On Efficiency Gains from Multiple Incomplete Sub-samples*

Saraswata Chaudhuri†

Current version: July 18, 2019. First version: March 8, 2013.

Abstract

Cost-effective survey methods such as multi(R)-phase sampling typically generate samples that are collections of monotonic sub-samples, i.e., the variables observed for the units in sub-sample r are also observed for the units in sub-sample $r+1$ for $r = 1, \ldots, R-1$. These sub-samples represent sub-populations that can be systematically different if the selection of a unit in each phase of sampling depends on the observed variables for that unit from past phases. Our paper is about optimally combining all the sub-samples for the efficient estimation of a finite dimensional parameter defined by moment restrictions on a generic target population that is an arbitrary union of these sub-populations. Only the R-th sub-sample is assumed to contain all the variables that are arguments of the moment function. Semiparametric efficiency bounds for estimation are obtained under a unified framework allowing for full generality of the selection on observables in the sampling design. Contribution of each sub-sample toward efficient estimation is analyzed; and this turns out to differ fundamentally from that in setups where the same collection of sub-samples are instead generated unplanned by unknown sampling. Uniquely, our setup enables all the sub-samples to contribute to the efficient estimation for all the target populations, which we show is not possible in other setups. Efficient estimation is standard. Simulation evidence of substantive efficiency gains from using all the sub-samples is provided for all the targets.

JEL Classification: C13; C14; C31.

Keywords: Planned-incompleteness, Incomplete sub-samples; Multi-phase sampling; Semiparametric efficiency; Generalized method of moments.

*I am very much grateful to the editor P.C.B. Phillips, the co-editor P. Guggenberger and three anonymous referees for their detailed insightful comments. The paper was circulated before as "A Note on Efficiency Gains from Multiple Incomplete Sub-samples" but the title was modified at the suggestion of the editor. Previous versions of the paper, some of which are available on the author's webpage, benefitted from the helpful comments of A. Prokhorov, C. Muris, D. Guilkey, D. Frazier, E. Renault, F. Lange, J. Hill, J. Haushofer, J. MacKinnon, J. Wooldridge, M. Carrasco, M. Chemin, P. Saha Chaudhuri, SJ. Lee, V. Zinde-Walsh, the seminar participants at Brown, Concordia, McGill (Econ and Biostat), Queen's, U. Canterbury, U. Montreal, U. New South Wales, UNC Chapel Hill, U. Sydney, West Virginia University and the Midwest Econometrics Group meetings (2013).

†Department of Economics, McGill University, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.

1 Introduction

Planned incompleteness in the data can be useful when conducting surveys under budget constraints. The basic idea behind planned incompleteness is that: when it is expensive to collect all the variables for all the units in the sample, the next best alternative could be to collect the less expensive variables for all the units in the sample but the more expensive variables only for a subset of these units.

A variable may be more expensive to collect for numerous reasons; e.g., a correct measurement may be expensive, it may require intensive follow-ups, it may require tracking of or offering incentives to respondents, etc. In all such cases, planned incompleteness cuts the cost of surveys by generating a sample in which only a subset of the units contains all the intended variables, while the rest contains various collections of only the less expensive variables. This happens by plan, i.e., sampling design, and, thus, the targeted use of survey resources eliminates or at least reduces unplanned non-response or mismeasurement that could have otherwise complicated subsequent analyses of the data.[1]

The idea of planned incompleteness is not new, and is more frequently employed in fields of research where the use of primary data is more prevalent than what has been typical in economics. The multiple-matrix sampling of Shoemaker (1973), the split-questionnaire design of Raghunathan and Grizzle (1995), the partial questionnaire design of Wacholder et al. (1994), the many-measurement-design and the multi-forms surveys discussed by Graham et al. (1996), Graham et al. (2006), etc. are standard examples of planned incompleteness designs that are widely used in biostatistics, epidemiology, psychology and statistics to reduce the burden on the respondents and the cost of surveys.

Survey-cost is of concern to economists too as the collection of primary data under tight budget gets more common. Consequently, similar sampling designs have also been used, perhaps on an ad hoc basis, in the laboratory and field experiments in economics where one often needs to contract the sampling frame in each phase of the experiment depending on the budget left and the survey time line; see, e.g., Holt and Laury (2002), Thornton (2008), Ashraf et al. (2010), Ashraf et al. (2014), Beaman et al. (2015), Beegle et al. (2012), etc. (Appendix A.1 discusses these references and others, and sketches an example to show how the analyses therein may benefit from the results in our paper.)

However, planned incompleteness involves loss of information in the data since not all the variables are observed for all the sample units. Hence, it imperative that any estimator using such data be as precise as possible. Unfortunately, efficient estimation using such data has rarely been considered.[2]

Our paper seeks to address this issue of efficient estimation using planned incomplete data by

[1] See Carroll et al. (1995), Little and Rubin (2002), etc. for methods of dealing with mismeasured or missing data.

[2] An exception is Chatterjee and Li (2010) who consider efficiency under the specific type of planned incompleteness design known as the partial questionnaire designs of Wacholder et al. (1994) [also see Chaudhuri and Guilkey (2016)].

taking the sampling design as given. We focus on monotonic multi-phase samplings and, for full flexibility in the design of the multiple phases, we maintain a general selection on observables, i.e., the missing at random (MAR), assumption. Close attention is paid to special cases of MAR.

To fix ideas consider the prototypical multi(R)-phase sampling that, along with its variations, falls under the premise of our paper. Suppose that a researcher intends to collect R sets of variables $Z_{(1)}, \ldots, Z_{(R)}$. In phase one she collects $Z_{(1)}$ for a random selection of units. Then, recursively, at each phase $r = 2, \ldots, R$, she collects $Z_{(r)}$ for a subset of the units from phase $r-1$, selecting the subset randomly with or without regard to the already available information on $Z_{(1)}, \ldots, Z_{(r-1)}$.[3] The resulting sample consists of R groups of units such that the r-th group contains only $Z_{(1)}, \ldots, Z_{(r)}$ (these are the units followed until phase r but dropped after that) where $r = 1, \ldots, R$. We refer to these R groups as R sub-samples. Only the R-th sub-sample is complete in the sense that it contains all the intended variables $Z_{(1)}, \ldots, Z_{(R)}$. These sub-samples are monotonic, i.e., variables observed for the units in sub-sample r are also observed for the units in sub-sample $r+1$ for $r = 1, \ldots, R-1$.

Note that, the underlying populations for the sub-samples, call them sub-populations, can differ systematically in terms of the joint distribution of the intended variables $Z_{(1)}, \ldots, Z_{(R)}$ if the selection of the units for *any* phase takes into account any information that is already available by then.

In this paper we consider efficient estimation of a finite dimensional parameter defined by generic moment restrictions on the joint distribution of $Z_{(1)}, \ldots, Z_{(R)}$ in a generic target population that is an arbitrary union of these sub-populations. Special cases of the target include the sub-populations and the full population (union of all sub-populations). We provide a unified presentation of all cases.

The main contribution of our paper is twofold. First, we obtain the efficient influence function and efficiency bound for the generic parameter of interest and analyze them closely to make explicit the role of each sub-sample toward efficient estimation. Second, we express the problem of efficient estimation in an alternative, equivalent way in the spirit of Brown and Newey (1998) and Graham (2011) rendering the actual efficient estimation a simple special case of Chamberlain (1992), Ai and Chen (2012), etc. Both contributions are driven by the two key features of our paper — planned incompleteness and monotonicity — and, throughout, we emphasize the novelty of their implications with respect to the literature, e.g., Robins et al. (1994), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Hahn (1998), Chen et al. (2008), Graham (2011), Barnwell and Chaudhuri (2018).[4]

[3] The two-phase sampling is a special case where $Z_{(3)}, \ldots, Z_{(R)}$ would also be collected with $Z_{(2)}$ in phase two, and the survey would end there. In turn, the variable probability (VP) sampling studied in Wooldridge (1999), Wooldridge (2007), etc. is a special type of two-phase sampling that would discard all the units for whom only $Z_{(1)}$ was collected. Thus, VP sampling unnecessarily losses information that has already been collected (paid for). The loss is naturally more severe if such a strategy is extended to multiple phases. Hence, we do not consider VP sampling in this paper.

[4] It must also be noted here that our presentation in the sequel is incomplete in the following sense. While the idea of planned incompleteness to reduce survey cost and the unintended consequences of unplanned incompleteness is

Our paper proceeds as follows. Section 2 describes our theoretical framework and presents our main theoretical results under MAR. Section 3 takes a closer look at two important special cases of MAR that could be more relevant in practice, and provides auxiliary results along with an analytical demonstration of efficiency gained from the optimal use of all the sub-samples for efficient estimation. Section 4 describes efficient estimation and presents the consistency and asymptotic normality of the efficient estimator under high level assumptions. Section 5 demonstrates the efficiency gains (discussed in Section 3) in finite samples using a Monte Carlo experiment. Section 6 concludes. Abridged versions of the proofs of the results presented in our paper are collected in an appendix.

There is also a Supplemental Appendix containing three sections: Appendix A, Appendix B and Appendix C. This is available at Cambridge Journals Online (journals.cambridge.org/ect). Appendix A elaborates on the statements from the main text when they require longer explanations (e.g., Appendices A.1 and A.2 cited above). Appendix B fills in the details in the proofs of the results for Sections 2 and 3. Appendix C expands on the material of Section 4 by: (i) relating our estimation framework to that of Ai and Chen (2012), (ii) filling in the details in the proofs of the results for Section 4, and (iii) describing (with proof) a computationally convenient one-step updating of any $\sqrt{n}$-consistent estimator that leads to asymptotic efficiency. Appendix C also presents Monte Carlo evidence of the good finite-sample properties of the efficient estimator under the setup of Section 5.

Lastly, we note that recent contributions to data combination in economics include, e.g., Ichimura and Martinez-Sanchis (2005), Ridder and Moffitt (2007), Devereux and Tripathi (2009), Tripathi (2009), Dardanoni et al. (2011), Muris (2016), Graham et al. (2016), Abrevaya and Donald (2017), and the many references therein. It is our focus on: (i) planned incompleteness, (ii) the allowance for a dynamically updating MAR condition, and (iii) the allowance for the parameter of interest to be defined in terms of arbitrary unions of sub-populations, that distinguishes our paper from the rest.

2 Framework and the Combination of Sub-samples

2.1 Framework

Let $Z := (Z'_{(1)}, \ldots, Z'_{(R)})'$ where $Z_{(r)}$ is a $d_r \times 1$ random vector for $r = 1, \ldots, R$, and $\sum_{r=1}^{R} d_r$ is finite. Following Tsiatis (2006), consider a scalar variable C with support $\mathcal{C} := \{1, \ldots, R\}$ and a transformation $T_C(Z)$ defined as $T_r(Z) := (Z'_{(1)}, \ldots, Z'_{(r)})'$ of dimension $(\sum_{s=1}^{r} d_s) \times 1$ for $r =$

intuitively appealing, our paper is silent about the optimality of such survey designs and instead takes a generic design as given. Indeed, to our knowledge, a general optimality theory is yet to be developed for survey designs with planned incompleteness and this perhaps needs to be addressed on a case-by-case basis [see Reilly (1996)]. While a broader exploration of optimality is the topic of our ongoing research, at this point we only present in Appendix A.2 simple and illustrative examples of the optimality of a planned incomplete survey design in two-phase sampling.

$1, \ldots, R$. The value of C determines $T_C(Z)$, i.e., how much of Z is observed for a sample unit.

Let $O := (C, T_C'(Z))'$ denote what is observable for a unit. The observed sample is $\{O_i := (C_i', T_{C_i}'(Z_i))'\}_{i=1}^n$. The r-th sub-sample is the collection of units with only $T_r(Z)$ observed; it is of size $n_r := \sum_{i=1}^n I(C_i = r)$ for $r = 1, \ldots, R$. Only the R-th sub-sample is complete, i.e., $T_R(Z) = Z$.

A natural consequence of our description of the multi-phase sampling is that it involves selection on observables. Note that, at the end of phase $r = 1, \ldots, R-1$, the researcher is left with the units $\{i = 1, \ldots, n : C_i \geq r\}$ and has observed $T_r(Z_i)$ for each of them. Now, the researcher decides the probability with which each such eligible unit continues to phase $r+1$. This probability can be equal, say $1-p_r$, for all eligible units, in which case $P(C = r | C \geq r, T_r(Z)) = p_r$; or it can be more involved if it depends on $T_r(Z)$. Regardless, the researcher cannot possibly incorporate in this decision the knowledge of $Z_{(r+1)}, \ldots, Z_{(R)}$ since she does not observe them by the end of phase r. We formalize this statement by maintaining a general selection on observables, i.e., the MAR condition that:

$$P(C = r | C \geq r, T_R(Z)) = P(C = r | C \geq r, T_r(Z)), \text{ equivalently, } P(C = r | T_R(Z)) = P(C = r | T_r(Z)) \tag{1}$$

for $r = 1, \ldots, R$. The equivalence in (1) follows from the invertible relation between hazard and probability mass functions [see Appendix A.3]. The second relation in (1) is the MAR condition in the sense of Rubin (1976) [see, e.g., Robins and Rotnitzky (1995), Tsiatis (2006)] and generalizes to the case of $R > 2$ the MAR assumption found in econometrics where the focus has traditionally been on $R = 2$ [see, e.g., Chen et al. (2005), Chen et al. (2008), Graham (2011), Graham et al. (2012)].

To signify that the incompleteness in the data is by plan/design, we maintain under (1) that:

$$P(C = r | C \geq r, T_r(Z)), \text{ equivalently, } P(C = r | T_r(Z)) \text{ is known for each } r = 1, \ldots, R. \tag{2}$$

The equivalence in (2) follows under (1) [see Appendix A.4]. The condition in (2) is a formality in this context since the researcher actually decides these probabilities as part of the sampling design.

Now, to define the parameter value of interest, consider a generic function $m(Z; \beta) : \text{Support}(Z) \times \mathcal{B} \mapsto \mathbb{R}^d$ of the parameter $\beta \in \mathcal{B} \subset \mathbb{R}^d$. (The case of over-identification is considered in older versions of this paper, but it does not provide any new insights.) For a given target population $\lambda \in \Lambda$ where $\Lambda := \text{Power-Set}(\mathcal{C})$ excluding the empty set, define the parameter value of interest β_λ^0 as:

$$E[m(Z; \beta) | C \in \lambda] = 0 \text{ for } \beta \in \mathcal{B} \Longleftrightarrow \beta = \beta_\lambda^0. \tag{3}$$

β_λ^0 is defined as a function of λ and may differ across targets $\lambda \in \Lambda$ if C and Z are dependent.

For a given β, the function $m(Z; \beta)$ can be evaluated from the observed sample only for the n_R

units in the complete sub-sample, i.e., $I(C = R)m(Z; \beta)$. However, point identification of β^0_λ is still possible by the Horvitz-Thompson re-weighting provided that $P(C = R|T_R(Z)) > 0$ almost surely. This is due to the following relation that holds identically in β [see Appendix A.5 for details]:

$$E\left[\frac{P(C \in \lambda|T_R(Z))}{P(C \in \lambda)} \frac{I(C = R)}{P(C = R|T_R(Z))} m(Z; \beta)\right] = E[m(Z; \beta)|C \in \lambda]. \tag{4}$$

All the terms inside the expectation on the left hand side (LHS) of (4) will be feasible under our assumptions because $P(C \in \lambda)$ will be trivially identified by the observed data under assumption (A1) below, whereas (1) and (2) already imply that $P(C \in \lambda|T_R(Z))$ and $P(C = R|T_R(Z))$ are known. Hence, the LHS of (4) can serve as the estimating function for β^0_λ. However, such estimation will be based solely on the complete sub-sample. We will focus on exploring the information contained in the incomplete sub-samples and demonstrating how that information can be combined with the information in the complete sub-sample for the purpose of efficient estimation of β^0_λ defined in (3).

The discussion of our framework concludes by listing an assumption that we maintain hereafter.

Assumption A

(A1) The observed sample units $\{O_i := (C_i, T'_{C_i}(Z_i))\}_{i=1}^n$ are i.i.d. copies of $O := (C, T'_C(Z))'$.

(A2) $(P(C = r|T_R(Z)))_{r=1}^{R-1} > 0$ and $P(C = R|T_R(Z)) > \underline{p}$ almost surely in $T_R(Z)$ for a fixed $\underline{p} \in (0, 1)$.

(A3) $M_\lambda := \left\{\frac{\partial}{\partial \beta'} E\left[m(Z; \beta)|C \in \lambda\right]\right\}_{\beta=\beta^0_\lambda}$ is a $d \times d$ finite nonsingular matrix.

Remark: (A1) is a standard assumption [see, e.g., Tsiatis (2006), Devereux and Tripathi (2009), Tripathi (2011), etc.]. $P(C = R|T_R(Z)) > \underline{p} > 0$ in (A2) is a strict version of the overlap assumption [see Khan and Tamer (2010)]. The restrictions $P(C = r|T_R(Z)) > 0$ for $r = 1, \ldots, R - 1$ are not strictly required but help to avoid more involved proofs peripheral to the main message. However, $P(C = r) > 0$ for $r = 1, \ldots, R$ is intrinsic to the R-level missing data model. (A3) allows for moment vectors $m(Z; \beta)$ that are not differentiable in β. We do, however, impose differentiability of $E[m(Z; \beta)|C \in \lambda]$ as in, e.g., Chen et al. (2003), Chen et al. (2008), Cattaneo (2010), etc.

The theoretical framework above is closely related to several well-known papers such as Robins and Rotnitzky (1995), Whittemore (1997), Holcroft et al. (1997), Chen et al. (2005), Chen et al. (2008), Cattaneo (2010), Dardanoni et al. (2011), Lee et al. (2012) and Abrevaya and Donald (2017). In Appendix A.6 we discuss in detail where we actually differ from them. Broadly speaking, the differences are one or more of the following: (i) allowance for a general R, (ii) expansion of the scope to all $(2^R - 1)$ sub-populations (including $\lambda = \mathcal{C}$), (iii) introduction of a dynamically updated sampling design via MAR, and (iv) emphasis on the new insights available only from letting $R > 2$.

2.2 Optimally combining the sub-samples for efficiency

To state our main result in Proposition 1 let us first, for a given $\lambda \in \Lambda$, define the following $d \times 1$ functions of the observed data O and the $d \times 1$ parameter β as:

$$\begin{aligned} \varphi_{r,\lambda}(O;\beta) &:= E\left[\frac{P(C \in \lambda | T_R(Z))}{P(C \in \lambda)} m(T_R(Z);\beta) \middle| T_r(Z) \right] \text{ for } r = 1, \ldots, R, & (5) \\ \varphi_\lambda(O;\beta) &:= \frac{I(C = R)}{P(C = R | T_R(Z))} \varphi_{R,\lambda}(O;\beta) \\ &+ \sum_{r=1}^{R-1} \left[\frac{I(C \geq R - r)}{P(C \geq R - r | T_{R-r}(Z))} - \frac{I(C \geq R - r + 1)}{P(C \geq R - r + 1 | T_{R-r+1}(Z))} \right] \varphi_{R-r,\lambda}(O;\beta). & (6) \end{aligned}$$

Proposition 1 *Let* (1), (2), (3) *and assumption A hold. Let the* $d \times d$ *matrix* $V_\lambda := Var(\varphi_\lambda(O;\beta_\lambda^0))$ *be finite and positive definite where* $\varphi_\lambda(O;\beta)$ *is defined in* (6) *and* β_λ^0 *is defined in* (3). *Then, the asymptotic variance lower bound for any regular estimator of* β_λ^0 *is given by* $\Omega_\lambda := M_\lambda^{-1} V_\lambda M_\lambda^{-1'}$. *A regular estimator whose asymptotic variance equals* Ω_λ *has the asymptotically linear representation:*[5]

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta_\lambda^0) = -M_\lambda^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_\lambda(O_i;\beta_\lambda^0) + o_p(1).$$

Remarks:

1. Chen et al. (2008)'s results are for $R = 2$ with $\lambda = \{1\}$ and $\lambda = \{1,2\}$. Proposition 1 generalizes Theorem 2 of Chen et al. (2008) to the case of a generic R and a generic target λ. To see this, let $R = 2$. Then, under (1), equations (5) and (6) imply that for $\lambda = \{1,2\}$ and $\{1\}$ respectively:

$$\begin{aligned} \varphi_{\{1,2\}}(O;\beta) &= \frac{I(C = 2)}{P(C = 2 | T_1(Z))} \left(m(T_2(Z);\beta) - E[m(T_2(Z);\beta) | T_1(Z)] \right) + E[m(T_2(Z);\beta) | T_1(Z)], \\ \varphi_{\{1\}}(O;\beta) &= \frac{P(C = 1 | T_1(Z))}{P(C = 1)} \varphi_{\{1,2\}}(O;\beta), \end{aligned}$$

giving exactly the same expressions as in Chen et al. (2008) (p. 830) [see Appendix A.7 for details].

Interestingly, however, pointing to the crux of the matter related to the planned incompleteness condition (2) is the case where $R = 2$ and $\lambda = \{2\}$. (This case is not considered in Chen et al. (2008).) In this case, our Proposition 1 implies that (following steps as in Appendix A.7):

$$\varphi_{\{2\}}(O;\beta) = \frac{I(C = 2)}{P(C = 2)} m(T_2(Z);\beta) + \left(\frac{P(C = 2 | T_1(Z))}{P(C = 2)} - \frac{I(C = 2)}{P(C = 2)} \right) E[m(T_2(Z);\beta) | T_1(Z)],$$

i.e., all the sub-samples still contribute toward efficient estimation (as evident from the first term inside parentheses on the RHS), a phenomenon that holds for the other targets (λ's) too. On the other hand, the generic result for unplanned incompleteness (i.e., without (2)) under MAR in Proposition

[5] Over-identified models were considered in older versions of the paper, but the proof for the corresponding result in those versions had logical gaps since the characterization of the tangent set was too general and did not take account of the over-identifying restrictions. We thank a referee for pointing out this problem with the older versions.

1 of Barnwell and Chaudhuri (2018) would imply that, when $R = 2$ and $\lambda = \{2\}$, then:

$$\varphi_{\{2\}[u]}(O;\beta) = \frac{I(C=2)}{P(C=2)} m(T_2(Z);\beta),$$

rendering the incomplete sub-sample useless. (The subscript $[u]$ stands for unknown/unplanned.) This comparison makes evident the important benefit of the planned incompleteness approach that makes all the sub-samples always usable irrespective of λ. It is also straightforward to see that:

$$Var\left(\varphi_{\{2\}}(O;\beta^0_{\{2\}})\right) = Var\left(\varphi_{\{2\}[u]}(O;\beta^0_{\{2\}})\right) - E\left[\frac{P(C=1|T_1(Z))P(C=2|T_1(Z))}{P^2(C=2)} q(T_1(Z))q'(T_1(Z))\right].$$

where $q(T_1(Z)) := E[m(T_2(Z);\beta^0_{\{2\}})|T_1(Z)]$. Hence, in this case, the difference between planned and unplanned incompleteness in terms of the efficiency bound boils down to the additional (relevant) information brought by the incomplete sub-sample, which is reflected by the last term on the RHS.

2. When $R > 2$, Chen et al. (2008)'s selection on observables assumption (Assumption 2) can be generalized as MAR in (1) (or its special cases noted in Section 3). Proposition 1 works under MAR. The result for a generic R under MAR and when the target is $\lambda = \mathcal{C}$ has been known since Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Robins et al. (1995), Holcroft et al. (1997).

On the other hand, the novelty in Proposition 1 is that it allows for any target λ. The key to obtaining this result under a unified framework is how we treat the term $P(C \in \lambda|T_R(Z))$ in (5) (immaterial when $\lambda = \mathcal{C}$ since $P(C \in \mathcal{C}|T_R(Z)) \equiv 1$). Our treatment simplifies to Chen et al. (2008)'s treatment when considering their verify-out-of-sample case, i.e., when $R = 2$ and $\lambda = \{1\}$, from which, however, an extension to the general case in our paper may not seem obvious *ex ante*.

3. $\varphi_\lambda(O;\beta)$ in (6) belongs to the class of AIPW (Augmented Inverse Probability Weighted) estimating functions of Robins et al. (1994). The first term $\varphi_{R,\lambda}(O;\beta)$ is the IPW term based on the complete sub-sample. The rest are the augmentations due to the incomplete sub-samples: the r-th term represents the contribution of the $(R-r+1)$-th sub-sample. Each of these R terms are themselves unbiased estimating function for β^0_λ but only the first one, i.e., the IPW term, is known without further assumptions [see below (4)]. The augmentation terms reduce the variability of the IPW estimating function and thereby deliver the efficient AIPW estimating function. More precisely:

$$Cov(\text{term}_1, \text{term}_r) = -Var(\text{term}_r) \text{ for } r = 2,\dots,R$$

$$Cov(\text{term}_s, \text{term}_r) = 0 \text{ for } s \neq r \neq 1,$$

$$\text{and hence } V_\lambda = Var\left(\sum_{r=1}^{R} \text{term}_r\right) = Var(\text{term}_1) - \sum_{r=2}^{R} Var(\text{term}_r).$$

The $(R-r+1)$-th sub-sample's contribution to the efficiency of estimation for β^0_λ rises with $Var(\text{term}_r)$ for $r>1$, countering $Var(\text{term}_1)$ to decrease the variance of the estimating function.[6]

2.3 Contribution of the observability of the $Z_{(r)}$'s toward efficiency

Let us now look into combining the sub-samples from an alternative viewpoint that stresses on how the observability of each $Z_{(r)}$ contributes toward efficiency. To this end, rearrange the terms on the RHS of (6) and rewrite $\varphi_\lambda(O;\beta)$ as:

$$\varphi_\lambda(O;\beta) = \varphi_{1,\lambda}(O;\beta) + \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r(Z))}\left[\varphi_{r,\lambda}(O;\beta) - \varphi_{r-1,\lambda}(O;\beta)\right] \tag{7}$$

to slice the contribution of the sub-samples differently. (Note that, $I(C\geq R)\equiv I(C=R)$.) Consider the r-th term on the RHS. $\varphi_{r,\lambda}(O;\beta)$ and $\varphi_{r-1,\lambda}(O;\beta)$ differ due to $Z_{(r)}$, which is only observed for all the $(R-r+1)$ sub-samples (i.e., for all the units $i=1,\ldots,n: C_i\geq r$) as is signified by the multiplier $I(C\geq r)$. Thus, the contribution of all the R sub-samples toward estimation is represented in this r-th term in an incremental fashion according to their ability in delivering an observable $Z_{(r)}$. This holds for each $r=1,\ldots,R$, i.e., including the first term on the RHS of (7). Note that the R terms on the RHS of (7) are uncorrelated. Therefore, V_λ is the sum of the variances of the R terms:

$$V_\lambda = Var(\varphi_{1,\lambda}(O;\beta^0_\lambda)) + \sum_{r=2}^{R} E\left[\frac{Var(\varphi_{r,\lambda}(O;\beta^0_\lambda)|T_{r-1}(Z))}{P(C\geq r|T_r(Z))}\right].$$

The variance inflation factor $1/P(C\geq r|T_r(Z))$ accounts for the observability of $Z_{(r)}$ by varying inversely with the conditional probability of observing $Z_{(r)}$. Naturally, there is no such variance inflation for the first term on the RHS of (7) since $Z_{(1)}$ is always observed.

Yet another way of looking at these incremental contributions is to design a set of extended moment restrictions whose information content, when combined optimally, equals that in Proposition 1. Accordingly, consider the estimation of β^0_λ based on the moment restrictions:

$$E\left[\phi_{R,\lambda}(O;\beta)\right] = 0 \text{ for } \beta\in\mathcal{B} \Longleftrightarrow \beta=\beta^0_\lambda, \tag{8}$$

$$E\left[\phi_{R-r}(O)|T_{R-r}(Z)\right] = 0 \text{ almost surely } T_{R-r}(Z) \text{ for } r=1,\ldots,R-1; \tag{9}$$

where $\phi_{R,\lambda}(O;\beta) := \dfrac{I(C=R)}{P(C=R|T_R(Z))}\varphi_{R,\lambda}(O;\beta)$ [the IPW term from (6)],

$$\phi_{R-r}(O) := I(C\geq R-r)\left[I(C\geq R-r+1) - P(C\geq R-r+1|C\geq R-r, T_{R-r}(Z))\right]$$

[6] While $Var(\varphi_{R-r+1,\lambda}(O;\beta)) \geq Var(\varphi_{R-r,\lambda}(O;\beta))$ (in a matrix sense), the order is not always preserved when comparing the relative contribution of $Var(\text{term}_r)$ and $Var(\text{term}_{r+1})$ for $r>1$ because these latter variances are affected by certain conditional probabilities in a nontrivial way as evident from the expression that $Var(\text{term}_r) = E\left[\frac{P(C=R-r+1|T_{R-r+1}(Z))}{P(C\geq R-r+1|T_{R-r+1}(Z))P(C\geq R-r+2|T_{R-r+2}(Z))}\varphi_{R-r+1,\lambda}(O;\beta^0_\lambda)\varphi'_{R-r+1,\lambda}(O;\beta^0_\lambda)\right]$ for $r=2,\ldots,R$. This is what complicates a general optimality theory for the survey design, which is the topic of our ongoing work [see footnote 4].

for $r = 1\ldots, R-1$. When considering the expression for $\phi_{R-r}(O)$, note that, by definition, $I(C \geq R-r) = I(C \geq 1) \equiv 1$ when $r = R-1$, and $I(C \geq R-r+1) = I(C \geq R) \equiv I(C = R)$ when $r = 1$.

Under (1) and (2), the moment restriction in (8) already identifies β^0_λ [see below (4)], and GMM estimation based on it using the complete sub-sample is the GMM-version of the Horvitz-Thompson method of obtaining IPW estimators [see, e.g., Wooldridge (2007)].

The key to our following discussion, on the other hand, is the moment restrictions in (9). These restrictions do not involve β but bring additional information due to the observability of the $Z_{(r)}$'s in the sub-samples. Under the monotonic structure of the observed data, this information is usable due to the MAR condition (1) and, importantly, the planned incompleteness condition (2).

(2) did not play a role in similar discussions in the literature, e.g., Graham (2011), Chaudhuri and Guilkey (2016), etc., since they focused on the full population, i.e., $\lambda = \mathcal{C}$, for which the efficiency bound is the same irrespective of (2). However, we also consider sub-populations; and, hence, (2) will play an important role here without which the contribution of the $Z_{(r)}$'s would be further attenuated.

Additionally, the monotonic structure also plays an important role in our discussion, as is apparent from a comparison with the results in pp. 686-687 of Chaudhuri and Guilkey (2016). The monotonicity is captured by the multiplier $I(C \geq r)$ for the r-th moment function $\phi_r(O)$ for $r = 1, \ldots, R-1$. This multiplier ensures that the corresponding moment restriction reflects the additional information that becomes available due to the observability of $Z_{(r)}$, which is observed if and only if $C \geq r$.

Proposition 2 *Denote $\phi_r(O)$ by ϕ_r for $r = 1, \ldots, R-1$, and define $\overline{Proj}_{T_r}(Y|\,\phi_r) := Y - Proj_{T_r}(Y|\,\phi_r)$ and $Proj_{T_r}(Y|\,\phi_r) := E[Y\phi_r|T_r(Z)]\left(E\left[\phi_r^2|T_r(Z)\right]\right)^{-1}\phi_r$ for any random variable Y whenever the relevant terms in the definition exist. Then, the following results hold.*
(i) If (1) and assumptions (A1) and (A2) hold, then $\varphi_\lambda(O;\beta)$ defined in (6) satisfies:

$$\varphi_\lambda(O;\beta) = \overline{Proj}_{T_1}\left(\overline{Proj}_{T_2}\left(\ldots\overline{Proj}_{T_{R-2}}\left(\overline{Proj}_{T_{R-1}}\left(\phi_{R,\lambda}(O;\beta)|\,\phi_{R-1}\right)\Big|\,\phi_{R-2}\right)\ldots\Big|\,\phi_2\right)\Big|\,\phi_1\right).$$

(ii) Let (1), (2) and assumption A hold. Then, the asymptotic variance lower bound under (8) and (9) for any regular estimator of β^0_λ is Ω_λ as defined in Proposition 1. A regular estimator with asymptotic variance Ω_λ has the same asymptotically linear representation as that in Proposition 1.

Remarks:

1. The results in (ii) under the moment restrictions (8)-(9) and the conditions (1) and (2) follow directly as a special case of Chamberlain (1992) and Ai and Chen (2012).[7]

[7] To match the sequential moment restrictions in Chamberlain (1992) and Ai and Chen (2012), define $T_0(Z)$ as a constant and consider the unconditional expectation in (8) equivalently as the expectation conditional on $T_0(Z)$.

2. The result in (i) is essentially a repeated application of equation (15) in Brown and Newey (1998) [see also Theorem 2.1 of Graham (2011)] facilitated by the monotonicity ($T_r(Z)$ nests $T_{r-1}(Z)$) of the conditioning sets in (9). As noted in pp. 686-687 of Chaudhuri and Guilkey (2016), who also refer to an earlier version of our current paper, a similar exercise under a non-monotonic structure would not lead to the efficient influence function except in very special cases [see their footnote 5].

3. The broad message of Proposition 2 is that the original problem of optimally combining the sub-samples can be boiled down to an equivalent problem of optimally combining a set of carefully chosen moments restrictions, a problem/idea that is perhaps more common in economics [see Appendix A.8 for further discussion and its relation with the calibration literature in survey sampling].

Graham (2011) was first to establish a similar result for the case where $R = 2$ and the target was $\lambda = \mathcal{C}$. Our setup is more involved and thus requires condition (2) and an adequately rich choice for the sequence of functions $(\phi_{R-r}(O))_{r=1}^{R-1}$ in (9) to establish this equivalence result that provides the alternative viewpoint to appreciate the contribution of the sub-samples toward efficient estimation.

4. It is important to note that the more involved nature of our setup is not just that we allow for $R > 2$, but also because we allow for the sub-populations to be the target λ. This latter feature helps to highlight a pertinent implication of the planned incompleteness condition in (2). To make this point, take $R = 2$ to match the setup of Hahn (1998), Chen et al. (2008), and, importantly, Graham (2011). Now, note that, under assumption (A2) (that now becomes Graham (2011)'s Assumption 1.4) our augmenting moment restriction (9) becomes $E[I(C = 2) - P(C = 2|T_1(Z))|T_1(Z)] = 0$ almost surely in $T_1(Z)$, i.e., the same as Graham (2011)'s [equation (5)] augmenting (auxiliary) moment restriction. However, when $\lambda = \{1\}$, Proposition 2 gives the efficiency result only under (2) but not under unplanned incompleteness, and this can be seen simply by comparing Case 1 in Theorems 1 and 2 of Chen et al. (2008) [see Appendix A.9 for details]. This point was moot in Graham (2011) because the efficiency results are identical under planned or unplanned incompleteness when $\lambda = \mathcal{C}$ $(\equiv \{1, 2\})$. Therefore, it is important to recognize that, *in general*, equivalence results such as Proposition 2 hold only under planned incompleteness (along with monotonicity; see Remark 2).

3 A closer look at two special cases of MAR: CMAR and INDEP

It is instructive to observe the simplifications in the efficient influence function and, thus, the efficiency bound if instead of the general MAR condition in (1), one maintains the following stronger conditions that rule out dynamically updated survey designs and makes it easier to plan ahead with the survey:

$$\text{CMAR:} \qquad P(C=r|T_R(Z)) = P(C=r|T_1(Z)) \;\; \text{for } r=1,\ldots,R, \tag{10}$$

$$\text{INDEP:} \qquad P(C=r|T_R(Z)) = P(C=r) \;\; \text{for } r=1,\ldots,R. \tag{11}$$

Convenient MAR (CMAR) sampling happens if the sampling design for the later phases is based only on the observed variables from the first phase (baseline). CMAR and MAR are trivially the same in the commonly studied case of $R=2$, i.e., both generalize Chen et al. (2008). Independent (INDEP) sampling happens if the sampling design is independent of Z. $\lambda = \mathcal{C}$ is the only target of interest under INDEP since β_λ^0 does not vary with λ; however, that is not the case under CMAR.

Given the results in Barnwell and Chaudhuri (2018), it is also instructive to compare the results thus obtained with those where one cannot maintain or, as in our paper, enforce by design the condition of planned incompleteness in (2). Both issues will be extensively analyzed in this section.

Under CMAR and INDEP, $P(C \in \lambda|T_R(Z))$ becomes $P(C \in \lambda|T_1(Z))$ and $P(C \in \lambda)$ respectively for all $\lambda \in \Lambda$. Since all sub-populations λ's are the same under INDEP but not under CMAR, the discussion under CMAR is going to be more involved. Hence, in the sequel, we primarily focus on the discussion under CMAR while complementing it with the associated results under INDEP.

3.1 Efficient influence functions and efficiency bounds

For brevity, we follow the spirit of the equivalent form of $\varphi_\lambda(O;\beta)$ in (7) and define, for all $\lambda \in \Lambda$,

$$\varphi_\lambda^{\text{CMAR}}(O;\beta) := \frac{P(C\in\lambda|T_1(Z))}{P(C\in\lambda)} \left\{ q(T_1(Z);\beta) + \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r)} \left[q(T_r(Z);\beta) - q(T_{r-1}(Z);\beta)\right] \right\} \tag{12}$$

where:

$$q(T_r(Z);\beta) := E\left[m(T_R(Z);\beta)|\, T_r(Z)\right] \;\; \text{for } r=1,\ldots,R. \tag{13}$$

It is straightforward to see using (7) that $\varphi_\lambda(O;\beta)$ in (6) boils down to $\varphi_\lambda^{\text{CMAR}}(O;\beta)$ under CMAR. (While this seems trivial, we will later point out by appealing to Barnwell and Chaudhuri (2018) that such nesting does not hold in general if the planned incompleteness condition in (2) is relaxed.)

Proposition 3 *Let* (2)*,* (3)*,* (10) *and assumption A hold. Let the* $d\times d$ *matrix* $V_\lambda := Var(\varphi_\lambda^{CMAR}(O;\beta_\lambda^0))$ *be finite and positive definite where* $\varphi_\lambda^{CMAR}(O;\beta)$ *is defined in* (12) *and* β_λ^0 *is defined in* (3)*. Then, the asymptotic variance lower bound for any regular estimator of* β_λ^0 *is given by* $\Omega_\lambda := M_\lambda^{-1} V_\lambda M_\lambda^{-1'}$*. A regular estimator whose asymptotic variance equals* Ω_λ *has the asymptotically linear representation:*

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta_\lambda^0) = -M_\lambda^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_\lambda^{CMAR}(O_i;\beta_\lambda^0) + o_p(1).$$

Proposition 4 *Let* (3), (10) *and assumption A hold. Assume* $P(C=r|T_1(Z)) = P(C=r|T_1(Z);\gamma^0)$ *for some* $\gamma^0 \in \Gamma \subset \mathbb{R}^{d_\gamma}$ *where* $P(C=r|T_1(Z);\gamma)$ *is known up to the finite-dimensional unknown* γ *for* $r=1,\ldots,R$. *Let* $S_\gamma(C|T_1(Z)) := \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|T_1(Z))} \frac{\partial}{\partial\gamma} P(C=r|T_1(Z);\gamma^0)$ *denote the score function for* γ *evaluated at* $\gamma=\gamma^0$, *and assume that* $E\left[S_\gamma(C|T_1(Z))S_\gamma(C|T_1(Z))'\right]$ *is positive definite. Define*

$$\varphi_{\lambda[pu]}^{CMAR}(O;\beta) := \varphi_\lambda^{CMAR}(O;\beta) + \Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)} E[m(T_R(Z);\beta)|T_1(Z)] \middle| S_\gamma(C|T_1(Z))\right)$$

where the subscript $[pu]$ *represents that* $P(C=r|T_1(Z))$ *is partially unknown, i.e., the finite dimensional parameter* γ *is unknown;* $\varphi_\lambda^{CMAR}(O;\beta)$ *is as in* (12)*; and for any variables* Y *and* X, *let* $\Pi(Y|X) := E[YX'](E[XX'])^{-1}X$ *denote the population least squares projection when it exists.*[8] *Let* $V_{\lambda[pu]} := Var(\varphi_{\lambda[pu]}^{CMAR}(O;\beta_\lambda^0))$ *be a* $d\times d$ *finite positive definite matrix. Then, the asymptotic variance lower bound for any regular estimator of* β_λ^0 *is given by* $\Omega_{\lambda[pu]} := M_\lambda^{-1} V_{\lambda[pu]} M_\lambda^{-1'}$. *A regular estimator whose asymptotic variance equals* $\Omega_{\lambda[pu]}$ *has the asymptotically linear representation:*

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta_\lambda^0) = -M_\lambda^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{\lambda[pu]}^{CMAR}(O_i;\beta_\lambda^0) + o_p(1).$$

Proposition 5 *Let* (3), (10) *and assumption A hold. Define*

$$\varphi_{\lambda[u]}^{CMAR}(O;\beta) := \frac{I(C\in\lambda)}{P(C\in\lambda)} q(T_1(Z);\beta) + \frac{P(C\in\lambda|T_1(Z))}{P(C\in\lambda)} \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r)} \left[q(T_r(Z);\beta) - q(T_{r-1}(Z);\beta)\right]$$

where the subscript $[u]$ *represents that* $P(C=r|T_1(Z))$ *is unknown; and* $q(T_r(Z);\beta)$ *is as defined in* (13) *for* $r=1,\ldots,R$. *Let* $V_{\lambda[u]} := Var(\varphi_{\lambda[u]}^{CMAR}(O;\beta_\lambda^0))$ *be a* $d\times d$ *finite positive definite matrix. Then, the asymptotic variance lower bound for any regular estimator of* β_λ^0 *is given by* $\Omega_{\lambda[u]} := M_\lambda^{-1} V_{\lambda[u]} M_\lambda^{-1'}$. *A regular estimator whose asymptotic variance equals* $\Omega_{\lambda[u]}$ *has the asymptotically linear representation:*

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta_\lambda^0) = -M_\lambda^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{\lambda[u]}^{CMAR}(O_i;\beta_\lambda^0) + o_p(1).$$

Remarks:

1. Proposition 3 turns out to be a special case of Proposition 1. Proposition 5 fully relaxes the planned incompleteness condition (2). Proposition 4 is an intermediate result partially relaxing (2).

[8] In terms of the notation for the conditional projection $\text{Proj}_{.}(.|.)$ in the statement of Proposition 2, $\Pi(Y|X) \equiv \text{Proj}_{T_0(Z)}(Y|X)$ where $T_0(Z)$ is defined as any constant, which makes the conditional projection an unconditional one.

2. It is straightforward to see (after some algebra) from Propositions 3-5 that:

$$\begin{aligned}
V_{\lambda[u]} &= E\left[\frac{P(C\in\lambda|T_1(Z))}{P^2(C\in\lambda)}q(T_1(Z);\beta^0_\lambda)q'(T_1(Z);\beta^0_\lambda)+\frac{P^2(C\in\lambda|T_1(Z))}{P^2(C\in\lambda)}\sum_{r=2}^{R}\frac{Var\left(q(T_r(Z);\beta^0_\lambda)|T_{r-1}(Z)\right)}{P(C\geq r|T_1(Z))}\right]\\
V_\lambda &= V_{\lambda[u]}-E\left[\frac{P(C\in\lambda|T_1(Z))(1-P(C\in\lambda|T_1(Z)))}{P^2(C\in\lambda)}q(T_1(Z);\beta^0_\lambda)q'(T_1(Z);\beta^0_\lambda)\right]\\
V_{\lambda[pu]} &= V_\lambda + B\left(E\left[S_\gamma(C|T_1(Z))S_\gamma(C|T_1(Z))'\right]\right)^{-1}B'\\
&= V_{\lambda[u]}-Var\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}q(T_1(Z);\beta^0_\lambda)-\Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}q(T_1(Z);\beta^0_\lambda)\middle| S_\gamma(C,T_1(Z))\right)\right)
\end{aligned}$$

where $B:=E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}q(T_1(Z);\beta^0_\lambda)S_\gamma(C|T_1(Z))'\right]=E\left[\frac{q(T_1(Z);\beta^0_\lambda)}{P(C\in\lambda)}\sum_{r\in\lambda}\frac{\partial}{\partial\gamma'}P(C=r|T_1(Z);\gamma^0)\right]$ ($=0$ if $\lambda=\mathcal{C}$). Therefore, $V_\lambda=V_{\lambda[pu]}=V_{\lambda[u]}$ if $\lambda=\mathcal{C}$. Otherwise $V_\lambda\leq V_{\lambda[pu]}\leq V_{\lambda[u]}$ in the matrix sense. (Proposition 4 is presented only for the purpose of this remark.) This ordering of the asymptotic variances shows that this well-known result for $R=2$ and $\lambda=\{1\}$, $\lambda=\{1,2\}$ from Chen et al. (2008) and Hahn (1998) also holds under CMAR for a generic R and a generic target (sub-)population λ.

3. For a generic $R>2$, CMAR is essentially an extreme dimension reduction assumption that helps to preserve the similarities of the results under planned and unplanned incompleteness. Of course, as we already saw, the forms of the efficient influence functions are different under these two cases; and there are still other important dissimilarities that we will note in the corollaries below.

However, all these dissimilarities are substantively mild compared to what would be the case under the general MAR condition in (1). Relaxing the planned incompleteness condition (2) and imposing restrictions on dimension reduction in MAR in (1), it follows from Proposition 1 in Barnwell and Chaudhuri (2018) that for sub-populations of interest such as $\lambda=\{a,a+1,\ldots,b\}$ where $a\in\{2,\ldots,b\}$ and $b\in\{2,\ldots,R\}$, the units from sub-samples $1,\ldots,a-1$ are *not at all* usable for efficient estimation.

By contrast, under planned incompleteness, units from all the sub-samples are usable for efficient estimation for any target λ irrespective of whether dimension reduction assumptions such as CMAR are allowed. In this sense, MAR in (1) properly nests all dimension reduction assumptions, with CMAR in (10) being the extreme one, only under planned incompleteness. On the other hand, in contrast to Remark 1 above, our Proposition 5 (under CMAR) is not a special case of Barnwell and Chaudhuri (2018) (under MAR) since neither imposes the planned incompleteness condition (2).

4. Lastly, note that if INDEP in (11) holds, then $P(C\in\lambda|T_R(Z))=P(C\in\lambda)$ for all λ. In this case, all the sub-populations are the same and hence there is only one population of interest $\lambda=\mathcal{C}$, for which our Proposition 1 (or Proposition 3 or 4 or 5, i.e., irrespective of (2)) implies that:

$$\varphi_{\lambda=\mathcal{C}}(O;\beta)=\varphi^{\text{INDEP}}(O;\beta):=q(T_1(Z);\beta)+\sum_{r=2}^{R}\frac{I(C\geq r)}{P(C\geq r)}\left[q(T_r(Z);\beta)-q(T_{r-1}(Z);\beta)\right]. \tag{14}$$

3.2 Efficiency gains from the existence of additional incomplete sub-samples

The efficiency gain for a generic target λ from using all the sub-samples instead of — (i) only the complete sub-sample, or (ii) the complete sub-sample and some but not all incomplete sub-samples — was evident from Remark 3 following Proposition 1 (and also from Proposition 2).[9] The underlying premise in that discussion is that all $R-1$ incomplete sub-samples exist and, hence, not using any sub-sample cannot be more beneficial (asymptotically) than using all the sub-samples.

The question that we ask in this subsection is different because it changes this premise. More precisely, we ask what is, if any, the benefit *from having an additional incomplete sub-sample*?

Care is required to avoid trivially positive answers by ensuring that the benefit is not entirely driven by the increase in sample size from the additional incomplete sub-sample, but rather incorporates the quality of information in this sub-sample that is actually relevant to the target population of interest (leading to zero benefits in certain cases). Accordingly, for a precise measure of "benefit", define the efficiency loss associated with the j-th element $\beta_{\lambda,j}$ from estimating β_λ based on a collection of sub-samples denoted by s instead of another collection of sub-samples denoted by s' as:

$$\mathrm{Loss}(\beta_{\lambda,j}; s, s') = \lim_{n\to\infty} \frac{\frac{1}{n_{\{s\}}}\mathrm{Avar}(\widehat{\beta}^{s}_{\lambda,j}) - \frac{1}{n_{\{s'\}}}\mathrm{Avar}(\widehat{\beta}^{s'}_{\lambda,j})}{\frac{1}{n_{\{s'\}}}\mathrm{Avar}(\widehat{\beta}^{s'}_{\lambda,j})} \quad \text{where } \lambda, s, s' \in \Lambda \text{ and } j = 1, \ldots, d. \tag{15}$$

$n_{\{l\}} := \sum_{r\in l} n_r = \sum_{r\in l}\sum_{i=1}^{n} I(C_i = r)$ is the size of the combined sub-samples in l for $l = s, s'$. $\widehat{\beta}^{l}_{\lambda,j}$ is the j-th element of $\widehat{\beta}^{l}_{\lambda}$ for $j = 1, \ldots, d$ and $l = s, s'$.

Crucially for the question posed here, $\widehat{\beta}^{l}_{\lambda}$ is the *efficient* estimator of β_λ based on the sub-samples in l. Hence, $\mathrm{Avar}(\widehat{\beta}^{l}_{\lambda,j})$ is the asymptotic variance ignoring the existence of the sub-samples not in l. For example, if $\lambda = \{1\}$ and $s = \{1, R\}$, then we replace $P(C = 1|T_1(Z))$ and $P(C = 1)$ in the result of Proposition 3 or 5 by $P(C = 1|T_1(Z), C \in \{1, R\})$ and $P(C = 1|C \in \{1, R\})$ respectively, as if only two sub-samples 1 and R exist (a substitution pattern as in multinomial/conditional logit).

Thus, the estimators not using all the sub-samples are not penalized for the sub-optimal use of the (available) information since they are actually efficient if the sub-samples they use were the only available sub-samples. Letting s be included in s', the loss in (15) thus reflects the usable incremental information brought in by the additional sub-samples that are included in s' but not in s.

Analytical expressions for this loss under INDEP in (11) and CMAR in (10) are intuitive, and are provided as corollaries to (14) and Proposition 3 in Corollaries 6 and 7. Analogous results without the planned incompleteness condition (2) are provided as corollary to Proposition 5 in Corollary 8.

[9] If the $(R-r+1)$-th sub-sample is not used then Remark 3 following Proposition 1 made it evident that the variance V_λ increases by $Var(\mathrm{term}_r)$ for $r = 2, \ldots, R$. Similar conclusions would follow from Propositions 3 and 5.

We take $R = 3$ and always include $\{R\}$ in s, s' for identification [see (4)]. Unless $\lambda = \mathcal{C}$, we include λ in s, s' as a convention. Unless $\lambda = \{3\}$, we do not consider $s = \{3\}$ for brevity (but do so in the Monte Carlo study in Section 5).[10]

For simplicity, let $d = 1$. For $l = s, s'$, let V_λ^l denote $Var(\varphi_\lambda(O; \beta_\lambda^0))$ when the latter is modified according to the discussion below (15). To avoid clutter, we write: $q(T_r(Z); \beta_\lambda^0)$ in (13) as q_r where λ is omitted from the latter but will be clear from the context; $T_r(Z)$ as T_r; $P(C = r)$ as p_r; $P(C = r|T_1(Z))$ as $p_r(T_1)$; $P(C \in \{r, t\})$ as p_{rt}; and $P(C \in \{r, t\}|T_1(Z))$ as $p_{rt}(T_1)$ for $r, t = 1, 2, 3$.

Corollary 6 *Let* (3), (11) *(i.e., INDEP) and assumption A hold. Under INDEP in* (11)*: (i)* β_λ^0 *is the same for all* $\lambda \in \Lambda$*, and (ii)* $p_r(T_1) = p_r$ *for all* $r = 1, \ldots, R$*. Thus, there is no distinction between planned versus unplanned incompleteness [see* (14)*], and hence* (2) *plays no role. Taking* $\lambda = \mathcal{C} := \{1, 2, 3\}$*, and assuming that the concerned variances exist, the following hold as* $n \to \infty$*:*

(a) $Loss(\beta_\lambda; s = \{3\}, s' = \{1,3\}) \times V_\lambda^{\{1,3\}} = \frac{p_1}{p_3} E\left[q_1^2\right]$.

(b) $Loss(\beta_\lambda; s = \{3\}, s' = \{2,3\}) \times V_\lambda^{\{2,3\}} = \frac{p_2}{p_3} E\left[q_2^2\right]$.

(c) $Loss(\beta_\lambda; s = \{1,3\}, s' = \{1,2,3\}) \times V_\lambda^{\{1,2,3\}} = \frac{p_2}{p_{13}} E\left[q_1^2\right] + \frac{p_2}{p_3 p_{23}} E\left[Var(q_2|T_1)\right]$.

(d) $Loss(\beta_\lambda; s = \{2,3\}, s' = \{1,2,3\}) \times V_\lambda^{\{1,2,3\}} = \frac{p_1}{p_{23}} E\left[q_1^2\right]$.

Corollary 7 *Let* (2), (3), (10) *(i.e., CMAR) and assumption A hold. Assuming that the concerned variances exist, the following hold as* $n \to \infty$*:*

(a) $Loss(\beta_{\{1\}}; s = \{1,3\}, s' = \{1,2,3\}) \times V_{\{1\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)p_2(T_1)}{p_1}\left\{\frac{q_1^2}{p_{13}(T_1)} + \frac{Var(q_2|T_1)}{p_3(T_1)p_{23}(T_1)}\right\}\middle| C = 1\right]$.

(b) $Loss(\beta_{\{2\}}; s = \{2,3\}, s' = \{1,2,3\}) \times V_{\{2\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)p_2(T_1)}{p_2 p_{23}(T_1)} q_1^2 \middle| C = 2\right]$.

(c1) $Loss(\beta_{\{3\}}; s = \{3\}, s' = \{1,3\}) \times V_{\{3\}}^{\{1,3\}} = E\left[\frac{p_{13}}{p_3} \frac{p_1(T_1)}{p_{13}(T_1)} q_1^2 \middle| C = 3\right]$.

(c2) $Loss(\beta_{\{3\}}; s = \{3\}, s' = \{2,3\}) \times V_{\{3\}}^{\{2,3\}} = E\left[\frac{p_{23}}{p_3} \frac{p_2(T_1)}{p_{23}(T_1)} q_2^2 \middle| C = 3\right]$.

(c3) $Loss(\beta_{\{3\}}; s = \{1,3\}, s' = \{1,2,3\}) \times V_{\{3\}}^{\{1,2,3\}} = E\left[\frac{p_2(T_1)p_3(T_1)}{p_3}\left\{\frac{q_1^2}{p_{13}(T_1)} + \frac{Var(q_2|T_1)}{p_3(T_1)p_{23}(T_1)}\right\}\middle| C = 3\right]$.

(c4) $Loss(\beta_{\{3\}}; s = \{2,3\}, s' = \{1,2,3\}) \times V_{\{3\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)p_3(T_1)}{p_3 p_{23}(T_1)} q_1^2 \middle| C = 3\right]$.

(d) $Loss(\beta_{\{1,3\}}; s = \{1,3\}, s' = \{1,2,3\}) \times V_{\{1,3\}}^{\{1,2,3\}} = E\left[\frac{p_2(T_1)p_{13}(T_1)}{p_{13}}\left\{\frac{q_1^2}{p_{13}(T_1)} + \frac{Var(q_2|T_1)}{p_3(T_1)p_{23}(T_1)}\right\}\middle| C \in \{1,3\}\right]$.

(e) $Loss(\beta_{\{2,3\}}; s = \{2,3\}, s' = \{1,2,3\}) \times V_{\{2,3\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)}{p_{23}(T_1)} q_1^2 \middle| C \in \{2,3\}\right]$.

(f1) $Loss(\beta_{\{1,2,3\}}; s = \{1,3\}, s' = \{1,2,3\}) \times V_{\{1,3\}}^{\{1,2,3\}} = E\left[\frac{p_2(T_1)}{p_{13}(T_1)} q_1^2 + \frac{p_2(T_1)}{p_3(T_1)p_{23}(T_1)} Var(q_2|T_1)\right]$.

(f2) $Loss(\beta_{\{1,2,3\}}; s = \{2,3\}, s' = \{1,2,3\}) \times V_{\{2,3\}}^{\{1,2,3\}} = E\left[\frac{p_1(T_1)}{p_{23}(T_1)} q_1^2\right]$.

[10] This is because the relevant comparison in such cases is rooted in the study of the sub-optimality of the asymptotic variance of standard IPW estimators that has already been studied extensively in the literature. On the other hand, our focus below is the comparison between two asymptotic variances each of which is optimal under its own assumption on the availability of the sub-samples as discussed in and around the definition of the loss in (15).

Remarks: Complementing the discussion on efficiency in Section 5 of Wooldridge (2007), let us note here that Corollaries 6 and 7 imply that there may not always be a loss in efficiency in the sense of (15) when one does not use the sub-samples that have been assumed in (15) to not exist (i.e., those in s' but not in s; see the discussion below (15)). For example, if $q_2 := E[m(Z;\beta^0_\lambda)|Z_{(1)}, Z_{(2)}] = 0$, then there is never any loss in all the above cases. Similarly, there is no loss in Corollary 6 (a), (d) and Corollary 7 (b), (c1), (c4), (e), (f2) under a weaker condition that $q_1 := E[m(Z;\beta^0_\lambda)|Z_{(1)}] = 0$.[11]

Corollary 8 *Let* (3), (10) *(i.e., CMAR) and assumption A hold, but* (2)*, i.e., planned incompleteness, does not hold. Assuming that the concerned variances exist, the following hold as* $n \to \infty$*:*

(a) $Loss(\beta_{\{1\}}; s = \{1,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{1\}} = E\left[\frac{p_1(T_1)p_2(T_1)}{p_1 p_3(T_1)p_{23}(T_1)} Var(q_2|T_1)\Big| C = 1\right]$.

(b) $Loss(\beta_{\{2\}}; s = \{2,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{2\}} = E\left[\frac{p_3(T_1)}{p_2 p_{23}(T_1)} Var(q_2|T_1)\Big| C = 2\right]$.

(c1) $Loss(\beta_{\{3\}}; s = \{3\}, s' = \{1,3\}) \times V^{\{1,3\}}_{\{3\}} = 0$.

(c2) $Loss(\beta_{\{3\}}; s = \{3\}, s' = \{2,3\}) \times V^{\{2,3\}}_{\{3\}} = 0$.

(c3) $Loss(\beta_{\{3\}}; s = \{3\}$ *or* $s = \{1,3\}$ *or* $s = \{2,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{3\}} = E\left[\frac{p_2(T_1)}{p_3 p_{23}(T_1)} Var(q_2|T_1)\Big| C = 3\right]$.

(d) $Loss(\beta_{\{1,3\}}; s = \{1,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{1,3\}} = E\left[\frac{p_2(T_1)p_{13}(T_1)}{p_{13}p_3(T_1)p_{23}(T_1)} Var(q_2|T_1)\Big| C \in \{1,3\}\right]$.

(e) $Loss(\beta_{\{2,3\}}; s = \{2,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{2,3\}} = 0$.

(f1) $Loss(\beta_{\{1,2,3\}}; s = \{1,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{1,3\}} = E\left[\frac{p_2(T_1)}{p_{13}(T_1)} q_1^2 + \frac{p_2(T_1)}{p_3(T_1)p_{23}(T_1)} Var(q_2|T_1)\right]$.

(f2) $Loss(\beta_{\{1,2,3\}}; s = \{2,3\}, s' = \{1,2,3\}) \times V^{\{1,2,3\}}_{\{2,3\}} = E\left[\frac{p_1(T_1)}{p_{23}(T_1)} q_1^2\right]$.

Remarks: First, it is not surprising that Corollaries 7 and 8 give identical results in (f1) and (f2) since, as evident from Propositions 3 and 5, there is no difference between planned and unplanned incompleteness when $\lambda = \mathcal{C}$. Second, the results in (a), (b), (c3) and (d) imply that leaving out incomplete sub-samples now results in zero loss under weaker conditions, i.e., $Var(q_2|T_1) = 0$ as opposed to the dual requirement of $Var(q_2|T_1) = 0$ and $q_1 := E[q_2|T_1] = 0$ under Corollary 7. Third, we note that such differences between planned and unplanned incompleteness can manifest more prominently if the additional sub-samples in s' that are not in s are of worse quality (in terms of the observability of the elements of Z) than each sub-sample in s. This is evident from comparing (c1), (c2) and (e) in Corollaries 7 and 8 respectively. Consequently, a comparison of (c1) or (c2) with (c3) in Corollary 8 shows that an identically zero loss when $R = 2$ need not imply the same when $R = 3$.

[11] A similar analysis of the loss in (15) with MAR in (1) under the premise of Section 5.2 is theoretically problematic. To see this, consider comparing the two cases $s = \{1,3\}$ and $s' = \{1,2,3\}$. For MAR in (1) to hold, a similar analysis demands $P(C = 3|Z) = 1 - P(C = 1|Z) = 1 - P(C = 1|Z_{(1)}) = P(C = 3|Z_{(1)})$ in the former case (i.e., if $\{2\}$ does not exist), whereas the latter case can still accommodate for $P(C = 3|Z) = P(C = 3|Z_{(1)}, Z_{(2)}) \neq P(C = 3|Z_{(1)})$ contradicting the requirement in the former. This obstructs our intended analytical comparison in a strict sense. Hence, such comparisons are not done here. Nevertheless, if one still proceeds with a similar, albeit theoretically problematic, comparison under MAR, then the way $P(C \in \lambda|T_R(Z))$ enters (5) and (6) [also see Remark 2 below Proposition 1] suggests that a zero loss under CMAR in Corollary 7 may not imply a zero loss under MAR.

4 Efficient Estimation

β^0_λ in (3) can be estimated efficiently by solving moment restrictions that depend on preliminary estimation of unknown nuisance parameters. In our case, these unknown nuisance parameters enter the moment restrictions in a way that leads to a certain orthogonality condition by virtue of which their estimation does not affect the asymptotic variance of the efficient estimator of β^0_λ; see, e.g., Andrews (1994) Newey and McFadden (1994), Chen et al. (2003), etc. We provide here a brief discussion of efficient estimation of β^0_λ by highlighting this orthogonality condition. Further details of efficient estimation, including its relation with Ai and Chen (2012), are presented in Appendix C.

To consolidate notation following Chen et al. (2003), and guided by (6), define a $d \times 1$ function:

$$g(O;\beta,h(\beta)) := \frac{I(C=R)}{P(C=R|T_R(Z))}\varphi_{R,\lambda}(O;\beta)+\sum_{r=1}^{R-1}\left[\frac{I(C\geq r)}{P(C\geq r|T_r(Z))}-\frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1}(Z))}\right]h_r(\beta) \tag{16}$$

where $h(\beta) = (h_1'(\beta), \ldots h_{R-1}'(\beta))'$ are the unknown nuisance parameters, and each $h_r(\beta)$ belongs to a class of functions $(Z,\beta) \mapsto \mathbb{R}^d$, call it $\mathcal{H}_r(\beta)$, for $r = 1, \ldots, R-1$. Let $\mathcal{H} := \{\mathcal{H}_1(\beta) \times \ldots \times \mathcal{H}_{R-1}(\beta) : \beta \in \mathcal{B}\}$ be a vector space endowed with a pseudo-metric $\|.\|_{\mathcal{H}}$, which is the sup-norm metric with respect to the argument β and a pseudo-metric with respect to the other arguments.

$g(O;\beta,h(\beta)) = \varphi_\lambda(O;\beta)$ defined in (6) if $h_r(\beta) = \varphi_{r,\lambda}(O;\beta)$ for $r = 1,\ldots,R-1$. Denote the true $h_r(\beta)$ as $h^0_r(\beta) := \varphi_{r,\lambda}(O;\beta)$ for $r = 1,\ldots,R-1$. While this suggests restricting $h_r(\beta)$ as $(T_r(Z),\beta) \mapsto \mathbb{R}^d$ for $r = 1,\ldots,R-1$, it turns out that letting $h_r(\beta)$ instead be a function of Z and β does not affect either consistency or asymptotic normality of the GMM estimator defined below.

In light of this discussion, now define the GMM average moment vector and its expectation as:

$$G_n(\beta,h(\beta)) := \frac{1}{n}\sum_{i=1}^n g(O_i;\beta,(h'_{1,i}(\beta),\ldots,h'_{R-1,i}(\beta))') \text{ and } G(\beta,h(\beta)) := E\left[G_n(\beta,h(\beta))\right].$$

Then, given any standard parametric or nonparametric estimator $\widehat{h}(\beta)$ for $h(\beta)$, the GMM estimator:

$$\widehat{\beta}_\lambda \text{ of } \beta^0_\lambda \text{ is defined as a (possibly approximate) solution of } G_n(\beta,\widehat{h}(\beta)) = 0. \tag{17}$$

The key feature of our setup is an orthogonality condition that can be represented as the following identity that for any $\beta \in \mathcal{B}$ and any $h(.) \in \mathcal{H}$ (that need not be $h(\beta)$):

$$G(\beta,h(.)) = E[\varphi_{R,\lambda}(O;\beta)] = E[m(Z;\beta)|C \in \lambda] \tag{18}$$

by (1), (4) and (16). That is, $G(\beta,h(.))$ does not depend on $h(.) \in \mathcal{H}$. Its main implications are:

(F1) $G(\beta^0_\lambda,h(.)) = 0$ for any $h(.) \in \mathcal{H}$ by also using (3). Also, for any $\beta \in \mathcal{B}$ and any $h(.),\bar{h}(.) \in \mathcal{H}$:
$$G(\beta,h(.)) - G(\beta^0_\lambda,\bar{h}(.)) = 0 \iff E[m(Z;\beta)|C\in\lambda] - E[m(Z;\beta^0_\lambda)|C\in\lambda] = 0 \iff \beta = \beta^0_\lambda.$$

(F2) The partial derivative of $G(\beta, h(\beta))$ with respect to β, denote it by $G_\beta(\beta, h(\beta))$, satisfies $G_\beta(\beta, h(\beta)) = M_\lambda(\beta) := \frac{\partial}{\partial \beta'} E\left[m(Z;\beta)|C \in \lambda\right]$, and it exists whenever $M_\lambda(\beta)$ exists.

(F3) $G(\beta, h(.)) - G(\beta, \bar{h}(.)) = 0$ for any $\beta \in \mathcal{B}$ and $h(.), \bar{h}(.) \in \mathcal{H}$. Thus, the pathwise derivative of $G(\beta, h(.))$ with respect to $h(.)$, denote it by $G_h(\beta, h(.))$, exists at all $h(.) \in \mathcal{H}$, in all directions $[\bar{h}(.) - h(.)]$ for $\{h(.) + \tau(\bar{h}(.) - h(.)) : \tau \in [0,1]\} \subset \mathcal{H}$, and satisfies $G_h(\beta, h(.))[\bar{h}(.) - h(.)] = 0$.

(F1) helps to verify the well-separability (of the true β) assumption for consistent estimation of β^0_λ by $\widehat{\beta}_\lambda$. It is even stronger since it indicates that $\widehat{h}(\beta)$ need not converge in probability to the true $h^0(\beta)$ but can converge to any $h^\dagger(\beta) \in$ interior($\mathcal{H}$) without affecting the consistency of $\widehat{\beta}_\lambda$ for β^0_λ [see Proposition 9]. (F2) simplifies the Jacobian formula in the asymptotic variance of $\widehat{\beta}_\lambda$ since it implies that $G_\beta(\beta^0_\lambda, h(\beta^0_\lambda)) = M_\lambda$. The final implication (F3) is stated in a way that helps to verify that the asymptotic variance of $\widehat{\beta}_\lambda$ is unaffected by the estimation of $h(\beta)$ even if $\widehat{h}(\beta)$ converges at a rate slower than $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(n^{-1/4})$; e.g., $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$ will suffice. It is worth noting that we do not even require that $h^\dagger(\beta) = h^0(\beta)$, the truth [see Proposition 10(i)]. Of course, semiparametric efficiency for $\widehat{\beta}_\lambda$ requires that $h^\dagger(\beta^0_\lambda) = h^0(\beta^0_\lambda)$, but the rate of convergence of the consistent $\widehat{h}(\beta)$ is still of no consequence as far as the first-order asymptotic properties of $\widehat{\beta}_\lambda$ are concerned [see Proposition 10(ii)]. Naturally, all these nice implications of (18) also provide flexibility in estimating the nuisance parameters – (i) parametrically based on misspecified models, e.g., giving linear projections rather than conditional expectations or (ii) nonparametrically under less than satisfactory conditions that might prevent a faster than $n^{1/4}$-rate convergence of the estimator. Based on this discussion above, Propositions 9 and 10 describe the asymptotic properties of $\widehat{\beta}_\lambda$ under high level conditions that are somewhat weaker than what one typically assumes in similar contexts.

For simplicity we follow Chen et al. (2003) and write $(\beta, h(\beta))$ as (β, h) unless confusing. Also, we write the Euclidean norm of a matrix/vector A as $\|A\| := \sqrt{\text{trace}(A'A)}$.

Proposition 9 *Let* (1), (3), *and assumptions (A1) and (A2) hold. Assume:*

(B1) $\|G_n(\widehat{\beta}_\lambda, \widehat{h})\| \leq \inf_{\beta \in \mathcal{B}} \|G_n(\beta, \widehat{h})\| + o_p(1)$ *where* $\mathcal{B}$ *is a compact subset of* $\mathbb{R}^d$*;*

(B2) $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$ *for some* $h^\dagger(\beta) \in$ *interior*$(\mathcal{H})$ *for all* β*, and where* $h^\dagger(\beta)$ *is not necessarily equal to* $h^0(\beta)$*;*

(B3) for all sequences of positive numbers $\{\delta_n\}$ *with* $\delta_n = o(1)$*,*

$$\sup_{\beta \in \mathcal{B}, \|h - h^\dagger(\beta)\|_{\mathcal{H}} \leq \delta_n} \frac{\|G_n(\beta, h) - G(\beta, h)\|}{1 + \|G_n(\beta, h)\| + \|G(\beta, h)\|} = o_p(1).$$

Then $\widehat{\beta}_\lambda - \beta^0_\lambda = o_p(1)$*.*

Proposition 10 *Let* (1), (3) *and assumptions A hold. Let* $\beta^0_\lambda \in interior(\mathcal{B})$ *and* $h^\dagger(\beta) \in interior(\mathcal{H})$ *for all* β*, and where* $h^\dagger(\beta)$ *is not necessarily equal to* $h^0(\beta)$*. For a small* $\delta > 0$ *define the neighborhoods* $\mathcal{B}_\delta := \{\beta \in \mathcal{B} : \|\beta - \beta^0_\lambda\| \leq \delta\}$ *and* $\mathcal{H}_\delta := \{h \in \mathcal{H} : \|h - h^\dagger(\beta)\|_{\mathcal{H}} \leq \delta\}$*. Let* $\widehat{\beta}_\lambda - \beta^0_\lambda = o_p(1)$ *and* $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$*. Assume:*

(C1) $\|G_n(\widehat{\beta}_\lambda, \widehat{h})\| \leq \inf_{\beta \in \mathcal{B}_\delta} \|G_n(\beta, \widehat{h})\| + o_p(n^{-1/2})$*;*

(C2) $G_\beta(\beta, h^\dagger)$ *exists for* $\beta \in \mathcal{B}_\delta$ *and is continuous at* $\beta = \beta^0_\lambda$*; (note that* $G_\beta(\beta^0_\lambda, h^\dagger)$ *is nonsingular by (A3) and (F2));*

(C3) for all sequences of positive numbers $\{\delta_n\}$ *with* $\delta_n = o(1)$*,*

$$\sup_{\beta \in \mathcal{B}_{\delta_n}, h \in \mathcal{H}_{\delta_n}} \frac{\|G_n(\beta, h) - G(\beta, h) - G_n(\beta^0_\lambda, h^\dagger)\|}{n^{-1/2} + \|G_n(\beta, h)\| + \|G(\beta, h)\|} = o_p(1);$$

(C4) $\sqrt{n} G_n(\beta^0_\lambda, h^\dagger) \xrightarrow{d} N(0, \Sigma)$ *where* $\Sigma := E\left[g(O; (\beta^0_\lambda, h^\dagger)) g(O; (\beta^0_\lambda, h^\dagger))'\right]$ *is finite.*

Then, the following results hold:

(i) $\sqrt{n}(\widehat{\beta}_\lambda - \beta^0_\lambda) = -M_\lambda^{-1} \sqrt{n} G_n(\beta^0_\lambda, h^\dagger) + o_p(1) \xrightarrow{d} N\left(0, M_\lambda^{-1} \Sigma M_\lambda^{-1'}\right)$*;*

(ii) if, additionally, $h^\dagger(\beta^0_\lambda) = h^0(\beta^0_\lambda)$ *then* $\Sigma = V_\lambda$ *as in Proposition 1 and:*

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta^0_\lambda) = -M_\lambda^{-1} \sqrt{n} G_n(\beta^0_\lambda, h^0) + o_p(1) \xrightarrow{d} N\left(0, \Omega_\lambda = M_\lambda^{-1} V_\lambda M_\lambda^{-1'}\right),$$

i.e., by Proposition 1, the estimator $\widehat{\beta}_\lambda$ *becomes semiparametrically efficient.*

Example: We conclude this section with a concrete example of the estimator $\widehat{\beta}_\lambda$ in (17). This example also informs our simulation study in Section 5. Consider a moment vector of the form $m(Z; \beta) = X(y - X'\beta)$. Let $R = 3$. For $i = 1, \ldots, n$, let $T_{ji} := T_j(Z_i)$ for $j = 1, 2, 3$, $a_{3i} := I(C_i = 3)/P(C = 3|T_{3i})$, $a_{2i} := I(C_i \geq 2)/P(C \geq 2|T_{2i}) - a_{3i}$, $a_{1i} := 1 - a_{2i} - a_{3i}$, $q := P(C \in \lambda|T_3(Z))$ and $q_i := P(C \in \lambda|T_{3i})$. Hence, ignoring the denominator $P(C \in \lambda)(\neq 0)$, (17) implies that $\widehat{\beta}_\lambda$ solves:

$$\begin{aligned} 0 &= \sum_{i=1}^n \Big(a_{3i} q_i X_i (y_i - X_i' \beta) + a_{2i} \widehat{E}[qX(y - X'\beta)|T_{2i}] + a_{1i} \widehat{E}[qX(y - X'\beta)|T_{1i}] \Big) \\ \Rightarrow \widehat{\beta}_\lambda &= \left(\sum_{i=1}^n \Big\{ a_{3i} q_i X_i X_i' + a_{2i} \widehat{E}\left[qXX'|T_{2i}\right] + a_{1i} \widehat{E}\left[qXX'|T_{1i}\right] \Big\} \right)^{-1} \\ &\quad \times \sum_{i=1}^n \Big\{ a_{3i} q_i X_i y_i + a_{2i} \widehat{E}\left[qXy|T_{2i}\right] + a_{1i} \widehat{E}\left[qXy|T_{1i}\right] \Big\} \end{aligned}$$

where $\widehat{E}$ denotes the estimated conditional expectations giving the estimated nuisance parameters as $\widehat{h}_2(\beta) := \widehat{E}[qX(y - X'\beta)|T_2]$ and $\widehat{h}_1(\beta) := \widehat{E}[qX(y - X'\beta)|T_1]$. Appendix C.4 provides a one-step updating efficient estimator that is useful when, unlike here, a closed form for $\widehat{\beta}_\lambda$ is unavailable.

5 Simulation Study

Now we numerically study the benefit, if any, of using all the sub-samples for efficient estimation of β_λ by estimating (15) in a standard linear regression setup using a small scale Monte Carlo experiment.

The following observation motivates our experiment. In their editorial introduction, McKenzie and Rosenzweig (2012) note how the different measurements of the same variables (e.g., consumption) can dramatically alter the conclusion of analyses using survey data. But, the "good" measures can be substantially more expensive.[12] Hence, under budget constraint, one could obtain the good but expensive measures for only subsets of units, and the other measures for larger subsets or everyone.

Accordingly, consider a linear regression of a random variable y on a constant and another random variable X. Let X_c and X_e be mismeasured X, possibly dependent also on y. Let $Z_{(1)} = (y, X_c')'$, $Z_{(2)} = X_e$ and $Z_{(3)} = X$, a data structure that can be justified if y and X_c ("c" for cheap) are cheap to observe, while X_e ("e" for expensive) is more expensive but still cheaper to observe than X (e.g., where X is true consumption). Now, defining $\vec{X} := (1, X)'$ and $\beta_\lambda := (\beta_{\lambda,1}, \beta_{\lambda,2})'$, our experiment involves efficient estimation of β_λ by taking the moment vector in (3) as $m(Z;\beta) = \vec{X}(y - \vec{X}'\beta)$.

The generic expression of the efficient estimator for any target λ relevant for this section is given at the end of Section 4. However, when using the collection of sub-samples in s, i.e., the nested collection, these estimators need to be adapted to the premise of the discussion in Section 3.2 [see below (15)]. Let us give an example to make this adaptation clear. Consider $\lambda = \{1\}$, $s' = \{1,2,3\}$ and $s = \{1,3\}$. Then, the efficient estimators using s' and s are as in (19) and (20) respectively.

$$\begin{aligned}\widehat{\beta}^{s'=\{1,2,3\}}_{\lambda=\{1\}} &= \left(\sum_{i=1}^{n} q_i \left\{a_{3i}\vec{X}_i\vec{X}_i' + a_{2i}\widehat{E}\left[\vec{X}\vec{X}'|T_2(Z_i)\right] + (1-a_{2i}-a_{3i})\widehat{E}\left[\vec{X}\vec{X}'|T_1(Z_i)\right]\right\}\right)^{-1} \\ &\quad \times \sum_{i=1}^{n} q_i \left\{a_{3i}\vec{X}_i + a_{2i}\widehat{E}\left[\vec{X}|T_2(Z_i)\right] + (1-a_{2i}-a_{3i})\widehat{E}\left[\vec{X}|T_1(Z_i)\right]\right\} y_i \qquad (19)\end{aligned}$$

where $\widehat{E}[.|.]$ denotes the estimated conditional expectation.[13] Under MAR, $a_{3i} := I(C_i = 3)/P(C = 3|T_2(Z_i))$, $a_{2i} := [1 - I(C_i = 1)]/[1 - P(C = 1|T_1(Z_i))] - a_{3i}$ and $q_i := P(C = \lambda|T_3(Z_i) = P(C = 1|T_1(Z_i))$ for $i = 1, \ldots, n$. The only difference under CMAR is that $a_{3i} := I(C_i = 3)/P(C = 3|T_1(Z_i))$

[12] For example, collecting consumption data through maintaining a personal diary can be 6 to 10 times more expensive than a 7-day recall [see the last 3 rows of column 6 in Table 10 of Beegle et al. (2012)]. On the other hand, the 7-day recall with a short list of aggregated consumption items can understate food consumption by 30% as compared to personal diaries [compare rows 4 and 8 of column 2 in Table 2 of Beegle et al. (2012)]. These figures are based on eight different measures of consumption, with accuracy varying inversely with cost, collected for non-monotonic sub-samples of 4029 sample units in Tanzania. For simplicity, and similarity with Section 3.2, we will consider (i) only three different measures and (ii) monotonic sub-samples. (ii) helps us to focus on efficient estimation and avoid ad hoc comparisons.

[13] In our simulation experiment below, these $\widehat{E}[.|T_r(Z)]$'s are series estimators for which we always use cubic polynomials of the elements of $(1, T_r(Z)')'$ *irrespective* of the sample size n. Hence, the resulting estimators for the parameters of interest β_λ's can alternatively be considered parametric in the sense of Ackerberg et al. (2012).

for $i = 1, \ldots, n$. Under INDEP, $a_{3i} := I(C_i = 3)/P(C = 3)$, $a_{2i} := I(C_i \in \{2, 3\})/P(C \in \{2, 3\}) - a_{3i}$, whereas $q_i := P(C = 1)$ (a constant, which cancels out making λ moot) for $i = 1, \ldots, n$. ((19) is simplified from the generic expression in Illustration 1 in Appendix C.5 that provides further details.)

$$\widehat{\beta}^{s=\{1,3\}}_{\lambda=\{1\}} = \left(\sum_{i=1}^{n} q_i \left\{ a_{3i} \vec{X}_i \vec{X}_i' + (1 - a_{3i}) \widehat{E} \left[\vec{X} \vec{X}' | T_1(Z_i) \right] \right\} \right)^{-1} \sum_{i=1}^{n} q_i \left\{ a_{3i} \vec{X}_i + (1 - a_{3i}) \widehat{E} \left[\vec{X} | T_1(Z_i) \right] \right\} y_i. \tag{20}$$

$a_{3i} := I(C_i = 3)/P(C = 3|C \in \{1, 3\}, T_1(Z_i))$ and $q_i := P(C = 1|C \in \{1, 3\}, T_1(Z_i))$ for $i = 1, \ldots, n$ under MAR and CMAR. While this conditioning does not affect the terms with $q_i a_{3i}$, this does affect the terms with $q_i(1 - a_{3i})$. Under INDEP, $a_{3i} := I(C_i = 3)/P(C = 3|C \in \{1, 3\})$ for $i = 1, \ldots, n$ (q_i is moot as before). Conditioning on the event $C \in \{1, 3\}$ adapts (20) to the premise of Section 3.2.

5.1 Simulation Design

We draw n i.i.d. copies of the concerned variables $Z = (y, X, X_c, X_e)'$ defined as follows.

$$y_i = \alpha + \delta X_i + \epsilon_i, \quad X_{ci} = X_i + I(y_i > 0)\sqrt{2}\epsilon_{ci}, \quad X_{ei} = X_i + I(y_i > 0)\epsilon_{ei}$$

where $\epsilon_i, \epsilon_{ci}, \epsilon_{ei}, X_i$ are mutually independent and i.i.d. $N(0, 1)$ for all $i = 1, \ldots, n$. We take $\alpha = \delta = 1$. While $E[X] = E[X_e] = E[X_c]$, X_e is relatively less variable than X_c as a measure of X because the former has smaller variance when $y > 0$ (precisely, $Var(X_c) - Var(X_e) = P(y > 0) = \Phi(1/\sqrt{2})$). The specification is arbitrary except only to motivate the data structure from a multi-phase sampling design — $Z_{(1)} = (y, X_c)'$, $Z_{(2)} = X_e$ and $Z_{(3)} = X$ — because the relative accuracy of X_e helps to envision X_e as more expensive to observe than X_c but less expensive to observe than X.

In the context of dependent sampling, without regard to the optimality of the sampling design, we generate the variable $C_i \in \mathcal{C} := \{1, 2, 3\}$ for $i = 1, \ldots, n$ as i.i.d. copies of C such that:

$$P(C = 1|Z_i) = F_{t_1}(\gamma_c(X_{ci} + y_i - 1)), \quad P(C = 2|C_i \geq 1, Z_i) = 1 - F_{t_1}(\gamma_e X_{ei} + \gamma_c(X_{ci} + y_i - 2)),$$

and $P(C = 3|Z_i) = 1 - P(C = 1|Z_i) - P(C = 2|Z_i)$ where $F_{t_1}(a)$ is the cumulative distribution function of a t_1-distributed random variable evaluated at $a \in \mathbb{R}$. (The fat tail of the t_1 distribution helps to partly offset problems with limited overlap [see assumption (A2) and Chaudhuri and Hill (2016)].) We design the selection mechanisms MAR in (1) and CMAR in (10) by taking $\gamma_c = \gamma_e = .25$ and $\gamma_c = .25, \gamma_e = 0$ respectively. Although $\gamma_c = \gamma_e = 0$ gives INDEP in (11), this hinders comparability with MAR and CMAR since this results in $P(C = 2) = P(C = 3) = .25$ while MAR and CMAR give $P(C=2) \approx n_2/n \approx .31$ and $P(C=3) \approx n_3/n \approx .19$ (obtained as average over 10,000 Monte Carlo trials). Hence, we directly design INDEP as $(n_1, n_2, n_3) \sim \text{Trinomial}(n, .5, .31)$.

The sub-samples are made incomplete by deleting X_i if $C_i \neq 3$ and X_{ei} if $C_i = 1$ for $i = 1, \ldots, n$. We take $n = 600, 1200, 1800$.

The true value of the parameters of interest β_1 (Intercept) and β_2 (Slope) is $(1,1)'$, i.e., $(\alpha, \delta)'$, under INDEP. The same holds under CMAR and MAR when $\lambda = \{1,2,3\}$. However, it is difficult to analytically obtain the true values under CMAR and MAR when $\lambda \neq \{1,2,3\}$. Since the study of bias is not our focus, we take the values listed in Table 1 as (roughly) the truth for the other β_λ's.

Target	CMAR Sampling					MAR Sampling				
λ	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,3\}$	$\{2,3\}$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,3\}$	$\{2,3\}$
Intercept	1.1375	0.7602	1.0087	1.1006	0.8652	1.1375	0.7624	0.9991	1.0985	0.8652
Slope	0.9630	0.9318	0.9562	0.9675	0.9685	0.9630	0.9239	0.9473	0.9628	0.9685

Table 1: Obtained as averages over 10,000 Monte Carlo trials of ordinary least squares estimates of Intercept and Slope from the regression of y on X using the correct (infeasible) sub-sample ($s = \lambda$) when $n = 1$ million.

5.2 Simulation Results

Tables 2 and 3 list the estimated loss (in percent) defined in (15) for various s with respect to $s' = \{1,2,3\}$ under INDEP, and CMAR and MAR respectively. Although, this particular demonstration may not be strictly correct theoretically under MAR [see footnote 11], we nevertheless report the MAR results to get a sense of the concerned loss. Incidentally, here the losses under MAR turn out to be quite close to those under CMAR, and hence are not given special attention.

Target Popln.	Used Sample	INDEP Sampling					
		Intercept			Slope		
λ	s	$n=600$	$n=1200$	$n=1800$	$n=600$	$n=1200$	$n=1800$
$\{1,2,3\}$	$\{3\}$	155	159	157	104	107	104
$\{1,2,3\}$	$\{1,3\}$	30	32	33	21	24	23
$\{1,2,3\}$	$\{2,3\}$	33	34	33	23	23	21

Table 2: Estimated Loss$(\beta_{\lambda,j}; s, s' = \{1,2,3\})$ (in percent) defined in (15) for $j = 1$ (Intercept) and $j = 2$ (Slope). Results are based on the analytically estimated Avar averaged over 10,000 Monte Carlo trials.

If all the sub-samples contained the same variables then these losses should more or less reflect the smaller than n size of the collection of sub-samples in s. For example, the first row of Table 2 would be $100 \times (1/n_3 - 1/n)/(1/n) \approx 100 \times (1/P(C=3) - 1) \approx 426$, and similarly the second and third rows would be approximately 45 and 100 respectively. The actual loss will invariably be much smaller in the first and third rows because the units in the additional sub-samples in $s' = \{1,2,3\}$ that are not in $s = \{3\}$ and $s = \{2,3\}$, i.e., the sub-samples $\{1,2\}$ and $\{1\}$ respectively, are uniformly worse in terms of their information content than those in s. This is however not true for the second

Target Popln. λ	Used Sample s	CMAR Sampling: Intercept, n = 600	1200	1800	CMAR Sampling: Slope, n = 600	1200	1800	MAR Sampling: Intercept, n = 600	1200	1800	MAR Sampling: Slope, n = 600	1200	1800
$\{1,2,3\}$	$\{3\}$	156	160	159	123	128	125	165	168	167	134	139	136
$\{1,2,3\}$	$\{1,3\}$	37	38	39	43	41	39	40	42	42	47	46	44
$\{1,2,3\}$	$\{2,3\}$	47	44	43	47	45	44	49	45	43	54	48	48
$\{1\}$	$\{3\}$	126	129	127	103	110	107	134	135	133	105	111	109
$\{1\}$	$\{1,3\}$	24	26	26	17	18	17	29	31	30	18	20	19
$\{2\}$	$\{3\}$	168	174	173	120	136	135	165	174	172	139	152	151
$\{2\}$	$\{2,3\}$	24	25	25	1	5	5	22	23	21	2	4	4
$\{3\}$	$\{3\}$	151	156	155	102	107	104	148	155	153	95	101	99
$\{3\}$	$\{1,3\}$	35	37	37	32	32	30	33	36	36	25	26	25
$\{3\}$	$\{2,3\}$	42	41	41	35	33	32	41	41	40	40	36	34
$\{1,3\}$	$\{3\}$	134	137	136	105	110	108	140	142	140	104	110	109
$\{1,3\}$	$\{1,3\}$	28	29	30	21	21	20	30	32	32	20	22	21
$\{2,3\}$	$\{3\}$	170	176	175	123	134	132	176	184	184	142	151	149
$\{2,3\}$	$\{2,3\}$	35	35	35	14	16	16	36	36	36	16	17	17

Table 3: Estimated Loss$(\beta_{\lambda,j}; s, s' = \{1,2,3\})$ (in percent) defined in (15) for $j = 1$ (Intercept) and $j = 2$ (Slope). Results are based on the analytically estimated Avar averaged over 10,000 Monte Carlo trials.

row since the extra sub-sample in s' is $\{2\}$, and a unit in it is more informative than a unit in the sub-sample $\{1\}$ but less so than a unit in the other sub-sample $\{3\}$ in s. Thus, it is not clear a priori in this case, i.e., $s = \{1,3\}$, if the actual loss will also be much smaller. All these intuitions are reflected in the tables, not only for INDEP (Table 2) but also for CMAR and MAR (Table 3).

There are cases like $\lambda = \{2\}$, $s = \{2,3\}$ under CMAR and MAR sampling where the loss for the Slope estimator is minimal and close to zero, and this is in spite of the fact that the estimator based on $s = \{2,3\}$ uses roughly half the number of observations used by the estimator based on $s' = \{1,2,3\}$. The loss can, however, be quite substantial in many cases and would be even larger if we had not restricted the definition of loss in (15) from penalizing the sub-optimal use of information by the sub-samples in s. Taken together, these simulation results show the obvious benefit of using all the sub-samples for estimation. Furthermore, comparing the first three (i.e., the only comparable) rows of Tables 2 and 3, it is evident that such benefits could be more under dependent sampling.

Appendix C.6 reports simulation evidence of reasonably good finite-sample properties of the efficient estimator used here under all the cases considered.[14] This lends credibility to the above simulation results on efficiency loss. In turn, the simulation results help to appreciate our detailed analytical exposition of efficiency gain/loss (from Section 3.2) by quantifying them numerically.

[14]These finite-sample properties and the magnitude of the losses remain stable when the same experiment is conducted with sample sizes such as $n = 1000, 2000, 5000$ (in older versions). For smaller sample sizes such as $n = 400$, the results for certain sub-populations, however, fluctuate without tail trimming. (The complete sub-sample is quite small, i.e., $n_3 \approx 76$, when $n = 400$.) While the use of adaptive negligible trimming to address the associated problem of limited overlap in related scenarios is a topic of our ongoing research [see Chaudhuri and Hill (2016) for estimation of effects of binary treatments], to our knowledge, a rigorous trimming-strategy with proper bias-correction is yet to be developed for more involved cases such as that in our present paper. Hence, no trimming is done in our experiment.

6 Conclusion

Our paper provided a thorough discussion of efficient estimation based on planned incomplete data obtained by sampling designs that generate monotonic incompleteness in the sample. Similar cost-effective sampling designs are widely used in biostatistics, epidemiology, psychology and statistics. Recently, such designs have also found use in the laboratory and field experiments in economics.

Generalizing the related literature, our theoretical framework allowed for a general $R \geq 2$ level of incompleteness in the data and a dynamically updating sampling design to provide more flexibility in designing and implementing the survey. However, similar to the related literature, our paper is actually silent about the optimality of the sampling design itself. A general optimality theory for planned incomplete designs turns out to be a difficult problem, and is a topic of our ongoing work.

Our results on efficient estimation are important because planned incompleteness, by definition, entails loss in information that makes it imperative that estimators based on such data use all available information in the data optimally. By obtaining the efficiency bound for estimation of a variety of target parameters, we established what "optimal" means in such contexts. Subsequently, we provided a thorough demonstration of how the sample units, that differ in their information content due to the planned incompleteness, contribute toward this optimality, and contrasted the characteristics and implications of this contribution from that found in the related literature.

The estimator presented in our paper makes the optimal use of available information in a direct way because we designed it based on the efficient influence function that we had obtained for the concerned target parameter. This estimator turned out to be a simple two-step estimator based on preliminary estimation of nuisance parameters, and its consistency and asymptotic normality were shown to hold under weaker than standard conditions.

We hope that our presentation of the analytical and simulation evidence of the benefits from this optimal use of information, and the simplicity of the underlying efficient estimation would encourage further research to facilitate the adoption of planned incomplete surveys in the face of budget constraints.

References

Abrevaya, J. and Donald, S. G. (2017). A GMM approach for dealing with missing data on regressors and instruments. Forthcoming in Review of Economics and Statistics.

Ackerberg, D., Chen, X., and Hahn, J. (2012). A Practical Asymptotic Variance Estimator For Two-Step Semiparametric Estimators. *The Review of Economics and Statistics*, 94: 481–498.

Ai, C. and Chen, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170: 442–457.

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

Ashraf, N., Berry, J., and Shapiro, J. M. (2010). Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia. *American Economic Review*, 100: 2383–2413.

Ashraf, N., Field, E., and Lee, J. (2014). Household Bargaining and Excess Fertility: An Experimental Study in Zambia. *American Economic Review*, 104: 2210–2237.

Barnwell, J. L. and Chaudhuri, S. (2018). Efficient estimation in sub and full populations with monotonically missing at random data. Technical report, McGill University.

Beaman, L., Karlan, D., Thusbaert, B., and Udry, C. (2015). Self-Selection into Credit Markets: Evidence from Agriculture in Mali. Mimeo.

Beegle, K., Weerdt, J. D., Friedman, J., and Gibson, J. (2012). Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics*, 98: 3 – 18.

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

Carroll, R., Ruppert, D., and Stefanski, L. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, New York.

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

Chamberlain, G. (1992). Comment: Sequential Moment Restrictions In Panel Data. *Journal of Business and Economic Statistics*, 10: 20–26.

Chatterjee, N. and Li, Y. (2010). Inference in Semiparametric Regression Models Under Partial Questionnaire Design and Nonmonotone Missing Data. *Journal of the American Statistical Association*, pages 787 – 797.

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

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

Chaudhuri, S. and Hill, J. B. (2016). Heavy Tail Robust Estimation and Inference for Average Treatment Effect. Technical report, University of North Carolina, Chapel Hill.

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.

Chen, X., Linton, O., and van Keilegom, I. (2003). Estimation of Semiparametric Models when the Criteria Function is not Smooth. *Econometrica*, 71: 1591–1608.

Dardanoni, V., Modica, S., and Peracchi, F. (2011). Regression with imputed covariates: A generalized missing-indicator approach. *Journal of Econometrics*, 162: 362–368.

Devereux, P. J. and Tripathi, G. (2009). Optimally combining censored and uncensored datasets. *Journal of Econometrics*, 151: 17–32.

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.

Graham, B. S., Pinto, C. C. D. X., and Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting. *Journal of Business and Economic Statistics*, 34: 288–301.

Graham, J. W., Hofer, S. M., and MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 31: 197–218.

Graham, J. W., Taylor, B. J., Olchowski, A. E., and Cumsille, P. E. (2006). Planned Missing Data Designs in Psychological Research. *Psychological Methods*, 11: 323–342.

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

Holcroft, C., Rotnitzky, A., and Robins, J. M. (1997). Efficient estimation of regression parameters from multistage studies with validation of outcome and covariates. *Journal of Statistical Planning and Inference*, 65: 349–374.

Holt, C. A. and Laury, S. K. (2002). Risk Aversion and Incentive Effects. *The American Economic Review*, 92: 1644–1655.

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

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

Lee, A. J., Scott, A. J., and Wild, C. J. (2012). Efficient estimation in multi-phase case-control studies. *Biometrika*, 97: 361–374.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. Wiley, Hoboken, NJ.

McKenzie, D. and Rosenzweig, M. (2012). Preface for symposium on measurement and survey design. *Journal of Development Economics*, 98: 1–2.

Muris, C. (2016). Efficient GMM Estimation with a General Missing Data Pattern. Technical report, Simon Frasier University.

Newey, W. K. and McFadden, D. L. (1994). Large Sample Estimation and Hypothesis Testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume IV, chapter 36, pages 2212–2245. Elsevier Science Publisher.

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

Raghunathan, T. E. and Grizzle, J. E. (1995). A Split Questionnaire Survey Design. *Journal of the American Statistical Association*, pages 54 – 63.

Reilly, M. (1996). Optimal Sampling Strategies for Two-Stage Studies. *American Journal of Epidemiology*, 143: 92–100.

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 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.

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

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

Shoemaker, D. M. (1973). *Principles and Procedures of Multiple Matrix Sampling.* Cambridge, MA: Ballinger.

Thornton, R. L. (2008). The Demand for, and Impact of, Learning HIV Status. *American Economic Review*, 98: 1829–1863.

Tripathi, G. (2009). Optimally combining censored and uncensored datasets. *Journal of Econometrics*, 151: 17–32.

Tripathi, G. (2011). Moment-based inference with stratified data. *Econometric Theory*, 27: 47–73.

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

Wacholder, S., Carroll, R. J., Pee, D., and Gail, M. H. (1994). The Partial Questionnaire Design For Case-Control Studies. *Statistics in Medicine*, 13: 623 – 634.

Whittemore, A. S. (1997). Multistage Sampling Designs and Estimating Equations. *Journal of Royal Statistical Society, Series B*, 59: 589–602.

Wooldridge, J. (1999). Asymptotic Properties of Weighted M-estimators for Variable Probability Samples. *Econometrica*, 69: 1385–1406.

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

Appendix: Proofs

The proofs are presented here in an abridged form, but all the details can be found in Appendix B for the proofs of the results for Sections 2 and 3, and in Appendix C for those for Section 4.

The proofs of Propositions 1, 3, 4 and 5 involve obtaining the semiparametric efficiency bound and the efficient influence function, under different assumptions, following Chen et al. (2008). They follow in two steps. Step 1 characterizes the tangent set for all regular parametric sub-models satisfying the semiparametric assumptions on the observed data. Step 2 obtains the efficient influence function and, thereby, the asymptotic variance lower bound as the expectation of its outer product. 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.

Proof of Proposition 1:

STEP 1: Consider a regular parametric sub-model indexed by a parameter θ for the distribution of the observed data $O = (C', T'_C(Z))'$. The log of the distribution can be expressed in terms of the full data $(C, Z')'$ as:

$$\log f_\theta(O) = \log f_\theta(Z_{(1)}) + \sum_{r=2}^{R} I(C \geq r) \log f_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)}) + \sum_{r=1}^{R} I(C = r) \log P(C = r|Z_{(1)}, \ldots, Z_{(r)}).$$

θ_0 is the unique value of θ such that $f_{\theta_0}(O)$ equals the true $f(O)$, and accordingly for all the quantities. The score function with respect to θ can then be written in terms of $(C, Z')'$ as:

$$S_\theta(O) = s_\theta(Z_{(1)}) + \sum_{r=2}^{R} I(C \geq r) s_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})$$

where $s_\theta(Z_{(1)}) := \frac{\partial}{\partial\theta} \log f_\theta(Z_{(1)})$ and $s_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)}) := \frac{\partial}{\partial\theta} \log f_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})$. (We will omit the subscript θ from the quantities evaluated at $\theta = \theta_0$.) The tangent set is the mean square closure of all d dimensional linear combinations of $S_\theta(O)$ for all such smooth parametric sub-models, and it takes the form:

$$\mathcal{T} := a_1(Z_{(1)}) + \sum_{r=2}^{R} I(C \geq r) a_r(Z_{(1)}, \ldots, Z_{(r)}), \tag{21}$$

where $a_1(Z_{(1)}) \in L_0^2(F(Z_{(1)}))$ and $a_r(Z_{(1)}, \ldots, Z_{(r)}) \in L_0^2(F(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)}))$.

STEP 2: Differentiating the moment conditions in (3) with respect to θ under the integral, and noting that $P(C \in \lambda|Z)$ (which is known) does not depend on θ but $P(C \in \lambda)$ (which is unknown) does, we obtain by using (1), (3) and assumption (A3) that:

$$\frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'} = -M_\lambda^{-1} E\left[m(Z; \beta_\lambda^0) \left\{ s(Z_{(1)})' + \sum_{r=2}^{R} s(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})' \right\} \middle| C \in \lambda \right].$$

Pathwise differentiability follows if we can find a $\psi(O) \in \mathcal{T}$ such that:

$$E[\psi(O)S(O)'] = \frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'}. \tag{22}$$

Let us conjecture that $\psi(O) = -M_\lambda^{-1}\varphi_\lambda(O;\beta_\lambda^0)$, and then verify (22) by equivalently showing that:

$$E[\varphi_\lambda(O;\beta_\lambda^0)S(O)'] = E\left[m(Z;\beta_\lambda^0)\left\{s(Z_{(1)})' + \sum_{r=2}^{R} s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})'\right\}\middle| C \in \lambda\right].$$

Consider the left hand side (LHS) and, in accordance with the partition of $\varphi_\lambda(O)$ (we work with the alternative specification in (7) for convenience), write it as $\sum_{q=1}^{R} B_q$ where, for $q = 2,\ldots,R$:

$$B_q := E\left[\frac{I(C \geq q)}{P(C \geq q|T_q(Z))}\left[\varphi_{q,\lambda}(O;\beta_\lambda^0) - \varphi_{q-1,\lambda}(O;\beta_\lambda^0)\right]S(O)'\right] \text{ while } B_1 := E\left[\varphi_{1,\lambda}(O;\beta_\lambda^0)S(O)'\right].$$

To avoid notational clutter, in the rest of STEP 2 we write $m(Z;\beta_\lambda^0)$ as m; $T_q(Z)$ as T_q; $\varphi_{q,\lambda}(O;\beta_\lambda^0)$ as $\varphi_{q,\lambda}$ for $q = 1,\ldots,R$; and also write $s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})$ as $s(Z_{(r)}|T_{r-1})$ for $r = 2,\ldots,R$.

First, we note that:

$$\begin{aligned} B_1 &= E\left[E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}m\middle|T_1\right]s(Z_{(1)})'\right] + \sum_{r=2}^{R} E\left[E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}m\middle|T_1\right]I(C\geq r)s(Z_{(r)}|T_{r-1})'\right] \\ &= E[ms(Z_{(1)})'|C\in\lambda] + 0 \end{aligned}$$

by using MAR in (1) and the fact that $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$ for $r = 2,\ldots,R$.

Now consider B_q. Since $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$, by using MAR in (1) we obtain that:

$$\begin{aligned} B_q &= \sum_{r=1}^{q-1} E\left[\frac{I(C\geq q)}{P(C\geq q|T_q)}(\varphi_{q,\lambda}-\varphi_{q-1,\lambda})s(Z_{(r)}|T_{r-1})'\right] + \sum_{r=q}^{R} E\left[\frac{I(C\geq r)}{P(C\geq q|T_q)}(\varphi_{q,\lambda}-\varphi_{q-1,\lambda})s(Z_{(r)}|T_{r-1})'\right] \\ &= 0 + E[ms(Z_{(q)}|T_{q-1})'|C\in\lambda] \end{aligned}$$

for $q = 2,\ldots,R$, and combining this with the B_1 obtained above verifies (22).

That $\psi(O) \in \mathcal{T}$ follows from matching terms as follows. (i) $-M_\lambda^{-1}\varphi_{1,\lambda}$ is a function of only T_1, and $E[\varphi_{1,\lambda}] = 0$ and, hence, satisfies the properties of $a_1(T_1)$ in (21). (ii) Using the definition of φ_r, the conditional on T_{r-1} expectation of the r-th term ($r = 2,\ldots,R$, without the multiplier $I(C \geq r)$) on the RHS of $\psi(O)$ is 0 by using (1). Therefore, this term, which is a function of only T_r, is also in $L_0^2(F(Z_{(r)}|T_{r-1}))$, and hence satisfies the properties of $a_r(T_r)$ in (21).

Therefore, we have now verified that the projection of the influence function $-M_\lambda^{-1}m(Z;\beta_\lambda^0)$ on to the tangent set $\mathcal{T}$ is $\psi(O) := -M_\lambda^{-1}\varphi_\lambda(O;\beta_\lambda^0)$. Hence, $\psi(O)$ is the efficient influence function and, therefore, the asymptotic variance lower bound is $E[\psi(O)\psi(O)'] = M_\lambda^{-1}V_\lambda M_\lambda^{-1'} =: \Omega_\lambda$. ■

Proof of Proposition 2:

(i) Let us start with $r = 1$, i.e., the residual from the projection, $\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\, \phi_{R-1}\right)$, inside the innermost parenthesis on the RHS. Then we apply induction arguments. For brevity, write $\varphi_{R,\lambda}(O;\beta)$ as $\varphi_{R,\lambda}$ and $T_r(Z)$ as T_r.

First, note that direct computation and MAR in (1) give:

$$\text{Proj}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\, \phi_{R-1}\right) = \left[\frac{I(C=R)}{P(C=R|T_R)} - \frac{I(C \geq R-1)}{P(C \geq R-1|T_{R-1})}\right] E[\varphi_{R,\lambda}|T_{r-1}],$$

which implies that:

$$\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\, \phi_{R-1}\right) = \frac{I(C=R)}{P(C=R|T_R)} \underbrace{\left(\varphi_{R,\lambda} - E[\varphi_{R,\lambda}|T_{R-1}]\right)} + \frac{I(C \geq R-1)}{P(C \geq R-1|T_{R-1})} E[\varphi_{R,\lambda}|T_{R-1}].$$

Consider the under-braced part in the RHS of the expression for $\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\, \phi_{R-1}\right)$. Using $T_{R-1} \setminus T_{R-2} = Z_{(R-1)}$ and (1), note that $E\left[\left(\varphi_{R,\lambda} - E[\varphi_{R,,\lambda}|T_{R-1}]\right)\phi_{R-2}|T_{R-2}\right]$ is a $d \times 1$ vector of zeros, and hence has no contribution in the successive projections. (Terms with no contribution in the successive projections are marked by under-braces in this proof.) On the other hand,

$$E\left[\frac{I(C \geq R-1)}{P(C \geq R-1|T_{R-1})} E[\varphi_{R,\lambda}|T_{R-1}]\phi_{R-2} \middle| T_{R-2}\right] = \frac{P(C = R-2|T_{R-2})}{P(C \geq R-2|T_{R-2})} E[\varphi_{R,\lambda}|T_{R-2}].$$

For our proof by induction, first assume that the following holds for a general $r \in \{2, \ldots, R-2\}$:

$$\begin{aligned}
&\overline{\text{Proj}}_{T_{R-r}}\left(\ldots \overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\, \phi_{R-1}\right)\ldots \middle|\, \phi_{R-r}\right) \\
= \;& \sum_{s=0}^{r-1} \frac{I(C \geq R-s)}{P(C \geq R-s|T_{R-s})} \underbrace{\left(E[\varphi_{R,\lambda}|T_{R-s}] - E[\varphi_{R,\lambda}|T_{R-s-1}]\right)} + \frac{I(C \geq R-r)}{P(C \geq R-r|T_{R-r})} E[\varphi_{R,\lambda}|T_{R-r}].
\end{aligned}$$

Now, once again using MAR in (1), we obtain that:

$$\begin{aligned}
&\left(E[\overline{\text{Proj}}_{T_{R-r}}\left(\ldots \overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\, \phi_{R-1}\right)\ldots \middle|\, \phi_{R-r}\right)\phi_{R-r-1}|T_{R-r-1}]\right)\left(E[\phi^2_{R-r-1}|T_{R-r-1}]\right)^{-1} \\
= \;& \left(\frac{P(C = R-r-1|T_{R-r-1})}{P(C \geq R-r-1|T_{R-r-1})} E[\varphi_{R,\lambda}|T_{R-r-1}]\right)\left(\frac{P(C \geq R-r-1|T_{R-r-1})}{P(C \geq R-r|T_{R-r})P(C = R-r-1|T_{R-r-1})}\right).
\end{aligned}$$

Hence, the proof follows by induction since the form is also valid for $r+1$, i.e.,

$$\begin{aligned}
&\overline{\text{Proj}}_{T_{R-r-1}}\left(\ldots \overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\, \phi_{R-1}\right)\ldots \middle|\, \phi_{R-r-1}\right) \\
= \;& \sum_{s=0}^{r} \frac{I(C \geq R-s)}{P(C \geq R-s|T_{R-s})} \left(E[\varphi_{R,\lambda}|T_{R-s}] - E[\varphi_{R,\lambda}|T_{R-s-1}]\right) + \frac{I(C \geq R-r-1)}{P(C \geq R-r-1|T_{R-r-1})} E[\varphi_{R,\lambda}|T_{R-r-1}].
\end{aligned}$$

(ii) The proof follows in the same way as that of Theorem 1 in Chamberlain (1992) or, more generally, as that of Theorem 1 of Ai and Chen (2012). ■

Proof of Proposition 3:

This proof follows in the same way as that of Proposition 1. The efficient influence function in this case turns out to be exactly the same as in Proposition 1 if CMAR is imposed on the latter. ■

We present the proofs of Propositions 4 and 5 in reverse order because the proof for the latter makes a reference to that for the former. Certain details of lesser importance are omitted below because they were already made explicit in the proof of Proposition 1.

To avoid notational clutter in these two proofs, when convenient, we write $m(Z;\beta_\lambda^0)$ as m; $T_r(Z)$ as T_r for $r=1,\dots,R$; and also write $s(Z_{(r)}|Z_{(1)},\dots,Z_{(r-1)})$ as $s(Z_{(r)}|T_{r-1})$ for $r=2,\dots,R$.

Proof of Proposition 5:

STEP 1: Consider a regular parametric sub-model indexed by θ for the joint distribution of the observed data $O=(C,T_C'(Z))'$. Because of CMAR in (10), the log of the distribution can be expressed in terms of the full data $(C,Z')'$ as:

$$\log f_\theta(O)=\sum_{r=1}^{R} I(C=r)\log P_\theta(C=r|Z_{(1)})+\sum_{r=1}^{R} I(C\geq r)\log f_\theta(Z_{(r)}|Z_{(1)},\dots,Z_{(r-1)})+\log f_\theta(Z_{(1)}).$$

(We use the same notation as before.) Then, the score function with respect to θ is:

$$S_\theta(O)=s_\theta(Z_{(1)})+\sum_{r=2}^{R} I(C\geq r)s_\theta(Z_{(r)}|Z_{(1)},\dots,Z_{(r-1)})+\sum_{r=1}^{R} I(C=r)\frac{\dot{P}_\theta(C=r|Z_{(1)})}{P_\theta(C=r|Z_{(1)})}$$

where $\dot{P}_\theta(C=r|Z_{(1)}):=\frac{\partial}{\partial\theta}P_\theta(C=r|Z_{(1)})$. The tangent space is characterized by functions of the form:

$$\mathcal{T}:=a_1(Z_{(1)})+\sum_{r=2}^{R} I(C\geq r)a_r(Z_{(1)},\dots,Z_{(r)})+\sum_{r=1}^{R} I(C=r)\frac{b_r(Z_{(1)})}{bb_r(Z_{(1)})}, \tag{23}$$

where $a_1(Z_{(1)})\in L_0^2(F(Z_{(1)}))$; $a_r(Z_{(1)},\dots,Z_{(r)})\in L_0^2(F(Z_{(r)}|Z_{(1)},\dots,Z_{(r-1)}))$ for $r=2,\dots,R$; $\sum_{r=1}^R b_r(Z_{(1)}))=0$, $\sum_{r=1}^R bb_r(Z_{(1)})=1$, and $\sum_{r=1}^R I(C=r)\frac{b_r(Z_{(1)})}{bb_r(Z_{(1)})}\in L_0^2(F(C|Z_{(1)}))$.

For a given $\lambda\in\Lambda$, the following relation obtained by two different factorization of the joint distribution of $(I(C\in\lambda),T_1(Z)\equiv Z_{(1)})$ helps us to switch between different factorizations:

$$\begin{aligned} & s(T_1)+I(C\in\lambda)\frac{\dot{P}(C\in\lambda|T_1)}{P(C\in\lambda|T_1)}+I(C\notin\lambda)\frac{\dot{P}(C\notin\lambda|T_1)}{P(C\notin\lambda|T_1)} \\ = \; & I(C\in\lambda)\left[\frac{\dot{P}(C\in\lambda)}{P(C\in\lambda)}+s(T_1|C\in\lambda)\right]+I(C\notin\lambda)\left[\frac{\dot{P}(C\notin\lambda)}{P(C\notin\lambda)}+s(T_1|C\notin\lambda)\right]. \end{aligned} \tag{24}$$

STEP 2: Differentiating (3) with respect to θ under the integral:

$$\frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'} = -M_\lambda^{-1} E\left[m \left\{ s(T_1|C\in\lambda)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})' \right\} \middle| C \in \lambda \right].$$

Then, as in the proof of Proposition 1, here we will need to correspondingly verify that:

$$E[\varphi_{\lambda[u]}^{\text{CMAR}}(O;\beta_\lambda^0) S(O)'] = E\left[m \left\{ s(T_1|C\in\lambda)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})' \right\} \middle| C \in \lambda \right]. \tag{25}$$

We do this term by term for $\varphi_{\lambda[u]}^{\text{CMAR}}(O;\beta_\lambda^0)$ and show equality of the terms on the LHS and RHS.

Consider the first term of $\varphi_{\lambda[u]}^{\text{CMAR}}(O;\beta_\lambda^0)$. Since $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$ for $r = 2, \ldots, R$, we can use (10) to take conditional expectations and, then using (24) to replace $s(T_1)$, write:

$$\begin{aligned}
E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)} E[m|T_1] S(O)'\right] &= E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)} E[m|T_1] \left\{ \frac{\dot{P}(C\in\lambda)}{P(C\in\lambda)} + s(T_1|C\in\lambda) - \frac{\dot{P}(C\in\lambda|T_1)}{P(C\in\lambda|T_1)} \right\}'\right] \\
&\quad + E\left[\frac{1}{P(C\in\lambda)} E[m|T_1] \dot{P}(C\in\lambda|T_1)'\right] \\
&= E\left[m|C\in\lambda\right] \frac{\dot{P}(C\in\lambda)'}{P(C\in\lambda)} + E\left[E[m|T_1] s(T_1|C\in\lambda)'|C\in\lambda\right] + 0 \\
&= 0 + E[ms(T_1|C\in\lambda)'|C\in\lambda] + 0
\end{aligned} \tag{26}$$

where the first zero in last line follows from (3). The second term follows by using (10) and noting that $E\left[E[m|T_1]s(T_1|C\in\lambda)'|C\in\lambda\right] = E\left[E[ms(T_1|C\in\lambda)'|T_1, C\in\lambda]|C\in\lambda\right] = E\left[ms(T_1|C\in\lambda)'|C\in\lambda\right]$.

Now consider the r-th term of $\varphi_{\lambda[u]}^{\text{CMAR}}(O;\beta_\lambda^0)$ for $r = 2, \ldots, R$. Taking expectation conditional on T_{r-1}, and using (10) and that $s(Z_{(s)}|T_{s-1}) \in L_0^2(F(Z_{(s)}|T_{s-1}))$ for $s = r, \ldots, R$, we obtain that:

$$\begin{aligned}
E\left[\frac{P(C\in\lambda|T_1)}{P(C\in\lambda)} (E[m|T_r] - E[m|T_{r-1}]) S(O)'\right] &= E\left[\frac{P(C\in\lambda|Z_1)}{P(C\in\lambda)} (E[m|T_r] - E[m|T_{r-1}]) \sum_{s=r}^{R} s(Z_{(s)}|T_{s-1})\right] \\
&= E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)} E[m|T_r] s(Z_{(r)}|T_{r-1})'\right] \\
&= E\left[ms(Z_{(r)}|T_{r-1})'|C\in\lambda\right].
\end{aligned} \tag{27}$$

Therefore, (26) and (27) verify (25). That $\varphi_{\lambda[u]}^{\text{CMAR}}(O;\beta_\lambda^0)$ belongs to $\mathcal{T}$ in (23) can be shown as follows. (i) Match the term $a(T_r)$ in $\mathcal{T}$ with the r-th term of $\varphi_{\lambda[u]}^{\text{CMAR}}(O;\beta_\lambda^0)$ for $r > 1$. (ii) Distribute the first term $s(Z_{(1)})$ in $\mathcal{T}$ according to the relation (24) and match the term $I(C\in\lambda)s(Z_{(1)}|C\in\lambda)$ with the first term of $\varphi_{\lambda[u]}^{\text{CMAR}}(O;\beta_\lambda^0)$ while keeping in mind that, by definition, $s(Z_{(1)}|C\in\lambda) \in L_0^2(F(Z_{(1)}|C\in\lambda))$. It is straightforward to verify that all the corresponding conditional expectations, as required by the definition in (23) and also (24), are zeros. The remaining terms in $\mathcal{T}$ (including the one due to distributing the terms in (ii)) are represented in $\varphi_{\lambda[u]}^{\text{CMAR}}(O;\beta_\lambda^0)$ by zeros. ■

Proof of Proposition 4:

The references in the steps of this proof are mainly to that of Proposition 3 (i.e., effectively to that of Proposition 1) and to that of Proposition 5.

As before, we obtain the score function for a parametric sub-model indexed by θ as:

$$S_\theta(O) = s_\theta(T_1) + \sum_{r=2}^{R} I(C \geq r) s_\theta(Z_{(r)}|T_{r-1}) + \sum_{r=1}^{R} \frac{I(C = r)}{P(C = r|T_1)} \left(\frac{\partial P(C = r|T_1; \gamma^0)}{\partial \gamma'} \frac{\partial \gamma^0}{\partial \theta'} \right)'.$$

Recall that $S_\gamma(C|T_1) := \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|T_1)} \frac{\partial}{\partial \gamma} P(C = r|T_1; \gamma^0)$. Let b denote constant matrices of dimension same as that of $\frac{\partial \gamma^0}{\partial \theta'}$. Then, the tangent set is characterized by the set of functions:

$$\mathcal{T} := a_1(T_1) + b' S_\gamma(C|T_1) + \sum_{r=2}^{R} I(C \geq r) a_r(T_r),$$

where $a_1(T_1) \in L_0^2(F(T_1))$, $S_\gamma(C|T_1) \in L_0^2(F(C|T_1))$ and $a_r(T_r) \in L_0^2(F(Z_{(r)}|T_{r-1}))$.

Recognizing that $P(C = r|T_1) = P(C = r|T_1; \gamma^0)$ is known up to the finite (d_γ) dimensional parameter γ, alters the relationship in (24) as follows:

$$\begin{aligned} & s(T_1) + \frac{\partial \gamma^{0\prime}}{\partial \theta} \left[I(C \in \lambda) \frac{\frac{\partial}{\partial \gamma} P(C \in \lambda|T_1; \gamma^0)}{P(C \in \lambda|T_1)} + I(C \notin \lambda) \frac{\frac{\partial}{\partial \gamma} P(C \notin \lambda|T_1; \gamma^0)}{P(C \notin \lambda|T_1)} \right] \\ = \;& I(C \in \lambda) \left[\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(T_1|C \in \lambda) \right] + I(C \notin \lambda) \left[\frac{\dot{P}(C \notin \lambda)}{P(C \notin \lambda)} + s(T_1|C \notin \lambda) \right]. \end{aligned}$$

As before, differentiating (3) (equivalently, (4)) with respect to θ, and using the above relationship gives:

$$\begin{aligned} & \frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'} \\ = \;& -M_\lambda^{-1} E\left[\frac{P(C \in \lambda|T_1)}{P(C \in \lambda)} m \left\{ s(T_1)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})' \right\} \right] - M_\lambda^{-1} E\left[E[m|T_1] \frac{\frac{\partial}{\partial \gamma'} P(C \in \lambda|T_1; \gamma^0)}{P(C \in \lambda)} \frac{\partial \gamma^0}{\partial \theta'} \right]. \end{aligned}$$

Therefore, utilizing the expression of the efficient influence function in Proposition 3 and its relation to that in Proposition 4, the verification of pathwise differentiability reduces to verifying that:

$$E\left[\Pi\left(\frac{I(C \in \lambda)}{P(C \in \lambda)} E[m|T_1)] \middle| S_\gamma(C|T_1) \right) S(O)' \right] = E\left[E[m|T_1] \frac{\frac{\partial}{\partial \gamma'} P(C \in \lambda|T_1; \gamma^0)}{P(C \in \lambda)} \frac{\partial \gamma^0}{\partial \theta'} \right].$$

Now, using that $E[S_\gamma(C|T_1)|T_1] = 0$, $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1})$ and CMAR in (1), it can be shown that $E\left[S_\gamma(C|T_1) \left\{ s(T_1)' + \sum_{r=2}^{R} I(C \geq r) s(Z_{(r)}|T_{r-1})' \right\} \right] = 0$. Subsequently, using the expressions for $S(O)$ and $S_\gamma(C|T_1)$, the verification of the above equality follows. (Certain steps that were omitted here for brevity may be less obvious. These are presented in detail in Appendix B.) ■

Proofs of Corollary 6, 7, 8: Straightforward but tedious manipulations of the results of Propositions 3 and 5 give Corollaries 7 and 8 respectively [see Chaudhuri (2014) for the proof of the latter]. Corollary 6 follows by imposing INDEP on the result of either Proposition 3 or Proposition 5. ■

Proof of Proposition 9: (F1) already implies the standard well-separability of β^0 by virtue of (3). Hence, for all $\delta > 0$ there exists $\epsilon(\delta) > 0$ such that $P(\|\widehat{\beta}_\lambda - \beta^0_\lambda\| > \delta) \leq P(\|G(\widehat{\beta}_\lambda, h^\dagger)\| \geq \epsilon(\delta))$.

Therefore, to establish that $\widehat{\beta}_\lambda \xrightarrow{P} \beta^0_\lambda$, it is sufficient to show that $\|G(\widehat{\beta}_\lambda, h^\dagger)\| = o_p(1)$. Assumption (B2) implies that $P(\widehat{h}(\beta) \in \mathcal{H}) \to 1$ uniformly in $\beta \in \mathcal{B}$ as $n \to \infty$. Hence, we work conditional on the event $\{\widehat{h}(\widehat{\beta}_\lambda) \in \mathcal{H}\}$, and, instead, show that $\|G(\widehat{\beta}_\lambda, h^\dagger)\| = o_p(1)$ conditional on $\{\widehat{h}(\widehat{\beta}_\lambda) \in \mathcal{H}\}$, since this would imply that $\|G(\widehat{\beta}_\lambda, h^\dagger)\| = o_p(1)$ unconditionally. To this end, first note that:

$$\begin{aligned}
\|G(\widehat{\beta}_\lambda, h^\dagger)\| &\leq \|G(\widehat{\beta}_\lambda, h^\dagger) - G(\widehat{\beta}_\lambda, \widehat{h})\| + \|G(\widehat{\beta}_\lambda, \widehat{h}) - G_n(\widehat{\beta}_\lambda, \widehat{h})\| + \|G_n(\widehat{\beta}_\lambda, \widehat{h})\| \\
&= \|G(\widehat{\beta}_\lambda, \widehat{h}) - G_n(\widehat{\beta}_\lambda, \widehat{h})\| + \|G_n(\widehat{\beta}_\lambda, \widehat{h})\|. \qquad (28)
\end{aligned}$$

The inequality holds by the triangle inequality (kept implicit hereafter). The equality holds by (F3).

Using (B3) and then (F3), we obtain that:

$$\|G(\widehat{\beta}_\lambda, \widehat{h}) - G_n(\widehat{\beta}_\lambda, \widehat{h})\| \leq o_p(1)\{1 + \|G_n(\widehat{\beta}_\lambda, \widehat{h})\| + \|G(\widehat{\beta}_\lambda, \widehat{h})\|\} \leq o_p(1)\{1 + \|G_n(\widehat{\beta}_\lambda, \widehat{h})\| + \|G(\widehat{\beta}_\lambda, h^\dagger)\|\}.$$

Using this along with (28) and (B1) gives:

$$\|G(\widehat{\beta}_\lambda, h^\dagger)\| \times (1 - o_p(1)) \leq o_p(1) + \|G_n(\widehat{\beta}_\lambda, \widehat{h})\| \times (1 + o_p(1)) \leq \inf_{\beta \in \mathcal{B}} \|G_n(\beta, \widehat{h})\|(1 + o_p(1)). \quad (29)$$

Now, as before, and by using (18), i.e., $G(\beta^0_\lambda, h) = 0$ for all $h \in \mathcal{H}$, note that:

$$\|G_n(\beta, \widehat{h})\| \leq \|G_n(\beta, \widehat{h}) - G(\beta, \widehat{h})\| + \|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| + \|G(\beta, h^\dagger) - G(\beta^0_\lambda, h^\dagger)\| \qquad (30)$$

where $\|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| = 0$ by (F3). Since $G(\beta^0_\lambda, h^\dagger) = 0$ and $\|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| = 0$, we have:

$$\|G_n(\beta, \widehat{h}) - G(\beta, \widehat{h})\| \leq o_p(1)\{1 + \|G_n(\beta, \widehat{h})\| + \|G(\beta, h^\dagger)\| + o_p(1)\} = o_p(1) + \|G_n(\beta, \widehat{h})\| \times o_p(1)$$

by using (B3). Therefore, (30) gives that:

$$\|G_n(\beta, \widehat{h})\| \leq o_p(1) + \|G_n(\beta, \widehat{h})\| \times o_p(1) + 0 + \|G(\beta, h^\dagger) - G(\beta^0_\lambda, h^\dagger)\| \times o_p(1).$$

Hence, $\|G_n(\beta, \widehat{h})\| \times (1 - o_p(1)) \leq o_p(1) + \|G(\beta, h^\dagger) - G(\beta^0_\lambda, h^\dagger)\| \times o_p(1)$ where all the $o_p(1)$ terms are uniform with respect to $\beta \in \mathcal{B}$. This implies that:

$$\inf_{\beta \in \mathcal{B}} \|G_n(\beta, \widehat{h})\| \leq \sup_{\beta \in \mathcal{B}} o_p(1) + \inf_{\beta \in \mathcal{B}} \|G(\beta, h^\dagger) - G(\beta^0_\lambda, h^\dagger)\| \times \sup_{\beta \in \mathcal{B}} o_p(1) = o_p(1)$$

since $\inf_{\beta\in\mathcal{B}} \|G(\beta, h^\dagger) - G(\beta^0_\lambda, h^\dagger)\| = 0$. Hence, if follows from (29) that $\|G(\widehat{\beta}_\lambda, h^\dagger)\| = o_p(1)$. ■

Proof of Proposition 10: (i) The proof consists of two parts. First, one shows the $\sqrt{n}$-consistency of $\widehat{\beta}_\lambda$, and then, using it, one shows the asymptotic normality of $\widehat{\beta}_\lambda$. Under the conditions of the proposition, $\sqrt{n}$-consistency of $\widehat{\beta}_\lambda$ follows if one shows that $\|G(\widehat{\beta}, h^\dagger)\| = O_p(n^{-1/2})$. This part of the proof is very similar to that in Chen et al. (2003) with minor adjustments made to it as we did in the proof of Proposition 9. Hence, we omit the first part and instead refer the reader to the same proof (under over-identification) for Proposition 14 in Appendix C where all the details are presented.

Subsequently, by taking the $\sqrt{n}$-consistency of $\widehat{\beta}_\lambda$ as given, we now proceed to show the asymptotic normality of $\widehat{\beta}_\lambda$. Define the linearization of $G_n(\beta^0_\lambda, h^\dagger)$ as $L_n(\beta) = G_n(\beta^0_\lambda, h^\dagger) + M_\lambda(\beta - \beta^0_\lambda)$. Note that the differences from the linearization in Chen et al. (2003) arise due to (F2) and (F3). This gives:

$$
\begin{aligned}
& \|G_n(\widehat{\beta}_\lambda, \widehat{h}) - L_n(\widehat{\beta}_\lambda)\| \\
= \; & \|G_n(\widehat{\beta}_\lambda, \widehat{h}) - G_n(\beta^0_\lambda, h^\dagger) - M_\lambda(\widehat{\beta}_\lambda - \beta^0_\lambda)\| \\
\leq \; & \|G_n(\widehat{\beta}_\lambda, \widehat{h}) - G_n(\beta^0_\lambda, h^\dagger) - G(\widehat{\beta}_\lambda, \widehat{h})\| + \|G(\widehat{\beta}_\lambda, \widehat{h}) - G(\widehat{\beta}_\lambda, h^\dagger)\| + \|G(\widehat{\beta}_\lambda, h^\dagger) - M_\lambda(\widehat{\beta}_\lambda - \beta^0_\lambda)\| \\
\leq \; & \|G_n(\widehat{\beta}_\lambda, \widehat{h}) - G_n(\beta^0_\lambda, h^\dagger) - G(\widehat{\beta}_\lambda, \widehat{h})\| + \|G(\widehat{\beta}_\lambda, h^\dagger) - M_\lambda(\widehat{\beta}_\lambda - \beta^0)\| \text{ [by (F3)]} \\
\leq \; & o_p(1) \times \{1 + \|G_n(\widehat{\beta}_\lambda, \widehat{h})\| + \|G(\widehat{\beta}_\lambda, \widehat{h})\|\} + \|G(\widehat{\beta}_\lambda, h^\dagger) - G(\beta^0_\lambda, h^\dagger) - M_\lambda(\widehat{\beta}_\lambda - \beta^0)\|
\end{aligned}
$$

where the term inside braces follows from (C3). Note that, the inclusion of $G(\beta^0_\lambda, h^\dagger)$ in the last term is innocuous since $G(\beta^0_\lambda, h^\dagger) = 0$. Now, by the definition of M_λ, assumptions (C2), (A3) and (F2), it follows that $\|G(\widehat{\beta}_\lambda, h^\dagger) - G(\beta^0_\lambda, h^\dagger) - M_\lambda(\widehat{\beta}_\lambda - \beta^0)\| = o_p(\|\widehat{\beta}_\lambda - \beta^0_\lambda\|)$, which is $o_p(n^{-1/2})$ since $\widehat{\beta}_\lambda - \beta^0_\lambda = O_p(n^{-1/2})$. On the other hand, one can also show that $\|G_n(\widehat{\beta}_\lambda, \widehat{h})\| \leq \inf_{\beta\in\mathcal{B}_\delta} \|G_n(\beta, \widehat{h})\| + o_p(n^{-1/2}) = O_p(n^{-1/2})$, for which all the details are given in the proof for Proposition 14 in Appendix C. Finally, since $\|G(\widehat{\beta}_\lambda, \widehat{h})\| \leq \|G(\widehat{\beta}_\lambda, \widehat{h}) - G(\widehat{\beta}_\lambda, h^\dagger)\| + \|G(\widehat{\beta}_\lambda, h^\dagger)\| = O_p(n^{-1/2})$ because the first term is 0 by (F3) and the second term is $O_p(n^{-1/2})$ from the first part of the proof, we obtain that $\|G_n(\widehat{\beta}_\lambda, \widehat{h}) - L_n(\widehat{\beta}_\lambda)\| \leq o_p(n^{-1/2})$. Similarly, for $\bar{\beta} := \arg\min_\beta \|L_n(\beta)\|$, that, by construction, satisfies $\sqrt{n}(\bar{\beta} - \beta^0_\lambda) = -M_\lambda^{-1}\sqrt{n}G_n(\beta^0_\lambda, h^\dagger)$, we can show that $\|G_n(\bar{\beta}, \widehat{h}) - L_n(\bar{\beta})\| \leq o_p(n^{-1/2})$. Now that the proximity of $G_n(\beta, \widehat{h})$ and $L_n(\beta)$ has been established at $\widehat{\beta}_\lambda$ and $\bar{\beta}$ respectively, the rest of the proof is to show that $\sqrt{n}(\bar{\beta} - \widehat{\beta}_\lambda) = o_p(1)$. As was the case in Chen et al. (2003), this does not involve anything particularly related to the key feature of our setup, and hence follows exactly in the same way as in the proof of Theorem 3.3 and Lemma 3.5 in Pakes and Pollard (1989).

(ii) This follows by noting that $g(O; \beta, h^0(O; \beta)) = \varphi_\lambda(O; \beta)$ defined in (6). ■

Online supplemental material for:

A Note on Efficiency Gains from Multiple Incomplete Sub-samples

Saraswata Chaudhuri[15]

This supplemental appendix contains three sections: Appendices A, B and C. Appendix A (A.1-A.9) contains clarifying or descriptive endnotes from Sections 1-3. Appendix B contains the detailed version of the proofs of the results in Sections 2 and 3 of our paper. Abridged versions of these proofs were presented in our paper. Appendix C (C.1-C.7) provides formal statements and their proofs for the asymptotic properties of the efficient estimator in Section 4. This presentation allows for overidentified models. Appendix C also reports simulation results describing the finite-sample properties of the efficient estimator in the context of the Monte Carlo experiment in Section 5.

Additionally, Appendix C describes a simple one-step updating of any $\sqrt{n}$-consistent estimator (e.g., IPW estimator) to obtain an estimator that is asymptotically equivalent to the efficient estimator. A sketch of the proof for this efficiency is provided under standard regularity conditions. This updating is computationally convenient and can be easily performed in standard statistical softwares. We provide two illustrations of the efficient estimator: (i) a linear regression as in Section 5 where a closed form efficient estimator is available (so, no updating is required), and (ii) a linear quantile regression where the updating is useful due to the unavailability of closed form expressions.

Index:

[15]Department of Economics, McGill University, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.

Appendix A: Descriptive endnotes

A.1 Planned incomplete design: examples from economics and other fields

Examples from other fields

The adoption of the planned incomplete survey design is common in other fields to the extent that there are even established terminologies to refer to the different types of planned incompleteness.

The two/many-measurement-design is used in psychology where it is common to encounter an expensive "gold standard" measure and other inexpensive but less accurate measures for behavioral traits [see, e.g., Graham et al. (2006)]. Then, the gold standard measure is typically employed only on a subset of the study subjects while the other measures are employed on all. In other contexts, planned missing waves for pre-selected sample units in a panel have been extensively used since MacArdle and Woodcock (1997) to cut the cost of estimation of key quantities in psychology.[16] In yet other contexts, the multiple matrix sampling of Shoemaker (1973), that requires most units to respond only to parts of the full survey questionnaire, was extended as the split-questionnaire design (SQD) by Raghunathan and Grizzle (1995) in statistics, as the partial questionnaire design (PQD) by Wacholder et al. (1994) in biostatistics and epidemiology, and as the multi-forms surveys discussed by Graham et al. (1996), Graham et al. (2006), and others in psychology and behavioral research.

Examples from economics

[16] While this example may appear less familiar than the other two types of examples, note that the structure of the sample due to missing waves is actually similar to that from rotating panels with a single rotation. Rotating panels such as the Current Population Survey are common in economics [see Nijman et al. (1991) for an influential study].

The common theme in all these references is the cost cutting of surveys, which also applies to the field of economics. This is even more relevant now as the use of primary data, often under tight budgets, gets more common among economists. However, in spite of the promising early work of DiNardo et al. (2006) who point to the benefits of planned incompleteness, systematic adoption of planned incompleteness seems nonexistent in economics. Ad hoc adoptions can be found in laboratory and field experiments, and we list below a small number of representative examples of both types.

(1) In a highly cited paper in experimental economics, Holt and Laury (2002) run a laboratory experiment to elicit risk aversion for studying its dependence on the size of the stake. The experiment involved planned incompleteness whereby the low-stake experiments were first run on all subjects (phase one) and then the high-stake experiments were run on subsets of these subjects.

(2) Field experiments also typically involve follow-up rounds. We provide three recent examples:

(2a) Thornton (2008) studies an experiment in rural Malawi where the subjects where tested for their HIV status and given incentives to learn the results from a nearby centre. After the respondents had a chance to learn about the result (some did not), a follow-up interview was conducted on 75% (so, 25% incompleteness by plan) of the original subjects to record their sexual behavior and their response to an offer to buy up to 5 packages of 3 condoms using the .30 USD that was paid to them.

(2b) Ashraf et al. (2010) run an experiment in Zambia to differentiate between the screening and sunk-cost effects measured by the usage of clorin (purchased from the experimenter) to purify drinking water. In the first phase (baseline), the experimenter measures, among other variables, the chemical concentration of clorin in the households' drinking water. In the second phase (marketing), the experimenter offers to sell a bottle of clorin to the concerned households at less than market price. In the third phase (follow-up), the experimenter again measures, among other things, the clorin concentration. The data are monotonic in terms of incompleteness — the third phase was conducted only on those households who could be reached in the second phase (planned incompleteness) and there was also high attrition, particularly, in the third phase (unplanned incompleteness).

(2c) Ashraf et al. (2014) run an experiment in Zambia to study household bargaining power in terms of eventual fertility and usage of contraceptives when women were given access to contraceptives in the presence and absence of their husbands. The first phase is a baseline survey on women that also provided them with information on contraception and prevention of STD, and distributed condoms. In the second phase (experiment) the respondents were reached either in the presence or absence of their husbands (reflecting two types of treatments) and vouchers for injectable contraceptives were provided. In the third phase (follow-up) information was collected on the women's use of contraceptives, sexual behavior, fertility, etc. Interestingly, beside a small number of rather balanced attrition (unplanned incompleteness), the monotonicity in this data resulted primarily from planned incompleteness because the second phase was conducted on a much smaller subset of the respondents from the first phase owing, in the authors' words, to "overwhelmingly...resource constraints on the part of the investigators and a strict timeline for completion of the study"/"Not enough budget".

Other types of planned incompleteness in economics

Another source of planned incompleteness (and eventual monotonicity) in Ashraf et al. (2014)'s data is the decision to collect new variables during the follow-up and an additional round but *only* in focus groups with subsets of participants. In other words, by design, both the original and the new variables are observed for a subset of units (those in focus groups) in the data, while only the original variables are observed for the remaining units in the data. Relatedly, there can be cases where such new variables might have less accurate counterparts in the original variables, making the former subset (in the last sentence) a validation sample. An example is Beaman et al. (2015) who use an input survey to obtain such data. An important consequence of this that we highlight in our paper is that the joint distribution of the more and less accurate variables that are jointly observed in the validation sample can often be useful for efficiency gains in subsequent estimation (although Beaman et al. (2015) did not need to exploit it). A similar example with more and less accurate measures of consumption, but unfortunately no joint observability (not needed for the stated purpose of their paper), is Beegle et al. (2012) [see our Section 5 for more on it]. This is also an example that does not involve a time dimension unlike the other references presented here.

Other types of cases where planned incompleteness could be useful include McKenzie (2012) and Allcott and Rogers (2014). Monotonicity is natural (at least, not unnatural) in both types of cases.

McKenzie (2012) draws on the clinical trial literature and provides an analysis of the benefit in precision gains from multiple follow-up measurements in field experiments over the standard practice of a single baseline and a single follow-up. His discussion focuses on the tradeoff in the choice of n (number of subjects) and T (number of measurements including baseline and follow-ups) at a given cost. Alternatively, one could keep both n and T large but measure the relevant variables only for a subset of subjects at each follow-up exactly like the prototypical multi-phase sampling.

Allcott and Rogers (2014) consider a treatment that was applied to subjects for varying duration. Specifically, the treatment was applied, i.e., a "home energy report" (containing personalized energy use, social comparisons, and energy conservation information) was sent to subjects, over a period of time but was discontinued (and not reinstated) for subsets of subjects during the tenure. The authors study the effect of this treatment on the energy consumption of the subjects. Note that, in such cases, the treatment administrator need not choose the subset of subjects "exogenously" but could conceivably incorporate the subjects' past responses to the treatment in the choice decision.

Relation with our framework

While the details of estimation vary, all the studies cited above involve estimating expectations and, sometimes, regression coefficients. For example, consider, without loss of generality, the instrumental variables (IV) regression in equation (2) (p. 1848) in Thornton (2008) (our Example (2a)) that was run on 75% of the full sample, namely, on the subjects from the districts of Rumphi and Balaka and not from Mchinji [see their Tables 6 and 7]. Assume in the spirit of Table 7 that the district-level heterogeneity is captured by the intercept, and extend this assumption to the full sample so that the regression continues to hold in the population of the full sample simply by adding a dummy D for Mchinji as a regressor. Denoting the instruments, endogenous regressors, exogenous regressors and dependent variable by W, X_1, X_2 and y respectively, define the (moment) function:

$$m(y, X_1, X_2, W; \beta_1, \beta_2) := (W', X_2')'(y - X_1\beta_1 - X_2\beta_2).$$

The planned incompleteness due to the selective follow-up here is a case of missing y. Now, while the coefficient of D (in X_2) is unidentified, the results in our paper imply that if interest lies in the population of all three districts then the optimal use of the full sample is possible using the modified moment vector: $\frac{(1-D)}{1-P(D=1)}m(y, X_1, X_2, W; \beta_1, \beta_2)+\left(1 - \frac{(1-D)}{1-P(D=1)}\right) E[m(y, X_1, X_2, W; \beta_1, \beta_2)|X_1, X_2, W]$ instead of $\frac{(1-D)}{1-P(D=1)}m(y, X_1, X_2, W; \beta_1, \beta_2)$ that is "close" to what was used in Table 7.[17,18] (Feasibility issues of the modified moment vector, which also arise in Example 1 below (Appendix A.2), are addressed in detail in the sequel and can be skipped for now in this introductory discussion.) Our paper explores such optimal uses of the sample for efficient estimation in more general contexts.

A.2 Planned incomplete design: examples of optimality of the design

Example 1: Minimizing variance of estimator subject to a given expected cost of survey

Let (Y, X) be scalar variables with finite means and variances. Let the parameter of interest be $\beta = E[Y - X]$. Consider two random samples $\mathcal{S}^\dagger = \{Y_j, X_j\}_{j=1}^{n^\dagger}$ and $\mathcal{S} = \{Y_i, D_i, D_iX_i\}_{i=1}^{n}$ where D is binary. We observe X in $\mathcal{S}$ only when $D = 1$. Assume that $P(D = 1|Y, X) = P(D = 1) = p$.[19]

[17] Standard IV conditions such as $E[m(y, X_1, X_2, W; \beta_1^0, \beta_2^0)|X_2, W] = 0$ or $E[m(y, X_1, X_2, W; \beta_1^0, \beta_2^0)] = 0$ do not imply that $E[m(y, X_1, X_2, W; \beta_1^0, \beta_2^0)|X_1, X_2, W] = 0$ where β_1^0 and β_2^0 are the true values of β_1 and β_2. Hence, the modification in the moment vector is not moot, and it reduces the variability of the estimating function for β_1 and β_2.

[18] We say "close" to mean asymptotically equivalent. Note that Tables 6 and 7 suggest that the first stage was run on the full sample since only y is missing, while the second stage was run on the sample where $D = 0$. While this gives more precise first stage estimates than what our latter representation above gives, under standard assumptions both approaches actually give asymptotically equivalent estimates of the parameters of interest β_1 and β_2 that, in turn, are less precise than what our former representation above with the modified moment vector does.

[19] While $n^\dagger$ and n are non-random quantities, we allow, here and throughout, D to be random. Hence $n_D := \sum_{i=1}^n D_i \sim Bin(n, p)$, i.e., the size of the complete sub-sample (the sub-sample containing all the variables required to estimate β) is random. This is in spirit similar to the familiar relationship between multinomial sampling and standard stratified sampling. It provides the technical convenience to consider a variety of cases under a unified framework.

The standard and, in this case, efficient estimator of β based on $\mathcal{S}^\dagger$ is:

$$\widehat{\beta}^\dagger = \sum_{j=1}^{n^\dagger} (Y_j - X_j)/n^\dagger \quad \text{with} \quad Var(\widehat{\beta}^\dagger) = \Delta/n^\dagger$$

where $\Delta := Var(Y - X)$. On the other hand, the result in this paper gives an infeasible version of the efficient estimator of β based on $\mathcal{S}$ as:

$$\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{D_i}{p}(Y_i - X_i) + \left(1 - \frac{D_i}{p}\right)(Y_i - E[X|Y_i])\right\} \text{ with } Var\left(\widehat{\beta}\right) = \frac{1}{n}\left[\Delta + \frac{1-p}{p}E[Var(X|Y)]\right].$$

$\widehat{\beta}$ is infeasible because $E[X|Y]$ is unknown in practice. A feasible version of $\widehat{\beta}$ plugs in an estimator $\widehat{E}[X|Y]$ for $E[X|Y]$ in the expression for $\widehat{\beta}$. An important and desirable feature of our results that is repeatedly emphasized in Appendix C is that as long as $\widehat{E}[X|Y]$ is consistent for $E[X|Y]$ uniformly in Support(Y), plugging $\widehat{E}[X|Y]$ in the expression for $\widehat{\beta}$ only makes the result asymptotic, i.e., (i) what is referred to as $Var(\widehat{\beta})$ turns out to be $(1/n)$ times the asymptotic variance of the feasible $\widehat{\beta}$, and (ii) the feasible $\widehat{\beta}$ is no longer unbiased but is asymptotically unbiased and normally distributed.

Now, let the cost of observing Y for a unit be 1 and that for X be c where $c > 1$. Let the allowed expected total cost for the sample be c^*. Thus, $n^\dagger = \lfloor c^*/(1+c)\rfloor$ and $n = \lfloor c^*/(1+pc)\rfloor$ for a given c, c^* and p, and where $\lfloor a \rfloor$ denotes the largest integer $\leq a$. Consider the problem of choosing p such that $Var(\widehat{\beta}) < Var(\widehat{\beta}^\dagger)$. By simple calculations: $Var(\widehat{\beta}) < Var(\widehat{\beta}^\dagger) \iff p > 1/(cq)$ provided that $cq > 1$ where $q = Var(Y - X)/E[Var(X|Y)] - 1$. No solution exists if $cq \leq 1$. However, if $cq > 1$ and $p > 1/(cq)$, then the sample $\mathcal{S}$ is strictly advantageous over the sample $\mathcal{S}^\dagger$ under the premise of the stated problem. (If Y and X are normally distributed with unit variance and correlation ρ then $q = (1-\rho)/(1+\rho)$.) If $cq > 1$ and $n = c^*/(1+pc)$, $Var(\widehat{\beta})$ is minimized when $p = 1/\sqrt{cq}$.

Example 2: Variance reduction through dependent as opposed to independent sampling

Consider estimating the parameter β from a regression model $Y = \alpha + \beta X + \epsilon$ where Y and X are scalar random variables. For simplicity, let $X \sim Bin(1, q)$ and let the model error $\epsilon \sim (0, \sigma^2)$ be independent of X. Let $\mathcal{S} = \{D_i, D_iY_i, X_i\}_{i=1}^n$ where D is a binary variable such that we observe Y in $\mathcal{S}$ only when $D = 1$. (We switch the missing variable from X to Y in this example, unlike in most of our paper, so that we can consider a simple unweighted estimator without bothering about bias due to the possible non-representativeness of the units with $D_i = 1$ [see Wooldridge (2007)].) Let $p(j) = E[D|X = j]$ for $j = 0, 1$. Then, $p := E[D] = qp(1) + (1-q)p(0)$ and $E[DX] = qp(1)$. The ordinary least squares estimator $\widehat{\beta}$ of β, based on sample units with $D_i = 1$, and the asymptotic variance of $\widehat{\beta}$ are, respectively:

$$\widehat{\beta} = \sum_{i=1}^{n} D_i X_i \left(Y_i - \sum_{j=1}^{n} D_j Y_j \Big/ \sum_{j=1}^{n} D_j \right) \Big/ \sum_{i=1}^{n} D_i X_i \left(X_i - \sum_{j=1}^{n} D_j X_j \Big/ \sum_{j=1}^{n} D_j \right)$$

and

$$\text{Avar} = \sigma^2 / E[DX] \left(1 - E[DX]/E[D]\right) = p\sigma^2 / [qp(1)(p - qp(1))] \, .$$

If $P(D = 1|Y, X) = P(D = 1) = p$, implying that $p(1) = p(0) = p$, then Avar $= \sigma^2/pq(1 - q)$. On the other hand, $p(1) = p/(2q)$ minimizes the general Avar and the minimized value is Avar $= 4\sigma^2/p$, which is strictly smaller than $\sigma^2/pq(1 - q)$ unless $q = 1/2$. Hence, by virtue of making D dependent on X, optimally, one could correct for the non-50-50 assignment of X in the population – the essential idea behind stratification – to minimize variance.

A.3 The equivalence relation in the MAR condition in (1)

Lemma 11 *Let $P(C = r|T_R(Z)) > 0$ for each $r = 1, \ldots, R$. Then, $P(C = r|C \geq r, T_R(Z)) = P(C = r|C \geq r, T_r(Z))$ for $r = 1, \ldots, R$ if and only if $P(C = r|T_R(Z)) = P(C = r|T_r(Z))$ for $r = 1, \ldots, R$.*

Proof: We assume only $P(C = r|T_R(Z)) > 0$ for each $r = 1, \ldots, R$ for simplicity to avoid cases with 0/0. The proof follows by induction. We first show the "if" part and then the "only if" part.

"if:" Let $P(C = r|T_R(Z)) = P(C = r|T_r(Z))$ for $r = 1, \ldots, R$. Therefore, $P(C = 1|C \geq 1, T_R(Z)) \equiv P(C = 1|T_R(Z)) = P(C = 1|T_1(Z)) \equiv P(C = 1|C \geq 1, T_1(Z))$. Now, suppose that $P(C = j|C \geq j, T_R(Z)) = P(C = j|C \geq j, T_j(Z))$ for $j = 1, \ldots, r$ for some $r = 1, \ldots, R - 1$. This will imply that $P(C = r + 1|C \geq r + 1, T_R(Z)) = P(C = r + 1|C \geq r + 1, T_{r+1}(Z))$ because:

$$\begin{aligned}
P(C = r + 1|C \geq r + 1, T_R(Z)) &= \frac{P(C = r + 1|T_R(Z))}{P(C \geq r + 1|T_R(Z))} \\
&= \frac{P(C = r + 1|T_R(Z))}{1 - \sum_{j=1}^{r} P(C = j|T_R(Z))} \\
&= \frac{P(C = r + 1|T_{r+1}(Z))}{1 - \sum_{j=1}^{r} P(C = j|T_j(Z))} \\
&= \frac{P(C = r + 1|T_{r+1}(Z))}{1 - \sum_{j=1}^{r} P(C = j|T_{r+1}(Z))} \\
&= \frac{P(C = r + 1|T_{r+1}(Z))}{P(C \geq r + 1|T_{r+1}(Z))} \\
&= P(C = r + 1|C \geq r + 1, T_{r+1}(Z))
\end{aligned}$$

where the first, second, fifth and sixth equalities follow by definition, and the third and fourth equalities follow from the assumed conditions once we note that $T_j(Z)$ is nested by $T_{j+1}(Z)$ for all $j = 1, \ldots, R - 1$.

"only if:" Let $P(C = r|C \geq r, T_R(Z)) = P(C = r|C \geq r, T_r(Z))$ for $r = 1, \ldots, R$. Therefore, $P(C = 1|T_R(Z)) \equiv P(C = 1|C \geq 1, T_R(Z)) = P(C = 1|C \geq 1, T_1(Z)) \equiv P(C = 1|T_1(Z))$. Now, suppose that $P(C = j|T_R(Z)) = P(C = j|T_j(Z))$ for $j = 1, \ldots, r$ for some $r = 1, \ldots, R-1$. This will imply that $P(C = r+1|T_R(Z)) = P(C = r+1|T_{r+1}(Z))$ because:

$$\begin{aligned}
P(C = r+1|T_R(Z)) &= P(C = r+1, C \geq r+1|T_R(Z)) \\
&= P(C = r+1|C \geq r+1, T_R(Z))P(C \geq r+1|T_R(Z)) \\
&= P(C = r+1|C \geq r+1, T_R(Z))\left(1 - \sum_{j=1}^{r} P(C = j|T_R(Z))\right) \\
&= P(C = r+1|C \geq r+1, T_{r+1}(Z))\left(1 - \sum_{j=1}^{r} P(C = j|T_j(Z))\right) \\
&= P(C = r+1|C \geq r+1, T_{r+1}(Z))P(C \geq r+1|T_r(Z)) \\
&= P(C = r+1|T_{r+1}(Z))
\end{aligned}$$

where the first three equalities follow by definition, the fourth equality follows by the assumed conditions, and the last two equalities are simply the reverse steps of the first three equalities coupled with the fact that $T_j(Z)$ is nested by $T_{j+1}(Z)$ for all $j = 1, \ldots, R-1$. ■

A.4 The equivalence relation in the planned incompleteness condition in (2)

Lemma 12 *Let* (1) *hold and also* $P(C = r|T_R(Z)) > 0$ *for each* $r = 1, \ldots, R$*. Then,* $P(C = r|C \geq r, T_r(Z))$ *is known for* $r = 1, \ldots, R$ *if and only if* $P(C = r|T_r(Z))$ *is known for* $r = 1, \ldots, R$*.*

Proof: The proof follows by induction exactly like the proof of Lemma 11. For the "if" part, when showing that the result holds for $r+1$ assuming that it holds for $j = 1, \ldots, r$, we have:

$$P(C = r+1|C \geq r+1, T_{r+1}(Z)) = \frac{P(C = r+1|T_{r+1}(Z))}{1 - \sum_{j=1}^{r} P(C = j|T_j(Z))}$$

as before due to (1). The RHS is known by the assumed conditions. Hence the LHS is known.

For the "only if" part, when showing that the result holds for $r+1$ assuming that it holds for $j = 1, \ldots, r$, we have:

$$P(C = r+1|T_{r+1}(Z)) = P(C = r+1|C \geq r+1, T_{r+1}(Z))\left(1 - \sum_{j=1}^{r} P(C = j|T_j(Z))\right)$$

as before due to (1). The RHS is known by the assumed conditions. Hence the LHS is known. ■

Remark: At this stage, it is important to list two useful relations that are both related to the steps in the proofs of Lemmas 11 and 12, and also used repeatedly in the proofs in Appendices A and B.

Relation 1: (1) implies that

$$P(C \geq r|T_R(Z)) = P(C \geq r|T_{r-1}(Z)). \tag{31}$$

This follows by noting that:

$$\begin{aligned} P(C \geq r|T_R(Z)) &= 1 - \sum_{j=1}^{r-1} P(C = j|T_R(Z)) \\ &= 1 - \sum_{j=1}^{r-1} P(C = j|T_j(Z)) \\ &= 1 - \sum_{j=1}^{r-1} P(C = j|T_{r-1}(Z)) \\ &= 1 - P(C \leq r-1|T_{r-1}(Z)) = P(C \geq r|T_{r-1}(Z)) \end{aligned}$$

where the first equality follows by definition, the second by (1), the third by (1) and the nested structure of $T_j(Z)$'s, while the fourth and the fifth by definition.

Note that, taking $R = 2$ in (31) implies that $P(C = 2|T_2(Z)) = P(C = 2|T_1(Z))$, the conventional MAR assumption found in the econometrics literature that has traditionally focused on $R = 2$ [see, e.g., Chen et al. (2005), Chen et al. (2008), Graham (2011), Graham et al. (2012)]. Looking at the complement events in (31) equivalently gives (31) as $P(C \leq r-1|T_R(Z)) = P(C \leq r-1|T_{r-1}(Z))$, which perhaps better indicates the generality of the selection on variables condition in our paper that can accommodate for all sorts of dimension reductions including the extreme reduction CMAR in (10) and the no reduction in Barnwell and Chaudhuri (2018).

Relation 2: For any function $\nu(Z)$ such that $E|\nu(Z)| < \infty$, (1) implies that:

$$\begin{aligned} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r(Z))}\nu(Z)\right] &= E\left[\frac{P(C \geq r|Z)}{P(C \geq r|T_r(Z))}\nu(Z)\right] \\ &= E\left[\frac{P(C \geq r|T_r(Z))}{P(C \geq r|T_r(Z))}\nu(Z)\right] = E[\nu(Z)] \end{aligned} \tag{32}$$

where the first equality follows by the law of iterated expectations and the second one by (1). As a consequence of (31), one can instead write (32) as:

$$\begin{aligned} E\left[\frac{I(C \geq r)}{P(C \geq r|T_{r-1}(Z))}\nu(Z)\right] &= E\left[\frac{P(C \geq r|T_r(Z))}{P(C \geq r|T_{r-1}(Z))}\nu(Z)\right] \\ &= E\left[\frac{P(C \geq r|T_{r-1}(Z))}{P(C \geq r|T_{r-1}(Z))}\nu(Z)\right] = E[\nu(Z)]. \end{aligned}$$

A.5 Intermediate steps in equation (4)

$$
\begin{aligned}
& E\left[\frac{P(C \in \lambda|T_R(Z))}{P(C \in \lambda)} \frac{I(C=R)}{P(C=R|T_R(Z))} m(Z;\beta)\right] \\
= \;& E\left[\frac{P(C \in \lambda|T_R(Z))}{P(C \in \lambda)} E\left[\left.\frac{I(C=R)}{P(C=R|T_R(Z))}\right| T_R(Z)\right] m(Z;\beta)\right] \\
= \;& E\left[\frac{P(C \in \lambda|T_R(Z))}{P(C \in \lambda)} m(Z;\beta)\right] \\
= \;& E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)} m(Z;\beta)\right] \\
= \;& E[m(Z;\beta)|C \in \lambda].
\end{aligned}
$$

The first and third equalities follow by the law of iterated expectations, and the rest by definition.

Importantly, note that, the MAR condition in (1) and the planned incompleteness condition in (2) are not required for this relation in (4) to hold. However, as noted in the discussion around equations (1) and (2) that led to (4), the MAR condition in (1), in particular, is required to implement this relation in practice for the estimation of β by the IPW or the efficient estimator.

A.6 Relation of the framework in Section 2 with closely related technical papers

We delineate the framework in Section 2 from the following not-too-old representative examples under the non-Bayesian paradigm. **(a)** Whittemore (1997) considers maximum likelihood and Horvitz-Thompson estimators with data obtained by multi-phase sampling (and seems to prefer the latter) where the target is the full population, i.e, $\lambda = \mathcal{C}$. **(b)** Robins and Rotnitzky (1995) and Holcroft et al. (1997) consider optimally using all the sub-samples under a framework similar to ours but with $\lambda = \mathcal{C}$. **(c)** Lee et al. (2012) consider efficient semiparametric likelihood-based estimation with $\lambda = \mathcal{C}$ in multi-phase case-control studies when $T_{R-1}(Z)$ has a finite number of support points. **(d)** While the multi-valued treatment framework with $\lambda = \mathcal{C}$ considered in Cattaneo (2010) is generally related, it also differs in an important way because we actually allow the entire random vector Z to be the argument for each element of the vectorial moment function $m(Z;\beta)$, and thus for each element there can be R levels of hierarchy in observability. This creates a major difference in terms of efficiency bounds, efficient influence functions, etc., and is discussed in details in Chaudhuri and Guilkey (2016) (p. 686). **(e)** Dardanoni et al. (2011) consider a multiple regression framework with regressors missing non-monotonically under an assumption that implies that the regression coefficients do not vary across the populations of the sub-samples. So, they focus on $\lambda = \mathcal{C}$ and, unlike in our paper and the references cited in (a)-(d) and (f) (below), use of their complete sub-sample without correction for selection does not cause any bias in estimation.[20] Similarly, if one extends Abrevaya and Donald (2017) to the case of multiple incomplete sub-samples, then each sub-population would still be representative of $\lambda = \mathcal{C}$. **(f)** Finally, Chen et al. (2005) and Chen et al. (2008) consider frameworks where β^0_λ is defined exactly as in (3) for $R = 2$ and $\lambda = \{1\}$ (sub-population) and $\{1, 2\}$ (full population).

By contrast in one way or the other to (a)-(f), our setup: (i) allows for a general R, (ii) expands the scope to all $(2^R - 1)$ sub-populations (including $\lambda = \mathcal{C}$), (iii) introduces a dynamically updated sampling design via MAR, and (iv) provides the new insights available only from letting $R > 2$.

In this regard, it is also important to recall that the references in (d)-(f) above or the well-known sampling designs like the SQD, PQD, etc. noted in Appendix A.1 either do not consider or do not have the scope to consider a key feature of our framework, namely, sampling designs that are dynamically updated using the newly available information from more than one phase.

A.7 Intermediate steps for Remark 1 following Proposition 1

When $R = 2$ and $\lambda = \{1, 2\}$, (5) and (6) give:

$$\begin{aligned}
\varphi_{\{1,2\}}(O;\beta) &= \frac{I(C=2)}{P(C=2|T_2(Z))} m(T_2(Z);\beta) \\
&\qquad + \left(\frac{I(C\geq 1)}{P(C \geq 1|T_1(Z))} - \frac{I(C=2)}{P(C=2|T_2(Z))} \right) E[m(T_2(Z);\beta)|T_1(Z)] \\
&= \frac{I(C=2)}{P(C=2|T_1(Z))} m(T_2(Z);\beta) + \left(1 - \frac{I(C=2)}{P(C=2|T_1(Z))} \right) E[m(T_2(Z);\beta)|T_1(Z)] \\
&= \frac{I(C=2)}{P(C=2|T_1(Z))} \left(m(T_2(Z);\beta) - E[m(T_2(Z);\beta)|T_1(Z)]\right) + E[m(T_2(Z);\beta)|T_1(Z)]
\end{aligned}$$

where the second equality follows from (31). The last line is the expression from Chen et al. (2008).

[20] Bias arises due to problems with the imputed values if the same estimation is done in the incomplete sub-samples by replacing the missing regressors with their imputed values. To improve the precision of the unbiased estimator based on the complete sub-sample, they recommend Bayesian model averaging using the unbiased and biased estimates. While this approach should be very useful in many cases, it is a difficult proposition to compare it with the results in our paper and the other references here that all solve a different optimization problem: minimize asymptotic variance for *asymptotically unbiased* estimators. We thank a referee for pointing out this useful reference that we had missed earlier.

When $R=2$ and $\lambda=\{1\}$, (5) and (6) give:

$$\begin{aligned}
\varphi_{\{1\}}(O;\beta) &= \frac{I(C=2)}{P(C=2|T_2(Z))}\frac{P(C=1|T_2(Z))}{P(C=1)}m(T_2(Z);\beta) \\
&\quad + \left(\frac{I(C\geq 1)}{P(C\geq 1|T_1(Z))} - \frac{I(C=2)}{P(C=2|T_2(Z))}\right) E\left[\left.\frac{P(C=1|T_2(Z))}{P(C=1)}m(T_2(Z);\beta)\right| T_1(Z)\right] \\
&= \frac{I(C=2)}{P(C=2|T_1(Z))}\frac{P(C=1|T_1(Z))}{P(C=1)}m(T_2(Z);\beta) \\
&\quad + \left(1 - \frac{I(C=2)}{P(C=2|T_1(Z))}\right) E\left[\left.\frac{P(C=1|T_1(Z))}{P(C=1)}m(T_2(Z);\beta)\right| T_1(Z)\right] \\
&= \frac{I(C=2)}{P(C=2|T_1(Z))}\frac{P(C=1|T_1(Z))}{P(C=1)}\left(m(T_2(Z);\beta) - E[m(T_2(Z);\beta)|T_1(Z)]\right) \\
&\quad + \frac{P(C=1|T_1(Z))}{P(C=1)}E[m(T_2(Z);\beta)|T_1(Z)]
\end{aligned}$$

where the second equality follows from (31) and (1). The RHS of the last equality is the expression from Chen et al. (2008).

A.8 Proposition 2's connection with the calibration and econometrics literature

The idea behind using the moment restrictions in (9) to augment the moment restriction (8), that already identifies β^0_λ and can be used to obtain a $\sqrt{n}$-consistent estimator [see, e.g., Wooldridge (2007)], and thus achieving efficiency gains is the same as the idea of calibration in the survey sampling literature [see, e.g., Deville and Sarndal (1992)]. The same idea, in more economics-centric ways, has appeared in the econometrics literature also: see Back and Brown (1993), Imbens and Lancaster (1994), Hellerstein and Imbens (1999), Devereux and Tripathi (2009), Tripathi (2011), Graham et al. (2012), etc. or Hellerstein and Imbens (1999), Nevo (2003), etc. in another context. To see the connection, first note that under our setup this means estimating β^0_λ by solving for β from $\sum_{i=1}^n \omega_i \varphi_{R,\lambda}(O_i,\beta) = 0$ where $\omega_i = I(C_i = R)/P(C = R|T_R(Z_i)) = \omega_{IPW,i}$, say, (instead of $1/n$ to reflect the non-representativeness of the complete sub-sample) if only (8) is used. On the other hand, if the calibration/augmenting/auxiliary restrictions in (9) are also utilized, then $\omega_i = \omega_{IPW,i} + \sum_{r=1}^{R-1} a_{r,i}$ for some appropriate (and complicated) set of random functions $a_{r,i}$'s. For example, if $R=2$, then $a_{1,i} = \omega_{IPW,i}\Upsilon'_{K_1}(T_1(Z_i))(\sum_{j=1}^n \Upsilon_{K_1}(T_1(Z_j))\Upsilon'_{K_1}(T_1(Z_j)))^{-1}\sum_{l=1}^n(1-\omega_{IPW,l})\Upsilon_{K_1}(T_1(Z_l))$ where $\Upsilon_{K_1}(T_1(Z))$ is a $K_1\times 1$ vector of some possibly orthogonalized series of functions (e.g., power series, splines, etc.) of $T_1(Z)$ with possibly $K_1\to\infty$ as $n\to\infty$ [see Graham et al. (2012)]. One could instead use $\bar{\omega}_i = \omega_i/\sum_j \omega_j$ as the weights so that they necessarily add up to one. However, there is no guarantee that $\bar{\omega}_i \in [0,1]$ for all i (indeed it can be outside $[0,1]$ for all i), which is not a desirable characteristic for weights. We do not pursue corrections for this undesirable characteristic of the weights since they are peripheral to the main message of our paper.

A.9 The importance of the planned incompleteness condition (2) in Proposition 2

This importance becomes evident when the target is *not* the full population. Consider $R = 2$ and $\lambda = \{1\}$, and note that Proposition 2 gives:

$$\begin{aligned}
\varphi_{\{1\}}(O;\beta) &= \overline{\mathrm{Proj}}_{T_1}\left(\phi_{2,\lambda}(O;\beta)|\,\phi_1\right) = \phi_{2,\lambda}(O;\beta) - \mathrm{Proj}_{T_1}\left(\phi_{2,\lambda}(O;\beta)|\,\phi_1\right) \\
&= \frac{I(C=2)}{P(C=2|T_1(Z))}\frac{P(C=1|T_1(Z))}{P(C=1)}m(Z;\beta) \\
&\quad - \left\{\frac{P(C=1|T_1(Z))}{P(C=1)P(C=2|T_1(Z))}E[m(Z;\beta)|T_1(Z)]\right\}(I(C=2) - P(C=2|T_1(Z))) \\
&= \frac{I(C=2)}{P(C=2|T_1(Z))}\frac{P(C=1|T_1(Z))}{P(C=1)}(m(Z;\beta) - E[m(Z;\beta)|T_1(Z)]) \\
&\quad + \frac{P(C=1|T_1(Z))}{P(C=1)}E[m(Z;\beta)|T_1(Z)].
\end{aligned}$$

On the other hand, it is known from Case 1 in Theorem 1 of Chen et al. (2008) (or plugging in $R = 2$ and $\lambda = \{1\}$ in our Proposition 5, or, equivalently, Barnwell and Chaudhuri (2018)'s Proposition 1) that the corresponding quantity without (2) would be:

$$\begin{aligned}
\varphi_{\{1\}[u]}(O;\beta) &= \frac{I(C=2)}{P(C=2|T_1(Z))}\frac{P(C=1|T_1(Z))}{P(C=1)}(m(Z;\beta) - E[m(Z;\beta)|T_1(Z)]) \\
&\quad + \frac{I(C=1)}{P(C=1)}E[m(Z;\beta)|T_1(Z)].
\end{aligned}$$

Of course, $\varphi_{\{1\}[u]}(O;\beta) \neq \varphi_{\{1\}}(O;\beta)$, i.e., Proposition 2 does not generally apply when targets are sub-populations unless the planned incompleteness condition in (2) holds.

Appendix B: Detailed version of the proofs of the results in Section 2 and 3

Appendix B provides the detailed version of the proofs of the results in Sections 2 and 3. Abridged version of the same proofs were presented in our paper.

The proofs of Propositions 1, 3, 4 and 5 involve obtaining the semiparametric efficiency bound and the efficient influence function, under different assumptions, following Chen et al. (2008). They follow in two steps. Step 1 characterizes the tangent set for all regular parametric sub-models satisfying the semiparametric assumptions on the observed data. Step 2 obtains the efficient influence function and, thereby, the asymptotic variance lower bound as the expectation of its outer product. 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.

Proof of Proposition 1:

STEP 1: Consider a regular parametric sub-model indexed by a parameter θ for the distribution of the observed data $O = (C', T'_C(Z))'$. The log of the distribution can be expressed in terms of the full data $(C, Z')'$ as:

$$\log f_\theta(O) = \log f_\theta(Z_{(1)}) + \sum_{r=2}^{R} I(C \geq r) \log f_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)}) + \sum_{r=1}^{R} I(C = r) \log P(C = r|Z_{(1)}, \ldots, Z_{(r)}).$$

To reflect our condition (2), i.e., $P(C = r|Z_{(1)}, \ldots, Z_{(r)})$ is known for $r = 1, \ldots, R$ and hence need not be accounted for in what follows, we do not index them by θ. (These quantities do not play a role in the proof of the present proposition but does so in the proof of our Propositions 4 and 5.)

θ_0 is the unique value of θ such that $f_{\theta_0}(O)$ equals the true $f(O)$, and accordingly for all the quantities. The score function with respect to θ can then be written in terms of $(C, Z')'$ as:

$$S_\theta(O) = s_\theta(Z_{(1)}) + \sum_{r=2}^{R} I(C \geq r) s_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})$$

where $s_\theta(Z_{(1)}) := \frac{\partial}{\partial\theta} \log f_\theta(Z_{(1)})$ and $s_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)}) := \frac{\partial}{\partial\theta} \log f_\theta(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})$. (We will omit the subscript θ from the quantities evaluated at $\theta = \theta_0$.) The tangent set is the mean square closure of all d dimensional linear combinations of $S_\theta(O)$ for all such smooth parametric sub-models, and it takes the form:

$$\mathcal{T} := a_1(Z_{(1)}) + \sum_{r=2}^{R} I(C \geq r) a_r(Z_{(1)}, \ldots, Z_{(r)}), \tag{33}$$

where $a_1(Z_{(1)}) \in L_0^2(F(Z_{(1)}))$ and $a_r(Z_{(1)}, \ldots, Z_{(r)}) \in L_0^2(F(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)}))$.

STEP 2: Differentiating the moment conditions in (3) with respect to θ under the integral, and noting that $P(C \in \lambda|Z)$ (which is known) does not depend on θ but $P(C \in \lambda)$ (which is unknown) does, we obtain by using (3) and (1) that:

$$0 = M_\lambda \frac{\partial \beta_\lambda^0(\theta_0)}{\partial\theta'} + E\left[m(Z; \beta_\lambda^0) \left\{ s(Z_{(1)})' + \sum_{r=2}^{R} s(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})' \right\} \middle| C \in \lambda \right].$$

Therefore, assumption (A3) now gives:

$$\frac{\partial \beta_\lambda^0(\theta_0)}{\partial\theta'} = -M_\lambda^{-1} E\left[m(Z; \beta_\lambda^0) \left\{ s(Z_{(1)})' + \sum_{r=2}^{R} s(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})' \right\} \middle| C \in \lambda \right].$$

Pathwise differentiability follows if we can find a $\psi(O) \in \mathcal{T}$ such that:

$$E[\psi(O)S(O)'] = \frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'}. \tag{34}$$

Let us conjecture that $\psi(O) = -M_\lambda^{-1}\varphi_\lambda(O;\beta_\lambda^0)$, and then verify (34) by equivalently showing that:

$$E[\varphi_\lambda(O;\beta_\lambda^0)S(O)'] = E\left[m(Z;\beta_\lambda^0)\left\{s(Z_{(1)})' + \sum_{r=2}^{R} s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})'\right\}\middle| C \in \lambda\right].$$

Consider the left hand side (LHS) and, in accordance with the partition of $\varphi_\lambda(O)$ (we work with the alternative specification in (7) for convenience), write it as $\sum_{q=1}^{R} B_q$ where, for $q = 2,\ldots,R$:

$$B_q := E\left[\frac{I(C \geq q)}{P(C \geq q|T_q(Z)}\left[\varphi_{q,\lambda}(O;\beta_\lambda^0) - \varphi_{q-1,\lambda}(O;\beta_\lambda^0)\right]S(O)'\right] \text{ while } B_1 := E\left[\varphi_{1,\lambda}(O;\beta_\lambda^0)S(O)'\right].$$

To avoid notational clutter, in the rest of STEP 2 we write $m(Z;\beta_\lambda^0)$ as m; $T_q(Z)$ as T_q; $\varphi_{q,\lambda}(O;\beta_\lambda^0)$ as $\varphi_{q,\lambda}$ for $q = 1,\ldots,R$; and also write $s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})$ as $s(Z_{(r)}|T_{r-1})$ for $r = 2,\ldots,R$.

Now, note that:

$$B_1 = E\left[E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}m\middle|T_1\right]s(Z_{(1)})'\right] + \sum_{r=2}^{R} E\left[E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}m\middle|T_1\right]I(C\geq r)s(Z_{(r)}|T_{r-1})'\right].$$

Using MAR in (1) in the first equality of the last line below and the fact that $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$ for $r > 1$ in the last equality of the last line below, we obtain that:

$$\begin{aligned}
&\sum_{r=2}^{R} E\left[E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}m\middle|T_1\right]I(C\geq r)s(Z_{(r)}|T_{r-1})'\right] \\
=\ &\sum_{r=2}^{R} E\left[E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}m\middle|T_1\right](1 - I(C\leq r-1))s(Z_{(r)}|T_{r-1})'\right] \\
=\ &\sum_{r=2}^{R} E\left[E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}m\middle|T_1\right]E[(1 - I(C\leq r-1))|T_{r-1}]E[s(Z_{(r)}|T_{r-1})'|T_{r-1}]\right] = 0.
\end{aligned}$$

This is the first observation. On the other hand, since $T_1 := Z_{(1)}$, we have the second observation:

$$E\left[E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}m\middle|T_1\right]s(Z_{(1)})'\right] = E\left[\frac{P(C\in\lambda|T_R)}{P(C\in\lambda)}ms(Z_{(1)})'\right] = E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}ms(Z_{(1)})'\right].$$

Combining the two observations it follows that $B_1 = E[ms(Z_{(1)})'|C\in\lambda]$.

Now, we consider B_q. (1) gives for $q = 2, \ldots, R$:

$$\begin{aligned} & B_q \\ = \; & \sum_{r=1}^{q-1} E\left[\frac{I(C \geq q)}{P(C \geq q|T_q)}(\varphi_{q,\lambda} - \varphi_{q-1,\lambda})s(Z_{(r)}|T_{r-1})'\right] + \sum_{r=q}^{R} E\left[\frac{I(C \geq r)}{P(C \geq q|T_q)}(\varphi_{q,\lambda} - \varphi_{q-1,\lambda})s(Z_{(r)}|T_{r-1})'\right]. \end{aligned}$$

Since $E[\varphi_{q,\lambda}|T_{q-1}] = \varphi_{q-1,\lambda}$, it follows by conditioning on T_{q-1} and from (32) that the first term on the RHS is 0. On the other hand, (31) and the fact that $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1}))$ implies that the second term is:

$$\begin{aligned} & \sum_{r=q}^{R} E\left[\frac{1 - I(C \leq r-1)}{1 - P(C \leq q-1|T_{q-1})}(\varphi_{q,\lambda} - \varphi_{q-1,\lambda})s(Z_{(r)}|T_{r-1})'\right] \\ = \; & E[\varphi_{q,\lambda}s(Z_{(q)}|T_{q-1})'] \\ = \; & E[ms(Z_{(q)}|T_{q-1})'|C \in \lambda]. \end{aligned}$$

Therefore, $B_q = E[ms(Z_{(q)}|T_{q-1})'|C \in \lambda]$ for $q = 2, \ldots, R$, combining which with B_1 verifies (34).

That $\psi(O) \in \mathcal{T}$ follows from matching terms as follows. (i) $-M_\lambda^{-1}\varphi_{1,\lambda}$ is a function of only $T_1 := Z_{(1)}$, and $E[\varphi_{1,\lambda}] = 0$ and, hence, satisfies the properties of $a_1(Z_{(1)})$ in (33). (ii) The r-th term ($r = 2, \ldots, R$, without the multiplier $I(C \geq r)$) on the RHS of $\psi(O)$ can be written as: $-\frac{1}{P(C \geq r|T_r)}M_\lambda^{-1}[\varphi_{r,\lambda} - \varphi_{r-1,\lambda}] = -\frac{1}{1-P(C \leq r-1|T_{r-1})}M_\lambda^{-1}[\varphi_{r,\lambda} - \varphi_{r-1,\lambda}]$ by (1) [also see (31)]. Hence, by definition of φ_r, taking expectation of the RHS of the above equation conditional on $T_{r-1} := (Z_{(1)}, \ldots, Z_{(r-1)})'$ gives 0. Therefore, this term is a function of only T_r and it is also in $L_0^2(F(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)}))$, and hence satisfies the properties of $a_r(Z_{(1)}, \ldots, Z_{(r)})$ in (33).

Therefore, we have now verified that the projection of the influence function $-M_\lambda^{-1}m(Z; \beta_\lambda^0)$ on to the tangent set $\mathcal{T}$ is $\psi(O) := -M_\lambda^{-1}\varphi_\lambda(O; \beta_\lambda^0)$. Hence, $\psi(O)$ is the efficient influence function and, therefore, the asymptotic variance lower bound is $E[\psi(O)\psi(O)'] = M_\lambda^{-1}V_\lambda M_\lambda^{-1'} =: \Omega_\lambda$. ■

Proof of Proposition 2:

(i) Let us start with $r = 1$, i.e., the residual from the projection, $\overline{\text{Proj}}_{T_{R-1}}(\phi_{R,\lambda}(\beta)|\,\phi_{R-1})$, inside the innermost parenthesis on the RHS. We will also consider $r = 2$ so that the pattern in the form of the residuals from the successive projections inside the first few innermost parentheses is clear to all. Then we apply induction arguments. For brevity, write $\varphi_{R,\lambda}(O; \beta)$ as $\varphi_{R,\lambda}$ and $T_r(Z)$ as T_r.

First, note that direct computation and (1) along with (31) give:

$$\text{Proj}_{T_{R-1}}(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}) = \left[\frac{I(C = R)}{P(C = R|T_R)} - \frac{I(C \geq R-1)}{P(C \geq R-1|T_{R-1})}\right] E[\varphi_{R,\lambda}|T_{r-1}],$$

which implies that:

$$\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right) = \frac{I(C=R)}{P(C=R|T_R)}\underbrace{(\varphi_{R,\lambda} - E[\varphi_{R,\lambda}|T_{R-1}])}+\frac{I(C\geq R-1)}{P(C\geq R-1|T_{R-1})}E[\varphi_{R,\lambda}|T_{R-1}].$$

Consider the under-braced part in the RHS of the expression for $\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right)$. Using $T_{R-1}\setminus T_{R-2} = Z_{(R-1)}$ and (1), note that $E\left[(\varphi_{R,\lambda} - E[\varphi_{R,,\lambda}|T_{R-1}])\,\phi_{R-2}|T_{R-2}\right]$ is a $d\times 1$ vector of zeros, and hence has no contribution in the successive projections. (Terms with no contribution in the successive projections are marked by under-braces in this proof.) On the other hand,

$$E\left[\frac{I(C\geq R-1)}{P(C\geq R-1|T_{R-1})}E[\varphi_{R,\lambda}|T_{R-1}]\phi_{R-2}\middle|\,T_{R-2}\right] = \frac{P(C=R-2|T_{R-2})}{P(C\geq R-2|T_{R-2})}E[\varphi_{R,\lambda}|T_{R-2}].$$

Thus, similar computation as above (and the use of (31)) gives for $r=2$:

$$\text{Proj}_{T_{R-2}}\left(\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right)\middle|\,\phi_{R-2}\right) = \left[\frac{I(C\geq R-1)}{P(C\geq R-1|T_{R-1})} - \frac{I(C\geq R-2)}{P(C\geq R-2|T_{R-2})}\right]E[\varphi_{R,\lambda}|T_{R-2}],$$

which implies that:

$$\begin{aligned}
&\overline{\text{Proj}}_{T_{R-2}}\left(\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right)\middle|\,\phi_{R-2}\right)\\
=\;&\sum_{s=0}^{1}\frac{I(C\geq R-s)}{P(C\geq R-s|T_{R-s})}\underbrace{(E[\varphi_{R,\lambda}|T_{R-s}] - E[\varphi_{R,\lambda}|T_{R-s-1}])}+\frac{I(C\geq R-2)}{P(C\geq R-2|T_{R-2})}E[\varphi_{R,\lambda}|T_{R-2}].
\end{aligned}$$

For our proof by induction, first assume that the following holds for a general $r\in\{2,\ldots,R-2\}$:

$$\begin{aligned}
&\overline{\text{Proj}}_{T_{R-r}}\left(\ldots\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right)\ldots\middle|\,\phi_{R-r}\right)\\
=\;&\sum_{s=0}^{r-1}\frac{I(C\geq R-s)}{P(C\geq R-s|T_{R-s})}\underbrace{(E[\varphi_{R,\lambda}|T_{R-s}] - E[\varphi_{R,\lambda}|T_{R-s-1}])}+\frac{I(C\geq R-r)}{P(C\geq R-r|T_{R-r})}E[\varphi_{R,\lambda}|T_{R-r}].
\end{aligned}$$

Now, once again using (31), note that:

$$E[\phi_{R-r-1}^2|T_{R-r-1}] = \frac{P(C\geq R-r|T_{R-r})P(C=R-r-1|T_{R-r-1})}{P(C\geq R-r-1|T_{R-r-1})},$$

and

$$\begin{aligned}
&E[\overline{\text{Proj}}_{T_{R-r}}\left(\ldots\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)|\,\phi_{R-1}\right)\ldots\middle|\,\phi_{R-r}\right)\phi_{R-r-1}|T_{R-r-1}]\\
=\;&\frac{P(C=R-r-1|T_{R-r-1})}{P(C\geq R-r-1|T_{R-r-1})}E[\varphi_{R,\lambda}|T_{R-r-1}].
\end{aligned}$$

Hence, the proof follows by induction since the form is also valid for $r+1$, i.e.,

$$\begin{aligned}&\overline{\text{Proj}}_{T_{R-r-1}}\left(\ldots\overline{\text{Proj}}_{T_{R-1}}\left(\phi_{R,\lambda}(\beta)\middle|\,\phi_{R-1}\right)\ldots\middle|\,\phi_{R-r-1}\right)\\ =&\sum_{s=0}^{r}\frac{I(C\geq R-s)}{P(C\geq R-s|T_{R-s})}\left(E[\varphi_{R,\lambda}|T_{R-s}]-E[\varphi_{R,\lambda}|T_{R-s-1}]\right)+\frac{I(C\geq R-r-1)}{P(C\geq R-r-1|T_{R-r-1})}E[\varphi_{R,\lambda}|T_{R-r-1}].\end{aligned}$$

(ii) The proof follows in the same way as that of Theorem 1 in Chamberlain (1992) or, more generally, as that of Theorem 1 of Ai and Chen (2012). Appendix B.1 makes the connection with Ai and Chen (2012) explicit. ■

Proof of Proposition 3:

This proof follows in the same way as that of Proposition 1. The efficient influence function in this case turns out to be exactly the same as in Proposition 1 if CMAR is imposed on the latter. ■

We present the proofs of Propositions 4 and 5 in reverse order because the proof for the latter makes a reference to that for the former. Certain details of lesser importance are omitted below because they were already made explicit in the proof of Proposition 1.

Proof of Proposition 5:

STEP 1: Consider a regular parametric sub-model indexed by θ for the joint distribution of the observed data $O=(C,T_C'(Z))'$. Because of CMAR in (10), the log of the distribution can be expressed in terms of the full data $(C,Z')'$ as:

$$\log f_\theta(O)=\sum_{r=1}^{R}I(C=r)\log P_\theta(C=r|Z_{(1)})+\sum_{r=1}^{R}I(C\geq r)\log f_\theta(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})+\log f_\theta(Z_{(1)}).$$

Let the true distribution be $f(O)=f_{\theta_0}(O)$ for some θ_0. Using the same notations as before, the score function with respect to θ can be written in terms of $(C,Z')'$ as:

$$S_\theta(O)=s_\theta(Z_{(1)})+\sum_{r=2}^{R}I(C\geq r)s_\theta(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})+\sum_{r=1}^{R}I(C=r)\frac{\dot{P}_\theta(C=r|Z_{(1)})}{P_\theta(C=r|Z_{(1)})}$$

where $\dot{P}_\theta(C=r|Z_{(1)}):=\frac{\partial}{\partial\theta}P_\theta(C=r|Z_{(1)})$. Thus, the tangent space is characterized by functions of the form:

$$\mathcal{T}:=a_1(Z_{(1)})+\sum_{r=2}^{R}I(C\geq r)a_r(Z_{(1)},\ldots,Z_{(r)})+\sum_{r=1}^{R}I(C=r)\frac{b_r(Z_{(1)})}{bb_r(Z_{(1)})},\tag{35}$$

where $a_1(Z_{(1)})\in L_0^2(F(Z_{(1)}))$; $a_r(Z_{(1)},\ldots,Z_{(r)})\in L_0^2(F(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)}))$ for $r=2,\ldots,R$; $\sum_{r=1}^{R}b_r(Z_{(1)}))=0$, $\sum_{r=1}^{R}bb_r(Z_{(1)})=1$, and $\sum_{r=1}^{R}I(C=r)\frac{b_r(Z_{(1)})}{bb_r(Z_{(1)})}\in L_0^2(F(C|Z_{(1)}))$.

To avoid notational clutter, in the rest of the proof we write $m(Z;\beta_\lambda^0)$ as m; $T_r(Z)$ as T_r for

$r = 1, \ldots, R$; and also write $s(Z_{(r)}|Z_{(1)}, \ldots, Z_{(r-1)})$ as $s(Z_{(r)}|T_{r-1})$ for $r = 2, \ldots, R$.

Unlike in Chen et al. (2008)'s proof we use the same factorization of the joint density of O for all λ. For a given $\lambda \in \Lambda$, the following relation obtained by two different factorization of the joint distribution of $(I(C \in \lambda), T_1(Z) \equiv Z_{(1)})$ helps us to switch between different factorizations:

$$
\begin{aligned}
& s(T_1) + I(C \in \lambda)\frac{\dot{P}(C \in \lambda|T_1)}{P(C \in \lambda|T_1)} + I(C \notin \lambda)\frac{\dot{P}(C \notin \lambda|T_1)}{P(C \notin \lambda|T_1)} \\
= \; & I(C \in \lambda)\left[\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(T_1|C \in \lambda)\right] + I(C \notin \lambda)\left[\frac{\dot{P}(C \notin \lambda)}{P(C \notin \lambda)} + s(T_1|C \notin \lambda)\right]. \qquad (36)
\end{aligned}
$$

STEP 2: Differentiating (3) with respect to θ under the integral:

$$
\frac{\partial \beta^0_\lambda(\theta_0)}{\partial \theta'} = -M_\lambda^{-1} E\left[m\left\{s(T_1|C \in \lambda)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})'\right\}\middle| C \in \lambda\right].
$$

Then, as in the proof of Proposition 1, here we will need to correspondingly verify that:

$$
E[\varphi^{\text{CMAR}}_{\lambda[u]}(O; \beta^0_\lambda)S(O)'] = E\left[m\left\{s(T_1|C \in \lambda)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})'\right\}\middle| C \in \lambda\right]. \qquad (37)
$$

We do this term by term for $\varphi^{\text{CMAR}}_{\lambda[u]}(O; \beta^0_\lambda)$ and show equality of the terms on the LHS and RHS.

Consider the first term of $\varphi^{\text{CMAR}}_{\lambda[u]}(O; \beta^0_\lambda)$. Since $s(Z_{(r)}|T_{r-1}) \in L^2_0(F(Z_{(r)}|T_{r-1}))$ for $r = 2, \ldots, R$ by definition, we can use (10) to take conditional expectations and then write:

$$
\begin{aligned}
E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)}E[m|T_1]S(O)'\right] &= E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)}E[m|T_1]\left\{s(T_1)' + \sum_{r=1}^{R} I(C = r)\frac{\dot{P}(C = r|T_1)'}{P(C = r|T_1)}\right\}\right] \\
&= E\left[\frac{I(C \in \lambda)}{P(C \in \lambda)}E[m|T_1]\left\{\frac{\dot{P}(C \in \lambda)}{P(C \in \lambda)} + s(T_1|C \in \lambda) - \frac{\dot{P}(C \in \lambda|T_1)}{P(C \in \lambda|T_1)}\right\}'\right] \\
&\quad + E\left[\frac{1}{P(C \in \lambda)}E[m|T_1]\dot{P}(C \in \lambda|T_1)'\right]
\end{aligned}
$$

where the second line follows by using (36) to replace $s(T_1)$. The last line follows since, by using (10), we obtain that:

$$
\begin{aligned}
E\left[I(C \in \lambda)\sum_{r=1}^{R} I(C = r)\frac{\dot{P}(C = r|T_1)}{P(C = r|T_1)}\middle| T_1\right] &= \sum_{r \in \lambda} P(C = r|T_1)\frac{\dot{P}(C = r|T_1)}{P(C = r|T_1)} \\
&= \sum_{r \in \lambda} \dot{P}(C = r|T_1) = \dot{P}(C \in \lambda|T_1).
\end{aligned}
$$

Hence, now by repeatedly using (10) (e.g., first term on RHS of second equality) we obtain that:

$$\begin{aligned}
E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1]S(O)'\right] &= E\left[E[m|T_1]|C\in\lambda\right]\frac{\dot{P}(C\in\lambda)'}{P(C\in\lambda)}+E\left[E[m|T_1]s(T_1|C\in\lambda)'|C\in\lambda\right]\\
&\quad - E\left[E[m|T_1]\frac{\dot{P}(C\in\lambda|T_1)'}{P(C\in\lambda)}\right]+E\left[E[m|T_1]\frac{\dot{P}(C\in\lambda|T_1)'}{P(C\in\lambda)}\right]\\
&= E\left[m|C\in\lambda\right]\frac{\dot{P}(C\in\lambda)'}{P(C\in\lambda)}+E\left[E[m|T_1]s(T_1|C\in\lambda)'|C\in\lambda\right]+0\\
&= 0+E[ms(T_1|C\in\lambda)'|C\in\lambda]+0 \qquad (38)
\end{aligned}$$

where the first zero in last line follows from (3). The second term follows by using (10) and noting that $E\left[E[m|T_1]s(T_1|C\in\lambda)'|C\in\lambda\right]=E\left[E[ms(T_1|C\in\lambda)'|T_1,C\in\lambda]|C\in\lambda\right]=E\left[ms(T_1|C\in\lambda)'|C\in\lambda\right]$.

Now consider the r-th term of $\varphi^{\mathrm{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ for $r=2,\ldots,R$. By taking expectation conditional on $T_{r-1}\equiv(Z_{(1)},\ldots,Z_{(r-1)})$, and using (10) we obtain that:

$$\begin{aligned}
&E\left[\frac{P(C\in\lambda|T_1)}{P(C\in\lambda)}(E[m|T_r]-E[m|T_{r-1}])S(O)'\right]\\
=\;&E\left[\frac{P(C\in\lambda|Z_1)}{P(C\in\lambda)}(E[m|T_r]-E[m|T_{r-1}])\sum_{s=r}^{R}s(Z_{(s)}|T_{s-1})\right]\\
=\;&E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_r]s(Z_{(r)}|T_{r-1})'\right]\\
=\;&E\left[ms(Z_{(r)}|T_{r-1})'|C\in\lambda\right] \qquad (39)
\end{aligned}$$

by using that $s(Z_{(s)}|T_{s-1})\in L^2_0(F(Z_{(s)}|T_{s-1}))$ for $s=r,\ldots,R$ by definition, and by (10).

Therefore, (38) and (39) verify (37). That $\varphi^{\mathrm{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ belongs to $\mathcal{T}$ in (35) can be shown as follows. (i) Match the term $a(Z_{(1)},\ldots,Z_{(r)})$ in $\mathcal{T}$ with the r-th term of $\varphi^{\mathrm{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ for $r>1$. (ii) Distribute the first term $s(Z_{(1)})$ in $\mathcal{T}$ according to the relation (36) and match the term $I(C\in\lambda)s(Z_{(1)}|C\in\lambda)$ with the first term of $\varphi^{\mathrm{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ while keeping in mind that, by definition, $s(Z_{(1)}|C\in\lambda)\in L^2_0(F(Z_{(1)}|C\in\lambda))$. It is straightforward to verify that all the corresponding conditional expectations, as required by the definition in (35) and also (36), are zeros. The remaining terms in $\mathcal{T}$ (including the one due to distributing the terms in (ii)) are represented in $\varphi^{\mathrm{CMAR}}_{\lambda[u]}(O;\beta^0_\lambda)$ by zeros. ■

Proof of Proposition 4:

The references in the steps of this proof are mainly to that of Proposition 3 (i.e., effectively to that of Proposition 1) and to that of Proposition 5. To avoid notational clutter, when convenient, we write $m(Z;\beta^0_\lambda)$ as m; $T_r(Z)$ as T_r for $r=1,\ldots,R$; and also write $s(Z_{(r)}|Z_{(1)},\ldots,Z_{(r-1)})$ as $s(Z_{(r)}|T_{r-1})$ for $r=2,\ldots,R$.

As before, we obtain the score function for a parametric sub-model indexed by θ as:

$$S_\theta(O) = s_\theta(T_1) + \sum_{r=2}^{R} I(C \geq r) s_\theta(Z_{(r)}|T_{r-1}) + \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|T_1)} \left(\frac{\partial P(C=r|T_1;\gamma^0)}{\partial \gamma'} \frac{\partial \gamma^0}{\partial \theta'} \right)'.$$

Recall that $S_\gamma(C|T_1) := \sum_{r=1}^{R} \frac{I(C=r)}{P(C=r|T_1)} \frac{\partial}{\partial \gamma} P(C = r|T_1;\gamma^0)$. Let b denote constant matrices of dimension same as that of $\frac{\partial \gamma^0}{\partial \theta'}$. Then, the tangent set for the model is characterized by the set of functions:

$$\mathcal{T} := a_1(T_1) + b' S_\gamma(C|T_1) + \sum_{r=2}^{R} I(C \geq r) a_r(T_r),$$

where $a_1(T_1) \in L_0^2(F(T_1))$, $S_\gamma(C|T_1) \in L_0^2(F(C|T_1))$ and $a_r(T_r) \in L_0^2(F(Z_{(r)}|T_{r-1}))$.

Recognizing that $P(C = r|T_1) = P(C = r|T_1;\gamma^0)$ is known up to the finite (d_γ) dimensional parameter γ, alters the relationship in (36) as follows:

$$\begin{aligned} & s(T_1) + \frac{\partial \gamma^{0\prime}}{\partial \theta} \left[I(C\in\lambda) \frac{\frac{\partial}{\partial \gamma} P(C\in\lambda|T_1;\gamma^0)}{P(C\in\lambda|T_1)} + I(C\notin\lambda) \frac{\frac{\partial}{\partial \gamma} P(C\notin\lambda|T_1;\gamma^0)}{P(C\notin\lambda|T_1)} \right] \\ = \;& I(C\in\lambda) \left[\frac{\dot{P}(C\in\lambda)}{P(C\in\lambda)} + s(T_1|C\in\lambda) \right] + I(C\notin\lambda) \left[\frac{\dot{P}(C\notin\lambda)}{P(C\notin\lambda)} + s(T_1|C\notin\lambda) \right]. \end{aligned}$$

As before, differentiating (3) (equivalently, (4)) under the integral with respect to θ, and using the above relationship gives:

$$\begin{aligned} & \frac{\partial \beta_\lambda^0(\theta_0)}{\partial \theta'} \\ = \;& -M_\lambda^{-1} E\left[\frac{P(C\in\lambda|T_1)}{P(C\in\lambda)} m \left\{ s(T_1)' + \sum_{r=2}^{R} s(Z_{(r)}|T_{r-1})' \right\} \right] - M_\lambda^{-1} E\left[E[m|T_1] \frac{\frac{\partial}{\partial \gamma'} P(C\in\lambda|T_1;\gamma^0)}{P(C\in\lambda)} \frac{\partial \gamma^0}{\partial \theta'} \right]. \end{aligned}$$

Therefore, utilizing the expression of the efficient influence function in Proposition 3 and its relation to that in Proposition 4, the verification of pathwise differentiability reduces to verifying that:

$$E\left[\Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)} E[m|T_1] \middle| S_\gamma(C|T_1) \right) S(O)' \right] = E\left[E[m|T_1] \frac{\frac{\partial}{\partial \gamma'} P(C\in\lambda|T_1;\gamma^0)}{P(C\in\lambda)} \frac{\partial \gamma^0}{\partial \theta'} \right].$$

Note that $E\left[S_\gamma(C|T_1) \left\{ s(T_1)' + \sum_{r=2}^{R} I(C\geq r) s(Z_{(r)}|T_{r-1})' \right\} \right] = 0$. This follows (term by term) by using $E[S_\gamma(C|T_1)|T_1] = 0$ for term one; and then, for the other terms $r = 2, \ldots, R$, by noting that (10) implies that $E\left[S_\gamma(C|T_1) I(C \geq r) s(Z_{(r)}|T_{r-1})'\right] = E\left[S_\gamma(C|T_1)(1 - I(C \leq r-1)) s(Z_{(r)}|T_{r-1})'\right] = E\left[S_\gamma(C|T_1)(1 - P(C \leq r-1|T_1)) s(Z_{(r)}|T_{r-1})'\right] = 0$ since $s(Z_{(r)}|T_{r-1}) \in L_0^2(F(Z_{(r)}|T_{r-1})$.

Therefore, using the expression for $S(O)$, it follows that in the above equation (that contains the equality relationship to be verified), the LHS simplifies as:

$$
\begin{aligned}
LHS &= E\left[\Pi\left(\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1]\middle| S_\gamma(C|T_1)\right)S_\gamma(C|T_1)'\right]\frac{\partial\gamma^0}{\partial\theta'}\\
&= E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1]S_\gamma(C|T_1)'\right]\frac{\partial\gamma^0}{\partial\theta'}\\
&= E\left[\frac{I(C\in\lambda)}{P(C\in\lambda)}E[m|T_1]\sum_{r=1}^{R}\frac{I(C=r)}{P(C=r|T_1)}\frac{\partial P(C=r|T_1;\gamma^0)}{\partial\gamma'}\right]\frac{\partial\gamma^0}{\partial\theta'}\\
&= E\left[\frac{1}{P(C\in\lambda)}E[m|T_1]\sum_{r\in\lambda}\frac{P(C=r|T_1)}{P(C=r|T_1)}\frac{\partial P(C=r|T_1;\gamma^0)}{\partial\gamma'}\right]\frac{\partial\gamma^0}{\partial\theta'}\\
&= E\left[\frac{1}{P(C\in\lambda)}E[m|T_1]\frac{\partial P(C\in\lambda|T_1;\gamma^0)}{\partial\gamma'}\right]\frac{\partial\gamma^0}{\partial\theta'}\\
&= RHS.\ \blacksquare
\end{aligned}
$$

Proofs of Corollary 6, 7, 8: Straightforward but tedious manipulations of the results of Propositions 3 and 5 give Corollaries 7 and 8 respectively [see Chaudhuri (2014) for the proof of the latter]. Corollary 6 follows by imposing INDEP on the result of either Proposition 3 or Proposition 5. $\blacksquare$

Appendix C: Generalized method of moments (GMM) estimation of β^0_λ

Sections C.1, C.4 and part of C.5 in Appendix C collect materials related to efficient estimation that were not presented in our paper. On the other hand, the materials in Sections C.2 and C.3 were presented in our paper but again presented here to make Appendix C self-contained. The proofs of all the results that were presented in abridged form in our paper are presented here in detail.

C.1 This GMM estimation is a special case of Ai and Chen (2012)

Recall that Proposition 2 shows that under (1), (2) and assumption A, the efficient influence function and the efficiency bound for the estimation of β^0_λ based on (3) are identical to those based on the sequential moment restrictions (8)-(9). Hence, one could perform the efficient GMM estimation of β^0_λ simply as a special case of the optimally weighted orthogonalized sieve minimum distance (SMD) estimator that was proposed by Ai and Chen (2012) in a more general context.

To see the connection with Ai and Chen (2012) more clearly, note that our unconditional moment restriction in (8) corresponds to equation (1) in Ai and Chen (2012) with their conditioning variable $X^{(1)}$ taken as a constant. Now, the simplifications for our setup follow because, unlike Ai and Chen (2012), we do not have any unknown nuisance parameters (thanks to (2)) and because in our setup β_λ only enters the unconditional moment restrictions. That is, in our setup the moment restrictions in (9) turn out to be truly auxiliary whose sole purpose is to assist in obtaining efficiency gains.

This results in equation (10) of Ai and Chen (2012) (using their notation) to become:

$$\begin{aligned} \alpha_0 &:= \inf_{\alpha \in \Theta} E\left\{m_1(X^{(1)}, \alpha)' \Sigma_{01}(X^{(1)})^{-1} m_1(X^{(1)}, \alpha)\right\}, \qquad (40) \\ \text{where} \quad m_1(X^{(1)}, \alpha) &:= E\left[\varepsilon_1(Z, \alpha) | X^{(1)}\right] = E\left[\varepsilon_1(Z, \alpha)\right], \\ \Sigma_{01}(X^{(1)}) &:= E\left[\varepsilon_1(Z, \alpha_0)\varepsilon_1(Z, \alpha_0)' | X^{(1)}\right] = E\left[\varepsilon_1(Z, \alpha_0)\varepsilon_1(Z, \alpha_0)'\right], \end{aligned}$$

i.e.,

$$\alpha_0 := \inf_{\alpha \in \Theta} E\left[\varepsilon_1(Z, \alpha)'\right]\left(E\left[\varepsilon_1(Z, \alpha_0)\varepsilon_1(Z, \alpha_0)' | X^{(1)}\right]\right)^{-1} E\left[\varepsilon_1(Z, \alpha)\right].$$

Now, note that $\varepsilon_1(Z, \alpha)$ is Ai and Chen (2012)'s sequentially orthogonalized moment vector, i.e.,

$$\varepsilon_1(Z, \alpha) := \rho_1(Z; \alpha) - \sum_{t=2}^{T} \Gamma_{1,t}(X^{(t)})\varepsilon_t(Z, \alpha)$$

where $\varepsilon_T(Z; \alpha) := \rho_T(Z, \alpha)$ and for $t = 2, \ldots, T-1$, $\varepsilon_t(Z, \alpha)$ are the orthogonalized residuals:

$$\begin{aligned} \varepsilon_t(Z, \alpha) &:= \rho_t(Z; \alpha) - \sum_{s=t+1}^{T} \Gamma_{t,s}(X^{(s)})\varepsilon_s(Z, \alpha), \\ \text{where} \quad \Gamma_{t,s}(X^{(s)}) &:= E\left[\rho_t(Z; \alpha_0)\varepsilon_s(Z; \alpha_0)' | X^{(s)}\right]\left(E\left[\varepsilon_s(Z; \alpha_0)\varepsilon_s(Z; \alpha_0)' | X^{(s)}\right]\right)^{-1}. \end{aligned}$$

Therefore, thanks to our Proposition 2, $\varepsilon_1(Z, \alpha)$ and $\Sigma_{01}(X^{(1)})$ in Ai and Chen (2012) are our $\varphi_\lambda(O; \beta)$ and $V_\lambda := Var(\varphi_\lambda(O; \beta^0))$ respectively. Accordingly, the optimally weighted orthogonalized SMD estimator in equation (11) of Ai and Chen (2012), that is based on the sample counterpart of (40), is identical to the GMM estimator that uses the average estimated $\varphi_\lambda(O; \beta)$ as the moment vector and an estimator of V_λ^{-1} as the weighting matrix. We say "estimated $\varphi_\lambda(O; \beta)$" because, as is clear from the definition of $\varepsilon_1(Z, \alpha)$ entering $m_1(X^{(1)}, \alpha) := E\left[\varepsilon_1(Z, \alpha)\right]$, this contains unknown conditional expectations (covariance and variances) as nuisance parameters that need to be estimated and, thereby, profiled out from the criterion function of the estimation of the parameter of interest.

The purpose of Section C.2 and C.3 below is to point out with some details that under this special case of Ai and Chen (2012) that is our setup, a key feature of $\varphi_\lambda(O; \beta)$ provides practically useful flexibility in the parametric or nonparametric estimation of these nuisance parameters.

This key feature is of independent interest even without any consideration of efficiency. Hence, for completeness, we will work under the setup of an over-identified model where β is $d_\beta \times 1$, $m(Z; \beta)$ is $d_m \times 1$ and $d_m \geq d_\beta$. However, it is important to remember that the results on efficiency bounds that were presented in our paper imposed the restriction that $d_\beta = d_m = d$ [see footnote 5]. In light of this, our discussion when $d_m > d_\beta$ can be viewed simply as highlighting the specialities of GMM estimation based on a special moment vector like $\varphi_\lambda(O;\beta)$. To avoid introducing new notation when $d_m > d_\beta$, we continue with the notation from our paper. It is easy to see how the expressions thus obtained below will simplify when $d_m = d_\beta$ to give the exact expressions presented in our paper.

C.2 Estimation framework and the key feature

To consolidate notation following Chen et al. (2003), and guided by (6), define a $d_m \times 1$ function:

$$g(O;\beta,h(\beta)) := \frac{I(C=R)}{P(C=R|T_R(Z))}\varphi_{R,\lambda}(O;\beta)+\sum_{r=1}^{R-1}\left[\frac{I(C\geq r)}{P(C\geq r|T_r(Z))}-\frac{I(C\geq r+1)}{P(C\geq r+1|T_{r+1}(Z))}\right]h_r(\beta) \tag{41}$$

where $h(\beta) = (h_1'(\beta),\dots h_{R-1}'(\beta))'$ are the unknown nuisance parameters, and each $h_r(\beta)$ belongs to a class of functions $(Z,\beta)\mapsto \mathbb{R}^{d_m}$, call it $\mathcal{H}_r(\beta)$, for $r = 1,\dots,R-1$. Let $\mathcal{H} := \{\mathcal{H}_1(\beta)\times\dots\times\mathcal{H}_{R-1}(\beta) : \beta\in\mathcal{B}\}$ be a vector space endowed with a pseudo-metric $\|.\|_{\mathcal{H}}$, which is the sup-norm metric with respect to the argument β and a pseudo-metric with respect to the other arguments.

$g(O;\beta,h(\beta)) = \varphi_\lambda(O;\beta)$ defined in (6) if $h_r(\beta) = \varphi_{r,\lambda}(O;\beta)$ for $r = 1,\dots,R-1$. Denote the true $h_r(\beta)$ as $h_r^0(\beta) := \varphi_{r,\lambda}(O;\beta)$ for $r = 1,\dots,R-1$. While this suggests restricting $h_r(\beta)$ as $(T_r(Z),\beta)\mapsto\mathbb{R}^{d_m}$ for $r = 1,\dots,R-1$, it turns out that letting $h_r(\beta)$ instead be a function of Z and β does not affect either consistency or asymptotic normality of the GMM estimator defined below.

In light of this discussion, now define the GMM average moment vector and its expectation as:

$$G_n(\beta,h(\beta)) := \frac{1}{n}\sum_{i=1}^{n} g(O_i;\beta,(h_{1,i}'(\beta),\dots,h_{R-1,i}'(\beta))') \text{ and } G(\beta,h(\beta)) := E\left[G_n(\beta,h(\beta))\right].$$

Then, given any standard parametric or nonparametric estimator $\widehat{h}(\beta)$ for $h(\beta)$ and any $d_m\times d_m$ symmetric weighting matrix W_n (possibly efficient), the GMM estimator $\widehat{\beta}_\lambda(W_n)$ of β_λ^0 is defined as:

$$\widehat{\beta}_\lambda(W_n) \approx \arg\min_{\beta\in\mathcal{B}} G_n(\beta,\widehat{h}(\beta))'W_nG_n(\beta,\widehat{h}(\beta)). \tag{42}$$

The key feature of our setup is the identity that for any $\beta\in\mathcal{B}$ and any $h(.)\in\mathcal{H}$ (that need not be $h(\beta)$):

$$G(\beta,h(.)) = E[\varphi_{R,\lambda}(O;\beta)] = E[m(Z;\beta)|C\in\lambda] \tag{43}$$

by (4), (1) and (41). That is, $G(\beta,h(.))$ does not depend on $h(.)\in\mathcal{H}$. Its main implications are:

(F1) $G(\beta_\lambda^0,h(.)) = 0$ for any $h(.)\in\mathcal{H}$ by also using (3). Also, for any $\beta\in\mathcal{B}$ and any $h(.),\bar{h}(.)\in\mathcal{H}$:

$$G(\beta, h(.)) - G(\beta^0_\lambda, \bar{h}(.)) = 0 \iff E[m(Z;\beta)|C \in \lambda] - E[m(Z;\beta^0_\lambda)|C \in \lambda] = 0 \iff \beta = \beta^0_\lambda.$$

(F2) The partial derivative of $G(\beta, h(\beta))$ with respect to β, denote it by $G_\beta(\beta, h(\beta))$, satisfies $G_\beta(\beta, h(\beta)) = M_\lambda(\beta) := \frac{\partial}{\partial \beta'} E\left[m(Z;\beta)|C \in \lambda\right]$, and it exists whenever $M_\lambda(\beta)$ exists.

(F3) $G(\beta, h(.)) - G(\beta, \bar{h}(.)) = 0$ for any $\beta \in \mathcal{B}$ and $h(.), \bar{h}(.) \in \mathcal{H}$. Thus, the pathwise derivative of $G(\beta, h(.))$ with respect to $h(.)$, denote it by $G_h(\beta, h(.))$, exists at all $h(.) \in \mathcal{H}$, in all directions $[\bar{h}(.) - h(.)]$ for $\{h(.) + \tau(\bar{h}(.) - h(.)) : \tau \in [0,1]\} \subset \mathcal{H}$, and satisfies $G_h(\beta, h(.))[\bar{h}(.) - h(.)] = 0$.

(F1) helps to verify the well-separability (of the true β) assumption for consistent estimation of β^0_λ by $\widehat{\beta}_\lambda(W_n)$. It is even stronger since it indicates that $\widehat{h}(\beta)$ need not converge in probability to the true $h^0(\beta)$ but can converge to any $h^\dagger(\beta) \in \text{interior}(\mathcal{H})$ without affecting the consistency of $\widehat{\beta}_\lambda(W_n)$ for β^0_λ [see Proposition 13]. (F2) simplifies the Jacobian formula (and its estimation) in the asymptotic variance of $\widehat{\beta}_\lambda(W_n)$ since it implies that $G_\beta(\beta^0_\lambda, h(\beta^0_\lambda)) = M_\lambda$. Finally, while it was already clear from (F1) that the asymptotic orthogonality condition, Assumption N(c), in Andrews (1994) is satisfied following his equations (4.9)-(4.11) if $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$ for any $h^\dagger(\beta) \in \text{interior}(\mathcal{H})$; (F3) is still stated in a way that makes it more convenient for us to verify condition (4.1.4) in Theorem 4.1 of Chen (2007). (Proofs of the results stated below proceed by verifying the conditions in Chen et al. (2003) or Chen (2007).) Hence, the asymptotic variance of $\widehat{\beta}_\lambda(W_n)$ is unaffected by the estimation of $h(\beta)$ even if $\widehat{h}(\beta)$ converges at a rate slower than $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(n^{-1/4})$; for example, $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$ will suffice. See Remark 2(iii) in Chen et al. (2003) and Theorem 5 in Cattaneo (2010). The scenario is actually stronger here since we do not even require that $h^\dagger(\beta) = h^0(\beta)$, the truth [see Proposition 14]. Of course, semiparametric efficiency for $\widehat{\beta}_\lambda(W_n)$ requires that $h^\dagger(\beta^0_\lambda) = h^0(\beta^0_\lambda)$, but the rate of convergence of the consistent $\widehat{h}(\beta)$ is still of no consequence as far as the first-order asymptotic properties of GMM estimators are concerned [see Corollary 15]. Naturally, all these nice implications of (43) also provide flexibility in estimating the nuisance parameters – (i) parametrically based on misspecified models, e.g., giving linear projections rather than conditional expectations or (ii) nonparametrically under less than satisfactory conditions that might prevent a faster than $n^{1/4}$-rate convergence of the estimator.

C.3 Asymptotic properties of the GMM estimator in (42)

For simplicity we follow Chen et al. (2003) and write $(\beta, h(\beta))$ as (β, h) unless confusing. Also, define $\|A\|_B := \sqrt{\text{trace}(A'BA)}$ for conformable matrices A and B. Write $\|A\| \equiv \|A\|_B$ if B is identity.

Proposition 13 *Let* (1), (3), *and assumptions (A1) and (A2) hold. Let* $\{W_n\}$ *be a* $d_m \times d_m$ *positive semidefinite matrix such that* $W_n = W + o_p(1)$ *where* W *is a constant positive definite matrix.*

Assume:

(B1) $\|G_n(\widehat{\beta}_\lambda(W_n),\widehat{h})\|_{W_n} \leq \inf_{\beta\in\mathcal{B}} \|G_n(\beta,\widehat{h})\|_{W_n} + o_p(1)$ *where* $\mathcal{B}$ *is a compact subset of* $\mathbb{R}^{d_\beta}$*;*

(B2) $\|\widehat{h}(\beta)-h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$ *for some* $h^\dagger(\beta) \in$ *interior*$(\mathcal{H})$ *for all* β*, and* $h^\dagger(\beta)$ *not necessarily equal to* $h^0(\beta)$*;*

(B3) for all sequences of positive numbers $\{\delta_n\}$ *with* $\delta_n = o(1)$*,*

$$\sup_{\beta\in\mathcal{B},\|h-h^\dagger(\beta)\|_{\mathcal{H}}\leq\delta_n} \frac{\|G_n(\beta,h)-G(\beta,h)\|}{1+\|G_n(\beta,h)\|+\|G(\beta,h)\|} = o_p(1).$$

Then $\widehat{\beta}_\lambda(W_n) - \beta^0_\lambda = o_p(1)$.

Proposition 14 *Let* (1), (3) *and assumptions A hold. Let* $\{W_n\}$ *be a* $d_m \times d_m$ *positive semidefinite matrix such that* $W_n = W + o_p(1)$ *where* W *is a constant positive definite matrix. Let* $\beta^0_\lambda \in$ *interior*$(\mathcal{B})$ *and* $h^\dagger(\beta) \in$ *interior*$(\mathcal{H})$ *for all* β*, but* $h^\dagger(\beta)$ *not necessarily equal to* $h^0(\beta)$*. For a small* $\delta > 0$ *define the neighborhoods* $\mathcal{B}_\delta := \{\beta \in \mathcal{B} : \|\beta-\beta^0_\lambda\| \leq \delta\}$ *and* $\mathcal{H}_\delta := \{h \in \mathcal{H} : \|h - h^\dagger(\beta)\|_{\mathcal{H}} \leq \delta\}$*. (Nothing changes if the sup-norm with respect to* β *in* $\|.\|_{\mathcal{H}}$ *is alternatively defined to be taken locally over* $\beta \in \mathcal{B}_\delta$ *instead* $\beta \in \mathcal{B}$*; see Chen et al. (2003).) Let* $\widehat{\beta}_\lambda(W_n)-\beta^0_\lambda = o_p(1)$ *and* $\|\widehat{h}(\beta)-h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$*. Assume:*

(C1) $\|G_n(\widehat{\beta}_\lambda(W_n),\widehat{h})\|_{W_n} \leq \inf_{\beta\in\mathcal{B}_\delta} \|G_n(\beta,\widehat{h})\|_{W_n} + o_p(n^{-1/2})$*;*

(C2) $G_\beta(\beta,h^\dagger)$ *exists for* $\beta \in \mathcal{B}_\delta$ *and is continuous at* $\beta = \beta^0_\lambda$ *(*$G_\beta(\beta^0_\lambda,h^\dagger)$ *is full column rank by (A3) and (F2));*

(C3) for all sequences of positive numbers $\{\delta_n\}$ *with* $\delta_n = o(1)$*,*

$$\sup_{\beta\in\mathcal{B}_{\delta_n},h\in\mathcal{H}_{\delta_n}} \frac{\|G_n(\beta,h)-G(\beta,h)-G_n(\beta^0_\lambda,h^\dagger)\|}{n^{-1/2}+\|G_n(\beta,h)\|+\|G(\beta,h)\|} = o_p(1);$$

(C4) $\sqrt{n}G_n(\beta^0_\lambda,h^\dagger) \xrightarrow{d} N(0,\Sigma)$ *where* $\Sigma := E\left[g(O;(\beta^0_\lambda,h^\dagger))g(O;(\beta^0_\lambda,h^\dagger))'\right]$ *is finite.*

Then, for $M_\lambda := M(\beta^0_\lambda)$ *defined in assumption (A3),* $R_\lambda := M'_\lambda W M_\lambda$ *and* $S_\lambda := M'_\lambda W\Sigma W M_\lambda$*,*

$$\sqrt{n}(\widehat{\beta}_\lambda(W_n)-\beta^0_\lambda) = -R^{-1}_\lambda M'_\lambda W\sqrt{n}G_n(\beta^0_\lambda,h^\dagger) + o_p(1) \xrightarrow{d} N\left(0, R^{-1}_\lambda S_\lambda R^{-1}_\lambda\right).$$

Remark: Propositions 13 and 14 respectively establish the consistency and asymptotic normality of the GMM estimator defined in (42). We focus on showing how the key feature (43) helps to satisfy some of the conditions from Theorem 1 in Chen et al. (2003) and Theorem 4.1 in Chen (2007). We assume their other conditions. Through its condition (4.1.4), as opposed to (4.1.4)', Theorem 4.1 in Chen (2007) broadens the scope of Theorem 2 in Chen et al. (2003). This is useful to highlight that Propositions 13 and 14 (and the subsequent results) do not depend on the rate of convergence $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$. Importantly, we allow $h^\dagger(\beta) \neq h^0(\beta)$ to emphasize that consistency and asymptotic unbiasedness of $\widehat{\beta}_\lambda(W_n)$ are robust to the estimation of the nuisance parameters $h(\beta)$ parametrically under misspecification or nonparametrically under less than satisfactory conditions.

Thus, the theoretical results confirm the intuitions from our discussion of the implications of the key feature, except for the final bit, i.e., on efficiency, that is to be confirmed by the following result.

Corollary 15 *Under the assumptions of Proposition 14:*

(1) if $W = \Sigma^{-1}$ then

$$\sqrt{n}(\widehat{\beta}_\lambda(W_n) - \beta^0_\lambda) = -\left(M'_\lambda \Sigma^{-1} M_\lambda\right)^{-1} M'_\lambda \Sigma^{-1} \sqrt{n} G_n(\beta^0_\lambda, h^\dagger) + o_p(1) \xrightarrow{d} N\left(0, \left(M'_\lambda \Sigma^{-1} M_\lambda\right)^{-1}\right);$$

(2) if, additionally, $h^\dagger(\beta^0_\lambda) = h^0(\beta^0_\lambda)$ then $\Sigma = V_\lambda$ as in Proposition 1, and letting $\widehat{\beta}_\lambda := \widehat{\beta}_\lambda(W_n)$,

$$\sqrt{n}(\widehat{\beta}_\lambda - \beta^0_\lambda) = -\left(M'_\lambda V_\lambda^{-1} M_\lambda\right)^{-1} M'_\lambda V_\lambda^{-1} \sqrt{n} G_n(\beta^0_\lambda, h^0) + o_p(1) \xrightarrow{d} N\left(0, \Omega_\lambda = \left(M'_\lambda V_\lambda^{-1} M_\lambda\right)^{-1}\right),$$

i.e., by Proposition 1, the estimator $\widehat{\beta}_\lambda$ becomes semiparametrically efficient when $d_\beta = d_m$.

Estimation of asymptotic variance: Consistent estimation of M_λ is simplified due to (F2) because one could completely ignore the unknown nuisance parameters and obtain an estimator by taking analytical derivative (if it exists) or numerical derivative only for the first term of $G_n(\beta, h)$. Consistency of $\widehat{M_\lambda}(\beta)$ for $M_\lambda(\beta)$ with numerical derivatives follows by Theorem 7.4 in Newey and McFadden (1994). Also see Section 5.3 of Cattaneo (2010).

Standard conditions, e.g., $g(O_i; (\beta, h))$ is continuous with probability approaching one in a neighborhood $\mathcal{N}$ of $(\beta^0_\lambda, h^\dagger)$ and $E\left[\sup_{(\beta,h)\in\mathcal{N}} \|g(O_i; (\beta, h))\|^2\right] < \infty$ [see Lemma 4.3 in Newey and McFadden (1994)], ensure that for any $\beta = \beta^0_\lambda + o_p(1)$ and $h(\beta)$ such that $\|h(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$ (suffices if the sup-norm in $\|.\|_{\mathcal{H}}$ with respect to β is only local), the estimator $\widehat{V}_\lambda(\beta, h) := \frac{1}{n}\sum_{i=1}^n g(O_i; (\beta, h)) g(O_i; (\beta, h))' = \Sigma + o_p(1)$. Thus, the estimator $\widehat{\Omega}_\lambda(\widehat{\beta}_\lambda, \widehat{h}) := \left(\widehat{M'_\lambda}(\widehat{\beta}_\lambda) \widehat{V}_\lambda^{-1}(\widehat{\beta}_\lambda, \widehat{h}) \widehat{M_\lambda}(\widehat{\beta}_\lambda)\right)^{-1}$ is consistent for the asymptotic variance in Corollary 15(1). If $h^\dagger(\beta^0_\lambda) = h^0(\beta^0_\lambda)$ then $\Sigma = V_\lambda$, and now $\widehat{\Omega}_\lambda(\widehat{\beta}_\lambda, \widehat{h})$ will be consistent for the asymptotic variance Ω_λ in Corollary 15(2). Any consistent (for the appropriate limit) estimator $(\widetilde{\beta}, \widetilde{h})$ ensures consistency of all these quantities.

C.4 One step from the IPW estimator gives asymptotic equivalence with $\widehat{\beta}_\lambda$

The presence of β in possibly highly nonlinear form in all the R additive terms of the average moment vector $G_n(\beta, \widehat{h}(\beta))$ should not ideally be a drawback for computational purpose. If the GMM estimator has a closed form (e.g., Illustration 1 below) then this is not an issue. However, if there is no closed form expression (e.g., Illustration 2 below), one could start with an easy to compute $\sqrt{n}$-consistent estimator for β^0_λ and then update it in one step to obtain an estimator with the same asymptotic distribution as the estimator $\widehat{\beta}_\lambda$ in Corollary 15(2). For example, an IPW estimator based on the complete sub-sample $\{i = 1, \ldots, n : C_i = R\}$ and with the identity (or some simple) weighting matrix is relatively easy to compute:

$$\begin{aligned} \widetilde{\beta}_\lambda &:= \arg\min_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{I(C_i = R)}{P(C = R|T_R(Z_i))} \varphi_{R,\lambda}(O_i;\beta) \right\| \\ &\equiv \arg\min_{\beta\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{I(C_i = R)}{P(C = R|Z_i)} \frac{P(C \in \lambda|Z_i)}{\widehat{P}(C \in \lambda)} m(Z_i;\beta) \right\|. \end{aligned} \tag{44}$$

It is consistent under the assumptions of Proposition 13 [see, e.g., Wooldridge (2002)]. Built-in routines in standard statistical softwares can be directly used or slightly modified to obtain this estimator for a wide variety of the moment vector $m(Z;\beta)$ (e.g., Illustration 2 below). Now a one-step estimator of β^0_λ can be obtained by updating $\widetilde{\beta}_\lambda$ as:

$$\widehat{\beta}_{1\text{step}} = \widetilde{\beta}_\lambda - \widehat{\Omega}^{-1}_\lambda(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda))\widehat{M}'_\lambda(\widetilde{\beta}_\lambda)\widehat{V}^{-1}_\lambda(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda))G_n(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda)) \tag{45}$$

where $\widehat{h}(\widetilde{\beta}_\lambda)$ is a consistent estimator of $h^0(\beta^0_\lambda)$, and $\widehat{M}_\lambda(\widetilde{\beta}_\lambda)$, $\widehat{V}_\lambda(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda))$ and $\widehat{\Omega}_\lambda(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda))$, defined below Corollary 15, are consistent estimators for M_λ, V_λ and Ω_λ respectively under the conditions noted therein.[21]

Proposition 16 *Let all the conditions of Corollary 15(2) hold for $\widehat{\beta}_\lambda$, i.e., for the GMM estimator with the efficient weighting matrix. Additionally, let there be a first step estimator $\widetilde{\beta}_\lambda$ satisfying: $\sqrt{n}(\widetilde{\beta}_\lambda - \beta^0_\lambda) = O_p(1)$, $\widehat{M}_\lambda(\widetilde{\beta}_\lambda) = M_\lambda + o_p(1)$, $\widehat{V}_\lambda(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda)) = V_\lambda + o_p(1)$ and $\widehat{\Omega}_\lambda(\widetilde{\beta}_\lambda, \widehat{h}(\widetilde{\beta}_\lambda)) = \Omega_\lambda + o_p(1)$. For simplicity, assume a slightly stronger version of the stochastic equicontinuity condition (C3) [see Proposition 14] as: $\sup_{\beta\in\mathcal{B}_{\delta_n}, h\in\mathcal{H}_{\delta_n}} \sqrt{n}\|G_n(\beta, h) - G(\beta, h) - G_n(\beta^0_\lambda, h^0)\| = o_p(1)$. Then, $\widehat{\beta}_{1step}$ defined in* (45) *satisfies: $\sqrt{n}\left(\widehat{\beta}_{1step} - \widehat{\beta}_\lambda\right) = o_p(1)$.*

C.5 Illustration of the GMM estimator when $R = 3$

To focus on the main components, we abstract from the weighting matrix W_n by taking $d_m = d_\beta$. We consider two cases where the moment vector respectively corresponds to: (1) a linear regression giving a closed form expression for the efficient estimator, and (2) a linear quantile regression where

[21] While the one-step estimator in (45) could be easily modified to allow for the possibility that $\widehat{h}(\widetilde{\beta}_\lambda)$ converges in probability to $h^\dagger(\beta^0_\lambda)$ instead of the truth $h^0(\beta^0_\lambda)$, we do not consider this here since, as evident from Corollary 15, semiparametric efficiency is usually not achieved unless $h^\dagger(\beta^0_\lambda) = h^0(\beta^0_\lambda)$.

the efficient estimator is computed in one step as in (45). As for a concrete scenario with $R = 3$, it may be useful to keep in mind the setup of our Monte Carlo experiment in Section 5.

Illustration 1: Linear regression in the target population λ

Consider a moment vector of the form $m(Z;\beta) = X(y - X'\beta)$. For $i = 1,\dots,n$, let $T_{ji} := T_j(Z_i)$ for $j = 1,2,3$, $a_{3i} := I(C_i = 3)/P(C = 3|T_{3i})$, $a_{2i} := I(C_i \geq 2)/P(C \geq 2|T_{2i}) - a_{3i}$, $a_{1i} := 1 - a_{2i} - a_{3i}$, $q := P(C \in \lambda|T_3(Z))$ and $q_i := P(C \in \lambda|T_{3i})$. Simple computations give a closed form expression for the estimator $\widehat{\beta}_\lambda$ in (42) as:

$$\begin{aligned}\widehat{\beta}_\lambda \quad = \quad & \left(\sum_{i=1}^{n}\left\{a_{3i}q_iX_iX_i' + a_{2i}\widehat{E}\left[qXX'|T_{2i}\right] + a_{1i}\widehat{E}\left[qXX'|T_{1i}\right]\right\}\right)^{-1} \\ & \times \sum_{i=1}^{n}\left\{a_{3i}q_iX_iy_i + a_{2i}\widehat{E}\left[qXy|T_{2i}\right] + a_{1i}\widehat{E}\left[qXy|T_{1i}\right]\right\}\end{aligned}$$

where $\widehat{E}$ denotes the estimated conditional expectation. While one could factor out y_i from all three terms inside the last pair of braces under the setup of Section 5 (where y is always observed), our experience is that estimating the conditional expectations, e.g., $E\left[qXy|T_{2i}\right]$ directly instead of using the form $E\left[qX|T_{2i}\right]y_i$ leads to smaller variance of the estimator $\widehat{\beta}_\lambda$ in small samples.

Illustration 2: Linear quantile regression in the target population λ

Consider a moment vector of the form $m(Z;\beta) = X\left(\tau - I(y - X'\beta < 0)\right)$ for some fixed $\tau \in (0,1)$. (The notation $a_{3i}, a_{2i}, a_{1i}, q_i$ and q remain the same as in Illustration 1.) For any (β, h) define:

$$g(O_i;(\beta,h)) = a_{3i}q_im(T_{3i};\beta) + a_{2i}E[qm(T_3;\beta)|T_{2i}] + a_{1i}E[qm(T_3;\beta)|T_{1i}],$$

and accordingly define $g(O_i;(\beta,\widehat{h}))$ and $G_n(\beta,\widehat{h})$ replacing the conditional expectations in $g(O_i;(\beta,h))$ by their estimators. (The ignored common denominator $P(C \in \lambda)$ will be adjusted for in the final step.) Let $\widetilde{\beta}_\lambda$ denote the inefficient but $\sqrt{n}$-consistent estimator of β_λ^0 obtained from (44) by using this particular choice of the moment vector $m(Z;\beta)$. It is simple to obtain $\widetilde{\beta}_\lambda$ since commonly used statistical softwares provide built-in routine for weighted quantile regression which automatically gives the estimator with $(a_{3i}q_i/\sum_j a_{3j}q_j)_{i=1}^n$ as weights. Estimate M_λ where $M_\lambda(\beta) = -(\partial/\partial\beta')E\left[XI(y - X'\beta < 0)|C \in \lambda\right]$ using $\widetilde{\beta}_\lambda$ [see below Corollary 15]. Therefore, since $d_m = d_\beta$, by using (45) we obtain the one-step estimator as: $\widehat{\beta}_{\text{1step}} = \widetilde{\beta}_\lambda - \widehat{M}_\lambda^{-1}(\widetilde{\beta}_\lambda)G_n(\widetilde{\beta}_\lambda,\widehat{h}(\widetilde{\beta}_\lambda))/\widehat{P}(C \in \lambda)$.

C.6 Simulation evidence from Section 5 of the finite-sample properties of $\widehat{\beta}_\lambda$

Besides the efficient estimators based on various sub-samples, we also consider the complete case (CC) and IPW [see (44)] estimators. The CC estimator is the default in the statistical softwares and is based only on the complete sub-sample ignoring its likely unrepresentative of the target population.

We consider certain finite-sample properties of all these estimators and report them in Table 4 under INDEP, Tables 5 for Intercept and 6 for Slope under CMAR, and Tables 7 for Intercept and 8 for Slope under MAR. We focus on the following quantities computed as averages over the 10,000 Monte Carlo trials: Mbias (deviation from the true values), Abias (absolute deviation from the true values), Std (standard deviation obtained as $\sqrt{(\text{estimated Avar})/(\text{size of the used sample})}$) and IQR (interquartile range). Mean squared error is not reported but follows directly as Mbias^2+ Std^2.

The CC and IPW estimators are numerically equivalent if $\lambda = \{3\}$ or under INDEP. Otherwise, as expected, CC can be badly biased (Mbias) since it does not recognize the sample-selection.

The other estimators are consistent under our assumptions, and their small Mbias and decreasing (with n) Std support this. The ordering of the variability of the estimators, as measured by Abias, Std and IQR, are as expected: always the largest when the used sample is $\{3\}$, and the smallest when the used sample is $\{1, 2, 3\}$.

Comparison between the two estimators based on the used samples $\{1, 3\}$ and $\{2, 3\}$ is possible under INDEP or under CMAR and MAR if $\lambda = \{3\}$ or $\lambda = \{1, 2, 3\}$. In these cases, it seems that in spite of the poorer quality of information in the units of $\{1, 3\}$, its larger sample size makes it more desirable than $\{2, 3\}$. (Under our premise, $\{1, 3\}$ could still be less expensive than $\{2, 3\}$ to observe.)

Overall, under our simulation design all the estimators display good properties in finite samples, and thus lend credibility to the encouraging simulation results on the efficiency loss in Section 5.

Used	$n = 600$				$n = 1200$				$n = 1800$			
Sample	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR
$\{3\}$	-.0002	.0748	.0933	.1250	.0011	.0529	.0661	.0895	-.0003	.0436	.0540	.0739
$\{1, 3\}$	.0005	.0560	.0667	.0947	.0007	.0388	.0473	.0666	.0002	.0313	.0388	.0530
$\{2, 3\}$	-.0001	.0584	.0673	.0986	.0008	.0392	.0475	.0661	.0003	.0317	.0388	.0534
$\{1, 2, 3\}$	.0003	.0523	.0584	.0878	.0006	.0346	.0411	.0589	.0003	.0278	.0337	.0475
$\{3\}$	.0004	.0773	.0927	.1296	.0001	.0527	.0659	.0885	.0002	.0434	.0539	.0737
$\{1, 3\}$	.0090	.0641	.0714	.1069	.0038	.0425	.0510	.0715	.0028	.0345	.0418	.0579
$\{2, 3\}$	.0062	.0667	.0720	.1106	.0019	.0432	.0507	.0739	.0013	.0347	.0415	.0586
$\{1, 2, 3\}$	.0082	.0631	.0649	.1044	.0030	.0403	.0458	.0686	.0021	.0320	.0377	.0545

Table 4: Bias (Mbias), absolute bias (Abias), standard deviation (Std) and interquartile range (IQR) of the estimators under INDEP sampling are reported based on the average over 10,000 Monte Carlo trials. Target population $\lambda = \{1, 2, 3\}$. Top panel: Intercept parameter $\beta_{\lambda,1}$. Bottom panel: Slope parameter $\beta_{\lambda,2}$.

CMAR Sampling. Parameter of interest is the Intercept ($\beta_{\lambda,1}$)

Target	Used	$n = 600$				$n = 1200$				$n = 1800$			
Popln. (λ)	Sample (s)	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR
$\{1\}$	$\{3\}$: CC	-.1283	.1353	.0917	.1274	-.1283	.1297	.0649	.0870	-.1288	.1291	.0530	.0713
$\{1\}$	$\{3\}$: IPW	-.0018	.0767	.0946	.1305	-.0005	.0542	.0670	.0907	-.0008	.0437	.0547	.0737
$\{1\}$	$\{1,3\}$	-.0027	.0617	.0701	.1025	-.0013	.0413	.0498	.0690	-.0008	.0335	.0408	.0561
$\{1\}$	$\{1,2,3\}$	.0001	.0579	.0629	.0964	.0001	.0378	.0443	.0642	-.0001	.0301	.0363	.0510
$\{2\}$	$\{3\}$: CC	.2490	.2491	.0917	.1274	.2490	.2490	.0649	.0870	.2485	.2485	.0530	.0713
$\{2\}$	$\{3\}$: IPW	.0040	.0814	.1006	.1393	.0023	.0577	.0714	.0977	.0012	.0469	.0583	.0788
$\{2\}$	$\{2,3\}$	.0036	.0596	.0685	.0998	.0017	.0398	.0481	.0673	.0004	.0322	.0394	.0542
$\{2\}$	$\{1,2,3\}$	-.0014	.0565	.0615	.0965	-.0014	.0366	.0431	.0616	-.0015	.0298	.0353	.0510
$\{3\}$	$\{3\}$: CC	.0005	.0742	.0917	.1274	.0005	.0523	.0649	.0870	.0000	.0423	.0530	.0713
$\{3\}$	$\{3\}$: IPW	.0005	.0742	.0402	.1274	.0005	.0523	.0285	.0870	.0000	.0423	.0233	.0713
$\{3\}$	$\{1,3\}$	-.0019	.0581	.0673	.0976	-.0011	.0389	.0475	.0650	-.0007	.0318	.0388	.0533
$\{3\}$	$\{2,3\}$	.0008	.0601	.0690	.1004	.0005	.0401	.0482	.0678	-.0003	.0320	.0394	.0539
$\{3\}$	$\{1,2,3\}$	.0001	.0518	.0579	.0863	-.0002	.0339	.0406	.0577	-.0004	.0274	.0332	.0463
$\{1,3\}$	$\{3\}$: CC	-.0914	.1074	.0917	.1274	-.0914	.0965	.0649	.0870	-.0919	.0938	.0530	.0713
$\{1,3\}$	$\{3\}$: IPW	-.0012	.0757	.0935	.1297	-.0002	.0535	.0662	.0893	-.0006	.0431	.0541	.0726
$\{1,3\}$	$\{1,3\}$	-.0025	.0603	.0690	.0999	-.0013	.0404	.0489	.0676	-.0008	.0328	.0401	.0554
$\{1,3\}$	$\{1,2,3\}$	-.0001	.0558	.0611	.0930	-.0001	.0364	.0430	.0618	-.0003	.0291	.0352	.0493
$\{2,3\}$	$\{3\}$: CC	.1440	.1483	.0917	.1274	.1440	.1447	.0649	.0870	.1435	.1436	.0530	.0713
$\{2,3\}$	$\{3\}$: IPW	.0026	.0770	.0952	.1317	.0016	.0544	.0675	.0919	.0008	.0441	.0551	.0746
$\{2,3\}$	$\{2,3\}$	.0018	.0583	.0673	.0978	.0009	.0389	.0472	.0659	-.0002	.0313	.0386	.0527
$\{2,3\}$	$\{1,2,3\}$	.0000	.0519	.0579	.0867	-.0004	.0339	.0406	.0579	-.0007	.0276	.0332	.0466
$\{1,2,3\}$	$\{3\}$: CC	.0092	.0744	.0917	.1274	.0092	.0528	.0649	.0870	.0087	.0428	.0530	.0713
$\{1,2,3\}$	$\{3\}$: IPW	.0004	.0761	.0939	.1308	.0007	.0536	.0666	.0894	.0001	.0434	.0544	.0733
$\{1,2,3\}$	$\{1,3\}$	-.0004	.0598	.0687	.0996	-.0002	.0400	.0486	.0670	-.0001	.0327	.0398	.0551
$\{1,2,3\}$	$\{2,3\}$	-.0013	.0619	.0711	.1033	-.0006	.0412	.0496	.0695	-.0013	.0328	.0404	.0549
$\{1,2,3\}$	$\{1,2,3\}$	.0000	.0531	.0587	.0884	-.0001	.0347	.0413	.0589	-.0004	.0280	.0338	.0472

Table 5: Bias (Mbias), absolute bias (Abias), standard deviation (Std) and interquartile range (IQR) of the estimators of the Intercept parameter ($\beta_{\lambda,1}$) under CMAR sampling are reported based on the average over 10,000 Monte Carlo trials. CC and IPW are different estimators.

CMAR Sampling. Parameter of interest is the Slope ($\beta_{\lambda,2}$)

Target	Used	$n = 600$				$n = 1200$				$n = 1800$			
Popln. (λ)	Sample (s)	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR
$\{1\}$	$\{3\}$: CC	-.0080	.0786	.0952	.1309	-.0050	.0546	.0677	.0924	-.0074	.0446	.0553	.0743
$\{1\}$	$\{3\}$: IPW	-.0027	.0846	.1008	.1422	.0014	.0592	.0727	.1005	-.0009	.0480	.0596	.0809
$\{1\}$	$\{1,3\}$	.0097	.0708	.0764	.1175	.0066	.0477	.0545	.0800	.0033	.0378	.0448	.0634
$\{1\}$	$\{1,2,3\}$	.0060	.0691	.0707	.1165	.0046	.0450	.0502	.0751	.0023	.0355	.0414	.0599
$\{2\}$	$\{3\}$: CC	.0232	.0801	.0952	.1309	.0262	.0582	.0677	.0924	.0238	.0480	.0553	.0743
$\{2\}$	$\{3\}$: IPW	-.0021	.0963	.1119	.1602	.0005	.0678	.0818	.1142	-.0018	.0546	.0675	.0911
$\{2\}$	$\{2,3\}$	.0068	.0726	.0760	.1218	.0045	.0480	.0545	.0803	.0028	.0386	.0451	.0647
$\{2\}$	$\{1,2,3\}$	.0031	.0748	.0755	.1260	.0018	.0486	.0533	.0821	.0007	.0385	.0440	.0648
$\{3\}$	$\{3\}$: CC	-.0012	.0782	.0952	.1309	.0018	.0544	.0677	.0924	-.0006	.0441	.0553	.0743
$\{3\}$	$\{3\}$: IPW	-.0012	.0782	.0417	.1309	.0018	.0544	.0297	.0924	-.0006	.0441	.0243	.0743
$\{3\}$	$\{1,3\}$	.0101	.0700	.0769	.1163	.0065	.0468	.0539	.0793	.0032	.0366	.0441	.0615
$\{3\}$	$\{2,3\}$	-.0009	.0729	.0777	.1224	.0010	.0467	.0543	.0786	-.0003	.0369	.0444	.0613
$\{3\}$	$\{1,2,3\}$	.0115	.0646	.0669	.1071	.0075	.0418	.0470	.0698	.0048	.0330	.0387	.0552
$\{1,3\}$	$\{3\}$: CC	-.0125	.0790	.0952	.1309	-.0095	.0550	.0677	.0924	-.0119	.0452	.0553	.0743
$\{1,3\}$	$\{3\}$: IPW	-.0023	.0820	.0984	.1381	.0016	.0572	.0706	.0973	-.0008	.0463	.0578	.0782
$\{1,3\}$	$\{1,3\}$	.0099	.0696	.0756	.1151	.0067	.0468	.0536	.0787	.0034	.0369	.0440	.0622
$\{1,3\}$	$\{1,2,3\}$	.0075	.0668	.0688	.1120	.0054	.0435	.0487	.0726	.0031	.0342	.0401	.0576
$\{2,3\}$	$\{3\}$: CC	-.0135	.0791	.0952	.1309	-.0105	.0551	.0677	.0924	-.0129	.0454	.0553	.0743
$\{2,3\}$	$\{3\}$: IPW	-.0017	.0854	.1016	.1426	.0011	.0596	.0733	.1010	-.0013	.0481	.0601	.0808
$\{2,3\}$	$\{2,3\}$	.0037	.0687	.0726	.1154	.0032	.0448	.0516	.0753	.0017	.0358	.0425	.0600
$\{2,3\}$	$\{1,2,3\}$	.0070	.0661	.0680	.1112	.0044	.0430	.0479	.0731	.0026	.0340	.0395	.0570
$\{1,2,3\}$	$\{3\}$: CC	-.0450	.0869	.0952	.1309	-.0420	.0643	.0677	.0924	-.0444	.0578	.0553	.0743
$\{1,2,3\}$	$\{3\}$: IPW	-.0025	.0801	.0967	.1335	.0012	.0558	.0692	.0947	-.0012	.0451	.0565	.0756
$\{1,2,3\}$	$\{1,3\}$	.0038	.0712	.0774	.1186	.0028	.0475	.0544	.0805	.0004	.0374	.0445	.0634
$\{1,2,3\}$	$\{2,3\}$	-.0037	.0745	.0786	.1240	-.0011	.0482	.0552	.0812	-.0021	.0382	.0453	.0641
$\{1,2,3\}$	$\{1,2,3\}$	.0066	.0628	.0648	.1048	.0047	.0409	.0458	.0691	.0027	.0322	.0377	.0548

Table 6: Bias (Mbias), absolute bias (Abias), standard deviation (Std) and interquartile range (IQR) of the estimators of the Slope parameter ($\beta_{\lambda,2}$) under CMAR sampling are reported based on the average over 10,000 Monte Carlo trials. CC and IPW are different estimators.

MAR Sampling. Parameter of interest is the Intercept ($\beta_{\lambda,1}$)

Target	Used	$n = 600$				$n = 1200$				$n = 1800$			
Popln. (λ)	Sample (s)	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR
{1}	{3}: CC	-.1380	.1438	.0926	.1236	-.1394	.1402	.0656	.0867	-.1380	.1382	.0536	.0735
{1}	{3}: IPW	-.0026	.0771	.0963	.1283	-.0024	.0541	.0683	.0906	-.0005	.0454	.0559	.0769
{1}	{1, 3}	-.0061	.0618	.0715	.1026	-.0030	.0426	.0510	.0720	-.0016	.0344	.0418	.0581
{1}	{1, 2, 3}	.0000	.0585	.0629	.0987	.0002	.0387	.0446	.0654	.0007	.0311	.0366	.0525
{2}	{3}: CC	.2371	.2376	.0926	.1236	.2357	.2357	.0656	.0867	.2371	.2371	.0536	.0735
{2}	{3}: IPW	.0038	.0849	.1043	.1432	.0010	.0587	.0742	.0981	.0011	.0490	.0607	.0828
{2}	{2, 3}	.0060	.0617	.0709	.1033	.0025	.0412	.0496	.0698	.0016	.0332	.0405	.0557
{2}	{1, 2, 3}	-.0016	.0595	.0641	.0998	-.0012	.0391	.0448	.0653	-.0015	.0311	.0368	.0525
{3}	{3}: CC	.0004	.0742	.0926	.1236	-.0010	.0517	.0656	.0867	.0004	.0435	.0536	.0735
{3}	{3}: IPW	.0004	.0742	.0405	.1236	-.0010	.0517	.0287	.0867	.0004	.0435	.0235	.0735
{3}	{1, 3}	-.0041	.0579	.0679	.0963	-.0021	.0397	.0480	.0667	-.0012	.0321	.0393	.0541
{3}	{2, 3}	.0011	.0626	.0699	.1037	.0003	.0408	.0488	.0687	.0010	.0333	.0399	.0558
{3}	{1, 2, 3}	-.0003	.0529	.0588	.0892	-.0003	.0348	.0411	.0590	.0001	.0281	.0337	.0476
{1, 3}	{3}: CC	-.0990	.1129	.0926	.1236	-.1004	.1038	.0656	.0867	-.0990	.1005	.0536	.0735
{1, 3}	{3}: IPW	-.0017	.0760	.0949	.1271	-.0020	.0532	.0673	.0895	-.0003	.0447	.0550	.0755
{1, 3}	{1, 3}	-.0056	.0602	.0700	.1010	-.0028	.0415	.0498	.0701	-.0015	.0335	.0408	.0562
{1, 3}	{1, 2, 3}	-.0002	.0564	.0613	.0956	.0000	.0373	.0433	.0632	.0005	.0300	.0355	.0507
{2, 3}	{3}: CC	.1343	.1411	.0926	.1236	.1329	.1339	.0656	.0867	.1343	.1345	.0536	.0735
{2, 3}	{3}: IPW	.0024	.0778	.0962	.1308	.0001	.0538	.0683	.0896	.0008	.0453	.0558	.0771
{2, 3}	{2, 3}	.0026	.0586	.0676	.0971	.0007	.0390	.0472	.0659	.0009	.0315	.0386	.0528
{2, 3}	{1, 2, 3}	.0003	.0515	.0579	.0861	-.0001	.0342	.0405	.0579	-.0002	.0274	.0331	.0463
{1, 2, 3}	{3}: CC	-.0005	.0742	.0926	.1236	-.0019	.0517	.0656	.0867	-.0005	.0435	.0536	.0735
{1, 2, 3}	{3}: IPW	.0000	.0767	.0955	.1276	-.0012	.0534	.0678	.0904	.0002	.0450	.0554	.0763
{1, 2, 3}	{1, 3}	-.0036	.0596	.0695	.0991	-.0018	.0411	.0493	.0695	-.0010	.0331	.0404	.0561
{1, 2, 3}	{2, 3}	-.0012	.0632	.0716	.1044	-.0011	.0416	.0498	.0707	.0000	.0335	.0406	.0567
{1, 2, 3}	{1, 2, 3}	.0000	.0531	.0587	.0896	.0000	.0354	.0414	.0602	.0002	.0282	.0339	.0478

Table 7: Bias (Mbias), absolute bias (Abias), standard deviation (Std) and interquartile range (IQR) of the estimators of the Intercept parameter ($\beta_{\lambda,1}$) under MAR sampling are reported based on the average over 10,000 Monte Carlo trials. CC and IPW are different estimators.

MAR Sampling. Parameter of interest is the Slope ($\beta_{\lambda,2}$)

Target	Used	$n = 600$				$n = 1200$				$n = 1800$			
Popln. (λ)	Sample (s)	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR	Mbias	Abias	Std	IQR
$\{1\}$	$\{3\}$: CC	-.0164	.0802	.0974	.1328	-.0146	.0573	.0692	.0949	-.0149	.0469	.0566	.0769
$\{1\}$	$\{3\}$: IPW	-.0019	.0843	.1015	.1428	.0004	.0596	.0729	.0997	.0004	.0484	.0599	.0816
$\{1\}$	$\{1,3\}$	.0109	.0714	.0771	.1193	.0064	.0475	.0550	.0795	.0039	.0379	.0452	.0639
$\{1\}$	$\{1,2,3\}$	.0072	.0696	.0709	.1167	.0040	.0452	.0502	.0753	.0021	.0359	.0414	.0612
$\{2\}$	$\{3\}$: CC	.0227	.0811	.0974	.1328	.0245	.0596	.0692	.0949	.0242	.0495	.0566	.0769
$\{2\}$	$\{3\}$: IPW	-.0023	.1096	.1238	.1859	.0008	.0777	.0920	.1307	-.0001	.0630	.0765	.1055
$\{2\}$	$\{2,3\}$	.0108	.0824	.0810	.1358	.0059	.0547	.0590	.0915	.0026	.0436	.0493	.0737
$\{2\}$	$\{1,2,3\}$	.0042	.0846	.0801	.1404	.0021	.0546	.0579	.0923	-.0001	.0437	.0483	.0740
$\{3\}$	$\{3\}$: CC	-.0007	.0791	.0974	.1328	.0011	.0561	.0692	.0949	.0008	.0454	.0566	.0769
$\{3\}$	$\{3\}$: IPW	-.0007	.0791	.0426	.1328	.0011	.0561	.0303	.0949	.0008	.0454	.0248	.0769
$\{3\}$	$\{1,3\}$	.0134	.0705	.0779	.1163	.0076	.0468	.0548	.0781	.0047	.0371	.0448	.0620
$\{3\}$	$\{2,3\}$	-.0014	.0788	.0826	.1281	-.0014	.0500	.0569	.0836	-.0015	.0395	.0465	.0662
$\{3\}$	$\{1,2,3\}$	.0136	.0679	.0698	.1121	.0075	.0436	.0488	.0727	.0046	.0347	.0401	.0581
$\{1,3\}$	$\{3\}$: CC	-.0162	.0802	.0974	.1328	-.0144	.0573	.0692	.0949	-.0147	.0469	.0566	.0769
$\{1,3\}$	$\{3\}$: IPW	-.0017	.0824	.0999	.1389	.0006	.0582	.0715	.0978	.0004	.0472	.0587	.0796
$\{1,3\}$	$\{1,3\}$	.0115	.0705	.0767	.1174	.0067	.0469	.0545	.0781	.0041	.0373	.0447	.0619
$\{1,3\}$	$\{1,2,3\}$	.0088	.0683	.0700	.1136	.0049	.0442	.0493	.0735	.0028	.0351	.0406	.0592
$\{2,3\}$	$\{3\}$: CC	-.0219	.0811	.0974	.1328	-.0201	.0584	.0692	.0949	-.0204	.0483	.0566	.0769
$\{2,3\}$	$\{3\}$: IPW	-.0015	.0906	.1067	.1511	.0011	.0642	.0777	.1087	.0004	.0520	.0641	.0882
$\{2,3\}$	$\{2,3\}$	.0055	.0727	.0740	.1204	.0030	.0477	.0530	.0801	.0012	.0381	.0439	.0644
$\{2,3\}$	$\{1,2,3\}$	.0085	.0696	.0686	.1152	.0047	.0448	.0490	.0737	.0022	.0359	.0406	.0608
$\{1,2,3\}$	$\{3\}$: CC	-.0534	.0904	.0974	.1328	-.0516	.0704	.0692	.0949	-.0519	.0631	.0566	.0769
$\{1,2,3\}$	$\{3\}$: IPW	-.0021	.0824	.0997	.1391	.0006	.0582	.0716	.0987	.0003	.0472	.0587	.0793
$\{1,2,3\}$	$\{1,3\}$	.0047	.0730	.0791	.1222	.0025	.0482	.0559	.0802	.0012	.0383	.0459	.0649
$\{1,2,3\}$	$\{2,3\}$	-.0037	.0787	.0808	.1283	-.0028	.0505	.0564	.0841	-.0028	.0402	.0465	.0670
$\{1,2,3\}$	$\{1,2,3\}$	.0079	.0647	.0652	.1075	.0045	.0416	.0463	.0696	.0024	.0333	.0382	.0566

Table 8: Bias (Mbias), absolute bias (Abias), standard deviation (Std) and interquartile range (IQR) of the estimators of the Slope parameter ($\beta_{\lambda,2}$) under MAR sampling are reported based on the average over 10,000 Monte Carlo trials. CC and IPW are different estimators.

C.7 Proofs

For simplicity, we write β_λ as β. We follow the steps of the proof for Theorems 1 and 2 in Chen et al. (2003) with adjustments for the weaker conditions that are consequences of (43) [see (F1)-(F3)]. The main adjustment is that we allow $\|\widehat{h} - h^\dagger\|_{\mathcal{H}} = o_p(1)$ where $h^\dagger \in \mathcal{H}$ need not be h^0.

Proof of Proposition 13: (F1) already implies the standard well-separability of β^0 by virtue of (3). Hence, for all $\delta > 0$ there exists $\epsilon(\delta) > 0$ such that $P(\|\widehat{\beta} - \beta^0\| > \delta) \leq P(\|G(\widehat{\beta}, h^\dagger)\| \geq \epsilon(\delta))$.

Therefore, to establish that $\widehat{\beta} \xrightarrow{P} \beta^0$, it is sufficient to show that $\|G(\widehat{\beta}, h^\dagger)\| = o_p(1)$. Assumption (B2) implies that $P(\widehat{h}(\beta) \in \mathcal{H}) \to 1$ uniformly in $\beta \in \mathcal{B}$ as $n \to \infty$. The rest of the proof works conditional on the sequence of events $\{\widehat{h}(\widehat{\beta}) \in \mathcal{H}\}$, i.e., we use the fact that:

$$\begin{aligned}
& P(\|G(\widehat{\beta}, h^\dagger)\| < \epsilon(\delta)) \\
= \; & P(\|G(\widehat{\beta}, h^\dagger)\| < \epsilon(\delta) | \widehat{h}(\widehat{\beta}) \in \mathcal{H}) P(\widehat{h}(\widehat{\beta}) \in \mathcal{H}) + P(\|G(\widehat{\beta}, h^\dagger)\| < \epsilon(\delta) | \widehat{h}(\widehat{\beta}) \notin \mathcal{H}) P(\widehat{h}(\widehat{\beta}) \notin \mathcal{H}) \\
= \; & P(\|G(\widehat{\beta}, h^\dagger)\| < \epsilon(\delta) | \widehat{h}(\widehat{\beta}) \in \mathcal{H}) + o(1) \qquad (46)
\end{aligned}$$

as $n \to \infty$ and, instead, show that $\|G(\widehat{\beta}, h^\dagger)\| = o_p(1)$ conditional on $\{\widehat{h}(\widehat{\beta}) \in \mathcal{H}\}$.

To this end, first note that:

$$\begin{aligned}
\|G(\widehat{\beta}, h^\dagger)\| & \leq \|G(\widehat{\beta}, h^\dagger) - G(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, \widehat{h}) - G_n(\widehat{\beta}, \widehat{h})\| + \|G_n(\widehat{\beta}, \widehat{h})\| \\
& = \|G(\widehat{\beta}, \widehat{h}) - G_n(\widehat{\beta}, \widehat{h})\| + \|G_n(\widehat{\beta}, \widehat{h})\|. \qquad (47)
\end{aligned}$$

The inequality holds by the triangle inequality (kept implicit hereafter). The equality holds by (F3).

Using (B3) and then (F3), we obtain:

$$\|G(\widehat{\beta}, \widehat{h}) - G_n(\widehat{\beta}, \widehat{h})\| \leq o_p(1)\{1 + \|G_n(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, \widehat{h})\|\} \leq o_p(1)\{1 + \|G_n(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, h^\dagger)\|\}.$$

Using this along with (47) gives:

$$\begin{aligned}
& \|G(\widehat{\beta}, h^\dagger)\| \times (1 - o_p(1)) \\
\leq \; & o_p(1) + \|G_n(\widehat{\beta}, \widehat{h})\| \times (1 + o_p(1)) \\
\leq \; & o_p(1) + \|G_n(\widehat{\beta}, \widehat{h})\|_{W_n} \times (1 + \|W_n^{-1} - W^{-1}\| + \|W^{-1} - I_{d_m}\|) \times (1 + o_p(1)) \\
= \; & o_p(1) + \|G_n(\widehat{\beta}, \widehat{h})\|_{W_n} \times (c + o_p(1)) \\
\leq \; & o_p(1) + \inf_{\beta \in \mathcal{B}} \|G_n(\beta, \widehat{h})\|_{W_n} \times (c + o_p(1)) \qquad (48)
\end{aligned}$$

where $c = 1 + \|W^{-1} - I_{d_m}\|$. The equality in the above equations follows since (i) $W_n - W = o_p(1)$ for a constant positive definite matrix W implies that W_n^{-1} exists with probability approaching one and $W_n^{-1} - W^{-1} = o_p(1)$, and hence $\|W_n^{-1} - W^{-1}\| = o_p(1)$ as d_m is finite, (ii) a finite and positive definite W and a finite d_m imply that $c(> 1)$ is finite. The last inequality in (48) is due to (B1).

Following similar steps again and letting $d = 1 + \|W - I_{d_m}\|$ (> 1 and finite), note that:

$$\begin{aligned} & \|G_n(\beta, \widehat{h})\|_{W_n} \\ \leq\ & \|G_n(\beta, \widehat{h})\| \times (d + o_p(1)) \\ \leq\ & \{\|G_n(\beta, \widehat{h}) - G(\beta, \widehat{h})\| + \|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| + \|G(\beta, h^\dagger) - G(\beta^0, h^\dagger)\|\} \times (d + o_p(1)) \end{aligned} \tag{49}$$

by using (43), i.e., $G(\beta^0, h) = 0$ for all $h \in \mathcal{H}$ (in the last term inside the braces). This is the special feature of our setup; whereas this holds only at $h = h^0$ in Chen et al. (2003). On the other hand, $\|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| = 0$ by (F3). Lastly, since $G(\beta^0, h^\dagger) = 0$ and $\|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| = 0$, we can use (B3) as before to obtain that:

$$\begin{aligned} \|G_n(\beta, \widehat{h}) - G(\beta, \widehat{h})\| &\leq o_p(1)\{1 + \|G_n(\beta, \widehat{h})\| + \|G(\beta, h^\dagger)\| + 0\} = o_p(1) + \|G_n(\beta, \widehat{h})\| \times o_p(1) \\ &= o_p(1) + \|G_n(\beta, \widehat{h})\|_{W_n} \times (c + o_p(1)) \times o_p(1) \end{aligned}$$

where the second line follows by the same argument as in (48). Therefore, (49) gives:

$$\begin{aligned} \|G_n(\beta, \widehat{h})\|_{W_n} &\leq \{o_p(1) + \|G_n(\beta, \widehat{h})\|_{W_n} \times (c + o_p(1)) \times o_p(1) + \|G(\beta, h^\dagger) - G(\beta^0, h^\dagger)\|\} \times (d + o_p(1)) \\ &= o_p(1) + \|G_n(\beta, \widehat{h})\|_{W_n} \times o_p(1) + \|G(\beta, h^\dagger) - G(\beta^0, h^\dagger)\| \times (d + o_p(1)) \end{aligned}$$

and hence $\|G_n(\beta, \widehat{h})\|_{W_n} \times (1 - o_p(1)) \leq o_p(1) + \|G(\beta, h^\dagger) - G(\beta^0, h^\dagger)\| \times (d + o_p(1))$ where all the $o_p(1)$ terms are uniform with respect to $\beta \in \mathcal{B}$. This implies that:

$$\inf_{\beta \in \mathcal{B}} \|G_n(\beta, \widehat{h})\|_{W_n} \leq \sup_{\beta \in \mathcal{B}} o_p(1) + \inf_{\beta \in \mathcal{B}} \|G(\beta, h^\dagger) - G(\beta^0, h^\dagger)\|_W \times (d + \sup_{\beta \in \mathcal{B}} o_p(1)) = o_p(1)$$

since $\inf_{\beta \in \mathcal{B}} \|G(\beta, h^\dagger) - G(\beta^0, h^\dagger)\|_W = 0$. So, by (46) and (48) it follows that $\|G(\widehat{\beta}, h^\dagger)\| = o_p(1)$. ∎

Proof of Proposition 14: First, we show $\sqrt{n}$-consistency of $\widehat{\beta}$, and then its asymptotic normality.

Since $\beta^0 \in \text{interior}(\mathcal{B})$, $h^\dagger(\beta) \in \text{interior}(\mathcal{H})$, $\widehat{\beta} - \beta = o_p(1)$ and $\|\widehat{h}(\beta) - h^\dagger(\beta)\|_{\mathcal{H}} = o_p(1)$, we can choose a positive sequence $\delta_n = o_p(1)$ such that $P((\widehat{\beta}, \widehat{h}) \in \mathcal{B}_{\delta_n} \times \mathcal{H}_{\delta_n}) \to 1$ as $n \to \infty$. For the δ in the statement of the proposition, $P(\mathcal{B}_{\delta_n} \times \mathcal{H}_{\delta_n} \subset \mathcal{B}_\delta \times \mathcal{H}_\delta) \to 1$ as $n \to \infty$. While to avoid repetition we do not make it explicit, it is important to keep in mind that as in the proof of Proposition 13, here also we work conditional on the event $\{(\widehat{\beta}, \widehat{h}) \in \mathcal{B}_{\delta_n} \times \mathcal{H}_{\delta_n}\}$ which occurs with probability approaching one, i.e., we implicitly use arguments similar to (46) throughout the proof.

(C2) implies that there exists a constant $a > 0$ such that $P(a\|\widehat{\beta} - \beta^0\| \leq \|G(\widehat{\beta}, h^\dagger)\|) \to 1$ as $n \to \infty$. Therefore, $\sqrt{n}$-consistency of $\widehat{\beta}$ follows if we can establish that $\|G(\widehat{\beta}, h^\dagger)\| = O_p(n^{-1/2})$.

To this end, note that:

$$\begin{aligned} \|G(\widehat{\beta}, h^\dagger)\| &\leq \|G(\widehat{\beta}, h^\dagger) - G(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, \widehat{h}) - G_n(\widehat{\beta}, \widehat{h}) + G_n(\beta^0, h^\dagger)\| + \|G_n(\widehat{\beta}, \widehat{h})\| + \|G_n(\beta^0, h^\dagger)\| \\ &= 0 + \|G(\widehat{\beta}, \widehat{h}) - G_n(\widehat{\beta}, \widehat{h}) + G_n(\beta^0, h^\dagger)\| + \|G_n(\widehat{\beta}, \widehat{h})\| + O_p(n^{-1/2}) \end{aligned} \tag{50}$$

where the first 0 follows from (F2) and the last $O_p(n^{-1/2})$ from (C4). Now, by (C3) for the first inequality below,

$$\begin{aligned} \|G(\widehat{\beta}, \widehat{h}) - G_n(\widehat{\beta}, \widehat{h}) + G_n(\beta^0, h^\dagger)\| &\leq o_p(1) \times \{n^{-1/2} + \|G_n(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, \widehat{h})\|\} \\ &\leq o_p(1) \times \{n^{-1/2} + \|G_n(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, \widehat{h}) - G(\widehat{\beta}, h^\dagger)\| + \|G(\widehat{\beta}, h^\dagger)\|\} \\ &= o_p(1) \times \{n^{-1/2} + \|G_n(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, h^\dagger)\|\} \end{aligned}$$

where the last line follows by (F3). Therefore, this along with (50) implies that:

$$\|G(\widehat{\beta}, h^\dagger)\| \leq o_p(1) \times \{n^{-1/2} + \|G_n(\widehat{\beta}, \widehat{h})\| + \|G(\widehat{\beta}, h^\dagger)\|\} + \|G_n(\widehat{\beta}, \widehat{h})\| + O_p(n^{-1/2})$$

which, further implies that (second inequality below follows using same arguments as in (48) with $c = 1 + \|W^{-1} - I_{d_m}\|$)

$$\begin{aligned} \|G(\widehat{\beta}, h^\dagger)\| \times (1 - o_p(1)) &\leq O_p(n^{-1/2}) + \|G_n(\widehat{\beta}, \widehat{h})\| \times (1 + o_p(1)) \\ &\leq O_p(n^{-1/2}) + \|G_n(\widehat{\beta}, \widehat{h})\|_{W_n} \times (c + o_p(1)) \\ &\leq O_p(n^{-1/2}) + \inf_{\beta \in \mathcal{B}_\delta} \|G_n(\beta, \widehat{h})\|_{W_n} \times (c + o_p(1)) \end{aligned} \tag{51}$$

where the last line follows by (C1). Now, for $d = 1 + \|W - I_{d_m}\|$, recall from the first line of (49) that $\|G_n(\beta, \widehat{h})\|_{W_n} \leq \|G_n(\beta, \widehat{h})\| \times (d + o_p(1))$. On the other hand,

$$\begin{aligned} \|G_n(\beta, \widehat{h})\| &\leq \|G_n(\beta, \widehat{h}) - G(\beta, \widehat{h}) - G_n(\beta^0, h^\dagger)\| + \|G(\beta, \widehat{h}) - G(\beta, h^\dagger)\| + \|G(\beta, h^\dagger)\| + \|G_n(\beta^0, h^\dagger)\| \\ &\leq o_p(1) \times \{n^{-1/2} + \|G_n(\beta, \widehat{h})\| + \|G(\beta, \widehat{h})\|\} + 0 + \|G(\beta, h^\dagger)\| + O_p(n^{-1/2}) \end{aligned}$$

where the first term in the last line follows from (C3), the third term, i.e., the 0, from (F3), and the last one from (C4). Therefore,

$$
\begin{aligned}
\|G_n(\beta,\widehat{h})\| \times (1-o_p(1)) &\leq \|G(\beta,\widehat{h})\| \times o_p(1) + \|G(\beta,h^\dagger)\| + O_p(n^{-1/2}) \\
&\leq \|G(\beta,\widehat{h}) - G(\beta,h^\dagger)\| \times o_p(1) + \|G(\beta,h^\dagger)\| \times (1+o_p(1)) + O_p(n^{-1/2}) \\
&= \|G(\beta,h^\dagger)\| \times (1+o_p(1)) + O_p(n^{-1/2}) \text{ [by (F3)]} \\
&\leq \|G(\beta,h^\dagger) - G(\beta^0,h^\dagger)\| \times (1+o_p(1)) + \|G(\beta^0,h^\dagger)\| \times (1+o_p(1)) + O_p(n^{-1/2}) \\
&= \|G(\beta,h^\dagger) - G(\beta^0,h^\dagger)\| \times (1+o_p(1)) + O_p(n^{-1/2})
\end{aligned}
$$

since $G(\beta^0,h^\dagger)=0$. Therefore, $\|G_n(\beta,\widehat{h})\|_{W_n} \leq \|G(\beta,h^\dagger)-G(\beta^0,h^\dagger)\| \times (d+o_p(1)) + O_p(n^{-1/2})$ where all the o_p and O_p terms are uniform with respect to $\beta \in \mathcal{B}_\delta$. Hence, as in the proof of Proposition 13, noting that $\inf_{\beta\in\mathcal{B}} \|G(\beta,h^\dagger) - G(\beta^0,h^\dagger)\| = 0$, it follows that $\inf_{\beta\in\mathcal{B}_\delta} \|G_n(\beta,\widehat{h})\|_{W_n} = O_p(n^{-1/2})$ and, therefore, (51) gives $\|G(\widehat{\beta},h^\dagger)\| = O_p(n^{-1/2})$ and, subsequently, $\widehat{\beta}-\beta^0 = O_p(n^{-1/2})$.

To establish asymptotic normality, define the linearization $L_n(\beta) = G_n(\beta^0,h^\dagger) + M_\lambda(\beta-\beta^0)$. Note that the differences from the linearization in Chen et al. (2003) arise due to (F2) and (F3). This gives:

$$
\begin{aligned}
& \|G_n(\widehat{\beta},\widehat{h}) - L_n(\widehat{\beta})\| \\
= \ & \|G_n(\widehat{\beta},\widehat{h}) - G_n(\beta^0,h^\dagger) - M_\lambda(\widehat{\beta}-\beta^0)\| \\
= \ & \|G_n(\widehat{\beta},\widehat{h}) - G_n(\beta^0,h^\dagger) - G(\widehat{\beta},\widehat{h}) + G(\widehat{\beta},\widehat{h}) + G(\widehat{\beta},h^\dagger) - G(\widehat{\beta},h^\dagger) - M_\lambda(\widehat{\beta}-\beta^0)\| \\
\leq \ & \|G_n(\widehat{\beta},\widehat{h}) - G_n(\beta^0,h^\dagger) - G(\widehat{\beta},\widehat{h})\| + \|G(\widehat{\beta},\widehat{h}) - G(\widehat{\beta},h^\dagger)\| + \|G(\widehat{\beta},h^\dagger) - M_\lambda(\widehat{\beta}-\beta^0)\| \\
\leq \ & \|G_n(\widehat{\beta},\widehat{h}) - G_n(\beta^0,h^\dagger) - G(\widehat{\beta},\widehat{h})\| + \|G(\widehat{\beta},h^\dagger) - M_\lambda(\widehat{\beta}-\beta^0)\| \text{ [by (F3)]} \\
\leq \ & o_p(1) \times \{1 + \|G_n(\widehat{\beta},\widehat{h})\| + \|G(\widehat{\beta},\widehat{h})\|\} + \|G(\widehat{\beta},h^\dagger) - G(\beta^0,h^\dagger) - M_\lambda(\widehat{\beta}-\beta^0)\|
\end{aligned}
$$

where the term inside braces follows from (C3) and the inclusion of $G(\beta^0,h^\dagger)$ in the last term is innocuous since $G(\beta^0,h^\dagger)=0$. Now, by the definition of M_λ, assumptions (C2), (A3) and (F2), it follows that $\|G(\widehat{\beta},h^\dagger) - G(\beta^0,h^\dagger) - M_\lambda(\widehat{\beta}-\beta^0)\| = o_p(\|\widehat{\beta}-\beta^0\|)$, which is $o_p(n^{-1/2})$ since $\widehat{\beta}-\beta^0 = O_p(n^{-1/2})$. On the other hand, the same steps from the top line of (51) until (almost) the end of the first part of the proof give $\|G_n(\widehat{\beta},\widehat{h})\| \leq \inf_{\beta\in\mathcal{B}_\delta} \|G_n(\beta,\widehat{h})\| + o_p(n^{-1/2}) = O_p(n^{-1/2})$. Finally, since $\|G(\widehat{\beta},\widehat{h})\| \leq \|G(\widehat{\beta},\widehat{h}) - G(\widehat{\beta},h^\dagger)\| + \|G(\widehat{\beta},h^\dagger)\| = O_p(n^{-1/2})$ because the first term is 0 by (F3) and the second term is $O_p(n^{-1/2})$ from the first part of the proof, we obtain that $\|G_n(\widehat{\beta},\widehat{h}) - L_n(\widehat{\beta})\| \leq o_p(n^{-1/2})$. Similarly, for $\bar{\beta} := \arg\min_\beta \|L_n(\beta)\|_W$, that, by construction, satisfies $\sqrt{n}(\bar{\beta}-\beta^0) =$

$-(M_\lambda' W M_\lambda)^{-1} M_\lambda' W \sqrt{n} G_n(\beta^0, h^\dagger)$, we can show that $\|G_n(\bar{\beta}, \widehat{h}) - L_n(\bar{\beta})\| \leq o_p(n^{-1/2})$. Now that the proximity of $G_n(\beta, \widehat{h})$ and $L_n(\beta)$ has been established at $\widehat{\beta}$ and $\bar{\beta}$ respectively, the rest of the proof is to show that $\sqrt{n}(\bar{\beta} - \widehat{\beta}) = o_p(1)$. As was the case in Chen et al. (2003), this does not involve anything particularly related to the key feature of our setup (it only works with the linearization), and hence follows exactly in the same way as in the proof of Theorem 3.3 and Lemma 3.5 in Pakes and Pollard (1989). ■

Proof of Corollary 15:

(1) This is standard and hence the proof is omitted.

(2) This follows by noting that $g(O; \beta, h^0(O; \beta)) = \varphi_\lambda(O; \beta)$ defined in (6). ■

Proof of Proposition 16: Define $L_n(\beta) := G_n(\beta^0, h^0) + M_\lambda(\beta - \beta^0)$ and note that $\sqrt{n} L_n(\widetilde{\beta}) = O_p(1)$ by assumptions (A3), (C4) and since $\sqrt{n}(\widetilde{\beta} - \beta^0) = O_p(1)$. Therefore, using (F1), (F3) and also the stochastic equicontinuity condition from the statement of the proposition, we obtain that:

$$
\begin{aligned}
& \sqrt{n}\|G_n(\widetilde{\beta}, \widehat{h}) - L_n(\widetilde{\beta})\| \\
= \; & \sqrt{n}\|\{G_n(\widetilde{\beta}, \widehat{h}) - G(\widetilde{\beta}, \widehat{h}) - G_n(\beta^0, h^0)\} + \{G(\widetilde{\beta}, \widehat{h}) - G(\widetilde{\beta}, h^0(\beta^0))\} + \{G(\widetilde{\beta}, h^0(\beta^0) - M_\lambda(\widetilde{\beta} - \beta^0)\| \\
\leq \; & \sup_{\beta \in \mathcal{B}_{\delta_n}, h \in \mathcal{H}_{\delta_n}} \sqrt{n}\|G_n(\beta, h) - G(\beta, h) - G_n(\beta^0, h^0)\| + \sqrt{n}\|G(\widetilde{\beta}, \widehat{h}) - G(\widetilde{\beta}, h^0(\beta^0))\| \\
& \qquad + \|\sqrt{n} G(\beta^0, h^0) + (M_\lambda + o_p(1) - M_\lambda)\sqrt{n}(\widetilde{\beta} - \beta^0)\| \\
= \; & o_p(1) + 0 + (0 + o_p(1)) = o_p(1).
\end{aligned}
$$

Now, the proof completes since under the conditions of the proposition, the definition in (45) gives:

$$
\begin{aligned}
\sqrt{n}\left(\widehat{\beta}_{\text{1step}} - \widetilde{\beta}\right) &= -\left(\Omega_\lambda^{-1} + o_p(1)\right)\left(M_\lambda' + o_p(1)\right)\left(V_\lambda^{-1} + o_p(1)\right)\left(\sqrt{n} L_n(\widetilde{\beta}) + o_p(1)\right) \\
&= -\Omega_\lambda^{-1} M_\lambda' V_\lambda^{-1}\left(\sqrt{n} G_n(\beta^0, h^0) + M_\lambda \sqrt{n}(\widetilde{\beta} - \beta^0)\right) + o_p(1) \\
&= \sqrt{n}\left(\widehat{\beta} - \beta^0\right) - \sqrt{n}\left(\widetilde{\beta} - \beta^0\right) + o_p(1) = \sqrt{n}\left(\widehat{\beta} - \widetilde{\beta}\right) + o_p(1). \; ■
\end{aligned}
$$

References

Abrevaya, J. and Donald, S. G. (2017). A GMM approach for dealing with missing data on regressors and instruments. Forthcoming in Review of Economics and Statistics.

Ai, C. and Chen, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170: 442–457.

Allcott, H. and Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, 104: 3003–3037.

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

Ashraf, N., Berry, J., and Shapiro, J. M. (2010). Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia. *American Economic Review*, 100: 2383–2413.

Ashraf, N., Field, E., and Lee, J. (2014). Household Bargaining and Excess Fertility: An Experimental Study in Zambia. *American Economic Review*, 104: 2210–2237.

Back, K. and Brown, D. (1993). Implied Probabilities in GMM estimators. *Econometrica*, 61: 971–976.

Barnwell, J. L. and Chaudhuri, S. (2018). Efficient estimation in sub and full populations with monotonically missing at random data. Technical report, McGill University.

Beaman, L., Karlan, D., Thusbaert, B., and Udry, C. (2015). Self-Selection into Credit Markets: Evidence from Agriculture in Mali. Mimeo.

Beegle, K., Weerdt, J. D., Friedman, J., and Gibson, J. (2012). Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics*, 98: 3 – 18.

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

Chamberlain, G. (1992). Comment: Sequential Moment Restrictions In Panel Data. *Journal of Business and Economic Statistics*, 10: 20–26.

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

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

Chen, X. (2007). Large Sample Sieve Estimation Of Semi-Nonparametric Models. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume VIB, chapter 76, pages 5550–5632. Elsevier Science Publisher.

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.

Chen, X., Linton, O., and van Keilegom, I. (2003). Estimation of Semiparametric Models when the Criteria Function is not Smooth. *Econometrica*, 71: 1591–1608.

Dardanoni, V., Modica, S., and Peracchi, F. (2011). Regression with imputed covariates: A generalized missing-indicator approach. *Journal of Econometrics*, 162: 362–368.

Devereux, P. J. and Tripathi, G. (2009). Optimally combining censored and uncensored datasets. *Journal of Econometrics*, 151: 17–32.

Deville, J. C. and Sarndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87: 376–382.

DiNardo, J., McCrary, J., and Sanbonmatsu, L. (2006). Constructive Proposals for Dealing with Attrition: An Empirical Example. NBER and University of Michigan.

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.

Graham, J. W., Hofer, S. M., and MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 31: 197–218.

Graham, J. W., Taylor, B. J., Olchowski, A. E., and Cumsille, P. E. (2006). Planned Missing Data Designs in Psychological Research. *Psychological Methods*, 11: 323–342.

Hellerstein, J. K. and Imbens, G. W. (1999). Imposing Moment Restriction from Auxiliary Data by Weighting. *The Review of Economics and Statistics*, 81: 1–14.

Holcroft, C., Rotnitzky, A., and Robins, J. M. (1997). Efficient estimation of regression parameters from multistage studies with validation of outcome and covariates. *Journal of Statistical Planning and Inference*, 65: 349–374.

Holt, C. A. and Laury, S. K. (2002). Risk Aversion and Incentive Effects. *The American Economic Review*, 92: 1644–1655.

Imbens, G. W. and Lancaster, T. (1994). Combining Micro and Macro Data in Microeconometric Models. *Review of Economic Studies*, 61: 655–689.

Lee, A. J., Scott, A. J., and Wild, C. J. (2012). Efficient estimation in multi-phase case-control studies. *Biometrika*, 97: 361–374.

MacArdle, J. J. and Woodcock, R. W. (1997). Expanding testretest designs to include developmental time-lag components. *Psychological Methods*, 2: 403–435.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, 99: 210–221.

Nevo, A. (2003). Using Weights to Adjust for Sample Selection When Auxiliary Information is Available. *Journal of Business and Economic Statistics*, 21: 43–52.

Newey, W. K. and McFadden, D. L. (1994). Large Sample Estimation and Hypothesis Testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume IV, chapter 36, pages 2212–2245. Elsevier Science Publisher.

Nijman, T., Verbeek, M., and van Soest, A. (1991). The efficiency of rotating-panel designs in an analysis-of-variance model. *Journal of Econometrics*, 49: 373–399.

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

Raghunathan, T. E. and Grizzle, J. E. (1995). A Split Questionnaire Survey Design. *Journal of the American Statistical Association*, pages 54 – 63.

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

Shoemaker, D. M. (1973). *Principles and Procedures of Multiple Matrix Sampling.* Cambridge, MA: Ballinger.

Thornton, R. L. (2008). The Demand for, and Impact of, Learning HIV Status. *American Economic Review*, 98: 1829–1863.

Tripathi, G. (2011). Moment-based inference with stratified data. *Econometric Theory*, 27: 47–73.

Wacholder, S., Carroll, R. J., Pee, D., and Gail, M. H. (1994). The Partial Questionnaire Design For Case-Control Studies. *Statistics in Medicine*, 13: 623 – 634.

Whittemore, A. S. (1997). Multistage Sampling Designs and Estimating Equations. *Journal of Royal Statistical Society, Series B*, 59: 589–602.

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.