Efficient estimation in sub and full populations with monotonically missing at random data*

Jean-Louis Barnwell† and Saraswata Chaudhuri‡

This version: March 25, 2018

Abstract

We consider estimation of a parameter defined by moment restrictions on a target population characterized by the missingness of monotonically missing at random data. Attrition or dropout to an absorbing state from a, say, R-period study/survey typically generates such data. In this case, a generic target is the underlying population of the sample units dropping out in contiguous periods $a, \ldots, b$ where $a \leq b \in \{1, \ldots, R\}$. The semiparametric efficiency bound and the efficient influence function are obtained for the parameter of interest from the generic target nesting the well-known special cases with $(a = 1, b = R)$ or $(R = 2, a = b = 1)$.

Our results, however, differ fundamentally from the existing literature in that the consideration of a generic target beyond those special cases provides new insights on the usability and contribution of the sample units toward efficient estimation. Efficient estimation turns out to be standard MINPIN estimation with asymptotic properties directly following from the well-known existing results. Further desirable properties follow since the concerned MINPIN estimating functions are also doubly robust to parametric misspecifications of their nonparametric nuisance components. A Monte Carlo study demonstrates all these nice properties of the efficient estimator and the t-test based on it. An empirical illustration using the Project STAR data demonstrates substantive improvement in precision over the standard but inefficient estimators.

Keywords: Attrition; Dropouts; Efficiency; GMM; Missing at random; Pattern-mixtures.

*We are very grateful to Daniel Farewell, Erica Moodie, Fabian Lange, Francesco Amodio, Marine Carrasco and Steven Lehrer for helpful comments and discussions. This version supersedes the previous version of the paper that can be found on the corresponding author's web page http://saraswata.research.mcgill.ca/research.html.

†Department of Economics, McGill University, Montreal, Canada. Email: jean-louis.barnwellmenard@mail.mcgill.ca.

‡Corresponding author. Department of Economics, McGill University and Cireq, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.

1 Introduction

Estimation based on monotonically missing data has received special attention in the missing data literature; see, e.g., the textbooks by Little and Rubin (2002), Tsiatis (2006), etc. Since the pioneering work of Robins et al. (1994, 1995), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Holcroft et al. (1997), etc., efficient estimation in such cases has generally been considered under the missing at random, i.e., a selection on observables, assumption.

In the econometrics literature, the study of monotone missingness under selection on observables is primarily focused on attrition in panel data. See Fitzgerald et al. (1996), Nicoletti (2006), Wooldridge (2002), Wooldridge (2010) (pages 840-845), etc., for influential studies. Hence, although our presentation and results apply more generally — e.g., they directly apply to the well-known pattern mixture models of Glynn et al. (1986), Little (1993, 1994), etc. thanks to an equivalence result of Molenberghs et al. (1998) (Theorem 1) — we will focus on attrition/drop-outs to fix ideas.[1]

Accordingly, as a running example, consider attrition to an absorbing state in an R-period study/survey where the decision to leave at the end of any period depends on the variables observed until then, which is what drives the selection on observables. The variables that should have been observed in periods $a+1, \ldots, R$ are missing for the sample units who leave at the end of a period, say, $a \in \{1, \ldots, R\}$. Due to the selection on observables, the underlying population, call it population a, of these sample units is generally different from that of the full-population, which is the underlying population of the full set of sample units who started in period 1 before any attrition occurred.

Our paper focuses on these different populations for the sample units. Without loss of generality, we maintain that all sample units have to leave by the end of period R. Thus, the populations $a, b \in \{1, \ldots, R\}$ and their unions are sub-populations of the full-population that is equivalently defined as the underlying population of the units who leave at the end of any period $1, 2, \ldots, R$.

In this paper, we consider efficient estimation of parameters defined by the joint distribution of the variables in unions of contiguous sub-populations $[a, b] := a, \ldots, b$ where $a \leq b \in \{1, \ldots, R\}$. This includes the full-population ($a = 1, b = R$) and the individual sub-populations ($a = b$) as special cases. Our main contribution is to obtain the efficient influence functions and the efficiency bounds for all such cases under a unified framework and make their novel implications explicit.

Our results enable us to design an efficient estimator for the parameters of interest such that: (i) it only involves routine MINPIN computations, and (ii) its asymptotic properties follow directly

[1] As in Diggle et al. (2007), we do not distinguish between attrition (subsequent outcomes unobserved) and dropouts (subsequent outcomes, but not the intended ones, observed), and treat both as attrition (c.f. footnote 1 in Heckman et al. (1998)). The distinction is important in many cases, but we abstract from it since it is already well understood.

from well-known results; e.g., Andrews (1994), Newey (1994), Chen et al. (2003). Further desirable properties follow since the underlying estimating function for this estimator has an augmented-inverse-probability-weighted (AIPW) structure that is doubly robust to parametric misspecification of its nonparametric nuisance components as in Robins et al. (1994), Scharfstein et al. (1999), Bang and Robins (2005), Tsiatis (2006), Tan (2007), Cattaneo (2010), Rothe and Firpo (2016), etc.

All these properties hold quite convincingly even in the small-sample setup of our Monte Carlo study where this efficient estimator easily outperforms other types of estimators in every aspect.

The same conclusion on the superiority of this efficient estimator also holds in our illustration with a real life example, the Project STAR data, where we consider estimation of parameters such as the expected counterfactual math and reading scores in grade 3 for students who left Project STAR from small and non-small classes respectively in grades 1, 2, etc., had they not left.

The practical relevance of these results in our paper is derived from the fact that attrition necessarily entails loss in data that makes it imperative to use any available information to increase the precision of the estimates of interest without introducing (much) bias. The asymptotically unbiased efficient estimator in our paper is able to do this by reaching the benchmark efficiency.

As demonstrated in our empirical illustration, this increased precision often helps to shrink the confidence intervals for various empirically interesting parameters by ruling out economically implausible (large) values that could not otherwise be ruled out at the conventional levels by other estimators such as the less precise IPW-type estimators considered in Fitzgerald et al. (1996), Wooldridge (2002), Wooldridge (2010), etc. Thus, our contribution could be seen as a step forward in the context of this related literature in econometrics on attrition with selection on observables.[2]

When considered in the context of the broader literature on missing (at random) data beyond econometrics, our contribution is precisely to focus and obtain the efficiency results on a variety of empirically interesting sub-populations. These results are new to our knowledge, and provide important insights on the usability and contribution of the sample units toward efficient estimation.

It must be noted that Hahn (1998) and Chen et al. (2008) also obtain efficiency results for sub-populations in the case $a = b = 1$, $R = 2$ (roughly speaking, focusing on the average treatment effect on the treated). Chaudhuri (2017) allows for $R > 2$ and obtains efficiency results on sub-populations either by assuming that the probabilities related to the selection on observables are

[2]There is a vast (old and new) literature on attrition with selection on unobservables in econometrics. Identification in such contexts is typically achieved by the existence of instruments, refreshment samples, etc.; and without these additional information it is typically not possible to distinguish between selection on observables and unobservables from the data itself. Our paper does not contribute to this literature. See Wooldridge (2010) for a textbook presentation of the relative and practical merits of selection on observables versus unobservables in similar contexts.

known, or by enforcing a dimension-reduction assumption on the unknown probabilities. However, as we will demonstrate, the new insights are gained precisely when we consider $R > 2$ and $a > 1$ and do not impose either the known probability or the dimension-reduction assumption, i.e., when we consider the most realistic setup for multi-period attrition. Our efficiency results, thus obtained, turn out to be substantively different from the aforementioned existing results, even qualitatively.

Our paper proceeds as follows. Section 2 establishes and extensively discusses our main result. Section 3 sketches the relevant MINPIN estimation and heuristically discusses the asymptotic properties of the estimator by referring to the well-known results. Section 4 is a small-scale simulation study of the finite-sample properties. Section 5 presents an empirical illustration using the Project STAR data. Section 6 concludes. Appendices A-D contain technical and supplemental materials.

2 The efficiency bound and the efficient influence function

2.1 Framework:

Let $Z := (Z_1', \ldots, Z_R')'$ where Z_r is a $d_r \times 1$ random vector and $\sum_{r=1}^{R} d_r$ is finite. Following Tsiatis (2006), let C be a random variable with support $\{1, \ldots, R\}$ and $T_C(Z)$ a transformation defined as $T_r(Z) := (Z_1', \ldots, Z_r')'$ with dimension $(\sum_{s=1}^{r} d_s) \times 1$ for $r = 1, \ldots, R$. In the context of the attrition example, Z_r are the variables specific to period r, while $T_r(Z)$ are all the variables observed for a unit that leaves at the end of period r, i.e., with $C = r$, for $r = 1, \ldots, R$. Formally, let $O := (C, T_C'(Z))'$ denote what is observed for a unit in the sample.

We maintain a general selection on observables, i.e., a missing at random (MAR), assumption:

$$P(C = r|Z) \equiv P(C = r|T_R(Z)) = P(C = r|T_r(Z)) \text{ for } r = 1, \ldots, R. \tag{1}$$

This is the MAR assumption as in, e.g., Robins and Rotnitzky (1995), Tsiatis (2006), etc. in the sense of Rubin (1976). It is important to recognize that (1) implies that for any $r = 2, \ldots, R$:

$$P(C \geq r|Z) = 1 - \sum_{j=1}^{r-1} P(C = j|T_j(Z)) = 1 - \sum_{j=1}^{r-1} P(C = j|T_{r-1}(Z)) = P(C \geq r|T_{r-1}(Z)) \tag{2}$$

only depends on $T_{r-1}(Z)$. This *does not* mean $P(C = r|T_r(Z)) = P(C = r|T_{r-1}(Z))$ unless $r = R$.

Under the selection on observables condition in (1), we consider sub-populations $[a, b]$, equivalently, $(a \leq C \leq b)$, for $a \leq b$ and $a, b \in \{1, \ldots, R\}$. If $a = b = r$ then, in the attrition example, this is the underlying sub-population from which the units who left at the end of period r can be

viewed as randomly drawn. If $a < b$, then this is the sub-population for the units who left in the periods $a, a+1, \ldots, b$. Thus, if $a = 1$ and $b = R$, then this is the full-population.

The distributions of Z, denote them by $F_{Z|(a\leq C\leq b)}(z)$, in these sub-populations are typically different. We will define the parameter of interest as a finite dimensional feature of $F_{Z|(a\leq C\leq b)}(z)$. Accordingly, consider a function $m(Z;\beta) : \text{Support}(Z) \times \mathcal{B} \mapsto \mathbb{R}^{d_m}$, $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$ where $d_\beta \leq d_m$, and, then, for a given $a, b \in \{1, \ldots, R\}$ $(a \leq b)$, define the parameter value of interest β^0 as:

$$E[m(Z;\beta) \mid a \leq C \leq b] = 0 \text{ for } \beta \in \mathcal{B} \text{ if and only if } \beta = \beta^0. \tag{3}$$

We also maintain the following standard assumptions as in Tsiatis (2006) and Chen et al. (2008).

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

(A3) The Jacobian $M := \frac{\partial}{\partial \beta'} E\left[m(Z;\beta^0)|a \leq C \leq b\right]$ is a $d_m \times d_\beta$ finite matrix of full column rank.

Remark: (A1) rules out dependence and heterogeneity across sample units once they are viewed as random draws from O. (A2) ensures that the efficiency bound is finite. (A3) allows for $m(Z;\beta)$ to be non-differentiable in β, but does require that its expectation be differentiable.

2.2 The main result:

The key quantity to describe the efficient influence function and the efficiency bound for β^0 is:

$$\begin{aligned}
\varphi(O;\beta) \;:=\; & \sum_{r=b+1}^{R} \frac{I(C \geq r)}{P(C \geq r \mid T_r)} \frac{P(a \leq C \leq b \mid T_b)}{P(a \leq C \leq b)} \left(E[m(T_R;\beta) \mid T_r] - E[m(T_R;\beta) \mid T_{r-1}]\right) \\
& + \sum_{r=a+1}^{b} \frac{I(C \geq r)}{P(C \geq r \mid T_r)} \frac{P(a \leq C \leq r-1 \mid T_{r-1})}{P(a \leq C \leq b)} \left(E[m(T_R;\beta) \mid T_r] - E[m(T_R;\beta) \mid T_{r-1}]\right) \\
& + \sum_{r=a}^{b} \frac{I(C = r)}{P(a \leq C \leq b)} E[m(T_R;\beta) \mid T_r].
\end{aligned} \tag{4}$$

In (4) and onward we use these equivalent notation exchangeably: $I(C \geq R) \equiv I(C = R)$, $T_r \equiv T_r(Z)$ (where $T_R \equiv Z$), and $m(Z;\beta) \equiv E[m(T_R;\beta) \mid T_R]$. If $b = R$, then the indices (e.g., in the sums) running from $b+1$ to R are void. If $a = b$ then similar indices running from $a+1$ to b are void, and those running from a to b contain only one term and it corresponds to a (equivalently b).

Proposition 1 *Let* (1) *hold but with a restriction on dimension reduction that* $P(C = r|Z) \neq P(C = r|T_s)$ *for any* $s < r = 2, \ldots, R-1$ *unless* $P(C = r|Z) = P(C = r)$ *for all* $r = 1, \ldots, R-1$. *Let* (3) *and assumption A hold. Let the* $d_m \times d_m$ *matrix* $V := Var(\varphi(O; \beta^0))$ *be finite and positive definite where* β^0 *and* $\varphi(O; \beta)$ *are as defined in* (3) *and* (4) *respectively. Then the asymptotic variance lower bound for* $\sqrt{n}(\widehat{\beta} - \beta^0)$ *of any regular estimator* $\widehat{\beta}$ *for* β^0 *is given by* $\Omega := (M'V^{-1}M)^{-1}$. *An estimator* $\widehat{\beta}$ *whose asymptotic variance equals* Ω *has the asymptotically linear representation:*

$$\sqrt{n}(\widehat{\beta} - \beta^0) = -\Omega M'V^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(O_i; \beta^0) + o_p(1). \ \blacksquare$$

2.3 Discussion of the main result:

2.3.1 Relation of Proposition 1 with the well-known literature on missing data:

The following special cases of this general result are already well-known in the literature.

(i) Taking $R = 2$ and $a = b = 1$ gives the result of Case (1) in Theorem 1 of Chen et al. (2008):

$$\varphi(O; \beta) = \frac{I(C = 2)}{P(C = 2|T_1)}\frac{P(C = 1|T_1)}{P(C = 1)}(m(T_2; \beta) - E[m(T_2; \beta)|T_1]) + \frac{I(C = 1)}{P(C = 1)}E[m(T_2; \beta)|T_1].$$

(ii) $R = 2$, $a = 1$ and $b = 2$ give Case (2) in Theorem 1 of Chen et al. (2008) (also see Robins et al. (1994)):

$$\varphi(O; \beta) = \frac{I(C = 2)}{P(C = 2|T_1)}(m(T_2; \beta) - E[m(T_2; \beta)|T_1]) + E[m(T_2; \beta)|T_1].$$

(iii) On the other hand, taking a general R and setting $a = 1, b = R$ give [see Appendix A.1]:

$$\varphi(O; \beta) = \sum_{r=2}^{R}\frac{I(C \geq r)}{P(C \geq r|T_r)}\left(E[m(T_R; \beta)|T_r] - E[m(T_R; \beta)|T_{r-1}]\right) + E[m(T_R; \beta)|T_1],$$

the well-known result for the full-population under monotone missing at random data as in Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Holcroft et al. (1997). Also see Tsiatis (2006).

2.3.2 Proposition 1 applied to individual sub-populations, i.e., $a = b$:

On the other hand, if we focus on any generic individual sub-population, i.e., $a = b$, then for a general R, the corresponding $\varphi(O; \beta)$ becomes:

$$\sum_{r=a+1}^{R}\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(C = a|T_a)}{P(C = a)}\left(E[m(T_R; \beta)|T_r] - E[m(T_R; \beta)|T_{r-1}]\right) + \frac{I(C = a)}{P(C = a)}E[m(T_R; \beta)|T_a].$$

Thus, $\varphi(O;\beta) = I(C = R)m(T_R;\beta)/P(C = R)$ in the special case when $a = b = R$. Considering the generic individual sub-population $a = b$ in the context of attrition, it is now clear that only those sample units who did not already leave before period a contribute to this efficient estimation.

2.3.3 Proposition 1 contrasted with dimension-reduction assumptions:

As we just noted, the terms involving $E[m(Z;\beta)|T_r]$ for $r < a$ do not appear in the expression for $\varphi(O;\beta)$ in (4). (This fact is immaterial if $a = 1$ as in Section 2.3.1(i).) This is because Proposition 1 imposes a restriction on dimension-reduction in (1): We do not allow $P(C = r|Z) = P(C = r|T_s)$ for any $s < r = 2, \ldots, R-1$. Such dimension-reductions would generally be unrealistic in cases of attrition and hence are not incorporated in Proposition 1. Roughly speaking, we do not let the most recent past become irrelevant to the selection process conditional on the less recent past.

To contrast, consider an extreme dimension-reduction assumption: $P(C = r|Z) = P(C = r|T_1)$ that pins the selection to period 1, i.e., the baseline. Under this, Proposition 9 in Chaudhuri (2017) establishes that the corresponding $\varphi(O;\beta)$ for the target sub-population $(a \leq C \leq b)$ is:

$$\begin{aligned}\varphi^{\dagger}_{[a,b]}(O;\beta) &= \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_1)} \frac{P(a \leq C \leq b|T_1)}{P(a \leq C \leq b)} (E[m(Z;\beta)|T_r] - E[m(Z;\beta)|T_{r-1}]) \\ &\quad + \frac{I(a \leq C \leq b)}{P(a \leq C \leq b)} E[m(Z;\beta)|T_1]. \end{aligned} \tag{5}$$

Hence, under this dimension-reduction, all $E[m(Z;\beta)|T_r]$ $(r = 1, \ldots, R)$ are usable irrespective of a, b except if $a = 2, b = R$. It is also noteworthy that the terms $E[m(Z;\beta)|T_r]$ $(r = a, \ldots, b)$, i.e., those corresponding to the sub-populations of interest, contribute differently in this dimension-reduction case than in (the last two lines of) (4) unless $a = 1, b = R$, i.e., unless interest lies in the full-population. So, Proposition 9 in Chaudhuri (2017) is not a special case of our Proposition 1.

Of course, the above considerations are moot under the so-called missing completely at random (MCAR) assumption, i.e., when $P(C = r|Z) = P(C = r)$ for $r = 1, \ldots, R$. Under this case, all the sub-populations have the same distribution $F_Z(z)$ for Z, i.e., there is no "selection" as far as the moment restrictions in (3) are concerned. Naturally, then the target of interest is $a = 1, b = R$. In this case, our Proposition 1 and Chaudhuri (2017)'s Proposition 9 both give the same result:

$$\varphi(O;\beta) = \varphi^{\dagger}_{[1,R]}(O;\beta) = \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r)} (E[m(Z;\beta)|T_r] - E[m(Z;\beta)|T_{r-1}]) + E[m(Z;\beta)|T_1].$$

It is also possible to obtain efficiency results for less extreme dimension-reductions (our Propo-

sition 1 represents no reduction). However, as evident from the discussion above, this needs to be done on a case by case basis since the entire structure of the efficient influence function changes depending on the specific dimension-reduction assumption made and the target $[a, b]$ of interest.

2.3.4 Proposition 1 contrasted with planned missingness assumptions:

This observation on dimension-reduction is in direct contrast to the case of planned missingness in the data. To see it let MAR in (1) hold, i.e., $P(C = r|Z) = P(C = r|T_r)$ for $r = 1, \ldots, R$. In this context, by planned missingness we mean that $P(C = r|T_r)$ is known for $r = 1, \ldots, R$. Under this, Proposition 1 in Chaudhuri (2017) establishes that the corresponding $\varphi(O; \beta)$ for $(a \leq C \leq b)$ is:

$$\varphi^*_{[a,b]}(O;\beta) = \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_r)} \left(E\left[\frac{P(a \leq C \leq b|T_R)}{P(a \leq C \leq b)} m(T_R;\beta) \middle| T_r \right] - E\left[\frac{P(a \leq C \leq b|T_R)}{P(a \leq C \leq b)} m(T_R;\beta) \middle| T_{r-1} \right] \right) + E\left[\frac{P(a \leq C \leq b|T_R)}{P(a \leq C \leq b)} m(T_R;\beta) \middle| T_1 \right]. \quad (6)$$

Thus, under planned missingness, nesting of the conditions $\{P(C = r|Z) = P(C = r|T_s) : s \leq r\}$ (along with MCAR) is automatic in the sense that all $E[m(Z; \beta)|T_r]$ $(r = 1, \ldots, R)$ are always usable irrespective of whether dimension-reduction is allowed or not. Hence, while Chaudhuri (2017) also contrasts unplanned and planned missingness under $P(C = r|Z) = P(C = r|T_1)$, as evident from (4) and (6), the same contrast in our paper is more prominent in terms of usability of the sample units since we maintain $P(C = r|Z) = P(C = r|T_r)$ without allowing for dimension-reduction.

2.3.5 Proposition 1 as a weighted average:

As evident from (5) and (6), in Chaudhuri (2017), we have that:

$$\varphi^\dagger_{[a,b]}(O;\beta) = \sum_{j=a}^{b} \frac{P(C=j)}{P(a \leq C \leq b)} \varphi^\dagger_{[j,j]}(O;\beta) \quad \text{and} \quad \varphi^*_{[a,b]}(O;\beta) = \sum_{j=a}^{b} \frac{P(C=j)}{P(a \leq C \leq b)} \varphi^*_{[j,j]}(O;\beta),$$

which reflect the standard relationship among the moment restrictions in (3) that:

$$E[m(Z;\beta)|a \leq C \leq b] = \sum_{j=a}^{b} \frac{P(C=j)}{P(a \leq C \leq b)} E[m(Z;\beta)|C = j].$$

This same weighted average representation also holds for $\varphi(O; \beta)$ defined in (4) [see Appendix A.2 for details]. This is intuitively appealing, and it presents a way of combining the efficient estimators for the individual sub-populations to obtain the efficient estimator for their contiguous unions.

2.3.6 Proposition 1 and double-robustness:

$\varphi(O;\beta)$ defined in (4) contains two sets of unknown nuisance parameters (NP)/ functions:

NP1. The conditional probabilities $P(a \leq C \leq r|T_r)$ $(= \sum_{j=a}^{r} P(C=j|T_j)$ by (1)) for $r = a, \ldots, b-1$, and $P(C \geq r|T_r)$ $(= 1 - P(C \leq r-1|T_{r-1})$ by (2)) for $r = a+1, \ldots, R$.

NP2. The conditional expectations $E[m(T_R;\beta)|T_r]$ for $r = a, \ldots, R-1$.

Interestingly, as shown in Appendix A.3, provided that the expectation exists, the expectation of $\varphi(O;\beta^0)$ is still zero when either NP1 or NP2 is replaced in $\varphi(O;\beta^0)$ by any function of T_r of conformable dimesion. That is, $\varphi(O;\beta)$ is doubly robust in the sense of Scharfstein et al. (1999) to parametric misspecifications of its nonparametric components NP1 and NP2. This leads to desirable properties for the estimator of β^0 based on the estimating function $\varphi(O;\beta)$ when NP1 and NP2 are estimated parametrically or nonparametrically; see Robins et al. (1994), Robins and Ritov (1997), Bang and Robins (2005), Tan (2007), Cao et al. (2009), Rothe and Firpo (2016), etc.

3 Efficient estimator of β, and estimation of asymptotic variance

Informed by Proposition 1 and Section 2.3.6, we use $\varphi(O;\beta)$ as the moment vector for the efficient GMM estimation of β^0. This requires preliminary estimation of NP1 and NP2 parametrically or nonparametrically; hence, MINPIN.[3] It is attractive since its computation and the derivation of its asymptotic properties are standard in all respect. Hence, for brevity, we discuss these only heuristically by referring to well-known results from (a rather incomplete set of) relevant references.

For notational brevity, denote generically the parameters in NP1 and NP2 by $p(.)$ and $q(.;\beta)$ respectively, their true values by $p_0(.)$ and $q_0(.;\beta)$ respectively, and their estimators by $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ respectively. Note by inspecting (4) and Section 2.3.6 that the estimation of any given individual element of $p(.)$ and $q(.;\beta)$ is only required when its arguments, i.e., the conditioning variables T_r (as appropriate for that individual element), are observed. Hence, no issue of infeasibility arises.

Let $g(O, p(.), q(.;\beta);\beta)$ be $\varphi(O;\beta)$ from (4) with $p_0(.)$ and $q_0(.;\beta)$ in $\varphi(O;\beta)$ replaced by the generic functions $p(.)$ and $q(.;\beta)$. That is, $g(O, p(.), q(.;\beta);\beta)$ is a function of O, β, $p(.)$ and $q(.;\beta)$ such that at the true values $p_0(.)$ and $q_0(.;\beta)$ of $p(.)$ and $q(.;\beta)$ we have $g(O, p_0(.), q_0(.;\beta);\beta) = \varphi(O;\beta)$. Now, define the average moment vector as:

$$\bar{g}_n(\beta, p(.), q(.;\beta)) := \frac{1}{n}\sum_{i=1}^{n} g(O_i, p(.), q(.;\beta);\beta).$$

[3]The unknown $P(a \leq C \leq b)$ in $\varphi(O;\beta)$ can be replaced by an estimator $\sum_{i=1}^{n} I(a \leq C_i \leq b)/n$ without any cost.

Given $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$, and a $d_m \times d_m$ weighting matrix W_n, define the GMM estimator $\widehat{\beta}_n(W_n)$ as:

$$\widehat{\beta}_n(W_n) \approx \arg\min_{\beta\in\mathcal{B}} \ \bar{g}_n(\beta, \widehat{p}(.), \widehat{q}(.;\beta))' \ W_n \ \bar{g}_n(\beta, \widehat{p}(.), \widehat{q}(.;\beta)). \tag{7}$$

3.1 When $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ are parametric estimators:

This is the strategy most commonly employed in practice. Appendix C contains a step-by-step description of the computation involved in the estimation of $\widehat{\beta}_n(W_n)$ and its asymptotic variance.

Standard extension of Holcroft et al. (1997) gives sufficient conditions for consistency and asymptotic normality of $\widehat{\beta}_n(W_n)$. The double-robustness property from Section 2.3.6 ensures that $\widehat{\beta}_n(W_n) \xrightarrow{P} \beta^0$ if the parametric model for either $p(.)$ or $q(.;\beta)$ is correctly specified and, subsequently, the standard conditions for the convergence of their corresponding maximum likelihood and nonlinear least squares estimations hold. Consistency of $\widehat{\beta}_n(W_n)$ for β^0 generally does not hold when models for $p(.)$ or $q(.;\beta)$ are both parametrically misspecified, and the asymptotic properties can be poor (see Kang and Schafer (2007)) as is typical in cases of two-step parametric estimation.

$\widehat{\beta}_n(W_n)$ is asymptotically efficient with asymptotic variance Ω, as in Proposition 1, if both parametric models are correctly specified and if the weighting matrix is efficient, i.e., if $W_n \xrightarrow{P} V^{-1}$.

When both these parametric models are correctly specified and W_n is efficient, the asymptotic variance of $\widehat{\beta}_n(W_n)$ can be estimated in the standard way. In particular, since (1) implies that:

$$E\left[\frac{P(a \le C \le b|T_R)}{P(a \le C \le b)} \frac{I(C=R)}{P(C=R|T_R)} m(T_R;\beta)\right] = E[m(T_R;\beta)|a \le C \le b],$$

one could, by virtue of (A3), estimate M by taking the (possibly numerical) derivative of

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{P}(a \le C \le b|T_R(Z_i))}{\widehat{P}(a \le C \le b)} \frac{I(C_i=R)}{\widehat{P}(C=R|T_R(Z_i))} m(T_R(Z_i);\beta) \tag{8}$$

with respect to β at $\widehat{\beta}_n(W_n)$. On the other hand, V can be estimated by averaging the outer product of $g(O_i, \widehat{p}(.), \widehat{q}(.;\widehat{\beta}_n(W_n)); \widehat{\beta}_n(W_n))$ while ignoring that $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ are actually estimated.

Ω is not the correct asymptotic variance when any of these parametric models is misspecified (although consistency of $\widehat{\beta}_n(W_n)$ for β^0 only requires that at least one parametric model is correctly specified). Then, for the estimation of V one should take the estimation of $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ into account as in Theorem 6.1 of Newey and McFadden (1994), and instead use the modification for two-step estimators as described in Section 6.3 of Newey and McFadden (1994). On the other hand,

(8) is not consistent for M when $p(.)$ is misspecified. Generally, it is safer to take the more tedious route that estimates M by directly taking the (possibly numerical) derivative of $\bar{g}_n(\beta, \widehat{p}(.), \widehat{q}(.;\beta))$ with respect to β. Section 2.3.6 implies that under standard conditions this estimator is doubly robust, i.e., the estimator is consistent for M if at least one parametric model is correctly specified.

We provide a working description of the computation of $\widehat{\beta}_n(W_n)$ and its asymptotic variance in Appendix C. The computations in our empirical illustration (Section 5) follow this description.

3.2 When $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ are nonparametric estimators:

If nonparametric estimators $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ are used then consistency of $\widehat{\beta}_n(W_n)$ (for β^0) and its asymptotic normality follow under the generic conditions of Theorems 1 and 2 of Chen et al. (2003). Chen et al. (2003) also provide the sufficient conditions under which the estimation of $\widehat{p}(.)$ and $\widehat{q}(.;\beta)$ has no effect on the asymptotic variance of $\widehat{\beta}_n(W_n)$. Conditions similar to Theorems 6 and 7 of Cattaneo (2010) give consistency of the asymptotic variance estimator obtained based on averaging the outer product of $g(O_i, \widehat{p}(.), \widehat{q}(.;\widehat{\beta}_n(W_n)); \widehat{\beta}_n(W_n))$ and using a possibly numerical derivative-based estimate of M. But also see Ackerberg et al. (2012) for important practical matters related to the estimation of the asymptotic variance. As usual, $W_n \xrightarrow{P} V^{-1}$ gives efficiency of $\widehat{\beta}_n(W_n)$. See Rothe and Firpo (2016) for a discussion of how the double-robustness of the estimating equations, as discussed in Section 2.3.6 and Appendix A.3, leads to nice higher-order properties of $\widehat{\beta}_n(W_n)$.

4 Monte Carlo Experiment

We consider a setup reflecting the decision to stay or leave dynamically over periods in the context of programs (e.g., smoking cessation, weight loss), school, employment, experiments, surveys, market, etc. (The selection mechanism considered below may reverse in some cases.) Hence, it also provides a plausible motivation for the empirical illustration with the Project STAR data in Section 5.

4.1 Simulation design:

The design and the data generating process (DGP) considered are as follows. For $t = 1, \ldots, T$, let:

$$Y_t = \frac{1}{2}Y_{t-1} + \frac{1}{4}Y_{t-2} + \frac{1}{4}X_t + e_t, \quad \text{where} \quad X_t = X_{t-1} + v_t. \tag{9}$$

e_t and v_t are the model errors. Take X_0, Y_{-1}, Y_0 independently $N(1,1)$ as the initial state. (We will not define the parameters of interest conditional on the initial state, but this could be done.)

Define $R := T$. For $r = 1, \ldots, R$, let Y_r be the outcome from staying until the end of the r-th period in the program, and X_r the other variables for the r-th period.

Let the individual's expectation for the outcome in the r-th period be Y_r^*. Suppose that the individual decides to leave the program at the end of the r-th period, conditional on staying until then, provided that the actual outcome exceeds the expectation, i.e., $Y_r^* < Y_r$. More precisely, let:

$$I(C = r) = I(Y_r^* < Y_r) \prod_{j=1}^{r-1} I(Y_j^* \geq Y_j) \text{ for } r = 1, \ldots, R-1, \text{ while } I(C = R) = 1 - \sum_{r=1}^{R-1} I(C = r). \tag{10}$$

We take $R = 3$. The analyst observes C but not Y_r^*. In our notation, this means: $Z_1 = (Y_{-1}, Y_0, Y_1, X_{-1}, X_0, X_1)'$, $Z_2 = (Y_2, X_2)'$ and $Z_3 = (Y_3, X_3)'$. So, the observables are $T_1 = Z_1$, $T_2 = (Z_1', Z_2')'$ and $T_3 = (Z_1', Z_2', Z_3')'$ for those with $(C = 1)$, $(C = 2)$ and $(C = 3)$ respectively.

Let e_t and v_t be i.i.d. $N(0, 1)$ for all t, while $u_r := Y_r^* - Y_r$ is i.i.d. $N(0, (2.5)^2)$ for all r. (1) is imposed by maintaining that $e_t, v_t, u_r, X_0, Y_{-1}, Y_0$ are mutually independent for all t, r. This results in roughly 50% of the individuals with $C = 1$, 26% with $C = 2$, and 24% with $C = 3$.

To define β^0, we take the moment function in (3) as $m(Z; \beta) = Y_3 - \beta$, and consider the six different target sub-populations $[a, b]$ with, respectively, $(a = 1, b = 3)$, $(a = b = 1)$, $(a = b = 2)$, $(a = b = 3)$, $(a = 1, b = 2)$ and $(a = 2, b = 3)$ giving six different parameters of interest. The first parameter is the average period 3 outcome for all the units. The second parameter is the average period 3 outcome for those who left at the end of period 1 had they instead continued till the end of period 3. And, so on for the other parameters. The last parameter is the average period 3 outcome for those who left at the end of period 2 or 3 had those who left at the end of period 2 instead continued till the end of period 3. We compute the "true value" of these parameters numerically by generating data from the above DGP with sample size 10 million, estimating the mean of Y_3 for each sub-population, and then averaging each mean over 10,000 Monte Carlo trials. Accordingly, the six different "true values", i.e., β^0's are: 1, 1.1709, .9617, .6858, 1.0994 and .8291 respectively. As evident from Table 1, the error in the above approximation is of a rather small order to seriously affect our subsequent analysis that is conducted with far smaller (than 10 million) sample size.

4.2 Simulation results:

We study the efficient estimation of the β^0's under this setup for sample sizes $n = 100, 200$ and 500 (also $n = 1000$ in Table 2). All results are reported below based on 10,000 Monte Carlo trials.

The efficient GMM estimator in (7) is computed ignoring the weighting matrix W_n since $d_m =$

Target Population $[a,b]$ for β	Descriptive Statistics					
	Mean	Std $= 10^{-3}\times$	Median	IQR	Min	Max
$a=1, b=3$	1	.4860	1	.0007	.9982	1.0017
$a=b=1$	1.1709	.6841	1.1709	.0009	1.1682	1.1735
$a=b=2$	.9617	.9430	.9617	.0013	.9581	.9648
$a=b=3$	.6858	.9769	.6858	.0013	.6817	.6895
$a=1, b=2$	1.0994	.5536	1.0994	.0007	1.0975	1.1012
$a=2, b=3$	.8291	.6800	.8291	.0009	.8265	.8316

Table 1: The "true" parameter value β^0 is approximated (column 2) for different target populations (column 1) based on averaging over 10,000 Monte Carlo trials the target-sample means obtained by using the same DGP and with sample size 10 million. Columns 3-7 list the standard deviation (Std), interquartile range (IQR), minimum (Min) and maximum (Max) of the estimator.

$d_\beta(=1)$. The nuisance parameters, i.e., the conditional expectations (of Y_3) and the conditional probabilities are estimated by least squares and probit regressions respectively. For both, we specify the regression function as linear in the conditioning variables and do not include interactions.

The parameter β for the sub-population $[a,b]$ with $(a=b=3)$ is not interesting for our purpose since the complete-case GMM estimator $\sum_{i=1}^n I(C_i=3)Y_{3i}/\sum_{j=1}^n I(C_j=3)$ already estimates it consistently and, by virtue of Proposition 1, efficiently (see Section 2.3.2). Table 2 briefly summarizes the performance of the complete-case GMM estimator, and, as expected, this is poor and misleading for all the target sub-populations except for the one with $(a=b=3)$.

Target Population	$n=100$		$n=200$		$n=500$		$n=1000$	
	Std $=.3140$		Std $=.2207$		Std $=.1386$		Std $=.0994$	
$[a,b]$ for β	Bias	Size	Bias	Size	Bias	Size	Bias	Size
$a=1, b=3$	-.3148	16.8	-.3151	29.8	-.3155	62.5	-.3148	88.7
$a=b=1$	-.4857	33.5	-.4860	59.1	-.4864	93.8	-.4857	99.8
$a=b=2$	-.2765	14.1	-.2768	24.1	-.2772	51.9	-.2765	79.7
$a=b=3$	-.0006	5.3	-.0009	5.2	-.0013	5.2	-.0006	4.8
$a=1, b=2$	-.4142	25.7	-.4145	46.8	-.4149	84.9	-.4142	98.7
$a=2, b=3$	-.1439	7.7	-.1442	9.9	-.1446	18.1	-.1439	30.8

Table 2: Results for the complete-case estimator are reported based on 10,000 Monte Carlo trials. Bias stands for the mean bias. The only representative population for the complete-case estimator is $(a=b=3)$. Std stands for the Monte Carlo standard deviation, and since the estimator is identical for all sub-populations (and hence the bias for the other targets), there is only a single Std reported for each sample size n. Size stands for the empirical size of the asymptotic 5% two-sided t-test.

For the other five parameters, the results for the efficient GMM estimator from (7) are presented in Table 3. In each Monte Carlo trial, we obtain the estimate for each β and subtract it from its "true value" given in Table 1. Then we average the mean, median, minimum and maximum of this difference over all the 10,000 trials and report them as Bias, MedBias, MinBias and MaxBias

respectively. The average of the absolute value of this difference is reported as AbsBias, the Monte Carlo standard deviation as Std, the average of the standard deviation based on the asymptotic variance formula as AStd, and the empirical size of the t-test based on Std and the asymptotic variance formula as Size and ASize respectively. The efficient GMM estimator performs quite well in all aspects and for all the target sub-populations even for relatively smaller sample sizes.

Target	$n = 100$								
$[a, b]$ for β	Bias	MedBias	MinBias	MaxBias	AbsBias	Std	AStd	Size	ASize
$a = 1, b = 3$	-.0037	-.0102	-1.4862	2.5896	.2544	.3254	.2512	5.3	13.8
$a = b = 1$	-.0041	-.0149	-2.6509	3.3278	.3177	.4092	.3336	5.1	11.8
$a = b = 2$	-.0068	-.0132	-2.4474	2.2725	.3484	.4501	.3986	5.3	8.2
$a = 1, b = 2$	-.0056	-.0147	-1.9473	2.8004	.2896	.3727	.2893	5.1	13.4
$a = 2, b = 3$	-.0034	-.0050	-1.6282	1.4262	.2492	.3187	.2803	5.1	8.5
	$n = 200$								
	Bias	MedBias	MinBias	MaxBias	AbsBias	Std	AStd	Size	ASize
$a = 1, b = 3$	-.0005	-.0042	-.9553	1.0449	.1650	.2078	.1884	5.1	8.1
$a = b = 1$	-.0009	-.0077	-1.7499	1.6389	.2015	.2551	.2436	5.0	6.4
$a = b = 2$	.0006	-.0021	-1.7145	3.1583	.2211	.2832	.2816	4.9	4.9
$a = 1, b = 2$	-.0009	-.0065	-1.2391	1.3064	.1836	.2321	.2139	4.9	7.5
$a = 2, b = 3$	.0001	-.0013	-.9836	1.7208	.1659	.2101	.2045	5.0	5.7
	$n = 500$								
	Bias	MedBias	MinBias	MaxBias	AbsBias	Std	AStd	Size	ASize
$a = 1, b = 3$	-.0020	-.0039	-.6830	.4939	.1009	.1269	.1208	5.0	6.7
$a = b = 1$	-.0024	-.0045	-.5654	.6081	.1225	.1540	.1546	5.0	5.3
$a = b = 2$	-.0015	-.0026	-1.6263	.7290	.1313	.1668	.1753	5.1	4.1
$a = 1, b = 2$	-.0023	-.0040	-.8461	.5669	.1103	.1392	.1358	5.2	6.0
$a = 2, b = 3$	-.0014	-.0016	-.9141	.4623	.1024	.1288	.1299	4.8	4.8

Table 3: Results for the efficient GMM estimator are reported based on 10,000 Monte Carlo trials. Bias, MedBias, MinBias, MaxBias and AbsBias stand for the mean, median, minimum, maximum and average absolute value respectively of the difference between the efficient estimator and the corresponding β^0. Std and AStd stand for the standard deviation based on Monte Carlo and the asymptotic variance formula respectively. Size and ASize stand for the empirical size of the asymptotic 5% two-sided t-test using the Monte Carlo and asymptotic standard deviation respectively. Roughly 50% of the individuals correspond to $C = 1$, 26% to $C = 2$, and 24% to $C = 3$.

To put the performance of the efficient GMM estimator into context, we report in Table 4 the same (except AStd and Asize) for the Horvitz and Thompson (1952)-type IPW estimator:

$$\sum_{i=1}^{n} \frac{I(C_i = 3)}{\widehat{P}(C = 3|T_{3i})} \frac{\widehat{P}(a \leq C_i \leq b|T_{bi})}{\sum_{j=1}^{n} I(a \leq C_j \leq b)} Y_{3i}. \tag{11}$$

The related literature in econometrics on attrition under selection of observables typically takes the full-population as the target and employs estimators of this genre (see, e.g., Fitzgerald et al. (1996), Wooldridge (2010), etc.). Our specifications for the conditional probabilities nests the truth.

Hence, this IPW estimator is also consistent and asymptotically unbiased by virtue of (1). This is reflected by the shrinking Bias and also the shrinking variability (AbsBias, Std) of the estimator.

Target	$n = 100$						
$[a, b]$ for β	Bias	MedBias	MinBias	MaxBias	AbsBias	Std	Size
$a = 1, b = 3$	-.0449	-.0823	-1.5733	7.9008	.3080	.4278	3.7
$a = b = 1$	-.0593	-.1386	-2.2333	16.3040	.4159	.6143	3.2
$a = b = 2$	-.0579	-.1168	-4.5318	6.3902	.3987	.5611	4.1
$a = 1, b = 2$	-.0589	-.1184	-2.0413	10.6195	.3653	.5187	3.5
$a = 2, b = 3$	-.0305	-.0469	-1.9653	3.8259	.2743	.3678	4.2
	$n = 200$						
	Bias	MedBias	MinBias	MaxBias	AbsBias	Std	Size
$a = 1, b = 3$	-.0143	-.0338	-1.2238	3.4379	.2006	.2701	4.0
$a = b = 1$	-.0209	-.0606	-1.1614	4.6535	.2697	.3698	4.1
$a = b = 2$	-.0141	-.0421	-2.0402	10.4536	.2611	.3839	3.4
$a = 1, b = 2$	-.0189	-.0505	-1.4550	4.4394	.2345	.3223	3.8
$a = 2, b = 3$	-.0075	-.0174	-1.2727	6.0802	.1840	.2527	3.6
	$n = 500$						
	Bias	MedBias	MinBias	MaxBias	AbsBias	Std	Size
$a = 1, b = 3$	-.0044	-.0088	-.5559	1.0035	.1192	.1527	4.6
$a = b = 1$	-.0060	-.0213	-.6293	1.3960	.1610	.2084	4.5
$a = b = 2$	-.0041	-.0180	-.9816	2.4238	.1542	.2016	4.7
$a = 1, b = 2$	-.0056	-.0152	-.6295	1.2874	.1377	.1779	4.7
$a = 2, b = 3$	-.0027	-.0061	-.5498	1.3350	.1116	.1417	4.8

Table 4: Results for the IPW estimator are reported based on 10,000 Monte Carlo trials. Bias, MedBias, MinBias, MaxBias, IQRBias and AbsBias stand for the mean, median, minimum, maximum and average absolute value respectively of the difference between the IPW estimator and the corresponding β^0. Std stands for the standard deviation based on Monte Carlo. Size stands for the empirical size of the asymptotic 5% two-sided t-test using the Monte Carlo standard deviation. Roughly 50% of the individuals correspond to $C = 1$, 26% to $C = 2$, and 24% to $C = 3$.

The efficiency gain due to the use of the efficient GMM estimator instead of the IPW estimator is large. Defining the gain as the difference between the Std of the IPW estimator and the efficient GMM estimator divided by the Std of the latter, it is evident from Tables 3 and 4 that the gain varies from 10% to 50% (averaging around 30%). Indeed, a comparison of the other dispersion-related characteristics clearly demonstrates that the efficient GMM estimator is much less dispersed than the IPW estimator in all aspects and for all sample sizes considered in our experiment.

5 Empirical Illustration with the Project STAR data

Tennessee's Student/Teacher Achievement Ratio experiment, also known as Project STAR, has been extensively studied to analyze the effect of small class size (student/teacher ratio) on various future outcomes for the students; see, among many others, Hanushek (1999), Krueger (1999), and

Ding and Lehrer (2010). In Project STAR, students enrolling in the kindergarten (grade K) of 80 participating schools in the 1985-1986 school year were randomly assigned to three types of classes: small classes (13-17 students per teacher), regular classes (22-25 students per teacher), and regular classes with a full-time teacher's aide (22-25 students per teacher).

In this section, we use the data from Project STAR, downloaded from the textbook-webpage https://economics.mit.edu/faculty/angrist/data1/mhe/krueger, for the purpose of an empirical illustration of the efficiency gains that can be obtained by the use of the efficient estimator in (7).

The data set contains characteristics of the schools, the teachers, demographic and socioeconomic characteristics of the students and their standardized math and reading scores from grade K up to grade 3 or up to a lower grade until which they stayed with a participating school.

Unfortunately, a sizable number of students did not stay in the participating schools until the end of grade 3. See Table 5. See the descriptive statistics by class types in Tables 8-9 (Appendix D) for an idea on the selection induced by attrition on the variables of our concern. Note that, we will hereafter refer to the class types — regular and regular with aide — jointly as non-small class.

The consequences of this attrition on the identification of various effects of interest related to the Project STAR have been studied in Krueger (1999), Ding and Lehrer (2010), etc. The latter, in particular, works under the selection on observables framework of Fitzgerald et al. (1996) (we do the same) and employs the IPW-principle of Horvitz and Thompson (1952) as in (11). However, since (11) does not use all available information in the sample, IPW is not efficient in general and, indeed, it is often far from being efficient as evident from our Monte Carlo experiment (Section 4).

This raises the question whether employing the efficient GMM estimator in (7) would result in substantive efficiency gain over the IPW estimator in the context of the Project STAR data.

We find the answer to be positive. We demonstrate this gain for a set of empirically relevant parameters. Since, ultimately, the quantities of interest in the related studies are sub-population means or their differences, for this empirical illustration we focus on the following parameters.

These are the mean math and reading scores in grade 3 for students who enrolled in grade K of a participating Project STAR school in 1985-1986 but left the participating school at the end of:

(a) grade K	(b) grade 1
(c) grade 2	(d) never left (i.e., continued until the end of grade 3)
(e) grade K or grade 1	(f) grade 1 or two
(g) grade 2 or never left	(h) grade K or grade 1 or grade 2
(i) grade 1 or grade 2 or never left	(j) grade K or grade 1 or grade 2 or never left,

if instead they had stayed in their initially assigned small or non-small class till the end of grade 3.

5.1 Estimation setup:

Therefore, we have 40 ($= 2 \times 2 \times 10$) parameters of interest: mean scores on 2 subjects (math and reading), for 2 class types (small and non-small), and for the 10 attrition categories (a)-(j).

(e) is the super-population of (a) and (b); (f) of (b) and (c); (g) of (c) and (d); (h) of (a), (b) and (c); (i) of (b), (c) and (d); and (j) of (a)-(d) (and, for that matter, of (a)-(i)). Indeed, given the initial randomization, (j) is not only the underlying population of those randomized to a small (respectively, non-small) class, but also that of the broader population of all students enrolled in grade K (small or non-small) of a Project STAR participating school in the 1985-1986 school year.

Given these parameters of interest, naturally, we drop the students who enrolled in a participating school only after grade K. See Ding and Lehrer (2010) for other reasons for not considering these students. We also drop those few students who switched between class types in any given grade. This is potentially problematic, and we would have preferred to avoid it. However, since the proportion (as well as the actual number) of switchers is very small — ranging from 1.8% to 5.7% of the eligible students (see Table 5) — it is difficult to systematically and explicitly take them into account in our analysis whose validity ultimately depends on the asymptotic theory.

	Randomized to small class			Randomized to non-small class		
After grade	Stayed in small	Left STAR school	Switched to non-small	Stayed in non-small	Left STAR school	Switched to small
K	1017	409	79	2241	1047	201
1	811	188	18	1682	484	75
2	683	105	23	1395	200	87

Table 5: Number of students by their switching or leaving dynamics at the end of each grade after they were initially randomized to small or non-small classes in grade K.

Finally, note that, only the parameters for (d) (i.e., never left) do not correspond to a counterfactual (sub-)population. Hence, they are trivially estimable by the complete-case estimator as the averages of the observed relevant grade 3 scores for these students. Indeed, for (d), the efficient GMM and IPW estimators in (7) and (11) are identical to the complete-case estimator (see Section 2.3.2). Hence, from Section 2.3.5, we know that (a), (b), and (c) (i.e., left at the end of grade K, 1, and 2 respectively) are the key quantities to be estimated (along with the variance-covariance matrix of the estimators) since the estimators for (e)-(j) are weighted averages of those for (a)-(d).

We estimate all these 40 parameters following the description presented in Appendix C. All estimators (except for (d)) involve preliminary estimation of nuisance parameters — the conditional probabilities NP1 and the conditional expectations NP2 — and we account for this to obtain the

standard deviations of the corresponding efficient GMM and IPW estimators. We posit parametric models for NP1 and NP2. In particular, considering the observed history of: (i) the student's scores relative to those in their school, and (ii) the student's socioeconomic status, we posit that:

NP1. The hazard of leaving at the end of any grade $j = K, 1, 2$ conditional on the observables until the end of grade j is logit with an index that is linear in the following three types of dummy variables for each grade prior to and including grade j. First, whether the student qualified for free lunch in school. Second (respectively, third) whether the student's total (math + reading) score exceeded the mean total score for all the students in the same grade in small (respectively, non-small) classes in the student's school. Thus, besides the intercept, there are three, six, and nine dummies entering the logit when $j = K, 1$, and 2 respectively.

NP2. The expectation of the counterfactual math and reading scores in grade 3 conditional on the observables until the end of grade $j = K, 1, 2$ is linear in the following five types of dummy variables for each grade prior to and including grade j. First, whether the student qualified for free lunch in school. Second/fourth (respectively, third/fifth) whether the student's math/reading score exceeded the mean of the same for all the students in the same grade in small (respectively, non-small) classes in the student's school. Thus, besides the intercept, there are five, ten, and fifteen dummies entering the specification when $j = K, 1$, and 2 respectively.

Irrespective of the students' own class type, NP1 and NP2 include comparison of their grades with students in both small and non-small classes in their school. These models are motivated by the framework laid out in Section 4. It is worth noting that the goal of these models is not to obtain the best possible fit but rather to allow us to control for the selection induced by attrition. Tables 10-12 (Appendix D) contain the results of the preliminary estimation of these nuisance parameters.

5.2 Estimation results:

Table 6 presents the estimates of the 40 parameters of interest: 4 parameters (2 class types and 2 subjects) for each of the 10 attrition categories (a)-(j) defined above.

The estimates of the parameters for (a), (b), and (c) are quite close and are lower than that for (d). This reflects the pattern induced by the attrition behavior on the descriptive statistics of the observed scores (e.g., in grade K) in Tables 8-9. However, the differences of our estimates for (d) with that for (a), (b), and (c) are actually lower than what simple (e.g., linear), unconditional extrapolations of these descriptive statistics of scores would suggest. The estimates for the other

parameters, including that for the full-population, i.e., (j), are weighted averages of those for (a)-(d).

The Eff and IPW estimates are very similar to each other, and this holds in all cases (a)-(j). The minimum p-value for the 20 Hausman tests (2 subjects and 10 attrition categories) comparing the differences between Eff and IPW estimates is .285 for small class and .215 for non-small class.[4]

Left STAR school at the end of grade	Randomized to small class				Randomized to non-small class			
	grade 3 math score		grade 3 read score		grade 3 math score		grade 3 read score	
	Eff	IPW	Eff	IPW	Eff	IPW	Eff	IPW
(a) K	616.93	617.59	619.95	620.28	611.85	610.43	609.90	608.44
	(1.30)	(1.84)	(1.11)	(1.83)	(0.74)	(1.26)	(0.69)	(1.11)
(b) 1	614.82	615.97	617.54	618.40	607.11	605.58	605.29	603.70
	(2.09)	(2.82)	(1.78)	(2.54)	(1.14)	(1.56)	(1.02)	(1.40)
(c) 2	616.08	619.91	617.71	621.47	614.29	616.67	612.69	614.00
	(2.72)	(3.68)	(2.33)	(3.30)	(1.85)	(2.39)	(1.77)	(2.25)
(d) 3 (never left)	633.44	633.44	632.62	632.62	626.56	626.56	623.68	623.68
	(1.52)	(1.52)	(1.43)	(1.43)	(1.05)	(1.05)	(0.97)	(0.97)
(e) K or 1	616.27	617.08	619.19	619.69	610.35	608.88	608.44	606.95
	(1.22)	(1.94)	(1.08)	(1.92)	(0.70)	(1.26)	(0.64)	(1.13)
(f) 1 or 2	615.27	617.38	617.60	619.50	609.21	608.78	607.45	606.71
	(1.71)	(2.68)	(1.49)	(2.45)	(1.00)	(1.51)	(0.93)	(1.38)
(g) 2 or 3	631.13	631.64	630.63	631.13	625.03	625.32	622.30	622.46
	(1.38)	(1.55)	(1.30)	(1.46)	(0.96)	(1.06)	(0.88)	(0.97)
(h) K or 1 or 2	616.24	617.50	618.97	619.95	610.80	609.78	608.93	607.76
	(1.19)	(2.00)	(1.07)	(1.97)	(0.69)	(1.27)	(0.63)	(1.14)
(i) 1 or 2 or 3	627.98	628.62	628.11	628.68	620.85	620.71	618.34	618.10
	(1.21)	(1.64)	(1.13)	(1.54)	(0.80)	(1.09)	(0.74)	(1.00)
(j) K or 1 or 2 or 3	624.72	625.36	625.70	626.20	617.84	617.27	615.51	614.86
	(0.99)	(1.56)	(0.92)	(1.53)	(0.62)	(1.07)	(0.58)	(0.96)

Table 6: The efficient GMM and IPW estimates from (7) and (11) of the expected grade 3 math and reading score by class types that the students were initially randomized to and their attrition period (grade). 40 parameters of interest are estimated. Standard deviations are inside parentheses.

Now, we turn to our main focus: the precision of estimates. Recall from Section 3.1 that Eff is not necessarily efficient unless assumption (1), and specifications NP1-NP2 are all correct; and this cannot be ensured or tested with real life data (see footnote 4). Nevertheless, comparing the standard deviations (inside parentheses) in Table 6, it is clear that Eff is much more precise than

[4]Four technical remarks are in order regarding these similarities before we can turn to the discussion of precision of the estimates. First, it is important to note that the underlying null hypothesis for each Hausman test is that the corresponding probability limits (assuming they exist) for the Eff and IPW estimators are the same. Second, even if all these null hypotheses were true, it does not necessarily imply that our selection on observables assumption (1), and our specifications NP1-NP2 are all correct. Therefore, Eff may not be efficient under these nulls, and hence the variance estimator in the Hausman tests is not simply the difference in the estimated variances of the IPW and the Eff estimators. Third, however, thanks to the double-robustness property of Eff (see Section 2.3.6), the results of these Hausman tests at least indicate that if NP2 and (1) were correct, then we cannot reject the correctness of NP1. Fourth, and finally, it would have been preferable to discuss the precision of estimates with the benefit of having even larger minimum p-values for such null hypotheses. The reason for this will be evident later in footnote 6 when we discuss the nesting behavior of the Eff and IPW confidence intervals while comparing their widths. Roughly speaking, the features that would have led to larger p-values would have also ensured the nesting of the Eff intervals by IPW.

IPW. Indeed, the efficiency gain that results from the use of Eff is substantive by most practical standards. Excluding the four parameters for (d) where Eff and IPW are necessarily identical, this efficiency gain, as defined below Table 4 in Section 4.2, is never less than 9.9%, between 25-50% for fourteen parameters, between 50-75% for twelve parameters, and over 75% for six parameters.

This substantive efficiency gain due to Eff is also useful for comparing between sub-populations of small and non-small classes to find the effect of small class on the students' performance. Although this section is only intended as an illustration of the increased precision and is not remotely a serious study of the effect of small class, let us now demonstrate the efficiency gains due to Eff in this context since such comparisons are of empirical interest and are routinely done in practice.

To get a sense of the magnitude of the differences between small and non-small classes, it is a common practice to scale the scores. We use the standard deviation (std), 40.11 and 37.13, of the observed grade 3 math and reading scores respectively as the scale in the respective comparisons. In terms of the columns of Table 6, all the comparisons considered below will be between:

1. For math: columns 2 and 6 in the case of Eff, and columns 3 and 7 in the case of IPW.
2. For reading: columns 4 and 8 in the case of Eff, and columns 5 and 9 in the case of IPW.

Typically, interest lies in the comparison of the parameters for (j) (i.e., full-population). Under our assumptions, the effect of attrition has been corrected for in the concerned estimates. Hence, the initial randomization would imply that the difference for (j) is the effect on the average grade 3 scores of being in a small class as opposed to a non-small class in consecutive grades from K to 3.

Using Eff and IPW, we find a 95% confidence interval of this effect on the math (reading) score to be $[.11, .23]$ and $[.11, .29]$ math std ($[.22, .33]$ and $[.21, .40]$ reading std) respectively. See Table 7, row (j). Not only are these Eff intervals nested in the IPW intervals, but also they help to substantially narrow the range of estimates at the upper end of the intervals by ruling out very large effects. The IPW interval is 61% and 66% wider than the Eff interval for math and reading.

Similar comparisons in Table 7 between small and non-small classes by the other attrition categories (a)-(i) reveal similar effectiveness of the Eff confidence intervals in terms of its narrowness both in absolute (unit: math std or reading std) and percentage terms.[5] (Of course, Eff and IPW intervals are necessarily identical for (d).) While the lower bound of an Eff interval can be a little smaller in many cases than that of the corresponding IPW interval (by a maximum of 2% of math

[5]We do not interpret the differences for (a),...,(i) as "effects" since these are conditional on the attrition (i.e., post randomization) behavior. Instead, we consider them as descriptive comparisons between counterfactuals and interpret them, e.g., in the case of (a) as the difference between what would have been the average grade 3 math (reading) scores of students who were initially randomized to small and non-small classes respectively but left the participating school at the end of grade K, had they instead continued in their initially assigned class until grade 3.

or reading std), the upper bound of the Eff interval is always much smaller — frequently by 10% of math or reading std — than that of IPW.[6] It is indeed noteworthy that Eff, but not IPW, can rule out these wide regions of economically large changes in scores at the upper end of the intervals.

Left STAR school	grade 3 math score (standardized)			grade 3 read score (standardized)		
at the end of grade	Eff	IPW	% wider	Eff	IPW	% wider
(a) K	$[0.05, 0.20]$	$[0.07, 0.29]$	49.08	$[0.20, 0.34]$	$[0.21, 0.43]$	64.55
(b) 1	$[0.08, 0.31]$	$[0.10, 0.42]$	35.01	$[0.22, 0.44]$	$[0.24, 0.55]$	41.56
(c) 2	$[-0.12, 0.21]$	$[-0.13, 0.30]$	33.51	$[-0.02, 0.29]$	$[-0.01, 0.41]$	36.66
(d) 3 (never left)	$[0.08, 0.26]$	$[0.08, 0.26]$	0	$[0.15, 0.33]$	$[0.15, 0.33]$	0
(e) K or 1	$[0.08, 0.22]$	$[0.09, 0.32]$	63.85	$[0.22, 0.36]$	$[0.23, 0.46]$	77.08
(f) 1 or 2	$[0.05, 0.25]$	$[0.06, 0.36]$	55.00	$[0.18, 0.37]$	$[0.20, 0.49]$	60.39
(g) 2 or 3	$[0.07, 0.23]$	$[0.07, 0.25]$	11.45	$[0.14, 0.31]$	$[0.14, 0.33]$	11.98
(h) K or 1 or 2	$[0.07, 0.20]$	$[0.08, 0.31]$	72.06	$[0.20, 0.34]$	$[0.21, 0.45]$	82.99
(i) 1 or 2 or 3	$[0.11, 0.25]$	$[0.10, 0.29]$	35.54	$[0.19, 0.33]$	$[0.19, 0.38]$	36.36
(j) K or 1 or 2 or 3	$[0.11, 0.23]$	$[0.11, 0.29]$	61.35	$[0.22, 0.33]$	$[0.21, 0.40]$	65.62

Table 7: 95% confidence intervals using Eff and IPW for the expected difference in the counterfactual math and reading scores between small and non-small classes by attrition category. "% wider" stands for the percentage by which an IPW interval is wider than the corresponding Eff interval.

6 Conclusion

The demonstrations in Sections 4 and 5 based on simulated and real (Project STAR) data respectively provided a remarkably encouraging picture for the overall performance of the efficient GMM estimator. As noted in Section 3, this estimator falls under the class of the well-known doubly robust AIPW estimators that have been extensively studied in the biostatistics, epidemiology, statistics and (recent) econometrics literatures. Importantly, this estimator and the estimator of its asymptotic variance do not pose any new theoretical or computational challenge to the practitioner.

On the other hand, our main result Proposition 1 showed that this estimator is actually semiparametrically efficient for a wide variety of target populations characterized by the missingness of monotonically missing at random data. The discussion in Section 2 showed that these targets nest those for which similar estimators have long been proposed and also fruitfully employed in practice. Hence, we hope that our main result, its implications, and the demonstration in our paper would encourage the use of the efficient GMM estimation in the broader setting of missing data, including pattern mixture models, i.e., for any target population that falls under the premise of our paper.

[6]The small margins at the lower end that prevent the nesting of the Eff intervals by the IPW intervals arise for the following reason. Although the Eff and IPW estimates are individually similar, Eff is a little smaller than IPW for small classes and a little larger (except for (c) and (g)) for non-small classes. Since the confidence intervals here are based on the differences between the concerned Eff (or IPW) estimates for small and non-small classes, these little dissimilarities with opposite signs now accumulate making the Eff estimates typically about .05 std smaller than the IPW estimates. So, the centers of the intervals are not aligned. The larger variance of IPW in this illustration is not sufficient to ensure the nesting of Eff by fully offsetting the nonaligned centers, albeit what is not offset is very small.

Appendix A: Technical endnotes for Section 2

A.1 The expression of $\varphi(O;\beta)$ in (4) when $a=1, b=R$ [see Section 2.3.1]:

$$\begin{aligned}
\varphi(O;\beta) &= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} P(C\leq r-1|T_{r-1}) \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad + \sum_{r=1}^{R} I(C=r) E[m(T_R;\beta)|T_r] \\
&= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad - \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} P(C\geq r|T_{r-1}) \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad + \sum_{r=1}^{R} I(C=r) E[m(T_R;\beta)|T_r] \\
&= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad - \sum_{r=2}^{R} I(C\geq r) \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) \text{ [by using (2)]} \\
&\quad + \sum_{r=1}^{R} I(C=r) E[m(T_R;\beta)|T_r] \\
&= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) \\
&\quad - \left\{\sum_{r=2}^{R} I(C=r) E[m(T_R;\beta)|T_r] - I(C\geq 2) E[m(T_R;\beta)|T_1]\right\} + \sum_{r=1}^{R} I(C=r) E[m(T_R;\beta)|T_r] \\
&= \sum_{r=2}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \left(E[m(T_R;\beta)|T_r] - E[m(T_R;\beta)|T_{r-1}]\right) + E[m(T_R;\beta)|T_1]. \ \blacksquare
\end{aligned}$$

A.2 The weighted average of sub-population $\varphi(O;\beta)$'s as discussed in Section 2.3.5:

Write $E[m(T_R;\beta)|T_r]$ as q_r for all r for brevity. Denote $\varphi(O;\beta)$ when $a=b=j$ by $\varphi_{[j,j]}(O;\beta)$ (see Section 2.3.2). Then, the weighted average with weights $\omega_j := P(C=j)/P(a\leq C\leq b)$ is:

$$\begin{aligned}
\sum_{j=a}^{b} \omega_j \varphi_{[j,j]}(O;\beta) &= \sum_{j=a}^{b} \left\{ \sum_{r=j+1}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \frac{P(C=j|T_j)}{P(a\leq C\leq b)} (q_r - q_{r-1}) + \frac{I(C=j)}{P(a\leq C\leq b)} q_j \right\} \\
&= \sum_{r=b+1}^{R} \frac{I(C\geq r)}{P(C\geq r|T_r)} \frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)} (q_r - q_{r-1}) \sum_{r=a}^{b} \frac{I(C=r)}{P(a\leq C\leq b)} q_r \\
&\quad + \sum_{j=a}^{b} \sum_{r=j+1}^{b} \frac{I(C\geq r)}{P(C\geq r|T_r)} \frac{P(C=j|T_j)}{P(a\leq C\leq b)} (q_r - q_{r-1})
\end{aligned}$$

where the first term follows by (1). Matching the first two terms with the terms on lines one and three of (4), the demonstration will be complete if the third term is equal to the term on the second line of (4). This follows by interchanging the order of the summations (allowed) and noting that:

$$\begin{aligned}
& \sum_{j=a}^{b} \sum_{r=j+1}^{b} \frac{I(C \geq r)}{P(C \geq r|T_r)} \frac{P(C = j|T_j)}{P(a \leq C \leq b)} (q_r - q_{r-1}) \\
= & \sum_{r=a+1}^{b} \sum_{j=a}^{r-1} \frac{I(C \geq r)}{P(C \geq r|T_r)} \frac{P(C = j|T_j)}{P(a \leq C \leq b)} (q_r - q_{r-1}) \\
= & \sum_{r=a+1}^{b} \frac{I(C \geq r)}{P(C \geq r|T_r)} \frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)} (q_r - q_{r-1}),
\end{aligned}$$

where the last line follows by (1). This expression is equal to the term on the second line of (4). ■

A.3 Double-robustness of the expression of $\varphi(O;\beta)$ in (4) [see Section 2.3.6]:

To see this, first replace the unknown $P(a \leq C \leq r|T_r)$ for $r = a+1, \ldots, b$ and $1/P(C \geq r|T_r)$ for $r = a+1, \ldots, R$ in (4) by any integrable scalar functions of T_r, and then note that $E[\varphi(O;\beta^0)] = 0$ by (3) applied to the last line of (4).

On the other hand, rearranging the terms in (4) gives $\varphi(O;\beta)$ alternatively as:

$$\begin{aligned}
& \frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)} \left[\frac{I(C \geq R)}{P(C \geq R|T_R)} m(T_R;\beta) \right. \\
& \qquad \left. + \sum_{r=b+1}^{R-1} \left(\frac{I(C \geq r)}{P(C \geq r|T_r)} - \frac{I(C \geq r+1)}{P(C \geq r+1|T_{r+1})} \right) E[m(T_R;\beta)|T_r] \right] \\
& + \sum_{r=a}^{b} \left\{ \left(\frac{I(C \geq r)}{P(C \geq r|T_r)} \frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)} - \frac{I(C \geq r+1)}{P(C \geq r+1|T_{r+1})} \frac{P(a \leq C \leq r|T_r)}{P(a \leq C \leq b)} \right) \right. \\
& \qquad \left. + \frac{I(C = r)}{P(a \leq C \leq b)} \right\} E[m(T_R;\beta)|T_r].
\end{aligned}$$

(In terms of the notation in this equation, $\{a \leq C \leq a-1\}$ should be considered a null event.)

Now, replacing in the above equation the unknown $E[m(T_R;\beta)|T_r]$ by any $d_m \times 1$ integrable functions of T_r for $r = 1, \ldots, R-1$, it still follows by (1) and (3) that $E[\varphi(O;\beta^0)] = 0$.

Therefore, taken together, we have now shown that $E[\varphi(O;\beta^0)] = 0$ even if either NP1 or NP2, i.e., any one set of nuisance parameters but not both, is misspecified inside the function $\varphi(O;\beta)$. This unbiasedness of $\varphi(O;\beta^0)$ for 0 leads to the doubly-robust estimating function. ■

Appendix B: Proof of Proposition 1

Notation: f and F denote the density and distribution functions, with the concerned random variables specified inside parentheses. Their conditional counterparts are denoted similarly. Let $L_0^2(F)$ denote the space of mean-zero, square integrable functions with respect to F.

We follow the three steps as in, e.g., Chen et al. (2008) to prove Proposition 1. 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 for a given rotation of $m(Z;\beta)$. Step 3 obtains the optimal rotation and, thereby, gives the efficiency bound as the expectation of the outer product of the efficient influence function.

STEP - 1: Consider a regular parametric sub-model indexed by a parameter θ for the distribution of the observed data $O = (C', T_C'(Z))'$. The log of the distribution is:

$$\log f_\theta(O) = \log f_\theta(Z_1) + \sum_{r=2}^{R} I(C \geq r) \log f_\theta(Z_r|Z_1,\ldots,Z_{r-1}) + \sum_{r=1}^{R} I(C = r) \log P_\theta(C = r|Z_1,\ldots,Z_r)$$

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

$$S_\theta(O) = s_\theta(Z_1) + \sum_{r=2}^{R} I(C \geq r) s_\theta(Z_r|Z_1,\ldots,Z_{r-1}) + \sum_{r=1}^{R} I(C = r) \frac{\dot{P}_\theta(C = r|Z_1,\ldots,Z_r)}{P_\theta(C = r|Z_1,\ldots,Z_r)}$$

where $s_\theta(Z_1) := \frac{\partial}{\partial\theta} \log f_\theta(Z_1)$, $s_\theta(Z_r|Z_1,\ldots,Z_{r-1}) := \frac{\partial}{\partial\theta} \log f_\theta(Z_r|Z_1,\ldots,Z_{r-1})$ for $r = 2,\ldots,R$, and $\dot{P}_\theta(C = r|Z_1,\ldots,Z_r) := \frac{\partial}{\partial\theta} P_\theta(C = r|Z_1,\ldots,Z_r)$ for $r = 1,\ldots,R$. For Step-2, it is useful to note from (1) and (2) that for any $r = 2,\ldots,R$:

$$\dot{P}_\theta(C \geq r|Z) = -\dot{P}_\theta(C \leq r-1|Z_1,\ldots,Z_{r-1}) = \dot{P}_\theta(C \geq r|Z_1,\ldots,Z_{r-1}). \tag{12}$$

The tangent set is the mean square closure of all d_β dimensional linear combinations of $S_\theta(O)$ for all such smooth parametric sub-models, and it can be generically defined as:

$$\mathcal{T} := \nu_1(Z_1) + \sum_{r=2}^{R} I(C \geq r) \nu_r(Z_1,\ldots,Z_r) + \sum_{r=1}^{R} I(C = r) \omega_r(Z_1,\ldots,Z_r), \tag{13}$$

where $\nu_1(Z_1) \in L_0^2(F(Z_1))$ and $\nu_r(Z_1,\ldots,Z_r) \in L_0^2(F(Z_r|Z_1,\ldots,Z_{r-1}))$ for $r = 2,\ldots,R$, and $\omega_r(Z_1,\ldots,Z_r)$ is any square integrable function of $Z_1,\ldots,Z_r$ for $r = 1,\ldots,R$.

STEP - 2: For brevity we write $m(Z;\beta^0)$ as m, and drop the subscript θ from all quantities evaluated at θ^0. The moment conditions in (3) for a given a, b are equivalent to the requirement that for any $d_\beta \times d_m$ matrix A, the following just-identified system of moment conditions holds:

$$AE[m|a \leq C \leq b] = AE\left[\frac{P(a \leq C \leq b|Z)}{P(a \leq C \leq b)}\frac{I(C=R)}{P(C=R|Z)}m\right] = 0$$

where the first equality follows from (1). Differentiating with respect to θ under the integral gives:

$$0 = AM\frac{\partial\beta^0(\theta_0)}{\partial\theta'} + AE\left[m\left\{s(Z)' + \frac{\dot{P}(a \leq C \leq b|Z)'}{P(a \leq C \leq b|Z)} - \frac{\dot{P}(a \leq C \leq b)'}{P(a \leq C \leq b)}\right\}\middle| a \leq C \leq b\right]$$

where $s(Z) := s(Z_1 + \sum_{r=2}^{R} s(Z_r|Z_1,\ldots,Z_{r-1})$ and $\dot{P}(a \leq C \leq b) := \frac{\partial}{\partial\theta}P_{\theta^0}(a \leq C \leq b)$. Taking a full row rank A along with (1), (3) and assumption (A3), gives:

$$\frac{\partial\beta^0(\theta_0)}{\partial\theta'} = -(AM)^{-1}A\left\{E\left[ms(Z)'|a \leq C \leq b\right] + \sum_{r=a}^{b} E\left[m\frac{\dot{P}(C=r|Z_1,\ldots,Z_r)'}{P(a \leq C \leq b)}\right]\right\}.$$

Now we establish that for this given A, $-(AM)^{-1}A\varphi(O;\beta^0)$ is the efficient influence function by showing that $E[-(AM)^{-1}A\varphi(O;\beta^0)S(O)'] = \frac{\partial\beta^0(\theta_0)}{\partial\theta'}$ and that $(AM)^{-1}A\varphi(O;\beta^0) \in \mathcal{T}$ defined in (13).

For this purpose, note by using (4) (and switching to the notation T_r for $(Z_1,\ldots,Z_r)$ when it helps brevity) that we can write $E[\varphi(O;\beta^0)S(O)'] = \sum_{i=1}^{3}\sum_{j=1}^{2} B_{ij}$ where:

$$\begin{aligned}
B_{11} &:= \sum_{r=b+1}^{R} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)D'\right],\\
B_{12} &:= \sum_{r=b+1}^{R} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)\sum_{k=1}^{R} I(C=k)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right],\\
B_{21} &:= \sum_{r=a+1}^{b} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)D'\right],\\
B_{22} &:= \sum_{r=a+1}^{b} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)\sum_{k=1}^{R} I(C=k)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right],\\
B_{31} &:= \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a \leq C \leq b)}E[m|T_r]D'\right],\\
B_{32} &:= \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a \leq C \leq b)}E[m|T_r]\sum_{k=1}^{R} I(C=k)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right],\\
D &:= s(Z_1) + \sum_{k=2}^{R} I(C \geq k)s(Z_k|T_{k-1}).
\end{aligned}$$

As noted above Proposition 1, we proceed with the understanding that if $b = R$ then $B_{11} = B_{12} = 0$, and if $a = b$ then $B_{21} = B_{22} = 0$. Also, for notational brevity define T_0 as any constant, so that $s(Z_1) \equiv s(Z_1|T_0)$. First, note that:

$$\begin{aligned}
B_{11} &= \sum_{r=b+1}^{R}\sum_{k=1}^{r} E\left[\frac{I(C \geq r)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right] \\
&\quad + \sum_{r=b+1}^{R}\sum_{k=r+1}^{R} E\left[\frac{I(C \geq k)}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right] \\
&= \sum_{r=b+1}^{R}\sum_{k=1}^{r} E\left[\frac{P(C \geq r|T_{r-1})}{P(C \geq r|T_{r-1})}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right] \\
&\quad + \sum_{r=b+1}^{R}\sum_{k=r+1}^{R} E\left[\frac{P(C \geq k|T_{k-1})}{P(C \geq r|T_r)}\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}\left(E[m|T_r] - E[m|T_{r-1}]\right)s(Z_k|T_{k-1})'\right] \\
&= \sum_{r=b+1}^{R} E\left[\frac{P(a \leq C \leq b|T_b)}{P(a \leq C \leq b)}E[m|T_r]s(Z_r|T_{r-1})'\right] + 0 \\
&= E\left[\frac{I(a \leq C \leq b)}{P(a \leq C \leq b)}ms(Z_R, \ldots, Z_{b+1}|T_b)'\right] = E\left[ms(Z_R, \ldots, Z_{b+1}|T_b)'|a \leq C \leq b\right] \qquad (14)
\end{aligned}$$

where the third and fourth lines follow by (2); the fifth line follows by noting that for all $k = 1, \ldots, r-1$: $E\left[(E[m|T_r] - E[m|T_{r-1}])s(Z_k|T_{k-1})'\right] = E\left[E[(E[m|T_r] - E[m|T_{r-1}])s(Z_k|T_{k-1})'|T_{r-1}]\right] = 0$ while for $k \geq r+1$: $E\left[E[m|T_r]s(Z_k|T_{k-1})'\right] = E\left[E[m|T_r]E[s(Z_k|T_{k-1})'|T_{k-1}]\right] = 0$; and the sixth (last) line follows by (1) and the definition of score.

Second, it now follows that:

$$B_{21} = \sum_{r=a+1}^{b} E\left[\frac{P(a \leq C \leq r-1|T_{r-1})}{P(a \leq C \leq b)}E[m|T_{r-1}]s(Z_r|T_{r-1})'\right]$$

exactly following the steps that led to the fifth line in the expression for B_{11} in (14) above. Therefore,

$$\begin{aligned}
B_{21} &= \sum_{r=a+1}^{b}\sum_{k=a}^{r-1} E\left[\frac{P(C = k|T_k)}{P(a \leq C \leq b)}ms(Z_r|T_{r-1})'\right] \\
&= \sum_{r=a+1}^{b}\sum_{k=a}^{r-1} E\left[ms(Z_r|T_{r-1})'|C = k\right]\frac{P(C = k)}{P(a \leq C \leq b)} \\
&= \sum_{k=a}^{b-1} E\left[m\sum_{r=k+1}^{b} s(Z_r|T_{r-1})'\middle| C = k\right]\frac{P(C = k)}{P(a \leq C \leq b)} \\
&= \sum_{k=a}^{b-1} E\left[ms(Z_b, \ldots, Z_{k+1}|T_k)'|C = k\right]\frac{P(C = k)}{P(a \leq C \leq b)} \qquad (15)
\end{aligned}$$

where the first line follows by (1); the second line follows by the same steps that gave the sixth line in (14); the third line follows by interchanging the order of summations (which is allowed here); and the fourth (last) line follows by the definition of score.

Third, we consider B_{31} and note that using the definition of score in the first line below and the same argument as before in the second (last) line below:

$$\begin{aligned} B_{31} &= \sum_{r=a}^{b}\sum_{k=1}^{r} E\left[\frac{I(C=r)}{P(a\leq C\leq b)}E[m|T_r]s(Z_k|T_{k-1})'\right] = \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(a\leq C\leq b)}E[m|T_r]s(T_r)'\right] \\ &= \sum_{r=a}^{b} E\left[ms(T_r)'|C=r\right]\frac{P(C=r)}{P(a\leq C\leq b)}. \end{aligned} \tag{16}$$

Now, we consider the terms B_{12}, B_{22} and B_{32} respectively. Accordingly, first note that:

$$\begin{aligned} B_{12} &= \sum_{r=b+1}^{R}\sum_{k=r}^{R} E\left[\frac{I(C=k)}{P(C\geq r|T_r)}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\frac{\dot{P}(C=k|T_k)'}{P(C=k|T_k)}\right] \\ &= \sum_{r=b+1}^{R} E\left[\frac{1}{P(C\geq r|T_r)}\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\sum_{k=r}^{R}\dot{P}(C=k|T_k)'\right] \\ &= \sum_{r=b+1}^{R} E\left[\frac{P(a\leq C\leq b|T_b)}{P(a\leq C\leq b)}\left(E[m|T_r]-E[m|T_{r-1}]\right)\frac{\dot{P}(C\geq r|T_{r-1})'}{P(C\geq r|T_{r-1})}\right] \\ &= 0 \end{aligned} \tag{17}$$

where the second line follows by (1); the third follows line by (1), (2) and (12); and the fourth (last) line follows by taking expectation conditional on T_{r-1} for the r-th term inside the summation. Exactly following the same steps as in the above (recall the analogy with B_{11} and B_{12}) we obtain:

$$B_{22} = 0. \tag{18}$$

Lastly, as before, note that:

$$B_{32} = \sum_{r=a}^{b} E\left[\frac{I(C=r)}{P(C=r|T_r)}\frac{E[m|T_r]\dot{P}(C=r|T_r)'}{P(a\leq C\leq b)}\right] = E\left[m\sum_{r=a}^{b}\frac{\dot{P}(C=r|T_r)'}{P(a\leq C\leq b)}\right]. \tag{19}$$

Therefore, (14)-(19) imply that $E[-(AM)^{-1}A\varphi(O;\beta^0)S(O)'] = \frac{\partial\beta^0(\theta_0)}{\partial\theta'}$. Finally, by matching the first set of terms in $-(AM)^{-1}A\varphi(O;\beta^0)$ (i.e., those that correspond to line one in (4)) to the terms corresponding to $\nu_{b+1}(Z_1,\ldots,Z_{b+1}),\ldots,\nu_R(Z_1,\ldots,Z_R)$ in $\mathcal{T}$; the second set of terms (i.e., those that correspond to line two in (4)) to the terms corresponding to $\nu_a(Z_1,\ldots,Z_a),\ldots,\nu_b(Z_1,\ldots,Z_b)$

in $\mathcal{T}$; and the third set of terms (i.e., those that correspond to line three in (4)) to the terms corresponding to $\omega_a(Z_1,\ldots,Z_a),\ldots,\omega_b(Z_1,\ldots,Z_b)$ in $\mathcal{T}$; while matching zeros with the remaining terms in $\mathcal{T}$, it follows that $-(AM)^{-1}A\varphi(O;\beta^0)$ is the efficient influence function given A.

STEP - 3: Standard arguments give that $A_* := \arg\inf_A Var((AM)^{-1}A\varphi(O;\beta^0)) = M'V^{-1}$. Thus $(A_*M)^{-1}A_*\varphi(O;\beta^0)$ is the efficient influence function, and Ω is the efficiency bound. ■

Appendix C: Working description for estimation and inference

Sort the units in the sample such that first n_1 units correspond to $C=1$, the next n_2 units to $C=2$, ..., while the last n_R units to $C=R$. For its relevance in practice, we focus on parametric estimation of the nuisance parameters in this discussion by assuming parametric models for the unknown conditional expectations and conditional probabilities (or hazards). Accordingly, let:

PM-NP1: For each $j=1,\ldots,R-1$, let there exist (i) a $d_{\gamma_j}\times 1$ vector $\gamma_j^0\in\Gamma_j$ where Γ_j is a compact subset of $\mathbb{R}^{d_{\gamma_j}}$, and (ii) a scalar function $p_j(T_j;\gamma_j)$ such that $P(C=j|T_j,C\geq j)=p_j(T_j;\gamma_j^0)$ for all $j=1,\ldots,R-1$. Let $\nabla_j p_j(T_j;\gamma_j) := \partial p_j(T_j;\gamma_j)/\partial\gamma_j$ exist and satisfy standard regularity conditions as in Newey and McFadden (1994). In the context of our Section 5:

$$p_j(T_j;\gamma_j)=\frac{\exp(T_j'\gamma_j)}{1+\exp(T_j'\gamma_j)} \quad \text{giving} \quad \nabla_j p_j(T_j;\gamma_j)=p_j(T_j;\gamma_j)(1-p_j(T_j;\gamma_j))T_j.$$

PM-NP2: For each $j=1,\ldots,R-1$, let there exist (i) a $d_{\delta_j}\times 1$ vector $\delta_j^0\in\Delta_j$ where Δ_j is a compact subset of $\mathbb{R}^{d_{\delta_j}}$, and (ii) a $d_m\times 1$ function $q_j(T_j;\delta_j,\beta)$ such that $E[m(Z;\beta^0)|T_j]=q_j(T_j;\delta_j^0,\beta^0)$ for all $j=1,\ldots,R-1$. Let $\nabla_j q_j(T_j;\delta_j,\beta) := \partial q_j(T_j;\delta_j,\beta)'/\partial\delta_j$ and $\nabla_\beta q_j(T_j;\delta_j,\beta) := \partial q_j(T_j;\delta_j,\beta)'/\partial\beta$ exist and satisfy standard conditions as in Newey and McFadden (1994). In Section 5, where β and Z are additively separable inside $m(Z;\beta)$, we take:

$$q_j(T_j;\delta_j,\beta)=T_j'\delta_j-\beta, \quad \text{giving} \quad \nabla_j q_j(T_j;\delta_j,\beta)=T_j \quad \text{and} \quad \nabla_\beta q_j(T_j;\delta_j,\beta)=-1.$$

Note that, using the relation between hazard and probability mass functions, PM-NP1 gives:

$$\begin{aligned}
P(C=1|T_1) &= p_1(T_1;\gamma_1^0), \\
P(C=j|T_j) &= p_j(T_j;\gamma_j^0)\prod_{r=1}^{j-1}\left(1-p_r(T_r;\gamma_r^0)\right) \quad \text{for} \quad j=2,\ldots,R-1, \\
P(C=R|T_R) &= 1-\sum_{r=1}^{R-1}P(C=r|T_r)=\prod_{r=1}^{R-1}\left(1-p_r(T_r;\gamma_r^0)\right),
\end{aligned} \tag{20}$$

and thereby giving its derivatives, denoted by $\nabla_k P(C = j|T_j) := \frac{\partial}{\partial \gamma_k} P(C = j|T_j)$ for all $k = 1, \ldots, R-1$ and $j = 1, \ldots, R$, as:

$$\begin{aligned}
\nabla_k P(C = j|T_j) &= 0 \text{ for } k = j+1, \ldots, R-1 \text{ and } j = 1, \ldots, R-1, \\
\nabla_1 P(C = 1|T_1) &= \nabla_1 p_1(T_1; \gamma_1^0), \\
\nabla_j P(C = j|T_j) &= \left(\nabla_j p_j(T_j; \gamma_j^0)\right) \prod_{r=1}^{j-1} \left(1 - p_r(T_r; \gamma_r^0)\right) \text{ for } j = 2, \ldots, R-1, \\
\nabla_k P(C = j|T_j) &= -\left(\nabla_k p_k(T_k; \gamma_k^0)\right) p_j(T_j; \gamma_j^0) \prod_{r=1}^{j-1} \frac{1 - p_r(T_r; \gamma_r^0)}{1 - p_k(T_k; \gamma_k^0)} \text{ for } k = 1, \ldots, j-1 \text{ and } j = 2, \ldots, R-1, \\
\nabla_k P(C = R|T_R) &= -\left(\nabla_k p_k(T_k; \gamma_k^0)\right) \prod_{r=1}^{R-1} \frac{1 - p_r(T_r; \gamma_r^0)}{1 - p_k(T_k; \gamma_k^0)} \text{ for } k = 1, \ldots, R-1.
\end{aligned}$$

It is also useful to note from these relationships that for $r = 2, \ldots, R$:

$$\begin{aligned}
P(C \geq r|T_r) &= 1 - P(C \leq r-1|T_{r-1}) = 1 - \prod_{j=1}^{r-1}(1 - p_j(T_j; \gamma_j^0)), \\
\text{and hence } \nabla_k P(C \geq r|T_r) &= 0 \text{ for } k = r, \ldots, R-1, \\
\nabla_k P(C \geq r|T_r) &= \left(\nabla_k p_k(T_k; \gamma_k^0)\right) \prod_{j=1}^{r-1} \frac{1 - p_j(T_j; \gamma_j^0)}{1 - p_k(T_k; \gamma_k^0)} \text{ for } k = 1, \ldots, r-1.
\end{aligned} \tag{21}$$

Now, consider the estimation of the nuisance parameters. The quasi-maximum likelihood estimator $\widehat{\gamma}_j$ for γ_j^0 for $j = 1, \ldots, R-1$ solves the score equations $0 = \sum_{i=1}^n S_{n,j,i}(\gamma_j)$ where, $i = 1, \ldots, n$,

$$S_{n,j,i}(\gamma_j) := I(C_i \geq j) \frac{I(C_i = j) - p_j(T_j(Z_i); \gamma_j)}{p_j(T_j(Z_i); \gamma_j)(1 - p_j(T_j(Z_i); \gamma_j))} \nabla_j p_j(T_j(Z_i); \gamma_j).$$

Let $H_{n,j} := (\partial/\partial\gamma) \sum_{i=1}^n S_{n,j,i}(\gamma_j)'/(n - \bar{n}_{j-1})$ at $\widehat{\gamma}_j$, and let $\bar{n}_j = \sum_{i=1}^j n_i$ for $j = 1, \ldots, R-1$. Let H_n be the block-diagonal matrix with $H_{n,j}$ as the j-th diagonal block for $j = 1, \ldots, R-1$.

The nonlinear least squares estimator $\widehat{\delta}_j$ for δ_j^0 for $j = 1, \ldots, R-1$ is $\widehat{\delta}_j := \widehat{\delta}_j(\widehat{\beta})$ where $\widehat{\beta}$ is the GMM estimator (to be defined below) while $\widehat{\delta}_j(\beta)$ is the solution of the estimating equations $0 = \sum_{i=1}^n L_{n,j,i}(\delta_j, \beta)$ where, for $j = 1, \ldots, R-1$ and $i = 1, \ldots, n$:

$$L_{n,j,i}(\delta_j, \beta) := I(C_i = R) \left(\nabla_j q_j(T_j(Z_i); \delta_j, \beta)\right) B_{n,j} \left(m(Z_i; \beta) - q_j(T_j(Z_i); \delta_j, \beta)\right).$$

For simplicity, we take the weighting matrix $B_{n,j} = I_{d_m}$ in the nonlinear least squares. Let $A_{n,j} := (\partial/\partial\delta_j) \sum_{i=1}^n L_{n,j,i}(\widehat{\delta}_j, \widehat{\beta})'/n_R$ at $\widehat{\delta}_j$ and $\widehat{\beta}$. Let A_n be a block-diagonal matrix with $A_{n,j}$ as its j-th diagonal block for $j = 1, \ldots, R-1$.

Now, consider the estimation of the actual parameter of interest, β. Note from Sections 2.3.5 and 3 that a consideration of the individual sub-populations is sufficient because the estimating equation for their contiguous unions is simply a weighted average of that for the individual sub-populations where the weight for sub-population $j \in \{a, \ldots, b\}$ can be estimated as $n_j/(\bar{n}_b - \bar{n}_{a-1})$.

Furthermore, to focus on the essential idea, abstract from the weighting matrix W_n in (7) by considering a just-identified system, i.e, $d_m = d_\beta$. Then, using the result from Section 2.3.2, note that, for the sub-population j: $\widehat{\beta}$ is the solution of $0 = \bar{g}_n(\beta, \widehat{p}(.), \widehat{q}(.;\beta))$ (see (7)), more precisely,

$$\begin{aligned} 0 &= \sum_{r=j+1}^{R} \frac{n}{n_j} \sum_{i=\bar{n}_{r-1}+1}^{n} \frac{\widehat{P}(C=j|T_j(Z_i))}{\widehat{P}(C \geq r|T_r(Z_i))} \left(\widehat{E}_n[m(Z;\beta)|T_r(Z_i)] - \widehat{E}_n[m(Z;\beta)|T_r(Z_i)] \right) \\ &\quad + \frac{n}{n_j} \sum_{i=\bar{n}_{j-1}+1}^{\bar{n}_j} \widehat{E}_n[m(Z;\beta)|T_j(Z_i)], \end{aligned}$$

where: $\widehat{E}_n[m(Z;\beta)|T_r(Z_i)] := q_r(T_r(Z_i), \widehat{\delta}_r(\beta), \beta)$ for $r = 1, \ldots, R-1$, while $\widehat{E}_n[m(Z;\beta)|T_R(Z_i)] := m(Z_i;\beta)$ by convention. On the other hand, $\widehat{P}(C=j|T_j(Z_i))$ and $\widehat{P}(C \geq r|T_r(Z_i))$ are obtained by (20) and (21) respectively with the corresponding γ_s^0's replaced by $\widehat{\gamma}_s$'s defined above for the appropriate $s = 1, \ldots, R-1$. To write these estimating equations for β more explicitly, stack the two sets of nuisance parameters as $\gamma = (\gamma_1', \ldots, \gamma_{R-1}')'$ and $\delta = (\delta_1', \ldots, \delta_{R-1}')'$, and then define:

$$\begin{aligned} \psi_i(\beta,\gamma,\delta) &:= \frac{n}{n_j} \frac{p_j(T_j(Z_i);\gamma_j) \prod_{r=1}^{j-1} (1 - p_r(T_r(Z_i);\gamma_r))}{\prod_{k=1}^{R-1}(1 - p_k(T_k(Z_i);\gamma_k))} (m(Z_i;\beta) - q_{R-1}(T_{R-1}(Z_i), \delta_{R-1}, \beta)) \\ &\quad + \frac{n}{n_j} \sum_{r=j+1}^{R-1} \frac{p_j(T_j(Z_i);\gamma_j) \prod_{r=1}^{j-1} (1 - p_r(T_r(Z_i);\gamma_r))}{1 - \prod_{k=1}^{r-1}(1 - p_k(T_k(Z_i);\gamma_k))} (q_r(T_r(Z_i), \delta_r, \beta) - q_{r-1}(T_{r-1}(Z_i), \delta_{r-1}, \beta)) \\ &\quad + \frac{n}{n_j} q_j(T_j(Z_i), \delta_j, \beta) \quad \text{for } i = 1, \ldots, n. \end{aligned}$$

Now, write the above estimating equations for β equivalently as: $0 = \bar{\psi}_n(\widehat{\beta}, \widehat{\gamma}, \widehat{\delta})$ where $\bar{\psi}_n(\beta,\gamma,\delta) := \sum_{i=1}^n \psi_i(\beta,\gamma,\delta)/n$. Let $\Psi_{n,\gamma} := (\partial/\partial\gamma)\bar{\psi}_n(\widehat{\beta}, \widehat{\gamma}, \widehat{\delta})'$, $\Psi_{n,\delta} := (\partial/\partial\delta)\bar{\psi}_n(\widehat{\beta}, \widehat{\gamma}, \widehat{\delta})'$ and $\Psi_{n,\beta} := (\partial/\partial\beta)\bar{\psi}_n(\widehat{\beta}, \widehat{\gamma}, \widehat{\delta})'$.

Define $\varepsilon_{n,i}(\gamma,\delta,\beta) := (S_{n,1,i}(\gamma_1)', \ldots, S_{n,R-1,i}(\gamma_{R-1})', L_{n,1,i}(\delta_1,\beta)', \ldots, L_{n,R-1,i}(\delta_{R-1},\beta)', \psi_i(\beta,\gamma,\delta)')'$ by stacking the estimating functions for $(\gamma', \delta', \beta')'$. Let $\Upsilon = \sum_{i=1}^n \varepsilon_{n,i}(\widehat{\gamma}, \widehat{\delta}, \widehat{\beta})\varepsilon_{n,i}(\widehat{\gamma}, \widehat{\delta}, \widehat{\beta})'/n$. Then, the estimator of the asymptotic variance for $\widehat{\beta}$ is the lower-right $d_\beta \times d_\beta$ block of the matrix:

$$\begin{bmatrix} H_n & 0 & 0 \\ 0 & A_n & 0 \\ \Psi_{n,\gamma} & \Psi_{n,\delta} & \Psi_{n,\beta} \end{bmatrix}^{-1'} \Upsilon \begin{bmatrix} H_n & 0 & 0 \\ 0 & A_n & 0 \\ \Psi_{n,\gamma} & \Psi_{n,\delta} & \Psi_{n,\beta} \end{bmatrix}^{-1}.$$

References

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.

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

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

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

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

Chaudhuri, S. (2017). A Note on Efficiency Gains from Multiple Incomplete Subsamples. Mimeo.

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.

Diggle, P., Farewell, D., and Henderson, R. (2007). Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *JRSS, Series C*, 56: 499–550.

Ding, W. and Lehrer, S. F. (2010). Estimating treatment effects from contaminated multiperiod education experiments: the dynamic impacts of class size reductions. *The Review of Economics and Statistics*, 92: 31–42.

Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1996). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. In *Working Paper Series*. NBER.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). *Selecion modeling versus mixture modeling with nonignorable nonresponses*, pages 115–142. Springer-Verlag, NY.

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

Hanushek, E. A. (1999). Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects. *Educational Evaluation and Policy Analysis*, 21: 143–63.

Heckman, J., Smith, J., and Taber, C. (1998). Accounting for Dropouts in Evaluations of Social Programs. *Review of Economics and Statistics*, LXXX: 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.

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

Kang, J. and Schafer, J. (2007). Demystifying Double Robustness :A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22: 523–539.

Krueger, A. B. (1999). Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics*, 114: 497–532.

Little, R, J. A. and Rubin, D. D. (2002). *Statistical Analysis with Missing Data*. Wiley - Interscience.

Little, R. J. A. (1993). Pattern-Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association*, 88: 125–134.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81: 471–483.

Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. (1998). Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica*, 52: 153–161.

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

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.

Nicoletti, C. (2006). Nonresponse in dynamic panel data models. *Journal of Econometrics*, 132: 461–489.

Robins, J. and Ritov, Y. (1997). Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theroy for Semi-Parametric Models. *Statistics in Medicine*, 16: 285–319.

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.

Rothe, C. and Firpo, S. (2016). Semiparametric Estimation and Inference Using doubly-Robust Moment Conditions. Mimeo.

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

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

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

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science*, 22: 560–568.

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

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

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Appendix D: Supplemental tables for Section 5

Observed variables	For students who left STAR school at the end of grade							
	K		1		2		3 (never left)	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
math score in grade K	473.01	51.14	478.50	50.09	481.75	46.70	507.94	46.07
reading score in grade K	430.74	32.14	432.36	30.57	433.63	27.37	450.87	32.63
free lunch in grade K	0.45	0.50	0.47	0.50	0.44	0.50	0.63	0.48
math in grade K > mean math in grade K(small)	0.37	0.48	0.36	0.48	0.43	0.50	0.59	0.49
math in grade K > mean math in grade K(non-small)	0.40	0.49	0.45	0.50	0.48	0.50	0.69	0.46
read in grade K > mean read in grade K(small)	0.33	0.47	0.38	0.49	0.36	0.48	0.58	0.49
read in grade K > mean read in grade K(non-small)	0.43	0.50	0.44	0.50	0.51	0.50	0.67	0.47
(math + read) in grade K > mean(math+read) in grade K(small)	0.35	0.48	0.35	0.48	0.40	0.49	0.58	0.49
(math + read) in grade K > mean(math+read) in grade K(non-small)	0.38	0.49	0.43	0.50	0.47	0.50	0.69	0.46
math score in grade 1			520.39	50.35	526.30	39.31	553.46	40.60
reading score in grade 1			510.21	62.78	513.75	50.96	548.76	53.28
free lunch in grade 1			0.41	0.49	0.46	0.50	0.62	0.49
math in grade 1 > mean math in grade 1(small)			0.35	0.48	0.38	0.49	0.60	0.49
math in grade 1 > mean math in grade 1(non-small)			0.46	0.50	0.48	0.50	0.68	0.47
read in grade 1 > mean read in grade 1(small)			0.31	0.47	0.32	0.47	0.54	0.50
read in grade 1 > mean read in grade 1(non-small)			0.39	0.49	0.47	0.50	0.65	0.48
(math + read) in grade 1 > mean(math+read) in grade 1(small)			0.29	0.45	0.37	0.49	0.58	0.49
(math + read) in grade 1 > mean(math+read) in grade 1(non-small)			0.44	0.50	0.49	0.50	0.67	0.47
math score in grade 2					568.70	53.35	600.43	43.38
reading score in grade 2					573.02	49.08	605.34	42.42
free lunch in grade 2					0.37	0.49	0.63	0.48
math in grade 2 > mean math in grade 2(small)					0.33	0.47	0.56	0.50
math in grade 2 > mean math in grade 2(non-small)					0.44	0.50	0.60	0.49
read in grade 2 > mean read in grade 2(small)					0.35	0.48	0.56	0.50
read in grade 2 > mean read in grade 2(non-small)					0.37	0.49	0.60	0.49
(math + read) in grade 2 > mean(math+read) in grade 2(small)					0.34	0.48	0.55	0.50
(math + read) in grade 2 > mean(math+read) in grade 2(non-small)					0.41	0.49	0.62	0.49
math score in grade 3							633.44	39.69
reading score in grade 3							632.62	37.48

Table 8: Mean and standard deviation (Std) of the observed variables by the attrition period (grade) for the students who were initially randomized to a small class.

Observed variables	For students who left STAR school at the end of grade							
	K		1		2		3 (never left)	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
math score in grade K	466.79	47.60	474.71	43.58	485.91	48.81	496.07	43.90
reading score in grade K	425.33	29.61	428.24	26.62	436.56	31.54	443.17	30.88
free lunch in grade K	0.42	0.49	0.39	0.49	0.40	0.49	0.64	0.48
math in grade K > mean math in grade K(small)	0.32	0.47	0.33	0.47	0.42	0.49	0.47	0.50
math in grade K > mean math in grade K(non-small)	0.36	0.48	0.40	0.49	0.54	0.50	0.59	0.49
read in grade K > mean read in grade K(small)	0.26	0.44	0.29	0.45	0.34	0.47	0.44	0.50
read in grade K > mean read in grade K(non-small)	0.34	0.47	0.34	0.47	0.51	0.50	0.55	0.50
(math + read) in grade K > mean(math+read) in grade K(small)	0.28	0.45	0.29	0.45	0.41	0.49	0.46	0.50
(math + read) in grade K > mean(math+read) in grade K(non-small)	0.35	0.48	0.36	0.48	0.54	0.50	0.58	0.49
math score in grade 1			507.06	44.23	523.93	39.58	542.86	39.70
reading score in grade 1			490.85	54.45	508.38	48.67	538.79	51.86
free lunch in grade 1			0.39	0.49	0.38	0.49	0.64	0.48
math in grade 1 > mean math in grade 1(small)			0.21	0.41	0.34	0.47	0.45	0.50
math in grade 1 > mean math in grade 1(non-small)			0.30	0.46	0.42	0.49	0.56	0.50
read in grade 1 > mean read in grade 1(small)			0.18	0.38	0.33	0.47	0.46	0.50
read in grade 1 > mean read in grade 1(non-small)			0.27	0.44	0.44	0.50	0.54	0.50
(math + read) in grade 1 > mean(math+read) in grade 1(small)			0.18	0.38	0.35	0.48	0.45	0.50
(math + read) in grade 1 > mean(math+read) in grade 1(non-small)			0.27	0.44	0.45	0.50	0.56	0.50
math score in grade 2					568.89	43.24	591.68	42.39
reading score in grade 2					569.01	44.69	597.68	43.28
free lunch in grade 2					0.39	0.49	0.64	0.48
math in grade 2 > mean math in grade 2(small)					0.35	0.48	0.46	0.50
math in grade 2 > mean math in grade 2(non-small)					0.41	0.49	0.51	0.50
read in grade 2 > mean read in grade 2(small)					0.38	0.49	0.48	0.50
read in grade 2 > mean read in grade 2(non-small)					0.41	0.49	0.51	0.50
(math + read) in grade 2 > mean(math+read) in grade 2(small)					0.35	0.48	0.45	0.50
(math + read) in grade 2 > mean(math+read) in grade 2(non-small)					0.38	0.49	0.51	0.50
math score in grade 3							626.56	39.41
reading score in grade 3							623.68	36.11

Table 9: Mean and standard deviation (Std) of the observed variables by the attrition period (grade) for the students who were initially randomized to a non-small class.

	From small class			From non-small class		
	Left STAR school at the end of grade					
Explanatory variables	K	1	2	K	1	2
free lunch in grade K	-0.44	0.15	-0.22	-0.48	-0.45	-0.25
	(0.12)	(0.25)	(0.34)	(0.08)	(0.17)	(0.26)
(math + read) in grade K > mean(math+read) in grade K(small)	0.09	-0.07	0.06	-0.04	0.16	0.07
	(0.18)	(0.24)	(0.31)	(0.12)	(0.16)	(0.21)
(math + read) in grade K > mean(math+read) in grade K(non-small)	-0.97	-0.50	-0.65	-0.66	-0.52	0.04
	(0.18)	(0.23)	(0.31)	(0.11)	(0.15)	(0.21)
free lunch in grade 1		-0.71	0.58		-0.35	-0.51
		(0.25)	(0.37)		(0.17)	(0.29)
(math + read) in grade 1 > mean(math+read) in grade 1(small)		-0.68	-0.23		-0.71	-0.08
		(0.24)	(0.32)		(0.17)	(0.22)
(math + read) in grade 1 > mean(math+read) in grade 1(non-small)		-0.15	-0.14		-0.45	-0.14
		(0.21)	(0.29)		(0.15)	(0.22)
free lunch in grade 2			-1.24			-0.41
			(0.38)			(0.27)
(math + read) in grade 2 > mean(math+read) in grade 2(small)			-0.19			0.00
			(0.34)			(0.23)
(math + read) in grade 2 > mean(math+read) in grade 2(non-small)			-0.22			-0.28
			(0.32)			(0.24)
intercept	-0.20	-0.49	-0.72	-0.15	-0.22	-1.16
	(0.10)	(0.15)	(0.20)	(0.06)	(0.09)	(0.14)
Number of Observations	1385	976	788	3126	2079	1595
Pseudo R^2	0.05	0.07	0.08	0.03	0.09	0.05

Table 10: Coefficients and standard deviation (in parentheses) of the explanatory variables in the logit model for the hazard of leaving the participating STAR school at the end of grade K, 1 and 2.

	From small class			From non-small class		
	Left STAR school at the end of grade					
Explanatory variables	K	1	2	K	1	2
free lunch in K	14.85	3.63	2.28	18.86	4.80	3.58
	(2.77)	(3.81)	(3.80)	(1.93)	(2.75)	(2.79)
math in K > mean math in K(small)	13.59	7.54	5.18	9.84	6.04	3.63
	(3.67)	(3.33)	(3.17)	(2.62)	(2.36)	(2.27)
math in K > mean math in K(non-small)	9.93	3.52	1.63	12.91	6.75	5.29
	(3.69)	(3.35)	(3.18)	(2.56)	(2.32)	(2.21)
read in K > mean read in K(small)	16.30	9.00	7.26	14.59	7.05	4.71
	(3.94)	(3.54)	(3.36)	(2.65)	(2.44)	(2.34)
read in K > mean read in K(non-small)	6.85	0.60	-1.51	3.98	-2.39	-4.11
	(3.92)	(3.52)	(3.35)	(2.56)	(2.36)	(2.26)
free lunch in grade 1		12.67	7.22		14.90	13.46
		(3.80)	(4.24)		(2.77)	(3.02)
math in 1 > mean math in 1(small)		16.14	6.13		15.61	10.76
		(3.43)	(3.44)		(2.25)	(2.18)
math in 1 > mean math in 1(non-small)		12.74	7.51		6.94	4.91
		(3.42)	(3.37)		(2.41)	(2.31)
read in 1 > mean read in 1(small)		18.17	15.15		12.60	6.94
		(3.55)	(3.38)		(2.28)	(2.25)
read in 1 > mean read in 1(non-small)		-2.39	-5.08		8.78	3.74
		(3.50)	(3.45)		(2.45)	(2.44)
free lunch in 2			5.58			0.62
			(3.97)			(2.84)
math in 2 > mean math in 2(small)			11.27			11.25
			(3.32)			(2.26)
math in 2 > mean math in 2(non-small)			9.13			6.39
			(3.41)			(2.43)
read in 2 > mean read in 2(small)			13.19			9.33
			(3.35)			(2.30)
read in 2 > mean read in 2(non-small)			2.62			6.24
			(3.35)			(2.38)
Intercept	594.49	582.89	580.36	593.44	583.23	581.53
	(3.05)	(2.95)	(2.82)	(1.93)	(1.83)	(1.77)
Number of Observations	683.00	683.00	683.00	1395.00	1395.00	1395.00
Adjusted R^2	0.25	0.41	0.47	0.24	0.39	0.45

Table 11: Coefficients and standard deviation (in parentheses) of the explanatory variables in the linear model for the expectation of the counterfactual grade 3 math score conditional on the variables observed until leaving the participating STAR school at the end of grade K, 1 and 2.

	From small class			From non-small class		
	Left STAR school at the end of grade					
Explanatory variables	K	1	2	K	1	2
free lunch in grade K	11.53	2.59	-0.11	18.03	4.83	1.57
	(2.72)	(3.83)	(3.79)	(1.76)	(2.48)	(2.44)
math in grade K > mean math in	10.57	5.28	4.20	6.02	3.64	1.15
grade K(small)	(3.61)	(3.35)	(3.16)	(2.40)	(2.13)	(1.98)
math in grade K > mean math in	12.94	5.55	3.07	10.53	4.43	3.12
grade K(non-small)	(3.63)	(3.37)	(3.17)	(2.34)	(2.09)	(1.93)
read in grade K > mean read in	7.79	1.76	0.38	7.63	1.80	0.94
grade K(small)	(3.87)	(3.56)	(3.35)	(2.42)	(2.20)	(2.05)
read in grade K > mean read in	6.98	1.85	-0.45	13.40	4.51	1.25
grade K(non-small)	(3.86)	(3.54)	(3.34)	(2.34)	(2.13)	(1.98)
free lunch in grade 1		8.73	3.36		13.54	10.17
		(3.82)	(4.23)		(2.49)	(2.64)
math in grade 1 > mean math in		5.36	-1.14		6.20	2.55
grade 1(small)		(3.45)	(3.44)		(2.02)	(1.91)
math in grade 1 > mean math in		21.54	15.18		11.36	8.89
grade 1(non-small)		(3.44)	(3.36)		(2.18)	(2.02)
read in grade 1 > mean read in		8.94	6.86		7.49	3.32
grade 1(small)		(3.57)	(3.38)		(2.05)	(1.97)
read in grade 1 > mean read in		5.78	-1.25		16.95	7.77
grade 1(non-small)		(3.52)	(3.44)		(2.20)	(2.14)
free lunch in grade 2			7.89			4.62
			(3.97)			(2.48)
math in grade 2 > mean math in			5.07			1.98
grade 2(small)			(3.32)			(1.98)
math in grade 2 > mean math in			14.65			10.19
grade 2(non-small)			(3.40)			(2.12)
read in grade 2 > mean read in			2.80			6.64
grade 2(small)			(3.34)			(2.01)
read in grade 2 > mean read in			13.30			16.34
grade 2(non-small)			(3.35)			(2.08)
Intercept	601.62	592.12	589.38	592.80	583.20	581.21
	(3.00)	(2.97)	(2.82)	(1.77)	(1.65)	(1.54)
Number of Observations	683.00	683.00	683.00	1395.00	1395.00	1395.00
Adjusted R^2	0.18	0.33	0.41	0.24	0.41	0.50

Table 12: Coefficients and standard deviation (in parentheses) of the explanatory variables in the linear model for the expectation of the counterfactual grade 3 reading score conditional on the variables observed until leaving the participating STAR school at the end of grade K, 1 and 2.