Economica

Economica (2019) **86**, 229–261
doi:10.1111/ecca.12286

Risky Sexual Behaviours: Biological Markers and Self-reported Data

By Lucia Corno† and Áureo de Paula‡

†*Cattolica University, Institute of Fiscal Studies, EUDN and LEAP* ‡*University College London, São Paulo School of Economics—FGV, CeMMAP and Institute of Fiscal Studies*

Final version received 12 June 2018.

High-risk sexual behaviours are generally unobserved and difficult to identify. In this paper, we investigate the accuracy of two risky-behaviour measures: biomarkers for sexually transmitted infections (STIs) and self-reported data. We build an epidemiological model to assess the relative performance of biomarkers versus self-reported data. We then suggest an econometric strategy that combines both types of measures to estimate *actual* unobserved risky sexual behaviours. Using data from the Demographic and Health Survey in 28 countries, we calibrate the model and provide conditions under which self-reported data are a better proxy for risky sexual behaviours than biomarkers. In countries with low STI prevalence, biomarkers have a higher probability of misclassification than self-reported answers. We apply our econometric strategy to the data and show that the probability of *actual* risky behaviour is much higher than the probability of self-reported risky behaviour and of testing positive for an STI.

Introduction

Unsafe sexual behaviour and the associated exposure to infection is one of the major causes of preventable mortality in low-income countries (after childhood underweight and unsafe water) (World Health Organization (WHO) 2009). It is the main mode of transmission of sexually transmitted infections (STIs), including human immunodeficiency virus HIV/AIDS, and human papillomavirus, which together annually kill more than one million people worldwide.[1] Correctly measuring risky sexual behaviour is therefore key to targeting disease prevention to vulnerable individuals and to designing effective strategies for improving global health.

However, sexual behaviours are largely private and it is difficult to measure them precisely. Early studies on the determinants of risky sexual behaviours mainly use self-reported data. Those have been criticized as unreliable because respondents may be uncomfortable disclosing their high-risk activities or may prefer to give socially desirable answers.[2] Underreporting of sex-related behaviour is a well-documented phenomenon in both the medical and economic literatures (Fenton *et al*. 2001; Ozler 2013; Gersovitz *et al*. 1998; de Paula *et al*. 2014). For example, de Walque (2007) shows that married people tend not to report extramarital sexual relationships, and Palen *et al*. (2008) find that adolescents may not report that they have already had sexual intercourse.[3] Recent research has therefore started to collect biological markers (or biomarkers) for STIs (i.e. syphilis, trichomoniasis, gonorrhea, etc.) as an objective measure of risky sexual behaviours (Juerges *et al*. 2013; Carrieri and Jones 2015), providing some evidence that self-reported answers might underestimate high-risk activities.[4] Given that STIs result only from unprotected sex with infected individuals, biomarkers for STIs are becoming a popular method to objectively measure unsafe sexual behaviours in experimental and non-experimental studies, and to identify riskier subpopulations in HIV/AIDS prevention programmes[5] (Cleland *et al*. 2004; Gallo *et al*. 2006; Mauck and Straten 2008; de Walque *et al*. 2012; Anderson *et al*. 2013; Gong 2015; Björkman Nyqvist *et al*.

2015; Tennekoon and Rosenman 2014). However, like other indirect records of risky behaviour, biomarkers themselves are not perfect: individuals who adopt risky sexual behaviours may remain uninfected and hence be misclassified as not having behaved in a risky manner.

The goal of this paper is twofold. First, we investigate, theoretically and empirically, the relative accuracy of two risky-behaviour measurements: biomarkers for STIs—both curable (e.g. syphilis) and non-curable (e.g. HIV)—and self-reported data. Second, we suggest an econometric method to estimate unobserved actual risky behaviour in a given population by combining data on biomarkers and self-reported responses.

To the best of our knowledge, no previous research has so far evaluated the reliability of biomarkers in the context of sexual behaviour: whether unsafe sexual behaviours are better measured by self-reported data or biomarkers or a combination of both remains an open question. Identifying the most rigorous method by which to measure sexual activities is crucial to targeting riskier individuals, to boosting safer sexual behaviours and ultimately to reducing HIV/AIDS prevalence. Among its most recent policy recommendations, UNAIDS argues that social behavioural programmes need to be more strategically focused on groups at higher risk of infection (UNAIDS 2013).[6]

First, we extend a standard epidemiological model to characterize the probability of misclassification of risky sexual behaviours with biomarkers. Risky sexual behaviours are misclassified by biomarkers when they are not detected by the presence of an STI (i.e. testing negative for an STI but having behaved in a risky manner). We use the model to compare the misclassification rate of biomarkers with self-reported data on risky sexual behaviour.[7]

In the model, the probability of transmission of an STI from an infected person to an uninfected person depends on the proportion of infected people in the population, the number of partners, the number of sexual contacts per partner, the probability of infection from an infected partner and, finally, on the likelihood of meeting an infected partner. By using STIs as a proxy for unsafe sexual behaviours, individuals who become infected in a given period are tagged as having behaved in a risky manner. The probability of correct classification using biomarkers for those who engage in risky sexual behaviour is therefore equal to the probability of transmitting disease from an infected individual to an uninfected individual. On the other hand, the probability of correct classification using self-reported sexual behaviours is defined as the likelihood of eliciting truthful answers in a survey. Since misclassification is possible in both cases, it is not *ex ante* clear that either one is a superior marker for risky behaviour.

Second, we suggest an econometric framework that combines both types of measures —biomarkers and self-reported data—to improve the estimation of correlates of risky sexual behaviours and to estimate the probability of true unobserved risky behaviour. Our strategy uses the information from biomarkers for STIs (i.e. STI-positive or STI-negative) to estimate the probability of correct classification with self-reported data. This strategy turns on the idea that the proportion of STI-infected people who report not having engaged in risky activities is informative about the rates of misclassification arising from self-reported behaviour. Building on previous work by Hausman *et al.* (1998), we provide a consistent estimator for the parameters of a binary outcome econometric model for risky behaviour. The estimator can be easily computed by a generalized method of moments (GMM), using commercially available packages.

Third, we calibrate the parameters of the epidemiological model using data on 28 countries from all the Demographic and Health Surveys with publicly available information on STIs and sexual behaviours, and show the conditions under which self-

reported data are a better proxy for risky sexual behaviours than biomarkers, and vice versa. Results show that in countries with low STI prevalence, the biomarkers have a higher probability of misclassification than self-reported answers.

Finally, we estimate the association between several individual characteristics and actual risky sexual behaviour using our proposed empirical strategy. We find a huge discrepancy in the fraction of high-risk individuals measured in three different ways: self-reported data, biomarkers, and the combination of the two measures with our proposed GMM strategy. In particular, our GMM strategy estimates that the probability of actual risky behaviour is 78.7%, while the estimated probability of self-reported risky behaviour is 17.4% and the predicted probability of testing positive for an STI is 4.3%.

We believe that our findings may help researchers and policymakers to estimate risky sexual behaviours in a more rigorous way. First, we show that when self-reported data and biomarker baselines are both available or easy to collect, a combination of these two measures would provide a more reliable proxy for risky sex. Precisely measuring risky sexual behaviour is important, for example, in cases when biomarkers for only one STI are available but not for other common ones, or in cases when risky sexual behaviours are correlated with other types of dangerous behaviours such as alcohol consumption, drug abuse and criminal activities (see, for example, Connor *et al*. 2015; Choudhry *et al*. 2014; Siegal *et al*. 1999). However, when baseline and endline biomarker data are available, they may still be a better proxy for risky sex compared to self-reported measures.[8]

Second, our econometric strategy can be applied to estimate risky behaviours in other contexts. Indeed, our estimator will be feasible as long as an assessment for the probability of truthful reporting is viable. One possibility is to infer the probability of truthful reporting using the proportion of individuals who admit the 'risky' behaviour among those who are positively marked for that behaviour. An example would be in the analysis of models for tax evasion or avoidance. Here, elicited responses on tax (non-) compliance would be the elicited behavioural response, and one could use reporting among those that do not have a registration with the tax authorities similarly to our use of self-reported behaviour among those infected with an STI. Another potential application would be in the context of list randomization, a technique developed to elicit truthful responses to sensitive questions. List randomization has been used to study, for example, microfinance loans (Karlan and Zinman 2012), illegal migration (McKenzie and Siegel 2013), and the use of condoms among Senegalese sex-workers (Treibich and Lépine 2016). The procedure relies on providing a list of non-sensitive questions with yes or no answers ('control group') and including the sensitive question for a randomly selected subset of respondents ('treatment group'). Each respondent reports only the count of affirmative answers in the list, but not which. If affirmative and negative answers for the non-sensitive questions are approximately balanced, respondents will be less likely to worry that their answer to the sensitive question can be identified. An estimate of the proportion of individuals (truthfully) answering yes to the sensitive question is then the difference in average counts between 'treatment' and 'control'. If the questionnaire also elicits direct responses for the sensitive question, it is possible to obtain an estimate for the probability of correct elicitation (given the behaviour of interest).

The remainder of the paper is organized as follows: Section I provides an epidemiological model to compare the probability of misclassification of risky sexual behaviour using biomarkers for STIs and self-reported data; Section II combines the two in a new estimator. Section III describes the data used for simulating the model, and in Section IV we show the results. Section V concludes and discusses the implications of our findings.

I. An Epidemiological Model for Biomarker Misclassification

Setup

In this section, we present a model of misclassification of risky sexual behaviours when they are measured by biomarkers for STIs. Risky sexual behaviours are misclassified when they are not detected if measured by the presence of an STI. We use the model to draw comparisons with the probability of self-reporting information on risky sexual behaviours in a survey, and generate suggestions to improve inferences on such behaviours.

We define a risky sexual intercourse as one that allows for the transmission of the STI. On the contrary, protected sexual contacts that preclude the transmission of the STI (by, for instance, using condoms) are not classified as risky behaviour in the model. [9]

Our theoretical framework is built on epidemiological models of disease dynamics (see, for example, Anderson and May 1991; Hethcote 2000; Hyman *et al.* 2001). The key output of the model is the transmission rate of the disease, λ, that is, the rate at which uninfected individuals are infected by infected partners.[10] We suppose that the (annual) transmission rate λ depends on the number of partners per individual (p), the total share of infected individuals in the population (I), the average number of sexual contacts per partner ($c(p)$), and the probability of infection by an infected partner ($\beta(p)$). Following Hyman *et al.* (2001), we model the average number of sexual contacts with each partner as a decreasing function in the number of partners per year:

$$c(p) = 104p^{-\eta} + 1, \tag{1}$$

where η is a positive parameter that controls how fast the number of sexual contacts decreases with the number of partners. As in Hyman *et al.* (2001), we set η equal to 1. With one partner per year ($p = 1$), the above functional form reasonably implies 105 instances of sexual intercourse in a year: roughly two per week. Using data from the Malawi Diffusion and Ideational Change Project for 2004, for example, about one-third of unmarried female respondents report having sexual intercourse at least twice a week (see Table A1 in the Appendix). With 20 partners per year ($p = 20$), for instance, the number of sexual contacts per partner would be 6.2 per year. So as the number of partners increases, the number of encounters per partner tends to 1. If the number of contacts per partner is not available from the data, one can use equation (1) to bound the probability of misclassification, since in the case above, $1 \leq c(p) \leq 105$.[11] Note that alternative functional forms may also be used to calibrate the number of sexual contacts per partner and to incorporate other observables (whenever those are available) such as the length of partnership.

Following Hyman *et al.* (2001), if someone has p partners and a given partner is infected, then the probability of infection from that partner, $\beta(p)$, depends on the average number of sexual contacts with a given partner, $c(p)$, and is given by the probability that the disease is transmitted in at least one sexual encounter:

$$\beta(p) = 1 - (1 - \xi)^{c(p)}, \tag{2}$$

where ξ is the probability of transmission from a single contact with an infected person, and $(1-\xi)^{c(p)}$ is the probability that an individual will avoid infection when he or she has $c(p)$ contacts with an infected partner.[12]

Finally, we assume that infected and uninfected individuals may have different numbers of sexual partners (p_I and p_U, respectively). We model the probability that an uninfected person meets an infected partner as

$$P_{UI} = \frac{\rho p_I I}{\rho p_I I + p_U(1-I)},$$

where $\rho \geq 0$ controls the degree of sorting between infected and uninfected individuals. When $\rho = 1$, the person meets every individual with equal and independent probability regardless of their infection status. There is perfect sorting when $\rho = 0$: an uninfected individual meets only other uninfected individuals. Finally, when $\rho \to \infty$, uninfected individuals meet only infected ones.[13]

The likelihood that an uninfected individual becomes infected during a given year is therefore

$$\lambda = 1 - \left(1 - \beta(p_U)\frac{\rho p_I I}{\rho p_I I + p_U(1-I)}\right)^{p_U}. \tag{3}$$

This probability is equal to the probability that the disease is transmitted by at least one of the sexual partners, which is obtained as one minus the probability that it is not transmitted by any of them. The probability that the disease is transmitted by each one of the sexual partners is given by the probability that this person meets an infected partner, P_{UI}, and the disease is transmitted by that partner, $\beta(p_U)$. Since there are p_U sexual partners, this is given by

$$\beta(p_U)\frac{\rho p_I I}{\rho p_I I + p_U(1-I)}.$$

One minus this term equals the probability that the disease is not transmitted. Raising this term to the power p_U gives the probability that the disease is not transmitted by any sexual partner. λ is the probability of the complementary event.

If $p_I = p_U = p$, then λ is given by

$$\lambda = 1 - \left(1 - \beta(p)\frac{\rho I}{\rho I + (1-I)}\right)^{p}. \tag{4}$$

A limitation of equations (3) and (4) is that since the information is generally self-reported, one may have only imperfect data on the number of sexual partners. If individuals overreport the number of sexual partners, then the calibrated value of λ will be larger than if the value were obtained with truthful reports. To see this, assume for simplicity that $p_U = p_I = p$. Then expressing λ as $1 - \exp(\ln(1-\lambda))$ shows that

$$\begin{aligned}\frac{\mathrm{d}\lambda}{\mathrm{d}p}; &= -(1-\lambda)\frac{\mathrm{d}\ln(1-\lambda)}{\mathrm{d}p} \\ &= -(1-\lambda)\frac{\mathrm{d}\left(1-\beta(p)\frac{\rho I}{\rho I+(1-I)}\right)}{\mathrm{d}p} \\ &= (1-\lambda)\frac{\rho I}{\rho I+(1-I)}\frac{\mathrm{d}\beta(p)}{\mathrm{d}p} > 0.\end{aligned}$$

The second equality uses equation (4), and the inequality stems from $d\beta(p)/dp > 0$. This is positive since an increase in p leads to a reduction in the number of sexual contacts per partner (equation (1)), which in turn leads to a higher probability of transmission in at least one sexual encounter $\beta(p)$ (equation (2)). An analogous argument states that when individuals underreport the number of sexual partners, λ is lower than if the number of partners is truthfully reported.

To partially address this limitation, in the empirical analysis we provide lower and upper bounds on the number of partners. Specifically, as a lower bound we considered the number of partners reported by married women—who generally tend to underreport them—and as an upper bound we look at the number of partners reported among single men—who tend to overreport them. The discrepancy in self-reported risky behaviour among married women and single men is well established in the literature (see, for example, Wellings *et al.* 2006; Nnko *et al.* 2004; Oster 2005). This evidence may be indicating underreporting or overreporting by both genders relative to their true measures of risky sex.

Of course, when panel data are available, it is possible to estimate the true value of λ by computing the rate at which individuals who were previously STI-negative contract an infection a year later. If direct measurement of λ is not possible, then an epidemiological model like the one described above allows for the translation of alternative available information (on prevalence, etc.) into the transmission rate λ.

Comparing biomarkers and self-reported measurements

Biomarkers can record whether or not an individual adopted a risky sexual behaviour (i.e. the 'extensive margin'), but they are less informative about its intensity (i.e. the 'intensive margin'). Hence, in our comparison between biomarkers and elicited measures of risky behaviours, it seems adequate to encode those into a binary variable.

Let Y^t describe whether an individual *truly* engaged in risky behaviour ($Y^t = 1$) or not ($Y^t = 0$). Denote by Y^e the variable indicating whether *elicited* behaviour (e.g. in a survey) is reported to be risky ($Y^e = 1$) or not ($Y^e = 0$). Finally, α denotes the probability of correct classification (i.e. marking someone who engaged in risky behaviour as having behaved in a risky way) using elicited sexual behaviour: $\alpha = \mathbb{P}(Y^e = 1 | Y^t = 1)$. Here, we assume that people have no incentives to report risky sexual behaviours if they did not engage in such behaviours (i.e. $\mathbb{P}(Y^e = 0 | Y^t = 0) = 1$, and consequently $\mathbb{P}(Y^e = 1 | Y^t = 0) = 0$).

On the other hand, by using biomarkers for STIs as a proxy for risky sexual behaviours, all those who become infected in a given period are tagged as having behaved in a risky manner. Here, we assume that all the biomarkers for STIs are able to detect the disease.[14]

The probability of correct classification using biomarkers for those who engage in risky sexual behaviour is then given by $\lambda \in [0,1]$. We also assume that those who do not engage in risky sexual behaviour are not infected (though see note 9). This means that the probability of correct classification using biomarkers for those who do not behave in a risky manner is 1.

Since misclassification is possible in both cases, it is not *ex ante* clear that either one is a superior marker for risky behaviour. For example, when the infection rate is low enough (either because of low prevalence or because of low transmission rates), the biomarkers would misclassify risky behaviour more often. This is formalized in the following result.

Proposition 1 If $\lambda < \mathbb{P}(Y^e = 1)$, then the biomarker has a higher probability of misclassification of risky behaviour than behaviour elicited by the survey questionnaire.

This result is easily established by noting that $\lambda\mathbb{P}(Y^t = 1) \leq \lambda$ and

$$\begin{aligned}\mathbb{P}(Y^e = 1) &= \mathbb{P}(Y^e = 1|Y^t = 1) \times \mathbb{P}(Y^t = 1) + \mathbb{P}(Y^e = 1|Y^t = 0) \times \mathbb{P}(Y^t = 0)\\ &= \mathbb{P}(Y^e = 1|Y^t = 1) \times \mathbb{P}(Y^t = 1)\\ &\equiv \alpha\mathbb{P}(Y^t = 1).\end{aligned}$$

If $\lambda < \mathbb{P}(Y^e = 1)$, then we get that $\lambda < \alpha\mathbb{P}(Y^t = 1)$. This implies that $\lambda\mathbb{P}(Y^t = 1) < \alpha\mathbb{P}(Y^t = 1)$ and thus $\lambda < \alpha$.

An interesting aspect of this proposition is that we can compare data on elicited sexual behaviour with reasonable values for λ to establish whether the above inequality holds.[15] Note also that it is possible to have $\lambda > \mathbb{P}(Y^e = 1) = \alpha\mathbb{P}(Y^t = 1)$ and $\lambda < \alpha$, and biomarkers may still be less reliable even if the transmission rate λ is higher than the elicited rate of risky sexual behaviour $\mathbb{P}(Y^e = 1)$.

For treatable STIs (e.g. syphilis), the calibrated rate λ is a measure of risky sexual behaviour over the reference period of interest (e.g. a year) and is therefore seen as an incidence rate. Note that the incidence rate for curable STIs with short cycles in a given population is usually close to the prevalence rate for curable STIs (WHO 2008). For non-treatable STIs (e.g. HIV) or treatable STIs not cured within the reference period, other measurements, such as the prevalence rate I, may also be used to assess risky behaviours over a longer horizon and as a lasting marker for past risky behaviours (Baird*et al.* 2014). In this case, one can just replace λ with I in Proposition 1.[16]

In the above result, a maintained assumption is that $\mathbb{P}(Y^e = 1|Y^t = 0) = 0$. This is a realistic assumption, but circumstances where individuals misreport a non-risky behaviour can be postulated. Whereas Proposition 1 would not accommodate this, our econometric strategy below can be adapted to such circumstances. In this case, one can estimate $\mathbb{P}(Y^e = 1|Y^t = 0)$ and test whether this probability is equal to zero.

II. An Econometric Model Combining Measurements

The epidemiological model developed in the previous section allows us to establish the best marker for risky sexual activities (see Proposition 1). In this section, we describe how the combination of biomarkers and self-reported data can be used to address the misclassification issue and more precisely estimate actual unobserved risky sexual behaviours. Specifically, we suggest an econometric framework to estimate a set of parameters θ that characterize the correlates of risky sexual behaviours by combining biomarkers and self-reported data. We then show how to estimate the probability of actual high-risk behaviours. Our estimation strategy builds on previous work by Hausman *et al.* (1998). But while they exploit non-linearities in the econometric model to obtain identification using one potentially misclassified measurement of the outcome of interest, we use two measurements of the same outcome.

We start by setting the relationship between risky sexual behaviours and observable covariates of interest (i.e. the determinants of risky sexual behaviours). In particular, we assume that the relationship between risky sexual behaviour and observable covariates can be summarized by the following conditional probability specification:

$$\mathbb{P}(Y^t = 1|\mathbf{X}) = F(\mathbf{X};\theta), \tag{5}$$

where $\mathbf{X}$ are the covariates of interest. Here, $F(\cdot)$ is known up to parameters θ and differentiable in θ (e.g. $F(\mathbf{X};\theta) = \Phi(\mathbf{X}^\top\theta)$ if the model corresponds to a probit, $F(\mathbf{X};\theta) = \Lambda(\mathbf{X}^\top\theta)$ if the model corresponds to a logit, or $F(\mathbf{X};\theta) = \mathbf{X}^\top\theta$ if the model corresponds to a linear probability model). Then the probability of reporting risky behaviour given $\mathbf{X}$ is

$$\begin{aligned}\mathbb{P}(Y^e = 1|\mathbf{X}) = {} & \mathbb{P}(Y^e = 1|Y^t = 1,\ \mathbf{X}) \times \mathbb{P}(Y^t = 1|\mathbf{X}) \\ & + \mathbb{P}(Y^e = 1|Y^t = 0,\ \mathbf{X}) \times \mathbb{P}(Y^t = 0|\mathbf{X}).\end{aligned}$$

Assume that $\mathbb{P}(Y^e = 1|Y^t = 0,\ \mathbf{X}) = 0$ and, for simplicity, that $\mathbb{P}(Y^e = 1|Y^t = 1,\ \mathbf{X}) = \mathbb{P}(Y^e = 1|Y^t = 1)$. Using equation (5), we obtain

$$\mathbb{P}(Y^e = 1|\mathbf{X}) = \mathbb{P}(Y^e = 1|Y^t = 1)F(\mathbf{X};\theta) = \alpha F(\mathbf{X};\theta). \tag{6}$$

This relationship implies that a model using elicited behaviour alone as a proxy for risky behaviour would lead to bias in the estimation of θ (see Hausman *et al.* 1998). Similar arguments establish that using biomarkers alone would also lead to bias in the estimation of θ.[17]

For example, in a linear probability model where $F(\mathbf{X};\theta) = \mathbf{X}^\top\theta$, the probability that $Y^e = 1$ given $\mathbf{X}$ is

$$\mathbb{P}(Y^e = 1|\mathbf{X}) = \mathbb{P}(Y^e = 1|Y^t = 1)\mathbf{X}^\top\theta = \alpha\theta^\top\mathbf{X},$$

and a linear probability model would estimate $\alpha\theta$, which is smaller in absolute value than θ (since $\alpha < 1$). Similarly with biomarkers:

$$\mathbb{P}(Y^{STI} = 1|\mathbf{X}) = \mathbb{P}(Y^{STI} = 1|Y^t = 1)\mathbf{X}^\top\theta = \lambda\theta^\top\mathbf{X},$$

where $\|\lambda\theta\| < \|\theta\|$). Consequently, estimation of θ based solely on elicited behaviour or biomarkers will suffer from attenuation bias (see Hausman *et al.* 1998).

Assume now that the probability of correct classification with elicited sexual behaviours can be inferred from the proportion of individuals who admit to having engaged in risky behaviours among those tested positive for the STI:

$$\mathbb{P}(Y^e = 1|Y^t = 1) = \mathbb{P}(Y^e = 1|Y^{STI} = 1). \tag{7}$$

Note that there are situations when the above assumption may be inappropriate. This would happen, for instance, if the subset of individuals who became infected misreported differently from those who did not (this would imply that $\mathbb{P}(Y^e = 1|Y^t = 1,\ Y^{STI} = 1)$ is not equal to $\mathbb{P}(Y^e = 1|Y^t = 1)$). When respondents are unaware of their status prior to the survey, this concern is less plausible. Another possibility is that someone reports (correctly) no risky sexual behaviour, but tests positive for the STI after being infected non-sexually (e.g. blood transfusion or needle-sharing). This would imply that

$\mathbb{P}(Y^e = 1|Y^{STI} = 1)$ is not necessarily equal to $\mathbb{P}(Y^e = 1|Y^t = 1,\ Y^{STI} = 1)$. Insofar as such non-sexual transmission channels are infrequent, again this would not be a cause for the inadequacy of equation (7) (Schmid *et al*. 2004; Hall *et al*. 2008; Tennekoon and Rosenman 2014). Note that equation (7) can also be made conditional on observable covariates (i.e. individual specific characteristics).

Substituting equation (7) into equation (6), we obtain

$$\mathbb{P}(Y^e = 1|\mathbf{X}) = \mathbb{P}(Y^e = 1|Y^{STI} = 1)F(\mathbf{X};\theta).$$

Since both $\mathbb{P}(Y^e = 1|\mathbf{X})$ and $\mathbb{P}(Y^e = 1|Y^{STI} = 1)$ are estimable, the above relationship can be used to consistently estimate θ. We also note that the probability of misclassification by self-reported risky behaviour can be made dependent on covariates, yielding a straightforward generalization of the equations above.[18]

Our approach is based on the regression model

$$Y^e = \mathbb{P}(Y^e = 1|Y^{STI} = 1)F(\mathbf{X};\theta) + u,$$

and uses the moment condition

$$\mathbb{E}[Y^e - \mathbb{P}(Y^e = 1|Y^{STI} = 1)F(\mathbf{X};\theta)|\mathbf{X}] = 0.$$

Let $\hat{\alpha}$ be the estimator for $\mathbb{P}(Y^e = 1|Y^{STI} = 1)$ based on the frequency of individuals for whom $Y^e = 1$ among STI-positive individuals. Then one estimator for θ is the non-linear least squares estimator, defined by the minimizer of the quadratic function

$$Q_N(\theta) \equiv \sum_{i=1}^{N} \left[y_i^e - \hat{\alpha}F(\mathbf{x}_i;\theta)\right]^2, \tag{8}$$

where i indexes the individual observations, and N is the sample size. The estimator can be framed as a GMM estimator and computed in standard packages (e.g. Stata). Standard textbook arguments deliver consistency and asymptotic normality for the above estimator. We summarize those results in the following proposition.

Proposition 2 Under random sampling, assume that

$$\mathbb{E}\left[\sup_{\theta} \|\partial_\theta F(\mathbf{X};\theta)\|\right] < \infty \quad \text{and} \quad \mathbb{E}\left[\sup_{\theta} \|\partial^2_{\theta\theta^\top} F(\mathbf{X};\theta)\|\right] < \infty,$$

and that the parameter space for θ is compact. Then the estimator $\hat{\theta} \equiv \arg\min_\theta Q_n(\theta)$ is ($\sqrt{n}$-) consistent for θ and asymptotically normal:

$$\sqrt{n}\left(\begin{bmatrix}\hat{\theta}\\ \hat{\alpha}\end{bmatrix} - \begin{bmatrix}\theta_0\\ \alpha_0\end{bmatrix}\right) \longrightarrow_d \mathcal{N}(0, \mathbf{G}^{-1}\mathbf{S}(\mathbf{G}^{-1})^\top), \tag{9}$$

where $\mathbf{G}$ and $\mathbf{S}$ are given by

$$\mathbf{G} = \mathbb{E}\begin{bmatrix} \alpha\partial_\theta F(\mathbf{X};\theta)\partial_\theta F(\mathbf{X};\theta)^\top & F(\mathbf{X};\theta)\partial_\theta F(\mathbf{X};\theta) \\ 0 & Y^{STI} \end{bmatrix}$$

and

$$\mathbf{S} = \mathbb{E}\begin{bmatrix} (\alpha F(\mathbf{X};\theta) - Y^e)^2\partial_\theta F(\mathbf{X};\theta)\partial_\theta F(\mathbf{X};\theta)^\top & (\alpha F(\mathbf{X};\theta) - Y^e)\, Y^{STI}(\alpha - Y^e)\partial_\theta F(\mathbf{X};\theta) \\ (\alpha F(\mathbf{X};\theta) - Y^e)\, Y^{STI}(\alpha - Y^e)\partial_\theta F(\mathbf{X};\theta)^\top & Y^{STI}(\alpha - Y^e)^2 \end{bmatrix}.$$

Proof The estimator can be cast as the GMM estimator minimizing

$$\left\| \sum_{i=1}^{N} g(\mathbf{Z}_i, (\theta, \alpha))_{\dim(\theta)+1\times 1} \right\|^2,$$

where

$$g(\mathbf{Z}_i, (\theta, \alpha)) = (\alpha F(\mathbf{X};\theta) - Y^e)\partial_\theta F(\mathbf{X};\theta)\, Y^{STI}(\alpha - Y^e)^\top.$$

This estimator is consistent and asymptotically normal with displayed asymptotic variance where $\mathbf{G} = \mathbb{E}[\partial_{(\theta,\alpha)} g(\mathbf{Z}_i, (\theta, \alpha))]$ and $\mathbf{S} = \mathrm{var}(g(\mathbf{Z}_i,(\theta,\alpha)))$ (see, for example, Hayashi 2001). The dominance condition $\mathbb{E}[\sup_\theta \|\partial_\theta F(\mathbf{X};\theta)\|] < \infty$ guarantees the dominance condition for consistency (see Hayashi 2001, prop. 7.7), and both dominance conditions guarantee the dominance condition for asymptotic normality (see Hayashi 2001, prop. 7.10).

For a linear probability model, the estimator for $\hat{\theta}$ can be simply computed as the ratio between the OLS estimates, where the dependent variable is Y^e, and $\hat{\alpha}$, the proportion of people reporting risky sex among the STI-positive respondents. In the linear model, standard errors need nevertheless to be adjusted as indicated in expression (9). More generally, this estimator can be estimated in most statistical packages (using, for example, the command `gmm` in Stata).

Our strategy can be employed in different contexts whenever the probability for truthful reporting or correct classification follows equation (7). That is when it is possible to infer the probability of truthful reporting using the proportion of individuals who admit the 'risky' behaviour among those who are positively marked for that behaviour. One example would be in analysing models for tax evasion or avoidance. More or less direct questions on individuals' attitudes toward tax avoidance or evasion are sometimes used as an indication for individual tax compliance. In this case, the outcome Y^e in equation (7) would indicate elicited tax non-compliance (i.e. equal to 0 if a firm is reporting to pay taxes, and 1 otherwise). In that equation, Y^{STI} could be whether a firm does not have a tax registration with the authorities. In this case, equation (7) would hold if the probability of (elicited) non-compliance given (true) non-compliance is the same as the probability of (elicited) non-compliance given that a firm does not have a registration with the authorities. Our estimation strategy could then be used to investigate the proportion of firms who truly do not pay taxes and the correlates of tax non-compliance.

Another potential use of our proposed estimator among applied economists is in the context of 'list randomization'. List randomization has been used in several papers (Karlan and Zinman 2012; McKenzie and Siegel 2013; Treibich and Lépine 2016) to obtain estimates of the fraction of individuals with an outcome for which direct elicitation may be prone to misclassification. The procedure relies on providing a list of non-sensitive questions with yes or no answers and including the sensitive question for a randomly selected subset of respondents. Each respondent reports only the count of affirmative answers in the list. If affirmative and negative answers for the non-sensitive questions are approximately balanced, then respondents will be less likely to worry that their answer to the sensitive question can be identified. An estimate of the proportion of individuals (truthfully) answering 'yes' to the sensitive question is then the difference in average counts between 'treatment' and 'control'. If the questionnaire also elicits direct responses for the sensitive question, then it is also possible to obtain an estimate for $\mathbb{P}(Y^e = 1|Y^t = 1)$. If one takes the difference between the averages of affirmative answers for 'treatment' and 'control' questionnaires among those that provide an affirmative answer to the elicitation question, then one gets an estimate for $\mathbb{P}(Y^t = 1|Y^e = 1)$, and multiplying this by the proportion of individuals who answer yes to the elicitation question, which is an estimate of $\mathbb{P}(Y^e = 1)$, one can then obtain an estimate for $\mathbb{P}(Y^e = 1 \text{ and } Y^t = 1)$. The difference between the average counts for 'treatment' and 'control' questionnaires gives an estimate for $\mathbb{P}(Y^t = 1)$, which can be combined with the previous estimate to obtain $\mathbb{P}(Y^e = 1|Y^t = 1)$. This can be used in the same way as the proxy in equation (7) to analyse the correlates of Y^t. Hence our strategy can also be seen as a way to incorporate survey instruments in (possibly non-linear) regression settings.

III. Data and Descriptive Statistics

The data used to calibrate the epidemiological model (described in Section I) and to estimate actual risky behaviour using our GMM procedure (described in Section II) come from the Demographic and Health Surveys (DHS). We include all the most recent DHS with available information on biomarkers for curable and non-curable STIs and on self-reported sexual activities, ending up with a sample of 28 countries. Table A2 in the Appendix reports the list of countries and the years included in the analysis.

The DHS are nationally representative, and the survey methodology is uniform across countries. Data are collected using a two-stage sampling design. In the first stage, a sample of clusters is selected from a list of enumeration areas from the latest national census of each country. In the second stage, a complete list of households is created in each cluster. In each randomly selected household, all women aged 15–49 and all men aged 15–59 who were either permanent residents or visitors present in the household on the night before the survey were eligible to be interviewed. Crucially for our purposes, the DHS, besides eliciting detailed information on respondents' sexual behaviour, include data on biomarkers for STIs: all women and men eligible to be interviewed were asked to voluntarily provide a blood sample to be tested for HIV in order to determine national prevalence rates, and in a subsample of every three households, a syphilis test was performed on eligible women and men who consented to be tested. In particular, biomarker data on a curable STI (syphilis) are available for Zambia and on a non-curable STI (HIV) are available for all the countries listed in Appendix Table A2. The STI tests were conducted after the survey. Table A2 also reports the share of individuals who voluntarily decided to take the HIV and syphilis tests, and those who replied to the

survey questions on sexual behaviours. On average, more than 60% of respondents have been tested for HIV, while given the sampling procedure, only 18% of respondents in Zambia have been tested for syphilis. Looking at the non-response rate on sexual behaviours, 44% of the respondents answered the question on condom use during the last intercourse, 63% reported the number of partners in the last 12 months, and 36% replied to the question on extramarital affairs.[19]

Our final sample, after discarding observations with missing information on biomarkers for HIV, includes 426,527 individuals, 53% of women and 47% of men. Table 1 reports the features of the sample. In panel A, we report sociodemographic characteristics. The average age of the individuals in the sample is nearly 30 years, and approximately 60% of them are currently married. Looking at the highest educational level, 25% of the respondents did not have any formal education, 31% have primary education, and about 44% have achieved secondary education or a college degree.

TABLE 1
DESCRIPTIVE STATISTICS

	Obs. (1)	Mean (2)	S.D. (3)	Min (4)	Max (5)
Panel A: Sociodemographic characteristics					
Female	435,375	0.53	0.50	0	1
Age	435,375	29.67	10.70	15	64
Married	435,375	0.60	0.49	0	1
No education	435,375	0.25	0.43	0	1
Primary education	435,375	0.31	0.46	0	1
Secondary education and above	435,375	0.44	0.50	0	1
Urban	435,375	0.40	0.49	0	1
Panel B: Biomarkers					
Syphilis-positive	2,392	0.04	0.20	0	1
HIV-positive	435,375	0.04	0.20	0	1
Panel C: Self-reported sexual behaviours					
Condom used last intercourse	304,818	0.15	0.36	0	1
Number of partners in last 12 months	304,786	1.15	0.87	1	60
Extramarital sex last intercourse	246,978	0.02	0.15	0	1
Risky sex	302,689	0.11	0.31	0	1

Notes
Pooled sample of all Demographic and Health Surveys data with available information on biomarkers for HIV, syphilis and sexual behaviour. In panel B, biomarker data on syphilis are available only for Zambia. In panel C, we include the sample of individuals who had sex at least once. 'Condom used last intercourse' is a binary variable equal to 1 if the respondent reported using a condom during the last intercourse, 0 otherwise, 'Number of partners in last 12 months' is the number of sexual partners the respondent reported to have had in the last 12 months for respondents who had at least one partner. 'Extramarital sex last intercourse' is a binary variable equal to 1 if a married/cohabiting respondent reported that the last sexual intercourse was not with spouse/cohabiting partner. 'Risky sex' is an indicator equal to 1 if a non-married respondent reported that a condom was not used the last time he or she had sexual intercourse, or a married respondent reported that the last sexual intercourse was not with spouse/cohabiting partner and no condom was used; the indicator equals 0 if a non-married respondent reported using a condom during last intercourse or a married respondent reported not having extramarital sex or having extramarital sex with a condom.
Source: Demographic and Health Surveys of the following countries: Burkina Faso 2010, Burundi 2010, Cambodia 2005, Cameroon 2011, Congo 2013, Ivory Coast 2011, Dominican Republic 2013, Ethiopia 2011, Gabon 2012, Ghana 2014, Guinea 2012, Haiti 2012, India 2005, Kenya 2008–9, Lesotho 2009, Liberia 2013, Malawi 2010, Mali 2012–13, Namibia 2013, Niger 2012, Rwanda 2010, Sao Tome and Principe 2008, Senegal 2010, Sierra Leone 2013, Swaziland 2006–7 , Togo 2013, Zambia 2007, Zimbabwe 2006.

Panel B reports the fraction of individuals testing positive for syphilis and HIV. The prevalence rate of syphilis is equal to 4%, and it is very similar across genders. Nearly 4% of the respondents have tested positive for HIV, 5% among women and about 3% among men. There is a large discrepancy in HIV prevalence across countries, with the lowest rates found in India (0.04%), Cambodia (0.05%), Niger (0.05%) and Senegal (0.07%), and the highest rates found in Swaziland (26%), Lesotho (22%), Zimbabwe (18%) and Zambia (15%). Looking at self-reported data on sexual behaviour (panel C), 15% of the respondents reported having used a condom during their last intercourse, and the average number of sexual partners in the last 12 months (for those who had at least one sexual partner) is approximately 1.15. Among the subsample of married or cohabiting respondents, 2% declared that the last sexual intercourse was not with their spouse/cohabiting partner. Finally, we construct an aggregate indicator of risky sexual behaviour: 'Risky sex' is a binary variable equal to 1 if an unmarried respondent reported that a condom was not used the last time that he or she had sexual intercourse, or a married respondent reported that his or her last sexual intercourse was not with the spouse/cohabiting partner and no condom was used; the variable is equal to 0 if a non-married respondent reported using a condom during last intercourse or a married respondent reported not having had extramarital sex or had extramarital sex with a condom.

In Table 2 we show the share of respondents reporting risky sexual behaviours by STI status to investigate any relationship between self-reported behaviour and the biomarkers data. In panel A, we focus on the sample of respondents interviewed and

TABLE 2
RISKY SEXUAL BEHAVIOURS AND BIOMARKERS

	Obs. (1)	Mean (2)	Obs. (3)	Mean (4)	(5)
Panel A: Curable STIs	Syphilis-positive (P)		Syphilis-negative (N)		*p*-value (P=N)
Condom used last intercourse	88	0.16	1,731	0.17	0.71
Number of partners in last 12 months	88	1.23	1,729	1.13	0.01
Extramarital sex last intercourse	75	0.04	1,302	0.02	0.14
Risky sex	88	0.14	1,718	0.15	0.79
Panel B: Non-curable STIs	HIV-positive (P)		HIV-negative (N)		*p*-value (P=N)
Condom used last intercourse	14,434	0.28	290,384	0.14	0.00
Number of partners in last 12 months	14,305	1.16	284,257	1.15	0.00
Extramarital sex last intercourse	10,352	0.04	236,626	0.02	0.00
Risky sex	14,267	0.16	288,422	0.11	0.00

Notes
The *p*-values tested the null hypothesis that the means between P and N are equal. Variables are defined as in Table 1.
Panel A includes Zambia only. Panel B includes Burkina Faso 2010, Burundi 2010, Cambodia 2005, Cameroon 2011, Congo 2013, Ivory Coast 2011, Dominican Republic 2013, Ethiopia 2011, Gabon 2012, Ghana 2014, Guinea 2012, Haiti 2012, India 2005, Kenya 2008–9, Lesotho 2009, Liberia 2013, Malawi 2010, Mali 2012–13, Namibia 2013, Niger 2012, Rwanda 2010, Sao Tome and Principe 2008, Senegal 2010, Sierra Leone 2013, Swaziland 2006–7, Togo 2013, Zambia 2007, Zimbabwe 2006.
Source: Demographic and Health Surveys.

tested in Zambia, and look at the correlation between the prevalence of syphilis and high-risk behaviour. We note that, on average, the fraction of individuals reporting risky sexual behaviours is higher among those who tested positive for syphilis. In particular, a lower fraction of respondents reported having used a condom in their last intercourse among those who tested positive for syphilis (16%) compared to those who declared having used a condom in their last intercourse among the STI-negative respondents (18%), but this difference is not statistically significant (p-value 0.66). Looking at the number of partners in the last 12 months, we observe a statistically significant difference among the number of partners declared by syphilis-positive respondents (1.07 partners) compared to the number reported by syphilis-negative individuals (0.86 partners). The probability of extramarital sex in the sample of syphilis-positive respondents is also higher: approximately 5% of syphilis-positive married individuals reported that their last sexual intercourse was not with their spouse, compared to about 2% of the STI-negative respondents (p-value 0.02). It is important to note that only 15% of the respondents who tested positive for syphilis reported unsafe behaviours (measured using the aggregate indicator for 'Risky sex'), suggesting that underreporting might be happening. Even more interesting, there is no statistically significant difference between the fractions of individuals reporting risky sexual behaviour in syphilis-positive and syphilis-negative respondents.

We should note that if the horizon over which the risky behaviour is elicited in the survey (e.g. the last 12 months) is longer than the cycle of the STI, then the STI test conducted at the end of the questionnaire may not be able to detect the risky behaviour. If this is the case, then the individual may correctly report a risky activity in the last 12 months but test negative. This happens because exposure may have been too recent for effective detection by the test or because the STI might have spontaneously resolved prior to the test. In the case of syphilis, for example, detection is difficult in the initial asymptomatic weeks after exposure. In subsequent stages, it can be cured with adequate treatment. If someone engaged in risky behaviour in the last 12 months and contracted the disease as a consequence, then it may still go undetected because of the initial latency period or because this person may have detected and treated the disease before the biomarker is collected. This may happen even if the person truthfully self-reports having engaged in risky behaviour. In practice, this will only exacerbate the mismeasurement by biomarkers and it highlights the need to carefully define the horizon over which the risky behaviour is elicited by the survey or revealed by a biomarker.

In panel B of Table 2, we look at the correlation between HIV prevalence and risky sexual behaviours. Once again, on average, the fraction of individuals reporting risky sexual behaviours is higher among those who tested positive for HIV. An exception is in the use of condoms. A higher fraction of HIV-positive individuals declared having used a condom during the last sexual intercourse (28%) compared to the HIV-negative respondents (14%). This may be driven by a higher willingness to reduce HIV transmission among infected individuals.

IV. Results

Comparing biomarkers and self-reported measurements

We next use the DHS data to fix a set of parameters to simulate the theoretical model described in Section I.

TABLE 3
PARAMETERS

Panel A: Curable STIs		
Share of syphilis-infected individuals in population	I	0.04
Probability of transmission from single contact with an infected person	ξ	0.20
Parameter that controls how number of contacts varies with number of partners	η	1
Parameter that controls for degree of sorting between infected and uninfected individuals	ρ	1
Number of sexual partners in last 12 months	p	1.14
Number of sexual partners in last 12 months for single men	p	1.41
Number of sexual partners in last 12 months for married women	p	1.01
Panel B: Non-curable STIs		
Share of HIV-infected individuals in population	I	0.04
Probability of transmission from single contact with an infected person	ξ	0.34
Parameter that controls how number of contacts varies with number of partners	η	1
Parameter that controls for degree of sorting between infected and uninfected individuals	ρ	1
Number of sexual partners in last 12 months	p	1.15
Number of sexual partners in last 12 months for single men	p	1.55
Number of sexual partners in last 12 months for married women	p	1.01

Notes
Samples as in Table 2.
Source: Demographic and Health Surveys for all parameters except ξ, which comes from Nelson and Masters Williams (2007, p. 978).

More precisely, we simulate λ, the probability of correct classification of risky sexual behaviours using biomarkers, with the parameter values reported in Table 3. In panel A we report the parameters used to estimate the probability of correct classification of risky behaviours using biomarkers for curable STIs, and in panel B we show the parameters for simulating the probability of correct classification using biomarkers for non-curable STIs. Recall that ξ is the probability of an STI being transmitted from an infected to an uninfected individual during a given unprotected sexual contact. Following Hyman *et al.* (2001), we set η, the parameter that controls for how the number of contacts varies with the number of partners, equal to 1. Finally, we also set ρ equal to 1, allowing for a person to meet every individual with an equal and independent probability regardless of their infection status. However, we show in Figures A1 and A2 of the Appendix how our results change by setting $\rho > 1$ and $\rho < 1$.

We then apply Proposition 1 and compare the simulated value of λ with the self-reported risky behaviours, measured by the aggregate variable 'Risky sex' described in Table 1.

Figure 1(a) shows the theoretical combinations of I—the curable STI prevalence rate in a given country in a given year—and λ—the probability of correct classification of risky sexual behaviours using biomarkers for curable STIs. The vertical dashed line indicates the value of syphilis prevalence, equal to 0.043, used to simulate λ. When $I = 0.043$, $p = 1.14$, $\rho = 1$ and $\eta = 1$, the simulated λ is equal to 0.047. From Figure 1(a), a positive relationship emerges between the probability of correctly measuring risky behaviours with biomarkers (λ) and the prevalence of STIs in a given country. Hence countries with a high prevalence of STIs are more likely to get a good proxy of risky sexual behaviours using biomarkers compared to countries where the prevalence is low. The intuition behind this result is very simple: if an STI is common in the population,

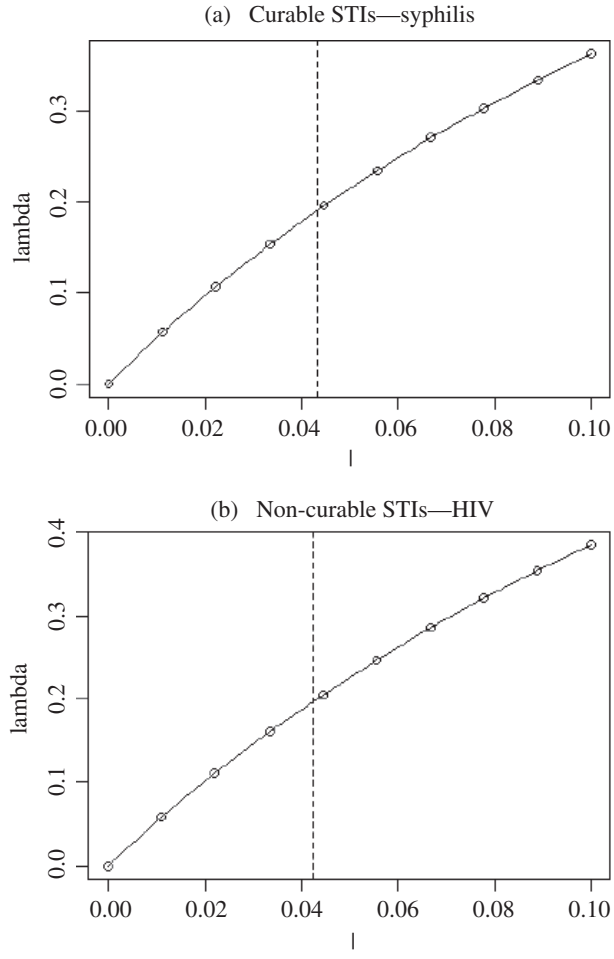

FIGURE 1. STI prevalence and probability of correct classification using biomarkers.
Notes: λ is the probability of correct classification of risky sexual behaviours using biomarkers for STIs, and I is syphilis prevalence in panel (a) and HIV prevalence in panel (b). In both panels, $\rho = 1$. Panel (a) includes the sample of respondents in Zambia. Panel (b) includes the sample of respondents in Burkina Faso 2010, Burundi 2010, Cambodia 2005, Cameroon 2011, Congo 2013, Ivory Coast 2011, Dominican Republic 2013, Ethiopia 2011, Gabon 2012, Ghana 2014, Guinea 2012, Haiti 2012, India 2005, Kenya 2008–9, Lesotho 2009, Liberia 2013, Malawi 2010, Mali 2012–13, Namibia 2013, Niger 2012, Rwanda 2010, Sao Tome and Principe 2008, Senegal 2010, Sierra Leone 2013, Swaziland 2006–7, Togo 2013, Zambia 2007, Zimbabwe 2006.
Source: Demographic and Health Surveys.

then the probability that an individual who engaged in risky sexual intercourse will be infected is higher, compared to settings where the same STI is less common. Thus if the probability of infection is higher, then the probability of detecting the infection with biomarkers is also higher. The picture looks very similar when we consider non-curable STIs (Figure 1(b)). The higher the prevalence of HIV, the higher the probability of correct classification using this biomarker's incidence rate λ.

Given that the number of partners used to estimate λ is self-reported, in Figure 2 we plot the relationship between λ and I for two different values of p. In particular, we consider the number of partners declared by married women, who are more likely to

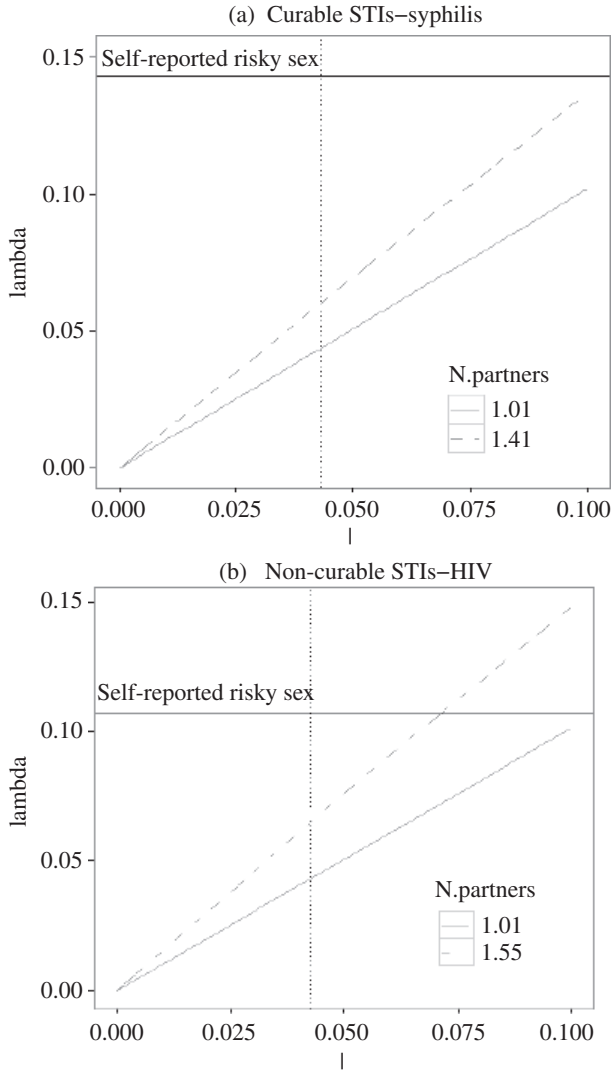

FIGURE 2. Comparison between biomarkers and self-reported risky behaviour for lower and upper bounds on the number of partners.
Notes: See Figure 1. 'Self-reported risky sex' is an indicator equal to 1 if a non-married respondent reported that a condom was not used the last time he or she had sexual intercourse, or a married respondent reported that the last sexual intercourse was not with spouse/cohabiting partner and no condom was used; the indicator equals 0 if a non-married respondent reported using a condom during last intercourse or a married respondent reported not having extramarital sex or having extramarital sex with a condom.
Source: Demographic and Health Surveys.

underreport them, as a lower bound for p, and the number of partners mentioned by single men, who are more likely to overreport them, as an upper bound for p (Oster 2005).[20] In the sample of respondents interviewed in Zambia (Figure 2(a)), the average number of partners in the last 12 months for single men is 1.41, while the average number of partners for married women is 1.01. In the pooled sample of countries for which HIV prevalence is available (Figure 2(b)), the average number of partners in the last 12

months for single men is 1.55, while the average number of partners for married women is 1.01. In Table A3 of the Appendix, we provide the average number of partners by country, splitting the sample between married women and single men. The number of partners reported by married women is consistently lower than that reported by single men. This discrepancy in self-reported risky behaviour among married women and single men is also well-documented in the previous literature (see, for example, Wellings *et al.* 2006; Nnko *et al.* 2004; Oster 2005).

We then compare the estimated value of λ with the horizontal lines in the graphs of Figure 2, indicating the fraction of people reporting risky sexual behaviours in the DHS ($\mathbb{P}(Y^e = 1)$), using the aggregate measures of risky sex reported in Table 1. We show that as the number of reported sexual partners increases for a given I, the probability of correct classification using biomarkers also increases. Thus in countries with a low prevalence of STIs and low average number of sexual partners per individual, λ is lower than $\mathbb{P}(Y^e = 1)$, and the biomarker has a higher probability of misclassification of risky behaviours than behaviours elicited by a survey questionnaire (as mentioned in Proposition 1). In other words, the intersection between the horizontal line indicating self-reported behaviour and the upward sloping curve describing the relationship between λ and I provides the threshold above which biomarkers are a better proxy for risky sexual behaviours. For values of I above this threshold, biomarkers have a smaller probability of misclassification compared to self-reported data. Figure 2(a) shows that in the case of Zambia, where syphilis prevalence is 4.3%, if the person meets every individual with equal and independent probability regardless of their infection status ($\rho = 1$), then biomarkers have a lower probability of correctly measuring risky sexual behaviours than self-reported responses. Positive assortative matching ($\rho < 1$), which is a plausible scenario,[21] leads to an even more favourable picture for elicited behaviour. We report the graphs for different values of ρ in Figures A1 and A2 of the Appendix.

These results can be an important tool to enable researchers and policymakers to find the best measure of risky behaviours before designing interventions or programmes aimed at targeting individuals at highest risk of HIV infection in a given country.

Combining biomarkers and self-reported measurements

Table 4 compares the parameters on the determinants of risky sexual behaviours using as dependent variable self-reported indicators for risky sex (columns (1)–(3)), the biomarker for syphilis (columns (4)–(6)), and the probability of actual risky sexual behaviour estimated using the GMM procedure described in Section II (columns (7)–(9)). For each dependent variable, we estimate a linear probability model, a probit model and a logit model.

Looking at columns (7)–(9) of Table 4, the probability of eliciting truthful answers among those who engage in risky behaviour (α) is estimated at 14.3%. We note that even in cases when this rate is less favourable compared to the probability of correct classification by the biomarkers, the combination of both sources allows one to more accurately estimate the correlates of risky sexual behaviour.

By comparing the determinants of risky sexual behaviour using self-reported information, biomarkers, and the combination of the two measures, we note a huge discrepancy in terms of both the magnitude and sign of the coefficients. For example, being a woman is negatively and significantly correlated with the probability of reporting risky sex (columns (1)–(3)), but is positively associated with the probability of testing positive for syphilis (although the coefficient is not statistically significant) (columns (4)–

TABLE 4
CORRELATES OF SELF-REPORTED RISKY BEHAVIOURS, CURABLE STIS AND ACTUAL RISKY BEHAVIOUR

	1 if self-reported risky sexual behaviours			1 if STI-positive (syphilis)			GMM 1 if actual risky behaviour		
	OLS	Probit	Logit	OLS	Probit	Logit	OLS	Probit	Logit
Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
α: probability of correct elicitation among those that engage in risky behaviour							0.136***	0.136***	0.136***
							[0.040]	[0.040]	[0.040]
Female	−0.057***	−0.209***	−0.437***	0.003	0.041	0.080	−0.460**	0.214	1.232
	[0.008]	[0.038]	[0.068]	[0.008]	[0.092]	[0.204]	[0.170]	[1.780]	[5.800]
Age	−0.010***	−0.051***	−0.104***	0.001**	0.008**	0.016**	−0.073***	−0.135**	−0.305
	[0.000]	[0.003]	[0.006]	[0.000]	[0.004]	[0.008]	[0.020]	[0.050]	[0.170]
Primary education	0.001	−0.026	0.023	−0.001	−0.010	−0.021	−0.339	−3.437	−6.359
	[0.013]	[0.075]	[0.140]	[0.017]	[0.178]	[0.394]	[0.240]	[2.101]	[5.320]
Secondary education and above	0.006	0.003	0.081	−0.002	−0.014	−0.039	0.147	−2.777	−4.960
	[0.015]	[0.079]	[0.146]	[0.018]	[0.188]	[0.412]	[0.250]	[2.000]	[4.820]
Urban	0.019**	0.093**	0.153**	0.007	0.074	0.163	0.221	0.641	1.597
	[0.009]	[0.040]	[0.072]	[0.010]	[0.103]	[0.228]	[0.151]	[0.531]	[1.340]
Constant	0.474***	0.484***	1.277***	0.019	−2.000***	−3.688***	3.752***	8.009***	16.255***
	[0.021]	[0.106]	[0.199]	[0.020]	[0.221]	[0.482]	[1.040]	[1.470]	[4.250]
Observations	8168	8168	8168	2392	2392	2392	1806	1806	1806

Notes
The dependent variable in columns (1)–(3) is described in Table 1.
***, **, * indicate $p<0.01$, $p<0.05$, $p<0.10$, respectively.
Source: Zambia Demographic and Health Survey 2007.

(6)). Looking at the GMM specifications in columns (7)–(9), the coefficient on female is negatively and significantly correlated with the dependent variable, but it is statistically significant only in the linear specification. Age is negatively and statistically significantly correlated with the probability of reporting risky sexual behaviour (columns (1)–(3)) and with the likelihood of true risky sexual behaviours (columns (7)–(9)), suggesting that older people are less likely to behave in a risky manner. But on the other hand, the coefficient on age turns out to be positive when looking at the probability of testing positive for syphilis. Both primary and secondary education are negatively correlated with the dependent variables in all the specifications (although statistically significant only on the probability of reporting risky sex), suggesting that more educated individuals are less likely to engage in risky behaviour. The coefficients on urban areas are not statistically significant. As expected, the magnitude of the coefficients estimated with the GMM is larger compared to those estimated in columns (1)–(6).

In Table 5, we do a similar exercise by using the sample of countries for which we have information on non-curable STIs. In particular, we compare the parameters on the determinants of high-risk sexual behaviours using as dependent variable self-reported indicators for risky sex (columns (1)–(3)), the biomarker for HIV (columns (4)–(6)), and the probability of true risky sexual behaviour estimated using the GMM procedure described in Section II (columns (7)–(9)). For each dependent variable, we estimate a linear probability model, a probit model and a logit model. Once again, we note a huge discrepancy in the importance of the determinants of risky behaviour using self-reported data or biomarkers. In columns (7)–(9) we show that the probability of eliciting truthful answers among those who engage in risky behaviour (α) is estimated at 15.5%.

The main point to be taken from Tables 4 and 5 is therefore that different measures of high-risk health behaviours provide completely different results on the determinants of such behaviours.

Using the GMM coefficients estimated in the previous tables, in Table 6 we compute the probabilities of true risky sexual behaviour for individuals with different sociodemographic characteristics. For example, a 30-year-old man with a secondary education who lives in an urban area has an 86% chance of adopting risky sexual behaviours. On the other hand, a 40-year-old woman with a primary education living in a rural area has only a 20% probability of behaving in a risky manner.

Table 7 shows the distribution of the predicted probability of self-reported risky behaviour, of biomarkers and of unobserved actual risky behaviour. The predicted probabilities have been computed with a probit regression using as dependent variables our aggregate measure of 'Risky sex' (column (1)), biomarkers for STIs (column (2)) and actual risky behaviour as estimated from the GMM specification described in equation (8) (column (3)), and 'Female', 'Age', 'Primary education', 'Secondary education and above' and 'Urban' as covariates. Panel A includes the sample of individuals tested for syphilis, and panel B includes individuals tested for HIV.

Looking at panel A of Table 7, we note that with an estimated probability of self-reported risky behaviour of 17.4% and a predicted probability of testing positive for syphilis of 4.3%, our GMM strategy estimates the probability of actual risky behaviour to be 78.7%. Hence the probability of true risky behaviour is higher, but much more dispersed (the standard deviation is very high and equal to 0.303) than the predicted probability using elicited risky behaviour and biomarkers. In addition, the predictions here do not take into account the statistical uncertainty from the estimation of the coefficients in the model.

TABLE 5
CORRELATES OF SELF-REPORTED RISKY SEXUAL BEHAVIOURS, NON-CURABLE STIS AND ACTUAL RISKY BEHAVIOUR

	1 if self-reported risky sexual behaviours			1 if STI-positive (HIV)			GMM 1 if actual risky behaviour		
Dependent variable	OLS (1)	Probit (2)	Logit (3)	OLS (4)	Probit (5)	Logit (6)	OLS (7)	Probit (8)	Logit (9)
α: probability of correct elicitation among those that engage in risky behaviour							0.155*** [0.000]	0.155*** [0.000]	0.155*** [0.000]
Female	−0.045*** [0.001]	−0.216*** [0.007]	−0.427*** [0.013]	0.027*** [0.001]	0.317*** [0.007]	0.700*** [0.016]	−0.288*** [0.010]	−0.552*** [0.030]	−0.985*** [0.050]
Age	−0.007*** [0.000]	−0.040*** [0.000]	−0.083*** [0.001]	0.002*** [0.000]	0.020*** [0.000]	0.042*** [0.001]	−0.044*** [0.000]	−0.117*** [0.000]	−0.205*** [0.000]
Primary education	0.040*** [0.001]	0.278*** [0.009]	0.537*** [0.018]	0.045*** [0.001]	0.549*** [0.011]	1.234*** [0.025]	0.261*** [0.010]	0.830*** [0.030]	1.448*** [0.060]
Secondary education and above	0.036*** [0.001]	0.251*** [0.009]	0.496*** [0.018]	0.039*** [0.001]	0.502*** [0.011]	1.141*** [0.026]	0.235*** [0.010]	0.763*** [0.030]	1.309*** [0.050]
Urban	0.011*** [0.001]	0.069*** [0.007]	0.123*** [0.013]	−0.002*** [0.001]	−0.020*** [0.008]	−0.057*** [0.017]	0.068*** [0.010]	0.379*** [0.030]	0.689*** [0.050]
Constant	0.320*** [0.003]	−0.160*** [0.016]	0.126*** [0.033]	−0.055*** [0.001]	−2.952*** [0.016]	−5.788*** [0.035]	2.069*** [0.050]	3.468*** [0.070]	6.004*** [0.130]
Observations	302,689	302,689	302,689	435,375	435,375	435,375	302,689	302,689	302,689

Notes
The dependent variable in columns (1)–(3) is described in Table 1.
***, **, * indicate $p<0.01$, $p<0.05$, $p<0.10$, respectively.
Source: Demographic and Health Surveys for Burkina Faso 2010, Burundi 2010, Cambodia 2005, Cameroon 2011, Congo 2013, Ivory Coast 2011, Dominican Republic 2013, Ethiopia 2011, Gabon 2012, Ghana 2014, Guinea 2012, Haiti 2012, India 2005, Kenya 2008–9, Lesotho 2009, Liberia 2013, Malawi 2010, Mali 2012–13, Namibia 2013, Niger 2012, Rwanda 2010, Sao Tome and Principe 2008, Senegal 2010, Sierra Leone 2013, Swaziland 2006–7 , Togo 2013, Zambia 2007, Zimbabwe 2006.

TABLE 6
PROBABILITY OF ACTUAL RISKY BEHAVIOUR, BY SOCIODEMOGRAPHIC CHARACTERISTICS

Female		Age		Primary education		Secondary education and above		Urban		Probability of actual risky behaviour
0	1	30	40	0	1	0	1	0	1	
Panel A: Sample of curable STIs										
×		×		×			×		×	88%
×			×	×			×		×	68%
×			×		×	×		×		20%
	×	×		×			×		×	98%
	×		×	×			×		×	76%
	×		×		×	×		×		27%
Panel B: Sample of non-curable STIs										
×		×		×			×		×	86%
×			×	×			×		×	47%
×			×		×	×		×		35%
	×	×		×			×		×	71%
	×		×	×			×		×	27%
	×	×			×	×		×		59%

Notes
The probability of actual risky behaviour reported in the final column, for panels A and B, is based on the probit estimated coefficients reported in Tables 4 and 5, respectively.

Similarly, in panel B of Table 7, 13.4% is the probability of self-reported risky behaviour, 4.2% is the probability of being HIV-positive, and 83.3% is the probability of engaging in truly risky behaviour. Underreporting of sex-related behaviours is a well-documented phenomenon in both the medical and economic literatures (see, for example, Fenton *et al*. 2001; Ozler 2013; Gersovitz *et al*. 1998; de Paula *et al*. 2014; de Walque 2007). People are understandably reluctant to admit personal transgressions. Anedoctal evidence from the USA in 2006 suggests that the probability of extramarital sex measured through anonymous telephone polls is about 18 times higher than the same probability measured through self-reported questions.[22,23] The sizeable discrepancy in the fraction of high-risk individuals measured in three different ways suggests that social scientists need to think carefully about the best proxy for risky sexual behaviour in the context of their study.

V. CONCLUSIONS

The conventional wisdom is that biomarkers for high-risk health behaviours are a superior measure to self-reported data. In this paper we challenge this notion in the context of risky sexual activities, the main cause of the spread of HIV/AIDS.

We build an epidemiological model to show that, as happens with self-reported data, misclassification of risky sexual behaviours is also possible when using biomarkers for sexually transmitted infection (STIs). We then suggest an econometric framework by proposing a new GMM estimator to precisely estimate correlates of risky sexual behaviours and unobserved actual risky behaviour, by combining biomarkers and self-reported data.

TABLE 7
DISTRIBUTION OF THE PREDICTED PROBABILITY OF RISKY BEHAVIOUR

	Self-reported sexual behaviour (1)	Biomarkers (2)	Actual risky behaviour (3)
Panel A: Curable STI (syphilis)			
Mean	0.174	0.043	0.787
S.D.	0.112	0.008	0.303
25%	0.078	0.037	0.668
50%	0.160	0.041	0.960
75%	0.264	0.047	0.997
Panel B: Non-curable STI (HIV)			
Mean	0.134	0.042	0.838
S.D.	0.086	0.027	0.277
25%	0.059	0.021	0.820
50%	0.122	0.037	0.988
75%	0.203	0.055	0.999

Notes
See Table 1 for the definitions of the variables.
The predicted probabilities in column (1) have been computed with a probit regression using as dependent variable 'Risky sex'. The predicted probability in column (2) uses as dependent variable a dummy equal to 1 if the respondent tested STI-positive. In column (3), actual risky behaviour has been estimated from the GMM specification described in equation (8) using 'Female', 'Age', 'Primary education', 'Secondary education', 'Urban' as covariates.
Samples as in Table 2.
Source: Demographic and Health Surveys.

Using the most recent Demographic and Health Survey for all countries with publicly available biomarkers and self-reported data on sexual activity, we calibrate the model and we find that in countries with a low prevalence of STIs and a low average number of sexual partners per individual, the biomarkers have a higher probability of misclassification than behaviours elicited by a survey questionnaire. We then apply our econometric framework to the DHS data and show a huge discrepancy in the prevalence of risky activities using three different measures: self-reported data, biomarkers, and the combination of the two with our proposed GMM. In particular, our GMM strategy estimates the probability of actual risky behaviour to be 78.7%, while the estimated probability of self-reported risky behaviour is 17.4% and the predicted probability of testing positive for an STI is 4.3%.

Our results and econometric framework have important implications for policy, especially given the growing number of studies on the HIV/AIDS epidemic which rely on STIs to infer risky sexual behaviour. First, they provide insights to policymakers and researchers on the most accurate measure of high-risk behaviours—biomarkers or self-reported data. Second, they have the potential to help in estimating the actual prevalence of unobserved risky behaviours in a population, a crucial step to design efficient programmes targeted to individuals with the highest health risks. We also believe that our findings open up new avenues for future research on risky-behaviour measurement.

APPENDIX

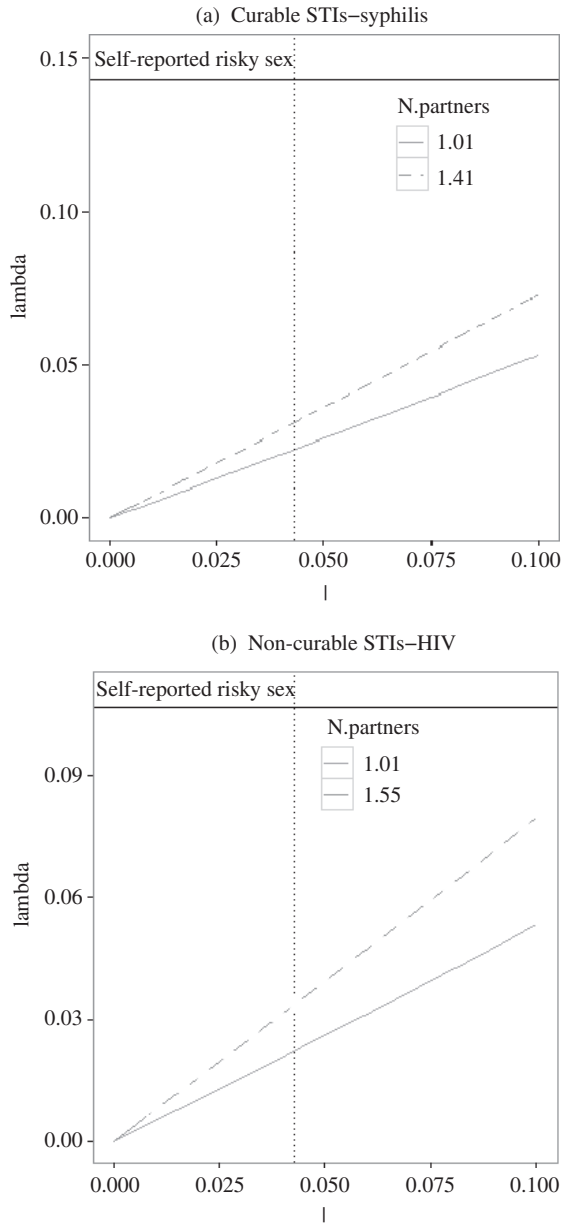

FIGURE A1. Comparison between biomarkers and self-reported risky behaviour, by number of partners ($\rho < 1$).
Notes: λ is the probability of correct classification of risky sexual behaviours using biomarkers for STIs, and I is syphilis prevalence in panel (a) and HIV prevalence in panel (b). In both panels, $\rho = 0.5$. See Figure 2 for the definition of 'Self-reported risky sex'. See Figure 1 for samples.
Source: Demographic and Health Surveys.

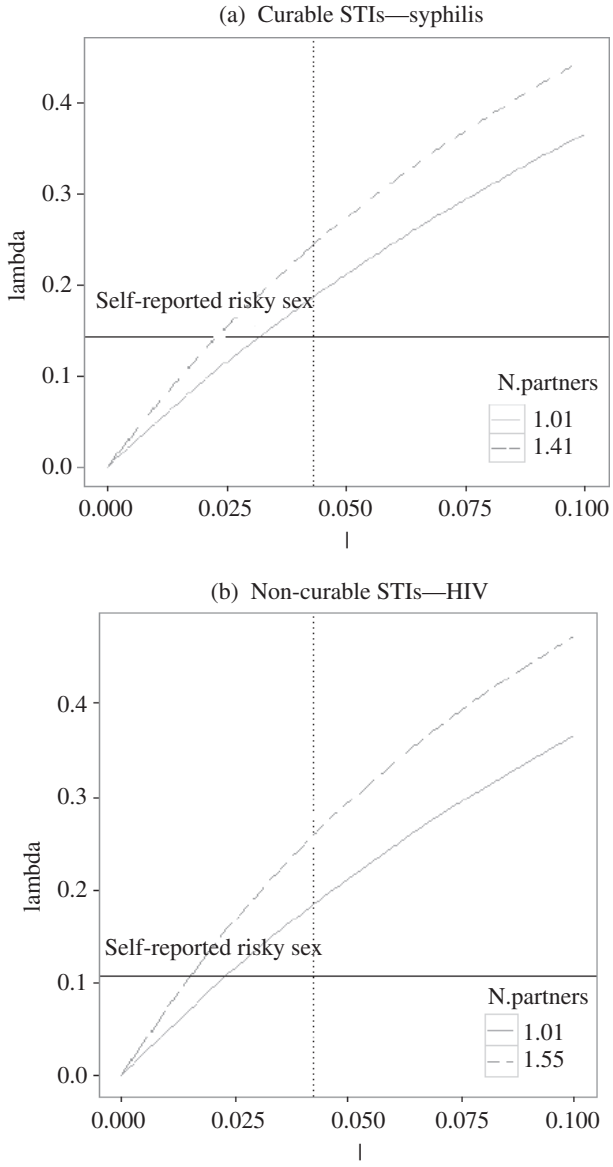

FIGURE A2. Comparison between biomarkers and self-reported risky behaviour, by number of partners ($\rho > 1$).
Notes: See Figure A1.
Source: Demographic and Health Surveys.

TABLE A1
FREQUENCY OF SEXUAL CONTACTS

Sexual contacts	Frequency	%
More than 3 per week	20	9.57
2 per week	46	22.01
2 per month	73	34.93
< 2 per month	63	30.14
Missing	7	3.35

Notes
Sample of unmarried women.
Source: Malawi Diffusion and Ideational Change Project (MDICP).

TABLE A2
COUNTRY LIST AND SHARES OF NON-MISSING DATA

Country	Waves	Number of respondents	% tested for HIV	% tested for curable STIs	% replied to condom used	% replied to number of partners	% replied to extramarital sex
Burkina Faso	2010	24,394	0.63	—	0.45	0.63	0.63
Burundi	2010	13,669	0.63	—	0.38	0.63	0.63
Cambodia	2005	23,554	0.62	—	0.39	0.62	0.62
Cameroon	2011	22,617	0.63	—	0.48	0.63	0.63
Congo	2013	27,483	0.64	—	0.51	0.64	0.64
Ivory Coast	2008	15,195	0.59	—	0.46	0.59	0.59
Dominican Republic	2011	19678	0.95	—	0.76	0.94	0.94
Ethiopia	2012	30,625	0.93	—	0.59	0.93	0.93
Gabon	2003	14,076	0.78	—	0.66	0.78	0.78
Ghana	2014	13,784	0.64	—	0.45	0.64	0.33
Guinea	2012	12,924	0.65	—	0.44	0.65	0.65
Haiti	2012	23,780	0.78	—	0.58	0.78	0.78
India	2005	198,754	0.52	—	0.33	0.52	0.52
Kenya	2008	11,909	0.58	—	0.42	0.58	0.58
Lesotho	2009	10,941	0.63	—	0.47	0.62	0.62
Liberia	2013	13,357	0.61	—	0.51	0.61	0.61
Malawi	2010	30,195	0.46	—	0.34	0.46	0.46
Mali	2012	14,823	0.60	—	0.43	0.54	0.54
Namibia	2013	14,499	0.61	—	0.43	0.61	0.61
Niger	2013	15,088	0.57	—	0.45	0.57	0.57
Rwanda	2010	20,000	0.66	—	0.38	0.66	0.66
Sao Tome and Principe	2008	4,911	0.96	—	0.76	0.95	0.95
Senegal	2010	20,617	0.48	—	0.30	0.48	0.48
Sierra Leone	2013	23,920	0.61	—	0.48	0.61	0.61
Swaziland	2006	9,143	0.90	—	0.58	0.89	0.89

TABLE A2
CONTINUED

Country	Waves	Number of respondents	% tested for HIV	% tested for curable STIs	% replied to condom used	% replied to number of partners	% replied to extramarital sex
Togo	2013	13,956	0.66	—	0.48	0.66	0.66
Zambia	2007	13,646	0.80	0.18	0.60	0.80	0.80
Zimbabwe	2006	32,164	0.81	—	0.53	0.81	0.81
Total		689,702	0.63	—	0.44	0.63	0.36

Notes
See Table 1 for the definitions of the variables.
Source: See Table 1.

TABLE A3
NUMBER OF SELF-REPORTED PARTNERS IN THE LAST 4 MONTHS

	All	Married women	Single men
Burkina Faso	1.14	1.00	1.26
Burundi	1.14	1.00	1.40
Cambodia	1.11	1.00	2.61
Cameroon	1.41	1.05	1.93
Congo	1.26	1.03	1.56
Ivory Coast	1.33	1.01	1.87
Domican Republic	1.44	1.03	2.09
Ethiopia	1.06	1.01	1.21
Gabon	1.35	1.09	1.58
Ghana	1.13	1.00	1.34
Guinea	1.19	1.03	1.34
Haiti	1.34	1.02	1.79
India	1.02	1.00	1.40
Kenya	1.11	1.01	1.33
Lesotho	1.21	1.09	1.44
Liberia	1.21	1.05	1.44
Malawi	1.08	1.01	1.22
Mali	1.10	1.01	1.26
Namibia	1.09	1.02	1.22
Niger	1.09	1.00	1.31
Rwanda	1.05	1.00	1.25
SaoTome and Principe	1.12	1.01	1.22
Senegal	1.13	1.00	1.27
Sierra Leone	1.24	1.06	1.43
Swaziland	1.13	1.02	1.37
Togo	1.15	1.00	1.31
Zambia	1.15	1.01	1.41
Zimbabwe	1.08	1.00	1.27
Average	1.15	1.01	1.55

Source: See Table 1.

ACKNOWLEDGMENTS

We thank Orazio Attanasio, Richard Blundell, Erick Gong, Bo Honoré and Imran Rasul, and seminar participants at the Institute of Fiscal Studies, the London School of Hygiene and Tropical Medicine and the AIES conference in Venice, for useful comments. De Paula acknowledges financial support from the National Science Foundation through award SES-1123990, the European Research Council through Starting Grant 338187, and the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001.

NOTES

1. Recent estimates indicate that approximately 1.1 million individuals worldwide died from HIV/AIDS in 2015 (UNAIDS 2016), and 270,000 women died from cervical cancer due to papillomavirus in 2012 (WHO 2018).
2. The issue of social desirability bias in self-reported answers has been discussed in settings other than health. For example, Baird and Ozler (2010), using data from a randomized cash transfer programme in Malawi,

compare self-reported data on school attendance with administrative records. They show that participants significantly overstate their school participation, and this overreporting is higher in the control group, thus producing biased impact estimates. In a more recent paper, Karlan and Zinman (2012) test the validity of self-reported data on loan expenditure for consumption or for investment purposes. They find that respondents were more likely to admit using their loan for household items and medical/educational expenses on an anonymous survey than they were in response to direct questioning.

3. To address the issue of misreporting risky behaviours during face-to-face interviews, audio computer-assisted self-interviewing techniques have been used to collect data in the USA (Tourangeau and Smith 1996; Hewitt 2002). Results from recent randomized evaluations show that indeed, there is a higher probability of reporting risky behaviours with a computerized mode of interviews than with face-to-face interviews (Hewett *et al*. 2008).
4. The discrepancies between biomarkers and self-reported answers have been tested in other domains. For example, Connor *et al*. (2009), using a meta-analysis of 67 studies, shows the divergence between self-reported smoking status and smoking status determined through measures of cotinine in biological fluids. Subar *et al*. (2003) assess dietary measurement error using self-reported dietary instruments and unbiased biomarkers of energy and protein intakes. Hellhammer *et al*. (2009) attempted to objectively test the level of psychological stress through salivary cortisol.
5. In experimental studies, the use of biomarkers on STIs as a marker for risky sexual behaviour is sometimes motivated by the fact that data on HIV incidence may be expensive to collect, given the sample size required to detect any effect of HIV/AIDS prevention interventions (Fishbein and Pequegnat 2000).
6. Recent studies on modes of transmission have documented new infections among high-risk populations to be an important component of the national prevalence. For example, globally, female sex workers are 13.5 times more likely to be living with HIV than other women, and a substantial proportion of new infections (10–30%) are estimated to occur as a result of sex work in Uganda, Swaziland and Zambia (UNAIDS 2013).
7. Although we are aware that unsafe sex is a multidimensional phenomenon, biological markers can record only whether or not an individual adopted a risky sexual behaviour. Hence, in our comparison between biological markers and elicited measures of risky behaviours, it seems adequate to encode those into a binary variable.
8. Among highly mobile populations such as migrants, who are above all considered a highly risky population for STI/HIV transmission, it is particularly challenging to collect both baseline and endline data (Corno and de Walque 2012).
9. If a non-risky sexual behaviour (e.g. the proper use of a condom during sexual intercourse) significantly reduces the probability of transmission of the STI but nonetheless still allows it, another type of misclassification would arise: being tagged as risky when the behaviour is non-risky. This would be the case, for example, with genital herpes, which can be transmitted when outbreaks occur in areas not protected by the condom but that still come into contact during the sexual act (Centers for Disease Control and Prevention 2014). Another example could be if a person contracts an STI from his or her spouse even if that person is faithful but the spouse is not.
10. Note that STI transmission can also happen through channels other than sexual intercourse (e.g. through the share of infected tools). However, data from WHO shows that in 2004, unsafe sex was estimated as being responsible for more than 99% of HIV infection in Africa. Elsewhere, the proportion of HIV/AIDS deaths due to unsafe sex ranges from around 50% in the low- and middle-income countries of the WHO Western Pacific Region to 90% in the low- and middle-income countries of the Americas.
11. This point is important given that most of the standard national representative surveys eliciting information on sexual behaviours, such as the Demographic and Health Surveys (DHS), do not include questions on the number of instances of sexual intercourse per partner or per month.
12. For simplicity, we assume that the rate of transmission is homogeneous across different stages of the STI. If transmission rates vary according to the stage of the infection, then stage-specific transmission rates could be employed and equation (2) would have $\beta(p)$ as a mixture over the proportion of individuals at each stage. This, of course, would involve an extra layer of calibrations for stage-specific transmission rates and the proportion of infected people at each stage.
13. Since our focus is on the misclassification of risky sexual behaviour for uninfected individuals, we abstract from the equilibrium characterization of the matching process.
14. STI testing today is quite accurate when it is guided by proper clinical information. However, no test is going to be always accurate all of the time. The probability of false negatives/positives may depend on the type of testing: whether it is conducted in the laboratory or with rapid tests, and by the type of rapid test itself. For example, Ripa and Nilsson (2007) investigate the reliability of different types of tests for chlamydia in Sweden, and find different prevalences depending on the test used. Repeat testing may therefore occasionally be needed. Björkman Nyqvist *et al*. (2015) conducted three tests to detect chlamydia and trichomonas vaginalis among patients in Lesotho as follows. If the first test result is negative, then the individual is considered STI-negative; if the individual tests positive, then the protocol requires a second confirmatory test. If a second test is positive, then the individual is considered STI-positive; if, however, the second test is negative after an initial positive result, then a third test is done for verification.

15. Note that Proposition 1 holds even when individuals overreport the number of partners, since in this case the calibrated λ will be larger than the one computed using the actual number of partners.
16. The most common treatable STIs are chancroid, chlamydia, crabs, gonorrhea, scabies, syphilis, trichomoniasis, yeast infection, vaginosis; the most common incurable STIs are hepatitis B, C and D, HPV, HSV2, HIV (Centers for Disease Control and Prevention 2014).
17. If misreporting of non-risky behaviour is possible, then the expression in (7) becomes $\mathbb{P}(Y^e = 1|\mathbf{X}) = \beta + (\alpha - \beta)F(\mathbf{X};\theta)$, where $\beta = \mathbb{P}(Y^e = 1|Y^t = 0)$. As pointed out previously, this can be accommodated in the estimation strategy described below by having β as an additional parameter to be estimated.
18. To properly distinguish covariates affecting the probability of misreporting and covariates directly affecting the probability of risky behaviour, the sets of covariates affecting one and the other should not completely coincide.
19. Problems like attrition and non-response are a separate issue from the one on which we focus in this paper, and may exist with or without misreporting. Those problems can sometimes be dealt with using selection methods (see, for example, Hausman and Wise 1979), and we believe that these remedies could also be employed in conjunction with our proposed strategy.
20. As indicated in Section I, because λ is overestimated when individuals overreport the number of sexual partners and underestimated when individuals underreport them, these bounds would be wider if those reports were not truthful.
21. For example, Dow and Philipson (1993) find that HIV-positive individuals in San Francisco are twice as likely to have an HIV-positive partner than an HIV-negative one.
22. Tom Smith, director of the General Social Survey (GSS) at the National Opinion Research Center (NORC) at the University of Chicago, mentioned that the best estimates in the USA in 2006 are that about 3–4% of currently married people have a sexual partner besides their spouse in a given year. GSS data are collected using self-reported questions. A new poll from Gallup—a famous data collection agency in the USA—in 2008 instead found a much higher figure, of around 55%. The difference between NORC and Gallup is that the second used telephone polls where respondents are less exposed to interviewers and are likely more comfortable at reporting on sensitive topics. This is, of course, not rigorous evidence, but if one believes the Gallup poll to be closer to actual behaviour, then this is about 18 times higher than the self-reported behaviour estimates collected in the NORC survey (Tourangeau and Smith 1996).
23. Also note that when we focus on one risky behaviour only, instead of looking at our aggregate measure 'Risky sex', the gap between the probability of self-reporting behaviour and actual behaviour may change. For example, the self-reported probability of not using a condom during the last sexual intercourse is equal to 82%, compared to the actual probability of not using a condom equal to 99% in the sample of individuals tested for HIV.

REFERENCES

Anderson, C., Gallo, M., Hylton-Kong, T., Steiner, M. J., et al. (2013). Randomized controlled trial on the effectiveness of counseling messages for avoiding unprotected sexual intercourse during sexually transmitted infection and reproductive tract infection treatment among female sexually transmitted infection clinic patients. *Sexually Transmitted Diseases*, **40**(2), 105–10.

Anderson, R. and May, R. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press.

Baird, S. and Ozler, B. (2010). Examining the reliability of self-reported data on school participation. *Journal of Development Economics*, **98**(1), 89–93.

———, Gong, E., McIntosh, C. and Ozler, B. (2014). The heterogeneous effects of HIV testing. *Journal of Health Economics*, **37**, 98–112.

Björkman Nyqvist, M., Corno, L., de Walque, D. and Svensson J.(2015). Using lotteries to incentivize safer sexual behavior: evidence from a randomized controlled trial on HIV prevention. Policy Research Working Paper no. 7215, World Bank.

Carrieri, V. and Jones, A. (2015). The income–health relationship beyond the mean: new evidence from biomarkers. Health Econometrics and Data Group Working Paper.

Centers for Disease Control and Prevention (2014). Genital herpes—CDC fact sheet; avaialbale online at www.cdc.gov/std/herpes/stdfact-herpes.htm (accessed 23 June 2018).

Choudhry, V., Agardh, A., Stafstrom, M. and Ostergren, P. (2014). Patterns of alcohol consumption and risky sexual behavior: a cross-sectional study among Ugandan university students. *BMC Public Health*, **14**(128).

Cleland, J., Boerma, M. and Weir, S. (2004). Monitoring sexual behaviour in general populations: a synthesis of lessons of the past decade. *Sexually Transmitted Infections*, **80**, 84–92.

Connor, J., Kydd, R. and Dickson, N. (2015). Alcohol involvement and adverse sexual health outcomes from 26 to 38 years of age. *Plos ONE*, **10**(8) ; available online at https://doi.org/10.1371/journal.pone.0135660 (accessed 23 June 2018).

Connor Gorber, S., Schofield-Hurwitz, S., Hardt, J., Levasseur, G. and Tremblay, M. (2009). The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine & Tobacco Research*, **1**(11), 12–24.

Corno, L. and de Walque, D. (2012). Mines, migration and HIV/AIDS in southern Africa. *Journal of African Economies*, **21**(3), 465–98.

de Paula, A., Shapira, G. and Todd, P. (2014). How beliefs about HIV affect risky behaviors: evidence from Malawi. *Journal of Applied Econometrics*, **29**(6), 944–64.

de Walque, D. (2007). Sero-discordant couples in five African countries: implications for prevention strategies. *Population and Development Review*, **33**(3), 501–23.

———, Dow, W. H., Nathan, R., et al. (2012). Incentivising safe sex: a randomised trial of conditional cash transfers for HIV and sexually transmitted prevention in rural Tanzania. *BMJ Open*, **2**(1), 1–10.

Dow, W. and Philipson, T. (1993). An empirical examination of the implications of assortative matching on the incidence of HIV. *Journal of Health Economics*, **15**, 735–49.

Fenton, K., Johnson, A., McManus, S. and Erens, B. (2001). Measuring sexual behaviour: methodological challenges in survey research. *Sexually Transmitted Infections*, **77**, 84–92.

Fishbein, M. and Pequegnat, W. (2000). Evaluating AIDS prevention interventions using behavioral and biological outcome measures. *Sexually Transmitted Disease*, **27**(2), 599–653.

Gallo, M. F., Behets, F. M., Steiner, M. J., et al. (2006). Prostate-specific antigen to ascertain reliability of self-reported coital exposure to semen. *Sexually Transmitted Infections*, **33**(8), 476.

Gersovitz, M., Jacoby, H., Dedy, S. and Goze Tape, A. (1998). Measuring sexual behaviour: methodological challenges in survey research. *Journal of the American Statistical Association*, **93**, 875–83.

Gong, E. (2015). HIV testing and risky sexual behaviour. *Economic Journal*, **125**, 32–60.

Hall, H. I., Song, R., Rhodes, P., Prejean, J., An, Q., Lee, L. M., Karon, J., Brookmeyer, R., Kaplan, E., McJenna, M. and Janssen, R. S. (2008). Estimation of HIV incidence in the United States. *Journal of the American Statistical Association*, **300**(5), 520–9.

Hausman, J. and Wise, D. (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica*, **47**(2), 455–73.

———, Abrevaya, J. and Scott-Morton, F. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, **87**(2), 239–69.

Hayashi, F. (2001). *Econometrics*Princeton, NJ: Princeton University Press.

Hellhammer, D. H., Wust, S. and Kudielka, B. M. (2009). Salivary cortisol as a biomarker in stress research. *Psychoneuroendocrinology*, **34**, 163–71.

Hethcote, H. (2000). The mathematics of infectious diseases. *SIAM Review*, **42**(4), 599–653.

Hewett, P., Mensch, B. and Ribeiro, M. (2008). Using sexually transmitted infection biomarkers to validate reporting of sexual behavior within a randomized, experimental evaluation of interviewing methods. *American Journal of Epidemiology*, **168**(2), 202–11.

Hewitt, M. (2002). Attitudes toward interview mode and comparability of reporting sexual behavior by personal interview and audio computer assisted self-interviewing: analyses of the 1995 National Survey of Family Growth. *Sociological Methods and Research*, **31**, 3–26.

Hyman, J., Li, J. and Stanley, E. (2001). The initialization and sensitivity of multigroup models for the transmission of HIV. *Journal of Theoretical Biology*, **208**, 227–48.

Juerges, H., Kruk, E. and Reinhold, S. (2013). The effect of compulsory schooling on health: evidence from biomarkers. *Journal of Population Economics*, **26**, 645–72.

Karlan, D. and Zinman, J. (2012). List randomization for sensitive behavior: an application for measuring use of loan proceeds. *Journal of Development Economics*, **98**(1), 71–5.

Mauck, C. and Straten, A. (2008). Using objective markers to assess participant behavior in HIV prevention trials of vaginal microbicides. *Journal of Acquired Immune Deficiency Syndromes*, **1**, 49–64.

McKenzie, D. and Siegel, M. (2013). Eliciting illegal migration rates through list randomization. *Migration Studies*, **1**(3), 253–7.

Nelson, K. and Masters Williams, C. (2007). *Infectious Disease Epidemiology: Theory and Practice*, 2nd edn. Boston, MA: Jones and Bartlett.

Nnko, S., Boerma, J. T., Urassa, M., Mwaluko, G. and Zaba, B. (2004). Secretive females or swaggering males? An assessment of the quality of sexual partnership reporting in rural Tanzania. *Social Science and Medicine*, **59**(2), 299–310.

Oster, E. (2005). Sexually transmitted infections, sexual behavior and the HIV/AIDS epidemic. *Quarterly Journal of Economics*, **120**(2), 467–515.

Ozler, B. (2013). *Economists have experiments figured out*. What's next? (Hint: it's measurement). World Bank, Development Impact, 14 January; available online at http://blogs.worldbank.org/impactevaluations/economists-have- experiments-figured- out-what-s-next-hint-it-s-measurement (accessed 23 June 2018).

Palen, L., Smith, E. A., Caldwell, L. L., Flisher, A. J., Wegner, L. and Vergnani, T. (2008). Inconsistent reports of sexual intercourse among South African high school students. *Journal of Adolescent Health*, **43**(3), 221–7.

Ripa, T. and Nilsson, P. (2007). A chlamydia trachomatis strain with a 377-bp deletion in the cryptic plasmid causing false-negative nucleic acid amplification tests. *Sexually Transmitted Diseases*, **34**(5), 255–6.

Schmid, G., Buve, A., Mugyenyi, P., Garnett, G. P., Hayes, R. J., Williams, B. G., Calleja, J. G., De Cock, K. M., Whitworth, J. A., Kapiga, S. H., Gjys, P., Hankins, C., Zaba, B., Heimer, R. and Boerma J. (2004). Transmission of HIV-1 infection in sub-Saharan Africa and effect of elimination of unsafe injections. *Lancet*, **363**, 482–8.

Siegal, H., Li, L., Leviton, L., Cole, P., Hook, E., Bachmann, L. and Ford, J. (1999). Under the influence: risky sexual behavior and substance abuse among driving under the influence offenders. *Sexually Transmitted Diseases*, **26**(2), 87–92.

Subar, A., Kipnis, V., Troiano, R. P., Midthune, D., Schoeller, D., et al. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the OPEN study. *American Journal of Epidemiology*, **1**(158), 1–13.

Tennekoon, V. and Rosenman, R. (2014). 'Behold, a virgin is with HIV!' Misreporting sexual behavior among infected adolescents. *Health Economics*, **3**, 345–58.

Tourangeau, R. and Smith, T. (1996). Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, **60**, 275–304.

Treibich, C. and Lépine, A. (2016). Estimating misreporting in sensitive health behaviours: evidence from condom use of female sex workers in Senegal. London School of Hygiene and Tropical Medicine Working Paper.

UNAIDS (2013). Global report: UNAIDS report on the global AIDS epidemic. Discussion Paper, Joint United Nations Programme on HIV/AIDS, Geneva.

UNAIDS (2016). UNAIDS Factsheet—latest statistics on the status of the AIDS epidemic; available online at www.unaids.org/en/resources/fact-sheet (accessed 23 June 2018).

Wellings, K., Collumbien, M., Slaymaker, M., Singh, S., Hodges, Z., Patel, D. and Bajos, N. (2006). Sexual behaviour in context: a global perspective. *Lancet*, **368**(9548), 1706–28.

World Health Organization (WHO) (2008). Global incidence and prevalence of selected curable sexually transmitted infections. Department of Reproductive Health and Research; available online at www.who.int/reproductivehealth/publications/rtis/stisestimates/en (accessed 26 June 2018).

World Health Organization (WHO) (2009). *Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks*. Geneva: WHO Publications.

World Health Organization (WHO) (2018). WHO Factsheet—human papillomavirus (HPV) and cervical cancer; available online at www.who.int/mediacentre/factsheets/fs380/en (accessed 18 February 2018).