The Prevalence of HIV in Botswana*

James Levinsohn
Yale University and NBER

Justin McCrary
University of California, Berkeley and NBER

January 6, 2010

Abstract

This paper implements five methods to correct naïve HIV seroprevalence rates for the selection bias induced by voluntary testing. We apply the methodologies to estimate the prevalence of HIV in Botswana.

I. Introduction

Household surveys increasingly include a request for a bio-specimen from some household members. Respondents sometimes refuse to provide the requested bio-specimen, raising issues of representativeness. This issue is especially relevant in the case of HIV. In several nationally representative surveys conducted to estimate HIV prevalence rates, opt-out rates of 25 to 50 percent are not uncommon. If opt-out is not random, then prevalence rates will be subject to selection bias.

In this paper, we use the 2004 Botswana AIDS Impact Survey (BAIS) to estimate prevalence rates corrected for selection. The richness of this particular data set makes it especially suitable for an analysis predicated on an assumption of selection on observables. Invoking a selection on observables assumption, we estimate a population prevalence rate based on the selected subsample of those who agreed to provide a bio-specimen.

We apply five methodologies to estimate the HIV prevalence rate as there is currently not a consensus view of the best way to address selection on observables. We implement a propensity score reweighting estimator (see for example (DiNardo, Fortin and Lemieux 1996) and (Hirano, Imbens and Ridder 2003)), a matching estimator (see for example (Heckman, Ichimura, Smith and Todd 1998)), a control function estimator (see for example (Heckman and Navarro-Lozano 2004)), the double robust approach (see (Robins and Rotnitzky 1995)), as well as an application of Blinder Oaxaca ((Blinder 1973) and (Oaxaca 1973)).

This results in both a robust estimate of the corrected-for-selection HIV prevalence rate (our principle focus) as well as a comparison of alternative econometric methodologies (a secondary issue.)

*We are grateful to Taryn Dinkelman and Zoe McLaren for research assistance. Levinsohn acknowledges support from NICHD.

Although the methods proposed address an issue that arises regularly in the context of HIV testing, they are potentially applicable to any study in which background information on respondents is available but some respondents decline to provide data on a variable that is key to the question at hand.

We find that all our methodologies result in HIV prevalence rates that are only slightly lower than the prevalence rate that is estimated when selection bias is ignored. We then apply non-parametric graphical methods to investigate why this is so.

In the next section, we introduce the data and use a bounds test to illustrate the bias that may result from sample selection when computing HIV prevalence rates. In section III, we explain the five approaches we implement. Section IV presents results, while Section V concludes.

II. Country Context, Data, and Bounded Prevalence Rates

Botswana is a southern African country with a population of about 1.63 million. According to UNAIDS, the overall HIV prevalence rate is 19%. For those in the 15-49 years old cohort, the prevalence rate is estimated to be 38.5% and by 2010 it is projected that 20% of all children will be orphans.[1] Few countries are as impacted by HIV/AIDS as Botswana. Only tiny Swaziland has a higher estimated overall prevalence rate.

In 2004, the second wave of the Botswana AIDS Impact Survey (BAIS-2) was collected. This survey was the first nationally representative survey in Botswana to include a test for HIV. After data cleaning, the sample included 7669 households, comprised of 27,521 individuals. The survey included detailed questions on sexual practices, knowledge of HIV/AIDS, as well as basic demographic information. Respondents 18 months and older were asked to provide an HIV test specimen.[2] 13,997 respondents provided a valid specimen, while 13,524 did not.[3]

With almost half of our sample opting out of the HIV test, there is substantial uncertainty about what the true HIV prevalence rate might be. We begin by computing the bounds on the true prevalence rate as well as the standard errors around those bounds using the approach of Imbens and Manski. Table 1 reports the upper bound (if all those who opted out were HIV+) and the lower bound (if all those who opted out were HIV-) as well as bootstrapped standard errors for both bounds. It is also possible to bound the prevalence rate that lies within these bounds. This is the Imbens-Manski confidence interval also reported in Table 1. That interval is [0.0811 0.5957] using a 95 percent confidence level.[4] With prevalence rates bounded by 8.1 percent to 59.6 percent, testing opt-out clearly may be quite important. Finally, we report the prevalence rate computed for those individuals who provided a bio-sample. We refer to this as the "naïve prevalence rate." That rate is 17.12 percent. We turn next to our multiple methodologies for estimating just where in the Imbens-Manski confidence interval the true prevalence rate lies.

[1] See: http://www.unbotswana.org.bw/unaids.html (accessed Feb. 3, 2009).

[2] HIV testing of respondents under 18 years old required consent of a guardian.

[3] Individuals who provided an "invalid" biosample and for whom HIV status is not measured are coded as not having been tested.

[4] See equations (6) and (7) in (Imbens and Manski 2004) for the derivation.

III. Methodology

Four of our methodologies (propensity score reweighting, pair matching, control function, and double robust) are based on the propensity score. After establishing notation, each is briefly discussed. Our fifth methodology is an application of the more established Blinder Oaxaca approach. We include this since it is simple to implement, has a long history, and hence provides a benchmark to which the propensity score-based methods can be compared.

A. Set-up

Denote an individual's characteristics by X. These include demographic attributes as well as information on sexual practices and knowledge about HIV/AIDS. D is an indicator variable set to one if a respondent eligible for the HIV test agreed to be tested. Y is an indicator variable set to one if a person tested positive for HIV, so Y is only observed for respondents with $D = 1$. There are n individuals in the sample. Denote the propensity score, a scalar, by $P(D = 1|X) \equiv p(X)$. Hence the propensity score is the predicted probability of taking the HIV test conditional on an individual's characteristics.

The naïve prevalence rate will be correct if seropositivity is independent of whether one opted out of the HIV test. This independence is denoted by:

$$Y \perp\!\!\!\perp D. \tag{1}$$

When (1) does not hold, naïve prevalence rates may be inaccurate. For the case at hand, even the sign of the bias when (1) does not hold is unclear. For example, if HIV-negative individuals are keenly aware of how HIV is transmitted and are hesitant to provide a stranger with a blood sample, the naïve prevalence rate is likely to be over-estimated. If individuals who believe in traditional religions that forbid giving a piece of one's self to a stranger (and whose lack of knowledge about HIV may make them more likely to be HIV+) opt out, then the naïve prevalence rate is likely to be under-estimated.

Each of the five employed methodologies assumes that selection is on observables, and this implies a conditional independence assumption (CIA) given by:

$$Y \perp\!\!\!\perp D \mid X. \tag{2}$$

That is, individual characteristics, X, "explain" the relationship between an individual's willingness to be tested, D, and that individual's HIV status, Y, so that when one conditions on X, the decision to be tested is independent of HIV status. The appropriateness of the CIA in practice will depend on the available X's. We return to the empirical plausibility of the CIA in the results section. We also adopt the common support assumption, that $p(X) > c$ for some $c > 0$.[5]

[5]This is sufficient to insure the root consistency of the estimator of reweighted prevalence rates. See Bussos, DiNardo and McCrary for details.

B. Propensity Score Reweighting

The first approach to estimating the HIV prevalence rate is propensity score reweighting. Given the CIA and the support assumption, it is possible to estimate the population prevalence rates using only a weighted average of data from those who agreed to take the HIV test. The proof of this result is straightforward and clarifies the content of the support and conditional independence assumptions. We have:

$$\begin{aligned}
E[Y] &= E[E[Y \mid X]] \\
&= E\left[E[Y \mid X] E[D \mid X] \frac{1}{p(X)}\right] \\
&= E\left[E[YD \mid X] \frac{1}{p(X)}\right] \\
&= E\left[E\left[YD\frac{1}{p(X)} \mid X\right]\right] \\
&= E\left[YD\frac{1}{p(X)}\right] \\
&= E\left[YD\frac{1}{p(X)} \mid D=1\right] p + E\left[YD\frac{1}{p(X)} \mid D=0\right](1-p) \\
&= E\left[Y\frac{p}{p(X)} \mid D=1\right]. \qquad (3)
\end{aligned}$$

where $p \equiv P(D=1)$. The first equality is the Law of Iterated Expectations. The second equality follows from the definition of the propensity score, $p(X)$ and makes use of the support assumption. Note that the support assumption rules out the possibility that there are some individuals for whom their background guarantees that they are unwilling to be tested. The third equality follows from the CIA. The fourth equality follows because $p(X)$ is a function of X, and the fifth is again the Law of Iterated Expectations. The sixth equality states that the overall average is the average over both $D=1$ types (those who took the HIV test) and $D=0$ types (those who did not take the HIV test). The last equality reflects the fact that only those with $D \neq 0$ matter for the calculation.

The final result shows that the overall prevalence rate $E[Y]$ can be computed by re-weighting the observed prevalence rate for those who agree to take the test. The appropriate weight is simply the unconditional probability relative to the propensity score, and the average is taken over all those who took the HIV test.

C. Matching

An alternative methodology, also based on the propensity score, is matching. There are several matching estimators including pair matching, nearest neighbor matching, kernel matching, local linear matching, and ridge matching.

With the matching estimators, the prevalence rate is given by:

$$E[Y] = \frac{1}{n}\sum_{i=1}^{n}\left[D_i Y_i + (1 - D_i)\hat{Y}_i\right] \tag{4}$$

and the different matching estimators define $\hat{Y}_i$ differently. We employ pair matching.

The intuition behind matching is to use data from individuals who did take the HIV test to infer the status of those who opted out. Pair matching simply infers the status of the opt-out from the status of the individual with the most similar propensity score.

D. Control Function

The third methodology, again based on the propensity score, is a control function approach. Because we maintain the selection on observables assumption, we do not impose an exclusion restriction. The control function approach is best motivated by a figure. Suppose one plotted the average prevalence rate on the vertical axis and each value of the propensity score on the horizontal axis. Hence, if such a plot was a flat line, there would be no relationship between the likelihood of taking the test and one's HIV status and hence no selection bias. The core idea of the control function (without an exclusion restriction) approach is to flexibly estimate the expected HIV status conditional on the estimated propensity score.[6] This estimation procedure can of course only be done for those for whom HIV status is observed ($D = 1$). One then predicts the HIV status for those who opted out based on the estimated flexible relationship. The estimated prevalence rate is then simply the average across all individuals of the observed and predicted HIV status. This is shown more formally below.

$$\begin{aligned} E[Y] &= E[E[Y \mid p(X)]] \\ &= E[E[Y \mid p(X), D = 1]] \\ &= \int E[Y \mid p(X) = q, D = 1] f_{p(X)}(q) dq \end{aligned} \tag{5}$$

The first line follows from the law of iterated expectations. The second line makes use of the CIA since having conditioned on $p(X)$, nothing is changed by additionally conditioning on $D = 1$. The third line computes the expectation by taking the integral of the density of propensity scores ($f_{p(X)}$, where q is a specific value of the propensity score and the integration is across q.

E. Double Robust

We also implement the double robust approach of Robins and Rotnitzky (1995). Imbens (2004) provides an intuitive discussion of this approach in the context of average treatment effect estima-

[6] In practice, we employ a fifth order polynomial of the propensity score.

tion. Essentially, this is a way of combining reweighting with regression adjustment. It is believed to be a more robust method, in the sense that if the propensity score model is misspecified, there is still a possibility that the regression model is correctly specified, and vice versa. It turns out that only one of these models need to be correct for the double robust estimator to deliver an accurate estimate.

In our context, the double robust approach involves estimating the propensity score weights for each observation, $\frac{p}{p(X)}$, and then regressing HIV status, Y, on those variables that appeared in the propensity score specification using the propensity score weights. This regression is then used to project HIV status for those who did not take the HIV test. The estimated prevalence rate is taken over the actual HIV status for those who were tested and the predicted HIV status for those who were not, as in (4).

F. Blinder-Oaxaca

Well before the propensity score literature developed, Blinder (1973) and Oaxaca (1973) developed a straightforward methodology to predict values that variables might take in the case of a well-specified counterfactual. In our context, we regress HIV status, Y, on those variables, X, that entered the propensity score and then use the estimated parameters to predict HIV status for those who did not take the test, $\hat{Y}_{D=0}$. In particular,

$$Y = X_{D=1}\hat{\beta} \tag{6}$$
$$\hat{Y}_{D=0} = X_{D=0}\hat{\beta} \tag{7}$$

The first regression is a simple OLS linear probability model run, by construction, only using those individuals who took the HIV test ($D = 1$.) The overall prevalence rate is then given by (4).

One potential advantage of the Blinder-Oaxaca approach is that it is not as dependent on the common support assumption required by the reweighting estimators. In particular, when the distributions of the propensity scores are not sufficiently overlapping between the treatment and control groups, some of the reweighting estimators are problematic. This is because reweighting results in division by zero (or near zero.) In this situation, the Blinder-Oaxaca estimator will be more robust, since it forecasts from the area for which there is common support all the way out to the boundary.

G. Standard Errors of Estimated Prevalence Rates

The naïve prevalence rate as well as the prevalence rates resulting from each of our five methodologies each have a standard error associated with it. All five of our methodologies, though, involve estimated regressions, and this introduces another source of noise that contributes to the standard error of these prevalence rates. (The propensity score is estimated for all but Blinder-Oaxaca and an HIV prediction equation is estimated for Blinder-Oaxaca.) Although for some estimators

(e.g. propensity score reweighting– see Hirano et al. (2003)) there are results showing that under particular circumstances it may be conservative to ignore this estimation error, we elect to report standard errors that account for it.

For propensity score reweighting, the control function approach, the double robust approach, and Blinder-Oaxaca, we report bootstrapped standard errors.[7] In the case of pair-matching, we use subsampling, as the bootstrap is known to be inconsistent (Abadie and Imbens 2006). Whereas the bootstrap involves repeatedly estimating the same estimator using n-choose-n resamples from the original data, drawn randomly with replacement, subsampling involves repeatedly estimating the same estimator using n-choose-b resamples from the original data, drawn randomly *without* replacement.

IV. Results

All of our approaches except Blinder-Oaxaca require, as a first step, estimating the propensity score. We begin, then, with a discussion of the estimated propensity score. Each propensity score-based approach also assumes that the CIA holds while reweighting, pair-matching, and double robust also require the support assumption.[8] We next investigate the appropriateness of these assumptions. We then present our estimated prevalence rates. We conclude this section with an investigation into the particular patterns of selection that are driving the results.[9]

A. Estimating the propensity score

The estimated propensity score is the predicted value of a logit regression of whether one took the HIV test on vectors of attributes, X. We adopt a relatively parsimonious specification and then check to see if the reweighted data are balanced. Results of the propensity score logit are given in Table 2, and predicted values of this regression comprise the propensity scores.

The first column of Table 2 gives the variables included in the propensity score.[10] The second column gives the excluded category for the case of categorical variables. The third column reports the coefficient from the logit and the standard error of the coefficient. These coefficients, though, are not readily interpretable. The fourth column reports the average change in the probability of taking the HIV test resulting from a specific counterfactual as well as the standard error of this difference. This is not the same as the marginal effects that are frequently computed by standard

[7] Results are reported for standard errors computed with 500 replications.

[8] It is not clear that the control function approach requires this assumption.

[9] Although our data are confidential, all programs used for estimation will be available on the web sites of each co-author.

[10] Not all respondents answered every question. For variables with missing values, a value was imputed when an observation was missing. An indicator variable was created for every variable with any missing observations. This indicator was set to one if a particular observation was imputed. This procedure allows one to still use the full sample for the logit (and hence propensity scores) while sweeping out the impact of the (arbitrary) imputation for the missing values. Such indicators were included in the logit for age, marital status, whether one had protected oneself from HIV, whether one had heard of HIV, whether one had previously taken an HIV test, and years of education. We do not report the coefficients on these variables for purposes of brevity.

statistical packages, since in the instance of variables that enter other than linearly and in the instance of categorical variables that take on more than two values, the standard marginal effects are difficult (or impossible) to interpret.

For the case of continuous variables (e.g. age and years of education), we increase the value of that variable by one year for everyone in the sample and then compute the average impact on the likelihood of testing taking into account the fact that the variable enters the propensity score quadratically.[11] Hence, there is only one value in column 4 for the continuous variables age and education. (Although standard statistical packages will compute a marginal effect for, say, age squared, doing so holding age constant makes little sense.)

For the case of categorical variables, we compute the average change in the probability of testing when the categorical variable is changed from its actual value to the value of the excluded characteristic. For example, there are six values that the variable for marital status can take. The excluded category is married. For the case of, say, living together but not married, we compute the change in the likelihood of testing by comparing the average probability of testing for the original logit with the average probability of testing when all individuals who were living together are instead coded as being married (the omitted category.) The computation is similar to that given in footnote 12. The key difference is in the set of observations over which the average change is computed. For a continuous variable such as age, the counterfactual has age increased for everyone so the average change is computed across all observations. For a categorical variable, only those observations coded as "1" (e.g. living together) are changed to zero (e.g. "married") so the average is only taken across those observations whose value changed in the counterfactual.

The first column of Table 2 lists the variables that enter the propensity score for the base case. These are the respondent's age and age squared as well as the guardian's age and age squared. (Minors do not give their permission to take the HIV test. Rather, their parent or guardian does, hence we include the attributes of the guardian for those respondents who are minors.) Next is an indicator variable for whether the respondent is female followed by indicator variables for five values of marital status. Next are indicator variables for whether the respondent lives in an urban area, is a citizen, has ever had sex, used protection against HIV, has heard of HIV and has previously had an HIV test. Finally, years of education and its square are included.

The last column reports the results in a form that is readily interpretable. For example, increasing the respondent's age by one year results on average in a 0.7 percentage point decrease in the likelihood of taking the HIV test. If one lived in an urban area ("Urban"), the coefficient in the logit is negative. Hence living in an urban area makes one less likely to take the HIV test. If we then take all urban residents and code them as rural (the counterfactual), these recoded respondents are 6.2 percentage points *more* likely to provide a bio-sample. This table also indicates that one is less likely to take the HIV test if one has already been tested and that more education makes one less likely to take the HIV test. While an entire literature has developed in the public health field around who agrees to be tested for HIV, this table is really only the first step to our analysis of

[11] This is done in 4 steps. Step one is to compute the vector of predicted probabilities using the specified logit. Step two is to increase the variable, say age, by one year for everyone in the sample, updating age squared accordingly and, using the coefficients from the original logit, predict the new vector of probabilities. Step 3 is to take the difference between the vector of new probabilities and the original probabilities. Step 4 is to compute the average of these differences across individuals and the standard error of that average. We use the paired (or nonparametric) bootstrap to compute the standard errors.

selection.

B. Evaluating the CIA

The CIA implies selection entirely on observables. We examine the appropriateness of this assumption by conducting balancing tests on variables in the logit (internal variables).[12] The intuition behind balancing tests is appealing. Selection means that the values for variables included in the logit are systematically different for those who provided a bio-specimen compared to those who opted out. If the CIA obtains, then reweighting variables by $\frac{p}{p(X)}$ as in (3), should erase these systematic differences.

To implement the balancing test, for every independent variable in the logit, we regress that variable on D using the original weights and then again weighting by $\frac{p}{p(X)}$.[13] Comparing results with the reweighted data to the results with the originally weighted data, we investigate whether the reweighting is addressing the selection. To the extent that the reweighting is effective, CIA obtains, and a comparison of regression results with reweighted data should show a coefficient on D that is insignificantly different from zero.

Results are reported in Table 3. The first row of each entry gives the difference in the means between those who provided a bio-sample and those who did not. We report the difference in the mean, the standard error of that difference, and the t-statistic for the null hypothesis that the mean is the same for those who provided a specimen as for those who did not. For example, females are 3.17 percent more likely to provide a bio-sample in the unweighted data and this difference is quite statistically significant with a t-statistic of 5.28. When the data are reweighted, the difference becomes statistically insignificant. This is true for every variable in the propensity score specification.[14]

We conclude that the logit specified in Table 3 is generally consistent with the CIA.[15]

C. The Distribution of Propensity Scores

Kernel density estimates of the distributions of propensity scores for those who provided a bio-specimen and for those who opted out are presented in Figure 1. This figure serves two functions. First, it informs one that a selection problem is present. This is evidenced by the fact that the two distributions in Figure 1 are not approximately coincident. (While Figure 1 shows that a selection

[12] In previous versions of this paper, we also conducted external balancing tests– that is, balancing tests using variables that themselves did not enter the propensity score. These results were also generally consistent with the CIA. They are not reported here for expositional ease.

[13] By using a regression framework, it is simple to do hypothesis tests for whether those tested differed in a given attribute from those not tested. There are of course other ways to equivalently test this.

[14] Age squared remains significantly different, but age does not. Taken together, age, flexibly specified, does not differ across those who provided a sample and those who did not in the reweighted data.

[15] There are of course more saturated logit models that could be used. In results not reported here, we experimented with adding up to 87 more variables to the logit and the estimated prevalence rates using propensity score reweighting did not change by more than about .001.

problem exists it does not speak directly to the empirical magnitude of the problem.) Second, Figure 1 provides empirical support for the support assumption.

D. Prevalence Rates

Estimated HIV prevalence rates are given in Table 4. The standard errors for each prevalence rate are based on bootstrapping for the case of all but pair matching and on sub-sampling for the case of pair matching.[16]

There are two striking findings in Table 4. First, all of the selection-corrected prevalence rates are very close to one another. Across four different methodologies based on the propensity score and one (Blinder-Oaxaca) that takes a different approach altogether, the results are mutually confirming. Furthermore, all of the selection-corrected prevalence rates are precisely measured. While it is straightforward to do hypothesis tests on whether the rates are statistically significant from one another, it is clear that most lie within two standard deviations of one another.

The second striking fact in Table 4 is that all the selection-corrected prevalence rates are very close to and only modestly lower than the naïve rate. Hence, even though the Imbens-Manski confidence interval given in Table 1 was quite large, correcting for selection into testing (five ways) makes almost no difference. The remaining question is "Why?" We investigate this question using non-parametric methods to investigate the empirical relationship between a respondent's propensity score and their actual HIV status.

If the likelihood of taking the HIV test (the propensity score) is independent of one's HIV status, then one can obtain accurate estimates of the population prevalence rate from any random sample (of sufficient size) of the survey respondents. In this case, a graph with HIV prevalence rate on the vertical axis and the propensity score on the horizontal axis would simply be a flat line– the likelihood of being HIV+ would not vary with the propensity score. If such a graph sloped upward, then the more likely one is to take the test, the more likely one is to be HIV+, so the actual prevalence rate would be less than the naïve rate. Conversely, if the graph sloped downward, those likely to take the test are more likely to be HIV-, and the actual prevalence rate would be greater than the naïve rate. We investigate this issue in our data in Figure 2.

Figure 2 presents a locally weighted linear regression of HIV status on the respondent's propensity score. The upper and lower lines give the bounds defined by two standard errors of the estimated regression coefficient. Figure 2 indicates that those who are especially likely to opt out are more likely to be HIV+ *and* those who are especially likely to test are also more likely to be HIV+. Were it only the case that those likely to opt out were more likely to be positive, the naïve prevalence rate would be lower than the corrected rate. Were it only the case that those likely to be tested were more likely to be positive, the naïve prevalence rate would be higher than the corrected rate. The fact that both those very likely *and* those very unlikely to take the test are more likely to be HIV positive illustrates that while selection is present, the two types of selection countervail one another so that, *in this instance*, the naïve and corrected rates are about the same.

[16] We drew 10,000 samples to compute standard errors.

An important message from Figure 2 is that it would be wrong to simply ignore selection bias *ex ante*. One can only determine whether the selection implied by the bi-modal pattern in Figure 2 results in a prevalence rate that differs from the rate by employing one or more of the methodologies used above.

V. Conclusions

Adopting a selection-on-observables assumption, we have illustrated five ways to correct HIV prevalence rates for non-random opt-out of testing. Using a nationally representative survey from Botswana, we show that opt-out is non-random. Indeed, individuals highly likely to be HIV+ are more likely to take to provide a bio-sample and individuals quite likely to be HIV- are more likely to provide a bio-sample. These potential sources of bias countervail one another and prevalence rates corrected for self-selection are only modestly less than the naïve rate.

References

Abadie, Alberto and Guido Imbens, "On the Failure of the Bootstrap for Matching Estimators," June 2006. NBER Working Paper No. T0325.

Blinder, Alan, "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 1973, *8* (4), 436–455.

Bussos, Mathias, John DiNardo, and Justin McCrary, "Testing for Parametric Consistency of Semiparametric Treatment Effects Estimators," 2008. University of Michigan, Working Paper.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux, "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, September 1996, *64* (5), 1001–1044.

Heckman, James and Salvador Navarro-Lozano, "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," *The Review of Economics and Statistics*, February 2004, *86* (1), 30–57.

Heckman, James J., Hidehiko Ichimura, Jeff Smith, and Petra Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica*, September 1998, *66* (5), 1017–1098.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, July 2003, *71* (4), 1161–1189.

Imbens, Guido and Charles F. Manski, "Confidence Intervals for Partially Identified Parameters," *Econometrica*, November 2004, *72* (6), 1845–1857.

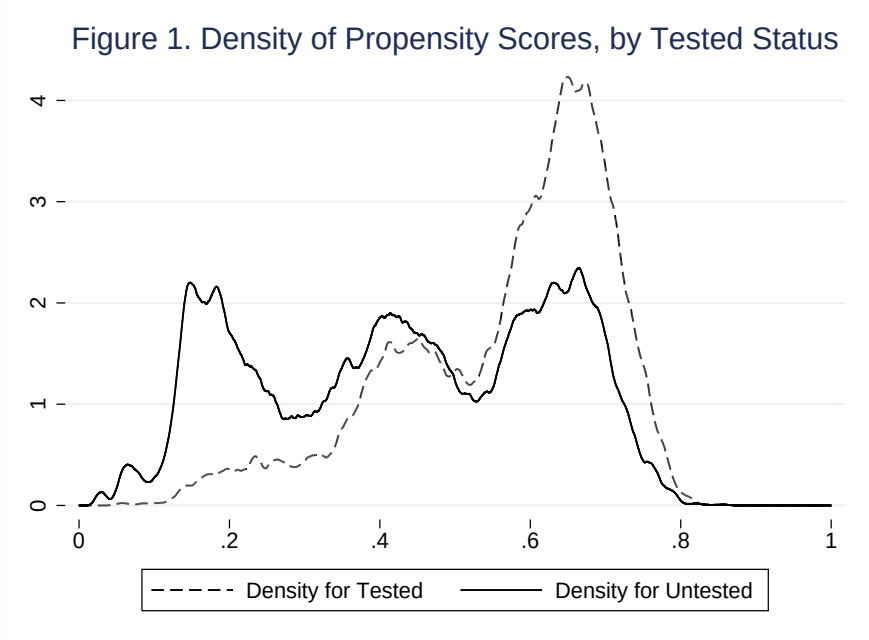
Imbens, G.W., "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 2004, *86* (1), 4–29.

Oaxaca, R., "Male-female wage differentials in urban labor markets," *International Economic Review*, 1973, *14*, 693–709.

Robins, J.M. and A. Rotnitzky, "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 1995, *90* (429), 122–129.

Table 1: Bounds for HIV Prevalence

Lower Bound	Upper Bound	Imbens-Manski Confidence Interval		Naïve Estimate
0.0858	0.5842	[0.0811	0.5957]	0.1712
(.0029)	(.0070)			(.0036)



The density for "accepted" is the line that initially lies below "refused" and then is above the "refused" line for higher values on the X-axis.

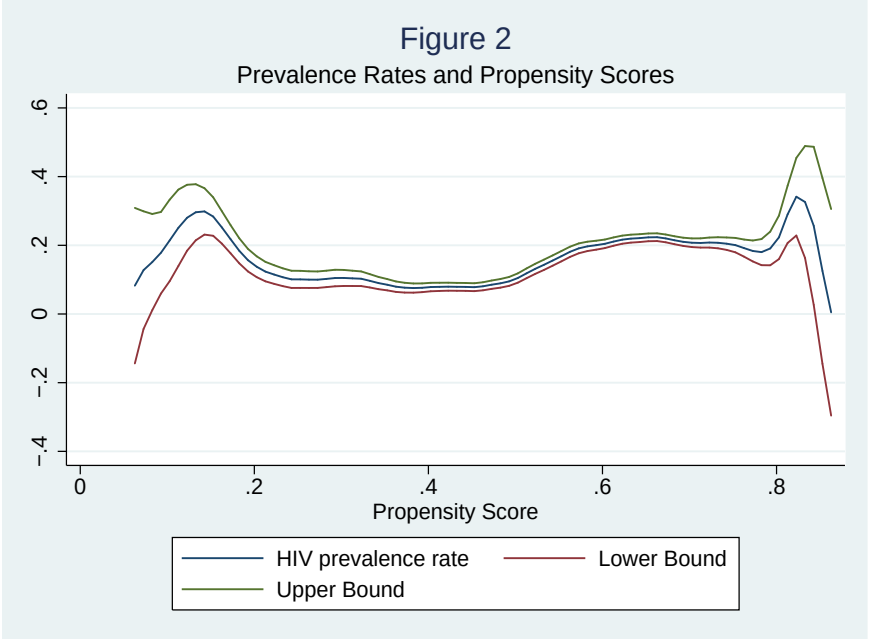


Table 2: Propensity Score Results

Variable	Excluded Category	Logit Coefficient	Impact of Specific Counterfactual
Age		-.077*	-.007*
		(.004)	(.0004)
Age^2		.0008*	
		(.00005)	
Guardian's Age		.011	.0004
		(.007)	(.0004)
Guardian's Age^2		-.0001	
		(.00008)	
Female	Male	.043	-.009
		(.029)	(.006)
Female Guardian	Male Guardian	.047	-.010
		(.050)	(.010)
Living Together	Married	-.064	.013
		(.067)	(.013)
Separated	Married	-.034	.007
		(.201)	(.042)
Divorced	Married	.212	-.043
		(.173)	(.035)
Widowed	Married	.0046	-.0009
		(.092)	(.019)
Never Married	Married	-.191*	.042
		(.061)	(.013)
Urban	Rural	-.290*	.062*
		(.050)	(.010)
Citizen	Non-citizen	.989*	-.208*
		(.099)	(.019)
Ever Had Sex	Never Had Sex	-.648	.130
		(.553)	(.101)
Used Protection	Didn't Protect	-.110*	.024*
		(.051)	(.011)
Heard of HIV	Not Heard	-.279	.060
		(.554)	(.119)
Guardian Heard of HIV	Not Heard	-.639	.133
		(1.554)	(.362)
Had a Previous HIV test	Not Tested	.184*	-.040*
		(.050)	(.011)
Guardian Previously Had HIV Test	Not Tested	.116	-.025
		(.064)	(.014)
Years of Educ.		-.131*	-.005*
		(.018)	(.001)
Years $Educ^2$		.006*	
		(.001)	
Guardian Years of Educ.		.064*	-.001
		(.033)	(.001)
Guardian Years of $Educ.^2$		-.0039*	

Notes: Standard errors are adjusted for household clusters. Dummy variables for imputed values were included in the regression but are not reported.

Table 3: Re-Balancing the Data

Variable	Coefficient	S.E.	t-Statistic
Age	-1.4877	.2418	-6.15
RW	.3128	.2489	1.26
Age^2	-186.35	18.747	-9.94
RW	38.627	19.46	1.98
Guardian Age	-1.8722	.2153	-8.69
RW	.2861	.2155	1.33
Guardian Age^2	-204.13	22.729	-8.98
RW	28.742	22.779	1.26
Female	.0317	.0060	5.28
RW	.0045	.0060	.761
Female Guardian	.0269	.0060	4.47
RW	-.0006	.0060	-.115
Living Together	.0271	.0041	6.61
RW	-.0001	.0040	-.03
Separated	.0005	.0008	.572
RW	.0001	.0008	.148
Divorced	.0020	.0010	1.86
RW	.00002	.0011	.0228
Widowed	-.0101	.0023	-4.27
RW	.0024	.0024	1.02
Never Married	.00005	.0056	.009
RW	-.0036	.0057	-.638
Urban	-.0366	.0060	-6.11
RW	-.0007	.0060	-.13
Citizen	.0393	.0026	15.1
RW	-.0019	.0026	-.718
Ever Had Sex	.2080	.0057	36.2
RW	.0066	.0059	1.12
Used Protection	.0993	.0046	21.3
RW	.0040	.0047	.851
Heard of HIV	.2688	.0058	46.3
RW	.0078	.0060	1.3
Guardian Heard	.0989	.0058	17
RW	-.0092	.0058	-1.58
Previously Tested	.0892	.0041	21.7
RW	.0032	.0041	.772
Guardian Tested	.0487	.0049	9.83
RW	-.0031	.0049	-.645
Years Educ.	.0841	.0412	2.04
RW	-.0452	.0414	-1.09
Year $Educ^2$	.9331	.6876	1.36
RW	-.7390	.6858	-1.08
Guardian Educ	-.1653	.0404	-4.09
RW	-.0067	.0406	-.167
Guardian $Educ^2$	-3.3738	.7415	-4.55
RW	-.1019	.7437	-.137

Notes: The first row of each entry gives the results without re-weighting while the second row gives results after re-weighting.

Table 4: HIV Prevalence Rates

Methodology	Prevalence Rate	Standard Error
Naïve	0.1712	0.0036
Propensity Score Reweighting	0.1599	0.0038
Pair Matching	0.1696	0.0050
Control Function	0.1630	0.0040
Double Robust	0.1633	0.0038
Blinder-Oaxaca	0.1648	0.0040