Accounting for non-response bias using participation incentives and survey design: An application using gift vouchers

McGovern, M., Canning, D., & Bärnighausen, T. (2018). Accounting for non-response bias using participation incentives and survey design: An application using gift vouchers. Economics Letters, 171, 239-244. https://doi.org/10.1016/j.econlet.2018.07.040

Published in:
Economics Letters

Document Version:
Publisher's PDF, also known as Version of record

Economics Letters 171 (2018) 239–244

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Accounting for non-response bias using participation incentives and survey design: An application using gift vouchers

Mark E. McGovern [a,b,c,*], David Canning [d], Till Bärnighausen [c,d,e]

[a] *CHaRMS — Centre for Health Research at the Management School, Queen's University Belfast, Belfast, Northern Ireland, United Kingdom*
[b] *Centre of Excellence for Public Health (Northern Ireland), United Kingdom*
[c] *Africa Health Research Institute, Somkhele, South Africa*
[d] *Harvard T. H. Chan School of Public Health, Boston, USA*
[e] *Institute of Public Health, Faculty of Medicine, Heidelberg University, Heidelberg, Germany*

HIGHLIGHTS

- Standard missing data approaches like imputation assume that data are missing at random (MAR).
- There are many contexts in which this MAR assumption is implausible.
- Heckman-type selection models can be used to test for MAR.
- Robustness to alternative selection variables and dependence structures strengthens the credibility of results.
- Randomized incentives or survey interventions provide ideal selection variables.

ARTICLE INFO

Article history:
Received 20 November 2017
Received in revised form 18 July 2018
Accepted 29 July 2018
Available online xxxx

JEL classification:
C34
J11
I13
C35

Keywords:
Participation incentives
Survey design
Selection bias
Non-ignorable missing data
Selection models
HIV

ABSTRACT

Standard corrections for missing data rely on the strong and generally untestable assumption of missing at random. Heckman-type selection models relax this assumption, but have been criticized because they typically require a selection variable which predicts non-response but not the outcome of interest, and can impose bivariate normality. In this paper we illustrate an application using a copula methodology which does not rely on bivariate normality. We implement this approach in data on HIV testing at a demographic surveillance site in rural South Africa which are affected by non-response. Randomized incentives are the ideal selection variable, particularly when implemented ex ante to deal with potential missing data. However, elements of survey design may also provide a credible method of correcting for non-response bias ex post. For example, although not explicitly randomized, allocation of food gift vouchers during our survey was plausibly exogenous and substantially raised participation, as did effective survey interviewers. Based on models with receipt of a voucher and interviewer identity as selection variables, our results imply that 37% of women in the population under study are HIV positive, compared to imputation-based estimates of 28%. For men, confidence intervals are too wide to reject the absence of non-response bias. Consistent results obtained when comparing different selection variables and error structures strengthen these conclusions. Our application illustrates the feasibility of the selection model approach when combined with survey metadata.

1. Introduction

Because of the implications for estimation, adjusting for missing data is an important component of program evaluation. Missingness can arise in various contexts including attrition in panel surveys (Thomas et al., 2001), mortality (Attanasio and Hoynes, 2000), and declining to answer particular survey questions or participate in auxiliary health or biomarker modules (Lillard et al., 1986). Adjustments are especially problematic because by definition we do not observe outcomes for non-respondents, so missing data mechanisms are generally not directly testable (Nicoletti, 2006). This has contributed to a reliance on methods which assume missing at random (MAR), often conditional on observables. These approaches include imputation (Conniffe and O'Neill, 2011), and inverse-probability weighting (Wooldridge, 2007). However, there are many contexts in which MAR may not be realistic.

Alternative selection model approaches (Heckman, 1979), simultaneously specify participation alongside the outcome without requiring MAR. However, Heckman-type selection models

[*] Correspondence to: Queen's Management School, Riddel Hall, 185 Stranmillis Road, Belfast BT95EE, Northern Ireland, United Kingdom.
E-mail address: m.mcgovern@qub.ac.uk (M.E. McGovern).

https://doi.org/10.1016/j.econlet.2018.07.040

have been criticized because alternative assumptions are necessary (Vytlacil, 2002). First, in practice an exclusion restriction is required, a variable which predicts participation but not the outcome. Plausible selection variables can be elusive (Madden, 2008), but model performance depends on their validity (Leung and Yu, 1996). Second, they can require parametric assumptions. The original formulation specified the joint distribution of the error terms in participation and outcome equations as bivariate normal. Extensions incorporate binary outcomes (Van de Ven and Van Praag, 1981) and semiparametric and nonparametric variants (Ahn and Powell, 1993; Das et al., 2003; Gallant and Nychka, 1987; Newey et al., 1990). The latter have larger data requirements and are less efficient than their parametric counterparts. Moreover, the intercept is often the quantity of interest (Heckman, 1990), and estimating the intercept in semiparametric or nonparametric selection models generally focuses on continuous outcomes and requires additional identification at infinity assumptions (Andrews and Schafgans, 1998; Schafgans and Zinde-Walsh, 2002). Reduced information inherent in binary (relative to continuous) data precludes estimation of the intercept without parametric restrictions (Klein et al., 2015).

Therefore, there is a trade-off between two sets of assumptions when attempting corrections for non-response. Lacking viable selection variables, it is understandable that researchers would proceed on the basis of MAR, even if objectively implausible. Alternative bounding approaches can be useful for avoiding this trade-off (Behaghel et al., 2015; Lee, 2009; Manski, 1990), but may not be informative when rates of non-response are high, resulting in too wide a range of possible estimates. Improving the methodology for implementing selection models therefore provides opportunities to avoid having to assume MAR.

Although well known that (quasi) experimental manipulation can solve for endogenous sorting into treatment groups, (quasi) experiments can also be used for dealing with non-response. Survey design affects participation (Hirano et al., 2001; Hill and Willis, 2001), and these findings have informed methods to reduce measurement error (Gibson et al., 2015). The resulting impact on participation has also been used to adjust for non-response bias, as these features of survey design can be used as selection variables in a Heckman-type framework. For example, Bhattacharya and Isen (2008) use a $5 gift certificate randomized to a subset of student respondents to adjust for non-response in a survey on willingness to pay for health care. Bailey (2017) examines sample selection in political surveys by randomly allocating some participants to a condition in which the political questions are asked after those on another topic. Interviewer identity is another selection variable which has been used to adjust for panel attrition (Van den Berg et al., 2007) and missing data in biomarker data (Reniers et al., 2009; Tchetgen and Wirth, 2017).

The ideal selection variable in this context is a randomized incentive or survey intervention because it is guaranteed to be unrelated to the outcome (in expectation) other than through any effect on participation. Because this approach is relatively rare, there are not many opportunities to leverage randomization to correct for missing data. However, there may be elements of survey design which are as good as random in some survey contexts ad therefore provide credible selection variables enabling this approach to be adopted more widely.

The contribution of this paper is to apply this methodology to data on HIV testing from demographic surveillance in South Africa, comparing standard approaches which assume MAR to selection model estimates. We adopt the copula-based framework developed in Marra et al. (2017) which allows flexible specification of unobserved dependence using various distributional forms. We build on this analysis by illustrating an application using two selection variables based on survey design; a food gift voucher and interviewer identity, which although not randomized, are plausibly exogenous in this survey context. We argue that showing results are robust to alternative exclusion restrictions and different distributional assumptions, as this framework allows, strengthens the conclusions from selection models.

2. Non-response in HIV research

Non-response is particularly concerning when there is an incentive not to participate. For example, people who are HIV positive may systematically opt out of testing because they fear disclosure of their status (Obare, 2010). However, accurate estimates of HIV prevalence are important because they provide information about the spread of the epidemic (Beyrer et al., 1999) and facilitate intervention evaluation (Baird et al., 2010).

Nationally representative household surveys and surveillance sites routinely include blood tests, and resulting prevalence estimates are considered the gold standard (Boerma et al., 2003). However, in some contexts less than half of eligible respondents participate (Larmarange et al., 2015). Trials are also affected; of 57 RCTs conducted before 2012 with HIV status outcomes, missing data ranged from 3% to 97% (mean 26%), with no study reporting their assumptions for managing non-response (Harel et al., 2012). Given the potential for HIV positive individuals to be systematically less likely to participate (Arpino et al., 2014), imposing an incorrect MAR assumption could result in substantial bias.

3. Participation incentives and survey design as selection variables

The Africa Health Research Institute (AHRI) cohort is a continuous survey of residents of a rural area in KwaZulu-Natal, South Africa. The main survey and HIV surveillance have provided valuable information on the epidemic for over a decade. Table 1 demonstrates that 45% of women participated in testing in 2010; compared to 33% of men. HIV prevalence among these participants was found to be 27% (women) and 16% (men). Potential implications of non-response are clear from nonparametric bounds, which, in this case, are too wide to be informative. In this paper we use the terms participation and consent to test interchangeably as relatively few individuals decline to participate in the survey (before consent to test for HIV is sought), but in other contexts they may need to be considered separately. Further information about the survey and cohort are presented in the supplementary material.

To increase participation, a gift voucher intervention was conducted in 2010. During the last 10 weeks of the surveillance, interviewers presented potential respondents with a food voucher worth 50 South African Rand to the first person they met in each household. 7% of those contacted in 2010 received a voucher, which was not conditional on consent. While not randomized, the intervention reflected concern among management about low participation in the first half of the surveillance, and apart from the timing, was not otherwise targeted. Previous evaluation found the gift voucher successfully raised participation by 25 percentage points (PP) (McGovern et al., 2016). Almost all those who received the voucher were living in households that were contacted in October or November. Once month of interview is controlled for, there is little evidence that the characteristics of those who received the voucher differ from those who did not receive the voucher (as shown Table A3 in the supplementary material). A joint test of covariates other than month yields an F statistic of 0.80, $p = 0.88$. Although this does not conclusively rule out a role for unobserved factors, and results should be interpreted with this in mind, it does provide some support for the hypothesis that the gift voucher was as good as randomly distributed (conditional on the timing of the interview). Given this, we use gift voucher receipt as a

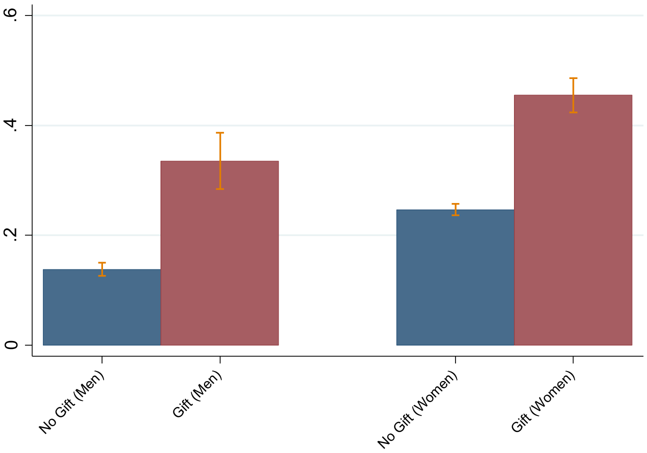

Fig. 1. AHRI 2010 HIV Prevalence by Gift Voucher Receipt. Note: For each group (received a gift voucher or did not receive a gift voucher), the HIV prevalence rate is calculated as the number of HIV positive respondents among those who participated in testing in that group divided by the number of respondents who participated in testing in that group. 95% confidence intervals are shown.

selection variable alongside an alternative based on survey design, interviewer identity. The interviewer a person has been allocated often strongly affects participation (Thomas et al., 2012), including in HIV testing (Janssens et al., 2014).

The assumption that interviewers are as good as randomly allocated may not always be appropriate. However, at AHRI, interviewers operate in teams who move from district to district as dictated by managers, attempting to contact all eligible households, and therefore the interviewer a respondent is allocated depends on survey procedure and is unlikely to be correlated with unobserved characteristics. Table A2 in the supplementary material presents results of a regression of interviewer success (which we define here as the proportion of each interviewer's interviewees who consent to test) on interviewee characteristics. However, because interviewer success will depend not only on interviewee characteristics but also on unobserved characteristics of the interviewer, it is difficult to interpret these associations. Nevertheless, we can assess the plausibility of the exclusion restriction indirectly by comparing results from two different selection variables.

Fig. 1 shows mean HIV prevalence (among participants) for AHRI residents in 2010 according to whether they received a voucher. Assuming gift receipt was exogenous, higher prevalence among those who received the incentive suggests those who would ordinarily refuse to test, but were persuaded by the gift voucher to test in this case, are more likely to be HIV positive. This is preliminary evidence of non-response bias as it indicates those who tend to decline to participate are more likely to be HIV positive.

4. Model

The standard selection model (Heckman, 1979) specifies participation and the outcome simultaneously as a function of covariates and a selection variable which enters only into the participation equation. For binary outcomes the model is a bivariate probit based on latent variables (Van de Ven and Van Praag, 1981):

$$Consent_i^* = X_i^T\beta_c + Z_j^T\gamma + u_i, i = 1, \dots, n,\ j = 1, \dots, J \qquad (1)$$

$$Consent_i = 1 \text{ if } Consent_i^* > 0, Consent_i = 0 \text{ otherwise}, \qquad (2)$$

The latent variable for whether person i with interviewer j consents to test, $Consent_i^*$, depends on individual and household characteristics, X_i, and interviewer effects, Z_j. When using the gift voucher, a vector indicating receipt replaces the interviewer effects. In addition, there is a random error term, u_i.

Similarly, the latent variable for the HIV status of person i with interviewer j, HIV_i^*, is given by:

$$HIV_i^* = X_i^T\beta_h + \epsilon_i \qquad (3)$$

$$HIV_i = 1 \text{ if } HIV_i^* > 0, HIV_i = 0 \text{ otherwise}, \qquad (4)$$

Where X_i is the same matrix of covariates included for $Consent_i^*$, and ϵ_i is a random error term. The selection variable, Z_j, does not enter into the outcome equation. We only observe HIV status conditional on consent:

$$HIV_i \text{ observed only if } Consent_i = 1, \text{ missing otherwise}. \qquad (5)$$

Previous implementations have assumed the joint error distribution is bivariate normal, $F(\epsilon_i, \mu_i) = \Phi_2(\epsilon_i, \mu_i; \rho)$, where Φ_2 is the standardized bivariate normal cumulative density function (CDF): $\Phi_2(\epsilon_i, \mu_i; \rho) = \int_{-\infty}^{\mu}\int_{-\infty}^{\epsilon} \frac{1}{2\pi\sqrt{1-\rho^2}} e^{\frac{1}{2\sqrt{1-\rho^2}}(s^2+t^2-2st\rho)} dsdt$. If the true error structure does not meet this assumption the model is misspecified, and estimates will be inconsistent (De Luca, 2008). In this paper we follow the methodology developed by Marra et al. (2017), who apply selection models to estimating HIV prevalence in three countries in sub-Saharan Africa. As described further in the supplementary material, we model the joint distribution of ϵ_i and u_i using copulae, a tractable way of specifying unobserved dependence structures while allowing margins to take a variety of different forms. Using this approach, Marra et al. (2017) find evidence of non-response bias in Zambia and Swaziland.

Dependence can be incorporated into the likelihood function through relevant copula functions, for example, $p_{11i} = P(Consent_i = 1, HIV_i = 1) = C(\Phi(X_i^T\beta_c + Z_j^T\delta), \Phi(X_i^T\beta_h); \theta)$, where Φ is the probability density function (PDF) of the bivariate normal distribution, defined as $\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]}$. The bivariate probit is equivalent to the Gaussian copula, with C then given by $C_g = \Phi_2(\Phi^{-1}(HIV), \Phi^{-1}(Consent); \theta)$, where Φ^{-1} is the quantile function of the standard univariate normal distribution.

Other copula applications have included selection models with continuous outcomes (Smith, 2003), and recursive models (Dancer et al., 2008; Murteira and Lourenço, 2011; Prieger, 2002; Winkelmann, 2012). We show results from the standard (bivariate normal) selection model along with those based on the copula selection model. An advantage of the copula approach is that it can be easily incorporated into the maximum likelihood framework, which allows the application of standard measures of model fit. As the Gaussian copula (which represents the bivariate normal case) is symmetric, this initial model can be used to identify the direction of dependence (using for instance the gamma rank association measure). If the association between error terms is estimated to be negative, as in this paper, this suggests a number of candidate copula (such as the Clayton or Joe rotated 90 or 270 degrees, as well as the Frank, which is symmetric). To determine which of these dependence structures best describe the data at hand, we identify the copula with the best fit as measured by the Akaike Information Criterion (AIC), with the lowest indicating the preferred model. In principle, other measures of fit, such as the Bayesian information criterion (BIC) could also be used for this purpose.

5. Results

HIV prevalence estimates are presented in Tables 2 (women) and 3 (men). Point estimates and confidence intervals based on respondents who consent to test (row 1), and an imputation (chained

Table 1
Participation in HIV Testing and HIV Prevalence at the 2010 AHRI Surveillance Cohort.

	Women		Men	
	No.	%	No.	%
Refused to Test	9,357	55	7,210	67
Consented to Test	7,590	45	3,527	33
Total	16,947	100	10,737	100
	Women		Men	
		%		%
Did not receive gift voucher — consented to test		42		31
Received gift voucher — consented to test		58		41
	Women		Men	
HIV Prevalence (%)	27		16	
95% CI for Nonparametric Bounds	12	68	5	73

Note: HIV prevalence estimates are based on those who participated in testing. Confidence intervals for nonparametric bounds are based on Horowitz and Manski (2006).

Table 2
Results for HIV Prevalence (Women).

Model	Selection Variable	HIV Prevalence	95% CI	Gamma Association	Copula
Complete Case		27	26–28		
Imputation		28	28–29		
Normal Selection Model	Interviewers	35	31–39	−0.42	Gaussian
Normal Selection Model	Gift voucher	33	23–42	−0.26	Gaussian
Normal Selection Model	Interviewers + Gift voucher	35	31–39	−0.40	Gaussian
Copula Selection Model	Interviewers	37	33–41	−0.49	Frank
Copula Selection Model	Gift voucher	39	31–47	−0.54	Frank
Copula Selection Model	Interviewers + Gift voucher	37	33–41	−0.46	Frank

Note: The following variables are included as predictors of consent to test for HIV and HIV status: age, month of interview, location of residence, urban/rural/peri-urban type of residence, distance to nearest clinic, distance to nearest secondary school, distance to nearest primary school, distance to nearest level 1 road, distance to nearest level 2 road, marital status, education, mother/father alive, electricity in home, fuel in home, toilet in home, water in home, and household asset index. The first row is the mean prevalence among the sample who consent to test and have a valid HIV test (complete case analysis). The second row imputes HIV prevalence for those who refused to test using the covariates described above. Row 3 implements a Heckman selection model for HIV status and consent to an HIV test using interviewer fixed effects. In row 4 the Heckman selection model uses a binary indicator for whether the respondent received the food gift voucher. The model in row 5 uses interviewers and the gift voucher intervention as exclusion restriction variables. 94 respondents who consented to test for HIV, but received indeterminate results were excluded from the procedure for estimating HIV prevalence. The copula model shown is the model with the best fit, as defined by the Akaike Information Criterion (AIC). Tables show the gamma rank association measure in column 5.

Table 3
Results for HIV Prevalence (Men).

Model	Selection Variable	HIV Prevalence	95% CI	Gamma Association	Copula
Complete Case		16	14–17		
Imputation		17	17–18		
Normal Selection Model	Interviewers	20	14–26	−0.22	Gaussian
Normal Selection Model	Gift voucher	28	12–43	−0.55	Gaussian
Normal Selection Model	Interviewers + Gift voucher	21	16–27	−0.29	Gaussian
Copula Selection Model	Interviewers	21	16–25	−0.20	Joe 90
Copula Selection Model	Gift voucher	33	19–46	−0.68	Frank
Copula Selection Model	Interviewers + Gift voucher	21	17–25	−0.24	Joe 90

Note: The following variables are included as predictors of consent to test for HIV and HIV status: age, month of interview, location of residence, urban/rural/peri-urban type of residence, distance to nearest clinic, distance to nearest secondary school, distance to nearest primary school, distance to nearest level 1 road, distance to nearest level 2 road, marital status, education, mother/father alive, electricity in home, fuel in home, toilet in home, water in home, and household asset index. The first row is the mean prevalence among the sample who consent to test and have a valid HIV test (complete case analysis). The second row imputes HIV prevalence for those who refused to test using the covariates described above. Row 3 implements a Heckman selection model for HIV status and consent to an HIV test using interviewer fixed effects. In row 4 the Heckman selection model uses a binary indicator for whether the respondent received the food gift voucher. The model in row 5 uses interviewers and the gift voucher intervention as exclusion restriction variables. 94 respondents who consented to test for HIV, but received indeterminate results were excluded from the procedure for estimating HIV prevalence. The copula model shown is the model with the best fit, as defined by the Akaike Information Criterion (AIC). Tables show the gamma rank association measure in column 5.

equations) model based on observed characteristics (row 2) are shown. We compare these to the standard bivariate normal and copula selection models. In rows 3 and 6, interviewer identity is the selection variable, while for rows 4 and 7 it is the gift voucher intervention. Rows 5 and 8 show estimates based on using both. All models include covariates and are stratified by sex to allow for differential selection effects.

Selection model estimates among women are substantially higher than imputation (28%). Point estimates are comparable using each of the selection variables, both together, and in normal and copula selection models. The Frank copula is the best fit for all models for women. Our new point estimate based on the copula selection model and both selection variables (37%) is 9 percentage points higher than the imputation model, a relative increase of 32%. There is therefore evidence that HIV positive women are less likely to participate.

For men, the evidence is less clear. Imputation estimates (17%) are lower than for normal and copula selection models using both exclusion restriction variables (21%). This 4 percentage point difference corresponds to a relative increase of 24% over imputation. However, confidence intervals are wide, which is likely to indicate too much uncertainty to rule out no selection bias, although a formal test would be required to assess the degree of statistical significance. These results are supported by analysis of the 2009 survey based on interviewers which also suggested non-response bias (McGovern et al., 2015).

6. Conclusions

Most standard approaches for dealing with missing data rely on assuming MAR, which may not be realistic if there are reasons to suspect participation is correlated with outcomes after controlling for observed characteristics. Previous research has shown that survey design can have a strong impact on participation (Hurd and Rohwedder, 2009). Here, we build on the flexible selection methodology developed in Marra et al. (2017) by demonstrating an application of how factors such as participation incentives or interviewer identity can be used to test for non-response bias. When survey metadata are combined with a copula approach, the usual assumption of bivariate normality can be relaxed, allowing for a wide variety of parametric distributions for characterizing the unobserved relationship between participation and outcome.

Our results illustrate the importance of testing the crucial assumption of MAR when non-response is substantial. Using data on HIV status, our prevalence estimates for the imputation-based MAR approach were almost identical to ignoring the missing data. For women, selection model point estimates indicate substantial non-response bias. For men, confidence intervals are overlapping and likely to be too wide to reject the hypothesis of no selection bias. However, if precisely estimated, the selection model point estimates would indicate a proportionally similar amount of non-response bias to women. Sex differences in participation in HIV testing and non-response are consistent with the hypothesis that women are more adversely affected by HIV status disclosure in this community. Given that the selection variables (interviewer identity and the gift voucher) are equally predictive of participation for men and women (i.e. women do not respond more strongly to the selection variable), this may suggest that HIV status may be a less important predictor of participating in HIV testing for men. Alternatively, overall consent rates for men are lower, which makes it harder to adjust for missing data using selection models, potentially resulting in wider confidence intervals. More efficient estimates based on additional data would be required to provide more concrete evidence on these hypotheses.

Other extensions of the model would also be interesting to pursue. For example, while the focus here has been on estimating the model intercept, the same framework can be adopted to examine the association between a predictor and an outcome. In the selection model, the association between household wealth and HIV status is weaker than that suggested by analysis of only those who participated. For example, among women being in the fifth household asset index quintile (compared to the first) is associated with a reduction of 6 percentage points in the probability of being HIV positive in the probit model, whereas it is smaller in magnitude (4 percentage points) and not statistically significant in the selection model. Similarly, when applied to an RCT affected by attrition, this approach can be used to estimate the selection-bias adjusted causal effect of the treatment. An important addition would be the development of a formal test assessing whether point estimates from alternative selection models with different selection variables, or other models assuming MAR, are statistically different. The equivalent of a Durbin–Wu–Hausman test (Nakamura and Nakamura, 1981) would be very useful in this context.

In the application in this paper we find very little difference between standard bivariate normal and copula selection models. However, as the true error structure is never observed, it is not possible to say whether this would be the case in other contexts. Given that the reliance on bivariate normality is commonly raised as a drawback of selection models (Bhattacharya and Isen, 2008; Vytlacil, 2002), the copula approach may provide valuable sensitivity analyses of alternative dependence relationships. Comparing estimates obtained using different exclusion restrictions can strengthen the credibility of results, as can demonstrating that results are not sensitive to any one parametric specification through flexible modeling of error structures.

Missing data correction is often implemented ex post, in which case the ideal randomized selection variable is unlikely to be available. Here we have used selection variables which we believe to be credible in our context, however, in the absence of randomization, careful case by case assessment of the exogeneity assumption will be necessary. Instead, if randomization of incentives and follow-up in surveys and trials could be built into the design ex ante, this would further strengthen the credibility of missing data and attrition adjustments based on the approach we outline in this paper.

Acknowledgments

We are grateful to Donal O'Neill and other seminar participants at Maynooth University, Harvard University, Queen's University Belfast, the Africa Health Research Institute, and the UNAIDS Reference Group on Estimates Modeling and Projections for comments. This project received financial assistance from the Program on the Global Demography of Aging and the National Institutes of Health [grant no. 1 P30 AG 024409-11], and the Wellcome Trust [grant no. 097410/Z/11/Z]. The funders had no role in the study design or analysis. Till Bärnighausen is funded by the Alexander von Humboldt Foundation through the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research. He is also supported by the Wellcome Trust, the European Commission, the European Union, Clinton Health Access Initiative and NICHD of NIH [grant no. R01-HD084233], NIAID of NIH [grant no. R01-AI124389 and grant no. R01-AI112339] and FIC of NIH [grant no. D43-TW009775].

Conflict of interest

None to declare.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econlet.2018.07.040.

References

Ahn, H., Powell, J.L., 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. J. Econometrics 58 (1), 3–29.

Andrews, D.W., Schafgans, M.M., 1998. Semiparametric estimation of the intercept of a sample selection model. Rev. Econom. Stud. 65 (3), 497–517.

Arpino, B., Cao, E.D., Peracchi, F., 2014. Using panel data for partial identification of human immunodeficiency virus prevalence when infection status is missing not at random. J. R. Stat. Soc. Ser. A 177 (3), 587–606.

Attanasio, O.P., Hoynes, H.W., 2000. Differential mortality and wealth accumulation. J. Hum. Resour. 35 (1), 1–29.

Bailey, M., 2017. Designing Surveys to Account for Endogenous Non-Response. Working Paper.

Baird, S., Chirwa, E., McIntosh, C., Ozler, B., 2010. The short-term impacts of a schooling conditional cash transfer program on the sexual behavior of young women. Health Econ. 19 (S1), 55–68.

Behaghel, L., Crépon, B., Gurgand, M., Le Barbanchon, T., 2015. Please call again: Correcting nonresponse bias in treatment effect models. Rev. Econ. Stat. 97 (5), 1070–1080.

Beyrer, C., Baral, S., Kerrigan, D., El-Bassel, N., Bekker, L.-G., Celentano, D.D., 1999. Expanding the space: Inclusion of most-at-risk populations in HIV prevention, treatment, and care services. J. Acquir. Immune Defic. Syndr. 57 (Suppl. 2), S96.

Bhattacharya, J., Isen, A., 2008. On inferring demand for health care in the presence of anchoring and selection biases. In: Forum for Health Economics & Policy, vol. 12, no. 2. Berkeley Electronic Press.

Boerma, J.T., Ghys, P.D., Walker, N., 2003. Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. Lancet 362 (9399), 1929–1931.

Conniffe, D., O'Neill, D., 2011. Efficient probit estimation with partially missing covariates. Adv. Econom. 27, 209–245.

Dancer, D., Rammohan, A., Smith, M.D., 2008. Infant mortality and child nutrition in Bangladesh. Health Econ. 17 (9), 1015–1035.

Das, M., Newey, W.K., Vella, F., 2003. Nonparametric estimation of sample selection models. Rev. Econom. Stud. 70 (1), 33–58.

De Luca, G., 2008. SNP and SML estimation of univariate and bivariate binary-choice models. Stata J. 8 (2), 190–220.

Gallant, A.R., Nychka, D.W., 1987. Semi-nonparametric maximum likelihood estimation. Econometrica 55 (2), 363–390.

Gibson, J., Beegle, K., De Weerdt, J., Friedman, J., 2015. What does variation in survey design reveal about the nature of measurement errors in household consumption? Oxford Bull. Econ. Stat. 77 (3), 466–474.

Harel, O., Pellowski, J., Kalichman, S., 2012. Are we missing the importance of missing values in HIV prevention randomized clinical trials? Review and recommendations. AIDS Behav. 16 (6), 1382–1393.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–161.

Heckman, J.J., 1990. Varieties of selection bias. Amer. Econ. Rev. 80 (2), 313–318.

Hill, D.H., Willis, R.J., 2001. Reducing panel attrition: A search for effective policy instruments. J. Hum. Resour. 36 (3), 416–438.

Hirano, K., Imbens, G.W., Ridder, G., Rubin, D.B., 2001. Combining panel data sets with attrition and refreshment samples. Econometrica 69 (6), 1645–1659.

Horowitz, J.L., Manski, C.F., 2006. Identification and estimation of statistical functionals using incomplete data. J. Econometrics 132 (2), 445–459.

Hurd, M., Rohwedder, S., 2009. Methodological innovations in collecting spending data: The HRS consumption and activities mail survey. Fiscal Stud. 30 (3–4), 435–459.

Janssens, W., van der Gaag, J., Rinke de Wit, T.F., Tanović, Z., 2014. Refusal bias in the estimation of HIV prevalence. Demography 51 (3), 1131–1157.

Klein, R., Shen, C., Vella, F., 2015. Estimation of marginal effects in semiparametric selection models with binary outcomes. J. Econometrics 185 (1), 82–94.

Larmarange, J., Mossong, J., Bärnighausen, T., Newell, M.L., 2015. Participation dynamics in population-based longitudinal HIV surveillance in rural South Africa. PLoS One 10 (4).

Lee, D.S., 2009. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. Rev. Econom. Stud. 76 (3), 1071–1102.

Leung, S.F., Yu, S., 1996. On the choice between sample selection and two-part models. J. Econometrics 72 (1), 197–229.

Lillard, L., Smith, J.P., Welch, F., 1986. What do we really know about wages? The importance of non-reporting and census imputation. J. Polit. Econ. 94 (3), 489–506.

Madden, D., 2008. Sample selection versus two-part models revisited: the case of female smoking and drinking. J. Health Econ. 27 (2), 300–307.

Manski, C.F., 1990. Nonparametric bounds on treatment effects. Amer. Econ. Rev. 80 (2), 319–323.

Marra, G., Radice, R., Bärnighausen, T., Wood, S., McGovern, M., 2017. A simultaneous equation approach to estimating HIV prevalence with non-ignorable missing responses. J. Amer. Statist. Assoc. 518 (12), 484–496.

McGovern, M.E., Herbst, K., Tanser, F., Mutevedzi, T., Canning, D., Gareta, D., Pillay, D., Bärnighausen, T., 2016. Do gifts increase consent to home-based HIV testing? A difference-in-differences study in rural KwaZulu-Natal, South Africa. Int. J. Epidemiol. 45 (6), 2100–2109.

McGovern, M.E., Marra, G., Radice, R., Canning, D., Newell, M.-L., Bärnighausen, T., 2015. Adjusting for non-participation bias at an HIV surveillance site in rural South Africa. J. Int. AIDS Soc. 18, 19954.

Murteira, J.M., Lourenço, Ó.D., 2011. Health care utilization and self-assessed health: specification of bivariate models using copulas. Empir. Econ. 41 (2), 447–472.

Nakamura, A., Nakamura, M., 1981. On the relationships among several specification error tests presented by Durbin, Wu, and Hausman. Econometrica 49 (6), 1583–1588.

Newey, W.K., Powell, J.L., Walker, J.R., 1990. Semiparametric estimation of selection models: Some empirical results. Amer. Econ. Rev. 80 (2), 324–328.

Nicoletti, C., 2006. Nonresponse in dynamic panel data models. J. Econometrics 132 (2), 461–489.

Obare, F., 2010. Nonresponse in repeat population-based voluntary counseling and testing for HIV in rural Malawi. Demography 47 (3), 651–665.

Prieger, J.E., 2002. A flexible parametric selection model for non-normal data with application to health care usage. J. Appl. Econometrics 17 (4), 367–392.

Reniers, G., Araya, T., Berhane, Y., Davey, G., Sanders, E.J., 2009. Implications of the HIV testing protocol for refusal bias in seroprevalence surveys. BMC Public Health 9 (1), 1–9.

Schafgans, M., Zinde-Walsh, V., 2002. On intercept estimation in the sample selection model. Econometric Theory 18 (01), 40–50.

Smith, M.D., 2003. Modelling sample selection using archimedean copulas. Econom. J. 6 (1), 99–123.

Tchetgen, E., Wirth, K., 2017. A general instrumental variable framework for regression analysis with outcome missing not at random. Biometrics 73 (4), 1123–1131.

Thomas, D., Frankenberg, E., Smith, J.P., 2001. Lost but not forgotten: Attrition and follow-up in the Indonesia family life survey. J. Hum. Resour. 36 (3), 556–592.

Thomas, D., Witoelar, F., Frankenberg, E., Sikoki, B., Strauss, J., Sumantri, C., Suriastini, W., 2012. Cutting the costs of attrition: Results from the Indonesia family life survey. J. Dev. Econ. 98 (1), 108–123.

Van den Berg, Gerard, Lindeboom, Maarten, Lopez, Marta, 2007. Interviewer Identities as Valid Instruments for Selective Panel Survey Attrition - An Evaluation with Matched Survey-Register Data. Working Paper.

Van de Ven, W.P., Van Praag, B., 1981. The demand for deductibles in private health insurance: A probit model with sample selection. J. Econometrics 17 (2), 229–252.

Vytlacil, E., 2002. Independence, monotonicity, and latent index models: An equivalence result. Econometrica 70 (1), 331–341.

Winkelmann, R., 2012. Copula bivariate probit models: With an application to medical expenditures. Health Econ. 21 (12), 1444–1455.

Wooldridge, J.M., 2007. Inverse probability weighted estimation for general missing data problems. J. Econometrics 141 (2), 1281–1301.

Accounting for Non-Response Bias using Participation Incentives and Survey Design: An Application Using Gift Vouchers

Mark E. McGovern, David Canning, Till Bärnighausen

Appendix

Selection Bias in HIV Testing

Most of what we know about the impact and spread of the HIV epidemic in low and middle income countries comes from data collected from blood tests taken from respondents in nationally representative household surveys and surveillance sites which track the residents of specific geographic areas. Household surveys in countries without developed health service infrastructure often include routine blood draws at the end of their standard interviews. For example, this type of testing is conducted in many of the Demographic and Health Surveys (Fabic et al., 2012). These representative datasets are important because they facilitate estimation of HIV prevalence and the change in HIV prevalence over time. These estimates are required for policy as they provide information about the spread of the HIV epidemic (Beyrer et al., 1999), allow for targeting of at risk communities (Tanser et al., 2013), inform about the factors which are protective against infection (De Walque, 2007; Case and Paxson, 2013), illustrate the impact of HIV and AIDS on government services and economic growth (Bloom and Mahal, 1997; Case and Paxson, 2011), and are used to evaluate the population effectiveness of HIV interventions (Baird et al., 2010). Data obtained from testing in HIV surveillance surveys have therefore been highly influential for informing our understanding of the HIV epidemic. However, they have the drawback that rates of participation in testing in these surveys can be low. Studies which use imputation to correct for this missing data tend to find little difference between imputed estimates and estimates based on analysis of cases without missing data (Mishra et al., 2008).

A key question is therefore whether the assumption of missing at random is reasonable in this context. Unfortunately, there are three reasons to be skeptical. First, respondents who are asked to provide blood for an HIV test have an incentive to decline to participate if they know or suspect they are HIV positive because the potential costs of disclosure are high (Parker and Aggleton, 2003). Second, when those who decline are asked to explain why, a high proportion give reasons related to having been tested previously or already knowing their status (Kranzer et al., 2008). Third, there are occasional opportunities to observe longitudinal information on HIV testing. These data support the hypothesis that those who are HIV positive are less likely to participate in testing (Obare, 2010; Bärnighausen et al., 2012). For example, data from Malawi found that HIV positive residents were more than 4 times more likely to refuse to test (Reniers and Eaton, 2009). Previous research based on Heckman-type selection models and cross-sectional data has also found evidence of selection bias in some HIV surveys (Reniers et al., 2009; Janssens et al., 2014).

Data

The Africa Health Research Institute (AHRI) cohort is a continuous survey of residents of a rural area approximately 434km^2 in KwaZulu-Natal, South Africa, which has been conducted since 2003. Regular HIV testing of the residents takes place, of whom there are around 90,000 in any given year. This predominantly Zulu-speaking region remains one of the poorest in South Africa, and has experienced very high rates of HIV prevalence, along with recent scale-up of antiretroviral treatment in the locality. The data have been highly influential in informing understanding of the impact of HIV and AIDS on individuals, families, and communities, and are publicly accessible (after registration) from www.ahri.org.

As part of the main survey, data are collected on a semi-annual basis from a key informant in each household (Tanser et al., 2008). The topics covered include the characteristics of individual household members and important events in their lives (such as births, deaths, and migration), the attributes of the household (such as assets and facilities), as well as data on the physical structures themselves. The HIV surveillance cohort is nested within the main household survey and is conducted on a subset of residents (since 2007 every resident aged 15 and over has been eligible). In some years of the surveillance (before 2007 and in 2009) results are made available to those who tested, however as very few residents attempt to obtain information on their HIV status through the surveillance testing, results are not provided to surveillance participants in all years, including in 2010. This community already has very good access to rapid HIV testing and results through public-sector HIV counselling and testing. When the annual HIV surveillance is in progress, those residents who are eligible are visited by teams of two trained interviewers. In accordance with WHO and UNAIDS guidelines, these interviewers approach potential participants and seek written consent for them to obtain a blood sample. If consent is given, they prepare a dried blood spot sample collected by finger prick. If eligible individuals are not present at the household when interviewers attempt to contact them for participation, the team makes three follow-up attempts to contact the individual by revisiting the household. Once a dried blood spot is obtained, there is no identifying information for that sample, only a unique numerical code is retained to link the HIV test result with the combined HIV surveillance and household surveys datasets.

Because the main survey collects information from one key informant per household, participation in the main survey is almost universal. However, the HIV surveillance survey collects information from individuals, and a limitation of the surveillance data is that participation rates are low (Larmarange et al., 2015), and participation rates are low mainly because residents decline consent for a blood test rather than not being found for contact. Given the evidence we discuss above, this raises concerns about the accuracy of estimates based on either analysis of cases without missing data (i.e. only residents who participate in testing, ignoring those who do not participate) or imputation. In this paper we focus on the 2010 HIV surveillance because during this year a participation incentive was offered to a subset of residents. In 2010, 40,789 residents were identified from the Africa Center database as being eligible for participation in the HIV surveillance. Of these, 7,400 were found to have migrated, become sick or disabled, or had died when consent was sought. A further 5,611 residents were found to be ineligible or could not be found due to incorrect demographic or contact information. Only 186 residents declined to participate in the surveillance before being asked to take a HIV test. 27,684 individuals were successfully contacted to participate in HIV testing and had a valid test result (a small number had an indeterminate reading). Descriptive statistics are shown in Table A1.

Interviewers as Predictors of Participation in HIV Testing

Interviewers have been found to be highly predictive of participation in HIV testing in previous studies, including at HIV surveillance sites such as AHRI (Clark and Houle, 2014; McGovern et al., 2015). Therefore, we also examine whether the interviewers influence the likelihood of their interviewees participating. Amongst female residents in 2010 there were 78 interviewers, and amongst male residents there were 72 interviewers. The median number of interviews conducted per interviewer (the number of residents from whom consent to test for HIV was sought by the interviewer) was 124.5 for women and 174 for men. The median participation rate per interviewer (the number of residents from whom consent to test for HIV was obtained by the interviewer divided by the number of residents from whom consent to test for HIV was sought by the interviewer) was 33% for men and 34% for women. Good interviewers were equally good at raising participation rates for both men and women. For example, the 25th percentile of interviewer consent is 18% for men and 13% for women, while the 75th percentile for interviewer consent is 45% for men and 50% for women. In order to summarize the effect of having a good interviewer on participation in HIV testing, we ran a logistic regression for participation on an indicator variable for having been interviewed by an interviewer who was over the 75th percentile for participation (interviewer consent rates in this regression were calculated as the leave-one-out rate where the individual interviewee was excluded from the numerator and denominator when calculating their interviewer's participation rate in order to avoid a mechanical correlation between the dependent variable and interviewer participation rates), adjusting for the other covariates used in the main analysis. We find an odds ratio for consent of having a good interviewer of 2.1, i.e. having a good interviewer doubled the probability that the interviewee would consent to participate in testing.

Without further data, it is difficult to further separate out the effectiveness of interviewers from the characteristics of their interviewees. In Table A2 we collapse the data at the interviewer level, and regress interviewer effectiveness (defined as the proportion of their interviewees who consented to test) on the characteristics of their interviewees to determine whether better interviewers are systematically associated with particular factors. Because of the simultaneity inherent in this relationship, as well as the fact that we only have 60 observations at the interviewer level (once we collapse those who conducted few interviews), we interpret the following results with caution. There are relatively few interviewees in the categories for marital status and household fuel type, leaving only rural location and some of the distance variables as statistically significant predictors of interviewer success. Collecting or making available additional metadata on interviewer characteristics (such as age, education, and experience) would be helpful for augmenting this analysis.

Table A3 presents a similar analysis for gift voucher receipt (a linear probability model for whether the respondent lived in a household which received a voucher as a function of individual-level characteristics.)

Table A1: 2010 AHRI Surveillance Cohort Descriptive Statistics

	Median	Mean	SD	N
Participated in HIV Testing	0	0.402	0.49	27,684
HIV Positive	0	0.236	0.425	11,117
Male	0	0.388	0.487	27,684
Received Gift Voucher	0	0.07	0.255	27,684
Household Has Piped Water	1	0.595	0.491	27,684
Household Has Flush Toiled	0	0.074	0.262	27,684

Type of Location	No.	%
Peri-Urban	8,404	30.36
Rural	17,205	62.15
Urban	2,075	7.5
Total	27,684	100
Month of Interview in 2010		
January	1,319	4.76
February	3,283	11.86
March	3,996	14.43
April	2,995	10.82
May	3,334	12.04
June	908	3.28
July	1,223	4.42
August	2,949	10.65
September	2,278	8.23
October	2,460	8.89
November	2,505	9.05
December	434	1.57
Total	27,684	100
Marital Status		
Married	4,099	14.81
Polygamous	579	2.09
Divorced/Separated/Widowed	2,827	10.21
Engaged	472	1.7
Never Married	14,987	54.14
Under Legal Age	4,315	15.59
Unknown/Other	405	1.46
Total	27,684	100
Mother is Alive		
Dead	22,464	81.14
Alive	4,734	17.1
Unknown	486	1.76
Total	27,684	100
Father is Alive		
Dead	21,861	78.97
Alive	5,207	18.81
Unknown	616	2.23
Total	27,684	100
Distance to Nearest Clinic		
0-1 Km	3,779	13.65
1-2 KM	6,383	23.06
2-3 KM	5,718	20.65
3-4 KM	4,948	17.87
4-5 KM	2,962	10.7
5 KM+	3,894	14.07
Total	27,684	100
Distance to Nearest Primary School		
0-1 Km	12,282	44.36
1-2 KM	12,087	43.66
2-3 KM	2,876	10.39
3 KM+	439	1.59
Total	27,684	100
Distance to Nearest Level 1 Road		
0-1 Km	7,268	26.25
1-2 KM	2,985	10.78
2-3 KM	1,410	5.09
3-4 KM	1,219	4.4
4-5 KM	1,370	4.95
5 KM+	13,432	48.52
Total	27,684	100

Household has Electricity	No.	%
Yes	17,452	63.04
No	5,156	18.62
N/A	4,483	16.19
Unknown	593	2.14
Total	27,684	100
Household Fuel Type		
Electricity	13,473	48.67
Coal or Wood	6,856	24.77
Gas	1,319	4.76
Other	918	3.32
Unknown	4,478	16.18
N/A	640	2.31
Total	27,684	100
Household Asset Index Quintile		
Lowest	4,402	15.9
2nd Lowest	4,520	16.33
Middle	4,649	16.79
2nd Highest	4,624	16.7
Highest	4,210	15.21
Missing	5,279	19.07
Total	27,684	100
Education		
None	2,903	10.49
Primary	2,817	10.18
Junior Secondary	4,480	16.18
Upper Secondary	10,255	37.04
Don't Know	2,112	7.63
Unknown	5,117	18.48
Total	27,684	100
Age Group		
15-19	5,247	18.95
20-24	4,434	16.02
25-29	3,185	11.5
30-34	2,246	8.11
35-39	2,005	7.24
40-44	1,790	6.47
45-49	1,835	6.63
50-54	1,717	6.2
55-59	1,221	4.41
60+	4,004	14.46
Total	27,684	100
Distance to Nearest Secondary School		
0-1 Km	6,717	24.26
1-2 KM	9,730	35.15
2-3 KM	6,913	24.97
3-4 KM	2,676	9.67
4-5 KM	1,005	3.63
5 KM+	643	2.32
Total	27,684	100
Distance to Nearest Level 2 Road		
0-1 Km	12,204	44.08
1-2 KM	7,716	27.87
2-3 KM	4,416	15.95
3-4 KM	2,282	8.24
4-5 KM	729	2.63
5 KM+	337	1.22
Total	27,684	100

Note to Table A2: An OLS model for the proportion of an interviewer's interviewees who consented to test is shown. The regression is at the interviewer level (one observation per interviewer, with interviewers who conducted fewer than 50 interviewers collapsed into one category). Categorical variables measure the proportion of each interviewer's interviewees in that category, while distance and age variables are measured in average KM and years, respectively.

Table A2: Predictors of Interviewer Success (Interviewer Level)

Variables	**Interviewer Consent Rate**	**Variables**	**Interviewer Consent Rate**
Male	0.175	**Education (Omitted=None)**	
	(0.142)	Primary	1.097
Marital Status (Omitted=Married)			(2.339)
Polygamous	-0.906	Junior Secondary	0.670
	(2.157)		(1.262)
Divorced/Separated/Widowed	0.675	Upper Secondary	-1.227
	(0.909)		(1.385)
Engaged	3.616**	Don't Know	-2.387
	(1.504)		(1.476)
Never Married	0.081	Missing	-3.264
	(0.533)		(1.911)
Under Legal Age	0.410		
	(0.897)	**Flush Toilet Access**	1.442
Missing/Other	-1.904		(1.164)
	(2.969)		
Household Location (Omitted=Peri-urban)		**Piped Water Access**	-0.594
Rural	0.734***		(0.486)
	(0.221)		
Urban	0.928	**Distance to Nearest Clinic (KM)**	-0.015
	(1.152)		(0.070)
Household Electricity (Omitted=Yes)		**Distance to Nearest Secondary School (KM)**	0.219*
No	-2.309		(0.110)
	(1.411)	**Distance to Nearest Primary School (KM)**	-0.615***
N/A	4.245		(0.153)
	(3.353)	**Distance to Nearest Level 1 Road (KM)**	-0.020
Missing	-9.670*		(0.014)
	(5.481)	**Distance to Nearest Level 2 Road (KM)**	0.314***
Household Fuel Type (Omitted=Electric)			(0.101)
Coal or Wood	-0.624		
	(0.477)	**Age**	-0.021
Gas	-1.325		(0.012)
	(1.402)		
Other	-3.512*	**Household Asset Index Quintile (Omitted=1)**	
	(1.806)	2	-0.263
Missing	-2.751		(0.985)
	(4.245)	3	-1.999
Unknown	8.466		(1.562)
	(6.152)	4	-0.600
			(1.615)
Constant	1.716	5	-1.840
	(2.492)		(2.082)
		Missing	-0.662
Observations	60		(4.246)
R-squared	0.923		

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note to Table A3: A linear probability model for whether the individual lived in a household which received the voucher is shown. Standard errors are clustered at the household level. Almost all vouchers were disbursed among households which were contacted in October and November.

Table A3: Predictors of Voucher Receipt (Individual Level)

Variables	Voucher Received	Variables	Voucher Received
Male	-0.001	**Education (Omitted=None)**	
	(0.001)	Primary	0.001
Marital Status (Omitted=Married)			(0.001)
Polygamous	-0.003	Junior Secondary	0.000
	(0.002)		(0.001)
Divorced/Separated/Widowed	-0.001	Upper Secondary	0.002
	(0.001)		(0.001)
Engaged	-0.004	Don't Know	-0.001
	(0.004)		(0.002)
Never Married	-0.001	Missing	-0.005
	(0.002)		(0.003)
Under Legal Age	-0.000	**Distance to Nearest Clinic (Omitted=0-1KM)**	
	(0.002)	1-2 KM	-0.001
Missing/Other	-0.014		(0.002)
	(0.012)	2-3 KM	0.001
Month Contacted (Omitted=January)			(0.001)
February	-0.000	3-4 KM	-0.000
	(0.000)		(0.001)
March	0.001	4-5 KM	0.001
	(0.001)		(0.001)
April	-0.000	5 KM+	0.001
	(0.001)		(0.001)
May	-0.000	**Distance to Nearest Secondary School (Omitted=0-1KM)**	
	(0.000)	1-2 KM	-0.001
June	-0.000		(0.001)
	(0.000)	2-3 KM	0.001
July	-0.000		(0.001)
	(0.001)	3-4 KM	0.001
August	0.000		(0.001)
	(0.000)	4-5 KM	-0.001
September	0.000		(0.001)
	(0.001)	5 KM+	-0.002*
October	1.000***		(0.001)
	(0.001)	**Distance to Nearest Primary School (Omitted=0-1KM)**	
November	0.999***	1-2 KM	0.001
	(0.001)		(0.001)
December	0.500***	2-3 KM	0.000
	(0.027)		(0.001)
Household Location (Omitted=Peri-urban)		3 KM +	0.003
Rural	-0.001		(0.003)
	(0.001)	**Distance to Nearest Level 1 Road (Omitted=0-1KM)**	
Urban	0.006	1-2 KM	-0.003
	(0.004)		(0.003)
Mother Alive(Omitted=No)		2-3 KM	-0.001
Alive	-0.001		(0.001)
	(0.001)	3-4 KM	0.000
Missing	0.010		(0.002)
	(0.012)	4-5 KM	-0.001
Father Alive(Omitted=No)			(0.002)
Alive	0.000	5 KM+	-0.001
	(0.001)		(0.002)
Missing	0.008	**Distance to Nearest Level 2 Road (Omitted=0-1KM)**	
	(0.008)	1-2 KM	-0.000
Household Electricity (Omitted=Yes)			(0.001)
No	0.000	2-3 KM	-0.001
	(0.001)		(0.001)
N/A	0.004*	3-4 KM	0.002
	(0.002)		(0.003)
Missing	-0.001	4-5 KM	0.006
	(0.004)		(0.005)
Household Fuel Type (Omitted=Electric)		5 KM+	-0.001
Coal or Wood	-0.001		(0.002)
	(0.001)	**Age Group (Omitted=¡20)**	
Gas	-0.000	20-24	0.000
	(0.001)		(0.002)
Other	-0.001	25-29	-0.001
	(0.002)		(0.002)
Missing	0.002	30-34	-0.003
	(0.002)		(0.002)
Unknown	0.000	35-39	-0.003
	(0.002)		(0.002)
Household Asset Index Quintile (Omitted=1)		40-44	0.001
2	-0.001		(0.002)
	(0.001)	45-49	0.001
3	0.001		(0.002)
	(0.001)	50-54	0.001
4	-0.000		(0.002)
	(0.002)	55-59	0.002
5	-0.002		(0.002)
	(0.002)	60+	0.002
Missing	-0.002*		(0.002)
	(0.001)		
		Constant	0.003
Flush Toilet Access	-0.004		(0.003)
	(0.004)		
Piped Water Access	0.000	**Observations**	27,684
	(0.001)	**R-squared**	0.973

Clustered standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Further Details of the Copula Approach

Following Marra et al. (2017), we model the joint distribution of the error terms using copulae. A major advantage of this approach is that these models can be estimated in a standard maximum likelihood framework, resulting in consistent, efficient and asymptotically normal estimators (Smith, 2003), with the log-likelihood (abstracting from the interviewer subscript) given by:

$$\sum_{i=1}^{n} Consent_i \times HIV_i log(p_{11i}) + Consent_i \times (1 - HIV_i) log(p_{01i}) + (1 - Consent_i) log(p_{0i}) \tag{1}$$

Where p_{0i} is the probability of declining to test, $P(Consent_i = 0)$, p_{11i} is the probability of being HIV positive and consenting to test, $P(Consent_i = 1, HIV_i = 1)$, p_{01i} is the probability of being HIV negative and consenting to test, $P(Consent_i = 1, HIV_i = 0)$. Copula functions can be incorporated into the likelihood function to map multivariate distributions to their marginal distributions. In this case, we are concerned with the copula mapping function C that links the two-dimensional cumulative density function for HIV status and consent to test to the relevant one dimensional margins, $F(HIV, Consent) = C(F_{HIV}(HIV), F_{Consent}(Consent); \theta)$, where θ is an association parameter indicating the degree of dependence.

A number of copulae have been proposed, each with different dependence structures, including, for example, the Frank, Gumbel, Clayton and Joe copulae. While the Frank copula is similar to the Gaussian, the Gumbel, Clayton and Joe are asymmetric, allowing the case where those who are most likely to be HIV negative do not have a greater dependence to test than those who are moderately likely to be HIV negative, while those who are most likely to be HIV positive are those who are the least likely to test. In addition, the rotated versions (90 degrees, 180 degrees, and 270 degrees) of these copula are easily obtained and allow for greater density in either tail of the distribution (Brechmann and Schepsmeier, 2013).

In the standard bivariate probit selection model, the error terms are assumed to be independent and identically distributed (i.i.d.) with means equal to zero, constant variances equal to one, and covariance (correlation coefficient) ρ. Therefore, ρ is the key parameter in the model (u_i and ϵ_i), and if those who are HIV positive are less likely to participate in testing (conditional on observed characteristics), we expect negative dependence. In the copula selection models, a nonparametric measure of association, such as Kendall's Tau (τ) or the gamma association measure (shown in Tables 2 and 3 in the main text), is more appropriate as the dependence modelled by copulae is typically non-linear.

By implementing the selection model allowing for a variety of different forms of dependence, the copula approach allows us to establish whether the results from the standard bivariate probit selection model are sensitive to the assumption of joint normality. Bivariate normality places restrictions on the dependence structure, for example it assumes that the propensity to participate given HIV status (conditional on covariates) is symmetric. Along with the difficulty finding viable exclusions restrictions, this parametric formulation is likely an impediment to the wider use of selection models for dealing with missing data because it cannot be verified and can be seen as arbitrary (Vytlacil, 2002). As these models are estimated under the standard maximum likelihood framework, we are able to use standard diagnostic tools for model fit, and choose our preferred copula specification based on information criteria.

Coupled with plausible exclusion restrictions based on survey design, relaxing the parametric assumptions required for identification in selection models for binary outcomes should remove an impediment to wider use of selection models for dealing with missing data without requiring the undesirable and untestable assumption of missing at random. Moreover, as we discuss in the main text, the copula approach is a viable means of relaxing the bivariate normality assumption when dealing with dichotomous dependent variables and the intercept is a parameter of interest. Even when the outcome of interest is continuous, there are important advantages to the copula method over the semi- and nonparametric approaches. Specifically, the latter require a much larger set of parameters to be estimated, potentially rendering them inefficient and prohibiting a comprehensive set of covariates, and typically involve numerical integration or complex simulation procedures. In contrast, copula models can be estimated under a conventional maximum likelihood framework, allowing, for example, model selection to be based on standard information criteria (Pigini, 2015).

References

S. Baird, E. Chirwa, C. McIntosh, and B. Özler. The short-term impacts of a schooling conditional cash transfer program on the sexual behavior of young women. *Health Economics*, 19(S1):55–68, 2010.

T. Bärnighausen, F. Tanser, A. Malaza, K. Herbst, and M.-L. Newell. HIV status and participation in HIV surveillance in the era of antiretroviral treatment: a study of linked population-based and clinical data in rural south africa. *Tropical Medicine & International Health*, 17(8):e103–e110, 2012.

C. Beyrer, S. Baral, D. Kerrigan, N. El-Bassel, L.-G. Bekker, and D. D. Celentano. Expanding the space: Inclusion of most-at-risk populations in HIV prevention, treatment, and care services. *Journal of Acquired Immune Deficiency Syndromes*, 57(Suppl 2):S96, 1999.

D. E. Bloom and A. S. Mahal. Does the AIDS epidemic threaten economic growth? *Journal of Econometrics*, 77(1):105–124, 1997.

E. C. Brechmann and U. Schepsmeier. Modeling dependence with c-and d-vine copulas: The r-package cdvine. *Journal of Statistical Software*, 52(3):1–27, 2013.

A. Case and C. Paxson. The impact of the AIDS pandemic on health services in Africa: evidence from Demographic and Health Surveys. *Demography*, 48(2):675–697, 2011.

A. Case and C. Paxson. HIV Risk and Adolescent Behaviors in Africa. *American Economic Review*, 103(3): 433–438, 2013.

S. J. Clark and B. Houle. Validation, Replication, and Sensitivity Testing of Heckman-Type Selection Models to Adjust Estimates of HIV Prevalence. *PLOS ONE*, 9(11):e112563, 2014.

D. De Walque. How does the impact of an HIV/AIDS information campaign vary with educational attainment? Evidence from rural Uganda. *Journal of Development Economics*, 84(2):686–714, 2007.

M. S. Fabic, Y. Choi, and S. Bird. A systematic review of Demographic and Health Surveys: data availability and utilization for research. *Bulletin of the World Health Organization*, 90(8):604–612, 2012.

W. Janssens, J. van der Gaag, T. F. Rinke de Wit, and Z. Tanović. Refusal bias in the estimation of HIV prevalence. *Demography*, 51(3):1131–1157, 2014.

K. Kranzer, N. McGrath, J. Saul, A. C. Crampin, A. Jahn, S. Malema, D. Mulawa, P. E. Fine, B. Zaba, and J. R. Glynn. Individual, household and community factors associated with HIV test refusal in rural Malawi. *Tropical Medicine & International Health*, 13(11):1341–1350, 2008.

J. Larmarange, J. Mossong, T. Bärnighausen, and M.-L. Newell. Participation Dynamics in Population-Based Longitudinal HIV Surveillance in Rural South Africa. *PLOS ONE*, 10(4), 2015.

G. Marra, R. Radice, T. Bärnighausen, S. Wood, and M. McGovern. A Simultaneous Equation Approach to Estimating HIV Prevalence with Non-Ignorable Missing Responses. *Journal of the American Statistical Association*, 518(12):484–496, 2017.

M. McGovern, G. Marra, R. Radice, D. Canning, M.-L. Newell, and T. Bärnighausen. Adjusting for Non-Participation Bias at an HIV Surveillance Site in Rural South Africa. *Journal of the International AIDS Society*, 18:19954, 2015.

V. Mishra, B. Barrere, R. Hong, and S. Khan. Evaluation of bias in HIV seroprevalence estimates from national household surveys. *Sexually Transmitted Infections*, 84(Suppl 1):i63–i70, 2008.

F. Obare. Nonresponse in repeat population-based voluntary counseling and testing for HIV in rural Malawi. *Demography*, 47(3):651–665, 2010.

R. Parker and P. Aggleton. HIV and AIDS-related stigma and discrimination: a conceptual framework and implications for action. *Social Science & Medicine*, 57(1):13–24, 2003.

C. Pigini. Bivariate non-normality in the sample selection model. *Journal of Econometric Methods*, 4(1), 2015.

G. Reniers and J. Eaton. Refusal bias in HIV prevalence estimates from nationally representative seroprevalence surveys. *AIDS*, 23(5):621, 2009.

G. Reniers, T. Araya, Y. Berhane, G. Davey, and E. J. Sanders. Implications of the HIV testing protocol for refusal bias in seroprevalence surveys. *BMC Public Health*, 9(1):163, 2009.

M. D. Smith. Modelling sample selection using archimedean copulas. *Econometrics Journal*, 6(1):99–123, 2003.

F. Tanser, V. Hosegood, T. Bärnighausen, K. Herbst, M. Nyirenda, W. Muhwava, C. Newell, J. Viljoen, T. Mutevedzi, and M.-L. Newell. Cohort profile: Africa centre demographic information system (ACDIS) and population-based HIV survey. *International Journal of Epidemiology*, 37(5):956–962, 2008.

F. Tanser, T. Bärnighausen, E. Grapsa, J. Zaidi, and M.-L. Newell. High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-natal, south africa. *Science*, 339(6122):966–971, 2013.

E. Vytlacil. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70 (1):331–341, 2002.