;\beta)]\right|.$$

Since $\mathbf{1}_N^q\mathbf{1}_N^p\left|1-\frac{D_i}{\widehat{p}(W_i)}\right| \le \max\left(1,\left|1-\frac{1}{\kappa}\right|\right)$ is bounded, using (4) and Assumption q(2b) we obtain

$$\sup_{\beta\in\mathcal{N}_{\delta_N}}\mathbf{1}_N^q\mathbf{1}_N^p\frac{|R_{2N}(\beta)|}{1+C\sqrt{N}|\beta-\beta^0|} \le C_1\delta_N^\epsilon\left[\frac{1}{N}\sum_{i=1}^{N}b(W_i)\right]$$

for some generic positive constant C_1, some non-negative measurable function $b(w)$ with $E[b(W)] < \infty$, and some $\epsilon > 0$. Letting $\delta_N \to 0$ and recalling that $\mathbf{1}_N^q \xrightarrow{P} 1$ and $\mathbf{1}_N^p \xrightarrow{P} 1$ (not depending on β) give $\sup_{\beta\in\mathcal{N}_{\delta_N}}|R_{2N}(\beta)|/[1 + C\sqrt{N}|\beta - \beta^0|] = o_p(1)$.

Finally consider $R_{3N}(\beta)$. By a mean-value expansion and then using Assumption q(2a) that allows for interchanging the order of integration and differentiation we obtain

$$\begin{aligned}
\sup_{\beta\in\mathcal{N}_{\delta_N}}\mathbf{1}_N^p|R_{3N}(\beta)| &\le \sup_{\beta\in\mathcal{N}_{\delta_N}}\sqrt{N}|\beta-\beta^0|\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_N^p\left|1-\frac{D_i}{\widehat{p}(W_i)}\right|\left|\frac{\partial}{\partial\beta'}[q(W_i;\beta) - E[q(W_i;\beta)]]\right|\\
\Rightarrow \sup_{\beta\in\mathcal{N}_{\delta_N}}\mathbf{1}_N^p\frac{|R_{3N}(\beta)|}{1+C\sqrt{N}|\beta-\beta^0|} &\le C_1\sup_{\beta\in\mathcal{N}_{\delta_N}}\frac{1}{N}\sum_{i=1}^{N}\left|\frac{\partial}{\partial\beta'}q(W_i;\beta) - E\left[\frac{\partial}{\partial\beta'}q(W_i;\beta)\right]\right|
\end{aligned}$$

for some generic positive constant C_1. The dominating integrable function in Assumption q(2a) also ensures

that $\frac{1}{N}\sum_{i=1}^{N}\left|\frac{\partial}{\partial\beta'}q(W_i;\beta)-E\left[\frac{\partial}{\partial\beta'}q(W_i;\beta)\right]\right|\xrightarrow{P}0$ uniformly in $\beta\in\mathcal{N}_{\delta_N}$. Recalling that $\mathbf{1}_N^p\xrightarrow{P}1$ (not depending on β) gives $\sup_{\beta\in\mathcal{N}_{\delta_N}}|R_{3N}(\beta)|/[1+C\sqrt{N}|\beta-\beta^0|]=o_p(1)$. ■

Lemma B: The following result holds under the conditions of Proposition-2.4 and its proof:

$$\bar{m}_N(\beta^0,\widehat{p},\widehat{q}(\beta^0))=\bar{m}_N(\beta^0,p,q(\beta^0))+o_p(N^{-1/2}).$$

where, $\bar{m}_N(\beta,p^*,q^*):=\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{D_i}{p^*(W_i)}\left[g(Z_i,W_i;\beta)-q^*(W_i;\beta)\right]+q^*(W_i;\beta)\right\}$ for some generic p^* and q^*.

Proof of Lemma B:

Note that $\sqrt{N}\left[\bar{m}_N(\beta^0,\widehat{p},\widehat{q}(\beta^0))-\bar{m}_N(\beta^0,p_1,q(\beta^0))\right]=A_{1N}+A_{2N}+A_{3N}$ where

$$\begin{aligned}
A_{1N} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{D_i}{p(W_i)\widehat{p}(W_i)}\left[g(Z_i,W_i;\beta^0)-q(W_i;\beta^0)\right]\left[p(W_i)-\widehat{p}(W_i)\right],\\
A_{2N} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{D_i}{p(W_i)}-1\right)\left[q(W_i;\beta^0)-\widehat{q}(W_i;\beta^0)\right],\\
A_{3N} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{D_i}{p(W_i)\widehat{p}(W_i)}\left[q(W_i;\beta^0)-\widehat{q}(W_i;\beta^0)\right]\left[p(W_i)-\widehat{p}(W_i)\right].
\end{aligned}$$

This is an exact relation in contrast to the proof of Theorem 8 in Cattaneo (2010). We will show one by one that $A_{1N}=o_p(1)$, $A_{2N}=o_p(1)$ and $A_{3N}=o_p(1)$.

We start with A_{1N}. Now, by (4), $E[\mathbf{1}_N^pA_{1N}|W_1,\ldots,W_N]=E[\mathbf{1}_N^pA_{1N}]=0$, and $Var(\mathbf{1}_N^pA_{1N}|W_1,\ldots,W_N)=O_p\left(\|\widehat{p}-p\|_\infty^2\right)$ further using Assumption T, and because for each $i\neq j$, we can condition on W_i,W_j to obtain

$$E\left[\frac{D_iD_j\left[p(W_i)-\widehat{p}(W_i)\right]\left[p(W_j)-\widehat{p}(W_j)\right]}{p(W_i)\widehat{p}(W_i)p(W_j)\widehat{p}(W_j)}\left[g(Z_i,W_i;\beta)-q(W_i;\beta)\right]\left[g(Z_j,W_j;\beta)-q(W_j;\beta)\right]'\right]=0$$

by (4). Therefore, Assumption T(2) gives $Var(\mathbf{1}_N^pA_{1N})=O\left(\|\widehat{p}-p\|_\infty^2\right)$ and hence $\mathbf{1}_N^pA_{1N}=O_p\left(\|\widehat{p}-p\|_\infty\right)$. Recalling that $\mathbf{1}_N^p\xrightarrow{P}1$, we obtain $A_{1N}=O_p\left(\|\widehat{p}-p\|_\infty\right)$ which is $o_p(1)$ by Lemma (B3) under (14).

Now consider $A_{2N}=B_{1N}+B_{2N}$ where

$$\begin{aligned}
B_{1N} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{D_i}{p(W_i)}-1\right)\left[q(W_i;\beta^0)-R_{K_q}(W_i)'\gamma_{K_q}^*\right],\\
B_{2N} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{D_i}{p(W_i)}-1\right)\left[R_{K_q}(W_i)'\gamma_{K_q}^*-\widehat{q}(W_i;\beta^0)\right].
\end{aligned}$$

Now, by (4), $E[B_{1N}|W_1,\ldots,W_N]=E[B_{1N}]=0$, and $Var(B_{1N}|W_1,\ldots,W_N)=O_p\left(\|q^0-R_{K_q}'\gamma_{K_q}^*\|_\infty^2\right)$. Therefore, $Var(B_{1N})=O_p\left(\|q^0-R_{K_q}'\gamma_{K_q}^*\|_\infty^2\right)$ by Assumption T(2) and using the same arguments as for A_{1N}. Therefore, $B_{1N}=O_p\left(\|q^0-R_{K_q}'\gamma_{K_q}^*\|_\infty\right)$ which is $o_p(1)$ by Lemma (B1) since $\frac{s_q}{d_w}>0$ under (14). On

the other hand,

$$|B_{2N}| = \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{D_i}{p(W_i)} - 1 \right) R_{K_q}(W_i)' \left(\gamma^*_{K_q} - \widehat{\gamma}^0_{K_q} \right) \right| \leq |\widehat{\gamma}^0_{K_q} - \gamma^*_{K_q}| \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{D_i}{p(W_i)} - 1 \right) R_{K_q}(W_i)' \right|.$$

$E\left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{D_i}{p(W_i)} - 1 \right) R_{K_q}(W_i)'\right] = 0$ and $Var\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{D_i}{p(W_i)} - 1 \right) R_{K_q}(W_i)'\right) = O\left(E|R_{K_q}(W)|^2\right) = O(K_q)$ by definition of $R_{K_q}(W)$ and using the same arguments as for A_{1N}. Hence $|B_{2N}| = O_p\left(|\widehat{\gamma}^0_{K_q} - \gamma^*_{K_q}| K_q^{1/2}\right)$, which is $o_p(1)$ by Lemma (B5) because (14) requires $v_q < \frac{1}{2}$ and $\frac{s_q}{d_w} > \frac{1}{2}$. Therefore, we obtain $A_{2N} = B_{1N} + B_{2N} = o_p(1)$.

Finally consider $A_{3N} = B_{3N} + B_{4N} + B_{5N} + B_{6N}$ where

$$\begin{aligned}
B_{3N} &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{D_i}{p(W_i)\widehat{p}(W_i)} \left[q(W_i;\beta^0) - R_{K_q}(W_i)'\gamma^*_{K_q} \right] \left[p(W_i) - L(R_{K_p}(W_i)'\pi^*_{K_p}) \right], \\
B_{4N} &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{D_i}{p(W_i)\widehat{p}(W_i)} \left[q(W_i;\beta^0) - R_{K_q}(W_i)'\gamma^*_{K_q} \right] \left[L(R_{K_p}(W_i)'\pi^*_{K_p}) - \widehat{p}(W_i) \right], \\
B_{5N} &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{D_i}{p(W_i)\widehat{p}(W_i)} \left[R_{K_q}(W_i)'\gamma^*_{K_q} - \widehat{q}(W_i;\beta^0) \right] \left[p(W_i) - L(R_{K_p}(W_i)'\pi^*_{K_p}) \right], \\
B_{6N} &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{D_i}{p(W_i)\widehat{p}(W_i)} \left[R_{K_q}(W_i)'\gamma^*_{K_q} - \widehat{q}(W_i;\beta^0) \right] \left[L(R_{K_p}(W_i)'\pi^*_{K_p}) - \widehat{p}(W_i) \right].
\end{aligned}$$

Note that $|B_{3N}| \leq \left[\frac{1}{N} \sum_{i=1}^{N} \frac{D_i}{p(W_i)\widehat{p}(W_i)}\right] \sqrt{N} \|q^0 - R'_{K_q}\gamma^*_{K_q}\|_\infty \|p - L(R'_{K_p}\pi^*_{K_p})\|_\infty$. Also, $\mathbf{1}^p_N \frac{1}{N} \sum_{i=1}^{N} \frac{D_i}{p(W_i)\widehat{p}(W_i)} = O_p(1)$. Hence $\frac{1}{N} \sum_{i=1}^{N} \frac{D_i}{p(W_i)\widehat{p}(W_i)} = O_p(1)$ recalling that $\mathbf{1}^p_N \xrightarrow{P} 1$. Thus, $|B_{3N}| \leq \sqrt{N} O_p(K_q^{-s_q/d_w}) O(K_p^{-s_p/d_w})$ by Lemmas (B1) and (B4). Since $v_p \frac{s_p}{d_w} + v_q \frac{s_q}{d_w} > \frac{1}{2}$ by (14), it follows that $|B_{3N}| = o_p(1)$.

Now denoting $\dot{L}(u) := \frac{\partial}{\partial u} L(u)$, a mean-value expansion gives for some mean-value $\tilde{\pi}$

$$B_{4N} = -\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{D_i}{p(W_i)\widehat{p}(W_i)} \left[q(W_i;\beta^0) - R_{K_q}(W_i)'\gamma^*_{K_q} \right] \dot{L}(R_{K_p}(W_i)'\tilde{\pi}) R_{K_p}(W_i)' \left(\widehat{\pi}_{K_p} - \pi^*_{K_p} \right).$$

Noting that $\dot{L}(u) = L(u)[1 - L(u)] \in (0,1)$ for all u, we obtain

$$\begin{aligned}
E|\mathbf{1}^p_N B_{4N}| &\leq \|q^0 - R'_{K_q}\gamma^*_{K_q}\|_\infty |\widehat{\pi}_{K_p} - \pi^*_{K_p}| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} E\left[\frac{\mathbf{1}^p_N D_i \dot{L}(R_{K_p}(W_i)'\bar{\pi})}{p(W_i)\widehat{p}(W_i)} |R_{K_p}(W_i)'| \right] \\
&\leq \|q^0 - R'_{K_q}\gamma^*_{K_q}\|_\infty |\widehat{\pi}_{K_p} - \pi^*_{K_p}| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sqrt{E\left[\frac{\mathbf{1}^p_N D_i}{p(W_i)\widehat{p}(W_i)} \right]^2} \sqrt{E\left[|R_{K_p}(W_i)'|^2 \right]}.
\end{aligned}$$

But $E\left[|R_{K_p}(W_i)'|^2\right] = K_p$ by definition of $R_{K_p}(W)$. Therefore, by Lemmas (B4), (B2), and (4) and Assumption M(2) respectively,

$$E|\mathbf{1}^p_N B_{4N}| \leq O(K_q^{-s_q/d_w}) O_p\left(K_p^{1/2} N^{-1/2} + K_p^{1/2} K_p^{-s_p/d_w} \right) \sqrt{N} O(1) \sqrt{K_p}.$$

Hence, $E|\mathbf{1}^p_N B_{4N}| \leq O_p\left(N^{v_p - v_q \frac{s_q}{d_w}} + N^{\frac{1}{2} + v_p - \left(v_p \frac{s_p}{d_w} + v_q \frac{s_q}{d_w}\right)} \right) = o_p(1)$ since $v_q \frac{s_q}{d_w} > v_p$ and $v_p \frac{s_p}{d_w} + v_q \frac{s_q}{d_w} > \frac{1}{2} + v_p$ by (14). This gives $|\mathbf{1}^p_N B_{4N}| = o_p(1)$. Recalling that $\mathbf{1}^p_N \xrightarrow{P} 1$, we obtain that $B_{4N} = o_p(1)$.

Following steps similar for B_{4N} we obtain

$$\begin{aligned} B_{5N} &= -\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{D_i}{p(W_i)\widehat{p}(W_i)}R_{K_q}(W_i)'\left(\widehat{\gamma}^0_{K_q}-\gamma^*_{K_q}\right)\left[p(W_i)-L(R_{K_p}(W_i)'\pi^*_{K_p})\right], \text{ and hence} \\ E|\mathbf{1}^p_N B_{5N}| &\leq \|p-L(R'_{K_q}\gamma^*_{K_p})\|_\infty|\widehat{\gamma}^0_{K_q}-\gamma^*_{K_q}|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}E\left[\frac{\mathbf{1}^p_N D_i}{p(W_i)\widehat{p}(W_i)}|R_{K_q}(W_i)'|\right] \\ &= O(K_p^{-s_p/d_w})O_p\left(K_q^{1/2}N^{-1/2}+K_q^{-s_q/d_w}\right)\sqrt{N}O(1)\sqrt{K_q} \end{aligned}$$

by Lemmas (B1), (B5), and (4) and Assumption M(2), and the definition of $R_{K_q}(W)$ respectively. Therefore, $E|\mathbf{1}^p_N B_{5N}| \leq O_p\left(N^{v_q-v_p\frac{s_p}{d_w}}+N^{\frac{1}{2}+\frac{v_q}{2}-\left(v_p\frac{s_p}{d_w}+v_q\frac{s_q}{d_w}\right)}\right)=o_p(1)$ since $v_p\frac{s_p}{d_w}>v_q$ and $v_p\frac{s_p}{d_w}+v_q\frac{s_q}{d_w}>\frac{1}{2}+\frac{v_q}{2}$ by (14). Hence, as before, it follows that $B_{5N}=o_p(1)$.

Finally, again denoting $\dot{L}(u):=\frac{\partial}{\partial u}L(u)$, a mean-value expansion gives for some mean-value $\tilde{\pi}$,

$$B_{6N} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{D_i}{p(W_i)\widehat{p}(W_i)}R_{K_q}(W_i)'\left(\widehat{\gamma}^0_{K_q}-\gamma^*_{K_q}\right)\dot{L}(R_{K_p}(W_i)'\tilde{\pi})R_{K_p}(W_i)'\left(\widehat{\pi}_{K_p}-\pi^*_{K_p}\right).$$

Using $\dot{L}(u)\in(0,1)$, note that (4) and Assumption M(2) give

$$\begin{aligned} |\mathbf{1}^p_N B_{6N}| &\leq |\widehat{\gamma}^0_{K_q}-\gamma^*_{K_q}||\widehat{\pi}_{K_p}-\pi^*_{K_p}|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\mathbf{1}^p_N D_i\dot{L}(R_{K_p}(W_i)'\bar{\pi})}{p(W_i)\widehat{p}(W_i)}|R_{K_q}(W_i)R_{K_p}(W_i)'| \\ &\leq |\widehat{\gamma}^0_{K_q}-\gamma^*_{K_q}||\widehat{\pi}_{K_p}-\pi^*_{K_p}|\frac{\sqrt{N}}{\kappa^2}\frac{1}{N}\sum_{i=1}^{N}|R_{K_q}(W_i)R_{K_p}(W_i)'|. \end{aligned}$$

$E\left[|R_{K_q}(W)R_{K_p}(W)'|\right]\leq E\left[|R_{K_q}(W)||R_{K_p}(W)'|\right]\leq\sqrt{E\left[|R_{K_q}(W)|^2\right]E\left[|R_{K_p}(W)'|^2\right]}=\sqrt{K_qK_p}$. Hence $\frac{1}{N}\sum_{i=1}^{N}|R_{K_q}(W_i)R_{K_p}(W_i)'|\leq O_p(\sqrt{K_qK_p})$. Therefore, by Lemmas (B5) and (B2),

$$|\mathbf{1}^p_N B_{6N}|\leq O_p\left(K_q^{1/2}N^{-1/2}+K_q^{-s_q/d_w}\right)O_p\left(K_p^{1/2}N^{-1/2}+K_p^{1/2}K_p^{-s_p/d_w}\right)\sqrt{N}\sqrt{K_pK_q},$$

giving $|\mathbf{1}^p_N B_{6N}|\leq O_p\left(N^{-\frac{1}{2}+v_p+v_q}+N^{v_p+v_q\left(\frac{1}{2}-\frac{s_q}{d_w}\right)}+N^{v_q+v_p\left(1-\frac{s_p}{d_w}\right)}+N^{\frac{1}{2}+v_p+\frac{v_q}{2}-\left(v_p\frac{s_p}{d_w}+v_q\frac{s_q}{d_w}\right)}\right)=o_p(1)$ since $v_p+v_q<\frac{1}{2}$, $v_q\frac{s_q}{d_w}>v_p+\frac{v_q}{2}$, $v_p\frac{s_p}{d_w}>v_p+v_q$ and $v_p\frac{s_p}{d_w}+v_q\frac{s_q}{d_w}>v_p+\frac{v_q}{2}+\frac{1}{2}$ by (14). $\mathbf{1}^p_N\xrightarrow{P}1$ implies $B_{6N}=o_p(1)$. Hence $A_{3N}=o_p(1)$. ■

Verification of the comment in Section 2 that the asymptotic variance of IPW-GMM equals the efficiency bound based on a smaller set of moment restrictions:

Toward the end of Section 2 we noted that: The asymptotic variance of the semiparametric IPW-GMM estimator in the general case equals the efficiency bound for estimation of β by combining the moment restrictions in (15) and (18) and instead considering a modified restriction:

$$E\left[\overline{\text{Proj}}_W\left(\phi(D_1,D_2,Z,W;\beta)|\phi_0\right)\right]=0 \text{ for } \beta\in\mathcal{B}\subset\mathbb{R}^{d_\beta} \text{ if and only if } \beta=\beta^0.$$

We verify this by obtaining the concerned efficiency bound. The idea is same as Theorem 2.1 and the discussion following it in Graham (2011). To convert the conditional (on W) restrictions into unconditional ones, we consider W with a known finite support $\mathcal{W} = \{w_1, w_2, \ldots, w_L\}$. This gets rid of the infinite dimensional nuisance parameters $p(W)$ that arises with an infinite support of W, and instead introduces a finite number of unknown nuisance parameters $\rho = (\rho_{10}', \rho_{01}', \rho_{11}')'$ where $\rho_{jk} = (\rho_{jk}(1) := P(D_1 = j, D_2 = k|W = w_1), \ldots, \rho_{jk}(L) := P(D_1 = j, D_2 = k|W = w_L))'$ for $j, k = 0, 1$. In Lemma C we obtain the Fisher information bound for β^0 in this model treating ρ as an unknown (finite dimensional) nuisance parameter. Since the bound does not depend on the multinomial assumption for W, the same arguments as in Graham (2011) (page 442) establish that this is the semiparametric efficiency bound β^0 under the moment restrictions (15) and (18).

Lemma C: Suppose that (i) the distribution of W has a known, finite support $\mathcal{W} = \{w_1, \ldots, w_L\}$, (ii) there is some $\beta^0 \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$ and $\rho^0 = (\rho_{10}^{0'}, \rho_{01}^{0'}, \rho_{11}^{0'})' \in R_1 \times \ldots \times R_L$ such that (15) and (18) hold. (iii) For each $l = 1, \ldots, L$ the space $R_l := \{(r_l(1), r_l(2), r_l(3)) : \text{ such that } r_l(1), r_l(2), r_l(3), 1 - (r_l(1) + r_l(2) + r_l(3)) \geq \kappa \in (0, 1)\}$ satisfies Assumption M(2). (iv) Other assumptions in Theorem 2.1 of Graham (2011) hold. Then the Fisher information bound for β^0 is $\left(G'[V + \Delta]^{-1}G\right)^{-1}$ [see (8)].

Proof of Lemma C: To simplify notations let β and ρ denote β^0 and ρ^0 unless explicitly stated otherwise. The result follows from the same three steps in the proof of Theorem 2.1 in Graham (2011).

Step 1: Let C be an $L \times 1$ vector with 1 in the l-th row if $W = w_l$ and 0 elsewhere, and $\tau_l := P(W = w_l)$. Exactly following Graham (2011), it can be established that the restrictions (15) and (18) are, in the multinomial case, equivalent to a finite number $(d_g + 3L)$ of unconditional moment restrictions:

$$E[m(D_1, D_2, Z_1, Z_2, W; \beta, \rho)] = E\begin{bmatrix} m_1(D_1, D_2, Z_1, Z_2, W; \beta, \rho) \\ m_2(D_1, D_2, W; \rho) \end{bmatrix} = 0,$$

$$\text{where} \qquad m_1(.; \beta, \rho) = \frac{D_1 D_2}{C'\rho_{11}} g(Z_1, Z_2, W; \beta) \text{ and } m_2(.; \rho) = C \otimes \begin{bmatrix} D_1(1 - D_2) - C'\rho_{10} \\ (1 - D_1)D_2 - C'\rho_{01} \\ D_1 D_2 - C'\rho_{11} \end{bmatrix}.$$

Step 2: Following Graham (2011) it can be shown that the variance bound for β under the sole restriction $E[m(D_1, D_2, Z_1, Z_2, W; \beta, \rho)] = 0$ is the upper (north-west) $d_\beta \times d_\beta$ block of the matrix $(M'\bar{V}^{-1}M)^{-1}$ where

$$\begin{aligned} M &= \left[M_\beta := E\left[\frac{\partial}{\partial \beta'} m(.; \beta, \rho)\right] = G, M_\rho := E\left[\frac{\partial}{\partial \rho'} m(.; \beta, \rho)\right]\right] \\ \bar{V} &= \begin{bmatrix} \bar{V}_{11} := E[m_1(.; \beta, \rho)m_1(.; \beta, \rho)'] & \bar{V}_{12} := E[m_1(.; \beta, \rho)m_2(.; \rho)'] \\ \bar{V}_{21} := E[m_2(.; \rho)m_1(.; \beta, \rho)'] & \bar{V}_{22} := E[m_2(.; \rho)m_2(.; \rho)'] \end{bmatrix} \end{aligned}$$

Since $m_2(.; \beta)$ does not involve β (meaning, $M_\beta = [G', 0]'$, i.e., the bottom $3L$ rows of M_β are identically 0), it follows after some algebra (shown below) that this bound is equal to

$$\left(G'\left(\bar{V}_{11} - \bar{V}_{12}\bar{V}_{22}^{-1}\bar{V}_{21}\right)^{-1} G\right)^{-1}. \tag{36}$$

This holds because the $d_\beta \times 3L$ block in the north-east of $(M'\bar{V}^{-1}M)^{-1}$ is a zero-block (and same for the $3L \times d_\beta$ block in the south-west). Under assumptions M(2)-(4), we show this below by equivalently showing that the $d_\beta \times 3L$ block in the north-east of $M'\bar{V}^{-1}M$ is a zero-block. This is equivalent to showing that the $d_\beta \times 3L$ matrix $G'\left(\bar{V}_{11} - \bar{V}_{12}\bar{V}_{22}^{-1}\bar{V}_{21}\right)^{-1} M_{1\rho} - G'\left(\bar{V}_{11} - \bar{V}_{12}\bar{V}_{22}^{-1}\bar{V}_{21}\right)^{-1}\bar{V}_{12}\bar{V}_{22}^{-1}M_{2\rho}$ is zero, where $M_{1\rho}$ and $M_{2\rho}$ respectively denote the first d_g and the last $3L$ rows of M_ρ. A sufficient condition for this is $M_{1\rho} = \bar{V}_{12}\bar{V}_{22}^{-1}M_{2\rho}$, and in the rest of Step 2 we verify that it holds. Define $A := \left[\frac{\tau_1}{\rho_{11}(1)}q(1), \ldots, \frac{\tau_L}{\rho_{11}(L)}q(L)\right]$ where $q(l) := E[g(Z,W;\beta^0)|W = w_l]$. Hence $M_{1\rho} = -[0,0,A]$. On the other hand, $M_{2\rho} = -[\tau_1 B(1)', \ldots, \tau_L B(L)']'$ where, for $l = 1, \ldots, L$, $B(l) := [e(l), e(L+l), e(2L+l)]'$ and $e(k)$ is a $3L \times 1$ unit vector with 1 in the k-th element and zeros elsewhere. Define $E[\phi(.;\beta^0)\phi_0(.)'|W] = H(W)'$ and $\left(E[\phi_0(.)\phi_0(.)'|W]\right)^{-1} = K(W)$ where

$$
\begin{aligned}
H(W)' &:= -\left[p_{10}(W), p_{01}(W), p_{11}(W) - 1\right] q(W), \\
K(W) &:= J(W)^{-1} = \begin{bmatrix} \frac{p_{00}(W)+p_{10}(W)}{p_{00}(W)p_{10}(W)} & \frac{1}{p_{00}(W)} & \frac{1}{p_{00}(W)} \\ \frac{1}{p_{00}(W)} & \frac{p_{00}(W)+p_{01}(W)}{p_{00}(W)p_{01}(W)} & \frac{1}{p_{00}(W)} \\ \frac{1}{p_{00}(W)} & \frac{1}{p_{00}(W)} & \frac{p_{00}(W)+p_{11}(W)}{p_{00}(W)p_{11}(W)} \end{bmatrix} \text{ and} \\
J(W) &:= \begin{bmatrix} p_{10}(W)(1-p_{10}(W)) & -p_{10}(W)p_{01}(W) & -p_{10}(W)p_{11}(W) \\ -p_{01}(W)p_{10}(W) & p_{01}(W)(1-p_{01}(W)) & -p_{01}(W)p_{11}(W) \\ -p_{11}(W)p_{10}(W) & -p_{11}(W)p_{01}(W) & p_{11}(W)(1-p_{11}(W)) \end{bmatrix},
\end{aligned}
$$

and for $W = w_l$ $(l = 1, \ldots, L)$ denote them by $H(l)$, $K(l)$ and $J(l)$. Therefore, some algebra gives

$$
\begin{aligned}
\bar{V}_{12} &= [\tau_1 H(1)', \tau_2 H(2)', \ldots, \tau_L H(L)'] = \bar{V}_{21}', \\
\bar{V}_{22} &= \text{diag}\{\tau_1 J(1), \tau_2 J(2), \ldots, \tau_L J(L)\}, \\
\text{and } \bar{V}_{22}^{-1} &= \text{diag}\left\{\frac{1}{\tau_1}K(1), \frac{1}{\tau_2}K(2), \ldots, \frac{1}{\tau_L}K(L)\right\}, \\
\text{and hence } \bar{V}_{12}\bar{V}_{22}^{-1} &= \left[H(1)'K(1), H(2)'K(2), \ldots, H(L)'K(L)\right].
\end{aligned}
$$

Letting $T_j(l)$ denote the j-th column of the $d_g \times 3$ matrix $H(l)'K(l)$ for $j = 1, 2, 3$ and $l = 1, \ldots, L$, we obtain

$$
\bar{V}_{12}\bar{V}_{22}^{-1}M_{2\rho} = -\left[\{\tau_1 T_1(1), \ldots, \tau_L T_1(L)\}, \{\tau_1 T_2(1), \ldots, \tau_L T_2(L)\}, \{\tau_1 T_3(1), \ldots, \tau_L T_3(L)\}\right] = M_{1\rho}
$$

by distributing the columns according to the selection elements in the matrices $B(1)$,...,$B(L)$. Therefore, the sufficient condition is verified and hence (36) gives the variance bound.[16]

Step 3: Since $\bar{V}_{11} = E[\phi(.;\beta^0)\phi(.;\beta^0)']$ and $\bar{V}_{12}\bar{V}_{22}^{-1}\bar{V}_{21} = \sum_{l=1}^{L}\tau_l H(l)'K(l)H(l) = E\left[E\left[\phi\phi_0'|W\right]\left(E\left[\phi_0\phi_0'|W\right]\right)^{-1}\right.$ $\left.E\left[\phi_0\phi'|W\right]\right]$, simple algebra gives $\bar{V}_{11} - \bar{V}_{12}\bar{V}_{22}^{-1}\bar{V}_{21} = V + \Delta$ [see (8)] and hence completes the proof . ■

[16]Recall that this sufficient condition has the important implication that knowing the nuisance parameters ρ^0 does not lead to more efficient estimator of β^0 under the current setup.