Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST)[1]

Bryan S. Graham,[◇] Cristine Campos de Xavier Pinto[+] and Daniel Egel[†]

Initial Draft: July 2006

This Draft: May 28, 2013

[1]We would like to 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 draft has benefited from comments by the co-editor, associate editor and three anonymous referees. We thank Jing Qin and Biao Zhang for assistance in replicating the Monte Carlo designs in Qin and Zhang (2008). We 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 paper previously appeared in Section 4 of an early draft of the NBER Working Paper 'Inverse probability tilting and missing data problems'. The latest revision of that paper 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.

[◇]Department of Economics, 530 Evans Hall, University of California - Berkeley, Berkeley, CA 94720-3880 and NBER. E-Mail: bgraham@econ.berkeley.edu, Web: https://emlab.berkeley.edu/~bgraham/

[+]Escola de Economia de São Paulo, FGV, Rua Itapeva 474, sala 1010, CEP: 01332-000. E-mail: cristinepinto@gmail.com. Web: http://sites.google.com/site/cristinepinto/.

[†]RAND Corporation. E-mail: Daniel_Egel@rand.org.

Abstract

We propose a locally efficient estimator for a class of semiparametric data combination problems. Our estimator also possesses a double robustness type property. 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 analyzed by Robins, Rotnitzky and Zhao (1994) (of which the Average Treatment Effect (ATE) estimand is a special case). Our procedure may be used to efficiently estimate, among other objects, the ATT, the two-sample instrumental variables model (TSIV), counterfactual distributions, and poverty maps. In an empirical application we use our procedure to characterize residual Black-White wage inequality after flexibly controlling for 'pre-market' differences in measured cognitive achievement as in Neal and Johnson (1996). We find that residual Black-White inequality is negligible at lower and higher quantiles of the Black wage distribution, but substantial at middle quantiles.

JEL Classification: C14, C21, C23

Key Words: Data Combination, Two-Sample Instrumental Variables (TSIV), Average Treatment Effect on the Treated (ATT), Poverty Maps, Earnings decompositions, Direct Standardization, Semiparametric Difference-in-Differences, Semiparametric Efficiency, Black-White Gap, Double Robustness

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 γ_0, and known function $\psi(z, \gamma)$ of the same dimension, we assume that

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

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 γ_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 paper we consider estimation of γ_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 Tarozzi (2008) show that identification of γ_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 'non-common' variables Y and X

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

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, Bound, Charles and Lupton (2002), poverty mapping as in Elbers, Lanjouw and Lanjouw (2003) and Tarozzi and Deaton (2007), 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), combine two distinct samples to estimate the effect of the National Supported Work (NSW) demonstration, a labor training program, on the post-intervention earnings of trainees. Their study sample consists of 185 NSW participants, while their auxiliary sample includes 2,490 non-participants drawn from the Panel Study of Income Dynamics (PSID). These two samples consist of random draws from distinct, non-overlapping, 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.

The theoretical statistics literature has emphasized differences between data combination and missing data problems. In an important paper 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) generalize 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. To our knowledge we are the first to propose a locally efficient estimator in the data combination context. Chen, Hong and Tarozzi (2008) propose a globally efficient estimator, but their procedure requires nonparametric modelling as opposed to the flexible parametric approach adopted here. Our methods provide a practical alternative to theirs when W is high dimensional (cf., Firpo and Rothe, 2013). Abadie (2005) develops a parametric propensity score reweighting (PSR) estimate of the ATT. Qin and Zhang (2008) show that Abadie's estimator can have low efficiency in some settings and

propose an alternative that uses empirical likelihood ideas. Qin and Zhang (2008) do not characterize the semiparametric efficiency or robustness properties of their ATT estimator, nor show how to extend it to the wider class of problems considered here. Hirano and Imbens (2001) also propose a modification of Abadie's (2005) estimator. They demonstrate that their modified estimator exhibits a double robustness property, but do not consider issues of semiparametric efficiency nor general data combination problems as we do.

In Section 2 we define the semiparametric data combination model. We also describe a number of specific data combination problems that arise frequently in applied statistics and econometrics. 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, Vardi, 1999). In Section 3 we define our estimator and formally characterize its large sample properties. Section 4 provides an empirical application and reports on the results of several Monte Carlo experiments.

2 Semiparametric data combination model

A formal definition of the data combination model is given by Assumption 2.1 below.

Assumption 2.1 ***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* $\gamma \neq \gamma_0$*,* $\gamma \in \mathcal{G} \subset \mathbb{R}^K$*,* $z \in \mathcal{Z} \subset \mathbb{R}^{\dim(Z)}$;

(ii) (CONDITIONAL DISTRIBUTIONAL EQUALITY) $F_s(x|w) = F_a(x|w)$ *and* $F_s(y|w) =$

$F_a(y|w)$ *for all* $w \in \mathcal{W} \subset \mathbb{R}^{\dim(W)}$, $x \in \mathcal{X} \subset \mathbb{R}^{\dim(X)}$ *and* $y \in \mathcal{Y} \subset \mathbb{R}^{\dim(Y)}$;

(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* $Q_0 \in (\kappa_0, 1-\kappa_0)$ *for* $0 < \kappa_0 < 1$ *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* $(i = 1, \ldots, N)$ *corresponds to a study population unit and* $D_i = 0$ *otherwise;*

(v) (Propensity score model) *There is a unique* $\delta_0 \in \mathcal{D} \subset \mathbb{R}^{1+M}$, *known vector* $r(W)$ *of linearly independent functions of* W *with a constant in the first row, and known function* $G(\cdot)$ *such that (i)* $G(\cdot)$ *is strictly increasing, differentiable and maps into the unit interval with* $\lim_{v \to -\infty} G(v) = 0$ *and* $\lim_{v \to \infty} G(v) = 1$, *(ii)* $\frac{f_s(w)}{f_a(w)} = \frac{1-Q_0}{Q_0} \frac{G(r(w)'\delta_0)}{1-G(r(w)'\delta_0)}$ *for all* $w \in \mathcal{W}$, *and (iii)* $0 < G(r(w)'\delta) \leq \kappa < 1$ *for all* $\delta \in \mathcal{D}$ *and* $w \in \mathcal{W}$.

The first part of Assumption 2.1 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 2.1 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. Formulating a model by imposing restrictions on a pseudo-population is somewhat awkward (cf., 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 $p_0(w)$

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

Under the merged population formulation of the problem it is clear that part (i) of Assumption 2.1 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) of the assumption 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 γ_0 follows from, using parts (ii) and (iii) of Assumption 2.1 and Equation (3), the equality

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

which is, by part one of Assumption 2.1, uniquely zero at $\gamma=\gamma_0$. See Lemma 3.1 of Abadie (2005) for a formal proof.

2.1 Examples

To give some idea of the range of problems to which our methods apply, we outline three examples (in addition to the program evaluation example discussed in the introduction). Additional examples are described in Chen, Hong and Tarozzi (2008) and Ridder and Moffitt (2007).

Two sample instrumental variables (TSIV) model: Ridder and Moffitt (2007) consider two sample instrumental variables (TSIV) models of the form

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

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 2.1 with the additional requirement that $F_s\left(w\right)=F_a\left(w\right)$. The TSIV model is of the form required by (2) with $\psi_s\left(y,w,\gamma\right)=f\left(Y;\gamma\right)e\left(W\right)$ and $\psi_a\left(x,w,\gamma\right)=g\left(X,W_1;\gamma\right)e\left(W\right)$. When $e\left(W\right)=W$, $f\left(Y;\gamma\right)=Y$ and $g\left(X,W_1;\gamma\right)=X'\alpha+W_1'\beta$

with $\gamma_0 = (\alpha_0, \beta_0')'$ we have the linear model analyzed by Angrist and Krueger (1992). Ridder and Moffitt (2007) show how one may estimate the Mixed Proportional Hazard (MPH) model under this setup, while Ichimura and Martinez-Sanchis (2004) discuss binary choice models.

A concrete example of a TSIV problem is provided by the work of Currie and Yelowitz (2000), who consider the model

$$\mathbb{E}_t \left[W \left(Y - X'\alpha_0 - W_1'\beta_0 \right) \right] = 0,$$

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 argue 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) observe Y and W for a random subsample of children drawn from the US Census. The Census, however, does not collect information on residence in public housing, X. This information is available in the US 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 Björkland 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 Theorem 3.1 below with Angrist and Krueger (1992, p. 331) or Ridder and Moffitt (2007, p. 5505)).

Poverty mapping: Let X be an indicator denoting whether a household's total outlay falls below a poverty line and W a vector of household characteristics. We seek to estimate the poverty rate in a specific study municipality as in Elbers, Lanjouw and Lanjouw (2003) and Tarozzi and Deaton (2007). Available is a random sample of N_s observations of W from this municipality; however, no poverty measurements are available in this sample. Also available is a random sample of size N_a of both X and W from the entire country. Our estimand is $\gamma_0 = \mathbb{E}_s[X]$ which corresponds to setting $\psi_s(Y, W, \gamma) = 0$ and $\psi_a(X, W, \gamma) = X - \gamma$. In this example part two of Assumption 2.1 implies that the conditional probability of begin poor given $W = w$ is the same in the entire country as it is in the specific municipality of interest.

Counterfactual distributions and direct standardization: We develop this example in our empirical application below. Let Y be wages of employed Black males and X those of White males. Let W be a vector of worker characteristics. A random

study sample of Black, and another auxiliary sample of White, workers are available. We seek to decompose differences in specific quantiles of the Black and White wage distributions into portions due to (i) differences in the distribution of characteristics, and (ii) differences in the mapping from those characteristics into wages, across the two populations. The latter difference is sometimes interpreted as a measure of labor market discrimination, although this interpretation is not assumption free (cf., Darity and Mason, 1998).

This decomposition requires knowledge of the distribution of White wages that would prevail under the Black distribution of worker characteristics. That is, what would the wage distribution look like in a hypothetical White population whose distribution of W coincided with the one in the actual Black population? The α^{th} quantile of this counterfactual distribution, $\gamma^{\alpha}_{W|B}$, is identified by

$$\mathbb{E}_s\left[\mathbf{1}(X \leq \gamma^{\alpha}_{W|B}) - \alpha\right] = 0,$$

which corresponds to setting $\psi_0(Y_0, X, \gamma) = \alpha - \mathbf{1}(X \leq \gamma^{\alpha}_{W|B})$ and $\psi_1(Y_1, X, \gamma)$ to a vector of zeros. The α^{th} quantiles of the actual Black and White earnings distributions are denoted by $\gamma^{\alpha}_{B|B}$ and $\gamma^{\alpha}_{W|W}$. A decomposition into wage structure and compositional effects is then given by

$$\gamma^{\alpha}_{B|B} - \gamma^{\alpha}_{W|W} = \left(\gamma^{\alpha}_{B|B} - \gamma^{\alpha}_{W|B}\right) - \left(\gamma^{\alpha}_{W|W} - \gamma^{\alpha}_{W|B}\right).$$

Barsky, Bound, Charles and Lupton (2002) and Fortin, Lemieux and Firpo (2010) survey alternative decomposition methods. For discretely-valued W these methods are similar to techniques used by demographers to standardize mortality rates across

localities (e.g., Kitagawa, 1964).

2.2 Efficiency bound

Hahn (1998, Theorem 1) calculated the semiparametric variance bound for the special case where γ_0 is the ATT and part (v) of Assumption 2.1 is not part of the prior restriction. Chen, Hong and Tarozzi (2008, Theorem 3) include 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).[2] To present this result we require some additional notation. Let

$$\Gamma_0(w) = \mathbb{E}\left[\left.\frac{\partial\psi(Z,\gamma_0)}{\partial\gamma'}\right| W = w\right], \quad p_0(w) = G(t(w)'\delta_0)$$

$$q_s(w) = \mathbb{E}[\psi_s(Y, W, \gamma_0)|\, W = w], \quad q_a(w) = \mathbb{E}[\psi_a(X, W, \gamma_0)|\, W = w]$$

$$\Sigma_s(w;\gamma_0) = \mathbb{V}(\psi_s(Y, W, \gamma_0)|\, W = w), \quad \Sigma_a(w;\gamma_0) = \mathbb{V}(\psi_a(X, W, \gamma_0)|\, W = w)$$

$$\mathbb{S}_\delta = \frac{D - G(r(W)'\delta_0)}{G(r(W)'\delta_0)[1 - G(r(W)'\delta_0)]} G_1(r(W)'\delta_0)\, r(W),$$

and

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

[2]Relative to Chen, Hong and Tarozzi (2008), our model includes an additional conditional independence assumption which influences the form of the efficiency bound.

Theorem 2.1 (SEMIPARAMETRIC VARIANCE BOUND) *Under Assumption 2.1 the maximal asymptotic precision with which γ_0 may be regularly estimated is given by the inverse of* $\mathcal{I}(\gamma_0) = \mathbb{E}\left[\frac{p_0(W)}{Q_0}\Gamma_0(W)\right]' \mathbb{E}[\Lambda(W)]^{-1} \mathbb{E}\left[\frac{p_0(W)}{Q_0}\Gamma_0(W)\right]$.

Proof. See the supplemental appendix. ■

It is easy to show that the information bound for γ_0 is smaller in the model which leaves $p_0(W)$ nonparametric (i.e., where part (v) of Assumption 2.1 is not part of the prior). Knowledge of the parametric form of the propensity score increases the large sample precision with which γ_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 better understand this difference consider estimation of the study population distribution of W. Since a random sample of W from the study population is available, an obvious estimate is the study sample empirical distribution function

$$\widehat{F}_s(w) = \frac{1}{N_s}\sum_{i=1}^{N_s} \mathbf{1}(W_i \leq w). \tag{6}$$

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. If nothing is known about the relationship between $F_s(w)$ and $F_a(w)$, as is true when the propensity score is left nonparametric, then this estimator is also efficient. However if the relationship between $F_s(w)$ and $F_a(w)$ is *a priori* restricted, as occurs when the propensity score is parametrically specified, a more efficient estimate can be constructed.

Let $G(r(w)'\widehat{\delta}_{ML})$ denote the conditional maximum likelihood estimate of the propensity score (based on the merged sample) and $\widehat{Q}_{ML} = \sum_{i=1}^{N} G(r(W_i)'\widehat{\delta}_{ML})/N$ that of Q_0, then the estimate

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

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 $\widehat{f}_s(W_i) = \widehat{\pi}_i^{\text{eff}}$, for $\widehat{\pi}_i^{\text{eff}}$ defined in (7). In contrast to (6), (7) uses *both* study and auxiliary units – linked via a parametric form for the propensity score – to efficiently estimate $F_s(w)$.

Parts (v) of Assumption 2.1 implies that we can view the auxiliary sample as a biased sampled from the study population of interest where the biasing function is known up to a finite dimensional parameter (cf., Qin, 1998; Gilbert, Lele and Vardi, 1999; Ridder and Moffitt, 2007). As is well known, 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. Equation (7) is a specific application of this general idea.

Since γ_0 involves integration over the study population distributions of (Y, W) and (X, W), these two distribution functions must be (implicitly) estimated in order to estimate γ_0. The estimator we propose in the next section improves the efficiency of these distribution function estimates by requiring them to share a finite number of moments of W in common with $\widehat{F}_s^{\text{eff}}(w)$. The idea of calibrating a distribution func-

tion estimate to information garnered from auxiliary sources arises in other contexts. Little and Wu (1991) discuss contingency table calibration to known margins and provide historical references (cf., Hellerstein and Imbens, 1999). Bickel, Ya'Acov and Wellner (1991) study estimation of linear functionals of probability measures with known marginals. Hirano, Imbens, Ridder and Rubin (2001) show 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) calibrates estimates of the two potential outcome distributions to features of the empirical distribution of always observed variables (cf., Qin and Zhang, 2007; Graham, Pinto and Egel, 2012). Recently Cheng, Small, Tan, and Ten Have (2009) apply related ideas to an instrumental variables model.

We calibrate our estimates of the study population distributions of (Y, W) and (X, W) to features of (7) (which is the most efficient estimate of $F_s(w)$ when the propensity score takes a parametric form). 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 (cf., Graham, 2011).

3 Auxiliary-to-Study Tilting

Our estimator for γ_0, which we call the auxiliary-to-study tilting (AST) estimator, is a sequential method of moments estimator, as surveyed by, for example, Newey

and McFadden (1994). In the first step we estimate the propensity score parameter δ by conditional maximum likelihood:

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

In the second step we compute a reweighting of both the study and auxiliary samples. Let $t\left(W\right)$ be vector of known linear independent functions of W with a constant 1 in the first row and λ_a and λ_s be 'tilting' parameters of the same dimension. We allow for $r\left(W\right)$ and $t\left(W\right)$ to include common elements or even coincide. Fixing δ at $\widehat{\delta}_{ML}$ and Q at $\widehat{Q}_{ML}=\sum_{i=1}^{N}G(t(W_i)'\widehat{\delta}_{ML})/N$ we choose $\widehat{\lambda}_a$ to solve:

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

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

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

for

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

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

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

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

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

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

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

We also compute an analogous tilt of the study sample

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

so that

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

for

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

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

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

Inspection of (13) indicates that our estimate of γ_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 $t\left(W_i\right)$, computed using the two tilts, coincide with the efficient estimate.

Our next two results provide formal descriptions of the asymptotic sampling properties of $\widehat{\gamma}_{AST}$ under different combinations of assumptions. We begin by introducing the following assumption.

Assumption 3.1 (MOMENT CEF) *For some unique pair of matrices* Π_s, Π_a *and vector of linear independent functions* $t\left(W\right)$ *with a constant in the first row, we have*

$$\mathbb{E}\left[\psi_s\left(Y, W, \gamma_0\right) | W\right] = \Pi_s t\left(W\right), \quad \mathbb{E}\left[\psi_a\left(X, W, \gamma_0\right) | W\right] = \Pi_a t\left(W\right).$$

Assumption 3.1 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 3.1 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.

Let $\mathbb{E}^*[Y|X]$ denote the mean squared error minimizing linear predictor of Y given X. If both Assumptions 2.1 and 3.1 hold the Appendix shows that $\widehat{\gamma}_{AST}$ is asymptotically linear with representation

$$\begin{aligned}\sqrt{N}(\widehat{\gamma}_{AST} - \gamma_0) &= -\Gamma_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\frac{D_i}{Q_0}\{\psi_s(Y_i, W_i, \gamma_0) - q_s(W_i)\}\right. \\ &\quad -\frac{1-D_i}{Q_0}\frac{p_0(W_i)}{1-p_0(W_i)}\{\psi_a(X_i, W_i, \gamma_0) - q_a(W_i)\} \\ &\quad +\frac{p_0(W)}{Q_0}\{q_s(W) - q_a(W)\} \\ &\quad \left.+\frac{1}{Q_0}\mathbb{E}^*\left[\left(\frac{D}{p_0(W)} - 1\right)p_0(W)(q_s(W) - q_a(W))\middle|\mathbb{S}_\delta\right]\right\} + o_p(1).\end{aligned} \tag{14}$$

Equation (14) then gives our asymptotic efficiency result.

Theorem 3.1 (Local Semiparametric Efficiency) *Consider the semiparametric data combination model defined by Assumption 2.1 and additional regularity conditions, then for $\widehat{\gamma}_{AST}$ the solution to (13), $\widehat{\gamma}_{AST}$ is locally efficient at Assumption 3.1 such that $\sqrt{N}(\widehat{\gamma}_{AST} - \gamma_0) \xrightarrow{D} \mathcal{N}\left(0, \mathcal{I}(\gamma_0)^{-1}\right)$ with $\mathcal{I}(\gamma_0)$ as defined in Theorem*

2.1.

Proof. See Appendix A. ■

Our efficiency bound calculation, Theorem 2.1, gives the information bound for γ_0 without imposing the additional auxiliary Assumption 3.1. 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 γ_0 more precisely. Our estimator is not efficient with respect to this augmented model. Rather it attains the bound provided by Theorem 2.1 if Assumption 3.1 *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 - 3) and Tsiatis (2006) discuss the concept of local efficiency in detail. In what follows we will, for brevity, say $\widehat{\gamma}_{AST}$ is locally efficient at Assumption 3.1.

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

Theorem 3.2 (DOUBLE ROBUSTNESS) *Under parts (i) to (iv) of Assumption 2.1,* $\widehat{\gamma}_{AST} \xrightarrow{p} \gamma_0$ *with a limiting normal distribution if* ***either*** *(a) part (v) of Assumption 2.1 also holds* ***or*** *(b) the analyst chooses* $G(v) = \exp(v) / [1 + \exp(v)]$, $t(W) = r(W)$, *and Assumption 3.1 holds.*

Proof. See Appendix A. ■

Theorem 3.2 indicates that the advantage of choosing $t(W)$ with Assumption 3.1 in mind is twofold. Under the baseline model defined by Assumption 2.1, Theorem 3.1 implies that $\widehat{\gamma}_{AST}$ 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 $t(w)$. This is the case covered by part (a) of the Theorem. Now consider the case where the analyst mispecifies the propensity score model, but Assumption 3.1 holds, part (b) of Theorem 3.2 indicates that $\widehat{\gamma}_{AST}$ will remain consistent for γ_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 3.2, relative to analogous results in the missing data literature (e.g., Tsiatis, 2006), is the requirement that the assumed propensity score take 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 γ_0. However when condition (b) of Theorem 3.2 holds we have, from the estimating equations for the propensity score parameter,

$$\frac{1}{N}\sum_{i=1}^{N}\left(D_i - G\left(t(W_i)'\widehat{\delta}_{ML}\right)\right)t(W_i) = 0. \tag{15}$$

Now consider the mean of $t(W_i)$ with respect to $\widehat{F}_s^{\text{eff}}(w)$. Using (15), and the fact that $t(W_i)$ contains a constant so that $\sum_{i=1}^{N} G(t(W_i)'\widehat{\delta}_{ML}) = \sum_{i=1}^{N} D_i$, we have the equalities

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

Therefore, under the conditions of part (b) of Theorem 3.2, $\sum_{i=1}^{N}\widehat{\pi}_i^{\text{eff}} t(W_i) \xrightarrow{p} \mathbb{E}_s[t(W)]$ 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 3.1 holds, $\widehat{\gamma}_{AST}$ will remain consistent for γ_0.

We note that, unlike in the missing data problem, where the propensity score is ancillary, it is surprising that *any* data combination estimator is consistent in the presence of propensity score misspecification since it enters the actual definition of γ_0 :

$$\mathbb{E}\left[\psi\left(Z,\gamma_0\right)|\, D=1\right] = Q_0^{-1}\int \psi\left(z,\gamma_0\right) p_0\left(w\right) f\left(z\right) \mathrm{d}z.$$

Collectively Theorems 3.1 and 3.2 provide a strong theoretical case for using AST in practice. If Assumption 3.1 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. Furthermore, if the propensity score is inadvertently misspecified, AST nevertheless remains consistent for γ_0 if Assumption 3.1 holds (and the analyst works with a logit form for $G\left(v\right)$). We acknowledge that in settings where the researcher is highly confident in Assumption 3.1 a direct imputation approach may be preferable (e.g., Kline, 2011; Chen, Hong and Tarozzi, 2008). Such an approach is valid under weaker support conditions than maintained by Assumption 2.1. A disadvantage of imputation is its sensitivity to violations of Assumption 3.1; this limitation is illustrated by our Monte Carlo experiments below.

4 Application and Monte Carlo experiments

Empirical application Neal and Johnson (1996) study the role of 'pre-market' (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) compute the least squares fit of the logarithm of hourly wages on a constant, a black dummy, age, and AFQT percentile score measured at age 16 to 18.[3] They find 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 argue 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, explains a substantial portion of subsequent Black-White wage inequality. We do not provide an assessment of this interpretation here, rather we are interested in the sensitivity of their statistical finding to their maintained (linear) functional form assumptions.

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

[3]The Armed Forces Qualification Test (AFQT) is used by the military for recruitment and job assignment purposes. It is widely used as a measure of cognitive achievement in social science research. The AFQT is a nationally normed test so that an individual's percentile score corresponds to her rank in the reference distribution.

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, Bound, Charles and Lupton (2002) and Fortin, Lemieux and Firpo (2010) survey the application of decomposition methods in economics.

Our sample closely resembles that used in Johnson and Neal (1998).[4] It includes 1,371 measurements of real wages, race, age and AFQT score 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 are 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 percent 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. For example, while over 45 percent of Blacks earn less than \$7.50 per hour in our sample, fewer than 30 percent of Whites do (Table 1, Row 3). Inspection of the CDF differences indicates that, while the

[4]We attempted to exactly reconstruct the Johnson and Neal (1998) sample by following the guidelines in their data appendix. Our sample differs form theirs negligibly, perhaps due to updates in the NLSY79 databases since their research was undertaken.

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, two year of birth dummies, a quadratic polynomial in transformed AFQT score, and twelve dummy variables for the transformed AFQT score lying respectively below $-2, -1.75, \ldots, 0.25, 0.5$. 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 2.1. In small samples the performance of the estimator is heavily dependent on the quality of the two CEF approximations. After adjusting for age and AFQT differences we find that, while a Black-White residual log wage CDF gap remains at lower wage values, it disappears at higher values. The average log wage gaps falls, after adjusting for age and AFQT differences, from -0.279 to -0.111.

Column 2 of Panel B implements the propensity score reweighting (PSR) estimator of Hirano and Imbens (2001) and Abadie (2005). We model the propensity

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

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

NOTES: Results based on an extract of 1,371 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.

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 $-2, -1.75, \ldots, 0.25, 0.5$ 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.

While the AST point estimates are similar to the corresponding CEP and PSR ones, their estimated sampling precision is uniformly superior (as Theorem 3.1 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 2.1 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

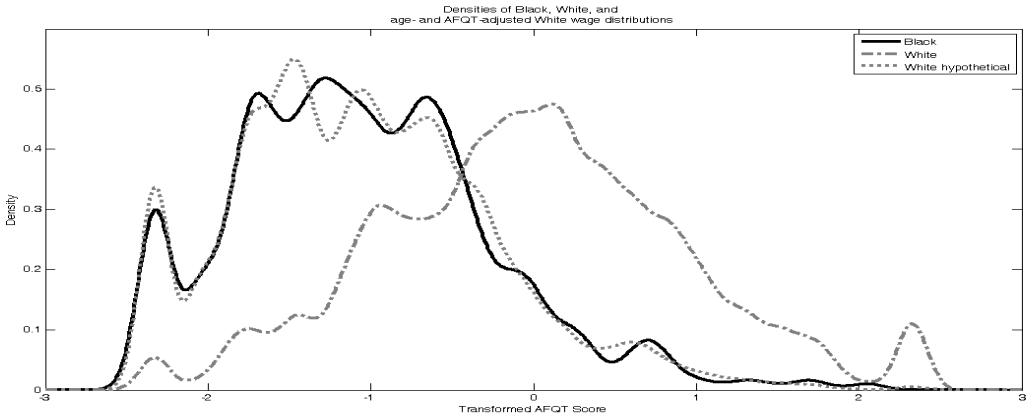

Figure 1: AFQT Densities
NOTES: The figure plots kernel density estimates of the actual Black and White AFQT score distributions as well as an estimate based on the auxiliary sample tilt. A Gaussian kernel is used with a bandwidth equal to 1/2 of Silverman's 'rule-of-thumb' choice. Undersmoothing highlights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

yield different point estimates depending on the true data generating process.

Our empirical application does generate new substantive findings relative to those of Neal and Johnson (1996). These are most easily described by reference to Figure 2.[5] Panel A of this figure plots differences in the quantiles of the unadjusted Black versus White log wage distributions. Panel B plots the same differences after adjusting for year of birth and AFQT differences using our AST procedure with $t(W)$ as described above (i.e., differences in the quantiles of the study versus auxiliary sample tilts). The shaded area in the two figures correspond to 95 percent pointwise confidence intervals. These intervals were computed using a percentile bootstrap with 1000 replications (sampling households with replacement). While the raw wage distributions differ significantly at all quantiles, after adjusting for year of birth and

[5]These quantiles are computed by numerically inverting the relevant AST distribution function estimate.

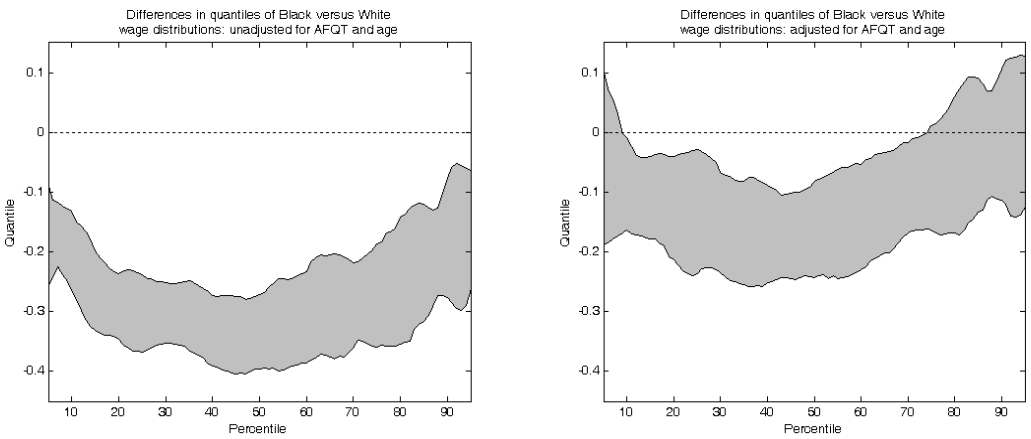

Figure 2: Actual and age- and AFQT-adjusted differences in the quantiles of the Black versus White log wage distributions
NOTES: Shaded areas correspond to 95 percent pointwise percentile bootstrap confidence intervals.

AFQT differences, they do not significantly differ for lower and higher quantiles. If we adopt the same interpretative perspective as Neal and Johnson (1996), our results are consistent with the conclusion that explicit labor market discrimination is less severe at the low and high ends of the Black wage distribution, and most pronounced in the middle of the wage distribution. The regression methods used by Neal and Johnson (1996) preclude the discovery of these heterogeneous effects. Indeed the average age and AFQT-adjusted wage gaps reported in row 1 of Table 1 are only two-thirds of the difference of medians reported in Figure 2.

Monte Carlo We now report on a number of Monte Carlo experiments we conducted to verify the theoretical properties described in Theorems 3.1 and 3.2. 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).

Table 2: Parameter values for Monte Carlo experiments

Design	(1)	(2)	(3)	(4)
σ_a^2	1	2/3	1	2/3
σ_Y^2	3.4823	2.6590	1.7496	0.9253
α_2	0	0	-1	-1

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 (μ_s, σ_s^2) and (μ_a, σ_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

$$\begin{aligned} Y \,|\, W, D &\sim \mathcal{N}\left(0, \sigma_Y^2\right) \\ X \,|\, W, D &\sim \mathcal{N}\left(\alpha_0 + \alpha_1\left(W - \mu_{W|D=1}\right) + \alpha_2\left[\left(W - \mu_{W|D=1}\right)^2 - \sigma_{W|D=1}^2\right], \sigma_X^2\right), \end{aligned}$$

where $\mu_{W|D=1}$ and $\sigma_{W|D=1}^2$ are the study population mean and variance of W (which differ from μ_s and σ_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 :

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

where β_0, β_1 and β_2 are functions of (μ_s, σ_s^2) and (μ_a, σ_a^2) (cf., 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 $N = 1,000$ 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 σ_A^2 and α_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 σ_Y^2 across designs to keep the variance bound fixed. Across each of our designs an efficient estimator (under Assumption 2.1) 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 three the reverse is true, while in Design four 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 γ_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 gives the median bias across 5,000 Monte Carlo replications (in both cases bias is scaled by the 'pencil and paper' asymptotic standard error reported in

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

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 percent confidence intervals for γ_0 (Column 6). By comparing columns 1 and 2 and columns 3 and 5, we see that – for the designs considered – the finite sample distributions of all of the estimators are very well approximated by their asymptotic counterparts.

Recently Qin and Zhang (2008) have proposed an empirical likelihood type estimator for the difference-in-differences program evaluation parameter (e.g., Abadie, 2005). This parameter may be viewed as a special case of the average treatment effect on the treated (ATT) parameter. Their procedure, like ours, calibrates estimates of the study population distributions of (Y, W) and (X, W) to features of $\widehat{F}_s^{\text{eff}}(w)$. They use empirical likelihood methods for this purpose, as opposed to our 'tilting' equations (9) and (11). In order to compare our method with the Qin and Zhang (2008) EL procedure we replicated a subset of their Monte Carlo experiments. Adapting their setup to our notation we let

$$W_1 \sim \mathcal{N}(0,1), \qquad W_2 | W_1 = w_1 \sim \mathcal{N}(1 + 0.6w_1, 1),$$

and

$$Y | W, D \sim \mathcal{N}\left(\mu_Y(W), W_2^2\right), \qquad X | W, D \sim \mathcal{N}\left(\mu_X(W), W_2^2\right).$$

They assume the propensity score takes a logit form with an index linear in $W = (W_1, W_2)'$ (this in turn induces the conditional distributions of W given $D = 0, 1$). The intercept in the logit index is set equal to one across all designs, while the two slope coefficients equal 0.1, 0.2 or 0.5 (corresponding to increasing selection bias).

The two conditional mean parameters are set equal to $\mu_Y(W) = 2 + 2W_1 + 2W_2$ and $\mu_X(W) = 2W_1 + 2W_2$ in Design (a) and $\mu_Y(W) = 2 + 2W_1^2 - W_2 + 3W_2^2$ and $\mu_X(W) = 2W_1^2 - W_2 + 3W_2^2$ in Design (b). Analogously to Qin and Zhang (2008) we choose two different specifications for $t(W)$. First, a 'linear' one of $t(W) = (1, W_1, W_2)'$. This corresponds to the locally efficient choice in Design (a). Second, a 'quadratic' one of $t(W) = (1, W_1^2, W_2^2)'$. This choice in not efficient in either design, but is expected to be more appropriate for Design (b). Across all designs the propensity score is correctly specified with $r(W) = (1, W_1, W_2)'$. We set $N = 1,000$ and perform $1,000$ Monte Carlo replications. The Monte Carlo statistics for the EL estimator are as reported in Table 2 of Qin and Zhang (2008, p. 341).

By Theorems 3.1 and 3.2 above, and Theorem 3 of Qin and Zhang (2008, p. 339), both the AST estimator and the EL estimator should be consistent and asymptotically normal across both designs and choices of $t(W)$. Our AST estimator should be efficient in Design (a) when $t(W)$ takes the linear form. (see Table 5 in the supplemental appendix).

In Design (a) the AST and EL estimator perform similarly in terms of bias (see Table 4). However, when $t(W)$ is (correctly) specified to be linear in W, AST has substantially less sampling variation that the EL estimator (consistent with Theorem 3.1). This effect is largest when selection bias is severe. In that case the sampling variation in the AST estimate is just over one half that of the EL one. When $t(W)$ is (incorrectly) specified to be quadratic, this efficiency ranking reverses. In Design (b) the EL estimate exhibits lower sample variation than the corresponding AST estimate when $t(W)$ is (incorrectly) specified to be linear. When $t(W)$ is quadratic, which more closely approximates the efficient choice, this ranking is reversed. As before, the efficiency gains are increasing in the degree of selection bias. In terms of

inference the AST Wald confidence intervals generally have actual coverage close to nominal coverage, while the corresponding EL ones tend to be conservative (Qin and Zhang (2008) suggest the use of boostrap confidence intervals in order to improve coverage).

While Qin and Zhang (2008) do not consider the semiparametric efficiency properties of their procedure, the results in Table 4 suggest that, in contrast to AST, their estimator is not Locally Efficient at Assumption 3.1 (although this is only a conjecture based on the Monte Carlo results). Evidently the comparison of the two estimators when Assumption 3.1 does not hold is more complicated.

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 2.1, 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 type 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 3.1) we can produce a locally efficient estimate of γ_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 γ_0 if Assumption 3.1 holds.

To our knowledge we are the first to propose a locally efficient estimator for the

Table 4: Monte Carlo results: Qin and Zhang (2008) designs with N =1,000

			(1) Mean Bias	(2) Sample Var.	(3) Mean Est. Var.	(4) RMSE	(5) Cov. of 95% CI
(β_1, β_2)		$t(W)$		**Design (a): Linear CEFs**			
$(0.1, 0.1)$	AST	Lin	-0.0004	0.0154	0.0151	0.1241	0.936
	AST	Qrd	-0.0083	0.0285	0.0513	0.1690	0.988
	EL	Lin	0.0038	0.0204	0.0311	0.1429	0.981
	EL	Qrd	0.0040	0.0241	0.0357	0.1553	0.978
$(0.2, 0.2)$	AST	Lin	-0.0065	0.0216	0.0195	0.1471	0.930
	AST	Qrd	-0.0039	0.0371	0.0555	0.1926	0.983
	EL	Lin	0.0031	0.0275	0.0402	0.1659	0.975
	EL	Qrd	-0.0009	0.0306	0.0430	0.1749	0.972
$(0.5, 0.5)$	AST	Lin	0.0024	0.0537	0.0428	0.2316	0.907
	AST	Qrd	0.0244	0.1015	0.0867	0.3193	0.920
	EL	Lin	0.0051	0.0900	0.7241	0.3000	0.912
	EL	Qrd	-0.0089	0.1103	0.5842	0.3322	0.891
				Design (b): Quadratic CEFs			
$(0.1, 0.1)$	AST	Lin	0.0009	0.3050	0.2856	0.5520	0.942
	AST	Qrd	-0.0011	0.0168	0.0174	0.1297	0.947
	EL	Lin	0.0347	0.1561	0.2003	0.3966	0.966
	EL	Qrd	0.0029	0.0226	0.1181	0.1504	0.995
$(0.2, 0.2)$	AST	Lin	0.0787	0.3620	0.3201	0.6065	0.916
	AST	Qrd	0.0078	0.0218	0.0217	0.1479	0.951
	EL	Lin	0.0477	0.1227	0.3790	0.3535	0.980
	EL	Qrd	0.0028	0.0309	0.4564	0.1758	0.998
$(0.5, 0.5)$	AST	Lin	0.1943	0.7010	0.4425	0.8591	0.817
	AST	Qrd	0.0095	0.0549	0.0429	0.2343	0.906
	EL	Lin	0.1969	0.2647	3.2656	0.5509	0.959
	EL	Qrd	0.0075	0.1026	2.1138	0.3204	0.993

class of data combination problems defined by Assumption 2.1. Our procedure also has a double robustness type 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 do 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 interesting to study data dependent methods for choosing $t(W)$.

A Proofs

Proof of Theorem 3.1: Let $m(Z_i, \theta_0)$ be the i^{th} unit's contribution to $\dim(r(W)) + 2\dim(t(W)) + \dim(\gamma_0)$ vector of estimating equations defined by (8), (9), (11) and (13) in the main text. Let $M = \mathbb{E}[\partial m(Z, \theta_0)/\partial\theta']$; a standard calculation gives the asymptotically linear representation

$$\sqrt{N}\left(\widehat{\theta} - \theta_0\right) = -M^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} m(Z_i, \theta_0)\right) + o_p(1). \tag{16}$$

The influence function for $\widehat{\gamma}_{AST}$ corresponds to the last K elements of (16). By tedious, but straightforward, calculation we can show that this subvector equals

$$\begin{aligned}\sqrt{N}(\widehat{\gamma} - \gamma_0) = \frac{-M_{44}^{-1}}{\sqrt{N}}\sum_{i=1}^{N}\Big\{ & m_4(Z_i, \delta_0, \lambda_{a0}, \lambda_{s0}, \gamma_0) - M_{41}M_{11}^{-1}m_1(Z_i, \delta_0) \\ & + M_{42}M_{22}^{-1}\left(M_{21}M_{11}^{-1}m_1(Z_i, \delta_0) - m_2(Z_i, \delta_0, \lambda_{a0})\right) \\ & + M_{43}M_{33}^{-1}\left(M_{31}M_{11}^{-1}m_1(Z_i, \delta_0) - m_3(Z_i, \delta_0, \lambda_{s0})\right)\Big\} + o_p(1).\end{aligned} \tag{17}$$

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 θ evaluated at $\theta = \theta_0$.

Under part (v) of Assumption 2.1 the Information Matrix equality gives $M_{11} = -\mathbb{E}[\mathbb{S}_\delta \mathbb{S}_\delta']$. Evaluating M_{21} yields, after some manipulation,

$$M_{21} = -\mathbb{E}\left[\left(\frac{1-D}{1-p_0(W)} - 1\right) p_0(W)\, t(W)\, \mathbb{S}_\delta'\right], \tag{18}$$

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

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

with $\mathbb{E}^*[Y|X]$ denoting the mean squared error minimizing linear predictor (LP) of Y given X. Evaluating M_{22} and M_{42} yields

$$M_{22} = \mathbb{E}\left[\frac{p_0(W)}{1-p_0(W)} G_1\left(r(W)'\delta_0\right) t(W)\, t(W)'\right] \tag{19}$$

$$M_{42} = -\frac{1}{Q_0}\mathbb{E}\left[\frac{p_0(W)}{1-p_0(W)} G_1\left(r(W)'\delta_0\right) \psi_a(X,W,\gamma_0)\, t(W)'\right] \tag{20}$$

Assumption 3.1 then gives $M_{42}M_{22}^{-1} = -\Pi_a/Q_0$. Similar calculations give

$$M_{31} = -\mathbb{E}\left[\left(\frac{D}{p_0(W)} - 1\right) p_0(W)\, t(W)\, \mathbb{S}_\delta'\right] \tag{21}$$

yielding

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

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

$$M_{33} = -\mathbb{E}\left[G_1\left(r\left(W\right)'\delta_0\right)t\left(W\right)t\left(W\right)'\right] \tag{22}$$

$$M_{43} = -\frac{1}{Q_0}\mathbb{E}\left[G_1\left(r\left(W\right)'\delta_0\right)\psi_s\left(X,W,\gamma_0\right)t\left(W\right)'\right]. \tag{23}$$

Assumption 3.1 then gives $M_{43}M_{33}^{-1} = \Pi_s/Q_0$. Tedious calculations give M_{41} equal to

$$M_{41} = \frac{1}{Q_0}\mathbb{E}\left[\frac{1-D}{1-p_0\left(W\right)}\psi_a\left(X,W,\gamma_0\right)\mathbb{S}_\delta'\right]. \tag{24}$$

Using this result, iterated expectations and part (ii) of Assumption 2.1 we then get

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

Substituting the above results into (17) and manipulating then gives (14).

Proof of Theorem 3.2: 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 $\widehat{\delta}$, $\widehat{\lambda}_a$,and $\widehat{\lambda}_s$ when part (v) of Assumption 2.1 fails to hold by, respectively δ_*, λ_{a*}, and λ_{s*}. Let $p_*\left(W\right) = G\left(r\left(W\right)'\delta_*\right)$ and $p_j\left(W\right) = G\left(r\left(W\right)'\delta_* + t\left(W\right)'\lambda_{j*}\right)$ for $j = s, a$. If $G\left(\cdot\right)$ takes the logit form, then $p_*\left(W\right)$ will satisfy the population restriction $\mathbb{E}\left[m_1\left(Z,\delta_*\right)\right] = \mathbb{E}\left[\left(D - p_*\left(W\right)\right)t\left(W\right)\right] = 0$ so that, using iterated expectations and rearranging, we have the equality.

$$\mathbb{E}\left[t\left(W\right)\middle|\,D=1\right] = \mathbb{E}\left[\frac{p_*\left(W\right)}{Q_0}t\left(W\right)\right]. \tag{25}$$

We also have $\mathbb{E}[m_2(Z, \delta_*, \lambda_{a*})] = \mathbb{E}[m_3(Z, \delta_*, \lambda_{s*})] = 0$, which, respectively multiplying by Π_a and Π_s (using Assumption 3.1), gives the additional equalities:

$$\mathbb{E}\left[\frac{1-D}{1-p_a(W)}p_*(W)\,q_a(W)\right] = \mathbb{E}[p_*(W)\,q_a(W)] \quad (26)$$

$$\mathbb{E}\left[\frac{D}{p_s(W)}p_*(W)\,q_s(W)\right] = \mathbb{E}[p_*(W)\,q_s(W)]. \quad (27)$$

Using (25), (26), (27), Assumption 3.1, iterated expectations, and part (ii) of Assumption 2.1 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 2.1 is uniquely zero at $\gamma = \gamma_0$.

References

[1] Abadie, Alberto. (2005). "Semiparametric difference-in-differences," *Review of Economic Studies* 72 (1): 1 - 19.

[2] Abadie, Alberto and Guido W. Imbens. (2006). "Large sample properties of matching estimators for average treatment effects," *Econometrica* 74 (1): 235 - 267.

[3] Anderson, J.A. (1982). "Logistic discrimination," *Handbook of Statistics* 2: 169 - 191 (P.R. Krishnaiah & L.N. Kanal, Eds.). Amsterdam: North-Holland.

[4] Angrist, Joshua D. and Alan B. Krueger. (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 (418): 328 - 336.

[5] Barsky, Robert, John Bound, Kerwin Ko' Charles and Joseph P. Lupton. (2002). "Accounting for the black-white wealth gap: a nonparametric approach," *Journal of the American Statistical Association* 97 (459): 663 - 673.

[6] Bickel, Peter J., Ya'Acov Ritov and Jon A. Wellner. (1991). "Efficient estimation of linear functionals of a probability measure P with known marginal distribitions," *Annals of Statistics* 19 (3): 1316 - 1346.

[7] Björklund, Anders and Markus Jäntti. (1997). "Intergenerational income mobility in Sweden compared to the United States," *American Economic Review* 87 (5): 1009 - 1018.

[8] Carroll, R. J. and M. P. Wand. (1991). "Semiparametric estimation in logistic measurement error models," *Journal of the Royal Statistical Society B* 53 (3): 573 - 585.

[9] Chen, Xiaohong, Han Hong and Alessandro Tarozzi. (2008). "Semiparametric efficiency in GMM models with auxiliary data," *Annals of Statistics* 36 (2): 808 - 843.

[10] Cheng, Jing, Dylan S. Small, Zhiqiang Tan, and Thomas R. Ten Have. (2009). "Efficient nonparametric estimation of causal effects in randomized trials with noncompliance," *Biometrika* 96 (1): 19 - 36.

[11] Currie, Janet and Aaron Yelowitz. (2000). "Are public housing projects good for kids?" *Journal of Public Economics* 75 (1): 99 - 124

[12] Darity; William A. and Patrick L. Mason. (1998). "Evidence on discrimination in employment: codes of color, codes of gender," *Journal of Economic Perspectives* 12 (2): 63 - 90.

[13] Dehejia, Rajeev H. and Sadek Wahba. (1999). "Causal effects in nonexperimental studies: reevaluating the evaluation of training programs," *Journal of the American Statistical Association* 94 (448): 1053 - 1062.

[14] Dinardo, John, Nicole M. Fortin, Thomas Lemieux. (1996). "Labor market institutions and the distribution of wages, 1973 - 1992: a semiparametric approach," *Econometrica* 64 (5): 1001 - 1044.

[15] Elbers, Chris, Jean O. Lanjouw and Peter Lanjouw. (2003). "Micro–level estimation of poverty and inequality," *Econometrica* 71 (1): 355 - 364.

[16] Firpo, Sergio and Cristoph Rothe. (2013). "Semiparametric estimation and inference using doubly robust moment conditions," *Mimeo.*

[17] Fortin, Nicole, and Thomas Lemieux, and Sergio Firpo. (2011). "Decomposition methods in economics," *Handbook of Labor Economics 4A*: 1 - 102 (O. Ashenfelter & D. Card, Eds.). Amsterdam: North-Holland.

[18] Gilbert, Peter B., Subhash R. Lele and Yehuda Vardi. (1999). "Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials," *Biometrika* 86 (1): 27 - 43.

[19] Graham, Bryan S. (2011). "Efficiency bounds for missing data models with semiparametric restrictions," *Econometrica* 79 (2): 437 - 452.

[20] Graham, Bryan S., Cristine Campos de Xavier Pinto, and Daniel Egel. (2012). "Inverse probability tilting for moment condition models with missing data," *Review of Economic Studies* 79 (3): 1053 - 1079.

[21] Hahn, Jinyong. (1998). "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica* 66 (2): 315 - 331.

[22] Heckman, James J. and Robb. (1985).

[23] Hellerstein, Judith K. and Guido W. Imbens. (1999). "Imposing moment restrictions from auxiliary data by weighting," *Review of Economics and Statistics* 81 (1): 1 - 14.

[24] Hirano, Keisuke and Guido W. Imbens. (2001). "Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization," *Health Services and Outcomes Research* 2 (3-4): 259 -278.

[25] Hirano, Keisuke, Guido W. Imbens and Geert Ridder. (2003). "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica* 71 (4): 1161 - 1189.

[26] Hirano, Keisuke, Guido W. Imbens, Geert Ridder, Donald B. Rubin. (2001). "Combining panel data sets with attrition and refreshment samples," *Econometrica* 69 (6): 1645 - 1659.

[27] Ichimura, Hidehiko and Elena Martinez-Sanchis. (2004). "Identification and estimation of GMM models by a combination of two data sets," *Mimeo.*

[28] Imbens, Guido W. (2004).

[29] Johnson, William R. and Derek A. Neal (1998). "Basic skills and the black-white earnings gap," *The Black-White Test Score Gap*: 480 - 500. (C. Jencks & M. Phillips, Eds.). Washington, D.C.: The Brookings Institution.

[30] Khan, Shakeeb and Elie Tamer (2010). "Irregular identification, support conditions, and inverse weight estimation," *Econometrica* 78 (6): 2021 - 2042.

[31] Kitagawa, Evelyn M. (1964). "Standardized comparisons in population research," *Demography* 1 (1): 296 - 315.

[32] Kline, Patrick. (2011). "Oaxaca-Blinder as a reweighting estimator," *American Economic Review Papers & Proceedings*, forthcoming.

[33] Lalonde, Robert J. (1986). "Evaluating the econometric evaluations of training programs," *American Economic Review* 76 (4): 604 - 620.

[34] Little, Roderick J.A. and Mei-Miau Wu. (1991). "Models for contingency tables with known margins when target and sampled populations differ," *Journal of the American Statistical Association* 86 (413): 87 - 95.

[35] Neal, Derek A. and William R. Johnson. (1996). "The role of premarket factors in black-white wage differences," *Journal of Political Economy* 104 (5): 869 - 895.

[36] Newey, Whitney K. (1990). "Semiparametric efficiency bounds," *Journal of Applied Econometrics* 5 (2): 99 - 135.

[37] Newey, Whitney K. and Daniel McFadden. (1994). "Large sample estimation and hypothesis testing," *Handbook of Econometrics* 4: 2111 - 2245 (R.F. Engle & D.L. McFadden, Eds.). Amsterdam: North-Holland.

[38] Qin, Jing. (1998). "Inferences for case-control and semiparametric two-sample density ratio models," *Biometrika* 85 (3): 619 - 630.

[39] Qin, Jing, and Biao Zhang. (2007). "Empirical-likelihood-based inference in missing response problems and its application in observational studies," *Journal of the Royal Statistical Society: Series B* 69 (1): 101 – 122.

[40] Qin, Jing, and Biao Zhang. (2008). "Empirical-likelihood-based difference-in-differences estimators," *Journal of the Royal Statistical Society B* 70 (2): 329 - 349.

[41] Ridder, Geert and Robert Moffitt. (2007). "The Econometrics of Data Combination," *Handbook of Econometrics* 6B: 5469 - 5547 (J.J. Heckman & E.E. Leamer, Ed.). Amsterdam: North-Holland.

[42] Robins, James M., Andrea Rotnitzky and Lue Ping Zhao. (1994). "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association* 89 (427): 846 - 866.