Econometric Theory, **0**, 2024, 1–34.
doi:10.1017/S0266466624000203

EFFICIENCY IN ESTIMATION UNDER MONOTONIC ATTRITION

JEAN-LOUIS BARNWELL
Analysis Group

SARASWATA CHAUDHURI
McGill University and CIREQ

Attrition is monotonic when agents leaving multi-period studies do not return. Under a general missing at random (MAR) assumption, we study efficiency in estimation of parameters defined by moment restrictions on the distributions of the counterfactuals that were unrealized due to monotonic attrition. We discuss novel issues related to overidentification, usability of sample units, and the information content of various MAR assumptions for estimation of such parameters. We propose a standard doubly robust estimator for these parameters by equating to zero the sample analog of their respective efficient influence functions. Our proposed estimator performs well and vastly outperforms other estimators in our simulation experiment and empirical illustration.

1. INTRODUCTION

Subjects/respondents often leave at various junctures of multi-period/phase studies/surveys. If they do not return, then that creates a monotonically missing dataset with respect to the original cohort of the study/survey. Monotonicity is reflected by the fact that the members of the original cohort that are observed in a later period are also observed in earlier periods. Equivalently, monotonicity is also reflected by the fact that the variables observed for those that left after an earlier period are also observed for those that left after later periods.

Attrition in the sample causes problems with statistical analysis. First, the sample's representativeness of the original population may be lost. Second, even if representativeness is restored by virtue of plausible assumptions such as missingness at random (selection on observables), the loss in data leads to imprecision in estimation; therefore, efficient estimation that optimally uses the remaining available information is of utmost importance.

We are very grateful to the Editor (Peter C. B. Phillips), the Co-Editor (Patrik Guggenberger), two anonymous referees, and Whitney Newey for their help in improving the paper. We also thank Francesco Amodio, Marine Carrasco, Daniel Farewell, Bryan Graham, Fabian Lange, Steven Lehrer, Thierry Magnac, Erica Moodie, Chris Muris, Tom Parker, Geert Ridder, Youngki Shin, and various conference and seminar participants for helpful comments. Earlier versions of the paper were circulated under different names; e.g., "A note on efficiency in estimation with monotonically missing at random data." The views presented in this work do not reflect those of Analysis Group. Analysis Group provided no financial support for this work. Address correspondence to Saraswata Chaudhuri, Department of Economics, McGill University and CIREQ, Montreal, QC, Canada; e-mail: saraswata.chaudhuri@mcgill.ca.

Our paper is about efficiency in estimation with monotonically missing at random (MAR) data. We build on the early work of Robins and Rotnitzky (1992), Robins, Rotnitzky, and Zhao (1995), Rotnitzky and Robins (1995), Fitzgerald, Gottschalk, and Moffitt (1996), Abowd, Crepon, and Kramarz (2001), Wooldridge (2002), Nicoletti (2006), Wooldridge (2010), etc. in the biostatistics and econometrics literature extending them to subpopulations defined by the monotone pattern of missingness. Such subpopulations are interesting because they reflect the attrition behavior of economic agents, e.g., agents left school or job or marriage after period one, after period two,..., never left.

To set the benchmark that any regular estimator should strive to reach, we obtain the efficiency bound for estimating parameters in general moment restrictions models. Our proposed estimator can reach this bound, and belongs in the class of two-step estimators satisfying double robustness with respect to the underlying nuisance parameters that can be estimated parametrically or nonparametrically. This class of estimators is well studied and known to be attractive in practice; see, e.g., Robins et al. (1994), Robins and Ritov (1997), Holcroft, Rotnitzky, and Robins (1997), Scharfstein, Rotnitzky, and Robins (1999), Bang and Robins (2005), Tsiatis (2006), Tan (2007), Cao, Tsiatis, and Davidian (2009), Rothe and Firpo (2019), and Chernozhukov et al. (2022).

Our results provide insights on the relation between the information content of the MAR assumption and the usability of the sample units toward estimation in subpopulations. The general (i.e., weakest) MAR assumption is that the hazard of leaving at any period does not depend on what would have happened afterward once we condition on "all" that has already happened. Under this general MAR assumption, we show that if interest lies on those that left at the end of, e.g., period four, then those that left before period four are not usable for estimation. By contrast, we show that if it is plausible to strengthen this general MAR assumption by restricting the "all" in its conditioning set as in, e.g., Chaudhuri (2020), then more (not all) sample units for whom the restricted "all" is observed and not just those that did not leave before period four become usable for estimation.

We also show that the efficiency bounds under the general MAR assumption coincide with those of particular augmented moment condition problems. A similar analysis in Chaudhuri (2020) was built on Graham (2011) that was based on an orthogonalization in Brown and Newey (1998). That cannot work for subpopulations in our setup because the key nuisance parameters—the conditional hazards of leaving—are unknown. (Known/unknown did not matter for Graham (2011) since he focused on the full population [see Hahn 1998].) Here, on the other hand, we need to use the orthogonalization in Newey (1994), Ackerberg et al. (2014), Chernozhukov et al. (2022), etc. for a unified treatment of full and subpopulations.

This orthogonalization implies that, in theory, the asymptotic variance of an inverse probability weighted (IPW) estimator based on nonparametrically estimated nuisance parameters will equal the inverse of the efficiency bound under the general MAR condition. However, our simulations suggest that this theory

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

could severely underestimate IPW's true variability (measured by Monte Carlo variance) even in very large samples when, unlike in Hirano, Imbens, and Ridder (2003), Chen, Hong, and Tarozzi (2008), Graham (2011), etc., we move beyond the single level of missingness. Our simulations also suggest that even the more conservative (in finite samples) formula for asymptotic variance in the spirit of Ackerberg, Chen, and Hahn (2012) can be a poor approximation underestimating IPW's true variability in small samples. Hence, IPW is not our recommended estimator. On the other hand, at least in our simulations, we do not see either of these two problems with our proposed estimator.

We also note that while this orthogonalization from Newey (1994), Ackerberg et al. (2014), Chernozhukov et al. (2022), etc. provides valid influence functions, it may not lead to semiparametric efficiency in general. Its claim to efficiency is solely based on a given moment function (e.g., IPW) involving unknown nuisance parameters that are nonparametrically exactly identified by a second set of moments, and on no additional information like the MAR assumption. In our setup, however, semiparametric efficiency is tied to the strength of the MAR assumption. While the general MAR assumption turns out to not contain any relevant information in this context, we show that when we strengthen that assumption then the said orthogonalization cannot reach the resulting efficiency bound. This suggests that while such orthogonalizations are obviously very useful, it is still important to consider all the available information to obtain the semiparametric efficiency bound that follows from it.

Finally, an important feature of our paper is that we obtain the efficiency results for parameters defined by overidentifying moment restrictions. This is not common in this literature; Chen et al. (2008) is among notable exceptions. To our understanding, the characterization of the tangent set in Chen et al. (2008) may be incomplete because overidentification is not explicitly used for that.[1] We show that the efficiency results in Chen et al. (2008) still hold. We also show that the efficiency results in Chaudhuri (2020) under (i) the general MAR with planned (known) conditional hazards or (ii) his convenient MAR can be extended to overidentifying moment restrictions. On the other hand, under our setup, it seems that a complete characterization of the tangent set hinders a seamless transition of the efficiency results for certain (not all) subpopulations between just- and overidentification. We provide a detailed treatment of this issue as it seems to be less appreciated (at least we did not know before an anonymous referee for Chaudhuri (2020) pointed it out).

Our paper proceeds as follows. Section 2 lays out the theoretical framework guided by an empirical motivation based on the attrition behavior of students from the widely studied, attrition-infested Project STAR experiment. Section 3 presents the core theory—efficiency bound, efficient influence function, overidentification, and the information content of the MAR assumption—by relating them to the

[1] We are very grateful to an anonymous referee for Chaudhuri (2020), and Patrik Guggenberger, and Whitney Newey for their help with this issue. Any error is of course only our responsibility.

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

literature. Section 4 presents the estimator and a sketch of its properties under parametric (mis)specification and nonparametric specification of the nuisance parameters. The asymptotic theory of such estimators is well studied and is certainly not our contribution; the sketch is presented only for completeness. Section 5 presents an elaborate empirical illustration of the benefits of the proposed estimator's precision in drawing substantive conclusions on the effect of small class size across dimensions induced by the attrition behavior of students from Project STAR. Section 6 concludes.

All the proofs are collected in Supplementary Appendix A. Complementing the theory in our paper, we present in Supplementary Appendix B a Monte Carlo experiment demonstrating excellent small-sample properties of our proposed estimator. The experiment also suggests that the promise of efficiency made by the theory for the competing IPW estimators based on nonparametric estimation of nuisance parameters may not realize even in very large samples.

2. EMPIRICAL MOTIVATION AND THE THEORETICAL FRAMEWORK

2.1. Empirical Motivation

Tennessee's Student/Teacher Achievement Ratio experiment, also known as Project STAR, has been extensively used to study the effect of small class size on future outcomes for the students; see, e.g., Hanushek (1999), Krueger (1999), Krueger and Whitmore (2001), Ding and Lehrer (2010), and Chetty et al. (2011). In Project STAR, students enrolling in grade K of 79 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). The literature on Project STAR typically does not differentiate between the latter two class types, and we will follow that here and refer to them jointly as "not-small" classes.

We use the well-known and publicly available Project STAR data (Achilles et al., 2008) containing characteristics of the schools, the teachers, demographic and socioeconomic characteristics of the students, and their normalized reading and math scores from grade K to grade 3 or to a lower grade until which they stayed with a Project STAR school.[2]

Many students—701 out of 1,493 (47%) from small classes and 1,725 out of 3,477 (49.6%) from not-small classes—entering Project STAR schools in grade K did not stay until the project ended, i.e., until the end of grade 3.[3] See Table 1.

[2]We work with normalized scores for the sake of interpretation. For example, the normalized reading score is the demeaned and standardized reading score of each student at each grade based on that grade's mean and standard deviation of reading scores of students across all participating Project STAR schools.

[3]In the original dataset, 917 out of 1,900 (48.3%) from small classes and 2,139 out of 4,425 (48.3%) from not-small classes entering Project STAR schools in grade K did not stay until the end of grade 3. For simplicity of the illustration, we construct our working sample by dropping from this original dataset students: (i) who did not enroll in Project STAR schools in grade K in 1985 but enrolled in grades 1–3 in the next 3 years, or (ii) who left Project STAR schools

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

TABLE 1. Number of students in our sample by their switching class type or leaving Project STAR dynamics at the end of each grade conditional on staying until the end of that grade in their initially assigned class. The switcher % inside the parentheses are with respect to the class-type-specific row total, e.g., $100 \times 79/(1,004 + 410 + 79) \approx 5.3$.

	Randomized to small class			Randomized to not-small class		
After grade	Stayed in small	Left STAR school	Switched to not-small	Stayed in not-small	Left STAR school	Switched to small
K	1,004	410	79 (5.3%)	2,230	1,047	200 (5.8%)
1	798	188	18 (1.8%)	1,674	481	75 (3.4%)
2	672	103	23 (2.9%)	1,392	197	85 (5.1%)

For simplicity of illustration, we further exclude from our sample the very small percentage of students who switched classes.[4]

This attrition makes the scores of a student in a grade unobserved/counterfactual if the student left before completing the grade. Consequently, many of the grade-specific average scores that researchers compare to estimate the effect of small classes are unavailable. To fix ideas, consider the reading scores reported in Table 2. Note that the grade-specific average reading scores in small or not-small classes are the weighted average of the elements of that grade's column in Table 2 with weights proportional to the corresponding number of students, e.g., for grade K in small class, it is $(-.19 \times 410 - .14 \times 188 - .09 \times 103 + .45 \times 672)/(410 + 188 + 103 + 672) \approx .14$. The grade-specific averages are unavailable (marked by "?") except in grade K because attrition starts after grade K.

Naively imputing these grade-specific averages by the "Never left" category would be extremely misleading for both small (.14 by .45) and not-small ($-.07$ by .20) classes in grade K. (We could compare since grade K scores are actually observed for all.) Therefore, naive imputation based on the Never left category would possibly be misleading as well for grades 1–3, where some sort of imputation is actually required. Interestingly, such imputations are less misleading when we compare the difference between the averages of grade K reading score in small and not-small classes: $(.45 - .20) - (.14 - (-.07)) = .25 - .21 = .04$—the effect of attrition largely cancels out, which can be seen as a type of "common trend" phenomenon.[5]

after grade K or 1 or 2 but came back in the subsequent years during the experiment, or (iii) with incidental missing (relevant) variables when the missingness is unrelated to attrition, or (iv) with invalid test scores (see, e.g., page 151 of Hanushek, 1999).

[4] Only 18 and 23 students switched from small class after grades 1 and 2, respectively. These numbers are too small for any analysis without extremely stringent restrictions on models for the switching behavior. We do not know enough to impose such stringent restrictions and hence exclude the switchers from our analysis.

[5] Similar observations have been made repeatedly in economics; see, e.g., the Special Issue: "Attrition in Longitudinal Surveys" in the *Journal of Human Resources* (1998) where one observes that big distortions of group means due to attrition often vanish in the results of regression, i.e., for difference in group means.

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

Table 2. Observed and unobserved normalized reading scores by attrition behavior of students from their initially assigned classes. If the full population is of interest, then the number of levels of missingness in any grade's score is the number of x's in that grade's column.

Left STAR	Randomized to small class					Randomized to not-small class				
school at the	Number of	Reading score in grade				Number of	Reading score in grade			
end of grade	students	K	1	2	3	students	K	1	2	3
K	410	−.19	x	x	x	1,047	−.36	x	x	x
1	188	−.14	−.19	x	x	481	−.27	−.54	x	x
2	103	−.09	−.14	−.23	x	197	−.00	−.22	−.31	x
3 (Never left)	672	.45	.50	.47	.44	1,392	.20	.33	.30	.22
Average score		.14	?	?	?		−.07	?	?	?

However, the investigation cannot end here as there are two outstanding questions. First, will the same phenomenon emerge from the scores in grades 1–3? Second, are any of those differences between small and not-small classes going to be statistically significant?

The first question is not answerable without assumptions on the mechanism of attrition because it involves comparing counterfactual means with the scores of the Never left category. We do not have anything original to say in this regard and will work under a very general MAR (selection on observables) assumption with a very flexible model specification for it.

On the other hand, our paper is about efficiency in estimation and is devised to address the second question. Under a general MAR assumption, we will estimate the counterfactual means with likely most precision and check if the concerned differences are statistically significant. (Section 5 will present strong evidence that such differences are significant.)

To efficiently estimate the grade-specific counterfactual mean, we will need to efficiently estimate the attrition-category-specific counterfactual means in each grade, i.e., the ones that are marked by "x" in Table 2. These are examples of what we mean by subpopulation-specific parameters where the subpopulations partition the full population by the attrition behavior of the students (population units). Such subpopulation-specific parameters are obviously important for many other purposes including as descriptive statistics, and we will make use of them in various ways in the empirical illustration in Section 5.

2.2. Theoretical 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. Let C be a random variable with support $\{1, \ldots, R\}$. Let $T_C(Z)$ be a transformation defined as $T_r(Z) := (Z_1', \ldots, Z_r')'$ for $r = 1, \ldots, R$. The notation is standard (see, e.g., Tsiatis, 2006). Z_j's may have common elements, e.g., time-invariant variables, and

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

empirical practice (coding, etc.) should ensure that they are not counted in the $T_r(Z)$'s more than once.

Let $O := (C, T_C'(Z))'$ denote what is observed for a unit in the sample.

Consider the Project STAR example. This is an $R = 4$ period study where grade K is period 1,..., and grade 3 is period 4. Z_r are the variables—characteristics of the schools, the teachers, demographic and socioeconomic characteristics of the students, and their normalized reading and math scores in period r—that are observed in period r. $T_r(Z)$ is the cumulative history of the Z_r variables (some of which may be time-invariant) observed until and including period r. If a unit (student) leaves after period $j \in \{1, \ldots, R\}$, then its $C = j$ and we only observe $T_j(Z)$ for it. $C = R$ is the same as never leaving (some denote this as $C = \infty$).

We maintain the general MAR (selection on observables) assumption that

$$P(C = r|T_R(Z), C \geq r) = P(C = r|T_r(Z), C \geq r) \text{ for } r = 1, \ldots, R-1. \quad \textbf{(1)}$$

Since $T_r(Z)$ is observable when $C \geq r$, (1) imposes that the conditional hazard $P(C = r|T_R(Z), C \geq r)$ at period r does not depend on the unobservables $Z_{r+1}, \ldots, Z_R$ once conditioned on the observables $T_r(Z)$. (1) is the MAR assumption in the sense of Rubin (1976).

Plausibility of MAR depends on the context. MAR has been widely used in studies on attrition especially if, as in our paper, the missingness is monotone.[6] Ding and Lehrer (2010) (and, less explicitly, Krueger (1999)) assumed MAR for attrition in the Project STAR data.

Generalizing the nomenclature introduced in Section 2.1, we refer to the underlying population of $O := (C, T_C'(Z))'$ as the full population. We refer to the partition of this full population by the values taken by C as subpopulations; e.g., subpopulation r is the underlying population from which units with $C = r$ can be viewed as randomly drawn. There are R unitary subpopulations indexed by $r = 1, \ldots, R$. Unions of unitary subpopulations form a composite subpopulation, e.g., $C \in \{1, 2\}$, or the full population $C \in \{1, \ldots, R\}$.

Under the general MAR condition in (1), the unconditional distribution of Z may not be the same as the distribution of Z conditional on $C = r$ for $r = 1, \ldots, R$, i.e., the subpopulations are possibly heterogeneous. In the example from Section 2.1 where the subpopulations are defined by the attrition categories based on the timing of attrition, this means that the distribution of the (potential) grade 3 reading scores may not be the same for those who left after grade K and those who left after grade 2 and those who never left.

We will work with a generic target subpopulation $C \in \{a, \ldots, b\}$, denoted for brevity by $a \leq C \leq b$ or $[a, b]$, for $a \leq b$ and $a, b \in \{1, \ldots, R\}$. If $a = b = r$, then this is the underlying unitary subpopulation from which the units who left at the end

[6] If the missingness is non-monotone, then MAR or selection on observables is unrealistic since the choice to return could depend on unobservables, i.e., on what happened when the individual was out of the study; see, e.g., Gill, van der Laan, and Robins (1997), Robins and Gill (1997), and Vansteelandt, Rotnitzky, and Robins (2007). That would be a case of selection on unobservables. Hoonhout and Ridder (2019) compare various selection on unobservables conditions with MAR in a multi-period context. We do not contribute to that literature.

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

of period r can be viewed as randomly drawn. If $a < b$, then this is the composite subpopulation for the units who left in the periods $a, a+1, \ldots, b$. If $a = 1$ and $b = R$, then this is the full population.

Denote the distribution of Z in the target population by $F_{Z|(a\leq C\leq b)}(z)$. This is the weighted average of the distributions of Z in subpopulations $a, \ldots, b$ with weights $P(C = j)/P(a \leq C \leq b)$ for $j = a, \ldots, b$. 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 $m(Z;\beta) : \text{Support}(Z) \times \mathcal{B} \mapsto \mathbb{R}^{d_m}$, $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$, and $d_\beta \leq d_m$. Then, for a given $a, b \in \{1, \ldots, R\}$ with $a \leq b$, define the parameter value of interest $\beta^0_{[a,b]}$ by an overidentifying system of moment restrictions as

$$E[m(Z;\beta) \mid a \leq C \leq b] = 0 \text{ for } \beta \in \mathcal{B} \text{ if and only if } \beta = \beta^0_{[a,b]}. \tag{2}$$

$m(Z;\beta)$ can depend on any element of Z; e.g., reading score in grade K or 1 or 2 or 3. If the least frequently observed element of Z that is involved in $m(Z;\beta)$ belongs in Z_k for some $k = 1, \ldots, R$, then exactly the same analysis in the sequel will still apply but with a different observability indicator $\bar{C}$ instead of C where $\bar{C} := k$ if $C \geq k$ and $\bar{C} := C$ otherwise.

We will also maintain the following assumptions that are standard in this literature.

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) $P(C = R|T_R(Z))$ is bounded away from zero almost surely $T_R(Z)$.

(A3) $M_{[a,b]}$ is a $d_m \times d_\beta$ finite matrix of full column rank where $M_{[a,b]} := M_{[a,b]}(\beta^0_{[a,b]})$ and $M_{[a,b]}(\bar{\beta}) := \left\{ \frac{\partial}{\partial \beta'} E[m(Z;\beta)|a \leq C \leq b] \right\}_{\beta=\bar{\beta}}$ at any $\bar{\beta} \in \mathcal{B}$ where it exists.

Remark. (A1) rules out dependence and heterogeneity across sample units when viewed as random draws from O. (A2) imposes the bounded away from zero condition instead of only $P(C = R|T_R(Z)) > 0$ to avoid the "limited overlap" problem (see, e.g., Khan and Tamer, 2010). (A3) gives local identification of $\beta^0_{[a,b]}$; it allows for non-smooth $m(Z;\beta)$ but requires the expectation to be differentiable with respect to β (see, e.g., Chen et al., 2008).

3. THE EFFICIENCY RESULTS

3.1. Efficiency Bound and Efficient Influence Function

Writing $T_r(Z)$ as T_r, for $r = 1, \ldots, R$, let us first introduce the key quantities for this section. Define

$$\varphi_{[a,b]}(O;\beta) := \sum_{j=a}^{b} \frac{P(C = j)}{P(a \leq C \leq b)} \varphi_{[j,j]}(O;\beta) \tag{3}$$

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

for the subpopulation $[a,b]$ as the weighted average of the unitary subpopulation quantities

$$\varphi_{[j,j]}(O;\beta) := \sum_{r=j+1}^{R} I(C \geq r)\omega_{r,j}(T_{r-1})\left(E[m(Z;\beta)|T_r] - E[m(Z;\beta)|T_{r-1}]\right) + \frac{I(C=j)}{P(C=j)}E[m(Z;\beta)|T_j] \tag{4}$$

that are feasible for each $\beta \in \mathcal{B}$ based on the observed data because of the equality in (5):

$$\omega_{r,j}(T_{r-1}) := \frac{P(C=j|T_j)}{P(C=j)P(C \geq r|T_{r-1})} = \frac{P(C=j|T_j, C \geq j)}{P(C=j)\prod_{k=j}^{r-1}[1-P(C=k|T_k, C \geq k)]}. \tag{5}$$

Under regularity conditions, the weighted average representation of $\varphi_{[a,b]}(O;\beta)$ implies

$$\begin{aligned}
\frac{\partial}{\partial\beta'}E\left[\varphi_{[a,b]}(O;\beta)\right] &= \sum_{j=a}^{b}\frac{P(C=j)}{P(a \leq C \leq b)}\frac{\partial}{\partial\beta'}E\left[\varphi_{[j,j]}(O;\beta)\right] \\
&= \sum_{j=a}^{b}\frac{P(C=j)}{P(a \leq C \leq b)}\frac{\partial}{\partial\beta'}E\left[m(Z;\beta)|C=j\right] \\
&= \frac{\partial}{\partial\beta'}E\left[m(Z;\beta)|a \leq C \leq b\right] \text{ and} \\
Var\left(\varphi_{[a,b]}(O;\beta)\right) &= \sum_{j=a}^{b}\sum_{k=a}^{b}\frac{P(C=j)P(C=k)}{P^2(a \leq c \leq b)}Cov\left(\varphi_{[j,j]}(O;\beta), \varphi_{[k,k]}(O;\beta)\right).
\end{aligned}$$

The covariance ($j \neq k$) terms in the composite (sub)populations simplify when $a=1, b=R$.

LEMMA 1. *In the case of the full population $a=1, b=R$, the above representation gives*

$$\begin{aligned}
\varphi_{[1,R]}(O;\beta) &= \sum_{r=2}^{R}\frac{I(C \geq r)}{P(C \geq r|T_{r-1})}\left(E[m(Z;\beta)|T_r] - E[m(Z;\beta)|T_{r-1}]\right) \\
&\quad + E[m(Z;\beta)|T_1], \\
Var\left(\varphi_{[1,R]}(O;\beta)\right) &= \sum_{r=2}^{R}E\left[\frac{Var\left(E[m(Z;\beta)|T_r]|T_{r-1}\right)}{P(C \geq r|T_{r-1})}\right] + Var\left(E[m(Z;\beta)|T_1]\right).
\end{aligned}$$

Equipped with these key quantities, we will now present the main result of our paper.

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

PROPOSITION 2. *Let the MAR condition in (1), the moment restrictions in (2), and Assumption A hold. Let* $V_{[a,b]} := Var(\varphi_{[a,b]}(O;\beta^0_{[a,b]}))$ *be a finite and positive definite matrix.*[7] *Then the semiparametric efficiency bound for* $\beta^0_{[a,b]}$ *is given by* $\Omega_{[a,b]} := M'_{[a,b]}V^{-1}_{[a,b]}M_{[a,b]}$*:*

(i) *when* $a = 1, b = R$ *(full population) or* $a = b$ *(unitary subpopulations);*

(ii) *when* $a, b \in \{1,\dots,R\}$ *with* $a \leq b$*, if additionally* $\beta^0_{[a,b]}$ *is just-identified, i.e.,* $d_m = d_\beta$.

A regular estimator $\widehat{\beta}_{[a,b]}$ *whose asymptotic variance equals* $\Omega^{-1}_{[a,b]}$ *has the asymptotically linear representation (with obvious cancellations giving* $\Omega^{-1}_{[a,b]}M'_{[a,b]}V^{-1}_{[a,b]} = M^{-1}_{[a,b]}$ *when* $d_m = d_\beta$*):*

$$\sqrt{n}(\widehat{\beta}_{[a,b]} - \beta^0_{[a,b]}) = -\Omega^{-1}_{[a,b]}M'_{[a,b]}V^{-1}_{[a,b]}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{[a,b]}(O_i;\beta^0_{[a,b]}) + o_p(1).$$

Remarks. First, Proposition 2 covers the well-known special cases found in the literature. $R = 2$ with $a = b = 1$ or $a = 1, b = 2$ covers Theorem 1 of Chen et al. (2008) (see also Robins et al., 1994). $a = 1, b = R > 2$ gives the full-population result like Robins and Rotnitzky (1992), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), and Holcroft, Rotnitzky, and Robins (1997).

Second, few papers in this literature allow for $d_m > d_\beta$, i.e., overidentifying restrictions for $\beta^0_{[a,b]}$. Chen et al. (2008) is among notable exceptions. However, it is possible that the characterization of the tangent set there (and similar papers) may be incomplete because overidentification is not explicitly used for that. Proposition 2(i) shows that Chen et al.'s (2008) results (Chen et al. (2008) worked with $R = 2$ with $a = b = 1$ or $a = 1, b = 2$) still hold. Additionally, in Section 3.2, we also extend the main efficiency results in Chaudhuri (2020) (also a generalization of Chen et al. (2008)) to the case of overidentifying restrictions.

Third, overidentification is not innocuous in our general framework. Under just identification, the efficiency bound in Proposition 2 applies to any $a, b \in \{1,\dots,R\}$ with $a \leq b$. However, in the case of overidentification, the efficiency bound result is for the full population ($a = 1, b = R$) and all R unitary subpopulations ($a = b$) but not for generic composite subpopulations $[a,b]$'s. Unlike in Chaudhuri (2020), here the overidentifying restrictions for $\beta^0_{[a,b]}$ impose restrictions on the tangent set that do not seem to be satisfied for generic $[a,b]$'s by the influence function

[7] While it is easier to think of primitive conditions for positive definiteness of $Var\left(\varphi_{[a,b]}(O;\beta)\right)$ when $a = b$ or $a = 1, b = R$, we maintain positive definiteness of $Var\left(\varphi_{[a,b]}(O;\beta^0_{[a,b]})\right)$ generally. Writing $\varphi_{[s,t]}(O;\beta)$ as $\varphi_{[s,t]}$ for $s,t = 1,\dots,R$ and $m(Z;\beta)$ as m for brevity, the components of $Var\left(\varphi_{[a,b]}\right)$ can be expressed as follows. For $j = a,\dots,b$ and $k = a,\dots,j-1$: $Var\left(\varphi_{[j,j]}\right) = \sum_{r=j+1}^{R} E\left[\Delta_{r,j}|C=j\right] + Var\left(\frac{I(C=j)}{P(C=j)}E[m|T_j]\right)$ and $Cov\left(\varphi_{[j,j]},\varphi_{[k,k]}\right) = E\left[\sum_{r=j+1}^{R}\Delta_{r,j} + \sum_{r=k+1}^{j}\nabla_{r,j,k}\middle|C=j\right] + Cov\left(\frac{I(C=j)}{P(C=j)}E[m|T_j], \frac{I(C=k)}{P(C=k)}E[m|T_k]\right)$ where, again for simplicity, we have used the notation $\Delta_{r,j} := \omega_{r,j}(T_{r-1})Var\left(E[m|T_r]|T_{r-1}\right)$ for $r = j+1,\dots,R$, and $\nabla_{r,j,k} := \omega_{r,k}(T_{r-1})E[m|T_j]\left(E[m|T_r] - E[m|T_{r-1}]\right)'$ for $r = k+1,\dots,j$. If, e.g., $a = b = j$ then primitive conditions for the positive definiteness of $Var\left(\varphi_{[j,j]}\right)$ can be guided by its expression above.

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

presented in the proposition. Since it seems less appreciated, we utilize Section 3.2 to be explicit about the restrictions imposed by overidentification.

Fourth, the weighted average representation of $\varphi_{[a,b]}(O_i;\beta^0_{[a,b]})$ in (3), that follows from the representation $E[m(Z;\beta)|a \le C \le b] = \sum_{j=a}^{b} \frac{P(C=j)}{P(a\le C\le b)} E[m(Z;\beta)|C=j]$, presents an easy way of combining the efficient estimators for the unitary subpopulations to obtain the efficient estimator for the composite subpopulation $[a,b]$ under just identification $d_m = d_\beta$:

$$\sqrt{n}\left(\widehat{\beta}_{[a,b]} - \sum_{j=a}^{b}\left[M^{-1}_{[a,b]}\frac{P(C=j)}{P(a\le C\le b)}M_{[j,j]}(\beta^0_{[a,b]})\right]\widehat{\beta}_{[j,j]}\right) = o_p(1),$$

where the weights for the $\widehat{\beta}_{[j,j]}$'s add up to the identity matrix since $M_{[a,b]} = \sum_{j=a}^{b}\frac{P(C=j)}{P(a\le C\le b)}M_{[j,j]}(\beta^0_{[a,b]})$.[8] Dardanoni, Modica, and Peracchi (2011), Abrevaya and Donald (2017), Muris (2020), and others also considered combining estimators or moment restrictions in similar contexts with missing data.

Fifth, each $\varphi_{[j,j]}(O;\beta^0_{[a,b]})$ is doubly robust to the misspecification of the two sets of unknown nuisance parameters: the conditional hazards $P(C=r|T_r, C\ge r)$ and the conditional expectations $E[m(Z;\beta)|T_r]$ for the various r's. Therefore, the representation of $\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ in (3) implies that $\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ also satisfies such double robustness. $\varphi_{[j,j]}(O;\beta^0_{[a,b]})$ is robust to the misspecification of the $P(C=r|T_r, C\ge r)$'s under (1) since if we take expectation after replacing each $P(C=r|T_r, C\ge r)$ in (4) (precisely, (5)) by arbitrary scalar functions of T_r, we still obtain $E[m(Z;\beta)|C=j]$ if the expectation exists. To see that $\varphi_{[j,j]}(O;\beta^0_{[a,b]})$ is also robust to the misspecification of the $E[m(Z;\beta)|T_r]$'s, rewrite (4) as

$$\begin{aligned}\varphi_{[j,j]}(O;\beta) :=\ & I(C=R)\omega_{R,j}(T_{R-1})m(Z;\beta)\\ &+\sum_{r=j+1}^{R-1}\left[\frac{I(C\ge r)}{P(C\ge r|T_{r-1})}-\frac{I(C\ge r+1)}{P(C\ge r+1|T_r)}\right]\frac{P(C=j|T_j)}{P(C=j)}E[m(Z;\beta)|T_r]\\ &+\left[\frac{I(C=j)}{P(C=j)}-\frac{I(C\ge j+1)}{P(C\ge j+1|T_j)}\frac{P(C=j|T_j)}{P(C=j)}\right]E[m(Z;\beta)|T_j],\end{aligned} \tag{6}$$

replace each $E[m(Z;\beta)|T_r]$ in (6) by arbitrary d_m-dimensional functions of T_r, take expectation while noting that $P(C\ge r|T_r) = P(C\ge r|T_{r-1})$ (see Lemma 9 in Supplementary Appendix A.1), and finally see that (1) gives the expectation as

[8] To see this result, write the weights $M^{-1}_{[a,b]}\frac{P(C=j)}{P(a\le C\le b)}M_{[j,j]}(\beta^0_{[a,b]})$ as A_j for brevity and note that

$$\begin{aligned}\sqrt{n}(\widehat{\beta}_{[a,b]}-\beta^0_{[a,b]}) &= -M^{-1}_{[a,b]}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{[a,b]}(O_i;\beta^0_{[a,b]})+o_p(1) = -M^{-1}_{[a,b]}\sum_{j=a}^{b}\frac{P(C=j)}{P(a\le C\le b)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{[j,j]}(O_i;\beta^0_{[a,b]})+o_p(1)\\ &= \sum_{j=a}^{b}A_j\left[-M^{-1}_{[j,j]}(\beta^0_{[a,b]})\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_{[j,j]}(O_i;\beta^0_{[a,b]})+o_p(1) = \sum_{j=a}^{b}A_j\sqrt{n}\left(\widehat{\beta}_{[j,j]}-\beta^0_{[a,b]}\right)+o_p(1).\end{aligned}$$

The result follows since $\beta^0_{[a,b]}$ on both sides cancels out as $\sum_{j=a}^{b}A_j = I_{d_m} = I_{d_\beta}$ implies $\beta^0_{[a,b]} = \sum_{j=a}^{b}A_j\beta^0_{[a,b]}$.

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

$E[I(C=R)\omega_{R,j}(T_{R-1})m(Z;\beta)]=E[m(Z;\beta)|C=j]$ (see Lemma 5) if the expectation exists. This is double robustness with respect to misspecification of nuisance parameters; see Robins et al. (1994), Robins and Ritov (1997), Scharfstein et al. (1999), Bang and Robins (2005), Tan (2007), Cao et al. (2009), Rothe and Firpo (2019), Chernozhukov et al. (2022), etc. We use this double robustness to motivate the estimating function for $\beta^0_{[a,b]}$ in Section 4 based on $\varphi_{[a,b]}(O;\beta^0_{[a,b]})$.

Sixth, the expression for the $\varphi_{[j,j]}(O;\beta)$'s in (4) or (6) tells us that if $a\geq 2$ then the units with $C<a$ do not contribute to the estimation of the target $\beta^0_{[a,b]}$.[9] We note that this is an artifact of the general MAR condition in (1). Units with $C<a$ can contribute to the efficient estimation of $\beta^0_{[a,b]}$ if it is plausible to strengthen the MAR condition. A concrete example can be found in Proposition 4 below adopted for extension from Chaudhuri (2020). (This example of strengthened MAR is revisited in Section 3.3 to caution against suboptimal use of sample units in case of overidentification of the nuisance conditional hazards.) In extreme contrast, Proposition 3 below adopted for extension from Chaudhuri (2020) shows that all sample units are usable for all target $\beta^0_{[a,b]}$'s if the conditional hazards are actually known.

3.2. Overidentification of $\beta^0_{[a,b]}$: Restriction on the Tangent Set

Let f and F denote the density and distribution functions, with the concerned random variables specified inside parentheses. Their conditional counterparts are denoted similarly. Let $L^2_0(F)$ denote the space of mean-zero, square integrable functions with respect to F.

We will first characterize the tangent set for all regular parametric submodels satisfying the semiparametric assumptions on the observed data $O=(C',T'_C(Z))'$. (We will then impose on it the restrictions due to overidentification.) Consider a regular parametric submodel indexed by a parameter η for the distribution of O. The log of this distribution is

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

in terms of $(C,Z')'$. The score function with respect to η is, in terms of $(C,Z')'$,

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

[9] Thus, generalizing the caption of Table 2, if Z_k is the least frequently observed element of Z that is involved in $m(Z;\beta)$, then the effective level of missingness is $\max\{0,k-a\}$ under the MAR condition in (1).

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

where $s_\eta(Z_1) := \frac{\partial}{\partial \eta} \log f_\eta(Z_1)$, $s_\eta(Z_r|Z_1,\ldots,Z_{r-1}) := \frac{\partial}{\partial \eta} \log f_\eta(Z_r|Z_1,\ldots,Z_{r-1})$, for $r = 2,\ldots,R$, and $\dot{P}_\eta(C = r|Z_1,\ldots,Z_r) := \frac{\partial}{\partial \eta} P_\eta(C = r|Z_1,\ldots,Z_r)$, for $r = 1,\ldots,R$. The tangent set $\mathcal{T}$ is the mean square closure of all d_β-dimensional linear combinations of $S_\eta(O)$ (see Newey (1990, pp. 105–106)) and can be expressed as

$$\mathcal{T} := \nu_1(Z_1) + \sum_{r=2}^{R} I(C \geq r)\nu_r(Z_1,\ldots,Z_r) + \sum_{r=1}^{R} I(C = r)\omega_r(Z_1,\ldots,Z_r), \tag{7}$$

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 $r = 1,\ldots,R$.

This is typically how the tangent set is characterized in this literature (e.g., Chen et al., 2008), but it does not take into account the additional restrictions imposed by overidentification of $\beta^0_{[a,b]}$. Apart from the incompleteness in the proofs due to such omissions, it does seem that the additional restrictions will matter in our general setup with generic subpopulations $[a,b]$. Hence, we will provide the details behind these additional restrictions.

For simplicity, we will drop the subscript η from all quantities (e.g., in (8) below) evaluated at η^0 where η^0 is the "true" submodel η, i.e., $f_{\eta^0}(O)$ is the actual distribution of the observed data. Note that the moment restrictions in (2) give the following identity in η for given a,b:

$$0 \equiv E_\eta[m(Z;\beta^0_{[a,b]})|a \leq C \leq b] \equiv E_\eta\left[\frac{P_\eta(a \leq C \leq b|Z)}{P_\eta(a \leq C \leq b)} m(Z;\beta^0_{[a,b]})\right].$$

Differentiate it with respect to η under the integral at $\eta = \eta^0$, and use (1) and (2) to get

$$0 = M_{[a,b]} \frac{\partial \beta^0_{[a,b]}(\eta_0)}{\partial \eta'} + E\left[m(Z;\beta^0_{[a,b]}) \left\{ s(Z) + \frac{\dot{P}(a \leq C \leq b|T_b)}{P(a \leq C \leq b|T_b)} \right\}' \middle| a \leq C \leq b \right], \tag{8}$$

where $s(Z) := s(Z_1) + \sum_{r=2}^{R} s(Z_r|T_{r-1})$ (with abuse, we briefly revert to the T_r notation for brevity). Now, we note that (2) also gives the following identity in η for given a,b:

$$0 \equiv AE_\eta[m(Z;\beta^0_{[a,b]})|a \leq C \leq b] \equiv AE_\eta\left[\frac{P_\eta(a \leq C \leq b|Z)}{P_\eta(a \leq C \leq b)} m(Z;\beta^0_{[a,b]})\right]$$

for any A that is a full row rank $d_\beta \times d_m$ matrix such that $AM_{[a,b]}$ is nonsingular. Such an A always exists under our assumptions; e.g., $A = M'_{[a,b]} V^{-1}_{[a,b]}$. Therefore, following the same steps as in (8) and then solving for $\frac{\partial \beta^0_{[a,b]}(\eta_0)}{\partial \eta'}$, we obtain that

$$\frac{\partial \beta^0_{[a,b]}(\eta_0)}{\partial \eta'} = -\left(AM_{[a,b]}\right)^{-1} AE\left[m(Z;\beta^0_{[a,b]}) \left\{ s(Z) + \frac{\dot{P}(a \leq C \leq b|T_b)}{P(a \leq C \leq b|T_b)} \right\}' \middle| a \leq C \leq b \right],$$

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

which when substituted for in (8) gives (noting that $s(Z) := s(Z_1) + \sum_{r=2}^{R} s(Z_r|T_{r-1})$):

$$0 = \left(I_{d_\beta} - M_{[a,b]}\left(AM_{[a,b]}\right)^{-1}A\right) \times E\left[m(Z;\beta^0_{[a,b]})\left\{s(Z_1) + \sum_{r=2}^{R} s(Z_r|T_{r-1}) + \frac{\dot{P}(a \le C \le b|T_b)}{P(a \le C \le b|T_b)}\right\}' \middle| a \le C \le b\right]. \tag{9}$$

Note that (9) is trivially true under just identification $d_m = d_\beta$ since then $M_{[a,b]}\left(AM_{[a,b]}\right)^{-1}A = I_{d_\beta}$ by the definition of inverse. However, under overidentification, (9) imposes restrictions on the quantities inside the expectations that must be reflected by the tangent set. Therefore, a complete characterization of the tangent set $\mathcal{T}$ in the case of overidentification would augment what is defined in (7) such that its components additionally satisfy (10) if $[a,b] = [1,R]$ and satisfy (11) if $[a,b] \neq [1,R]$. Letting $B_{[a,b]} := \left(I_{d_\beta} - M_{[a,b]}\left(AM_{[a,b]}\right)^{-1}A\right)$,

- if the moment restrictions in (2) hold for $[a,b] = [1,R]$, then

$$0 = B_{[1,R]}E\left[m(Z;\beta^0_{[1,R]})\sum_{r=1}^{R}\nu_r(Z_1,\ldots,Z_r)'\right] \tag{10}$$

  as $\dot{P}_\eta(1 \le C \le R|Z) = 0$ in (9) since obviously $P_\eta(1 \le C \le R|Z) \equiv 1$ for all η;[10]
- if the moment restrictions in (2) hold for $[a,b] \neq [1,R]$, then

$$0 = B_{[a,b]}E\left[m(Z;\beta^0_{[a,b]})\left\{\sum_{r=1}^{R}\nu_r(Z_1,\ldots,Z_r) + \sum_{r=a}^{b}\frac{P(C=r|Z_1,\ldots,Z_r)}{P(a \le C \le b|Z_1,\ldots,Z_b)}\omega_r(Z_1,\ldots,Z_r)\right\}' \middle| a \le C \le b\right]. \tag{11}$$

Hence, $-\Omega^{-1}_{[a,b]}M'_{[a,b]}V^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ has to satisfy the restriction (10) or (11), as appropriate, to belong in $\mathcal{T}$ that is necessary for it to be the efficient influence function. Generalizing the literature, Proposition 2(i) showed that $-\Omega^{-1}_{[a,b]}M'_{[a,b]}V^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ satisfies the restriction when focus lies on the full population, i.e., $[a,b] = [1,R]$, or on the unitary subpopulations, i.e., $a = b$. Curiously, however, $-\Omega^{-1}_{[a,b]}M'_{[a,b]}V^{-1}_{[a,b]}\varphi_{[a,b]}(O;\beta^0_{[a,b]})$ does not seem to satisfy the restriction when $a \neq b$ but $[a,b] \neq [1,R]$, i.e., for composite subpopulations that are not the full population, and hence it is not efficient in that case although it remains

[10] We have not imposed enough structure on the $\omega_r(Z_1,\ldots,Z_r)$'s to write (10) as a special case of (11). Other than here—restriction on tangent set due to overidentification (that to the best of our knowledge has not been covered in the MAR literature)—we presented full and subpopulation analysis under the same framework instead of treating them separately as in, e.g., Hahn (1998), Hirano et al. (2003), and Chen et al. (2008).

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

a valid influence function since it satisfies the so-called "pathwise derivative" condition.

For completeness, we note that a similar characterization of the tangent set allows us to extend the main efficiency results in Chaudhuri (2020) to the case of overidentification. Those results work with strengthened MAR conditions but apply to remarkably more general target (sub)populations λ. Precisely, Propositions 3 and 4 will concern a β_λ^0 defined by the following moment restrictions: For any λ that is a subset of $\{1,\ldots,R\}$ including the full set, let

$$E[m(Z;\beta) \mid C \in \lambda] = 0 \text{ for } \beta \in \mathcal{B} \text{ if and only if } \beta = \beta_\lambda^0. \tag{12}$$

PROPOSITION 3. *Let the MAR condition in (1) and the moment restrictions in (12) hold. Let Assumption A hold with $M_{[a,b]}$ in A3 replaced by $M_\lambda := E[\partial m(Z;\beta_\lambda^0)/\partial\beta' \mid C \in \lambda]$. Let $\bar{V}_\lambda := Var(\bar{\varphi}_\lambda(O;\beta_{[a,b]}^0))$ be a finite and positive definite matrix where*

$$\bar{\varphi}_\lambda(O;\beta_\lambda^0) := \bar{\varphi}_{1,\lambda} + \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_{r-1})}\left(\bar{\varphi}_{r,\lambda} - \bar{\varphi}_{r-1,\lambda}\right) \text{ with}$$

$$\bar{\varphi}_{r,\lambda} := E\left[\frac{P(C \in \lambda|T_r)}{P(C \in \lambda)} m(Z;\beta_\lambda^0)\middle| T_r\right],$$

for $r = 1,\ldots,R$. If we additionally assume that $P(C = r|T_r, C \geq r)$ is known for $r = 1,\ldots,R-1$, i.e., the incompleteness is planned, then the semiparametric efficiency bound for β_λ^0 is given by $\bar{\Omega}_\lambda := M_\lambda' \bar{V}_\lambda^{-1} M_\lambda$ and the efficient influence function is $-\bar{\Omega}_\lambda^{-1} M_\lambda' \bar{V}_\lambda^{-1} \bar{\varphi}_\lambda(O;\beta_{[a,b]}^0)$.

The planned monotonic incompleteness condition was motivated in Chaudhuri (2020) as a cost cutting measure in survey designs. Another condition considered in Chaudhuri (2020) is a strengthened version of MAR, referred to as convenient MAR (CMAR), whereby

$$P(C = r|Z, C \geq r) = P(C = r|T_1, C \geq r) \text{ for } r = 1,\ldots,R. \tag{13}$$

PROPOSITION 4. *Let the moment restrictions in (12) and the CMAR condition in (13) hold. Let Assumption A hold with $M_{[a,b]}$ in A3 replaced by $M_\lambda := E[\partial m(Z;\beta_\lambda^0)/\partial\beta' \mid C \in \lambda]$. Let $V_\lambda^{CMAR} := Var(\varphi_\lambda^{CMAR}(O;\beta_\lambda^0))$ be a finite and positive definite matrix where*

$$\varphi_\lambda^{CMAR}(O;\beta_\lambda^0) := \frac{I(C \in \lambda)}{P(C \in \lambda)} E[m(Z;\beta_\lambda^0)|T_1] + \sum_{r=2}^{R} \frac{I(C \geq r)}{P(C \geq r|T_1)} \frac{P(C \in \lambda|T_1)}{P(C \in \lambda)} \left(E[m(Z;\beta_\lambda^0)|T_r] - E[m(Z;\beta_\lambda^0)|T_{r-1}]\right).$$

Then the semiparametric efficiency bound for β_λ^0 is given by $\Omega_\lambda^{CMAR} := M_\lambda'[V_\lambda^{CMAR}]^{-1} M_\lambda$ and the efficient influence function is $-[\Omega_\lambda^{CMAR}]^{-1} M_\lambda'[V_\lambda^{CMAR}]^{-1} \varphi_\lambda^{CMAR}(O;\beta_{[a,b]}^0)$.

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

3.3. IPW, Variance Adjustment, and Information Content of MAR

Returning to the general MAR condition in (1), it is clear from (6) that $\varphi_{[j,j]}(O;\beta)$ is an augmented inverse probability weighted (AIPW) moment vector where the first term on the right-hand side (RHS) of (6) is the IPW term, while the other terms on the RHS are the augmentation. Therefore, the weighted average representation in (3) implies that $\varphi_{[a,b]}(O;\beta)$ is also another AIPW moment vector, but concerning a different set of moments.

Lemma 5 summarizes in the current context the idea behind the Narain (1951)–Horvitz and Thompson (1952)–Hajek (1971) IPW principle under the general MAR condition in (1). For each $\beta \in \mathcal{B}$, this IPW principle enables identification of $E[m(Z;\beta)|a \leq C \leq b]$ whose sample version is infeasible, based on a quantity whose sample version is feasible.

LEMMA 5. *If $P(C = R|T_R) > 0$ almost surely T_R, then the general MAR condition in (1) implies that $E[m(Z;\beta)|a \leq C \leq b] = E\left[I(C = R)\omega^{IPW}_{[a,b]}m(Z;\beta)\right]$ for each $\beta \in \mathcal{B}$ where*

$$\omega^{IPW}_{[a,b]} := \frac{\sum_{j=a}^{b} P(C = j|T_j, C \geq j)\prod_{r=1}^{j-1}[1 - P(C = r|T_r, C \geq r)]}{\prod_{r=1}^{R-1}[1 - P(C = r|T_r, C \geq r)]\,P(a \leq C \leq b)}$$
$$= \sum_{j=a}^{b} \frac{P(C = j)}{P(a \leq C \leq b)}\omega_{R,j}(T_{R-1})$$

and where $\omega_{R,j}(T_{R-1})$ is defined in (5), and indeed $\omega_{R,j}(T_{R-1}) = \omega^{IPW}_{[j,j]}$ for $j = 1,\ldots,R$.

For brevity, we used the convention that if $a = 1$ then $\prod_{r=1}^{a-1}(1 - P(C = r|T_r, C \geq r)) = 1$.

Lemma 5 gives the foundation for IPW estimation based on an estimator of $E[I(C = R)\omega^{\text{IPW}}_{[a,b]}m(Z;\beta)]$, namely,

$$\frac{1}{n}\sum_{i=1}^{n} I(C_i = R)\widehat{\omega}^{\text{IPW}}_{[a,b]}m(Z_i;\beta) \tag{14}$$

as the GMM sample moment vector, where $\widehat{\omega}^{\text{IPW}}_{[a,b]}$ is an estimator of $\omega^{\text{IPW}}_{[a,b]}$ obtained by replacing each conditional hazard by its parametric or nonparametric estimator. In this section, our discussion of variance adjustment and efficiency in the context of the information content of the general MAR condition in (1) will correspond to nonparametric estimation of $\omega^{\text{IPW}}_{[a,b]}$.

PROPOSITION 6. (i) *The "limited information" efficient GMM estimator of $\beta^0_{[a,b]}$ based on the moment restrictions*

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

$$E\left[I(C=R)\omega_{[a,b]}^{IPW}m(Z;\beta_{[a,b]}^{0})\right]=0, \tag{15}$$

where for each $r=a\ldots,R-1$ *the* $P(C=r|T_r,C\geq r)$*'s in* $\omega_{[a,b]}^{IPW}$ *solve for the* $p_r(T_r)$*'s from*

$$E[I(C\geq r)\{I(C=r)-p_r(T_r)\}\mid T_r]=0 \text{ almost surely } T_r, \tag{16}$$

has asymptotic variance equal to the inverse of the semiparametric information bound for $\beta_{[a,b]}^{0}$ *under the "full information" contained jointly in the restrictions (15) and (16).*

(ii) *Furthermore, this asymptotic variance from (i) is equal to* $\Omega_{[a,b]}^{-1}$ *where* $\Omega_{[a,b]}:=M_{[a,b]}'V_{[a,b]}^{-1}M_{[a,b]}$ *is defined in the statement of Proposition 2.*

Proposition 6(i) applies Theorem 1 of Ackerberg et al. (2014) to show that the "limited information" and "full information" (using their terminology) efficient GMM estimation of $\beta_{[a,b]}^{0}$ based on (15) and (16) are equivalent in terms of the asymptotic variance of the estimator of $\beta_{[a,b]}^{0}$. Concretely, this "limited information" estimator is the efficient GMM estimator based on the IPW GMM sample moment vector in (14), i.e.,

$$\begin{aligned}\widehat{\beta}_{[a,b]}^{IPW}(W_n):=\arg\min_{\beta\in\mathcal{B}}&\left(\frac{1}{n}\sum_{i=1}^{n}I(C_i=R)\widehat{\omega}_{[a,b]}^{\mathrm{IPW}}m(Z_i;\beta)\right)'\\&\times W_n\left(\frac{1}{n}\sum_{i=1}^{n}I(C_i=R)\widehat{\omega}_{[a,b]}^{\mathrm{IPW}}m(Z_i;\beta)\right)\end{aligned} \tag{17}$$

when W_n^{-1} is consistent for the asymptotic variance of the moment vector, accounting for the estimation of the nuisance conditional hazards $P(C=r|T_r,C\geq r)$'s involved in $\omega_{[a,b]}^{\mathrm{IPW}}$.

The equivalence in asymptotic variance in Proposition 6(i) holds because the conditional hazards $P(C=r|T_r,C\geq r)$'s that constitute $\omega_{[a,b]}^{\mathrm{IPW}}$ are "exactly identified" by (16).

Proposition 6(ii) shows that this asymptotic variance in Proposition 6(i) reaches the semiparametric efficiency bound that was obtained in Proposition 2 under the general MAR condition in (1) for $\beta_{[a,b]}^{0}$ defined by (2). Thus, in the spirit of Graham (2011), we say that the moment restrictions (15) and (16) exhaust all available information about $\beta_{[a,b]}^{0}$ under the general setup of Proposition 2. Similar results with $R=2$ are known from Hirano et al. (2003), Chen et al. (2008), Graham (2011), etc., but the case of $R>2$ will help us to get further insights into this result and the information content of the MAR assumption.

We spend the rest of this section discussing Proposition 6(ii) with the following remarks.

First, there are two different semiparametric efficiency bounds present in Proposition 6: in (i) it is the bound based on the system (15) and (16), whereas in (ii) it is the bound based on our general framework (1) and (2). The result on

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

semiparametric efficiency in Newey (1994), Ackerberg et al. (2014), etc. of the "limited information" approach concerns the first efficiency bound, i.e., the result in Proposition 6(i). On the other hand, the second efficiency bound is traditionally established independently in this literature, albeit in simpler contexts. Graham (2011) established the equality of these two bounds when $R = 2$ and $a = 1, b = R$, and $d_m = d_\beta$; however, his result was based on the Brown and Newey (1998) orthogonalization that is not applicable here if interest lies on subpopulations.

Second, we find the equality of the two efficiency bounds remarkable in the case of $R > 2$ considering how much information the general MAR condition in (1) has and how little of it is used by the moment restrictions (15) and (16) leading to the first efficiency bound. In fact, (1) does not play any direct role in Proposition 6. (1)'s only role would be in the background ensuring that (15) holds, and Proposition 6(i) takes (15) as given.[11] The general MAR condition in (1) has no role to play in (16)—it contains no information about the unknown parameters in (16) since these moment restrictions simply follow from the definition of the conditional hazards and thus the parameters involved there are what is variously called "nonparametrically identified," "exactly identified," or "locally just identified"; see Newey (1994), Ackerberg et al. (2014), Chen and Santos (2018), Chernozhukov et al. (2022), etc.

Third, we point out that an equivalence result like Proposition 6(ii) will not hold if the general MAR condition in (1) is strengthened. The limited or full information approach will "pay a price" in terms of efficiency for not considering the (strengthened) MAR condition. For a clean demonstration of paying a price, Lemma 7(ii) strengthens (1) by imposing an extreme dimension reduction on the conditioning set leading to the CMAR condition in (13).

LEMMA 7. (i) *The efficient GMM estimator of* $\beta^0_{[a,b]}$ *based on the moment restrictions*

$$E\left[\sum_{j=a}^{b} \frac{P(C=j)}{P(a \le C \le b)} \frac{I(C=R)}{\prod_{r=j}^{R-1}(1-P(C=r|T_1, C \ge r))} \frac{P(C=j|T_1, C \ge j)}{P(C=j)} m(Z; \beta^0_{[a,b]})\right] = 0,$$

where for each $r = a \ldots, R-1$ *the* $P(C=r|T_1, C \ge r)$*'s solve for the* $p_r(T_1)$*'s from* $E[I(C \ge r)\{I(C=r) - p_r(T_1)\} \mid T_1] = 0$ *almost surely* T_1,

has the same asymptotic variance $\left(M'_{[a,b]}\left[V^\dagger_{[a,b]}\right]^{-1} M_{[a,b]}\right)^{-1}$ *under both the "limited and full information" approaches under regularity conditions where* $V^\dagger_{[a,b]} := E\left[\varphi^\dagger_{[a,b]}\varphi^{\dagger\prime}_{[a,b]}\right]$ *and*

[11] For (15) to hold, it only requires the part of MAR with $r = a, \ldots, R-1$. The part with $r = 1, \ldots, a-1$ is unused since only the $P(C=r|T_r, C \ge r)$'s for $r = a, \ldots, R-1$ appear in the weight $\omega^{IPW}_{[a,b]}$; see (5).

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

$$\varphi^{\dagger}_{[a,b]} = \frac{I(C=R)}{P(C=R|T_1)} \frac{P(a \le c \le b|T_1)}{P(a \le C \le b)} \left[m(Z;\beta^0_{[a,b]}) - E[m(Z;\beta^0_{[a,b]})|T_1] \right] + \frac{I(a \le C \le b)}{P(a \le C \le b|T_1)} E[m(Z;\beta^0_{[a,b]})|T_1].$$

(ii) *The inverse of the semiparametric information bound for* $\beta^0_{[a,b]}$ *in Proposition 4 that works under the CMAR condition in (13) cannot exceed this asymptotic variance* $\left(M'_{[a,b]} \left[V^{\dagger}_{[a,b]} \right]^{-1} M_{[a,b]} \right)^{-1}$ *because* $V^{\dagger}_{[a,b]} - V^{CMAR}_{[a,b]}$ *is positive semi-definite since* $V^{\dagger}_{[a,b]} - V^{CMAR}_{[a,b]}$ *is given by*

$$\sum_{r=2}^{R} E\left[\frac{P(a \le C \le b|T_1)}{P(a \le C \le b)} \left\{ \frac{1}{P(C \ge R|T_1)} - \frac{1}{P(C \ge r|T_1)} \right\} \times Var\left(E[m(Z;\beta^0_{[a,b]})|T_r] \mid T_{r-1} \right) \middle| a \le C \le b \right].$$

The moment function for $\beta^0_{[a,b]}$ in Lemma 7(i) could be more compactly written as the weighted average of the $I(C=R)\omega_{R,j}(T_1)m(Z;\beta^0_{[a,b]})$'s where $\omega_{R,j}(.)$ is defined in (5). However, the more elaborate form in the lemma helps to better visualize where/how the variance adjustment, as in Newey (1994), Ackerberg et al. (2014), Chernozhukov et al. (2022), etc., works in this IPW estimation. It works in Lemma 7(i) because there the conditional hazards are still exactly identified by the respective conditional moment restrictions. By contrast, the variance adjustment of IPW does not (and should not) work in Lemma 7(ii) in a way that leads to the efficient influence function and efficiency bound from Proposition 4.

The variance adjustment of IPW does not lead to the efficiency bound from Proposition 4 because the variance adjustment is based only on the given moment restrictions and no other information such as the MAR or CMAR conditions.[12] Note, e.g., that the CMAR condition in (13) contains additional information $P(C=r|Z,C \ge r) = \cdots = P(C=r|T_r,C \ge r) = \cdots = P(C=r|T_1,C \ge r)$ about the nuisance conditional hazards $P(C=r|T_1,C \ge r)$ in Lemma 7(i), thus providing a sequence of additional feasible moment restrictions

$$E\left[I(C \ge r) \{ I(C=r) - p_r(T_1) \} \mid T_j \right] = 0 \text{ almost surely } T_j, \text{ for } j = 1, \ldots, r,$$

to solve for the $p_r(T_1)$'s in Lemma 7(i). Lemma 7(i) does not use this additional information and hence the IPW variance adjustment fails to reach the efficiency bound in Proposition 4.

While CMAR is a strong assumption, other types of strengthening of MAR—e.g., $P(C=r|Z,C \ge r) = P(C=r|Z_r,C \ge r)$, i.e., with conditioning set involving only period r's observables and not the entire history—could be more plausible.

[12] This did not matter in Proposition 6(ii) that worked under the MAR condition (1) because MAR did not have any information on the nuisance conditional hazards $P(C=r|T_r,C \ge r)$ in Proposition 6(i). MAR's information $P(C=r|Z,C \ge r) = P(C=r|T_r,C \ge r)$ cannot be feasibly used based on the observed data for efficiency via overidentification of $P(C=r|T_r,C \ge r)$. This led to the equivalence in Proposition 6(ii).

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

In general, the common empirical practice of any kind of variable selection also leads to an implicit strengthening of the MAR condition by imposing exclusion restrictions. It is likely that in such cases the IPW variance adjustment will also not lead to the efficiency bound like in the CMAR example.

Finally, we note that despite Proposition 6 and the elegant theory in the literature behind the variance adjustment of IPW estimators based on nonparametric estimation of the conditional hazards, IPW is not our recommended estimator even under MAR. The theory depends crucially on proper conditioning on the conditioning sets T_r's. However, the dimension of the conditioning set T_r increases with r, and in practice it is difficult to condition on all those variables especially if they are continuous. This makes the theory of nonparametric variance adjustment less reflective of the finite-sample behavior even in very large samples when $R > 2$, which is a key feature of our paper. Simulations in Supplementary Appendix B suggest that nonparametric variance adjustment can underestimate IPW's true variability (measured by Monte Carlo variance) even in very large samples when $R > 2$, while parametric variance adjustment in (17) in the spirit of Newey (1994) or Ackerberg et al. (2012) can reflect the true variability in moderately large samples. This issue with IPW is distinct from the problems with IPW (primarily concerning bias) that have been noted in the recent double robustness literature; see Rothe and Firpo (2019), Chernozhukov et al. (2022), etc.

4. ESTIMATOR OF $\beta^0_{[a,b]}$ AND ITS ASYMPTOTIC PROPERTIES

Our proposed estimator for $\beta^0_{[a,b]}$ will utilize the doubly robust structure of $\varphi_{[a,b]}(O;\beta)$ that was highlighted in remark 5 following Proposition 2. We know from (3)–(5) that $\varphi_{[a,b]}(O;\beta)$ depends on the unknown conditional hazards and conditional expectations. Denote the true value of these nuisance parameters by $p^0(T_{R-1})$ and $q^0(T_{R-1};\beta)$ where

$$p^0(T_{R-1}) := (P(C = R-1|T_{R-1}, C \geq R-1), \ldots, P(C = a|T_a, C \geq a))',$$
$$q^0(T_{R-1};\beta) := (E[m(Z;\beta)|T_{R-1}]', \ldots, E[m(Z;\beta)|T_a]')'.$$

Let $p(T_{R-1})$ and $q(T_{R-1};\beta)$ be generic functions of the same dimension as $p^0(T_{R-1})$ and $q^0(T_{R-1};\beta)$.

Define the function $g(O;\beta, p(T_{R-1}), q(T_{R-1};\beta))$ as $\varphi_{[a,b]}(O;\beta)$ with the conditional hazards and conditional expectations replaced by the concerned elements of $p(T_{R-1})$ and $q(T_{R-1};\beta)$, respectively. Note that $g(O;\beta, p^0(T_{R-1}), q^0(T_{R-1};\beta)) \equiv \varphi_{[a,b]}(O;\beta)$ for all β.

We will use the following $d_m \times 1$ GMM sample moment vector to estimate the $d_\beta \times 1 \beta^0_{[a,b]}$:

$$\bar{g}_n(\beta, \widehat{p}(T_{R-1}), \widehat{q}(T_{R-1},\beta)) := \frac{1}{n}\sum_{i=1}^{n} g(O_i;\beta, \widehat{p}(T_{R-1,i}), \widehat{q}(T_{R-1,i},\beta)), \tag{18}$$

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

where $\widehat{p}(T_{R-1,i})$ and $\widehat{q}(T_{R-1,i},\beta)$ are nonparametric or parametric estimators of $p^0(T_{R-1,i})$ and $q^0(T_{R-1,i};\beta)$ for $i=1,\ldots,n$; see Robins and Rotnitzky (1992), Robins et al. (1994), etc. For a $d_m \times d_m$ weighting matrix W_n, we will define the GMM estimator of $\beta^0_{[a,b]}$ as

$$\widehat{\beta}(W_n) := \arg\min_{\beta\in\mathcal{B}} \left[\bar{g}_n(\beta,\widehat{p}(T_{R-1}),\widehat{q}(T_{R-1},\beta))\right]' W_n \left[\bar{g}_n(\beta,\widehat{p}(T_{R-1}),\widehat{q}(T_{R-1},\beta))\right]. \tag{19}$$

Practitioners often use flexible parametric models to estimate the nuisance parameters. If there is "promise" to make the models more flexible when sample size increases, then such estimators can be considered as nonparametric, otherwise they are parametric; see, e.g., Newey (1994, p. 1369), Ackerberg et al. (2012), etc. We adopt this convention in our paper and provide a brief heuristic discussion of the properties of $\widehat{\beta}(W_n)$ by considering both parametric and nonparametric estimation of the nuisance parameters under a unified framework. Some generality is lost due to the unified presentation; but these results are already well known and our presentation here is only for the sake of completeness.

First, consider the conditional hazards. Let the parametric model, e.g., logit/probit, for $P(C=r|T_r, C\geq r)$ be $p_r(T_r;\gamma_r)$ where γ_r is a $d_{\gamma_r}\times 1$ unknown vector for $r=a,\ldots,R-1$. We obtain the quasi-maximum likelihood estimator $\widehat{\gamma}_r$ of γ_r solving the score equations:

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} S_r(O_i;\widehat{\gamma}_r) \text{ for } r=a,\ldots,R-1, \text{ where for } i=1,\ldots,n, \\ S_r(O_i;\gamma_r) &:= I(C_i\geq r)\frac{I(C_i=r)-p_r(T_{r,i};\gamma_r)}{p_r(T_{r,i};\gamma_r)(1-p_r(T_{r,i};\gamma_r))}\left\{\frac{\partial}{\partial\gamma_r}p_r(T_{r,i};\gamma_r)\right\}. \end{aligned} \tag{20}$$

Now, consider the conditional expectations. Let the parametric model, e.g., linear model, for the jth element $E[m_j(Z;\beta)|T_r]$ of $E[m(Z;\beta)|T_r]$ be $q_{r,j}(T_r;\beta,\lambda_{r,j}(\beta))$. Let $q_r(T_r;\beta,\lambda_r(\beta)) = (q_{r,1}(T_r;\beta,\lambda_{r,1}(\beta)),\ldots,q_{r,d_m}(T_r;\beta,\lambda_{r,d_m}(\beta)))'$ where $\lambda_r(\beta) = (\lambda'_{r,1}(\beta),\ldots,\lambda'_{r,d_m}(\beta))'$ and $\lambda_{r,j}(\beta)$ is a $d_{\lambda_{r,j}}\times 1$ unknown vector for $r=a,\ldots,R-1$, $j=1,\ldots,d_m$. We obtain the least squares estimator $\widehat{\lambda}_{r,j}(\beta)$ of $\lambda_{r,j}(\beta)$ for $j=1,\ldots,d_m$ as functions of β solving the normal equations:

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} L_{r,j}(O_i;\beta,\widehat{\lambda}_{r,j}(\beta)) \text{ for } r=a,\ldots,R-1, \text{ where for } i=1,\ldots,n, \\ L_{r,j}(O_i;\beta,\lambda_r) &:= I(C_i=R)\left\{\frac{\partial}{\partial\lambda_{r,j}}q_{r,j}(T_{r,i};\beta,\lambda_{r,j})\right\}\left(m_j(T_{R,i};\beta)-q_{r,j}(T_{r,i};\beta,\lambda_{r,j})\right). \end{aligned} \tag{21}$$

In empirical work, the $p_r(T_r;\gamma_r)$'s are typically logit/probit with index $\xi'_{d_{\gamma_r}}(T_r)\gamma_r$, and the $q_{r,j}(T_r;\beta,\lambda_{r,j}(\beta))$'s are typically linear $\pi'_{d_{\lambda_{r,j}}}(T_r)\lambda_{r,j}(\beta)$ where the $\xi_{d_{\gamma_r}}(T_r)$'s and $\pi_{d_{\lambda_{r,j}}}(T_r)$'s are possibly the first d_{γ_r} and $d_{\lambda_{r,j}}$ terms of some basis function, e.g., powers. We consider the estimator $\widehat{p}(T_{R-1}) = (p_{R-1}(T_{R-1};\widehat{\gamma}_{R-1}),\ldots,$

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

$p_a(T_a;\widehat{\gamma}_a))'$ for $p^0(T_{R-1})$ and the estimator $\widehat{q}(T_{R-1};\beta) = (q'_{R-1}(T_{R-1};\beta,\widehat{\lambda}_{R-1}(\beta)), \ldots, q'_a(T_a;\beta,\widehat{\lambda}_a(\beta)))'$ for $q^0(T_{R-1};\beta)$ as parametric if the d_{γ_r}'s and $d_{\lambda_{r,j}}$'s are fixed, and as nonparametric if the d_{γ_r}'s and $d_{\lambda_{r,j}}$'s increase with n.

Assumption CH. The conditional hazard (CH) models are correct, i.e., there exists a $\gamma^0 = (\gamma_a^{0'}, \ldots, \gamma_{R-1}^{0'})'$ such that $p_r(T_r;\gamma_r^0) = P(C = r|T_r, C \geq r)$ for $r = a, \ldots, R-1$.

Assumption CE. The conditional expectation (CE) models are correct, i.e., there exists a $\lambda^0 = (\lambda_a^{0'}, \ldots, \lambda_{R-1}^{0'})'$ such that $q_r(T_r;\beta_{[a,b]}^0, \lambda_r^0) = E[m(Z;\beta_{[a,b]}^0)|T_r]$ for $r = a, \ldots, R-1$.

Assumptions CH and CE can be assumed to hold approximately arbitrarily well if $\widehat{p}(T_{R-1})$ and $\widehat{q}(T_{R-1};\beta)$ are nonparametric. But assumptions CH and CE may not hold if $\widehat{p}(T_{R-1})$ and $\widehat{q}(T_{R-1};\beta)$ are parametric. We will assume that $\|\widehat{p} - p^*\| = o_p(1)$ and $\|\widehat{q} - q^*\| = o_p(1)$ (at suitable rates and with respect to suitable metrics in suitable function spaces) for some pseudo-true functions $p^*(T_{R-1})$ and $q^*(T_{R-1};\beta)$ where $p^*(T_{R-1}) = p^0(T_{R-1})$ if CH holds and $q^*(T_{R-1};\beta_{[a,b]}^0) = q^0(T_{R-1};\beta_{[a,b]}^0)$ if CE holds. If both CH and CE fail to hold, then there is no protection of double robustness and the GMM moment for $\beta_{[a,b]}^0$ may be misspecified. Then, in case of overidentification ($d_m > d_\beta$), there may be no solution to the GMM population moment restriction and the probability limit of $\widehat{\beta}(W_n)$, if it exists, may depend on the limiting behavior of W_n; see, e.g., Hall and Inoue (2003). Such probability limits may not be of interest in the related empirical literature where the focus is on the true value $\beta_{[a,b]}^0$ and not the pseudo true values. Therefore, in our heuristic discussion below of the asymptotic properties of $\widehat{\beta}(W_n)$, we will maintain that assumptions CH and CE cannot be jointly false.

First, consistency. Double robustness implies (see remark 5 following Proposition 2) that

$$\begin{aligned} &E[g(O;\beta,p^*(T_{R-1}),q^0(T_{R-1};\beta))] \\ &\quad = E[g(O;\beta,p^0(T_{R-1}),q^*(T_{R-1};\beta))] = E[m(Z;\beta)|a \leq C \leq b]. \end{aligned}$$

Therefore, consistency $\widehat{\beta}(W_n) \xrightarrow{p} \beta_{[a,b]}^0$ follows under standard conditions (see, e.g., Theorem 1 of Chen, Linton, and van Keilegom (2003)) if CH and CE are not jointly false.

Now, the asymptotic distribution of $\widehat{\beta}(W_n)$. We can see that the same double robustness property also implies that the $M_{[a,b]}$ defined in assumption A3 satisfies

$$\begin{aligned} M_{[a,b]} &= \frac{\partial}{\partial\beta'} E[g(O;\beta_{[a,b]}^0,p^0(T_{R-1}),q^*(T_{R-1};\beta))] \\ &= \frac{\partial}{\partial\beta'} E[g(O;\beta_{[a,b]}^0,p^*(T_{R-1}),q^0(T_{R-1};\beta_{[a,b]}^0))]. \end{aligned}$$

Let $G_p(\beta,p,q)[v_p]$ and $G_q(\beta,p,q)[v_q]$ be the pathwise derivatives of $E[g(O;\beta,p,q)]$ at p and q in the directions v_p and v_q such that $p + \tau v_p$ and $q + \tau v_q$ for $\tau \in [0,1]$

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

belong in the respective function spaces. We can see that the same double robustness property also implies that

$$G_p(\beta^0_{[a,b]}, p^*(T_{R-1}), q^0(T_{R-1}, \beta^0_{[a,b]})) = 0 \text{ and } G_q(\beta^0_{[a,b]}, p^0(T_{R-1}), q^*(T_{R-1}, \beta)) = 0. \tag{22}$$

Let $W_n \xrightarrow{p} W$. If $\widehat{\beta}(W_n) \xrightarrow{p} \beta^0_{[a,b]}$ as we just noted above, then it now follows under standard conditions (see, e.g., Theorem 2 of Chen et al. (2003)) that

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) = &-\left(M'WM\right)^{-1} M'W\sqrt{n}\left[\bar{g}_n(\beta^0, p^*, q^*)\right. \\ &\left. + G_p(\beta^0, p^*, q^*)[\widehat{p} - p^*] + G_q(\beta^0, p^*, q^*)[\widehat{q} - q^*]\right] + o_p(1),\end{aligned}$$

writing the triple $\beta, p(T_{R-1}), q(T_{R-1}; \beta)$ as β, p, q, and dropping the subscript $[a, b]$ for brevity.

Therefore, if assumption CH holds, then $p^*(T_{R-1}) = p^0(T_{R-1})$ and hence by (22)

$$\begin{aligned}&\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) \\ &\quad = -\left(M'WM\right)^{-1} M'W\sqrt{n}\left[\bar{g}_n(\beta^0, p^0, q^*) + G_p(\beta^0, p^*, q^*)[\widehat{p} - p^0]\right] + o_p(1).\end{aligned}$$

So the estimation of the unknown conditional expectations $E[m(Z; \beta^0)|T_r]$'s has no effect on the asymptotic distribution of $\widehat{\beta}(W_n)$ if the conditional hazard models are correct.

Similarly, if assumption CE holds, then $q^*(T_{R-1}; \beta^0) = q^0(T_{R-1}; \beta^0)$ and hence by (22)

$$\begin{aligned}&\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) \\ &\quad = -\left(M'WM\right)^{-1} M'W\sqrt{n}\left[\bar{g}_n(\beta^0, p^*, q^0) + G_q(\beta^0, p^0, q^0)[\widehat{q} - q^0]\right] + o_p(1).\end{aligned}$$

So the estimation of the unknown conditional hazards $P(C = r|T_r; C \geq r)$'s has no effect on the asymptotic distribution of $\widehat{\beta}(W_n)$ if the conditional expectation models are correct.

Finally, if both assumptions CH and CE hold, then we have $p^*(T_{R-1}) = p^0(T_{R-1})$ and $q^*(T_{R-1}; \beta^0) = q^0(T_{R-1}; \beta^0)$ and hence by (22)

$$\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) = -\left(M'WM\right)^{-1} M'W\sqrt{n}\bar{g}_n(\beta^0, p^0, q^0) + o_p(1).$$

Now consider efficiency in the sense of Proposition 2. If $W^{-1} = Var\big(g(O; \beta^0, p^*(T_{R-1}), q^*(T_{R-1}, \beta^0))\big) =: V(\beta^0, p^*, q^*)$, which when CH and CE hold jointly is denoted by $V(\beta^0, p^0, q^0)$, then

$$\begin{aligned}&\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) \\ &\quad = -\left(M'[V(\beta^0, p^0, q^0)]^{-1}M\right)^{-1} M'[V(\beta^0, p^0, q^0)]^{-1}\sqrt{n}\bar{g}_n(\beta^0, p^0, q^0) + o_p(1)\end{aligned}$$

when CH and CE hold jointly. Now, since the moment vector $g(O; \beta, p, q)$ was defined such that $g(O; \beta^0, p^0, q^0) \equiv \varphi_{[a,b]}(O; \beta^0)$ (and hence $V(\beta^0, p^0, q^0) \equiv V_{[a,b]}$),

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

it follows that

$$\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) = -\Omega_{[a,b]}^{-1} M'_{[a,b]} V_{[a,b]}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{[a,b]}(O_i; \beta^0) + o_p(1), \tag{23}$$

where the non-$o_p(1)$ term on the RHS is the influence function from Proposition 2 which was shown to be efficient for any $[a,b]$ when $d_m = d_\beta$ and for $a = b$ or $a = 1, b = R$ when $d_m > d_\beta$. Under the conditions maintained in Proposition 2, it follows from (23) that

$$\sqrt{n}\left(\widehat{\beta}(W_n) - \beta^0\right) \xrightarrow{d} N\left(0, \Omega_{[a,b]}^{-1}\right).$$

The related literature on the doubly or locally robust moment functions using nonparametric $\widehat{p}$ and $\widehat{q}$, or even parametric $\widehat{p}$ and $\widehat{q}$ but without allowing for the violation of CH or CE, focuses solely on (23) and takes $\Omega_{[a,b]}^{-1}$ as the asymptotic variance of $\widehat{\beta}(W_n)$ when $W_n \xrightarrow{p} V_{[a,b]}^{-1}$.

However, assumption CH or CE may not hold if $\widehat{p}$ and $\widehat{q}$ are parametric. Then the above asymptotically linear representations of $\widehat{\beta}(W_n)$ are not practically useful to obtain the asymptotic variance of $\widehat{\beta}(W_n)$ without more structure on $\widehat{p}$ and $\widehat{q}$. The usual solution is to exploit the parametric structure of $\widehat{p}$ and $\widehat{q}$, and obtain the asymptotic variance of $\widehat{\beta}(W_n)$ based on the standard stacked representation of the moment vectors for β, $\gamma := (\gamma_a', \ldots, \gamma_{R-1}')'$ and $\lambda := (\lambda_a', \ldots, \lambda_{R-1}')'$ where $\lambda_r := (\lambda_{r,1}', \ldots, \lambda_{r,d_m}')'$ for $r = a, \ldots, R-1$. Accordingly, consider the $\left(d_m + \sum_{r=a}^{R-1} d_{\gamma_r} + \sum_{r=a}^{R-1} \sum_{j=1}^{d_m} d_{\lambda_{r,j}}\right) \times 1$ stacked moment vector:

$$\psi(O_i; \beta, \gamma, \lambda) := \begin{bmatrix} g(O_i; \beta, p(Z_i; \gamma), q(Z_i; \beta, \lambda)) \\ \underline{S}(O_i; \gamma) \\ \underline{L}(O_i; \beta, \lambda) \end{bmatrix} \text{ where } \underline{S}(O_i; \gamma) := \begin{bmatrix} S_a(O_i; \gamma_a) \\ \vdots \\ S_{R-1}(O_i; \gamma_{R-1}) \end{bmatrix},$$

$$\underline{L}(O_i; \beta, \lambda) := \begin{bmatrix} \underline{L}_a(O_i; \beta, \lambda_a) \\ \vdots \\ \underline{L}_{R-1}(O_i; \beta, \lambda_{R-1}) \end{bmatrix}, \text{ and } \underline{L}_r(O_i; \beta, \lambda_r) := \begin{bmatrix} L_{r,1}(O_i; \beta, \lambda_{r,1}) \\ \vdots \\ L_{r,d_m}(O_i; \beta, \lambda_{r,d_m}) \end{bmatrix}$$

for $r = a, \ldots, R-1$. We will obtain the GMM estimator $\widehat{\beta}$ using the usual two-step GMM.

We will refer to $\widehat{\beta}$ as EFF (as in efficient). In step one, we use the identity matrix as the GMM weighting matrix to obtain the first step estimators $\bar{\beta}, \bar{\gamma}$ and $\bar{\lambda}$ for β, γ and λ, and estimate the efficient weighting matrix as $\widehat{\Sigma}_n^{-1}(\bar{\beta}, \bar{\gamma}, \bar{\lambda})$ where $\widehat{\Sigma}_n(\beta, \gamma, \lambda) := \sum_{i=1}^{n} \psi(O_i; \beta, \gamma, \lambda)\psi'(O_i; \beta, \gamma, \lambda)/n$. Step one is not needed if $d_m = d_\beta$. In step two, we obtain the efficient GMM estimators $\widehat{\beta}, \widehat{\gamma}$, and $\widehat{\lambda}$ by minimizing with respect to β, γ, λ the GMM objective function based on the efficient weighting matrix. Finally, we estimate the asymptotic variance of $\widehat{\beta}$, i.e., EFF, as the first $d_\beta \times d_\beta$ block diagonal of the GMM asymptotic variance matrix $\left(\widehat{\Psi}_n'(\widehat{\beta}, \widehat{\gamma}, \widehat{\lambda})\widehat{\Sigma}_n^{-1}(\widehat{\beta}, \widehat{\gamma}, \widehat{\lambda})\widehat{\Psi}_n(\widehat{\beta}, \widehat{\gamma}, \widehat{\lambda})\right)^{-1}$ where $\widehat{\Psi}_n(\widehat{\beta}, \widehat{\gamma}, \widehat{\lambda})$ is the (possibly

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

numerical) derivative of $\sum_{j=1}^{n} \psi(O_j; \beta, \gamma, \lambda)/n$ with respect to β, γ, and λ at $\widehat{\beta}, \widehat{\gamma}$, and $\widehat{\lambda}$.

The asymptotic theory for EFF with parametric (fixed) nuisance models is simple. When CH and CE are not jointly false, the interesting structure described in the text between equations (22) and (23) is preserved by the influence function of EFF (and hence its asymptotic variance) thanks to the double robustness to the misspecification of the parametric nuisance models. If the parametric nuisance models are not fixed but "promise" to become sufficiently flexible with the increase in sample size, then, as shown in Ackerberg et al. (2012) (see also Newey, 1994), EFF can be interpreted as semiparametric and the estimator of its asymptotic variance obtained above can be consistent for the benchmark variance $\Omega^{-1}_{[a,b]}$.

5. EMPIRICAL ILLUSTRATION BASED ON PROJECT STAR

We continue with the motivating example from Section 2.1 of attrition in Project STAR. We wish to illustrate the possible benefits of the efficiency gains due to our proposed estimator EFF in drawing substantive conclusion from this experiment on the effect of small class size on students' performance. As a reference to EFF, we also present the same results using the IPW estimator from (14) that is the reweighted Hajek (1971) version of IPW.

To this end, it is useful to first ask which effects an "ideal" Project STAR experiment would have generated with the subjects/students entering grade K in 1985 if there was no subsequent attrition or other implementation-related compromises (see, e.g., Hanushek, 1999). The answer is that, since there was no protocol to randomly assign the class types of students except at the beginning, an "ideal" Project STAR experiment would have generated in grades K, 1, 2, and 3 the effect of continued presence in small classes with respect to continued presence in not-small classes. Our illustration will focus on the "ideal" experiment.

We first formally define these effects that the "ideal" experiment would have generated. We view attrition—a compromise to the ideal experiment—as a mitigating action by students in response to the treatment (class type) that they perceived as unhelpful to them. To gain a better understanding of this mitigating action, we then decompose these effects by the attrition behavior of students from small and not-small classes.

For brevity of this illustration, we present only the results for (normalized) reading scores.[13] Let Y^{s}(grade j read) be the potential grade j reading score of a student *had* (s)he stayed in the small class at least until the end of grade j for $j = K, 1, 2, 3$ after being initially randomized to a small class in grade K. Similarly, with superscript "ns" denoting not-small, define the potential scores

[13]To streamline our empirical illustration, we ignore the compromises other than attrition to the experiment, e.g., students who enrolled after grade K or the few students (1.8%–5.8% in the respective grades; see Table 1) who switched their assigned class types. Some of these compromises can be accommodated in this illustration at the cost of strong modeling assumptions and messier notation that we want to avoid here for simplicity.

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

Y^{ns}(grade j read) for $j = K, 1, 2, 3$. These scores are not observed for a student in grade j if the student left the participating school before grade $j = 1, 2, 3$.

As noted above, we focus on two treatment regimes—a continued presence in small classes and a continued presence in not-small classes over the 4 years of Project STAR. Denote the average difference between the outcomes of these two regimes at each grade $j = K, 1, 2, 3$ as

$$\mu_j^{\text{read}} := E[Y^{s}(\text{grade } j \text{ read}) - Y^{ns}(\text{grade } j \text{ read})].$$

5.1. Evolution of the Effect of Small Classes

First, consider the trajectory of μ_j^{read} for $j = K, 1, 2, 3$ to see how the effect of the small-class regime with respect to the not-small class regime evolved over continued presence in these regimes. Their EFF and IPW estimates are plotted in Figure 1a.[14] The EFF and IPW estimates of the trajectory are quite similar. Consistent with the literature, we observe that the initial effect μ_K^{read} is very large compared to the "value added" (e.g., $\mu_j^{\text{read}} - \mu_K^{\text{read}}$ for $j = 1, 2, 3$) in the subsequent grades 1, 2, and 3. However, our value added estimates are not as pessimistic as Hanushek's (1999) that led him to question the justification of the huge cost of prolonged operation of small classes, but are more in line with Krueger (1999).

We conjecture that the correction for attrition makes our estimates less pessimistic than Hanushek's (1999). This would happen under asymmetric selection, e.g., if the students leaving not-small classes left because they were going to score badly had they stayed, whereas the students leaving small classes left under other concerns or lesser concerns of bad scores.

Following up on our conjecture, as proxies to Hanushek's (1999) annual and 4-year samples, respectively, we also plot in Figure 1a the "In grade" and "Never left" estimates of the trajectory. These are based on the average observed score of the students who took the tests at the end of the respective grades (for In grade estimates) and the students who continued in Project STAR until the end of grade 3 (for Never left estimates).[15] Note that Never left actually estimates $\nu_{j,3}^{\text{read}}$, while In grade estimates $\nu_{j,j}^{\text{read}}$ for $j = K, 1, 2, 3$ where

[14] We obtain these estimates following Section 4 using parametric models specified for the conditional hazards and conditional expectations. The conditional hazard of leaving small (resp. not-small) classes after grade j (= K, 1, 2) is modeled as logit with a linear index of a constant, dummies for race, sex, types of school (inner city, urban, and rural), the share of students on free lunch in school, dummies for all grades (present and past) where the student was on free lunch, where the student's teacher had bachelor's degree, and the difference in each of the past grades between the student's normalized math and reading scores from, respectively, the average normalized math and average normalized reading scores in small classes and also in not-small classes in their school. The differences between the student's and the average scores are continuous variables, and we also include their quadratic and cubic terms in the index. The conditional expectations of the grade j (= 1, 2, 3) scores in small (resp. not-small) classes are modeled linearly with exactly the same set of variables. These estimation results are not reported but are available from us.

[15] "In grade" and "Never left" are those that correspond to the so-called "available cases" and "complete cases," respectively, in the parlance of the missing data literature. To fix ideas, consider Table 2. Never left has its own row in the table, while In grade for each grade is composed of the non-x entries in the column for that grade. In grade is

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

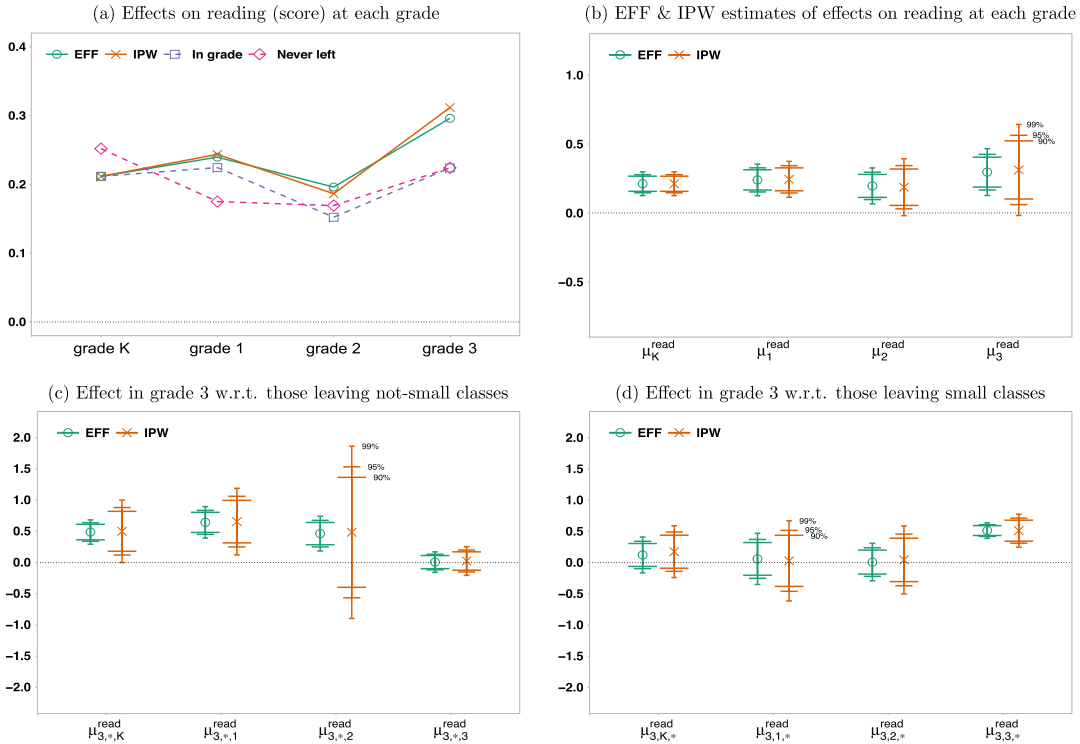

FIGURE 1. **(a)** EFF, IPW, In grade, and Never left estimates of effect on reading score at each grade. **(b)** EFF and IPW estimates and confidence intervals (90%, 95%, and 99%) of $\mu_K^{\text{read}}, \mu_1^{\text{read}}, \mu_2^{\text{read}}, \textit{and}\ \mu_3^{\text{read}}$. The 90%, 95%, 99% EFF, and IPW confidence intervals for the decomposition of μ_3^{read} by comparing: **(c)** the entirety of small classes with different attrition categories from not-small classes and **(d)** the different attrition categories from small classes with the entirety of not-small classes.

$$\nu^{\text{read}}_{j,l} := E[Y^{\text{s}}(\text{grade } j \text{ read})|B^{\text{s}}_l] - E[Y^{\text{ns}}(\text{grade } j \text{ read})|B^{\text{ns}}_l] \quad \text{for } j, l = K, 1, 2, 3$$

and B^{s}_l is the event that a student assigned to small class in grade K does not leave before the end of grade l for $l = K, 1, 2, 3$; and similarly B^{ns}_l is the event for the not-small class.[16]

Supporting our conjecture, visual inspection of In grade and Never left estimates reveals that without correction for attrition the value added estimates would indeed be pessimistic.

5.2. Does Attrition Matter?

But, beyond this visual inspection, does the correction for attrition matter statistically as well? More precisely, since we observed that the attrition-corrected estimates (EFF and IPW) are larger than the attrition-uncorrected estimates (In grade, which is typically favored to Never left), it is natural to ask if this is entirely due to sampling variation or is there systematic evidence for this in the population. That is, one would want to test the null hypothesis $H_{0,j} : \mu^{\text{read}}_j = \nu^{\text{read}}_{j,j}$ against the alternative $H_{1,j} : \mu^{\text{read}}_j > \nu^{\text{read}}_{j,j}$ for grades $j = 1, 2, 3$.

The p-values for these tests using EFF and IPW estimates of μ^{read}_j for grades $j = 1, 2, 3$ are as follows:

- 21% using EFF and 26.3% using IPW for $H_{0,1} : \mu^{\text{read}}_1 = \nu^{\text{read}}_{1,1}$ against $H_{1,1} : \mu^{\text{read}}_1 > \nu^{\text{read}}_{1,1}$. (Note that grade 1 score has a single level of missingness; see the caption of Table 2.)
- 5.5% using EFF and 30.7% using IPW for $H_{0,2} : \mu^{\text{read}}_2 = \nu^{\text{read}}_{2,2}$ against $H_{1,2} : \mu^{\text{read}}_2 > \nu^{\text{read}}_{2,2}$.
- 6.6% using EFF and 23.3% using IPW for $H_{0,3} : \mu^{\text{read}}_3 = \nu^{\text{read}}_{3,3}$ against $H_{1,3} : \mu^{\text{read}}_3 > \nu^{\text{read}}_{3,3}$.

The EFF p-values for $H_{0,2}$ and $H_{0,3}$ are small and not sufficient in practice to take for granted the reliability of the attrition-uncorrected In grade estimates for the true effect μ^{read}_2 and μ^{read}_3. On the other hand, the IPW p-values are quite a bit larger for $H_{0,2}$ and $H_{0,3}$. It is, however, not prudent (and possibly misleading) to take $H_{0,2}$ and $H_{0,3}$ for granted because, as we will see below, the large IPW p-values are entirely due to the imprecise nature of the IPW estimates. By contrast, EFF helps

preferred in practice (not always correctly) to Never left as a representative of the full population since it contains Never left and also units from various subpopulations of the full population.

[16] While we have deviated from the C-notation for attrition category to better reflect the sequencing $K, 1, 2, 3$ of grades, in this 4-period experiment: $B^{\text{s}}_K \equiv \{C \geq 1\}$ and $B^{\text{s}}_l \equiv \{C \geq l+1\}$ if $l = 1, 2, 3$ for small class, and similarly B^{ns}_l for not-small class. We hope that this switch from C to B notation is not confusing.

Equipped with this notation, let us now recall and generalize the motivating discussion below Table 2 in Section 2.1 on the problem of selection. $\nu^{\text{read}}_{K,K} = \mu^{\text{read}}_K$ obviously as attrition started only after the end of grade K. However, in general, $\nu^{\text{read}}_{j,l} \neq \mu^{\text{read}}_j$ for $j = K, 1, 2, 3$ and $l = 1, 2, 3$ unless suitable mean independence assumptions hold or, by happenstance, the biases for small and not-small classes cancel out, i.e., $E[Y^{\text{s}}(\text{grade } j \text{ read})|B^{\text{s}}_l] - E[Y^{\text{s}}(\text{grade } j \text{ read})] = E[Y^{\text{ns}}(\text{grade } j \text{ read})|B^{\text{ns}}_l] - E[Y^{\text{ns}}(\text{grade } j \text{ read})]$.

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

to avoid this possibly misleading confidence in $H_{0,2}$ and $H_{0,3}$ and points toward the possibility that attrition does matter here.

5.3. Do Attrition-Corrected Estimates Give Substantive Conclusions on the Effects?

Attrition-correction will be of limited use to practitioners if it does not lead to precisely estimated (zero or nonzero) effects. To explore if that is the case here, we plot in Figure 1b the 90%, 95%, and 99% two-sided confidence intervals around the EFF and IPW estimates for $\mu_K^{\text{read}}, \mu_1^{\text{read}}, \mu_2^{\text{read}}$, and μ_3^{read}. The EFF intervals turn out to be subsets of the IPW intervals.

Specifically, while the EFF and IPW intervals are identical for μ_K^{read} by definition and are similarly precise for μ_1^{read} (one level of missingness), the EFF intervals are much more precise than the IPW intervals for μ_2^{read} and μ_3^{read} (more than one level of missingness).

EFF rejects a zero or negative value of μ_j^{read} for all $j = K, 1, 2, 3$ at all conventional levels, but IPW fails to reject it for $j = 2, 3$ at the 1% level. (The EFF p-values do not exceed even .01%.) Small classes are an expensive policy proposition. Hence, the fact that EFF can rule out with extreme confidence any negative evidence against continued presence in small classes for every duration 1–4 years (after starting in grade K) has serious policy implications.

5.4. Attrition as a Mitigating Action against Unhelpful Class Type Assignment

Students were randomly assigned to small and not-small classes when they enrolled in a Project STAR school in grade K. Many students did not score well in their randomly assigned class type. Leaving the Project STAR school was an important course of mitigating action available to these students. If attrition in Project STAR was primarily due to this mitigating action, then, given the initial random assignment, we would expect that students who stayed scored better than what students who left would have scored had they stayed instead.

This is exactly what we observe in our estimates for each grade 1, 2, and 3. For brevity, we report here only the results for grade 3 since it is the terminal period of the experiment, and compare those who never left with each of the other attrition categories. Table 3 reports the EFF and IPW estimates of $\alpha_3^{\text{s,read}} - \alpha_j^{\text{s,read}}$ and $\alpha_3^{\text{ns,read}} - \alpha_j^{\text{ns,read}}$ for $j = K, 1, 2$ where

$$\alpha_j^{\text{s,read}} := E[Y^{\text{s}}(\text{grade 3 read}) \mid A_j^{\text{s}}], \text{ and } \alpha_j^{\text{ns,read}} := E[Y^{\text{ns}}(\text{grade 3 read}) \mid A_j^{\text{ns}}]$$

and A_j^{s} is the event that a student assigned to small class in grade K leaves exactly at the end of grade j; and similarly A_j^{ns} is the event for not-small classes.[17]

[17] This switch from the C to A notation in this 4-period experiment is trivial: $A_K^{\text{s}} \equiv \{C = 1\}$ and $A_j^{\text{s}} \equiv \{C = j+1\}$ if $j = 1, 2, 3$ for small class, and similarly A_j^{ns} for not-small class. As in footnote 16, this switch better reflects the sequencing $K, 1, 2, 3$ of grades and does so in small and not-small classes separately.

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

Table 3. EFF and IPW estimates and standard errors (in parentheses) for $\alpha_3^{t,read} - \alpha_j^{t,read}$ for $t = s, ns$ and $j = K, 1, 2$. *, **, and *** signify if the null that the parameter is zero is rejected against the alternative that it is greater than zero at the 10%, 5%, and 1% levels, respectively.

	$\alpha_3^{s,read} - \alpha_j^{s,read}$		$\alpha_3^{ns,read} - \alpha_j^{ns,read}$	
j	EFF	IPW	EFF	IPW
K	0.39***	0.34***	0.48***	0.48***
	(0.11)	(0.14)	(0.05)	(0.18)
1	0.45***	0.48**	0.64***	0.63***
	(0.16)	(0.24)	(0.08)	(0.19)
2	0.51***	0.47***	0.46***	0.46
	(0.11)	(0.20)	(0.09)	(0.53)

EFF and IPW estimates are very similar, but EFF is much more precise than IPW. Consequently, EFF confirms with a higher level of confidence in all cases the intuition that students who stayed scored better on average than what students who left would have scored had they stayed instead. By contrast, IPW fails to confirm at conventional levels of significance this intuition behind the choice to leave not-small classes at the end of grade 2.

Relatedly, consider the two decompositions of the effect μ_3^{read} of small classes by attrition categories

$$\mu_3^{read} = \sum_{j=K,1,2,3} \mu_{3,j,*}^{read} \times P\left(A_j^s\right) = \sum_{j=K,1,2,3} \mu_{3,*,j}^{read} \times P\left(A_j^{ns}\right)$$

based on the attrition from small and not-small classes, respectively, where, for $j = K, 1, 2, 3$,

$$\mu_{3,j,*}^{read} = E[Y^s(\text{grade } j \text{ read}) | A_j^s] - E[Y^{ns}(\text{grade } j \text{ read})],$$
$$\mu_{3,*,j}^{read} = E[Y^s(\text{grade 3 read})] - E[Y^{ns}(\text{grade 3 read}) | A_j^{ns}].$$

EFF and IPW estimates of these two decompositions, along with the 90%, 95%, and 99% two-sided confidence intervals, are reported in Figure 1c,d, showing the relative contribution of each attrition category from small and not-small classes, respectively, toward the overall effect. Given the large number of students who left, it is important to understand what the effect would have been with respect to students leaving at various junctures of the experiment. $\mu_{3,*,j}^{read}$ and $\mu_{3,j,*}^{read}$ for $j = K, 1, 2, 3$ are those effects on the grade 3 reading scores.

Figure 1c reveals that if we compare a randomly chosen student assigned to small class with a randomly chosen student assigned to not-small class who never left not-small class, then there is no benefit of small classes on the grade 3 reading score. The benefit on the grade 3 reading score is driven by the comparison of the

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

TABLE 4. EFF and IPW estimates of expected (counterfactual) reading scores in grade 3 by the student's attrition period are presented under the class types to which they were initially randomized. Standard deviations are presented in parentheses. All results in this empirical illustration are based on such parameters, and the standard errors of those results were computed by noting that the estimates in this table across the two class types are independent but are correlated within class types. Row (d), i.e., Never left, involves nothing unobserved, and hence both IPW and EFF estimates are equal to the simple group averages.

Left STAR school	Randomized to small class		Randomized to not-small class	
at the end of grade	EFF	IPW	EFF	IPW
(a) K	0.05	0.10	−0.27	−0.26
	(0.11)	(0.13)	(0.05)	(0.17)
(b) 1	−0.02	−0.04	−0.42	−0.42
	(0.16)	(0.23)	(0.08)	(0.19)
(c) 2	−0.07	−0.03	−0.24	−0.24
	(0.11)	(0.19)	(0.09)	(0.53)
(d) 3 (Never left)	0.44	0.44	0.22	0.22
	(0.04)	(0.04)	(0.03)	(0.03)

former student with randomly chosen students assigned to not-small class who left not-small class after grade K, 1, or 2.

Figure 1d reveals that if we compare a randomly chosen student assigned to not-small class with randomly chosen students assigned to small class who left small class after grade K or 1 or 2, then there is no harm to the grade 3 reading score due to not-small classes. The harm to the grade 3 reading score due to not-small classes is driven by the comparison of the former student with a randomly chosen student assigned to small class who never left small class. Thus, attrition was clearly a mitigating action against unhelpful class assignment.

These decompositions reveal such interesting patterns telling us which group of students (by attrition category) are driving the overall effect of small classes in the terminal period grade 3, and by how much. While the EFF estimates of the decompositions are very similar to the IPW estimates, the precision of EFF provides more statistical confidence toward confirming the contribution of each group of students to the overall effect of small classes.[18]

Lastly, as we noted in Section 2.1, these EFF-based inferences are precise mainly because the subpopulation-specific components of the effects are estimated more precisely by EFF. Table 4 reports the results for EFF and IPW estimation of a subset

[18]Note that, for each $j = K, 1, 2$, the estimands from Table 3 are related to these decompositions as follows: $\alpha_3^{\text{s,read}} - \alpha_j^{\text{s,read}} = \mu_{3,3,*}^{\text{read}} - \mu_{3,j,*}^{\text{read}}$ while $\alpha_3^{\text{ns,read}} - \alpha_j^{\text{ns,read}} = -(\mu_{3,*,3}^{\text{read}} - \mu_{3,*,j}^{\text{read}})$.

Therefore, going back to Table 3, we see that it suggests that EFF rejects the null $\mu_{3,*,3}^{\text{read}} = \mu_{3,*,j}^{\text{read}}$ against $\mu_{3,*,3}^{\text{read}} < \mu_{3,*,j}^{\text{read}}$ and the null $\mu_{3,3,*}^{\text{read}} = \mu_{3,j,*}^{\text{read}}$ against $\mu_{3,3,*}^{\text{read}} > \mu_{3,j,*}^{\text{read}}$ for each $j = K, 1, 2$ even at the 1% level. IPW cannot do that, and moreover it does not reject $\mu_{3,3,*}^{\text{read}} = \mu_{3,2,*}^{\text{read}}$ at any conventional level of significance.

https://doi.org/10.1017/S0266466624000203 Published online by Cambridge University Press

of such components. Rows (a)–(c) correspond to the components marked with "x" in the columns for the grade 3 score in Table 2 that was presented in Section 2.1 as an empirical motivation behind the theoretical contribution of our paper. The gain in precision due to EFF is clear in all cases.

6. CONCLUSION

Our paper provided a comprehensive presentation of efficiency in estimation of parameters defined by the missingness pattern of monotonically MAR data. The efficiency results on the parameters for generic subpopulations are new, and extend the well-known results on the treatment effects on the treated or the untreated or the parameters from the so-called "verify-out-of-sample" case in various empirically relevant directions.

We saw in the empirical illustration that such parameters are, among other things, fundamental to our understanding of the economic agent's mitigation behavior when faced with unhelpful situations, e.g., leaving a school where a class-assignment is perhaps not working well for the student. Our proposed estimator for such parameters is a standard two-step doubly robust estimator. We saw that its computation is standard, and its precision may help to draw substantive conclusions when the standard estimators fail to do so. The excellent performance of our proposed estimator in our simulation experiment (Supplementary Appendix B) and, by contrast, the poor performance of its competitors give credibility to the results obtained by our proposed estimator, and we hope that encourages its use in practice.

We now conclude by recalling two important technical features of our paper. First, we clearly characterized the additional restrictions that were imposed on the tangent set for the underlying semiparametric model by the overidentification of the parameters of interest. To the best of our knowledge, this characterization was missing from the related literature on missing data. In the process, we validated and extended various existing results.

Second, we analyzed the information content (strength) of the MAR assumption linking it to the usability of sample units toward efficient estimation in subpopulations. This allowed us to contrast between the efficiency bound that is reached by the variance adjustment due to the estimation of exactly identified nuisance parameters and the efficiency bound that is obtained under the model assumptions involving the strength of the MAR assumption.

To the best of our knowledge, these two technical features distinguish our paper from the related literature on missing data, and are possibly of independent interest for future work on semiparametric efficiency bounds in broader contexts.

SUPPLEMENTARY MATERIAL

Barnwell and Chaudhuri (2024): Supplement to "Efficiency in estimation under monotonic attrition," *Econometric Theory* Supplementary Material. To view, please visit https://doi.org/10.1017/S0266466624000203.

REFERENCES

Abowd, J. M., Crepon, B., & Kramarz, F. (2001). Moment estimation with attrition: An application to economic models. *Journal of the American Statistical Association*, 96, 1223–1231.

Abrevaya, J., & Donald, S. G. (2017). A GMM approach for dealing with missing data on regressors and instruments. *Review of Economics and Statistics*, 99, 657–662.

Achilles, C., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J., & Word, E. (2008). Tennessee's Student Teacher Achievement Ratio (STAR) project.

Ackerberg, D., Chen, X., & Hahn, J. (2012). A practical asymptotic variance estimator for two-step semiparametric estimators. *The Review of Economics and Statistics*, 94, 481–498.

Ackerberg, D., Chen, X., Hahn, J., & Liao, Z. (2014). Asymptotic efficiency of semiparametric two-step GMM. *Review of Economic Studies*, 81, 919–943.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–972.

Brown, B., & Newey, W. (1998). Efficient semiparametric estimation of expectations. *Econometrica*, 66, 453–464.

Cao, W., Tsiatis, A., & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96, 723–734.

Chaudhuri, S. (2020). On efficiency gains from multiple incomplete subsamples. *Econometric Theory*, 36, 488–525.

Chen, X., Hong, H., & Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Annals of Statistics*, 36, 808–843.

Chen, X., Linton, O., & van Keilegom, I. (2003). Estimation of semiparametric models when the criteria function is not smooth. *Econometrica*, 71, 1591–1608.

Chen, X., & Santos, A. (2018). Overidentification in regular models. *Econometrica*, 86, 1771–1817.

Chernozhukov, V., Escanciano, J.-C., Ichimura, H., Newey, W., & Robins, J. (2022). Locally robust semiparametric estimation. *Econometrica*, 90, 1501–1535.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, 126, 1593–1660.

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

Ding, W., & 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., & Moffitt, R. (1996). *An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics* [NBER Working paper].

Gill, R. D., van der Laan, M. J., & Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures and counterexamples. In D. Y. Lin, & T. R. Fleming (Eds.), *Proceedings of the first Seattle symposium in biostatistics: Survival analysis*. Lecture Notes in Statistics (pp. 255–294). Springer.

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*, 79, 437–452.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66, 315–331.

Hajek, J. (1971). Comment on a paper by D. Basu. In V. R. Godambe, & D. A. Sprott (Eds.), *Foundations of statistical inference* (p. 236). Holt, Rinehert and Winston.

Hall, A. R., & Inoue, A. (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics*, 114, 361–394.

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, 143–163.

Hirano, K., Imbens, G., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity scores. *Econometrica*, 71, 1161–1189.

Holcroft, C., Rotnitzky, A., & 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.

Hoonhout, P., & Ridder, G. (2019). Nonignorable attrition in multi-period panels with refreshment samples. *Journal of Business and Economic Statistics*, 37, 377–390.

Horvitz, D., & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663–685.

Khan, S., & Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78, 2021–2042.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114, 497–532.

Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *The Economic Journal*, 111, 1–28.

Muris, C. (2020). Efficient GMM estimation with incomplete data. *Review of Economics and Statistics*, 102, 518–530.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of Indian Soc. Agricultural Statistics*, 3, 169–174.

Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62, 1349–1382.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5, 99–135.

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

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

Robins, J. M., & Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16, 285–319.

Robins, J. M., & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In N. Jewell, K. Dietz, & V. T. Farewell (Eds.), *AIDS epidemiology: Methodological issues* (pp. 297–331). Birkhliuser.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of American Statistical Association*, 90, 122–129.

Robins, J. M., Rotnitzky, A., & Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association*, 427, 846–866.

Robins, J. M., Rotnitzky, A., & 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., & Firpo, S. (2019). Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. *Econometric Theory*, 35, 1048–1087.

Rotnitzky, A., & Robins, J. M. (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., & 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*. Springer.

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

Wooldridge, J. M. (2002). Inverse probability weighted *M*-estimation for sample selection, attrition, and stratification. *Portuguese Economic Journal*, 1, 117–139.

Wooldridge, J. M. (2010). *Econometric analysis of cross section & panel data*. MIT Press.