Inverse probability tilting for moment condition models with missing data, supplemental material: Additional proofs, Monte Carlo experiments, and further details on the empirical application

This supplemental web appendix contains additional proofs, summarizes the results of a series of Monte Carlo experiments, and provides further details on the empirical application. All notation is as defined in the main text unless explicitly defined otherwise. Equation and Table numbering continues in sequence with that established in the main text. To simplify notation let β denote the true parameter value β_0 unless explicitly stated otherwise (similarly the '0' subscript is removed from other objects, such as the propensity score, when doing so does not cause confusion).

Proof of Proposition 2.1 Wooldridge (2007) proves consistency and asymptotic normality of IPW M-estimators. Here we derive the variance expression given in Proposition 2.1. Let $\beta = (\gamma', \delta')'$ and define the moment vector and derivative matrix

$$m_i(\beta) = \begin{pmatrix} \frac{D_i}{G_i(\delta)}\psi_i(\gamma) \\ \frac{D_i - G_i(\delta)}{G_i(\delta)(1-G_i(\delta))}G_{1i}(\delta)\, t_i \end{pmatrix}, \; M_i(\beta) = \begin{bmatrix} \frac{D_i}{G_i(\delta)}\frac{\partial \psi_i(\beta)}{\partial \gamma'} & -\frac{D_i G_{1i}(\delta)}{G_i(\delta)^2}\psi_i(\gamma)\, t_i' \\ 0 & -J_i(\delta) \end{bmatrix}, \tag{52}$$

where $J_i(\delta)$ is the i^{th} unit's contribution to the Hessian of the log-likelihood for the propensity score parameter δ. The solution to $\sum_{i=1}^{N} m_i\left(\widehat{\beta}\right)/N = 0$ corresponds to the IPW estimate. The covariance of m_i is given by

$$\Omega = \begin{pmatrix} \mathbb{E}\left[\frac{\psi\psi'}{G}\right] & \mathbb{E}\left[\frac{G_1}{G}\psi t'\right] \\ \mathbb{E}\left[\frac{G_1}{G}t\psi'\right] & \mathcal{I}(\delta) \end{pmatrix}, \tag{53}$$

while the population mean of M_i equals

$$M = \begin{pmatrix} \Gamma & -\mathbb{E}\left[\frac{G_1}{G}\psi t'\right] \\ 0 & -\mathcal{I}(\delta_0) \end{pmatrix}, \tag{54}$$

with $\mathcal{I}(\delta_0)$ the Fisher information for δ_0.

Standard results on GMM imply that $\sqrt{N}\left(\widehat{\beta} - \beta_0\right)$ has a limiting sampling vari-

ance of

$$M^{-1}\Omega M^{-1\prime} = \begin{pmatrix} \Gamma^{-1}\mathbb{E}\left[\frac{\psi\psi'}{G}\right]\Gamma^{-1\prime} - \Gamma^{-1}\mathbb{E}\left[\frac{G_1}{G}\psi t'\right]\mathcal{I}\left(\delta_0\right)^{-1}\mathbb{E}\left[\frac{G_1}{G}t\psi'\right]\Gamma^{-1\prime} & 0 \\ 0 & \mathcal{I}\left(\delta_0\right)^{-1} \end{pmatrix}. \tag{55}$$

An insightful rearrangement of the upper-left-hand block of (55) is (see the detailed calculations supplement)

$$\mathcal{I}\left(\gamma_0\right)^{-1} + \Gamma^{-1}\mathbb{E}\left[\left(\left(\frac{D}{G}-1\right)q - \Pi_S S_\delta\right)\left(\left(\frac{D}{G}-1\right)q - \Pi_S S_\delta\right)'\right]\Gamma^{-1\prime}, \tag{56}$$

with Π_S as defined in the statement of Proposition 2.1. Part (ii) of the result follows by inspection.

The asymptotic variance of three-step AIPW estimators Here we summarize the first order asymptotic properties of a class of three-step AIPW estimators under Assumptions 1.1 to 1.5. This class includes the estimator proposed by Robins, Rotnitzky and Zhao (1994). As well as the variants proposed by Newey (1994), Hirano and Imbens (2001), and Cao, Tsiatis and Davidian (2009). While the first order properties of AIPW are well-known, we include the results below as they will prove useful for the higher order bias calculations.

We begin by developing some notation. Let $\beta = (\gamma', \delta')'$ and define the $K + 2\left(1+M\right)\times 1$ moment vector and derivative matrix

$$m_i\left(\beta\right) = \begin{pmatrix} \frac{D_i}{G_i(\delta)}\psi_i\left(\gamma\right) \\ \left(\frac{D_i}{G_i(\delta)}-1\right)t_i \\ \frac{D_i - G_i(\delta)}{G_i(\delta)\left(1-G_i(\delta)\right)}G_{1i}\left(\delta\right)t_i \end{pmatrix},\ M_i\left(\beta\right) = \begin{pmatrix} \frac{D_i}{G_i(\delta)}\frac{\partial\psi_i(\beta)}{\partial\gamma'} & -\frac{D_i}{G_i(\delta)}\frac{G_{1i}(\delta)}{G_i(\delta)}\psi_i\left(\gamma\right)t_i' \\ 0 & -\frac{D_i}{G_i(\delta)}\frac{G_{1i}(\delta)}{G_i(\delta)}t_i t_i' \\ 0 & J_i\left(\delta\right) \end{pmatrix}, \tag{57}$$

with $J_i\left(\delta\right)$ as defined previously. Further define the weight matrix

$$V_i\left(\beta\right) = \begin{bmatrix} \frac{D_i}{G_i(\delta)^2}\psi_i\left(\gamma\right)\psi_i\left(\gamma\right)' & \frac{D_i}{G_i(\delta)}\omega_i\left(\delta\right)\psi_i\left(\gamma\right)t_i' & 0 \\ \frac{D_i}{G_i(\delta)}\omega_i\left(\delta\right)t_i\psi_i\left(\gamma\right)' & \nu_i\left(\delta\right)\omega_i\left(\delta\right)t_i t_i' & 0 \\ 0 & \frac{D_i}{G_i(\delta)}\frac{G_{1i}(\delta)}{G_i(\delta)}t_i t_i' & -J_i\left(\delta\right) \end{bmatrix}, \tag{58}$$

where $\omega_i\left(\delta\right) = \omega\left(X_i, \delta\right)$ and $\nu_i\left(\delta\right) = \nu\left(D_i, X_i, \delta\right)$ are known, scalar-valued, weight functions (the latter with the property that $\mathbb{E}\left[\nu_i\left(\delta_0\right)|\, X\right] = 1$).

Finally let

$$\overline{M}(\beta) = \frac{1}{N}\sum_{i=1}^{N} M_i(\beta), \quad \overline{V}(\beta) = \frac{1}{N}\sum_{i=1}^{N} V_i(\beta), \quad \overline{m}(\beta) = \frac{1}{N}\sum_{i=1}^{N} m_i(\beta). \tag{59}$$

Our asymptotic analysis of three-step AIPW estimators exploits their representation as particular iterated GMM estimators.

Lemma A.1 (Iterated GMM Representation of AIPW) *The AIPW estimate $\widehat{\gamma}$ which solves (14) in the main text is numerically identical to the iterated GMM estimate $\widehat{\beta} = \left(\widehat{\gamma}', \widehat{\delta}'\right)'$ which solves*

$$\overline{M}\left(\widehat{\beta}\right)\overline{V}\left(\widehat{\beta}\right)^{-1}\overline{m}\left(\widehat{\beta}\right) = 0. \tag{60}$$

The proof of Lemma A.1 is omitted but may be found online in the 'detailed calculations' supplement at `https://files.nyu.edu/bsg1/public/.` Invoking Lemma A.1 we proceed to characterize the large sample properties of (60). The population mean of the derivative of $m_i(\beta_0)$, as defined in (57) above, equals

$$M = \begin{bmatrix} \Gamma & -\mathbb{E}\left[\frac{G_1}{G}\psi t'\right] \\ 0 & -\mathbb{E}\left[\frac{G_1}{G}tt'\right] \\ 0 & -\mathcal{I}(\delta_0) \end{bmatrix}. \tag{61}$$

The probability limit of the weight matrix (58) is given by

$$V = \begin{bmatrix} \mathbb{E}\left[\frac{\psi\psi'}{G}\right] & E_\omega & 0 \\ E_\omega' & F_\omega & 0 \\ 0 & \mathbb{E}\left[\frac{G_1}{G}tt'\right] & \mathcal{I}(\delta_0) \end{bmatrix}, \tag{62}$$

with $E_\omega = \mathbb{E}[\omega\psi t']$ and $F_\omega = \mathbb{E}[\omega tt']$.

The covariance of the moment vector $m_i = m_i(\beta_0)$ is

$$\Omega = \begin{pmatrix} \mathbb{E}\left[\frac{\psi\psi'}{G}\right] & E_0 & \mathbb{E}\left[\frac{G_1}{G}\psi t'\right] \\ E_0' & F_0 & \mathbb{E}\left[\frac{G_1}{G}tt'\right] \\ \mathbb{E}\left[\frac{G_1}{G}t\psi'\right] & \mathbb{E}\left[\frac{G_1}{G}tt'\right] & \mathcal{I}(\delta_0) \end{pmatrix}, \tag{63}$$

with E_0 and F_0 as defined in the main text.

Using (61), (62) and (63) we get a limiting sampling variance for $\sqrt{N}\left(\widehat{\beta}-\beta_0\right)$ of

$$\begin{aligned}&\left(M'V^{-1}M\right)^{-1}M'V^{-1}\Omega V^{-1}M\left(M'V^{-1}M\right)^{-1\prime} \qquad (64)\\ &=\begin{pmatrix}\begin{pmatrix}\Gamma^{-1}\left(\mathbb{E}\left[\frac{\psi\psi'}{G}\right]-E_0F_0^{-1}E_0'-\Delta_\omega\mathcal{I}\left(\delta_0\right)^{-1}\Delta_\omega'\right)\Gamma^{-1\prime}\\ +\Gamma^{-1}\left(E_0F_0^{-1}-E_\omega F_\omega^{-1}\right)F_0\left(E_0F_0^{-1}-E_\omega F_\omega^{-1}\right)'\Gamma^{-1\prime}\end{pmatrix} & 0\\ 0 & \mathcal{I}\left(\delta_0\right)^{-1}\end{pmatrix},\end{aligned}$$

with

$$\Delta_\omega=\mathbb{E}\left[\frac{D}{G}\left\{\psi-E_\omega F_\omega^{-1}t\right\}S_\delta'\right].$$

If Assumption 2.1 also holds (64) simplifies to $diag\left\{\mathcal{I}\left(\gamma_0\right)^{-1},\mathcal{I}\left(\delta_0\right)^{-1}\right\}$.

Derivation of the higher order bias of AIPW (Theorem 3.1) Let $\theta=(\gamma',\delta',\lambda')$ be the $T=2K+3\left(1+M\right)$ vector of parameters of interest. Let $\widehat{\theta}$ be the solution to (35) with

$$b_i\left(\theta\right)=-\begin{bmatrix}M_i\left(\beta\right)'\lambda\\ m_i\left(\beta\right)+V_i\left(\beta\right)'\lambda\end{bmatrix}, \qquad (65)$$

and $M_i\left(\beta\right)$ and $V_i\left(\beta\right)$ as defined by (57) and (58) above. The AIPW estimate corresponds to the first $K\times 1$ components of $\widehat{\theta}$ since $\overline{V}\left(\widehat{\beta}\right)^{-1}\overline{m}\left(\widehat{\beta}\right)=-\widehat{\lambda}$ so that $\overline{M}\left(\widehat{\beta}\right)'\widehat{\lambda}=\overline{M}\left(\widehat{\beta}\right)'\overline{V}\left(\widehat{\beta}\right)^{-1}\overline{m}\left(\widehat{\beta}\right)=0$ which, by Lemma A.1, is the first order condition for the AIPW estimator.

Similar to the treatment of two-step and iterated GMM by Newey and Smith (2004, cf., Lemmas A.5 and A.6), we derive the higher order bias properties of the AIPW estimate of γ_0 by considering those of a simplified $O_p\left(N^{-3/2}\right)$ equivalent estimate.

We have, by Lemma A.4 of Newey and Smith (2004), $\widehat{\theta}-\theta_0=\widetilde{\phi}/\sqrt{N}+O_p\left(1/N\right)$ with $\widetilde{\phi}=\sum_{i=1}^N\phi_i/\sqrt{N}$ for ϕ_i as defined in (37) above. This means that for $q=1,\ldots,K+1+M$ we have $\widehat{\beta}-\beta_{q0}=e_q'\widetilde{\phi}/\sqrt{N}+O_p\left(1/N\right)$. Now consider the mean

value expansion

$$\begin{aligned}\overline{V}\left(\widehat{\beta}\right) &= \overline{V}\left(\beta_0\right) + \sum\nolimits_{q=1}^{K+1+M} \frac{\partial \overline{V}\left(\overline{\beta}\right)}{\partial \beta_q}\left(\widehat{\beta} - \beta_{q0}\right) \\ &= \overline{V}\left(\beta_0\right) + \sum\nolimits_{q=1}^{K+1+M} \mathbb{E}\left[\frac{\partial V_i\left(\beta_0\right)}{\partial \beta_q}\right]\left(\widehat{\beta} - \beta_{q0}\right) + O_p\left(1/N\right) \\ &= \overline{V}\left(\beta_0\right) + \sum\nolimits_{q=1}^{K+1+M} \mathbb{E}\left[\frac{\partial V_i\left(\beta_0\right)}{\partial \beta_q}\right] e_q' \widetilde{\phi}/\sqrt{N} + O_p\left(1/N\right),\end{aligned}$$

with $\overline{V}\left(\beta\right)$ and $V_i\left(\beta\right)$ as defined in (59) and (58) above; $\overline{\beta}$ is mean value between β_0 and $\widehat{\beta}$. Let $\xi_i = V_i - V + \sum_{q=1}^{K+1+M} \mathbb{E}\left[\frac{\partial V_i(\beta_0)}{\partial \beta_q}\right] e_q' \phi_i$ with V as given in (62) above. This gives $\overline{V}\left(\widehat{\beta}\right) = V + \frac{1}{N}\sum_{i=1}^{N} \xi_i + O_p\left(1/N\right)$. Now consider the solution to $\overline{b}^*\left(\widehat{\theta}\right) = \frac{1}{N}\sum_{i=1}^{N} b_i^*\left(\widehat{\theta}\right) = 0$ with

$$b_i^*\left(\theta\right) = -\left[\begin{array}{c} M_i\left(\beta\right)'\lambda \\ m_i\left(\beta\right) + \left[V + \xi_i\right]'\lambda \end{array}\right]. \tag{66}$$

Using the definitions above we have

$$\begin{aligned} 0 &= \overline{b}^*\left(\widehat{\theta}\right) \\ &= \overline{b}\left(\widehat{\theta}\right) + \left(\begin{array}{c} 0 \\ \left[\overline{V}\left(\widehat{\beta}\right) - V - \frac{1}{N}\sum_{i=1}^{N}\xi_i\right]'\widehat{\lambda} \end{array}\right) \\ &= \overline{b}\left(\widehat{\theta}\right) + O_p\left(N^{-3/2}\right), \end{aligned} \tag{67}$$

since $\overline{V}\left(\widehat{\beta}\right) - V - \frac{1}{N}\sum_{i=1}^{N}\xi_i = O_p\left(1/N\right)$ and $\widehat{\lambda} = \lambda_0 + O_p\left(1/\sqrt{N}\right)$. Appealing to the equivalence implicit in the last line of (67) we henceforth analyze the bias properties of the solution to $\overline{b}^*\left(\widehat{\theta}\right) = 0$. In what follows we redefine $b_i\left(\theta\right)$ to be equal to $b_i^*\left(\theta\right)$ as given in (66).

The terms defined in (37) are given by, recalling that $\lambda_0 = 0$,

$$B = -\left(\begin{array}{cc} 0 & M' \\ M & V \end{array}\right), \quad A_i = -\left(\begin{array}{cc} 0 & \left(M_i - M\right)' \\ \left(M_i - M\right) & \xi_i \end{array}\right),$$

with M and V given by (61) and (62) above.

The partitioned inverse formula gives

$$B^{-1} = -\begin{pmatrix} -\Upsilon & H \\ H' & L \end{pmatrix}, \tag{68}$$

with

$$\Upsilon = \left(M'V^{-1}M\right)^{-1}, \quad H = \Upsilon M'V^{-1}, \quad L = V^{-1} - V^{-1}MH. \tag{69}$$

Manipulating we get

$$\Upsilon = \begin{pmatrix} \Gamma^{-1}\Lambda\Gamma^{-1\prime} & 0 \\ \Pi_S'\Gamma^{-1\prime} & \mathcal{I}\left(\delta_0\right)^{-1} \end{pmatrix} \tag{70}$$

$$H = \begin{pmatrix} \Gamma^{-1} & -\Gamma^{-1}\Pi_0 & 0 \\ 0 & 0 & -\mathcal{I}\left(\delta_0\right)^{-1} \end{pmatrix} \tag{71}$$

$$L = \begin{pmatrix} 0 & 0 & -\Lambda^{-1}\Pi_S \\ 0 & F_\omega^{-1} & \Pi_0'\Lambda^{-1}\Pi_S \\ \Pi_S'\Lambda^{-1} & -\mathcal{I}\left(\delta_0\right)^{-1}\mathbb{E}\left[\frac{G_1}{G}tt'\right]\left(F_\omega - E_\omega'\mathbb{E}\left[\frac{\psi\psi'}{G}\right]^{-1}E_\omega\right)^{-1} & 0 \end{pmatrix}. \tag{72}$$

From (37) and the expressions above we also have $\phi_i = -\begin{pmatrix} H' & L' \end{pmatrix}' m_i$, where we evaluate at the population value of θ.

The first part of the AIPW bias formula is given by the first K rows of $\frac{-B^{-1}}{N}\mathbb{E}\left[A_i\phi_i\right]$. Manipulating, using the expressions given above, we have

$$-B^{-1}\mathbb{E}\left[A_i\phi_i\right] = -\mathbb{E}\left[\begin{pmatrix} -HM_iH + \Upsilon M_i'L - H\xi_iL \\ -LM_iH - H'M_i'L - L\xi_iL \end{pmatrix} m_i\right]. \tag{73}$$

We require expressions for the first K rows of the matrix $\mathbb{E}\left[\left(HM_iH - \Upsilon M_i'L + H\xi_iL\right)m_i\right]$.

Let $\{A\}_{1:K,:}$ denote rows 1 to K of a matrix. Very tedious calculations give

$$\{H\mathbb{E}\left[M_i H m_i\right]\}_{1:K,:} = \Gamma^{-1}\mathbb{E}\left[\frac{1}{G}\frac{\partial\psi\left(\beta\right)}{\partial\gamma'}\Gamma^{-1}\left(\psi - \Pi_0 t\right)\right] + \Gamma^{-1}\mathbb{E}\left[\frac{\partial\psi\left(\beta\right)}{\partial\gamma'}\Gamma^{-1}\Pi_0 t\right] \tag{74}$$

$$-\{\Upsilon\mathbb{E}\left[M_i' L m_i\right]\}_{1:K,:} = \mathcal{I}\left(\gamma_0\right)^{-1}\mathbb{E}\left[\frac{D}{G}\left(\frac{\partial\psi}{\partial\gamma'}\right)'\Lambda^{-1}\Pi_S S_\delta\right] \tag{75}$$

$$\begin{aligned}\{H\mathbb{E}\left[\xi_i L m_i\right]\}_{1:K,:} = {} & \Gamma^{-1}\Pi_0\mathbb{E}\left[\omega\left(\frac{D}{G}-\nu\right)\left(\frac{D}{G}-1\right)tt'F_0^{-1}t\right] \\ & - \Gamma^{-1}\mathbb{E}\left[\frac{D}{G}\psi\left(\frac{D}{G}\psi - \omega\Pi_0 t_0\right)'\Lambda^{-1}\Pi_S S_\delta\right] \\ & + \Gamma^{-1}\mathbb{E}\left[\omega\left(\frac{D}{G}-\nu\right)\Pi_0 tt'\Pi_0\Lambda^{-1}\Pi_S S_\delta\right].\end{aligned} \tag{76}$$

Collecting these terms yields a bias contribution of

$$\begin{aligned}C_L = {} & \frac{\Gamma^{-1}}{N}\mathbb{E}\left[\frac{1}{G}\frac{\partial\psi\left(\beta\right)}{\partial\gamma'}\Gamma^{-1}\left(\psi - \Pi_0 t\right)\right] + \frac{\Gamma^{-1}}{N}\mathbb{E}\left[\frac{\partial\psi\left(\beta\right)}{\partial\gamma'}\Gamma^{-1}\Pi_0 t\right] \\ & + \frac{\mathcal{I}\left(\gamma_0\right)^{-1}}{N}\mathbb{E}\left[\frac{D}{G}\left(\frac{\partial\psi}{\partial\gamma'}\right)'\Lambda^{-1}\Pi_S S_\delta\right] \\ & + \frac{\Gamma^{-1}}{N}\Pi_0\mathbb{E}\left[\omega\left(\frac{D}{G}-\nu\right)\left(\frac{D}{G}-1\right)tt'F_0^{-1}t\right] \\ & - \frac{\Gamma^{-1}}{N}\mathbb{E}\left[\frac{D}{G}\psi\left(\frac{D}{G}\psi - \omega\Pi_0 t_0\right)'\Lambda^{-1}\Pi_S S_\delta\right] \\ & + \frac{\Gamma^{-1}}{N}\mathbb{E}\left[\omega\left(\frac{D}{G}-\nu\right)\Pi_0 tt'\Pi_0\Lambda^{-1}\Pi_S S_\delta\right].\end{aligned} \tag{77}$$

To compute the second component of (36) for $\widehat{\gamma}_{AIPW}$ we require some additional notation and results. Recall that $B_q = \mathbb{E}\left[\partial^2 b_i\left(\theta\right)/\partial\theta_q\partial\theta'\right]$. For $q = 1, \ldots, K+1+M$ we have this term equal to

$$B_q = -\begin{pmatrix} 0 & \mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right] \\ \mathbb{E}\left[\frac{\partial M_i}{\partial\beta_q}\right] & 0 \end{pmatrix}, \tag{78}$$

while for $q = K+1+M+1, \ldots, 2K+3(1+M)$ is it given by

$$B_q = -\begin{pmatrix} \mathbb{E}\left[\frac{\partial^2 m_{q-K-1-M}(Z_i,\beta)}{\partial\beta\partial\beta'}\right] & 0 \\ 0 & 0 \end{pmatrix}. \tag{79}$$

We also have

$$\mathbb{E}\left[\phi_i\phi_i'\right] = \begin{pmatrix} H\Omega H' & H\Omega L' \\ L\Omega H' & L + L(\Omega - V)L' \end{pmatrix}, \tag{80}$$

with Ω as given by (63) above. Note that $LVL' = L$ so that $L\Omega L' = LVL' + L(\Omega - V)L' = L + L(\Omega - V)L'$.

Using these expressions we evaluate the second component of (36) as follows:

$$-\frac{B^{-1}}{2N}\mathbb{E}\left[\sum_{q=1}^{T}\phi_{q,i}B_q\phi_i\right] = -\frac{1}{2N}\sum_{q=1}^{K+1+M}\begin{pmatrix} -\Upsilon\mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right]L\Omega H' + H\mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right]H\Omega H' \\ +H'\mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right]L\Omega H' + L\mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right]H\Omega H' \end{pmatrix} e_q \tag{81}$$

$$-\frac{1}{2N}\sum_{q=1}^{K+2(1+M)}\begin{pmatrix} -\Upsilon\mathbb{E}\left[\frac{\partial^2 m_q(Z_i,\beta)}{\partial\beta\partial\beta'}\right]H\Omega L' \\ +H'\mathbb{E}\left[\frac{\partial^2 m_q(Z_i,\beta)}{\partial\beta\partial\beta'}\right]H\Omega L' \end{pmatrix} e_q.$$

The first K rows of (81) contribute to the bias expression for $\widehat{\gamma}_{AIPW}$. To determine the form of the first K rows of (81) we only require expressions for the first K rows of the two matrices in parentheses to the right of the equality in (81).

After tedious calculation we have, for $q = 1, \ldots, K$,

$$-\left\{\Upsilon\mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right]L\Omega H'\right\}_{1:K,:} = \begin{pmatrix} 0 & -\mathcal{I}(\gamma_0)^{-1}\mathbb{E}\left[\frac{\partial^2\psi}{\partial\gamma_q\partial\gamma}\right]'\Lambda^{-1}\Pi_S \end{pmatrix} \tag{82}$$

$$\left\{H\mathbb{E}\left[\frac{\partial M_i}{\partial\beta_q}\right]H\Omega H'\right\}_{1:K,:} = \begin{pmatrix} \Gamma^{-1}\mathbb{E}\left[\frac{\partial^2\psi}{\partial\gamma_q\partial\gamma}\right]\mathcal{I}(\gamma_0)^{-1} & -\Gamma^{-1}\mathbb{E}\left[\frac{D}{G}\frac{\partial\psi}{\partial\gamma_q}S_\delta'\right]\mathcal{I}(\delta_0)^{-1} \end{pmatrix} \tag{83}$$

Similarly, for $q = K+1, \ldots, K+1+M$, we have

$$-\left\{\Upsilon\mathbb{E}\left[\frac{\partial M_i'}{\partial\beta_q}\right]L\Omega H'\right\}_{1:K,:} = \begin{pmatrix} 0 & \mathcal{I}(\gamma_0)^{-1}\mathbb{E}\left[\frac{D}{G}\frac{\partial\psi}{\partial\gamma'}S_{\delta,q-K}'\right]'\Lambda^{-1}\Pi_S \end{pmatrix} \tag{84}$$

$$\left\{H\mathbb{E}\left[\frac{\partial M_i}{\partial\beta_q}\right]H\Omega H'\right\}_{1:K,:} = \begin{pmatrix} -\Gamma^{-1}\mathbb{E}\left[\frac{D}{G}\frac{\partial\psi}{\partial\gamma'}S_{\delta,q-K}'\right]\mathcal{I}(\gamma_0)^{-1} & 0 \end{pmatrix} \tag{85}$$

Using (81) and (82) to (85) we get, after some manipulation, a bias contribution of

$$C_{NL1} = -\frac{1}{2N}\sum\nolimits_{q=1}^{K}\Gamma^{-1}\mathbb{E}\left[\frac{\partial^2\psi}{\partial\gamma_q\partial\gamma}\right]\mathcal{I}\left(\gamma_0\right)^{-1}e_q \\ -\frac{\mathcal{I}\left(\gamma_0\right)^{-1}}{2N}\mathbb{E}\left[\frac{D}{G}\left(\frac{\partial\psi}{\partial\gamma'}\right)'\Lambda^{-1}\Pi_S S_\delta\right]. \tag{86}$$

Now consider the second part of (81). For $q = 1, \ldots, K$ we have

$$-\left\{\Upsilon\mathbb{E}\left[\frac{\partial^2 m_q\left(Z_i,\beta\right)}{\partial\beta\partial\beta'}\right]H\Omega L'\right\}_{1:K,:} = \left(\begin{array}{ccc} \mathcal{I}\left(\gamma_0\right)^{-1}\mathbb{E}\left[\frac{D}{G}\frac{\partial\psi_q}{\partial\gamma}S_\delta'\right]\Pi_S'\Lambda^{-1} & \text{(NR)} & \text{(NR)}\end{array}\right), \tag{87}$$

while for $q = K+1, \ldots, K+1+M$

$$-\left\{\Upsilon\mathbb{E}\left[\frac{\partial^2 m_q\left(Z_i,\beta\right)}{\partial\beta\partial\beta'}\right]H\Omega L'\right\}_{1:K,:} = \left(\begin{array}{ccc} 0 & 0 & 0\end{array}\right), \tag{88}$$

and finally for $q = K+1+M+1, \ldots, K+2\left(1+M\right)$

$$-\left\{\Upsilon\mathbb{E}\left[\frac{\partial^2 m_q\left(Z_i,\beta\right)}{\partial\beta\partial\beta'}\right]H\Omega L'\right\}_{1:K,:} = \left(\begin{array}{ccc} 0 & 0 & 0\end{array}\right). \tag{89}$$

Using (81) and (87) to (89) we get, after some manipulation, a bias contribution of

$$C_{NL2} = -\frac{\mathcal{I}\left(\gamma_0\right)^{-1}}{2N}\mathbb{E}\left[\frac{D}{G}\left(\frac{\partial\psi}{\partial\gamma'}\right)'\Lambda^{-1}\Pi_S S_\delta\right]. \tag{90}$$

Equation (23) is given by the sum of (77), (86), and (90).

Comparison of small sample properties of IPT with leading alternatives We compare the small sample performance of IPT with that of several alternative missing data estimators. Specifically we consider the parametric inverse probability weighting estimator described by Wooldridge (2007), henceforth IPW; parametric linear imputation as in Rubin (1977), henceforth PI; the non-parametric IPW estimator of Hirano, Imbens and Ridder (2003), henceforth HIR; the nonparametric imputation estimator of Imbens, Newey and Ridder (2005) (with their data-dependent choice of smoothing parameter), henceforth INR; the nonparametric conditional expectation projection estimator of Chen, Hong and Tarozzi (2008), henceforth CHT; and the augmented inverse probability weighting estimator of Robins, Rotnitzky and Zhao

(1994) as described by Tsiatis (2006), henceforth AIPW-RRZ.[24]

We assume that Y_1, the outcome of interest, is generated according to

$$Y_1 = \alpha_0 + \alpha_1 X + \alpha_2 \Phi\left(\frac{X-a}{b}\right) + U, \quad U|\, X \sim \mathcal{N}\left(0, \sigma^2\right),$$

where $\Phi(\cdot)$ is the CDF of a standard normal random variable and $X \sim \mathcal{U}(-1,1)$. Observations of Y_1 are missing at random with Y_1 observed if $D = 1$ where

$$D = \mathbf{1}\left(\beta_0 + \beta_1 X + \beta_2 \Phi\left(X/c\right) - V\right),$$

with $V|\, X, Y_1$ logistic.

We consider four different data generating processes (DGPs). Across all designs we set $a = 1/2$, $b = 1/5$, $c = 3/20$, and $\alpha_1 = -1/4$. We vary α_0, α_2 and β_0, β_2 to induce nonlinearity in, respectively, the CEF of Y_1 given X and the index of the propensity score model.

Each design is calibrated such that the propensity score ranges from 0.1 to 0.9 with a marginal probability of missingness equal to one half. The target estimand is $\gamma_0 = \mathbb{E}[Y]$, which is identically equal to zero for each design. For each experiment the sample size is set equal to $N = 1,000$ with $5,000$ Monte Carlo replications. We vary σ across designs in order to keep the variance bound constant such that, from the perspective of first order efficiency theory, each data generating process is equally 'difficult'.

The values of the varying parameters are listed in Table 4. In the first design both the outcome CEF and the propensity score are smooth in X. In the second, the outcome CEF is inhomogenous, while the propensity score remains smooth. In the third, the CEF is smooth, while the propensity score is now inhomogenous. In the fourth design both the CEF and propensity score are inhomogenous.

The IPW, AIPW-RRZ and IPT estimators are based upon a logistic model for the propensity score with X entering the index linearly. The HIR estimator is based on a series logistic model for the propensity score with X entering quadratically.[25]

[24] Matlab replication files for our Monte Carlo experiments as well as a technical Appendix, describing in detail our implementation of each estimator, is available on the first author's webpage.

[25] Hirano, Imbens and Ridder's (2003) theoretical results suggest using a series logit estimator with $N^{1/9}$ times some constant polynomial terms. For our designs $1000^{1/9} \approx 2.15$ which we round up to 3, yielding the quadratic specification.

Table 4: Parameter values for the four Monte Carlo experiments.

	Design 1: 'smooth-smooth'	Design 2: 'rough-smooth'	Design 3: 'smooth-rough'	Design 4: 'rough-rough'
α_0	0.00000	-0.12510	0.00000	-0.12510
α_2	0.00000	0.50000	0.00000	0.50000
β_0	0.00000	0.00000	2.00000	2.00000
β_1	2.19722	2.19722	4.19722	4.19722
β_2	0.00000	0.00000	-4.00000	-4.00000
σ	0.16183	0.17327	0.17410	0.18640
$\sqrt{\mathcal{I}(\gamma_0)^{-1}/N}$	0.01000	0.01000	0.01000	0.01000

NOTES: The square root of Hahn's (1998) variance bound for each design (divided by $N^{1/2} = \sqrt{1,000}$) is reported in the last row of the table.

The PI model is based on a linear model for $\mathbb{E}[Y|X]$. The INR estimator is based on a polynomial series estimate of $\mathbb{E}[Y|X]$ with the number of series terms chosen to minimize estimated mean square error of $\widehat{\gamma}$ (see Imbens, Newey and Ridder, 2005). The CHT estimator is also based on a polynomial series estimate of $\mathbb{E}[Y|X]$, but with the number of series terms fixed at 3 (i.e., a quadratic approximation).

Our designs are chosen to highlight the strengths and weakness of each estimator. In the first design we expect all estimators to perform well. In the second design we expect the IPW and HIR estimators to perform acceptably and the parametric imputation (PI) estimator to perform poorly. In principal the nonparametric imputation estimators, INR and CHT, are consistent and efficient in this design, although we expect them to perform poorly in practice due to the inhomogeneity of $\mathbb{E}[Y|X]$. In design 3 we expect a pattern opposite to that of design 2. The AIPW-RRZ and IPT estimators, due to their double-robustness attribute, should perform well in both designs 2 and 3 (as well as design 1). In design 4 we expect all estimators to perform poorly.

The HIR, INR and CHT estimators are consistent and attain Hahn's (1998) bound in all designs. IPW is never efficient but is consistent in designs 1 and 2. PI is consistent and efficient in designs 1 and 3. AIPW-RRZ and IPT are consistent in designs 1, 2 and 3. In design 1 they attain Hahn's (1998) bound. In design 2 their large sample variance lies above Hahn's (1998) bound but below that of parametric

IPW. In design 3 their large sample variance is actually smaller than Hahn's (1998) bound.[26]

For purposes of comparison we also include an infeasible oracle estimator: the mean of Y across all missing as well as non-missing observations. This estimator calibrates the cost of missingness in each of our designs.

Table 5 summarizes the results for the first design. Column 1 of the Table reports the asymptotic bias of each estimator and Column 2 the median bias across Monte Carlo replications (both scaled by the relevant estimator's asymptotic standard error). Column 3 reports each estimator's asymptotic standard error, Column 4 the median estimated standard error across Monte Carlo replications and Column 5 the standard deviation of the point estimates across Monte Carlo replications. Column 6 reports the coverage of a nominal 95 percent confidence interval. As expected each estimator performs well in design 1, with small sample properties well-approximated by asymptotic distribution theory. An approximate standard error for the Column 2 Monte Carlo scaled bias estimates is $\sqrt{(\pi/2)/5,000} \approx 0.0177$; differences in median bias are not significant across estimators for this design.[27]

Table 6 reports results from the second design. In this design the imhomogeneity of $\mathbb{E}[Y|X]$ creates problems for PI as well as both the INR and CHT imputation estimators, all of which exhibit significant median bias. Unsurprisingly, given the underlying smoothness of the propensity score, both the IPW and HIR estimators do well. In this design the HIR procedure involves an overfit of the propensity score. Consistent with the theoretical results of their paper, such overfitting results in improved precision (see Graham (2011) for a related discussion). The Monte Carlo sampling standard deviation of the HIR point estimates are about 5 percent lower than the corresponding IPW estimates (cf., Column 5).

Both AIPW-RRZ and IPT estimators also perform well in design 2. When both $\mathbb{E}[Y|X]$ and $p(X)$ are correctly modelled the two estimators are first-order equiva-

[26] Bang and Robins (2005, p. 966) comment that, when the propensity score is misspecified, but the outcome CEF is not, AIPW is often 'nearly' as efficient as parametric imputation. This assessment is based on their Monte Carlo experiments. In the designs whose results are reported in the lower half of Table 2 (p. 966) and the upper and lower portions of Table 4 (p. 970) of their paper, the AIPW estimator with a misspecified propensity score has a Monte Carlo sampling variance that lies below that of the AIPW estimator based on a correct model for the propensity score, but above that of the parametric imputation estimator. Our Design 3 results replicate this ordering.

[27] The standard error of the median bias estimates are approximately $[(\pi/2)/5000]^{1/2} \simeq 0.01772454$ since for N and B, the number of Monte Carlo replications, large enough $\widehat{\gamma}^2/[I(\gamma_0)^{-1}/N]^{1/2} \overset{D}{\simeq} N(Bias_A, 1)$ and $\sqrt{B}(Bias_{MC} - Bias_A) \overset{D}{\simeq} N(0, \pi/2)$.

Table 5: Monte Carlo results for Design 1: both p(x) and E[Y|X=x] smooth

	(1) Asymptotic Bias	(2) Median Bias	(3) Asymptotic Std. Err.	(4) Median Std. Err.	(5) Standard Deviation	(6) Coverage of 95% CI
Oracle	0.0000	-0.0398*	0.0069	0.0069	0.0068	0.951
			Parametric estimators			
IPW	0.0000	-0.0128	0.0107	0.0106	0.0108	0.947
PI	0.0000	-0.0129	0.0096	0.0096	0.0097	0.951
			Double robust parametric estimators			
AIPW	0.0000	-0.0195	0.0100	0.0099 [0.0099]	0.0101	0.948 [0.947]
IPT	0.0000	-0.0205	0.0100	0.0099	0.0101	0.946
			Nonparametric estimators			
HIR	0.0000	-0.0163	0.0100	0.0100	0.0103	0.945
INR	0.0000	-0.0109	0.0100	0.0099	0.0100	0.948
CHT	0.0000	-0.0271	0.0100	0.0099	0.0101	0.948

NOTES: Each row corresponds to a specific estimator as described in the text. Column 1 reports the scaled large sample bias of each estimator (i.e., its probability limit minus the true parameter divided by the square root of its large sample variance, $(AVar(\widehat{\gamma})/N)^{1/2}$). Column 2 reports the median Monte Carlo bias of each estimator scaled by its asymptotic standard error (a $*$ (+) next to these estimates denotes that they are significantly different from zero at the 5 (10) percent level). Column 1 calibrates the scale of inconsistency for each estimator, while a comparison of Columns 1 and 2 allows for an assessment of whether an estimator's actual sampling distribution is centered at is probability limit. Column 3 gives the large sample standard error of each estimator (i.e., $(AVar(\widehat{\gamma})/N)^{1/2}$), Column 4 the median estimated standard error and column 5 the standard deviation of the point estimates across the 5,000 Monte Carlo replications. Column 6 reports the actual coverage of a 95 percent Wald-based confidence interval (the asymptotic variance estimators are described in a technical appendix available online). Two standard errors are provided for the AIPW-RRZ estimator. The first assumes that the working models for both $E[Y|X]$and $p(X)$ are correctly specified. This is the covariance estimator typically used in applied work. The second, given in the square brackets, treats the AIPW-RRZ estimator as a sequential method-of-moments estimator with standard errors calculated accordingly. These standard errors are valid whenever AIPW-RRZ is consistent. The mean number of series terms selected by the INR estimator is 2.26, while the median is 2.

Table 6: Monte Carlo results for Design 2: p(x) smooth and E[Y|X=x] rough

	(1) Asymptotic Bias	(2) Median Bias	(3) Asymptotic Std. Err.	(4) Median Std. Err.	(5) Standard Deviation	(6) Coverage of 95% CI
Oracle	0.0000	-0.0231	0.0063	0.0063	0.0061	0.957
			Parametric estimators			
IPW	0.0000	-0.0136	0.0106	0.0105	0.0107	0.941
PI	-3.0095	-3.0599*	0.0093	0.0114	0.0111	0.278
			Double robust parametric estimators			
AIPW	0.0000	-0.0246	0.0114	0.0124 [0.0113]	0.0116	0.961 [0.943]
IPT	0.0000	-0.0067	0.0104	0.0103	0.0105	0.942
			Nonparametric estimators			
HIR	0.0000	-0.0089	0.0100	0.0099	0.0103	0.941
INR	0.0000	0.1226*	0.0100	0.0104	0.0106	0.945
CHT	0.0000	0.0461*	0.0100	0.0100	0.0102	0.951

NOTES: The mean number of series terms selected by the INR estimator is 3.07, while the median is 5. See notes to Table 5 for additional information.

lent. However under partial misspecification they are no longer first-order equivalent. In design 2, while both are consistent, their asymptotic variances differ, with IPT being more precisely determined (Columns 1 and 3 of Table 6). The small sample properties of the two estimators mirror their large sample ones. Both are approximately median unbiased, but the sampling variability of the IPT estimator is about 10 percent lower than that of AIPW (Columns 2 and 5 of Table 6).

Table 7 reports results from the third design. In this case the IPW and HIR perform extremely poorly, reflecting their fragility vis-a-vis misspecification of the propensity score. The remaining estimators all perform well. Importantly, the AIPW-RRZ and IPT estimators, while based on misspecified models for the propensity score, are nevertheless consistent. However, as in design 2, their asymptotic variances differ with IPT's again being smaller, albeit negligibly so. These large sample properties are reflected in small samples. In design 3 both AIPW-RRZ and IPT are approximately median unbiased with similar sampling variances (Columns 2 and 5 of Table 7).

Table 8 reports results from the our fourth design. In this design the AIPW-RRZ and IPT estimators are inconsistent as are IPW and PI. While HIR, INR and CHT are consistent and efficient in principle, in practice they exhibit significant median

Table 7: Monte Carlo results for Design 3: p(x) rough and E[Y|X=x] smooth

	(1) Asymptotic Bias	(2) Median Bias	(3) Asymptotic Std. Err.	(4) Median Std. Err.	(5) Standard Deviation	(6) Coverage of 95% CI
Oracle	0.0000	0.0178	0.0072	0.0072	0.0072	0.949
			Parametric estimators			
IPW	-0.3204	-0.3270*	0.0093	0.0092	0.0094	0.937
PI	0.0000	-0.0078	0.0093	0.0093	0.0095	0.944
			Double robust parametric estimators			
AIPW	0.0000	-0.0106	0.0094	0.0093 [0.0093]	0.0095	0.943 [0.944]
IPT	0.0000	-0.0115	0.0094	0.0093	0.0095	0.944
			Nonparametric estimators			
HIR	0.0000	-0.3028*	0.0100	0.0092	0.0094	0.937
INR	0.0000	-0.0002	0.0100	0.0093	0.0099	0.934
CHT	0.0000	-0.0016	0.0100	0.0093	0.0096	0.942

NOTES: The mean number of series terms selected by the INR estimator is 3.08, while the median is 2. See notes to Table 5 for additional information.

Table 8: Monte Carlo results for Design 4: p(x) rough and E[Y|X=x] rough

	(1) Asymptotic Bias	(2) Median Bias	(3) Asymptotic Std. Err.	(4) Median Std. Err.	(5) Standard Deviation	(6) Coverage of 95% CI
Oracle	0.0000	-0.0081	0.0066	0.0066	0.0067	0.947
			Parametric estimators			
IPW	0.0614	0.0271	0.0096	0.0095	0.0096	0.945
PI	-0.7273	-0.7635*	0.0089	0.0100	0.0099	0.9006
			Double robust parametric estimators			
AIPW	-0.0487	-0.0859*	0.0098	0.0098 [0.0097]	0.0095	0.946 [0.946]
IPT	0.0499	0.0224	0.0096	0.0096	0.0097	0.946
			Nonparametric estimators			
HIR	0.0000	0.0324^{+}	0.0100	0.0092	0.0092	0.947
INR	0.0000	0.1821*	0.0100	0.0091	0.0102	0.919
CHT	0.0000	0.6957*	0.0100	0.0092	0.0094	0.873

NOTES: The mean number of series terms selected by the INR estimator is 5, while the median is 5. See notes to Table 5 for additional information.

bias.

The performance of IPT across designs 2 and 3 highlights the practical value of using a 'doubly robust' estimator. If either the propensity score or outcome CEF is well approximated by the implicit parametric models used, then IPT will perform well. While the HIR, INR and CHT estimators have attractive large sample properties irrespective of the form of the propensity score and outcome CEF, in practice their small performance is highly sensitive to the smoothness of one of these two functions. The HIR estimator performs well when the propensity score is smooth, but very poorly when it is not. The INR and CHT imputation estimators, in contrast, perform best when the outcome response is smooth.

The Monte Carlo experiments also highlight differences between the AIPW-RRZ estimator of Robins, Rotnitzky and Zhao (1994) and our IPT procedure. Under misspecification of either $\mathbb{E}[Y|X]$ or $p(X)$, but not both simultaneously, they remain consistent but have different large sample variances. In the designs considered here IPT has a smaller asymptotic sampling variance than AIPW-RRZ in such cases.

Monte Carlo 'verification' of Higher Order bias expressions in Theorem 3.1 To assess the small sample relevance of the higher-order bias expressions given in Theorem 3.1 we conducted two additional Monte Carlo experiments. Consider the DGP

$$Y_1 = \left(\gamma_0 - \sigma_X^2\right) + \Pi h(X) + U, \qquad U|X \sim \mathcal{N}\left(0, \sigma_U^2\right), \tag{91}$$

where $\Pi = (0,1)$ and $h(X) = (X, X^2)'$ with $X \sim \mathcal{N}(0, \sigma_X^2)$. It is straightforward to show that the variance bound for $\gamma_0 = \mathbb{E}[Y] = 0$ when Y_1 is MCAR with probability Q_0 is equal to

$$\mathcal{I}(\gamma_0)^{-1} = \frac{\sigma_U^2}{Q_0} + 2\sigma_X^4.$$

We also have $\mathbb{V}(h(X)) = diag\{\sigma_X^2, 2\sigma_X^4\}$ so that evaluating the bias expressions given in Theorem 3.1 yields

$$\text{Bias}\left(\widehat{\gamma}_{IPT}\right) = 0, \qquad \text{Bias}\left(\widehat{\gamma}_{AIPW-NEWEY}\right) = -\frac{1}{N}\frac{1-Q_0}{Q_0}6\sigma_X^2.$$

When the propensity score and the conditional variance of Y_1 are both constant in X, the biases of the other AIPW estimators listed in Table 1 coincide with that of IPT. We therefore focus on a $\widehat{\gamma}_{IPT}$ and $\widehat{\gamma}_{AIPW-NEWEY}$ comparison in what follows.

Table 9: Higher order bias comparisons of IPT and AIPW-NEWEY with data MCAR: Design 1

	IPT			AIPW-NEWEY		
Q	Bias_A	Bias_{MC}	Coverage of 95% CI	Bias_A	Bias_{MC}	Coverage of 95% CI
0.8	0.0000	-0.0100	0.9520	-0.0300	-0.0390	0.7802
0.6	0.0000	-0.0080	0.9482	-0.0800	-0.0879	0.9010
0.4	0.0000	0.0045	0.9440	-0.1800	-0.1750	0.9676
0.2	0.0000	-0.0445	0.9460	-0.4800	-0.4667	0.9858

Notes: The table reports Monte Carlo estimates of the scaled median bias ($Bias_{MC}$) of the IPT and AIPW-NEWEY estimates of γ_0 (i.e., the median bias of $\widehat{\gamma}$ divided by its asymptotic standard error, $[I(\gamma_0)^{-1}/N]^{1/2}$). The DGP is as specified in equation (91) and the text immediately following. Reported statistics are based on $5,000$ Monte Carlo draws. The column labelled $Bias_A$ gives the asymptotic bias of $\widehat{\gamma}$ – again divided by $[I(\gamma_0)^{-1}/N]^{1/2}$ – as calculated analytically using the expressions given in Theorem 3.1 of the main text.

In moderately sized samples, $\widehat{\gamma}_{AIPW-NEWEY}$ may exhibit non-negligble bias when σ_X^2 is large relative to σ_U^2 and/or Q_0 is small. This bias, which arises despite the fact that $h(X)$ is low dimensional, comes from skewness in the distribution of X^2.

Table 9 reports the asymptotic bias, and Monte Carlo estimates of the median bias, of $\widehat{\gamma}_{IPT}$ and $\widehat{\gamma}_{AIPW-NEWEY}$ (scaled by their asymptotic standard error $\sqrt{\mathcal{I}(\gamma_0)^{-1}/N}$) for the above DGP with $\sigma_X^2 = 4$, $Q = 0.8, 0.6, 0.4$ and 0.2, and σ_U^2 varied across designs in order to keep the variance bound fixed at 40 (this implies an asymptotic standard error for $\widehat{\gamma}$ of $1/5$). We set $N = 1,000$ and perform $5,000$ Monte Carlo replications. Each of these designs is first-order equivalent, as are $\widehat{\gamma}_{IPT}$ and $\widehat{\gamma}_{AIPW-NEWEY}$.

Across each of these designs IPT is median unbiased (up to simulation error), consistent with Theorem 3.1. In contrast AIPW-NEWEY is very biased, particularly for the designs with a high probability of missingness. Across each design Theorem 3.1 does a good job of predicting the small sample median bias of AIPW-NEWEY.

A second example illustrates how the asymptotic bias of AIPW-NEWEY may grow with the dimension of set of conditioning variables. Assume that

$$Y_1 = \gamma_0 + \Pi X + U, \qquad U|\, X \sim \mathcal{N}\left(0, \sigma_U^2\right), \tag{92}$$

where X is an $M \times 1$ vector of independent standardized χ^2_1 random variables (i.e., $X = \sqrt{1/2}\,(W-1)$ with $W_m \sim \chi^2_1$ for $m = 1, \ldots, M$) and Y_1 is missing completely at random with a probability of observation equal to $Q_0 = 1/2$. The target parameter is $\gamma_0 = \mathbb{E}[Y_1] = 0$. Let Π be a column vector of ones such that $\Pi\Pi' = M$, it is straightforward to show that the variance bound for γ_0 is equal to

$$\mathcal{I}(\gamma_0)^{-1} = \frac{\sigma_U^2}{Q_0} + M.$$

Under this DGP evaluating the bias expressions given in Theorem 3.1 yields

$$\text{Bias}(\widehat{\gamma}_{IPT}) = 0, \qquad \text{Bias}(\widehat{\gamma}_{AIPW-NEWEY}) = -\frac{M}{N}\sqrt{8},$$

so that $\widehat{\gamma}_{IPT}$ is asymptotically unbiased while $\widehat{\gamma}_{AIPW-NEWEY}$ has a higher order bias which increases linearly with the dimension of X.

We consider six designs corresponding to $M = 1, 5, 10, 15, 20$ and 25. In all designs $N = 1,000$. Again, we vary σ_U^2 across designs in order to keep the variance bound fixed at 40.

Asymptotic bias and the median bias of the Monte Carlo estimates (again based on 5,000 replications), both scaled by the asymptotic standard error, for IPT and AIPW-NEWEY are reported in Table 10. As theory would predict the median bias of $\widehat{\gamma}_{AIPW-NEWEY}$ is increasing in the dimension of X, while $\widehat{\gamma}_{IPT}$ is median unbiased across all designs. In all cases our Monte Carlo median bias estimates are within simulation error of their theoretical counterparts. For the $M = 25$ design the median bias of $\widehat{\gamma}_{AIPW-NEWEY}$ is about one third of its asymptotic standard error.

Additional details of empirical application Our initial goal was to reconstruct the exact NLSY79 extract used by Johnson and Neal (1998). A preliminary inspection of the data, however, revealed that a preadolescence test score was available for only a handful of Hispanic respondents. We therefore decided to exclude Hispanics from our analysis. We targeted all non-Hispanic male respondents in the cross-sectional sample, as well as in the supplemental Black sample, born during or after 1962 ($N = 1,612$). These individuals were aged 16 to 18 when they took the NLSY79's administration of the Armed Services Vocational Aptitude Battery (ASVAB) from which the AFQT score is constructed.

Table 10: Higher order bias comparisons of IPT and AIPW-NEWEY with data MCAR: Design 2

	IPT			AIPW-NEWEY		
M	Bias_A	Bias_{MC}	Coverage of 95% CI	Bias_A	Bias_{MC}	Coverage of 95% CI
1	0.0000	0.0126	0.9518	-0.0141	-0.0015	0.9512
5	0.0000	-0.0126	0.9512	-0.0707	-0.0863	0.9480
10	0.0000	0.0026	0.9484	-0.1414	-0.1317	0.9430
15	0.0000	-0.0092	0.9528	-0.2121	-0.2118	0.9446
20	0.0000	-0.0208	0.9460	-0.2828	-0.3061	0.9292
25	0.0000	0.0228	0.9546	-0.3536	-0.3259	0.9276

Notes: The table reports Monte Carlo estimates of the scaled median bias ($Bias_{MC}$) of the IPT and AIPW-NEWEY estimates of γ_0 (i.e., the median bias of $\widehat{\gamma}$ divided by its asymptotic standard error, $[I(\gamma_0)^{-1}/N]^{1/2}$). The DGP is as specified in equation (92) and the text immediately following. Reported statistics are based on $5,000$ Monte Carlo draws. The column labelled $Bias_A$ gives the asymptotic bias of $\widehat{\gamma}$ – again divided by $[I(\gamma_0)^{-1}/N]^{1/2}$ – as calculated analytically using the expressions given in Theorem 3.1 of the main text.

Of the $1,612$ targeted individuals 159 were missing 1990 to 1993 average hourly wage data, 58 a valid AFQT score, and 24 both these items.[28] Our base sample therefore consists of the $1,371$ individuals with valid wage and AFQT data. As in Johnson and Neal (1998) we conditioned on this non-response.

AFQT scores are reported in terms of normed percentiles (i.e., relative to population of American youths aged 18 to 23).[29] We transformed these scores onto the real line using the inverse normal CDF. The residual associated with the least squares fit of the transformed scores onto a vector of birth year dummies is our AFQT measure.[30] The average hourly wage for Whites in the base sample is $12.32 in 1993

[28] Hourly wages are equal to their average over the 1990 to 1993 survey waves. In cases where an individual was not employed or interviewed in a given year, the average is over those years with wage information. Reported wage values less that $1 and greater than $75 per hour are discarded. Wages are measured in 1993 dollars. As in Johnson and Neal (1998) we exclude any AFQT score where the testing protocol was non-standard.

[29] We used the 1989 scoring of the AFQT.

[30] Neal and Johnson (1996) appear to have used the residual associated with the least squares fit of AFQT *percentiles* onto a set of birth year dummies (see their Figure 2, p. 886). Even after standardizing this variable to have mean zero and unit variance its distribution is non-normal. Our approach results in an AFQT score distribution that is indistinguishable from a normal one.

prices. Blacks earn on average $2.87 less per hour. The mean AFQT score for Whites is 0.160, while that for Blacks is 1.031 standard deviations lower.

For 11 percent of the respondents ($N = 144$) in our base sample an IQ test score from between the ages of 7 and 12 is available. This score is recorded as a (nationally normed) percentile. As with the AFQT scores we transformed these scores onto the real line using the inverse normal CDF. The residual associated with the least squares fit of these transformed scores onto a vector of birth year and age when tested (in years) dummies is our EARLYTEST measure. We view this score as measure of acquired cognitive skills not of innate ability (cf., Fryer and Levitt, forthcoming).[31]

Columns 1, 2 and 3 of Table 11 respectively the report average values of our main variables across the full sample, the complete case subsample ($D = 1$), and the subsample with missing early test scores ($D = 0$). Column 4 gives the Column 2 versus Column 3 difference. While the full and complete case samples are similar in terms of wages and age, they are rather different in terms of racial composition and AFQT.

Table 12 provides a synopsis of the results given in the main paper. Conditioning on age alone the Black-White gap in wages is 28 percent (Column 1). Conditioning on AFQT, a measure of cognitive skills acquired by age 18, this gap falls to 11 percent. Conditioning on our early test measure, a measure of cognitive skills acquired by age 12, the gap is 18 percent. We interpret this result as implying that approximately two thirds of the pre-market effect found by Neal and Johnson (1996) reflects skill differences already present by age 12. This interpretation is justified by two additional pieces of evidence. First, when we include both AFQT and EARLYTEST simultaneously in our model we find that the coefficient on the latter is insignificantly different from zero, while that on the former is insignificantly different from the estimate which conditions on AFQT alone (Column 4 of Table 12). This suggests that AFQT and EARLYTEST measure similar types of skills, with the later measure being relevant for labor market outcomes. A similar interpretation is suggested by Table 13 which shows that roughly two thirds of the AFQT gap across Blacks and Whites can be accounted for by skill differences present by age 12 (i.e., our EARLYTEST variable).

Table 14 reports the CMLE and IPT estimates of the propensity score parameter.

[31] If a respondent's record includes multiple test scores from the age 7 to 12 period, we use the average percentile score across all available tests. A *STATA* dictionary file of the NLSY79 extract used in our analysis as well as a do file which replicates the data manipulations described here is available online at `https://files.nyu.edu/bsg1/public/`.

Table 11: Comparison of the full sample with the complete case subsample

	(1) Full Sample	(2) $D = 1$	(3) $D = 0$	(4) Difference
LogWage		6.999	6.982	0.0167 (0.0433)
YearOfBirth		62.91	62.99	−0.0842 (0.0824)
Black		0.1113	0.1603	−0.0490 (0.0218)*
AFQT		0.1960	-0.0246	0.2205 (0.0953)*
N	1,371	144	1,227	

NOTES: Samples are as described in text. The 1979 baseline sampling weights are used when computing all summary statistics. A '**', '*' and '+' denotes that the Column 4 difference is significantly different from zero at the 1, 5 and 10 percent levels. Standard errors (in parentheses) allow for arbitrary patterns of heteroscedasticity and dependence across units residing in the same household at baseline.

References

[1] Fryer, Roland G. and Steven D. Levitt. (forthcoming). "Testing for racial differences in the mental ability of young children," *American Economic Review.*

[2] Imbens, Guido W., Whitney K. Newey and Geert Ridder (2005). "Mean-square-error calculations for average treatment effects," *IEPR Working Paper 05.34.*

[3] Rubin, Donald B. (1977). "Assignment to treatment group on the basis of a covariate," *Journal of Educational Statistics* 2 (1): 1 - 26.

Table 12: Pre-adolescent skills and adult wages: a synopsis

	(1) *OLS*	(2) *OLS*	(3) *IPT*	(4) *IPT*
YearOfBirth	−0.0458	−0.0466	−0.0537	−0.0443
	(0.0151)**	(0.0147)**	(0.0162)**	(0.0147)**
Black	−0.2776	−0.1079	−0.1837	−0.1132
	(0.0261)**	(0.0284)**	(0.0356)**	(0.0293)**
AFQT	–	0.1645	–	0.1866
		(0.0146)**		(0.0284)**
EarlyTest	–	–	0.1112	−0.0332
			(0.0296)**	(0.0374)

Notes: Samples are as described in the main text. The 1979 baseline sampling weights are used when computing all estimates. A '**', '*' and '+' denotes that a coefficient is significantly different from zero at the 1, 5 and 10 percent levels. Standard errors (in parentheses) allow for arbitrary patterns of heteroscedasticity and dependence across units residing in the same household at baseline.

Table 13: Understanding the AFQT gap at age 18

	(1) *OLS*	(2) *IPT*
YearOfBirth	0.0050	−0.0502
	(0.0330)	(0.0493)
Black	−1.0314	−0.3780
	(0.0519)**	(0.1205)**
EarlyTest	–	0.7739
		(0.0790)**

Notes: Samples are as described in the main text. The 1979 baseline sampling weights are used when computing all estimates. A '**', '*' and '+' denotes that a coefficient is significantly different from zero at the 1, 5 and 10 percent levels. Standard errors (in parentheses) allow for arbitrary patterns of heteroscedasticity and dependence across units residing in the same household at baseline.

Table 14: Propensity score estimates and balance assessment (Part 1 of 2)

	Panel A: P-Score		**Panel B: Balance**		
	(1) CMLE	(2) IPT	(1) FS	(2) CC-IPW	(3) CC-IPT
Black	−51.737	108.174	0.155	0.171	0.155
	(62.573)	(90.367)			
BornIn1963	−67.995	−292.099	0.325	0.314	0.325
	(96.750)	(83.740)**			
BornIn1964	−118.245	−370.900	0.330	0.326	0.330
	(67.851)	(96.265)**			
Black×BornIn1963	0.717	1.029	0.052	0.057	0.052
	(0.602)	(0.596)+			
Black×BornIn1964	−0.566	−0.721	0.052	0.065	0.052
	(0.689)	(0.590)			
AFQT	1.012	−1.650	0.000	-0.029	0.000
	(2.615)	(0.945)+			
LogWage	−7.109	−3.414	6.984	6.845	6.984
	(6.882)	(5.392)			
LogWage×AFQT	−0.095	0.297	0.175	0.004	0.175
	(0.376)	(0.145)*			
$AFQT^2$	−0.040	−0.094	0.975	0.926	0.975
	(0.128)	(0.074)			
$LogWage^2$	0.484	0.202	48.965	47.829	48.965
	(0.490)	(0.381)			
Black×AFQT	−7.836	10.380	-0.135	-0.149	-0.135
	(6.337)	(12.650)			
Black×LogWage	13.680	−31.672	1.045	1.139	1.045
	(17.777)	(25.935)			
Black×LogWage×AFQT	1.046	−1.570	-0.892	-0.979	-0.892
	(0.880)	(1.771)			
Black×$AFQT^2$	−0.052	0.117	0.215	0.211	0.215
	(0.284)	(0.387)			
Black×$LogWage^2$	−0.908	2.305	7.082	7.641	7.082
	(1.260)	(1.863)			

NOTES: See notes to Part 2 of the table.

Table 15: Propensity score estimates and balance assessment (Part 2 of 2)

	Panel A: P-Score		**Panel B: Balance**		
	(1) CMLE	(2) IPT	(1) FS	(2) CC-IPW	(3) CC-IPT
BornIn1963×AFQT	−5.723	−52.878	0.000	-0.006	0.000
	(6.238)	(17.479)**			
BornIn1963×LogWage	18.208	78.757	2.276	2.206	2.276
	(28.007)	(22.800)**			
BornIn1963×LogWage×AFQT	0.815	7.467	0.061	0.045	0.061
	(0.898)	(2.483)**			
BornIn1963×AFQT^2	−0.111	−1.045	0.332	0.329	0.332
	(0.210)	(0.358)**			
BornIn1963×LogWage^2	−1.221	−5.295	15.991	15.554	15.991
	(2.023)	(1.545)**			
BornIn1964×AFQT	−1.716	−13.540	0.000	-0.027	0.000
	(5.026)	(6.657)*			
BornIn1964×LogWage	35.051	109.870	2.286	2.223	2.286
	(19.618)+	(28.233)**			
BornIn1964×LogWage×AFQT	0.209	1.904	0.049	-0.135	0.049
	(0.734)	(0.953)*			
BornIn1964×AFQT^2	−0.052	−0.185	0.289	0.271	0.289
	(0.218)	(0.130)			
BornIn1964×LogWage^2	−2.592	−8.115	15.893	15.220	15.893
	(1.417)	(2.064)**			
Constant/Sum of Weights	24.079	12.077	1.000	0.983	1.000
	(24.145)	(19.099)			
N	1,371	1,371	1,371	144	144

NOTES: Panel A of the table reports conditional maximum likelihood and IPT propensity score parameter estimates. The 1979 baseline sampling weights are used when computing both estimates. A '**', '*' and '+' denotes that a point estimate is significantly different from zero at the 1, 5 and 10 percent levels. Standard errors (in parentheses) allow for arbitrary patterns of dependence across units residing in the same household at baseline. Column 1 of Panel B gives the full sample unweighted mean (i.e., only sampling weights used) of each of the propensity score regressors. Columns 2 and 3 give the corresponding complete case inverse probability weighted means using, respectively, the CML and IPT propensity score estimates.