




## Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)

Bryan S. Graham, Cristine Campos de Xavier Pinto & Daniel Egel


To cite this article: Bryan S. Graham, Cristine Campos de Xavier Pinto & Daniel Egel (2016) Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST), Journal of Business & Economic Statistics, 34:2, 288-301, DOI: [10.1080/07350015.2015.1038544](https://doi.org/10.1080/07350015.2015.1038544)

To link to this article: <http://dx.doi.org/10.1080/07350015.2015.1038544>


 View supplementary material [↗](#)

 Accepted author version posted online: 07 May 2015.  
Published online: 17 Mar 2016.

 Submit your article to this journal [↗](#)

 Article views: 61

 View Crossmark data [↗](#)

 Citing articles: 1 View citing articles [↗](#)

# Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)

**Bryan S. GRAHAM**

Department of Economics, University of California-Berkeley, Berkeley, CA 94720-3880 and NBER  
([bgraham@econ.berkeley.edu](mailto:bgraham@econ.berkeley.edu))

**Cristine Campos de Xavier PINTO**

Escola de Economia de São Paulo, FGV, 01332-000 SP, Brazil ([cristinepinto@gmail.com](mailto:cristinepinto@gmail.com))

**Daniel EGEL**

RAND Corporation, Santa Monica, CA 90401-3208 ([Daniel.Egel@rand.org](mailto:Daniel.Egel@rand.org))

We propose a locally efficient estimator for a class of semiparametric data combination problems. A leading estimand in this class is the average treatment effect on the treated (ATT). Data combination problems are related to, but distinct from, the class of missing data problems with data missing at random (of which the average treatment effect (ATE) estimand is a special case). Our estimator also possesses a double robustness property. Our procedure may be used to efficiently estimate, among other objects, the ATT, the two-sample instrumental variables model (TSIV), counterfactual distributions, poverty maps, and semiparametric difference-in-differences. In an empirical application, we use our procedure to characterize residual Black–White wage inequality after flexibly controlling for “premarket” differences in measured cognitive achievement. Supplementary materials for this article are available online.

**KEY WORDS:** Average treatment effect on the treated (ATT); Double robustness; Earnings decompositions; Propensity score; Semiparametric difference-in-differences; Two-sample instrumental variables (TSIV).

## 1. INTRODUCTION

Let  $Z = (W', X', Y)'$  denote a random vector drawn from some *study population* of interest with distribution function  $F_s$ . For some unique  $\gamma_0$ , and known function  $\psi(z, \gamma)$  of the same dimension, we assume that

$$\mathbb{E}_s[\psi(Z, \gamma_0)] = 0, \quad (1)$$

where  $\mathbb{E}_s[\cdot]$  denotes expectations taken with respect to the study population. If a random sample of  $Z$  is available, then consistent estimation of  $\gamma_0$  (under regularity conditions) is straightforward (e.g., Newey and McFadden 1994). Many statistical models of interest can be represented in terms of moment restrictions like (1); see Wooldridge (2002) for a textbook exposition.

In this article, we consider estimation of  $\gamma_0$  when a random sample of  $Z$  is unavailable. Instead two separate samples are available. The first is drawn from the study population and contains  $N_s$  measurements of  $(Y, W)$ . The second is drawn from an *auxiliary population* (with distribution function  $F_a$ ;  $\mathbb{E}_a[\cdot]$  denotes expectations taken with respect to this distribution) and contains  $N_a$  measurements of  $(X, W)$ . While the variable  $W$  is common to the two-samples,  $X$  and  $Y$  are not. Hahn (1998) and Chen, Hong, and Tarozi (2008) showed that identification of  $\gamma_0$  follows if (i) the conditional distributions of  $X$  given  $W$  in the two populations coincide (although their marginal distributions for  $W$  may differ), (ii) the support of  $W$  in the auxiliary population is at least as large as that in the study population, and (iii)  $\psi(z, \gamma_0)$  is separable in the components depending on the “noncommon” variables  $Y$  and  $X$

$$\psi(Z, \gamma_0) = \psi_s(Y, W, \gamma_0) - \psi_a(X, W, \gamma_0). \quad (2)$$

Examples of statistical problems to which the above setup applies include the two-sample instrumental variables (TSIV) model of Angrist and Krueger (1992) and Ridder and Moffitt (2007), the average treatment effect on the treated (ATT) estimand from the program evaluation literature (e.g., Heckman and Robb 1985; Imbens 2004), counterfactual earnings/wealth decompositions as in Dinardo, Fortin, and Lemieux (1996) and Barsky et al. (2002), poverty mapping as in Elbers, Lanjouw, and Lanjouw (2003) and Tarozi and Deaton (2009), direct standardization methods used in demography (e.g., Kitagawa 1964), and models with mismeasured regressors and validation samples (e.g., Carroll and Wand 1991).

To help fix ideas consider the ATT example. Here,  $Y$  denotes an individual’s potential outcome under active treatment, say earnings given participation in a job training program,  $X$  denotes her outcome under control (earnings in the absence of training) and  $W$  is a vector of baseline covariates. Available is a random sample of  $(Y, W)$  from the population assigned active treatment (i.e., “the treated”). A separate sample of measurements of  $(X, W)$  is drawn from a population of controls. The ATT,  $\gamma_0 = \mathbb{E}_s[Y - X]$ , is given by the solution to (1) with  $\psi_s(Y, W, \gamma_0) = Y$  and  $\psi_a(X, W, \gamma_0) = X + \gamma_0$ .

Dehejia and Wahba (1999), revisiting earlier work by LaLonde (1986), combined two distinct samples to estimate the effect of the National Supported Work (NSW) demonstration, a labor training program, on the post-intervention earn-

ings of trainees. Their study sample consists of 185 NSW participants, while their auxiliary sample includes 2490 non-participants drawn from the Panel Study of Income Dynamics (PSID). These two-samples consist of random draws from distinct, nonoverlapping, populations. The two-sample feature of their analysis distinguishes it from one seeking to estimate a population average treatment effect (ATE). In that case, the researcher generally bases her analysis on a random sample from the population of interest, where some units happen to be treated, and others not (e.g., Rosenbaum and Rubin 1983). There the inferential problem is usefully conceptualized as one of missing data and the general theory of Robins, Rotnitzky, and Zhao (1994) directly applies.

*Relationship Between Data Combination and Missing Data Problems.* One perspective is that data combination problems are nothing more than a particular class of “missing data” problems in which the auxiliary sample is collected independently, and from a different population than that, of the study sample. Our use of the term missing data is more technical, referring, in particular, to the family of problems analyzed by Robins, Rotnitzky, and Zhao (1994, sec. 8.1). In this family, both the study and auxiliary samples are random ones from the population of interest. It turns out that this difference has statistical content with, as we emphasize here (and others have before us), implications for estimator formulation and properties. In an important article, Hahn (1998) showed that while prior restrictions on the form of the propensity score do not lower the semiparametric variance bound for the ATE, they do lower the corresponding bound for the ATT. Chen, Hong, and Tarozzi (2008) generalized this result, showing that, unlike in the missing data context (their “verify-in-sample” case), knowledge of the form of the propensity score is asymptotically valuable in data combination problems (their “verify-out-of-sample” case).

Our contribution is to develop a flexible parametric estimator for general data combination problems with good efficiency and robustness properties. Similar to the augmented inverse probability weighting (AIPW) estimator for missing data problems due to Robins, Rotnitzky, and Zhao (1994), our data combination procedure is locally efficient and possesses a double robustness property. This latter property, given the nonancillarity of the propensity score in the data combination problem, is surprising.

To our knowledge, we are the first to propose a locally efficient estimator in the data combination context. Chen, Hong, and Tarozzi (2008) proposed a globally efficient estimator, but their procedure requires nonparametric modeling as opposed to the flexible parametric approach adopted here. Our methods provide a practical alternative to theirs when  $W$  is high dimensional (see Firpo and Rothe 2013). Abadie (2005) developed a parametric propensity score reweighting (PSR) estimate of the ATT. Qin and Zhang (2008) showed that Abadie’s estimator can have low efficiency in some settings and proposed an alternative that uses empirical likelihood ideas. Qin and Zhang (2008) did not characterize the semiparametric efficiency or robustness properties of their ATT estimator, nor showed how to extend it to the wider class of problems considered here. Hirano and Imbens (2001) also proposed a type of propensity score reweighting estimator for the ATT. Their estimator exhibits a double robustness property, but they did not consider issues of semiparametric efficiency nor general data combination prob-

lems as we do. Besides its robustness and efficiency properties, our estimator is simple to compute and is suitable for many applied problems, like the estimation of the ATT, two-sample instrumental variables and others cited above.

In Section 2, we define the semiparametric data combination model. Modestly extending the work of Chen, Hong, and Tarozzi (2008), we calculate the semiparametric efficiency bound for our model. We relate our efficiency bound analysis to prior work on distribution function estimation based on a random sample from the population of interest and a second, biased, sample from the same population (e.g., Qin 1998; Gilbert, Lele, and Vardi 1999). This discussion motivates the form of our AST estimator, which we introduce in Section 3, where we also formally characterize its large sample properties. Our key results are Theorems 2 to 4. Section 4 provides an illustrative empirical application and reports on the results of several Monte Carlo experiments. Proofs of our main results are contained in the Appendix. The supplemental Web appendix contains additional proof details, extra examples of data combination problems, and additional Monte Carlo results. An algorithm for computing our estimator, that we have found to work well in practice, is also described in the supplemental Web appendix.

## 2. SEMIPARAMETRIC DATA COMBINATION MODEL

A formal definition of the data combination model is given by Assumption 1. Let  $A \in \mathbb{R}^P$  denote a compact subset of  $\mathbb{R}^P$ .

*Assumption 1. Semiparametric Data Combination Model*

- (i) (*Identification*). For some  $\psi(z, \gamma) = \psi_s(y, w, \gamma) - \psi_a(x, w, \gamma)$ , Equation (1) holds with  $\mathbb{E}_s[\psi(Z, \gamma)] \neq 0$  for all  $\gamma \neq \gamma_0$ ,  $\gamma \in \mathcal{G} \in \mathbb{R}^K$ ,  $z \in \mathcal{Z} \in \mathbb{R}^{\dim(\mathcal{Z})}$ .
- (ii) (*Conditional distributional equality*).  $F_s(x|w) = F_a(x|w)$  and  $F_s(y|w) = F_a(y|w)$  for all  $w \in \mathcal{W} \in \mathbb{R}^{\dim(W)}$ ,  $x \in \mathcal{X} \in \mathbb{R}^{\dim(X)}$ , and  $y \in \mathcal{Y} \in \mathbb{R}^{\dim(Y)}$ .
- (iii) (*Weak overlap*). Let  $S_j = \{w : f_j(w) > 0\}$  for  $j = s, a$ , then  $S_s \subset S_a$ .
- (iv) (*Multinomial sampling*). With probability  $Q_0 \in (\xi, 1 - \xi)$  for  $0 < \xi < 1$ , we draw a unit at random from  $F_s$  and record its realizations of  $Y$  and  $W$ , otherwise we draw a unit at random from  $F_a$  and record its realizations of  $X$  and  $W$ . Let  $D_i = 1$  if the  $i$ th draw ( $i = 1, \dots, N$ ) corresponds to a study population unit and  $D_i = 0$  otherwise.
- (v) (*Propensity score model*). There is a unique  $\delta_0 \in \mathcal{D} \in \mathbb{R}^{\dim(\delta)}$ , known vector  $r(W)$  of linearly independent functions of  $W$  with a constant in the first row, and known function  $G(\cdot)$  such that (a)  $G(\cdot)$  is strictly increasing, differentiable, and maps into the unit interval with  $\lim_{v \rightarrow -\infty} G(v) = 0$  and  $\lim_{v \rightarrow \infty} G(v) = 1$ , (b)  $\frac{f_s(w)}{f_a(w)} = \frac{1-Q_0}{Q_0} \frac{G(r(w)'\delta_0)}{1-G(r(w)'\delta_0)}$  for all  $w \in \mathcal{W}$ , and (c)  $0 < G(r(w)'\delta) \leq \kappa < 1$  for all  $\delta \in \mathcal{D}$  and  $w \in \mathcal{W}$ .

The first part of Assumption 1 implies global identifiability of the complete data model. The second part implies that the distributions of  $(Y, W)$  and  $(X, W)$  in the two populations differ only in terms of their marginal distributions for the always

measured variable,  $W$ . The third part ensures that, in large samples, for each unit in the study sample there will be matching units with similar values of  $W$  in the auxiliary sample. The fourth part of Assumption 1 allows us to treat the *merged sample*

$$\left\{ (D_i, W_i, (1 - D_i) X_i', D_i Y_i') \right\}_{i=1}^N,$$

“as if” it were a random one from a pseudo *merged population* with distribution function  $F$  (let  $\mathbb{E}[\cdot]$  denote expectations taken with respect to this distribution). The semiparametric data combination model is typically defined by specifying properties of the merged population (e.g., Hahn 1998; Chen, Hong, and Tarozzi 2008). We prefer the formulation given above because it (i) emphasizes that the problem is fundamentally one of combining two datasets and (ii) in many applications the merged population does not correspond a real world population. Neither (i) or (ii) are features of standard missing data problems (i.e., Robins, Rotnitzky, and Zhao 1994). We also note that formulating a model by imposing restrictions on a pseudo-population is somewhat awkward (see the discussion in Abadie and Imbens 2006, p. 239).

The sampling distribution induced by the multinomial scheme,  $F$ , has density

$$f(z, d) = Q_0^d (1 - Q_0)^{1-d} f_s(z)^d f_a(z)^{1-d},$$

such that  $f(z|d = 1) = f_s(z)$  and  $f(z|d = 0) = f_a(z)$ . Now consider the conditional probability given  $W = w$  that a unit in the merged sample corresponds to a draw from the study population. Let  $\mathbb{E}[D|W = w] = p_0(w)$  denote this “propensity score,” by Bayes’ Law we can define a relationship between the study and auxiliary densities of  $W$  in terms of  $p_0(w)$

$$f_s(w) = f_a(w) \left\{ \frac{1 - Q_0}{Q_0} \frac{p_0(w)}{1 - p_0(w)} \right\}. \quad (3)$$

Under the merged population formulation of the problem it is clear that part (i) of Assumption 1 corresponds to requiring that  $\mathbb{E}[\psi(Z, \gamma_0)|D = 1] = 0$ , part (ii) to conditional independence restrictions on the merged population distribution function of  $F(y|w, d = 1) = F(y|w, d = 0)$  and  $F(x|w, d = 1) = F(x|w, d = 0)$ , and parts (iii) and (iv) to assuming that  $p_0(w)$  is bounded away from one. Part (v) implies that the density ratio  $f_s(w)/f_a(w)$  takes a parametric form or, equivalently, that the propensity score is known up to a finite dimensional parameter.

Identification of  $\gamma_0$  follows from, using parts (ii) and (iii) of Assumption 1 and Equation (3), the equality

$$\mathbb{E}_s[\psi(Z, \gamma)] = \mathbb{E} \left[ \frac{D}{Q_0} \psi_s(Y, W, \gamma) - \mathbb{E} \left[ \frac{1 - D}{Q_0} \frac{p_0(W)}{1 - p_0(W)} \psi_a(X, W, \gamma) \right] \right], \quad (4)$$

which is, by part (i) of Assumption 1, uniquely zero at  $\gamma = \gamma_0$ . See Lemma 3.1 of Abadie (2005) for a formal proof.

### 2.1 Example: Two-Sample Instrumental Variables (TSIV)

To give some idea of the range of problems to which our methods apply, we elaborate on one common data combination problem in detail: the two-sample instrumental variable model

(TSIV). This model is widely used by empirical researchers in economics (see Inoue and Solon 2010). Our observation that TSIV is a special case of the model defined by Assumption 1 is a new one, with empirically relevant implications. In particular, the auxiliary-to-study (AST) estimator we propose below is both (i) more efficient and (ii) consistent under a wider, and empirically relevant, set of assumptions, than, for example, the estimators of Angrist and Krueger (1992) and Ridder and Moffitt (2007).

Additional examples of data combination problems are outlined in the supplemental Web appendix. Chen, Hong, and Tarozzi (2008), Ridder and Moffitt (2007), and Abadie (2005) provided further examples.

Following Ridder and Moffitt (2007), consider two-sample instrumental variables (TSIV) models of the form

$$\mathbb{E}_s[\{f(Y; \gamma) - g(X, W_1; \gamma)\} e(W)] = 0,$$

with  $W = (W_0', W_1')$ . The first sample consists of measurements of  $(Y, W)$  and the second of  $(X, W)$ . They assume that both samples are random ones from the study population (i.e., the samples are “compatible”). This corresponds to augmenting Assumption 1 with the additional requirement that  $F_s(w) = F_a(w)$ . The TSIV model is of the form required by (2) with  $\psi_s(y, w, \gamma) = f(Y; \gamma)e(W)$  and  $\psi_a(x, w, \gamma) = g(X, W_1; \gamma)e(W)$ . When  $e(W) = W$ ,  $f(Y; \gamma) = Y$  and  $g(X, W_1; \gamma) = X'\alpha + W_1'\beta$  with  $\gamma_0 = (\alpha_0, \beta_0)'$  we have the linear model analyzed by Angrist and Krueger (1992). Ridder and Moffitt (2007) showed how one may estimate the mixed proportional hazard (MPH) model under this setup, while Ichimura and Martinez-Sanchis (2004) discussed binary choice models.

A concrete example of a TSIV problem is provided by the work of Currie and Yelowitz (2000), who considered the model  $\mathbb{E}_s[W(Y - X'\alpha_0 - W_1'\beta_0)] = 0$ , where  $Y$  is an indicator for whether a school-aged child has repeated a grade,  $X$  an indicator for residence in public housing,  $W_0$  equals the number of male siblings in the household, and  $W_1$  equals the overall number of siblings and also contains other household characteristics;  $W = (W_0', W_1')$ . Their interest centers on the causal effect of residence in public housing on human capital acquisition. The number of male siblings changes the probability of residence in public housing since, conditional on the overall number of siblings, families with a mixture of boys and girls qualify for larger units and hence higher (implicit) housing subsidies. Currie and Yelowitz (2000) additionally argued that, conditional on the total number of one’s siblings, their gender mix should not influence schooling independently of any effect mediated by exposure to public housing. Hence,  $W_0$  may serve as an instrumental variable for  $X$ .

Currie and Yelowitz (2000) observed  $Y$  and  $W$  for a random subsample of children drawn from the U.S. Census. The Census, however, does not collect information on residence in public housing,  $X$ . This information is available in the U.S. Current Population Survey (CPS), which also includes measurements of  $W$  (but not  $Y$ ). They treat both the Census and CPS samples as random ones from their study population (school-aged children living in the United States) and use a variant of Angrist and Krueger’s (1992) method to estimate  $\gamma_0 = (\alpha_0, \beta_0)'$ .

In applications of the TSIV model, like Currie and Yelowitz's (2000), it is often found that the sample moments of the common variables  $W$  differ significantly across the two datasets being combined (see also Björklund and Jäntti 1997). This suggests that full compatibility may fail in practice (i.e.,  $F_s(w) \neq F_a(w)$ ). The estimator presented below does not require full compatibility and is generally more efficient than the one proposed by Angrist and Krueger (1992) (compare Theorems 2 and 3 with Angrist and Krueger (1992, p. 331) or Ridder and Moffitt (2007, p. 5505)).

## 2.2 Efficiency Bound

Hahn (1998, Theorem 1) calculated the semiparametric variance bound for the special case where  $\gamma_0$  is the ATT and part (v) of Assumption 1 is not part of the prior restriction. Chen, Hong, and Tarozzi (2008, Theorem 3) included part (v) in their prior, but assume that  $\psi_s(Y, W, \gamma) = 0$ . The following result generalizes that of Chen, Hong, and Tarozzi (2008) to the case where the moment function is of the form given in (2). To present this result, we require some additional notation. Let  $\mathbb{E}^*[Y|X]$  denote the mean squared error minimizing linear predictor of  $Y$  given  $X$  and define

$$\begin{aligned}\Gamma_0(w) &= \mathbb{E} \left[ \frac{\partial \psi(Z, \gamma_0)}{\partial \gamma'} \Big| W = w \right], \\ \Gamma_0 &= \mathbb{E}[\Gamma_0(W)], \quad p_0(w) = G(r(w)' \delta_0) \\ q_s(w) &= \mathbb{E}[\psi_s(Y, W, \gamma_0) | W = w], \\ q_a(w) &= \mathbb{E}[\psi_a(X, W, \gamma_0) | W = w] \\ \Sigma_s(w; \gamma_0) &= \mathbb{V}(\psi_s(Y, W, \gamma_0) | W = w), \\ \Sigma_a(w; \gamma_0) &= \mathbb{V}(\psi_a(X, W, \gamma_0) | W = w) \\ \mathbb{S}_\delta &= \frac{D - G(r(W)' \delta_0)}{G(r(W)' \delta_0)[1 - G(r(W)' \delta_0)]} G_1(r(W)' \delta_0) r(W)\end{aligned}$$

with  $G_1(v) = \partial G(v)/\partial v$  and

$$\begin{aligned}\Lambda(W) &= \left( \frac{p_0(W)}{Q_0} \right)^2 \left\{ \frac{\Sigma_s(W; \gamma_0)}{p_0(W)} + \frac{\Sigma_a(W; \gamma_0)}{1 - p_0(W)} \right. \\ &\quad \left. + [q_s(W) - q_a(W)][q_s(W) - q_a(W)]' \right\} \\ &\quad + \mathbb{E} \left[ \left( \frac{D}{p_0(W)} - 1 \right) \frac{p_0(W) \{q_s(W) - q_a(W)\}}{Q_0} \mathbb{S}'_\delta \right] \\ &\quad \times \mathbb{E}[\mathbb{S}_\delta \mathbb{S}'_\delta]^{-1} \mathbb{E} \left[ \left( \frac{D}{p_0(W)} - 1 \right) \right. \\ &\quad \left. \times \frac{p_0(W) \{q_s(W) - q_a(W)\}}{Q_0} \mathbb{S}'_\delta \right]'. \quad (5)\end{aligned}$$

*Theorem 1 (Semiparametric Variance Bound).* Under Assumption 1(i) the maximal asymptotic precision with which  $\gamma_0$  may be regularly estimated is given by the inverse of  $\mathcal{I}(\gamma_0) = \Gamma_0' \mathbb{E}[\Lambda(W)]^{-1} \Gamma_0$  and (ii) the efficient influence

function is

$$\begin{aligned}\phi^{\text{eff}}(Z, \gamma_0) &= -\Gamma_0^{-1} \times \left[ \frac{D}{Q_0} \{\psi_s(Y, W, \gamma_0) - q_s(W)\} \right. \\ &\quad \left. - \frac{1 - D}{Q_0} \frac{p_0(W)}{1 - p_0(W)} \{\psi_a(X, W, \gamma_0) - q_a(W)\} \right. \\ &\quad \left. + \frac{p_0(W)}{Q_0} \{q_s(W) - q_a(W)\} \right. \\ &\quad \left. + \frac{1}{Q_0} \mathbb{E}^* \left[ \left( \frac{D}{p_0(W)} - 1 \right) \right. \right. \\ &\quad \left. \left. \times p_0(W) \{q_s(W) - q_a(W)\} \mathbb{S}_\delta \right] \right]. \quad (6)\end{aligned}$$

*Proof.* The proof, which involves a modest extension of the analysis of Chen, Hong, and Tarozzi (2008, Theorem 3), is in the supplemental Web appendix.  $\square$

It is easy to show that the information bound for  $\gamma_0$  is smaller in the model, which leaves  $p_0(W)$  nonparametric (i.e., where part (v) of Assumption 1 is not part of the prior). Knowledge of the parametric form of the propensity score increases the large sample precision with which  $\gamma_0$  may be estimated. In contrast, in semiparametric missing data problems, it is well known that parametric restrictions on the propensity score do not shift the efficiency bound (e.g., Robins, Rotnitzky, and Zhao 1994; Hahn 1998). The value of prior restrictions on the propensity score distinguishes the data combination problem from the missing data one.

To understand this difference, we use the well-known result that a biased sample may be combined with a random one to form a more efficient distribution function estimate as long as the biasing function is known or parametrically specified. Parts (v) of Assumption 1 implies that we can view the auxiliary sample as a biased sample from the study population of interest where the biasing function is known up to a finite-dimensional parameter (see Qin 1998; Gilbert, Lele, and Vardi 1999; Ridder and Moffitt 2007).

Here, and in what follows, we assume without loss of generality that the merged sample is arranged such that its first  $N_s$  units correspond to study population draws, and its remaining  $N_a$  units to auxiliary sample draws. Let  $G(r(w)'\hat{\delta}_{\text{ML}})$  denote the conditional maximum likelihood estimate of the propensity score (based on the merged sample), then the estimate

$$\begin{aligned}\hat{F}_s^{\text{eff}}(w) &= \sum_{i=1}^N \hat{\pi}_i^{\text{eff}} \mathbf{1}(W_i \leq w), \\ \hat{\pi}_i^{\text{eff}} &= \frac{G(r(W_i)'\hat{\delta}_{\text{ML}})}{\sum_{i=1}^N G(r(W_i)'\hat{\delta}_{\text{ML}})} \quad (7)\end{aligned}$$

efficiently uses the information in both the study and auxiliary samples to estimate  $F_s(w)$ . To understand (7) note that Bayes' law gives  $f_s(W_i) = f(W_i | D_i = 1) = p_0(W_i) f(W_i) / Q_0$ ; replacing  $p_0(W_i)$  and  $Q_0$  with their maximum likelihood estimates, and  $f(W_i)$  with the empirical measure of the merged sample,  $1/N$ , gives  $\hat{f}_s(W_i) = \hat{\pi}_i^{\text{eff}}$ , for  $\hat{\pi}_i^{\text{eff}}$  defined in (7). Equation (7) uses both study and auxiliary units—linked via a parametric form for the propensity score—to efficiently estimate  $F_s(w)$ .

In contrast, in missing data problems the population of interest corresponds to what we have termed the merged population.

The most efficient estimate of the merged population distribution function of  $W$  is the merged sample empirical distribution function. This is true irrespective of the form of the propensity score. This provides one intuition for why prior knowledge of the form of the propensity score is not valuable in the missing data context (see Graham 2011).

### 3. AUXILIARY-TO-STUDY TILTING

In this section, we present our AST estimator and characterize its large sample properties under different sets of assumptions. Since the parameter of interest,  $\gamma_0$ , involves integration over the study population distributions of  $(Y, W)$  and  $(X, W)$ , these two distribution functions must be (implicitly) estimated to estimate  $\gamma_0$ . The AST estimator uses distribution function estimates that share a finite number of moments of  $W$  in common with  $\widehat{F}_s^{\text{eff}}(w)$ . That is, we calibrate our estimates of the study population distributions of  $(Y, W)$  and  $(X, W)$  to features of (7) (which is a semi-parametrically efficient estimate of  $F_s(w)$  when the propensity score takes a parametric form). This, as we explain below, is the source of the efficiency gains associated with our procedure.

The idea of calibrating a distribution function estimate to information garnered from auxiliary sources arises in other contexts. Little and Wu (1991) discussed contingency table calibration to known margins and provided historical references (see Hellerstein and Imbens 1999). Bickel, Ritov, and Wellner (1991) studied estimation of linear functionals of probability measures with known marginals. Hirano et al. (2001) showed how calibration to marginal information from refreshment samples may be used to correct for certain types of nonignorable attrition in panel data. In the context of average treatment effect estimation, Tan (2006) calibrated estimates of the two potential outcome distributions to features of the empirical distribution of always observed variables (see Qin and Zhang 2007; Graham, de Xavier Pinto, and Egel 2012). Recently, Cheng et al. (2009) applied related ideas to an instrumental variables model.

#### 3.1 Outline of the AST Estimator

Our estimator for  $\gamma_0$ , which we call the AST estimator, is a sequential method of moments estimator. In the first step, we estimate the propensity score parameter  $\delta$  by conditional maximum likelihood:

$$\frac{1}{N} \sum_{i=1}^N \frac{D_i - G(r(W_i)' \widehat{\delta}_{ML})}{G(r(W_i)' \widehat{\delta}_{ML}) [1 - G(r(W_i)' \widehat{\delta}_{ML})]} \times G_1(r(W_i)' \widehat{\delta}_{ML}) r(W_i) = 0. \tag{8}$$

In the second step, we compute a reweighting of both the study and auxiliary samples. Let  $t(W)$  be vector of known linearly independent functions of  $W$  with a constant 1 in the first row and  $\lambda_a$  and  $\lambda_s$  be “tilting” parameters of the same dimension. We allow for  $r(W)$  and  $t(W)$  to include common elements or even coincide. Fixing  $\delta$  at  $\widehat{\delta}_{ML}$  and  $Q$  at  $\widehat{Q}_{ML}$ , we choose  $\widehat{\lambda}_a$  to

solve

$$\frac{1}{N} \sum_{i=1}^N \left( \frac{1 - D_i}{1 - G(r(W_i)' \widehat{\delta}_{ML} + t(W_i)' \widehat{\lambda}_a)} - 1 \right) \times \frac{G(r(W_i)' \widehat{\delta}_{ML})}{\widehat{Q}_{ML}} t(W_i) = 0. \tag{9}$$

To understand this method of choosing  $\widehat{\lambda}_a$ , it is helpful to rearrange (9) to get

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \frac{1 - D_i}{\widehat{Q}_{ML}} \frac{G(r(W_i)' \widehat{\delta}_{ML}) t(W_i)}{1 - G(r(W_i)' \widehat{\delta}_{ML} + t(W_i)' \widehat{\lambda}_a)} \\ = \frac{1}{N} \sum_{i=1}^N \frac{G(r(W_i)' \widehat{\delta}_{ML}) t(W_i)}{\widehat{Q}_{ML}} \\ \sum_{i=N_s+1}^N \widehat{\pi}_i^a t(W_i) = \sum_{i=1}^N \widehat{\pi}_i^{\text{eff}} t(W_i), \end{aligned} \tag{10}$$

for

$$\widehat{\pi}_i^a = \frac{G(r(W_i)' \widehat{\delta}_{ML})}{\sum_{i=1}^N G(r(W_i)' \widehat{\delta}_{ML})} \frac{1}{1 - G(r(W_i)' \widehat{\delta}_{ML} + t(W_i)' \widehat{\lambda}_a)}, \quad i = N_s + 1, \dots, N,$$

and where the second line of (10) is equivalent to the first. The term to the right of the equality in (10) is an estimate of  $\mathbb{E}_s[t(W_i)]$ —the study population mean of  $t(W_i)$ —based on the *efficient* distribution function estimate (7). It is consequently an efficient estimate of  $\mathbb{E}_s[t(W_i)]$ . The solution to (9)—our estimate of  $\lambda_a$ —is chosen to form a reweighting of the auxiliary sample such that  $\sum_{i=1}^{N_s} \widehat{\pi}_i^a t(W_i)$  is *numerically identical* to the efficient estimate of  $\mathbb{E}_s[t(W_i)]$  based on  $\widehat{F}_s^{\text{eff}}(w)$ .

To better understand (10) recall that, as shown by Abadie (2005) and others, the propensity score reweighting-type estimator

$$\widehat{F}_s^{\text{PSR}}(x, w) = \frac{1}{N} \sum_{i=1}^N \frac{1 - D_i}{\widehat{Q}_{ML}} \frac{G(r(W_i)' \widehat{\delta}_{ML})}{1 - G(r(W_i)' \widehat{\delta}_{ML})} \times \mathbf{1}(X_i \leq x, W_i \leq w),$$

is consistent for the study population distribution function of  $(X, W)$ . Our AST estimator replaces  $\widehat{F}_s^{\text{PSR}}(x, w)$  with the more efficient tilted version

$$\widehat{F}_s^{\text{AST}}(x, w) = \sum_{i=N_s+1}^N \widehat{\pi}_i^a \mathbf{1}(X_i \leq x, W_i \leq w).$$

This tilted distribution estimate, unlike  $\widehat{F}_s^{\text{PSR}}(x, w)$ , is guaranteed to integrate to one and shares a finite number of moment in common with  $\widehat{F}_s^{\text{eff}}(w)$ .

We also compute an analogous tilt of the study sample

$$\frac{1}{N} \sum_{i=1}^N \left( \frac{D_i}{G(r(W_i)' \widehat{\delta}_{ML} + t(W_i)' \widehat{\lambda}_s)} - 1 \right) \times \frac{G(r(W_i)' \widehat{\delta}_{ML})}{\widehat{Q}_{ML}} t(W_i) = 0, \tag{11}$$

so that

$$\sum_{i=1}^{N_s} \widehat{\pi}_i^s t(W_i) = \sum_{i=1}^N \widehat{\pi}_i^{\text{eff}} t(W_i), \quad (12)$$

for

$$\widehat{\pi}_i^s = \frac{G(r(W_i)' \widehat{\delta}_{\text{ML}})}{\sum_{i=1}^N G(r(W_i)' \widehat{\delta}_{\text{ML}})} \frac{1}{G(r(W_i)' \widehat{\delta}_{\text{ML}} + t(W_i)' \widehat{\lambda}_s)},$$

$$i = 1, \dots, N_s.$$

With the auxiliary and study sample tilts in hand, we then choose  $\widehat{\gamma}_{\text{AST}}$  to solve, holding  $\widehat{\lambda}_a$  and  $\widehat{\lambda}_s$  fixed at their second step values,

$$\sum_{i=1}^{N_s} \widehat{\pi}_i^s \psi_s(Y_i, W_i, \widehat{\gamma}_{\text{AST}}) - \sum_{i=N_s+1}^N \widehat{\pi}_i^a \psi_a(X_i, W_i, \widehat{\gamma}_{\text{AST}}) = 0. \quad (13)$$

Inspection of (13) indicates that our estimate of  $\gamma_0$  is based on two separate estimates of the study population distribution function. The first, corresponding to the *study tilt*  $\{\widehat{\pi}_i^s\}_{i=1}^{N_s}$  is an estimate of the study population distribution of  $(Y_i, W_i)$ , the second, corresponding to the *auxiliary tilt*,  $\{\widehat{\pi}_i^a\}_{i=N_s+1}^N$ , is an estimate of the study population distribution of the  $(X_i, W_i)$ . Neither of these two estimates coincide with the efficient estimate of the study population distribution of  $W_i$  alone (i.e., with (7)), but they do share important features with it. Specifically, they are constructed so that the means of  $t(W_i)$ , computed using the two tilts, coincide with the efficient estimate.

### 3.2 Large-Sample Properties

Our next three results provide formal descriptions of the asymptotic sampling properties of  $\widehat{\gamma}_{\text{AST}}$  under different combinations of assumptions. We begin with a characterization of the sampling properties of  $\sqrt{N}(\widehat{\gamma}_{\text{AST}} - \gamma_0)$  under our baseline model (i.e., Assumption 1). We then outline our local semiparametric efficiency and double robustness results.

To state our first result, we require some additional notation. Let

$$q_a^*(W) = \Pi_a^* t(W), \quad q_s^*(W) = \Pi_s^* t(W),$$

be weighted projections of  $\psi_s(Y, W, \gamma_0)$  and  $\psi_a(X, W, \gamma_0)$  onto the space spanned by  $t(W)$ , with projection coefficients of

$$\begin{aligned} \Pi_s^* &= \mathbb{E} \left[ G_1(r(W)' \delta_0) q_s(W) t(W)' \right] \\ &\quad \times \mathbb{E} \left[ G_1(r(W)' \delta_0) t(W) t(W)' \right]^{-1}, \\ \Pi_a^* &= \mathbb{E} \left[ \frac{p_0(W)}{1-p_0(W)} G_1(r(W)' \delta_0) q_a(W) t(W)' \right] \\ &\quad \times \mathbb{E} \left[ \frac{p_0(W)}{1-p_0(W)} G_1(r(W)' \delta_0) t(W) t(W)' \right]^{-1}. \end{aligned} \quad (14)$$

Also define  $R(D, W) = R_s(D, W) - R_a(D, W)$  with

$$R_s(D, W) = \frac{1}{Q_0} \left\{ \left( \frac{D}{p_0(W)} - 1 \right) p_0(W) \{q_s^*(W) - q_s(W)\} \right. \quad (15)$$

$$\left. - \mathbb{E}^* \left[ \left( \frac{D}{p_0(W)} - 1 \right) p_0(W) \{q_s^*(W) - q_s(W)\} \middle| \mathbb{S}_\delta \right] \right\}$$

$$R_a(D, W) = \frac{1}{Q_0} \left\{ \left( \frac{1-D}{1-p_0(W)} - 1 \right) p_0(W) \{q_a^*(W) - q_a(W)\} \right. \quad (16)$$

$$\left. - \mathbb{E}^* \left[ \left( \frac{1-D}{1-p_0(W)} - 1 \right) p_0(W) \{q_a^*(W) - q_a(W)\} \middle| \mathbb{S}_\delta \right] \right\}.$$

**Theorem 2 (Asymptotic Distribution).** Suppose that Assumption 1 and additional regularity conditions hold, then (i)  $\widehat{\gamma}_{\text{AST}} \xrightarrow{P} \gamma_0$ , (ii)

$$\sqrt{N}(\widehat{\gamma}_{\text{AST}} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, I(\gamma_0)^{-1} + \Sigma_{\text{RR}}(\gamma_0)),$$

with

$$\Sigma_{\text{RR}}(\gamma_0) = \Gamma_0^{-1} \mathbb{E} [R(D, W) R(D, W)'] \Gamma_0^{-1'}$$

and (iii) the asymptotic efficiency of  $\sqrt{N}a'(\widehat{\gamma}_{\text{AST}} - \gamma_0)$ , for any vector of constants  $a$ , is bounded below by

$$\begin{aligned} & \text{ae} \left( \sqrt{N}a'(\widehat{\gamma}_{\text{AST}} - \gamma_0) \right) \\ & \geq \frac{a'I(\gamma_0)^{-1}a}{a'I(\gamma_0)^{-1}a + \frac{\epsilon^2}{Q_0^2} \mathbb{E} \left[ \frac{p_0(W)}{1-p_0(W)} \right] a'(\Gamma_0^{-1}t)(\Gamma_0^{-1}t)'a}, \end{aligned} \quad (17)$$

where  $\epsilon = \max(\epsilon_s, \epsilon_a)$  with

$$\epsilon_s = \sup_{w \in \mathcal{W}} \|q_s(w) - \Pi_s^* t(w)\|_\infty,$$

$$\epsilon_a = \sup_{w \in \mathcal{W}} \|q_a(w) - \Pi_a^* t(w)\|_\infty,$$

and  $\|\cdot\|_\infty$  denoting the maximum absolute row sum norm.

Theorem 2 indicates that under Assumption 1 our AST estimator is consistent and asymptotically normal, but inefficient relative to a semiparametrically efficient estimator (see Theorem 1). Some insight into the degree of AST's inefficiency is provided by the bound (17). First, the term,  $\epsilon^2$ , indicates that the AST estimator performs better when  $q_s(w)$  and  $q_a(w)$  are well approximated by a linear combination of the elements of  $t(w)$ . We discuss the nature of this approximation further below. Second, the performance of the AST estimator will, in general, be sensitive to the degree of overlap. If the expected value of the propensity score weight,  $p_0(W)/(1-p_0(W))$ , used to reweight auxiliary units is large, as may be true if  $\kappa$ , the upper bound on  $p_0(W)$ , is close to one (see part (v) of Assumption 1), then the performance of the AST estimator may be poor (see Khan and Tamer 2010).

More generally the form of  $R(D, W)$  indicates that the relative efficiency of  $\widehat{\gamma}_{\text{AST}}$  depends on the quality of the linear approximations  $q_s(W) \simeq \Pi_s^* t(W)$  and  $q_a(W) \simeq \Pi_a^* t(W)$ . This is easiest to see in the special case where  $r(W) \subset t(W)$ , in which case (see (A.13) in the Appendix), defining  $U_s^* = q_s(W) - \Pi_s^* t(W)$ ,  $U_a^* = q_a(W) - \Pi_a^* t(W)$ , and  $U^* = (1-p_0(W))U_s^* + p_0(W)U_a^*$ :

$$\Sigma_{\text{RR}}(\gamma_0) = \frac{1}{Q_0^2} \Gamma_0^{-1} \mathbb{E} \left[ \frac{p_0(W)}{1-p_0(W)} U^* U^{*'} \right] \Gamma_0^{-1'}$$

so that the degree of inefficiency depends on a (weighted) expectation of the squares and cross products of a linear combination

of the approximation errors. The form of  $\Sigma_{RR}(\gamma_0)$  indicates that  $\hat{\gamma}_{AST}$  will have high relative efficiency whenever  $q_s(W)$  and  $q_a(W)$  are well approximated by a linear combination of the elements of  $t(W)$ . This will be particularly true when overlap is good such that the weight  $\frac{p_0(W)}{1-p_0(W)}$  does not take on extreme values.

Our next result, which characterizes when  $\hat{\gamma}_{AST}$  will be efficient, is anticipated by the discussion above. Consider the assumption:

*Assumption 2 (Moment CEF).* For some unique pair of matrices  $\Pi_s, \Pi_a$ , and vector of linear independent functions  $t(W)$  with a constant in the first row, we have

$$\begin{aligned} \mathbb{E}[\psi_s(Y, W, \gamma_0) | W] &= \Pi_s t(W), \\ \mathbb{E}[\psi_a(X, W, \gamma_0) | W] &= \Pi_a t(W). \end{aligned}$$

Assumption 2 posits a working model for the conditional expectation functions (CEFs) of  $\psi_s(Y, W, \gamma_0)$  and  $\psi_a(X, W, \gamma_0)$  given  $W$ . The substantive content of this assumption is, of course, model and application specific. The ATT example discussed in the introduction provides a simple illustration. In that case, Assumption 2 implies that the CEFs of the potential outcomes given active and control treatment,  $Y$  and  $X$ , are linear in  $t(W)$ . Thus, if the object of interest is the ATT, the analyst should pick the elements of  $t(W)$  so as to provide a good approximation to these two CEFs. For the two-sample instrumental variables (TSIV) model, it is possible to show that the correct  $t(W)$  is an implication of the structure of the first-stage relationship between the endogenous right-hand side variable,  $X$ , and the instrument vector,  $W$ .

If both Assumptions 1 and 2 hold, the Appendix shows that  $\hat{\gamma}_{AST}$  is asymptotically linear with representation

$$\sqrt{N}(\hat{\gamma}_{AST} - \gamma_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \phi^{\text{eff}}(Z_i, \gamma_0) + o_p(1)$$

from which our next theorem directly follows.

*Theorem 3 (Local Semiparametric Efficiency).* Suppose that Assumption 1 and additional regularity conditions hold, then for  $\hat{\gamma}_{AST}$  the solution to (13),  $\hat{\gamma}_{AST}$  is locally efficient at Assumption 2 such that  $\sqrt{N}(\hat{\gamma}_{AST} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \mathcal{I}(\gamma_0)^{-1})$  with  $\mathcal{I}(\gamma_0)$  as defined in Theorem 1.

*Proof.* See the Appendix. □

Our efficiency bound calculation, Theorem 1, gives the information bound for  $\gamma_0$  without imposing the additional auxiliary Assumption 2. This assumption imposes restrictions on the joint distribution of the data not implied by the baseline model. If these restrictions are added to the prior used to calculate the efficiency bound, then it may be possible to estimate  $\gamma_0$  more precisely. Our estimator is not efficient with respect to this augmented model. Rather it attains the bound provided by Theorem 1 if Assumption 2 happens to be true in the population being sampled from, but is not part of the prior restriction used to calculate the bound. Newey (1990, p. 114), Robins, Rotnitzky, and Zhao (1994, p. 852–853), and Tsiatis (2006) discussed the concept of local efficiency in detail. In what follows we will, for brevity, say  $\hat{\gamma}_{AST}$  is locally efficient at Assumption 2. The

form of the variance bound when semiparametric, or parametric (as in Assumption 2), restrictions on  $q_s(w)$  and  $q_a(w)$  are maintained as part of the prior restriction is unknown. Graham (2011) studied such restrictions in the missing data context.

Next we give our double robustness result. Here, our result is slightly less general than similar results in the missing data literature, but nevertheless may be useful in practice.

*Theorem 4 (Double Robustness).* Under parts (i) to (iv) of Assumption 1,  $\hat{\gamma}_{AST} \xrightarrow{P} \gamma_0$  with a limiting normal distribution if either (a) part (v) of Assumption 1 also holds or (b) the analyst chooses  $G(v) = \frac{\exp(v)}{1+\exp(v)}$  and Assumption 2 holds.

*Proof.* See the Appendix. □

Theorem 4 indicates that the advantage of choosing  $t(W)$  with Assumption 2 in mind is two-fold. Under the baseline model defined by Assumption 1, Theorem 3 implies that  $\hat{\gamma}_{AST}$  will have low sampling variation if  $q_s(w) = \mathbb{E}[\psi_s(Y, W, \gamma_0) | W = w]$  and  $q_a(w) = \mathbb{E}[\psi_a(X, W, \gamma_0) | W = w]$  are approximately linear in  $t(w)$  (see also part (iii) of Theorem 2). This is the case covered by part (a) of the theorem. Now consider the case where the analyst misspecifies the propensity score model, but Assumption 2 holds, part (b) of Theorem 4 indicates that  $\hat{\gamma}_{AST}$  will remain consistent for  $\gamma_0$  in this case if the analyst chooses  $G(v)$  to take the logit form. We emphasize that the true propensity score model may or may not be of the logit form.

The peculiar feature of Theorem 4, relative to analogous results in the missing data literature (e.g., Tsiatis 2006), is the requirement that the assumed propensity score takes the logit form. To understand this requirement, note that, in general, (7) will be an inconsistent estimate of the study population distribution of  $W$  when the propensity score is misspecified. Calibrating the study and auxiliary tilts to moments of this distribution will therefore typically produce an inconsistent estimate of  $\gamma_0$ . However, when condition (b) of Theorem 4 holds, we have, from the estimating equations for the propensity score parameter,

$$\frac{1}{N} \sum_{i=1}^N (D_i - G(t(W_i)'\hat{\delta}_{ML})) t(W_i) = 0. \tag{18}$$

Now consider the mean of  $t(W_i)$  with respect to  $\hat{F}_s^{\text{eff}}(w)$ . Using (18), and the fact that  $t(W_i)$  contains a constant so that  $\sum_{i=1}^N G(t(W_i)'\hat{\delta}_{ML}) = \sum_{i=1}^N D_i$ , we have the equalities

$$\begin{aligned} \sum_{i=1}^N \hat{\pi}_i^{\text{eff}} t(W_i) &= \sum_{i=1}^N \frac{G(t(W_i)'\hat{\delta}_{ML})}{\sum_{j=1}^N G(t(W_j)'\hat{\delta}_{ML})} \\ t(W_i) &= \frac{\sum_{i=1}^N D_i t(W_i)}{\sum_{i=1}^N D_i}. \end{aligned}$$

Therefore, under the conditions of part (b) of Theorem 4,  $\sum_{i=1}^N \hat{\pi}_i^{\text{eff}} t(W_i) \xrightarrow{P} \mathbb{E}_s[t(W)]$  irrespective of whether the propensity score is correctly model. This implies that the study and auxiliary tilts will be correctly calibrated such that, when Assumption 2 holds,  $\hat{\gamma}_{AST}$  will remain consistent for  $\gamma_0$ . Note that this estimate of  $\mathbb{E}_s[t(W)]$  will not be efficient when the propensity score is misspecified.

Although the propensity score is not ancillary in the data combination problem, our estimator remains consistent in the presence of propensity score misspecification when  $G(v)$  takes



the logit form. It is an open question where there exist a locally efficient *and* doubly robust estimator under nonlogit parametric forms for the propensity score.

The alternative estimator, which replaces maximum likelihood (ML) propensity score fit, computed in the first step of our procedure with the method of moments (MM), one

$$\frac{1}{N} \sum_{i=1}^N (D_i - G(t(W_i)' \hat{\delta}_{MM})) t(W_i) = 0,$$

will be double robust but not locally efficient (unless a logit form for  $G(v)$  is maintained as part of Assumption 1, in which case the ML and MM fits coincide). More generally there is a tension between efficiency, which requires using the MLE of the propensity score for reweighting, and robustness to propensity score misspecification.

*Implications for Practitioners.* Collectively, Theorems 2 to 4 suggest several useful guidelines for empirical researchers. First, when overlap is good, or equivalently the propensity score weights  $p_0(W)/(1 - p_0(W))$  do not take very large values, Theorems 2 to 4 provide a very strong theoretical case for using AST in practice. If Assumption 2 happens to be true in the sampled populations, then AST will be more efficient than the propensity score reweighting approach of Abadie (2005). This result is analogous to the enhanced efficiency of the augmented inverse probability weighting (AIPW) estimator of Robins, Rotnitzky, and Zhao (1994) relative to conventional inverse probability weighting (IPW) in the missing data context. In practice, high levels of precision will be observed whenever  $q_s(w)$  and  $q_a(w)$  are reasonably well approximated by a linear combination of the elements of  $t(w)$ . A further advantage of the AST procedure is that, if the propensity score is inadvertently misspecified, AST will nevertheless remain consistent for  $\gamma_0$  if Assumption 2 holds (and the analyst works with a logit form for  $G(v)$ ).

In settings with poor overlap, the AST estimator may be highly variable and, in extreme cases, may not even exist. To understand this last observation consider the case where  $G(v)$  takes the logit form. In that case, the computation of the auxiliary tilt requires that the study sample mean of  $t(W)$  lies within the convex hull of the auxiliary sample. If the study and auxiliary distributions of  $W$  are very different from one another, this convex hull condition may fail in practice even if Assumption 1 holds in the population. We do not view this as a weakness of our procedure, rather such situations alert the researcher to the fragility of identification when overlap is poor (see Khan and Tamer 2010). When overlap is poor, direct imputation approach may be preferable (e.g., Chen, Hong, and Tarozzi 2008; Kline 2011). However, imputation will be very sensitive to violations of Assumption 2; this limitation is illustrated by our Monte Carlo experiments below.

The computational algorithm detailed in the supplemental Web appendix is designed to work well in situations where the convex hull condition is “nearly” violated and we recommend its routine use. For covariance matrix estimation, we recommend use the textbook formulas for the GMM estimator based on the moment vector implied by (8), (9), (11), and (13) above and explicitly defined in the Appendix.

#### 4. APPLICATION AND MONTE CARLO EXPERIMENTS

*Empirical Application.* Neal and Johnson (1996) studied the role of “premarket” (i.e., acquired prior to age 18) differences in cognitive achievement in explaining differences in earnings between young Black and White men. Using a sample of employed Black and White males drawn from the National Longitudinal Survey of Youth 1979 (NLSY79), Neal and Johnson (1996) computed the least-square fit of the logarithm of hourly wages on a constant, a black dummy, age, and Armed Forces Qualification Test (AFQT) percentile score measured at age 16 to 18. They found that the coefficient on the black dummy variable drops by two-thirds to three quarters when AFQT score is included as a covariate. On the basis of this finding, they argued that differences in the rate of cognitive skill acquisition across Blacks and White prior to age 18, due to differences in family background, school quality, and neighborhood characteristics, explain a substantial portion of subsequent Black–White wage inequality. We do not provide an assessment of this interpretation here, rather our goal is to illustrate the use of AST in a familiar setting.

Let  $Y$  denote real average wages from 1990 to 1993 for a random draw from the population of Black men aged 16 to 18 in 1979 and residing in the United States. This population corresponds to our study population of interest. Let  $X$  denote real wages for a random draw from the population of White men aged 16 to 18 in 1979 and residing in the United States. This corresponds to our auxiliary population. Let  $W$  be a vector including year of birth and AFQT score (we transform the percentile scores used by Neal and Johnson (1996) onto the real line using the inverse standard normal CDF). We compare features of the observed distribution of Black wages with those of a hypothetical White population whose age and AFQT distribution *coincides* with that of the Blacks (i.e., with study population’s). These types of hypothetical comparisons underlie Oaxaca decompositions, as used in labor and health economics, and similar exercises undertaken in demography (e.g., Kitagawa 1964). Barsky et al. (2002) and Fortin, Lemieux and Firpo (2011) surveyed the application of decomposition methods in economics.

Our sample closely resembles that used in Johnson and Neal (1998). It includes 1371 measurements of real wages, race, age, and AFQT scores drawn from the NLSY79. Throughout we replace the empirical measure of our sample with the NLSY79 base year sampling weights (although this adjustment has little effect on our results). The age distributions for Blacks and Whites in the merged sample are, as would be expected, quite similar. The distribution of AFQT scores across the two groups is quite different. The mean Black score is approximately one standard deviation lower than the mean White score. The two distributions also substantially differ in their second, third, and fourth moments (not reported).

Panel A of Table 1 reports estimates of mean log Wages for Blacks (Column 1), as well as the Black–White average difference (Column 2). On average, Blacks earn almost 28% less per hour than Whites in our sample. Panel A also reports estimates of the CDF of the Black wage distribution at selected points, and the corresponding Black–White CDF differences.

Table 1. Raw and adjusted differences in Black versus White hourly wages

	Panel A		Panel B		
	(1) Black	(2) B–W	(1) CEP	(2) PSR	(3) AST
Average (log(Wage))	6.749 (0.021)	–0.279 (0.026)	–0.1108 (0.0348)	–0.1072 (0.0303)	–0.1052 (0.0298)
Pr(Wage ≤ \$5.00)	0.0801 (0.0125)	0.0566 (0.0135)	0.0243 (0.0216)	0.0246 (0.0193)	0.0278 (0.0187)
Pr(Wage ≤ \$7.50)	0.4505 (0.0244)	0.2948 (0.0275)	0.1780 (0.0391)	0.1737 (0.0355)	0.1757 (0.0350)
Pr(Wage ≤ \$10.00)	0.6590 (0.0244)	0.2691 (0.0300)	0.0987 (0.0406)	0.0964 (0.0358)	0.0903 (0.0353)
Pr(Wage ≤ \$12.50)	0.8020 (0.0198)	0.2001 (0.0265)	0.0417 (0.0328)	0.0386 (0.0288)	0.0348 (0.0284)
Pr(Wage ≤ \$15.00)	0.8896 (0.0153)	0.1426 (0.0219)	0.0176 (0.0238)	0.0129 (0.0203)	0.0109 (0.0202)

NOTES: Results based on an extract of 1371 Black and White men ages 16 to 18 in 1979 from the NLSY79. Estimated standard errors, which account for within-household dependence in outcomes across siblings, are reported in parentheses.

For example, while over 45% of Blacks earn less than \$7.50 per hour in our sample, fewer than 30% of Whites do (Table 1, row 3). Inspection of the CDF differences indicates that while the distributions are most different at the lower wage levels, differences exist across the entire support of wages.

Panel B of Table 1 reports average wage differences between Blacks and a hypothetical population of Whites whose distribution of age and AFQT score coincides with the Black distribution. This allows for a comparison between Black and White wages that flexibly controls for differences between the two populations in age and AFQT score.

In Column 1 of Panel B, we report age- and AFQT-adjusted differences in mean wages and wage CDFs based on the conditional expectation projection (CEP) estimator of Chen, Hong, and Tarozzi (2008). Our implementation of their procedure models the conditional expectation functions (CEFs) of  $Y$  and  $X$  given  $W$  as a separable functions of a constant, 2 year of birth dummies, a quadratic polynomial in transformed AFQT score, and 12 dummy variables for the transformed AFQT score lying, respectively, below  $-2$ ,  $-1.75$ ,  $\dots$ ,  $0.25$ ,  $0.5$ . Let  $t(W)$  be the vector containing all these functions of  $W$ . In principle, if the dimension of the approximating model is allowed to grow with the sample size, the Chen, Hong, and Tarozzi (2008) estimator is consistent for, and efficient under, all data-generating processes satisfying parts (i) to (iv) of Assumption 1. In small samples, the performance of the estimator is heavily dependent on the quality of the two CEF approximations.

Column 2 of Panel B implements the propensity score reweighting (PSR) estimator of Hirano and Imbens (2001) and Abadie (2005). We model the propensity score as a logit function with an index linear in  $t(W)$  as defined above for the CEP estimator. The PSR estimates are very close in magnitude and precision to the CEP estimates.

Column 3 of Panel B implements our AST procedure using the same choice of  $t(W)$  and  $r(W) = t(W)$ . This choice ensures that the study and auxiliary sample tilts share the following features with the efficient distribution function estimate of  $W$ : (i) the marginal year of birth distributions coincide, (ii) the

means and variances of the transformed AFQT score coincide, (iii) the probability masses assigned to the intervals defined by the  $-2, -1.75, \dots, 0.25, 0.5$  grid of AFQT score intervals coincide. Figure 1 plots undersmoothed kernel density estimates of the actual Black and White AFQT score densities; the two distributions are very different from one another. The figure also plots a density estimate based on the auxiliary sample tilt. This corresponds to the AFQT score density in the hypothetical comparison population of Whites. As is evident from the figure, our choice of  $t(W)$  is rich enough to closely match this density with its target Black one.

After adjusting for age and AFQT differences, we find that while a Black–White residual log wage CDF gap remains at middle parts of the wage distribution, it disappears at the low and high ends of this distribution. The average log wage gaps falls, after adjusting for age and AFQT differences, from  $-0.279$  to  $-0.111$ .

While the AST point estimates are similar to the corresponding CEP and PSR ones, their estimated sampling precision is uniformly superior (as Theorem 3 would suggest). The close correspondence between the CEP, PSR, and AST point estimates in our application likely reflects a combination of two factors. First, while the AFQT distributions across Blacks and Whites differ dramatically, the support of the Black distribution is clearly contained within that of the White distribution. Hence part (iii) of Assumption 1 is well satisfied. Second the approximating models underlying each of the estimators are quite flexible. In settings where overlap is weaker, and/or the approximating models more parsimonious (as would be required when the dimension of  $W$  is large), we would expect the three estimators to more often yield different point estimates depending on the true data-generating process.

*Monte Carlo.* We now report on a number of Monte Carlo experiments we conducted to verify the theoretical properties described in Theorems 2 to 4. In particular, we wish to assess the relevance of our theoretical robustness and efficiency results. To do this, we consider a stylized program evaluation setting. The analyst wishes to estimate the average treatment effect on the treated (ATT).

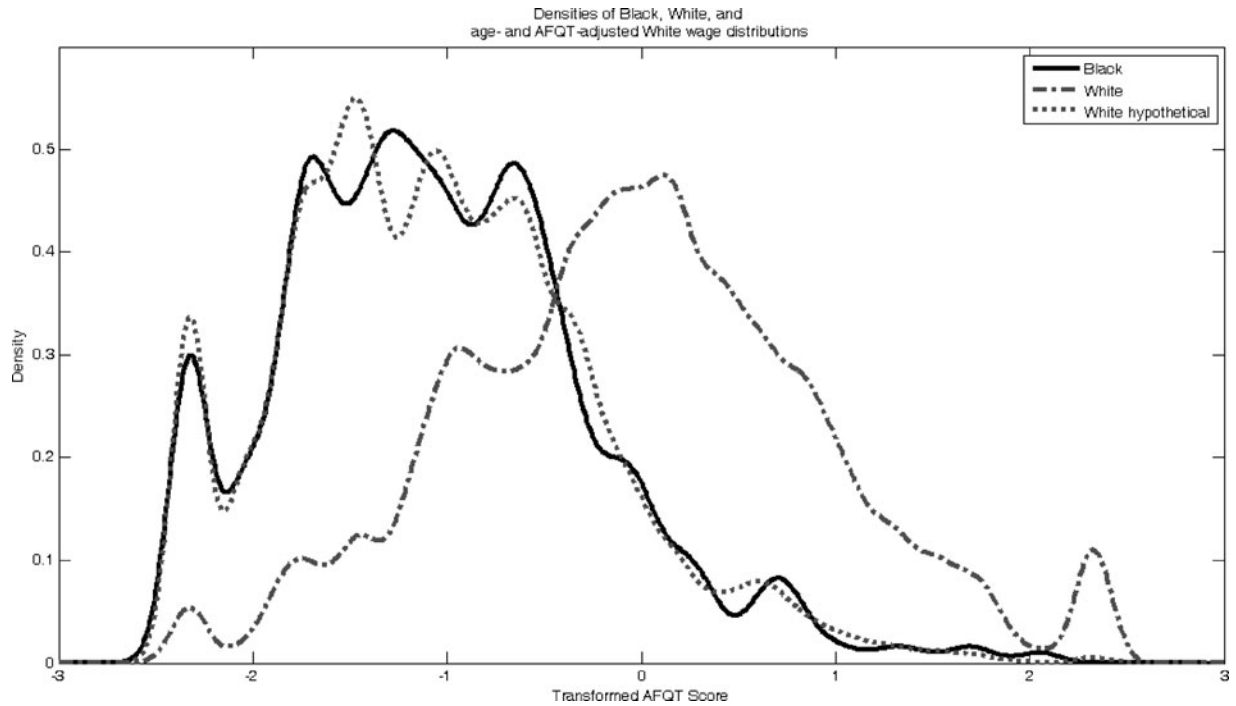


Figure 1. AFQT densities. The figure plots kernel density estimates of the actual Black and White AFQT score distributions as well as an estimate based on the auxiliary sample tilt. A Gaussian kernel is used with a bandwidth equal to 1/2 of Silverman’s “rule-of-thumb” choice. Undersmoothing highlights the ability of the auxiliary tilt to match local features of the Black AQFT density.

In each of our first set of experiments, we assume that  $W$  is distributed according to a truncated normal distribution, with support  $[-c, c]$ , in both the study (treated) and auxiliary (control) populations. The location and scale parameters of these two distributions, respectively,  $(\mu_s, \sigma_s^2)$  and  $(\mu_a, \sigma_a^2)$ , may differ. We assume a multinomial sampling scheme: with probability  $Q_0 = 1/2$  a draw of  $(Y, W)$  is taken at random from the study (treated) population, otherwise a draw of  $(X, W)$  is taken from the auxiliary (control) population. Finally, we assume that  $Y$  and  $X$ , which play the roles of the outcome under treatment and control, are generated according to

$$Y|W, D \sim \mathcal{N}(0, \sigma_Y^2)$$

$$X|W, D \sim \mathcal{N}(\alpha_0 + \alpha_1(W - \mu_{W|D=1}) + \alpha_2[(W - \mu_{W|D=1})^2 - \sigma_{W|D=1}^2], \sigma_X^2),$$

where  $\mu_{W|D=1}$  and  $\sigma_{W|D=1}^2$  are the study population mean and variance of  $W$  (which differ from  $\mu_s$  and  $\sigma_s^2$  due to truncation).

The target parameter is  $\gamma_0 = \mathbb{E}_s[Y - X] = \alpha_0$ . The propensity score induced by these designs is of the logit form with an index quadratic in  $W$ :

$$p_0(w) = [1 + \exp(-\beta_0 - \beta_1 W - \beta_2 W^2)]^{-1},$$

where  $\beta_0, \beta_1$ , and  $\beta_2$  are functions of  $(\mu_s, \sigma_s^2)$  and  $(\mu_a, \sigma_a^2)$  (see Anderson 1982). When the study and auxiliary population distributions of  $W$  have different means, but a common variance, the logit index will be linear in  $W$ . When both the means and variances differ, then the index will generally be nontrivially quadratic in  $W$ .

Across all designs, we assume a sample size of  $N = 1000$  and set  $\mu_s = 0, \sigma_s^2 = 1, \mu_a = -1/2, \alpha_0 = 0, \alpha_1 = 1/2, \sigma_X^2 = 1$ , and  $c = 3$ . We vary  $\sigma_a^2$  and  $\alpha_2$  across designs to, respectively,

induce nonlinearity in the (index of) the propensity score and  $\mathbb{E}[\psi_a(X, W, \gamma_0)|W] = q_a(W)$ . We vary  $\sigma_Y^2$  across designs to keep the variance bound fixed. Across each of our designs an efficient estimator (under Assumption 1) will have an asymptotic standard error of  $\sqrt{\mathcal{I}(\gamma_0)^{-1}}/1000 = 1/10$ .

Table 2 gives the parameter configurations for each of four Monte Carlo designs. In the first design both the propensity score,  $p_0(w)$  and  $q_a(w)$  are “linear” in  $w$  (for  $p_0(w)$  “linear” means linear in the logit index). In the second design, the propensity score is quadratic in  $w$ , while  $q_a(w)$  remains linear. In design 3 the reverse is true, while in design 4 both objects are “quadratic.” Across each design we implement the AST estimator with  $G(\cdot)$  being the logit function and  $r(W) = t(W) = (1, W)'$ . For the conditional expectation projection (CEP) estimator, we proceed “as if”  $\mathbb{E}[X|W]$  were linear in  $W$ , while our implementation of propensity score reweighting (PSR) uses a logit propensity score with a linear index.

Our AST estimator is consistent for  $\gamma_0$  in designs 1 through 3. CEP is consistent in designs 1 and 2, but inconsistent in design 3. The PSR estimator is consistent in designs 1 and 3, but inconsistent in design 2. All estimators are inconsistent in design 4 due to the nonlinearity of both  $p_0(w)$  and  $q_a(w)$ . Table 3 reports the results of our experiments. Column 1 lists a “pencil and paper” asymptotic bias calculation, while Column 2

Table 2. Parameter values for Monte Carlo experiments

Design	(1)	(2)	(3)	(4)
$\sigma_a^2$	1	2/3	1	2/3
$\sigma_Y^2$	3.4823	2.6590	1.7496	0.9253
$\alpha_2$	0	0	-1	-1

Table 3. Monte Carlo results

	(1) Asym. bias	(2) Med. bias	(3) Asym. SE.	(4) Median SE.	(5) Std. dev.	(6) Cov. of 95% CI	(7) RMSE
Design 1: $p_0(w)$ linear, $q_a(w)$ linear							
CEP	0.0000	0.0097	0.0997	0.0996	0.0986	0.9526	0.0986
PSR	0.0000	0.0164	0.1007	0.1006	0.1005	0.9506	0.1005
AST	0.0000	0.0055	0.0100	0.0998	0.0998	0.9540	0.0997
Design 2: $p_0(w)$ quadratic, $q_a(w)$ linear							
CEP	0.0000	0.0137	0.0925	0.0924	0.0947	0.9480	0.0947
PSR	0.5053	0.5437	0.0905	0.0911	0.0912	0.9126	0.1039
AST	0.0000	0.0169	0.0941	0.0931	0.0941	0.9470	0.0942
Design 3: $p_0(w)$ linear, $q_a(w)$ quadratic							
CEP	-1.6125	-2.0082	0.1309	0.1296	0.1627	0.6204	0.3111
PSR	0.0000	-0.0137	0.1063	0.1037	0.1068	0.9420	0.1068
AST	0.0000	-0.0266	0.1076	0.1054	0.1081	0.9416	0.1081
Design 4: $p_0(w)$ quadratic, $q_a(w)$ quadratic							
CEP	-4.6038	-6.7095	0.1192	0.1157	0.1728	0.0010	0.8196
PSR	-3.0049	-3.1031	0.0847	0.0821	0.0858	0.1694	0.2670
AST	-2.8789	-2.9313	0.0941	0.0873	0.0953	0.1726	0.2908

gives the median bias across 5000 Monte Carlo replications (in both cases bias is scaled by the “pencil and paper” asymptotic standard error reported in Column 3). As predicted, AST is median unbiased (up to simulation error) in designs 1 through 3. In contrast, PSR is severely biased in design 2 and CEP in design 3. As expected, all estimators perform poorly in design 4. These bias properties are reflected in the coverage of standard, Wald-based, 95% confidence intervals for  $\gamma_0$  (Column 6). By comparing Columns 1 and 2 and Columns 3 and 5, we see that—for the designs considered here—the finite sample distributions of all of the estimators are very well approximated by their asymptotic counterparts.

### 5. SUMMARY

When the propensity score is parametrically specified, information in both the study and auxiliary samples may be used to form an efficient estimate of  $W$ , the variable common to both datasets. An intuition for this insight follows from recognizing that, under part (v) of Assumption 1, the auxiliary sample is equivalent to a biased sample from the study population with the biasing function known up to a finite-dimensional parameter. Using this efficient distribution function estimate, we tilt the propensity score reweighting (study population) distribution function estimates of  $(Y, W)$  and  $(X, W)$  so that they share certain moments in common. By choosing these moments carefully (i.e., with reference to Assumption 2), we can produce a locally efficient estimate of  $\gamma_0$ . Even if the parametric relationship between the study and auxiliary populations, as embodied in the propensity score model, is misspecified, AST remains consistent for  $\gamma_0$  if Assumption 2 holds.

To our knowledge, we are the first to propose a locally efficient estimator for the class of data combination problems defined by Assumption 1. Our procedure also has a double robustness property. Our results provide a useful complement to the work of Robins, Rotnitzky, and Zhao (1994), Tan (2006),

and others for missing data problems. Relative to Chen, Hong, and Tarozzi (2008), who did provide explicit results for data combination problems (their so-called “verify-out-of-sample” case), our approach may be useful when  $W$  is high dimensional such that their method, which requires nonparametric estimation of  $q_s(w)$  and  $q_a(w)$ , is impractical.

In future work, it would be useful to study data-dependent methods for choosing  $t(W)$ . Similarly, it would be interesting to construct a locally efficient estimator with minimal variance across all estimators based on the linear approximating models  $q_s(W) \simeq \Pi_s t(W)$  and  $q_a(W) \simeq \Pi_a t(W)$ . In the missing data context, such estimators are called “improved locally efficient” (e.g., Tan 2010).

### APPENDIX: PROOFS

*Proof of Theorem 2.* The AST procedure coincides with a just identified GMM estimator based on the  $\dim(r(W)) + 2 \dim(t(W)) + \dim(\gamma_0)$  vector of moment functions  $m(Z_i, \theta_0)$  with  $\theta = (\delta', \lambda'_a, \lambda'_s, \gamma')'$ . This vector of moment functions is composed of the subvectors:

$$\begin{aligned}
 m_1(Z, \delta_0) &= \frac{D - G(r(W)' \delta_0)}{\dim(r(W)) \times 1} G(r(W)' \delta_0) (1 - G(r(W)' \delta_0)) \\
 &\quad \times G_1(r(W)' \delta_0) r(W) \\
 m_2(Z, \delta_0, \lambda_{a0}) &= \frac{1}{Q_0} \left( \frac{1 - D}{1 - G(r(W)' \delta_0 + t(W)' \lambda_{a0})} - 1 \right) \\
 &\quad \times G(r(W)' \delta_0) t(W) \\
 m_3(Z, \delta_0, \lambda_{s0}) &= \frac{1}{Q_0} \left( \frac{D}{G(r(W)' \delta_0 + t(W)' \lambda_{s0})} - 1 \right) \\
 &\quad \times G(r(W)' \delta_0) t(W)
 \end{aligned}$$

$$m_4(Z, \delta_0, \lambda_{a0}, \lambda_{s0}, \gamma_0) = \frac{D}{Q_0} \frac{G(r(W)' \delta_0)}{G(r(W)' \delta_0 + t(W)' \lambda_{s0})} \psi_s(Y, W, \gamma_0) - \frac{1-D}{Q_0} \frac{G(r(W)' \delta_0)}{1 - G(r(W)' \delta_0 + t(W)' \lambda_{a0})} \times \psi_a(X, W, \gamma_0).$$

Let  $M = \mathbb{E}[\partial m(Z, \theta_0) / \partial \theta']$ ; a standard argument (e.g., Newey and McFadden 1994) gives, under regularity conditions, the asymptotically linear representation

$$\sqrt{N}(\hat{\theta} - \theta_0) = -M^{-1} \left( \frac{1}{\sqrt{N}} \sum_{i=1}^N m(Z_i, \theta_0) \right) + o_p(1). \quad (\text{A.1})$$

The influence function for  $\hat{\gamma}_{\text{AST}}$  corresponds to the last  $K$  elements of (A.1). By tedious, but straightforward, calculation, we can show that this subvector equals

$$\begin{aligned} \sqrt{N}(\hat{\gamma} - \gamma_0) &= \frac{-M_{44}^{-1}}{\sqrt{N}} \sum_{i=1}^N \{m_4(Z_i, \delta_0, \lambda_{a0}, \lambda_{s0}, \gamma_0) \\ &\quad - M_{41} M_{11}^{-1} m_1(Z_i, \delta_0) \\ &\quad + M_{42} M_{22}^{-1} (M_{21} M_{11}^{-1} m_1(Z_i, \delta_0) - m_2(Z_i, \delta_0, \lambda_{a0})) \\ &\quad + M_{43} M_{33}^{-1} (M_{31} M_{11}^{-1} m_1(Z_i, \delta_0) - m_3(Z_i, \delta_0, \lambda_{s0}))\} \\ &\quad + o_p(1), \end{aligned} \quad (\text{A.2})$$

where  $M_{kl}$  equals the expected value of the derivative of the  $k$ th subvector of  $m(Z, \theta)$  with respect to the  $l$ th subvector of  $\theta$  evaluated at  $\theta = \theta_0$ .

Under part (v) of Assumption 1, the information matrix equality gives  $M_{11} = -\mathbb{E}[\mathbb{S}_\delta \mathbb{S}_\delta']$ . Calculation gives  $M_{41}$  equal to

$$M_{41} = \frac{1}{Q_0} \mathbb{E} \left[ \frac{1-D}{1-p_0(W)} \psi_a(X, W, \gamma_0) \mathbb{S}_\delta' \right]. \quad (\text{A.3})$$

Using this result, iterated expectations, and part (ii) of Assumption 1, we then get

$$-M_{41} M_{11}^{-1} m_1(Z, \delta_0) = \frac{1}{Q_0} \mathbb{E}^* \left[ \frac{1-D}{1-p(W)} q_a(W) \middle| \mathbb{S}_\delta \right].$$

Evaluating  $M_{21}$  yields, after some manipulation,

$$M_{21} = -\frac{1}{Q_0} \mathbb{E} \left[ \left( \frac{1-D}{1-p_0(W)} - 1 \right) p_0(W) t(W) \mathbb{S}_\delta' \right], \quad (\text{A.4})$$

where  $p_0(W) = G(r(W)' \delta_0) = G(r(W)' \delta_0 + t(W)' \lambda_{a0})$ . These results imply that

$$M_{21} M_{11}^{-1} m_1(Z, \delta) = \frac{1}{Q_0} \mathbb{E}^* \left[ \left( \frac{1-D}{1-p_0(W)} - 1 \right) p_0(W) t(W) \middle| \mathbb{S}_\delta \right].$$

Similar calculations give

$$M_{31} = -\frac{1}{Q_0} \mathbb{E} \left[ \left( \frac{D}{p_0(W)} - 1 \right) p_0(W) t(W) \mathbb{S}_\delta' \right], \quad (\text{A.5})$$

yielding

$$M_{31} M_{11}^{-1} m_1(Z, \delta_0) = \frac{1}{Q_0} \mathbb{E}^* \left[ \left( \frac{D}{p_0(W)} - 1 \right) p_0(W) t(W) \middle| \mathbb{S}_\delta \right].$$

Evaluating  $M_{22}$  and  $M_{42}$  yields

$$M_{22} = \frac{1}{Q_0} \mathbb{E} \left[ \frac{p_0(W)}{1-p_0(W)} G_1(r(W)' \delta_0) t(W) t(W)' \right] \quad (\text{A.6})$$

$$M_{42} = -\frac{1}{Q_0} \mathbb{E} \left[ \frac{p_0(W)}{1-p_0(W)} G_1(r(W)' \delta_0) \psi_a(X, W, \gamma_0) t(W)' \right]. \quad (\text{A.7})$$

Using (A.6) and (A.7) and iterated expectations, we get

$$\begin{aligned} M_{42} M_{22}^{-1} &= -\mathbb{E} \left[ \frac{p_0(W)}{1-p_0(W)} G_1(r(W)' \delta_0) q_a(W) t(W)' \right] \\ &\quad \times \mathbb{E} \left[ \frac{p_0(W)}{1-p_0(W)} G_1(r(W)' \delta_0) t(W) t(W)' \right]^{-1} \\ &\stackrel{\text{def}}{=} -\Pi_a^*, \end{aligned}$$

as defined in (14) of the main text.

Now consider  $M_{33}$  and  $M_{43}$ ; we have

$$M_{33} = -\frac{1}{Q_0} \mathbb{E} [G_1(r(W)' \delta_0) t(W) t(W)'] \quad (\text{A.8})$$

$$M_{43} = -\frac{1}{Q_0} \mathbb{E} [G_1(r(W)' \delta_0) \psi_s(X, W, \gamma_0) t(W)']. \quad (\text{A.9})$$

Using (A.8) and (A.9) and iterated expectations, we get

$$\begin{aligned} M_{43} M_{33}^{-1} &= \mathbb{E} [G_1(r(W)' \delta_0) q_s(W) t(W)'] \\ &\quad \times \mathbb{E} [G_1(r(W)' \delta_0) t(W) t(W)']^{-1} \stackrel{\text{def}}{=} \Pi_s^*, \end{aligned}$$

also as defined in (14) of the main text.

Recalling the definitions,  $q_a^*(W) = \Pi_a^* t(W)$  and  $q_s^*(W) = \Pi_s^* t(W)$ , substituting the expressions derived immediately above into (A.2), and rearranging yield the form of the influence function stated in the theorem.

Now recall the definitions of  $R_s(D, W)$  and  $R_a(D, W)$  given in (15) and (16) of the main text. By the definition of the LP operator, we have that  $\mathbb{E}[R_a(D, W) \mathbb{S}_\delta'] = \mathbb{E}[R_s(D, W) \mathbb{S}_\delta'] = 0$ . This follows since  $R_a(D, W)$  and  $R_s(D, W)$  are linear prediction errors, with  $\mathbb{S}_\delta$  the vector of predictors. The conditional mean zero property of the score function also yields the restrictions  $\mathbb{E}[R_a(D, W) | W] = \mathbb{E}[R_s(D, W) | W] = 0$ . From these properties and direct calculation, we have that

$$\mathbb{E}[\phi^{\text{eff}}(Z, \gamma_0) \{R_s(D, W) - R_a(D, W)\}] = 0$$

from which the claimed form of the asymptotic variance of  $\gamma_0$  follows.

Let  $a$  be a vector of constants. By linearity of the LP operator, the Cauchy-Schwartz inequality, and recalling that  $R(D, W) = R_s(D, W) - R_a(D, W)$ , we have that

$$\begin{aligned} a' \nabla (R(D, W)) a &\leq a' \nabla \left( \frac{1}{Q_0} \left( \frac{1-D}{1-p_0(W)} - 1 \right) p_0(W) \{q_a^*(W) - q_a(W)\} \right. \\ &\quad \left. + \frac{1}{Q_0} \left( \frac{D}{p_0(W)} - 1 \right) p_0(W) \{q_s^*(W) - q_s(W)\} \right) a. \end{aligned}$$

This bound will hold with equality if  $r(W) \subset t(W)$  since, by the definitions of  $\Pi_s^*$  and  $\Pi_a^*$ , we will have (in that case) the zero covariance results

$$\begin{aligned} \mathbb{E} \left[ (q_a(W) - q_a^*(W)) \frac{p_0(W)}{1-p_0(W)} G_1(r(W)' \delta_0) r(W)' \right] &= 0 \\ \mathbb{E} [\{q_a^*(W) - q_a(W)\} G_1(r(W)' \delta_0) r(W)'] &= 0. \end{aligned} \quad (\text{A.10})$$

Iterated expectations and (A.10) then give

$$\begin{aligned} \mathbb{E} \left[ \left( \frac{D}{p_0(W)} - 1 \right) p_0(W) \{q_a^*(W) - q_a(W)\} \mathbb{S}_\delta' \right] \\ = \mathbb{E} [\{q_a^*(W) - q_a(W)\} G_1(r(W)' \delta_0) r(W)'] = 0 \end{aligned} \quad (\text{A.11})$$

and also

$$\begin{aligned} & \mathbb{E} \left[ \left( \frac{1-D}{1-p_0(W)} - 1 \right) p_0(W) \{q_a^*(W) - q_a(W)\} S'_0 \right] \\ &= -\mathbb{E} \left[ \{q_a^*(W) - q_a(W)\} \frac{(D-p_0(W))^2}{(1-p_0(W))^2} G_1(r(W)' \delta_0) r(W)' \right] = 0. \end{aligned} \tag{A.12}$$

Equations (A.11) and (A.12) imply that, if  $r(W) \subset t(W)$ , then, after manipulation,

$$\mathbb{V}(R(D, W)) = \frac{1}{Q_0^2} \mathbb{E} \left[ \frac{p_0(W)}{1-p_0(W)} U^* U^{*'} \right], \tag{A.13}$$

with the right-hand side of (A.13) an upper bound otherwise. Recalling the definition of  $\epsilon$  given in the statement of the Theorem, and making use of the various compact support conditions embedded in Assumption 1, we get the bound

$$a' \mathbb{V}(R(D, W)) a \leq \frac{\epsilon^2}{Q_0^2} \mathbb{E} \left[ \frac{p_0(W)}{1-p_0(W)} \right] a' u' a,$$

from which (17) follows directly.

*Proof of Theorem 3.* Under Assumption 2, we have that  $\Pi_s^* = \Pi_s$  and  $\Pi_a^* = \Pi_a$ . This implies that  $R_s(D, W)$  and  $R_a(D, W)$  are identically equal to zero. The result then follows directly from Theorem 1.

*Proof of Theorem 4.* Asymptotic normality follows from standard results. Consistency under part (a) is a consequence of Equation (4) in the main text. Showing consistency under part (b) is more complicated. Denote the probability limits of  $\hat{\delta}$ ,  $\hat{\lambda}_a$ , and  $\hat{\lambda}_s$  when part (v) of Assumption 1 fails to hold by, respectively,  $\delta_*$ ,  $\lambda_{a*}$ , and  $\lambda_{s*}$ . Let  $p_*(W) = G(r(W)' \delta_*)$  and  $p_j(W) = G(r(W)' \delta_* + t(W)' \lambda_{j*})$  for  $j = s, a$ . If  $G(\cdot)$  takes the logit form, then  $p_*(W)$  will satisfy the population restriction  $\mathbb{E}[m_1(Z, \delta_*)] = \mathbb{E}[(D - p_*(W))t(W)] = 0$  so that, using iterated expectations and rearranging, we have the equality

$$\mathbb{E}[t(W)|D=1] = \mathbb{E} \left[ \frac{p_*(W)}{Q_0} t(W) \right]. \tag{A.14}$$

We also have  $\mathbb{E}[m_2(Z, \delta_*, \lambda_{a*})] = \mathbb{E}[m_3(Z, \delta_*, \lambda_{s*})] = 0$ , which, respectively, multiplying by  $\Pi_a$  and  $\Pi_s$  (using Assumption 2), gives the additional equalities:

$$\mathbb{E} \left[ \frac{1-D}{1-p_a(W)} p_*(W) q_a(W) \right] = \mathbb{E}[p_*(W) q_a(W)] \tag{A.15}$$

$$\mathbb{E} \left[ \frac{D}{p_s(W)} p_*(W) q_s(W) \right] = \mathbb{E}[p_*(W) q_s(W)]. \tag{A.16}$$

Using (A.14), (A.15), (A.16), Assumption 2, iterated expectations, and part (ii) of Assumption 1 yields

$$\begin{aligned} \mathbb{E}[m_4(Z, \delta_*, \lambda_{a*}, \lambda_{s*}, \gamma)] &= \mathbb{E} \left[ \frac{p_*(W)}{Q_0} \{q_s(W) - q_a(W)\} \right] \\ &= (\Pi_s - \Pi_a) \mathbb{E} \left[ \frac{p_*(W)}{Q_0} t(W) \right] \\ &= \mathbb{E}[q_s(W) - q_a(W)|D=1] \\ &= \mathbb{E}[\psi(Z, \gamma)|D=1], \end{aligned}$$

which by part (i) of Assumption 1 is uniquely zero at  $\gamma = \gamma_0$ .

### SUPPLEMENTARY MATERIALS

The supplemental Web appendix contains additional proof details, extra examples of data combination problems, and additional Monte Carlo results. An algorithm for computing our estimator, that we have

found to work well in practice, is also described in the supplemental Web appendix.

### ACKNOWLEDGMENTS

The authors thank David Card, Stephen Cosslett, Jinyong Hahn, Michael Jansson, Patrick Kline, Richard Smith, Tom Rothenberg, and members of the Berkeley Econometrics Reading Group for helpful discussions. We are particularly grateful to Gary Chamberlain, Guido Imbens, Justin McCrary, Geert Ridder, Enrique Sentana, and Leonard Stefanski for detailed comments on earlier drafts. This article has benefited from comments by the co-editor, associate editor, and three anonymous referees. The authors thank Jing Qin and Biao Zhang for assistance in replicating the Monte Carlo designs in Qin and Zhang (2008). The authors also acknowledge feedback and suggestions from participants in seminars at the University of Pittsburgh, Ohio State University, University of Southern California, University of California - Riverside, University of California - Davis, University of Maryland, Georgetown University, Duke University, University of California - Berkeley, CEMFI (Madrid), Pontifícia Universidade Católica do Rio de Janeiro, and the 2013 North American Summer Meeting of the Econometric Society. Preliminary portions of the current article previously appeared in Section 4 of an early draft of the NBER Working Paper “Inverse probability tilting and missing data problems.” The published version of that article excludes the material reported here. A supplemental appendix with proofs and additional details regarding computation may be found on the first author’s web page. All the usual disclaimers apply.

[Received January 2014. Revised March 2015.]

### REFERENCES

Abadie, A. (2005), “Semiparametric Difference-in-Differences,” *Review of Economic Studies*, 72, 1–19. [289,290,292,295,296]

Abadie, A., and Imbens, G. W. (2006), “Large Sample Properties of Matching Estimators for Average Treatment Effects,” *Econometrica*, 74, 235–267. [290]

Anderson, J. A. (1982), “Logistic Discrimination,” in *Handbook of Statistics* (Vol. 2), eds. P. R. Krishnaiah and L. N. Kanal, Amsterdam: North-Holland, pp. 169–191. [297]

Angrist, J. D., and Krueger, A. B. (1992), “The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables With Moments From Two-Samples,” *Journal of the American Statistical Association*, 87, 328–336. [288,290]

Barsky, R., Bound, J., Charles, K. K., and Lupton, J. P. (2002), “Accounting for the Black-White Wealth Gap: A Nonparametric Approach,” *Journal of the American Statistical Association*, 97, 663–673. [288,295]

Bickel, P. J., Ritov, Y., and Wellner, J. A. (1991), “Efficient Estimation of Linear Functionals of a Probability Measure P With Known Marginal Distributions,” *Annals of Statistics*, 19, 1316–1346. [292]

Björklund, A., and Jäntti, M. (1997), “Intergenerational Income Mobility in Sweden Compared to the United States,” *American Economic Review*, 87, 1009–1018. [291]

Carroll, R. J., and Wand, M. P. (1991), “Semiparametric Estimation in Logistic Measurement Error Models,” *Journal of the Royal Statistical Society, Series B*, 53, 573–585. [288]

Chen, X., Hong, H., and Tarozzi, A. (2008), “Semiparametric Efficiency in GMM Models With Auxiliary Data,” *Annals of Statistics*, 36, 808–843. [288,289,290,291,295,296,298]

Cheng, J., Small, D. S., Tan, Z., and Have, T. R. T. (2009), “Efficient Nonparametric Estimation of Causal Effects in Randomized Trials With Noncompliance,” *Biometrika*, 96, 19–36. [292]

Currie, J., and Yelowitz, A. (2000), “Are Public Housing Projects Good for Kids?” *Journal of Public Economics*, 75, 99–124. [290]

Dehejia, R. H., and Wahba, S. (1999), “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs,” *Journal of the American Statistical Association*, 94, 1053–1062. [288]

Dinardo, J., Fortin, N. M., and Lemieux, T. (1996), “Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach,” *Econometrica*, 64, 1001–1044. [288]

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003), “Micro-Level Estimation of Poverty and Inequality,” *Econometrica*, 71, 355–364. [288]

Firpo, S., and Rothe, C. (2013), “Semiparametric Estimation and Inference Using Doubly Robust Moment Conditions,” Mimeo. [289]

Fortin, N., Lemieux, T., and Firpo, S. (2011), “Decomposition Methods in Economics,” in *Handbook of Labor Economics* (Vol. 4A), eds. O. Ashenfelter and D. Card, Amsterdam: North-Holland, pp. 1–102. [295]

- Gilbert, P. B., Lele, S. R., and Vardi, Y. (1999), "Maximum Likelihood Estimation in Semiparametric Selection Bias Models With Application to AIDS Vaccine Trials," *Biometrika*, 86, 27–43. [289,291]
- Graham, B. S. (2011), "Efficiency Bounds for Missing Data Models With Semiparametric Restrictions," *Econometrica*, 79, 437–452. [292,294]
- Graham, B. S., de Xavier Pinto, C. C., and Egel, D. (2012), "Inverse Probability Tilting for Moment Condition Models With Missing Data," *Review of Economic Studies*, 79, 1053–1079. [292]
- Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. [288,289,290,291]
- Heckman, J. J., and Robb, R. (1985), "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, eds. J. Heckman and B. Singer, Cambridge: Cambridge University Press, pp. 156–245. [288]
- Hellerstein, J. K., and Imbens, G. W. (1999), "Imposing Moment Restrictions From Auxiliary Data by Weighting," *Review of Economics and Statistics*, 81, 1–14. [292]
- Hirano, K., and Imbens, G. W. (2001), "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization," *Health Services and Outcomes Research*, 2, 259–278. [289,296]
- Hirano, K., Imbens, G. W., Ridder, G., and Rubin, D. B. (2001), "Combining Panel Data Sets With Attrition and Refreshment Samples," *Econometrica*, 69, 1645–1659. [292]
- Ichimura, H., and Martinez-Sanchis, E. (2004), "Identification and Estimation of GMM Models by a Combination of Two Data Sets," Mimeo. [290]
- Imbens, G. W. (2004), "Nonparametric Estimation of Average Treatment Effect Under Exogeneity: A Review," *Review of Economic and Statistics*, 86, 4–29. [288]
- Inoue, A., and Solon, G. (2010), "Two-Sample Instrumental Variables Estimators," *Review of Economics and Statistics*, 92, 557–561. [290]
- Johnson, W. R., and Neal, D. A. (1998), "Basic Skills and the Black-White Earnings Gap," in *The Black-White Test Score Gap*, eds. C. Jencks and M. Phillips, Washington, DC: The Brookings Institution, pp. 480–500. [295]
- Khan, S., and Tamer, E. (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042. [293,295]
- Kitagawa, E. M. (1964), "Standardized Comparisons in Population Research," *Demography*, 1, 296–315. [288,295]
- Kline, P. (2011), "Oaxaca-Blinder as a Reweighting Estimator," *The American Economic Review*, 101, 532–537. [295]
- Lalonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604–620. [288]
- Little, R. J. A., and Wu, M. M. (1991), "Models for Contingency Tables With Known Margins When Target and Sampled Populations Differ," *Journal of the American Statistical Association*, 86, 87–95. [292]
- Neal, D. A., and Johnson, W. R. (1996), "The Role of Premarket Factors in Black-White Wage Differences," *Journal of Political Economy*, 104, 869–895. [295]
- Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135. [294]
- Newey, W. K., and McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics* (Vol. 4), eds. R. F. Engle and D. L. McFadden, Amsterdam: North-Holland, pp. 2111–2245. [288,299]
- Qin, J. (1998), "Inferences for Case-Control and Semiparametric Two-Sample Density Ratio Models," *Biometrika*, 85, 619–630. [289,291]
- Qin, J., and Zhang, B. (2007), "Empirical-Likelihood-Based Inference in Missing Response Problems and its Application in Observational Studies," *Journal of the Royal Statistical Society, Series B*, 69, 101–122. [292]
- (2008), "Empirical-Likelihood-Based Difference-in-Differences Estimators," *Journal of the Royal Statistical Society, Series B*, 70, 329–349. [289,300]
- Ridder, G., and Moffitt, R. (2007), "The Econometrics of Data Combination," in *Handbook of Econometrics* (Vol. 6B), eds. J. J. Heckman and E. E. Leamer, Amsterdam: North-Holland, pp. 5469–5547. [288,290,291]
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors are not Always Observed," *Journal of the American Statistical Association*, 89, 846–866. [289,290,291,294,295,298]
- Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [289]
- Tan, Z. (2006), "A Distributional Approach for Causal Inference Using Propensity Scores," *Journal of the American Statistical Association*, 101, 1619–1637. [292,298]
- (2010), "Bounded, Efficient and Doubly Robust Estimation With Inverse Weighting," *Biometrika*, 97, 661–682. [298]
- Tarozzi, A., and Deaton, A. (2009), "Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas," *Review of Economics and Statistics*, 91, 773–792. [288]
- Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer. [294]
- Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press. [288]