ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2015)

Diagnostics for respondent-driven sampling

Krista J. Gile,

University of Massachusetts, Amherst, USA

Lisa G. Johnston

Tulane University, New Orleans, and University of California, San Francisco, USA

and Matthew J. Salganik

Microsoft Research, New York, and Princeton University, USA

[Received August 2013]

Summary. Respondent-driven sampling (RDS) is a widely used method for sampling from hard-to-reach human populations, especially populations at higher risk for human immunodeficiency virus or acquired immune deficiency syndrome. Data are collected through a peer referral process over social networks. RDS has proven practical for data collection in many difficult settings and has been adopted by leading public health organizations around the world. Unfortunately, inference from RDS data requires many strong assumptions because the sampling design is partially beyond the control of the researcher and not fully observable. We introduce diagnostic tools for most of these assumptions and apply them in 12 high risk populations. These diagnostics empower researchers to understand their RDS data better and encourage future statistical research on RDS sampling and inference.

Keywords: Acquired immune deficiency syndrome; Diagnostics; Exploratory data analysis; Hard-to-reach populations; Human immunodeficiency virus; Link tracing sampling; Non-ignorable design; Respondent-driven sampling; Social networks; Survey sampling

1. Introduction

Many problems in social science, public health and public policy require detailed information about 'hidden' or 'hard-to-reach' populations. For example, efforts to understand and control the human immunodeficiency virus (HIV) and acquired immune deficiency syndrome epidemic require information about the disease prevalence and risk behaviours in populations at higher risk of HIV exposure: female sex workers, FSW, illicit drug users, DU, and men who have sex with men, MSM (Magnani *et al.*, 2005). Respondent-driven sampling (RDS) is a recently introduced link tracking network sampling technique for collecting such information (Heckathorn, 1997). Because of the pressing need for information about populations at higher risk and the weaknesses of alternatives approaches, RDS has already been used in hundreds of HIV-related studies in dozens of countries (Malekinejad *et al.*, 2008; Montealegre *et al.*, 2013) and has been adopted by leading public health organizations, such as the US Centers for Disease Control and Prevention (Lansky *et al.*, 2007; Barbosa Júnior *et al.*, 2011; Wesnert *et al.*, 2012) and the World Health Organization (Johnston, Chen, Silva-Santisteban and Raymond, 2013).

Address for correspondence: Krista J. Gile, Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003-9305, USA.
E-mail: gile@math.umass.edu

Collectively, these previous studies demonstrate that RDS can generate large samples in a wide variety of hard-to-reach populations. However, the quality of estimates derived from these data has been challenged in many recent references (Heimer, 2005; Scott, 2008; Poon *et al.*, 2009; Bengtsson and Thorson, 2010; Goel and Salganik, 2010; Gile and Handcock, 2010; White *et al.*, 2012; McCreesh *et al.*, 2012; Salganik, 2012; Burt and Thiede, 2012; Nesterko and Blitzstein, 2014; Mouw and Verdery, 2012; Mills *et al.*, 2012; Rudolph *et al.*, 2013; Yamanis *et al.*, 2013). A major source of concern is that inference from RDS data requires many strong assumptions, which are widely believed to be untrue and yet are rarely examined in practice. The widespread use of RDS for important public health problems, combined with its reliance on untested assumptions, creates a pressing need for exploratory and diagnostic techniques for RDS data.

RDS data collection begins when researchers select, in an *ad hoc* manner, typically 5–10 members of the target population to serve as 'seeds'. Each seed is interviewed and provided a fixed number of coupons (usually three) that they use to recruit other members of the target population. These recruits are in turn provided with coupons that they use to recruit others. In this way, the sample can grow through many waves, resulting in recruitment trees like those shown in Fig. 1 (created by using NetDraw; Borgatti (2002)). Respondents are encouraged to participate and recruit through the use of financial and other incentives (Heckathorn, 1997). The fact that the majority of participants are recruited by other respondents and not by researchers makes RDS a successful method of data collection. However, the same feature also inherently complicates inference because it requires researchers to make assumptions about the recruitment process and the structure of the social network connecting the target population.

There are three interrelated approaches to addressing the assumptions underlying inference from RDS data. First, researchers can identify assumptions whose violations significantly impact estimates, either analytically or through computer simulation. Second, researchers can develop new estimators that are less sensitive to these assumptions. Third, researchers can develop methods to detect the violation of assumptions in practice. This third approach is the primary focus of this paper, but we hope that our results will help to motivate and inform research of the first two types.

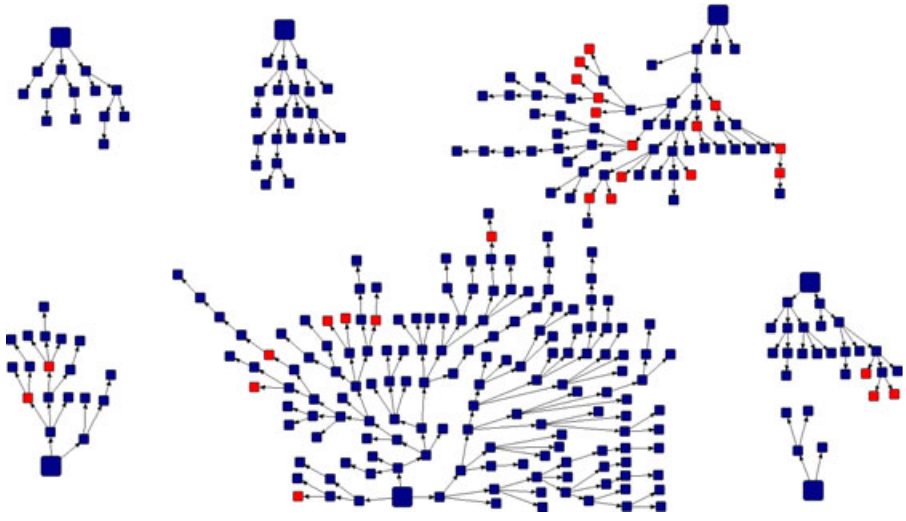


Fig. 1. Recruitment trees plot from the sample of men who have sex with men in Higuey: ■, self-identification as 'heterosexual'

This paper makes two main contributions. First, we review and develop diagnostics for most assumptions underlying statistical inference from RDS data. One reason for the relative dearth of RDS diagnostics is that the same conditions that complicate inference from RDS data also complicate formal diagnostic tests. In particular, the unknown dependence between recruiters and recruits renders most standard tests invalid. Therefore, when possible, we develop diagnostic approaches that are intuitive, graphical and not reliant on statistical testing. Further, when possible, we emphasize approaches that can be used while data collection is occurring so that some problems can be investigated and potentially resolved while researchers are still in the field. In addition, our diagnostics frequently take advantage of three specific features of RDS studies that are not typically utilized: information about the time sequences of responses, contact with respondents who visit the study site twice and the multiple seeds that are used to begin the sampling process. The second main contribution of our paper is to deploy these diagnostics in 12 RDS studies conducted in accordance with the national strategic HIV surveillance plan of the Dominican Republic. We believe that these case-studies—which include samples of female sex workers, FSW, drug users, DU, and men who have sex with men, MSM, in four cities—are reasonably reflective of the way that RDS is used in many countries. Therefore, we believe that our empirical results have broad applicability for RDS practitioners and researchers who wish to develop improved methods of RDS data collection and inference.

The remainder of the paper is organized as follows: in Section 2 we briefly review the assumptions underlying RDS estimation and in Section 3 we describe the data from 12 studies in the Dominican Republic that will be used throughout the paper. Sections 4–8 present diagnostics, including extensions of previous approaches as well as wholly new approaches. In Section 9 we discuss the results and conclude with suggestions for future research. We also include on-line supporting information with additional results and approaches.

The code used in these analyses can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Assumptions of respondent-driven sampling

Estimation from RDS data requires many assumptions about the sampling process, the underlying population and respondent behaviour. These assumptions are outlined in Table 1 and described fully in Gile and Handcock (2010). In particular, these assumptions are required by the estimator that was proposed by Volz and Heckathorn (2008). Other available estimators require similar assumptions, especially pertaining to respondent behaviour.

Each row of Table 1 includes assumptions according to their roles in allowing for estimation. The first row ('Random-walk model') corresponds to assumptions that are required to allow the sampling process to be approximated by a random walk on the nodes. Critically, the random-walk model requires with-replacement sampling, whereas the true sampling process is known to be without replacement. We, therefore, first consider diagnostics that are designed to detect impacts of the without-replacement nature of the sampling (Section 4).

The second row ('Remove seed dependence') contains assumptions that are required to reduce the influence of the initial sample—the seeds—on the final estimates. Because the initial sample is usually a convenience sample, RDS is intended to be carried out for many sampling waves through a well-connected population to minimize the effect of the seed selection process. Therefore, we consider diagnostics that are designed to detect seed bias that may remain because of an insufficient number of sample waves (Section 5).

The final row of Table 1 ('Respondent behaviour') contains assumptions that are related to

Table 1. Assumptions of the Volz–Heckathorn estimator†

	Network structure assumptions	*Sampling assumptions*
Random-walk model	Network size large ($N \gg n$)	*With-replacement sampling (4)* Single non-branching chain
Remove seed dependence	*Homophily sufficiently weak (5)* *Bottlenecks limited (5)* Connected graph	*Enough sample waves (5)*
Respondent behaviour	*All ties reciprocated (6)*	*Degree accurately measured (7)* *Random referral (8)*

†Assumptions in italics are considered in this paper, with section numbers given. A version of this table appeared in Gile and Handcock (2010).

respondents' behaviour. In RDS, unlike in traditional survey sampling, respondents' decision making plays a significant role in the sampling process, and, therefore, assumptions about these decisions are needed for estimating inclusion probabilities. In particular, we consider the assumptions that all network ties are reciprocated, that degree (also referred to as the number of contacts or personal network size) is accurately reported, and that future participation is random among contacts in the target population (Sections 6–8).

3. Case-study: 12 sites in the Dominican Republic

We employ these diagnostics in a case-study of 12 parallel RDS studies conducted in the spring of 2008 by using standard RDS methods (Johnston, 2008, 2013; Johnston, Caballero, Dolores and Values, 2013). As part of the national strategic HIV surveillance plan of the Dominican Republic, data were collected from female sex workers, FSW, drug users, DU, and men who are gay, transsexual or have sex with men, MSM, in four cities: Santo Domingo (denoted by 'SD'), Santiago (denoted by 'SA'), Barahona (denoted by 'BA') and Higuey (denoted by 'HI'). These studies are typical of the way that RDS is used in national HIV surveillance around the world. Eligible people were 15 years old or older and lived in the province under study. Eligible FSW were females who had exchanged sex for money in the previous 6 months, DU were females or males who had used illicit drugs in the previous 3 months, and MSM were males who had had anal or oral sexual relations with another man in the previous 6 months. Seeds were purposively selected through local non-governmental organizations or through the use of peer outreach workers. Each city had a fixed interview site where respondents enrolled in the survey.

During the initial visit, consenting respondents were screened for eligibility, completed a face-to-face interview, received HIV pretest counselling and provided blood samples that were tested for HIV, hepatitis B and C, and syphilis. Before leaving the study site, respondents were encouraged to set an appointment to return 2 weeks later for a follow-up visit during which they would receive HIV post-test counselling, collect infection test results and, if necessary, be referred to a nearby health facility for care and treatment. During the follow-up visit respondents also completed a follow-up questionnaire and received secondary incentives for any peers whom they recruited; respondents were compensated the equivalent of US $9.00 for completing the initial survey and US $3.00 for each successful recruitment (up to a maximum of three). To ensure confidentiality, respondents' coupons, questionnaires and biological tests were identified by using a unique study identification number; no personal identifying information was collected. The studies ranged in sample size from 243 to 510 with a total sample size of 3866 people, of

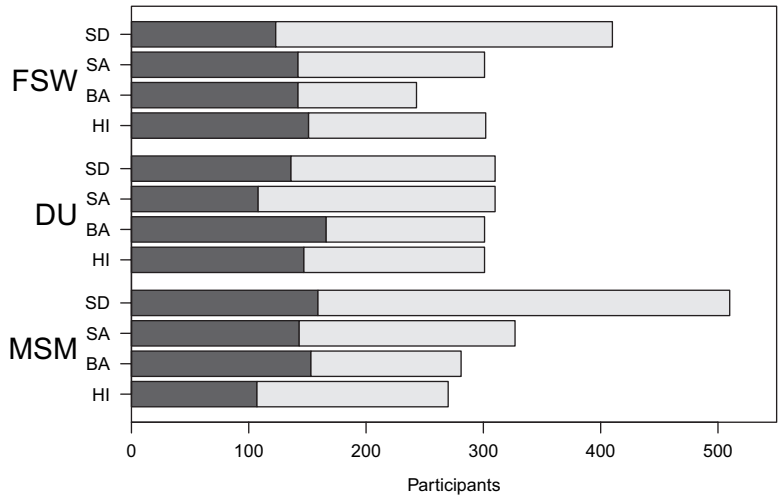


Fig. 2. Sample sizes from the 12 studies (in total, 3866 people participated, of whom 1677 (43%) completed a follow-up survey): ■, initial and follow-up: □, initial only

whom 1677 (43%) completed a follow-up survey (Fig. 2). In the on-line supporting information we compare the characteristics of respondents who did and did not complete the follow-up survey. We found that follow-up respondents tended to recruit more often and to participate earlier in the study than follow-up non-respondents. For more information, see section S6. Where relevant to our conclusions, the effects of these results are noted.

We analyse data from the 12 studies by using the estimator that was introduced in Volz and Heckathorn (2008) because it has been used in most of the recent evaluations of RDS methodology (Wejnert, 2009; Goel and Salganik, 2010; Gile and Handcock, 2010; Tomas and Gile, 2011; Nesterko and Blitzstein, 2014; Lu *et al.*, 2012; McCreesh *et al.*, 2012; Yamanis *et al.*, 2013). The estimator of the proportion of the population with a specific trait (e.g. HIV infection) is

$$\hat{p} = \frac{\sum_{j \in I} 1/d_j}{\sum_{j \in S} 1/d_j}, \tag{1}$$

where S is the full sample, I is the infected sample members and d_j is the self-reported 'degree', or number of contacts of respondent j. Equation (1), which is sometimes called the RDS II estimator or the Volz–Heckathorn (VH) estimator, is a generalized ratio estimator of a population mean, with inverse probability weighting and sampling weights proportional to degree.

4. With-replacement sampling

Many estimators for RDS data are based on the assumption that the sample can be treated as a with-replacement random walk on the social network of the target population. In particular, respondents are assumed to choose freely whom of their contacts to recruit into the study. In practice, sampling is *without* replacement; respondents are not allowed to recruit people who have already participated. This restriction may lead to inaccurate estimates of sampling probabilities and biased estimates, as described in Gile (2011).

In a very large, highly connected population, it is possible that respondents can pass coupons

much as they would in a with-replacement sample and, in such cases, the with-replacement approximation is probably adequate. In contrast, indications that earlier respondents influenced subsequent sampling decisions would suggest potentially problematic violation of the sampling-with-replacement assumption. Previous samples may affect sampling in two ways: locally, when members of a small well-connected subgroup are sampled at a high rate, influencing the future referral choices of other subgroup members, and globally, when the target population as a whole is sampled at a sufficiently high rate that later samples are influenced by earlier samples. In this section, we examine the with-replacement sampling assumption in several ways. First, we use three types of evidence to detect local and global effects of previous samples. Next, we assess the effect of global without-replacement sampling on estimates. Finally, we compare the methods and conclude with recommendations.

4.1. Failure to attain sample size

One apparently straightforward indication of global finite population effects on sampling is a failure to attain the study's target sample size due to the inability of participants to recruit additional members of the target population. Three of our studies, FSW–BA, MSM–BA and MSM–HI, failed to reach their target sample sizes, suggesting that they may have exhausted the available portions of their respective populations. As a diagnostic, however, this indicator has three primary limitations. First, it cannot be assessed until the study is complete. Second, although failure to attain the sample size is an indication that the available population has been exhausted, the absence of such failure is not an indication that those effects are absent. Finally, the *available* population could be dramatically smaller than the *target* population because of insufficient respondent incentives, inadequate network connections in the populations or negative perception of the study in the target community. Given these limitations, however, failure to attain the target sample size could be an indication that the target population is nearly fully sampled, which would suggest that estimates treating the sample as a small fraction of the population are not appropriate.

4.2. Failed recruitment attempts

If the sampling process were not influenced by the previous sample, each respondent could distribute coupons without considering whether contacts had already participated in the study. Therefore, respondents who returned for a follow-up survey were asked

> '(A) How many people did you try to give a coupon but they had already participated in the study?'.

Responses to this question are summarized in Fig. 3. Rates of failed coupon distributions varied widely by site, with the most failures among drug users in Higuey, with over half of follow-up respondents reporting a failed attempt to distribute a coupon, and the fewest failures being among the DU in Santo Domingo and Santiago, and among the FSW in Santiago and Barahona, with 3% or fewer respondents reporting failed coupon distributions. In six of the 12 sites, at least 25% of respondents participating in follow-up interviews indicated that they had attempted to give coupons to at least one person who had already participated in the study. Note that these rates of failed recruitment attempts may be underestimates, as follow-up respondents had significantly more successful recruitments than those who did not respond to follow-up (see section S6 in the on-line supporting information). Where present, these reported failures provide direct evidence that respondents' recruiting decisions were affected by earlier parts of the sample. Where absent, they can either indicate a lack of such influence or accurate knowledge of which alters have already participated in the study.

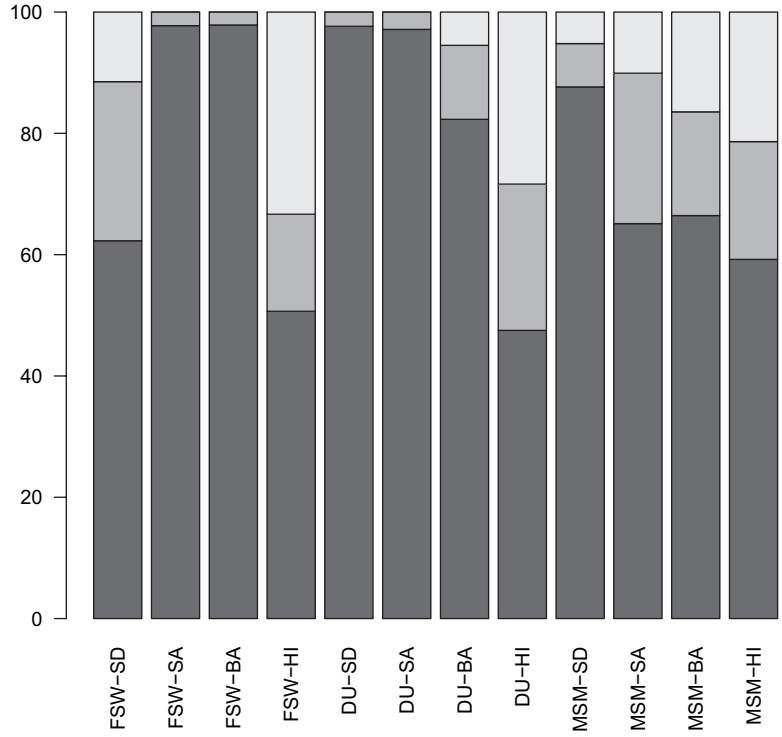


Fig. 3. Percentage of respondents reporting 0 (■), 1–3 (■) or four or more(■) failed recruitment attempts: in six sites, at least 25% of respondents reported at least one failed recruitment attempt

4.3. Contacts participated

Respondents' coupon passing choices could also be influenced by the contacts they know who have already participated in the study. To assess this possibility, respondents were asked the following question (see also McCreesh *et al.* (2012)):

> '(B) How many other MSM/DU/FSW do you know that have already participated in this study, without counting the person who gave a coupon to you?'.

Across all 12 data sets, only 30% of respondents answered '0', and the mean proportion of alters reported to have already participated was 36%. This result suggests that previously sampled population members may indeed impact the alters who are available for the passing of coupons. Note that about 10% of respondents (347 out of 3866) reported knowing more people who had already participated than they reported knowing (which was collected in question (F); see Section 7). Throughout this section, we truncate responses at one less than the reported number of people known.

If the rate of known participants is uniform across the sampling process, it may be partially explained by measurement error or low level local clustering with minimal connection to global finite population effects. An increase in this effect over the course of the sample, however, suggests that the population is becoming increasingly depleted, such that previously sampled alters constrain the choices of later respondents more than those of earlier respondents. In looking for evidence of a time trend, we fitted a simple linear model relating the sample order to the proportion of alters who had already participated. Results using survey time or more complex models were similar. To serve as a conservative flagging criterion, in a setting where formal

testing is likely to be invalid, we flag any cases with positive trends over time. We find positive trends in probability of having been previously sampled for increasing survey order in eight of the 12 populations (DU–SD, DU–SA, FSW–SA, FSW–BA, FSW–HI, MSM–SA, MSM–BA and MSM–HI), suggestive of potential finite population effects. In the on-line supporting information, we consider two approaches to visualizing these effects.

4.4. Assessing effects on estimates

The results in Sections 4.1–4.3 focused on detecting the effect of previous samples on the sampling process. Next, we turn to detecting global effects on estimates by using an approach that requires knowing or estimating the size of the target population. If the target population is very large compared with the sample size, then global exhaustion is unlikely to be of concern. If the target population is small, however, then a bias may be induced, but the magnitude of estimator bias will depend on the relative degree distributions of the groups of interest (such as infected and uninfected people): the greater the systematic difference in degrees, the greater the potential bias in estimates (Gile, 2011; Gile and Handcock, 2010). Such biases can be mitigated by using estimators that are designed to account for finite population effects, such as the estimator that is based on successive sampling (SS) introduced in Gile (2011) and implemented in the R (R Core Team, 2012) packages `RDS` (Handcock *et al.*, 2009) and `RDS Analyst` (Handcock *et al.*, 2013). Note that the SS estimator differs from the VH estimator only in that the former uses finite population adjusted sampling weights. Therefore, although both may be affected by other sampling anomalies, the effects of other factors will be nearly identical. Thus, a comparison between the results of the SS estimator and the VH estimator can serve as a sensitivity analysis to global population exhaustion. If the two estimators are nearly identical for reasonable estimates of the population size, then global exhaustion is probably not inducing bias into estimates.

To undertake this sensitivity analysis, and as described in greater detail in the on-line supporting information, we estimated the size of our target populations by using two different approaches:

(a) drawing on meta-analysis of related studies and
(b) the approach that was introduced in Handcock *et al.* (2012) and implemented in the package `size` (Handcock, 2011), which uses information in the degree sequence in the RDS sample.

Using these estimated population sizes, we then compared the SS and VH estimators in all 12 target populations for all characteristics described in section S2 in the supporting information (120 trait–site combinations). In most cases, the two estimates were within 0.01 of each other (see Fig. 4, as well as Fig. S4 in the supporting information); Table S1 in the supporting information lists all traits with differences larger than 0.01. Overall, therefore, this analysis suggests that there were not large finite population effects on the VH estimator in these studies. In the supporting information we also illustrate *population size sensitivity plots* that can be made for individual study sites.

4.5. Comparison of approaches and current recommendations

Table 2 summarizes all the sampling process indicators across study sites. Failure to attain the desired sample size (studies FSW–BA, MSM–BA and MSM–HI) is an indication that the earlier samples impacted the later sampling decisions. Consistent with this result, the MSM sites in Barahona and Higuey showed evidence of without-replacement sampling effects on all three of these proposed indicators. Nearly all sites, however, had evidence of finite population effects on at least one indicator. Together, these indicators show that without-replacement

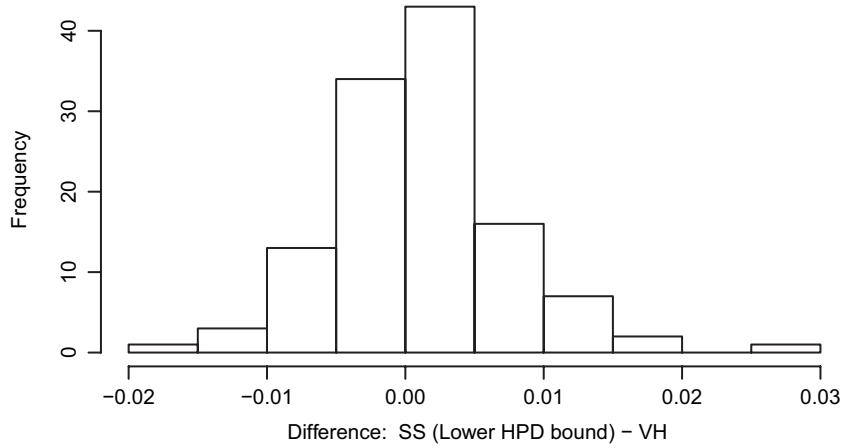


Fig. 4. Histogram of the difference between SS and VH estimators, over many traits: the successive sampling estimates are based on a 'worst-case' small approximated population size, based on the lower bound of the highest posterior density interval generated by the population size estimation method in Handcock *et al.* (2012)

Table 2. Summary of indicators of violations of the with-replacement sampling assumption†

	Results for the FSW				*Results for the DU*				*Results for the MSM*			
	SD	*SA*	*BA*	*HI*	*SD*	*SA*	*BA*	*HI*	*SD*	*SA*	*BA*	*HI*
Failed to attain sample size			×								×	×
Failed attempts > 25%	×			×				×		×	×	×
Increasing participants known		×	×	×	×	×		×		×	×	×

†The first row indicates sites which could not attain the intended sample sizes (Section 4.1). The second indicates at least 25% of follow-up respondents reporting that they attempted to give coupons to at least one person who had already participated in the study (Section 4.2). The third indicates a positive coefficient of sample order in the linear regression model for the probability that an alter is in the study (Section 4.3).

effects on sampling were frequent and that reasonable diagnostic approaches for detecting them can produce different results. These differences between indicators can either be the result of random variation or the result of different indicators reflecting different features of the underlying recruitment process. A fourth sampling process indicator, decreasing degree sequence in the sample, is described in the on-line supporting information. Contrary to our expectations based on Gile (2011), direct evaluation of the trend in degree over time suggested little evidence of population exhaustion on sampling.

The most effective diagnostic of global effects on estimates is the comparison of the VH and SS estimators. Unlike the other indicators, this indicator measures the direct effect on the estimate. It is possible that global population exhaustion influences sampling (as indicated by one of the earlier indicators) but does not induce bias in the estimator because of other features of the network, such as similar degree distributions between the two subpopulations of interest. This is the case, for example, among the FSW in Barahona and MSM in Higuey, which do not exhibit worrisome effects on estimates, despite failing to reach their intended sample sizes.

Among the MSM in Barahona, however, the large sample fraction may be influencing estimates. One challenge in implementing this diagnostic is that the SS estimator requires an estimate of the size of the target population, and these size estimates can be difficult to construct (UNAIDS, 2010; Bernard *et al.*, 2010; Salganik *et al.*, 2011; Handcock *et al.*, 2012; Johnston, Prybylski, Raymond, Mirzazadeh, Manopaiboon and McFarland, 2013).

In future studies, questions about failed recruitments and numbers of known participants (questions (A) and (B)) can be helpful in diagnosing local effects, and should be collected. Further, when diagnostics suggest large global effects of previous samples, researchers should use estimators that do not depend on the sampling-with-replacement assumption (e.g. Gile (2011) and Gile and Handcock (2011)), or minimally these estimators should be used for sensitivity analysis as in Section 4.4. Methods for inference in the presence of local effects of previous samples are not yet available.

5. Assessing seed dependence

In RDS studies the seeds are not selected from a sampling frame; instead, they are an *ad hoc* convenience sample. In general, the seed selection mechanism has not concerned RDS researchers because of asymptotic results suggesting that the choice of seeds does not affect the final estimate (Heckathorn, 1997, 2002; Salganik and Heckathorn, 2004). However, these asymptotic results hold only as the sample size approaches ∞ and, in practice, samples may not be sufficiently large to justify this approximation. Therefore, a natural question is whether a given sample is sufficiently large to overcome the potential biases that are introduced during seed selection.

There are some apparent similarities between this problem and the monitoring of convergence of computer-based Markov chain Monte Carlo simulations. Standard Markov chain Monte Carlo methods, unfortunately, cannot be directly applied here. First, single-chain methods, such as that of Raftery and Lewis (1992), are not applicable because we have multiple trees created by the multiple seeds. Further, multiple-chain methods, such as that of Gelman and Rubin (1992), are not directly applicable because RDS trees are of different lengths. Finally, these standard approaches typically rely on sample chains that are far longer than are available in RDS data; the longest tree in these studies had a maximum of 16 waves.

The currently used diagnostic for assessing whether the RDS sample is no longer affected by the seeds is to compare the length of the longest tree with the estimated number of waves required for the sampling process to approximate its stationary distribution under a first-order Markov chain model on group membership (Heckathorn *et al.*, 2002). This approach is now standard in the field (Johnston, Malekinejad, Kendall, Iuppa and Rutherford, 2008; Malekinejad *et al.*, 2008; Montealegre *et al.*, 2013). However, it is not ideal for four reasons:

(a) it is typically interpreted as allowing sampling to stop after the supposed stationary distribution has been attained rather than collecting most of the sample from the stationary distribution,
(b) it is based on a model for the sampling process which is different from that assumed in most RDS estimators,
(c) it is focused on the sample composition and not the estimates, and
(d) it has generated much confusion (see for example, Ramirez-Valles *et al.* (2005a, b), Heimer (2005) and Wejnert and Heckathorn (2008)).

Here we propose a series of graphical approaches that help to assess whether there are lingering effects of the choice of seeds on the estimates.

First, we suggest that researchers examine the dynamics of the RDS estimate. Roughly, the more the estimate changes as we collect more data, the more concern we should have that the choice of seeds is still influencing the estimate (see also Bengtsson *et al.* (2012) for a similar approach). More concretely, let $\hat{p}_t$ be the estimated trait prevalence by using the first t observations (where we exclude all seeds). To assess the possible lingering effect of seed selection, we plot $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n$ and see whether the estimates seem to stabilize. Fig. 5(a) shows a *convergence plot* for the proportion of DU in Barahona who report using drugs every day. The estimate is increasing over time, suggesting that the seeds and early respondents were atypical in their drug use frequency. This constant and sharp increase in estimates actually underrepresents the differences between the early and late parts of the sample because the estimate is cumulative. For example, on the basis of the first 50 respondents we would estimate that 8% of the population use drugs every day, but from the final 50 respondents we would estimate that 67% use drugs every day. Compare these dynamics with Fig. 5(b), which plots the estimated proportion of DU in Barahona who reported engaging in unprotected sex in the previous 30 days. This estimate appears to be stable for the second half of the sample. Note that both of these estimates arise from the same sample and, therefore, highlight the fact that convergence is a property of an estimate not a sample.

Convergence plots, however, can mask important differences between seeds. Therefore, we also recommend creating *bottleneck plots* that show the dynamics of the estimates from each seed individually. For example, the convergence plot for the estimated proportion of MSM in Santo Domingo who have had sex with a women in the previous 6 months appears to be stable (Fig. 6(a)). However, despite this aggregate stability there are large differences between the data from the different seeds (Fig. 6(b)). More generally, a large difference in estimates between seeds suggests bottlenecks in the underlying social network that can substantially increase the effect of the seeds on the estimates (Goel and Salganik, 2009).

Bottleneck plots, although showing differences between seeds, obscure the fact that the trees can be started at different times and grow at different speeds. Therefore, we suggest an additional plot called the *all points plot*, which plots the unweighted characteristics of respondents by seed and sample order (Fig. 7(c)). To demonstrate how these three plots can work together, Fig. 7 plots the estimated proportion of MSM in Higuey who self-identify as heterosexual, which is a key 'bridge group' because they can spread infection between the high risk MSM group and the larger heterosexual population. The convergence plot (Fig. 7(a)) shows that there were no self-identified heterosexuals in the first 100 observations (header), but over time the sample started to reach people who identified as heterosexual. Since the estimate has not clearly stabilized, we should be worried that the final estimate of $\hat{p} = 0.12$ might be unduly influenced by the choice of seeds. Further, the bottleneck plot shows that the self-identified heterosexuals were reached only within certain trees, suggesting a possible problem with bottlenecks (Fig. 7(b)). Finally, the all points plot (Fig. 7(c)) reveals that self-identified heterosexuals were unusual in that they both appeared in the sample late and only in a small number of trees, a fact that is difficult to infer from the previous two plots.

5.1. Current recommendations

We recommend creating convergence plots, bottleneck plots and all points plots for all traits of interest during data collection. Evidence of unstable estimates from convergence plots (e.g. Fig. 5(a)) should be taken as an indication that results may be suspect and that more data should be collected. If additional data collection is not possible, researchers may need to use more advanced estimators that are designed to correct for features such as seed bias (e.g. Gile and Handcock (2011)). Evidence of bottlenecks (e.g. Fig. 6(b)) should be taken as an indication that

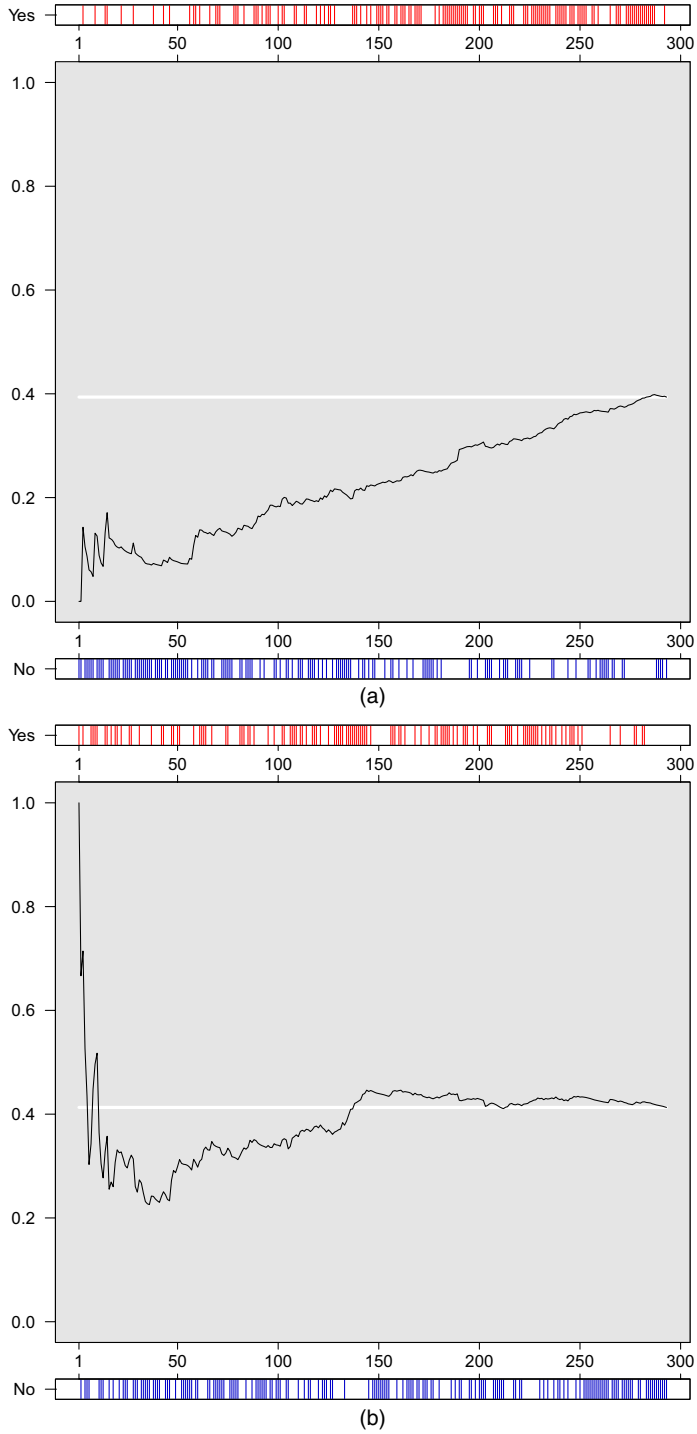


Fig. 5. Convergence plots showing $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n$ (the headers and footers plot the sample observations with and without the trait, and the white line shows the estimate based on the complete sample, $\hat{p}_n$): (a) DU–BA, use drugs every day; (b) DU–BA, risky sex

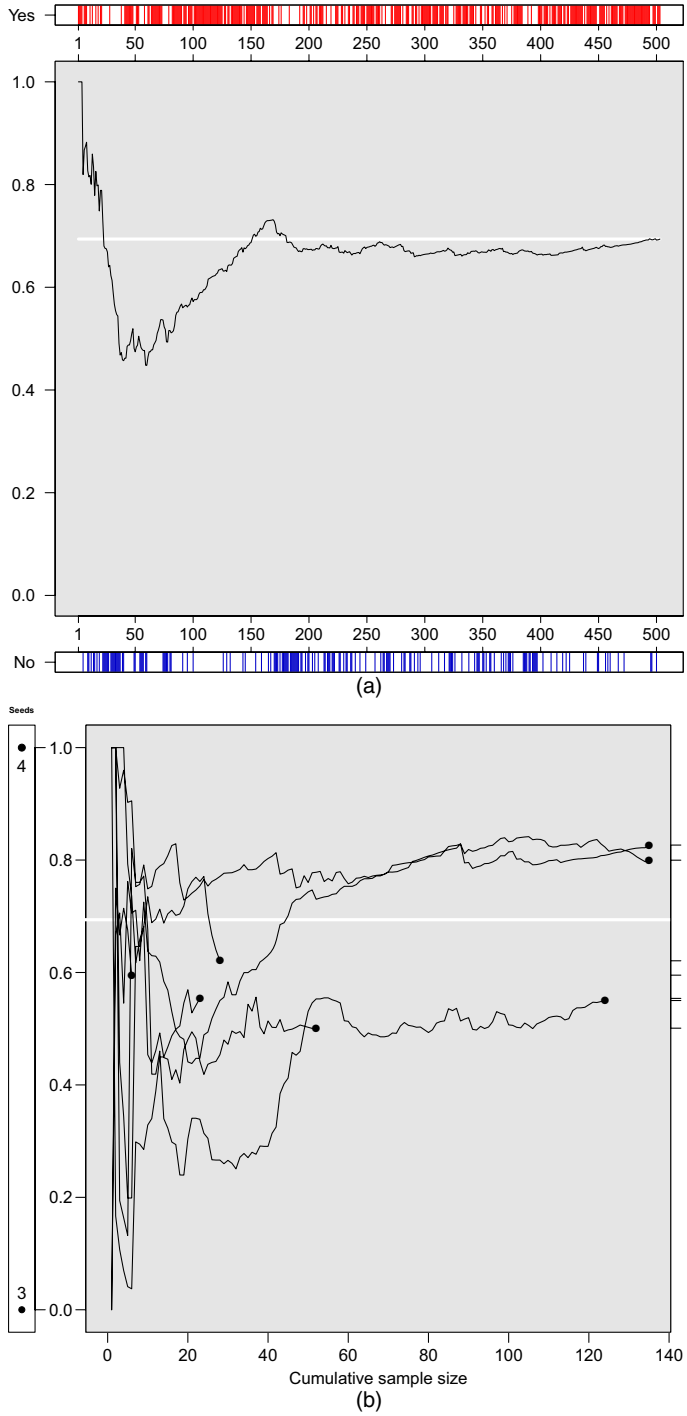


Fig. 6. (a) Convergence plot and (b) bottleneck plot for the proportion of MSM in Santo Domingo who have had sex with a woman in the previous 6 months: the convergence plot masks important differences between the seeds that are revealed by the bottleneck plot; in both plots, the white line shows the estimate based on the complete sample, $\hat{p}_n$

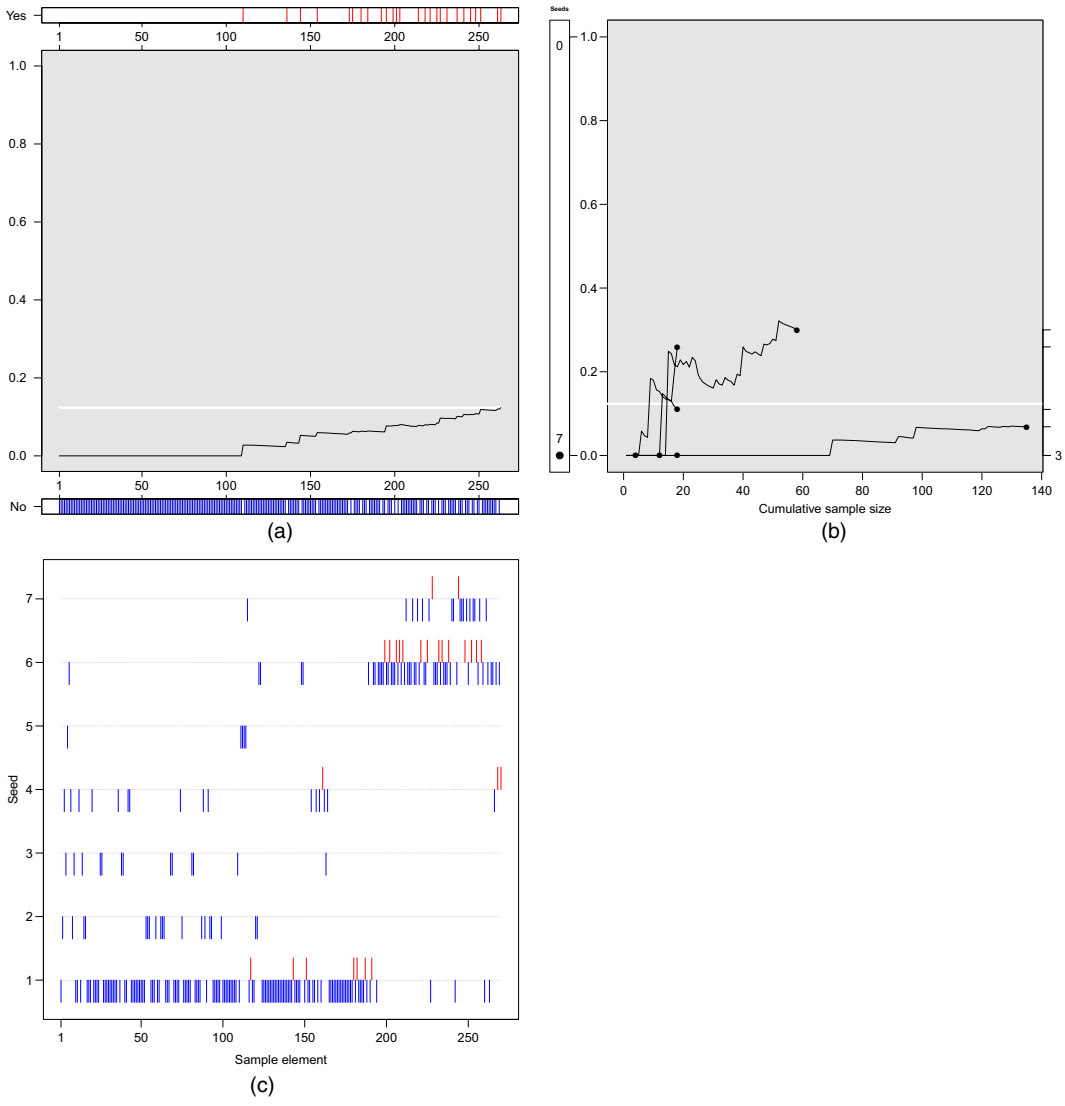


Fig. 7. Three diagnostic plots for estimates of MSM in Higuey who self-identify as heterosexual: (a) convergence plot showing that data collected late in the sample differ from data collected early in the sample; (b) bottleneck plot showing that the chains explored different subgroups, suggesting a problem with bottlenecks; (c) all points plot showing that the self-identified heterosexuals (|) were unusual in that they both arrived in the sample late and arrived from a small number of chains, a fact that is difficult to infer from (a) and (b)

estimates may be unstable and that more data should be collected. If additional data collection is not possible, researchers should consider presenting estimates for each tree individually rather than trying to combine them into an overall estimate, and researchers should be aware that standard RDS confidence intervals will be too small (Goel and Salganik, 2010). If it is not possible to create these plots during data collection, we suggest that they should still be made, used to consider alternative estimators and possibly presented with published results. Further, if it is not possible to monitor all of these plots closely during data collection—as might occur

in a large multisite study with many traits of interest—we suggest using flagging criteria such as those developed and applied in the on-line supporting information.

We wish to emphasize that there are cases where the convergence plots and bottleneck plots could fail to detect real problems. For example, in some situations the estimate might appear stable (Fig. 5(b)), but then the sample could move to a previously unexplored part of the target population, yielding very different estimates. Further, the bottleneck plot can fail in the presence of extremely strong bottlenecks and very unbalanced seed selection. For example, if there is a strong bottleneck between brothel-based and street-based sex workers and all the seeds are brothel based, the sample may never include street-based sex workers and the bottleneck plots would not be able to alert researchers to this problem.

6. Reciprocation

Most current RDS estimators use self-reported degree to estimate inclusion probabilities based on the assumption that all ties are reciprocated. Current best practice monitors this feature by asking respondents during their initial visit about their relationship with the person who recruited them, typically choosing from a set of categories (e.g. acquaintance, friend, sex partner, spouse, other relative, stranger or other) (Heckathorn, 2002; Lansky *et al.*, 2012). Here, we present responses to a slightly different question and, in the on-line supporting information (section S3), we present further discussion of additional approaches that are aimed at studying the reciprocation patterns in the broader social network.

On the follow-up questionnaire, for each coupon given out, respondents were asked

> '(C) Do you think that the person to whom you gave a coupon would have given you a coupon if you had not participated in the study first?'.

Table 3 shows the results of this question, separated by population and site. Overall, about 88% of responses indicated reciprocation, but there are notable differences across the populations and sites. Reciprocation rates in Santiago were considerably higher than in the other cities, and reciprocation rates of DU were lower than for other populations. The reciprocation rates among the DU were especially low in Higuey, and also in Barahona, where participants may have been selling coupons (for more on coupon selling, see Scott (2008), and also Broadhead (2008) and Ouellet (2008)).

6.1. Current recommendations

The reciprocity assumption requires that the recruiter and the recruit are known to each other and that both people would be willing to recruit each other. Therefore, we recommend that

Table 3. Percentage of affirmative responses to the question (C), 'Do you think that the person to whom you gave a coupon would have given you a coupon if you had not participated in the study first?'

	Results for the FSW				*Results for the DU*				*Results for the MSM*			
	SD	*SA*	*BA*	*HI*	*SD*	*SA*	*BA*	*HI*	*SD*	*SA*	*BA*	*HI*
% reciprocated	87	98	87	89	86	96	74	79	87	98	91	91

on the initial survey researchers should collect information about the relationship between the recruiter and recruit (see for example Heckathorn (2002)) and information directly assessing the possibility of recruitment (similar to question (C); see also Rudolph *et al.* (2013)). Researchers should calculate reciprocity rates as defined by both questions during data collection. Low rates of reciprocation by either measure could be used to improve field procedures (e.g. training respondents about how to recruit others) and alert researchers to potential problems (e.g. coupon selling). Further, high rates of non-reciprocation may require alternative RDS estimators; see Lu *et al.* (2013) for one such approach.

7. Measurement of degree

The VH estimator weights respondents on the basis of their self-reported degree (see equation (1)), and the fact that the estimates can depend critically on self-reported degree has troubled some RDS researchers (Frost *et al.*, 2006; Wejnert, 2009; Iguchi *et al.*, 2009; Goel and Salganik, 2009; Bengtsson and Thorson, 2010; Rudolph *et al.*, 2013) because of the well-documented problems with self-reported social network data in general (Bernard *et al.*, 1984; Marsden, 1990; Brewer, 2000). However, despite the widespread concern about degree measurement, the issue is rarely explored empirically in RDS studies (for important exceptions, see Wejnert and Heckathorn (2008), Wejnert (2009) and McCreesh *et al.* (2012)). Here we present several methods of assessing the measurement of degree and the resulting effects on estimates.

In this study, respondents were asked a series of four questions to measure degree (Johnston, Malekinejad, Kendall, Iuppa and Rutherford, 2008) (versions for the DU; the others are analogous):

> '(D) How many people do you know who have used illegal drugs in the past three months?
> (E) How many of them live or work in this province?
> (F) How many of them [repeat response from E] are 15 years old or older?
> (G) How many of them [repeat response from F] have you seen in the past week?'.

The response to the fourth question (G) was the degree used for estimation. Respondents were also asked

> '(H) If we were to give you as many coupons as you wanted, how many of these drug users [repeat the number in F] do you think you could give a coupon to by this time tomorrow?
> (I) If we were to give you as many coupons as you wanted, how many of these drug users [repeat the number in F] do you think you could give a coupon to by this time next week?'.

During the follow-up visit, the series of four main degree questions ((D), (E), (F) and (G)) was repeated, and respondents were also asked how quickly they distributed each of their coupons. We use these responses, along with data on the number of days between recruiter and recruit interviews, to evaluate three features of the degree question: validity of the 1-week timeframe that was used in question (G), test–retest reliability of responses and the possible effect of inconsistent reporting on estimates.

7.1. Validity of time window

A timeframe of 1 week was used in the key degree question (G) because previous qualitative experience with RDS suggested that most coupons were distributed within that time. In the on-line supporting information, therefore, we provide detailed examination of recruitment time dynamics and conclude that respondents reported that a high proportion (92%) of their alters could be reached within 1 week (Fig. S9 in the on-line supporting information), respondents

reported distributing most (95%) of their coupons within 1 week (Fig. S10) and the number of days between the interviews of the recruiter and recruit was usually less than 1 week (79% of the time; Fig. S11). We conclude, therefore, that for these studies the 1-week time window was reasonable.

7.2. Test–retest reliability

For participants who completed a follow-up survey (about half the participants; see Fig. 2), we have a measure of the consistency, but not accuracy, of their degree responses. The median difference between degree at the initial and follow-up visits was 0 (Fig. S12 in the on-line supporting information) suggesting that there was nothing systematic about the two visits that led to different answers on the questionnaire (e.g. different location or different length of interview). However, the responses of many individuals differed, in some cases substantially. The association between the measurements is affected by a small number of outliers, so we use the more robust Spearman's rank correlation to measure the association between the visits. The rank correlations range from 0.17 to 0.47 with a median correlation for the FSW of 0.33 and a median correlation for the DU and MSM of 0.41 (Fig. S13(a)). Although these low measures of test–retest reliability match our expectation, they differ substantially from a recent study of FSW in Shanghai, China, which found a test–retest reliability of network size of 0.98 (Yamanis *et al.*, 2013). Further research will be needed to understand the source of this difference.

7.3. Effect on estimates

Finally, we studied the robustness of our estimates by calculating disease prevalence estimates by using degree as measured in the initial and follow-up interviews, for those responding to both surveys (Fig. 8). The differences in disease prevalence estimates are generally small in an absolute sense, ranging from 0 to 0.08 (8 percentage points) with a median difference of 0.01. When broken down by disease, the HIV estimate had the largest median absolute difference, 0.031, followed by that for syphilis, 0.017. The hepatitis B and C estimates had median absolute differences in prevalence of essentially 0, possibly driven by the fact that these diseases are very rare in these populations. Note that these differences probably slightly overestimate the sensitivity of RDS estimates, as these estimates are restricted to respondents who completed both surveys, and so have smaller sample sizes.

In addition to comparing these differences in absolute units, we also consider the differences in relative units, $(|\hat{p} - \hat{p}'|)/\hat{p}$. The difference between the two estimates is more than 50% of the original estimate in about a quarter of the cases. In public health disease surveillance, an estimated increase in disease prevalence of 50% is probably a cause for concern, even if the estimated prevalences themselves were quite low. These data show that measurement error with respect to degree could introduce a change this large when prevalence is low.

7.4. Current recommendations

When collecting data, the time period that is used to elicit self-reported degree should be reflective of the time in which coupons are likely to be distributed. These results suggest that the 1-week period which was used in these studies was reasonable, but this should be checked in future studies with different populations. We also recommend that researchers collect degree at both the initial and the follow-up visits to assess test–retest reliability. Finally, in future studies, when considering which measure of degree to use, it is important to recall that these measures are being used to approximate the relative probability of inclusion of respondents. To the extent that there are other things about a respondent, such as social class or geographic location, that make him or her more or less likely to participate, the probability of inclusion is no longer

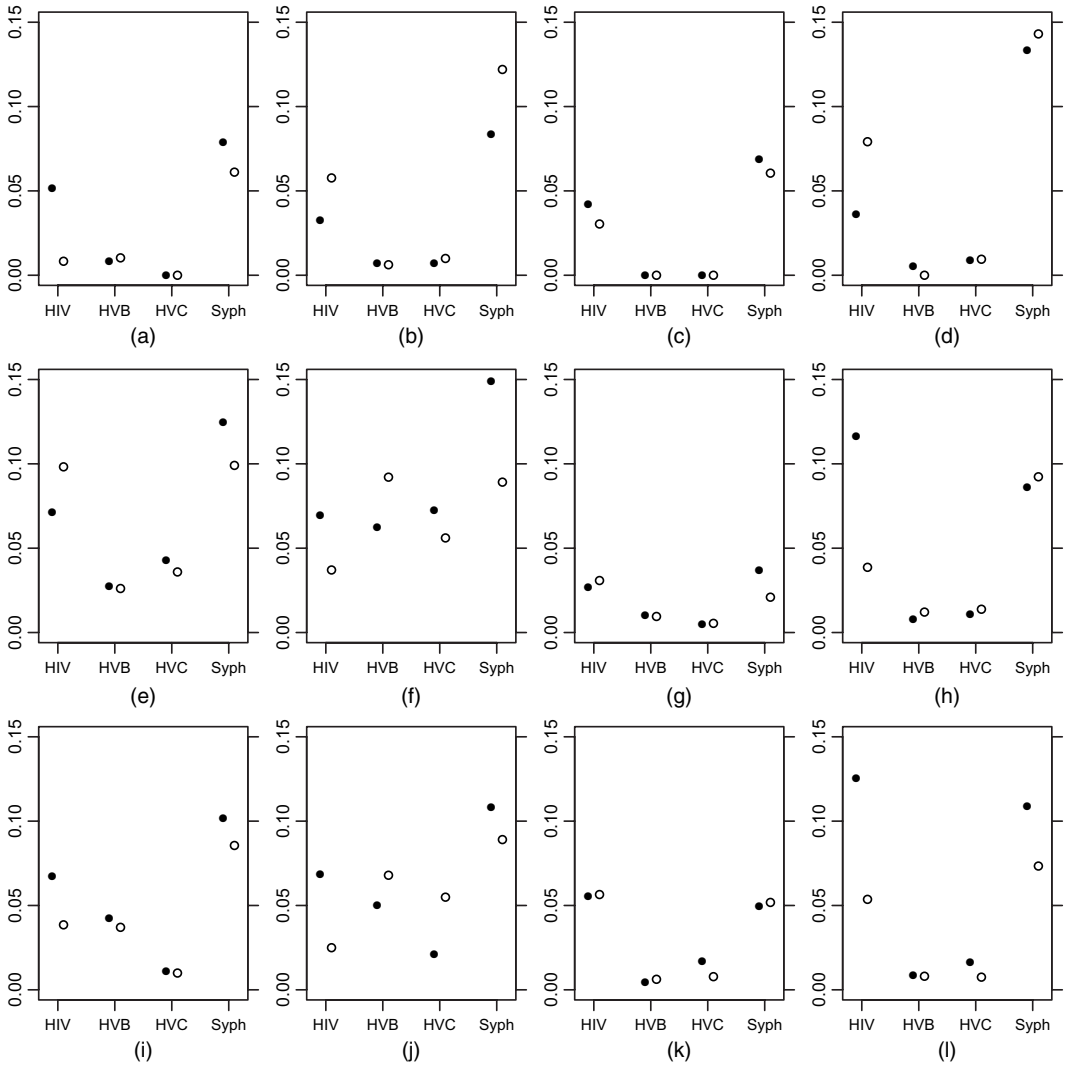


Fig. 8. Disease prevalence estimates from 12 studies for four diseases, using question (G) at enrolment (●) and follow-up (○) (the plots include only people who participated in both the initial and the follow-up survey; see Fig. 2 for sample sizes): (a) FSW, SD; (b) FSW, SA; (c) FSW, BA; (d) FSW, HI; (e) DU, SD; (f) DU, SA; (g) DU, BA; (h) DU, HI; (i) MSM, SD; (j) MSM, SA; (k) MSM, BA; (l) MSM, HI

proportional to degree. Any other features that are thought to be related to the probability of particiption should also be collected.

8. Participation bias

RDS estimation relies on the assumption that recruits represent a simple random sample from the contacts of each recruiter. Limited ethnographic evidence, however, suggests that recruitment decisions can be substantially more complex than is assumed in standard RDS statistical models (Scott, 2008; Ouellet, 2008; Broadhead, 2008; Bengtsson and Thorson, 2010; Kerr *et al.*, 2011; McCreesh *et al.*, 2012, 2013). For example, a study of MSM in Brazil found

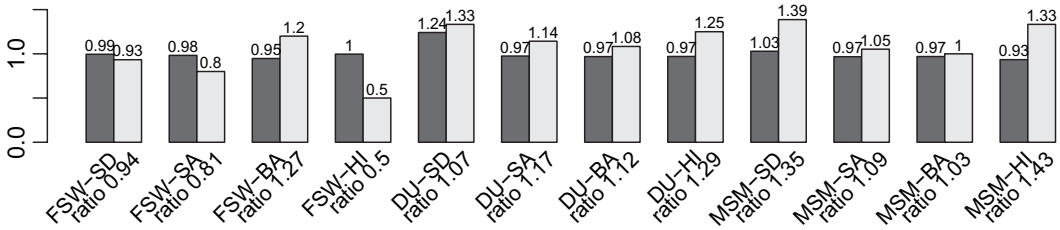


Fig. 9. Recruitment effectiveness plot: average recruits for HIV positve (■) and HIV negative (□) respondents by site (the ratio is provided under the bars; differential recruitment effectiveness can lead to bias in the RDS estimates under some conditions (Tomas and Gile, 2011))

that some people tended to recruit their riskiest friends because they were thought to need safe sex counselling (de Mello *et al.*, 2008). Further, the same study found that some MSM refused to participate when recruited because they were worried about revealing their sexual orientation. Such selective recruitment and participation could lead to non-response bias.

We find it helpful to consider the process of a new person entering the sample as the product of three decisions:

(a) the decision by the recruiter to pass coupons (how many and to whom);
(b) the decision by the recruit to accept a coupon;
(c) the decision by the recruit to participate in the study given that they have accepted a coupon.

Biases at any of these steps could result in systematic overrepresentation or underrepresentation of certain subgroups in the sample, resulting in biased estimates. We assess these possible biases in four ways. The first two, recruitment effectiveness and recruitment bias, address the cumulative effects of all three decisions on the quantity and characteristics of recruits; the third addresses two forms of non-response corresponding to steps (b) and (c); and the final analysis examines a respondent's motivation for participation.

8.1. Recruitment effectiveness

Systematic differences in recruitment effectiveness can lead to biased estimates under some conditions (Tomas and Gile, 2011). For example, if respondents with HIV have systematically more recruits and are also more likely to have contact with others with HIV, then people with HIV will be overrepresented in the sample. In Fig. 9, we present the mean numbers of recruits by HIV status for each site. We call this plot a *recruitment effectiveness plot*. In a single study, paired bars might represent differential recruitment effectiveness by many traits. In these 12 studies, the most dramatic difference is among the FSW in Higuey, where respondents with HIV recruit at only half the rate of those without HIV.

8.2. Recruitment bias

Recruitment bias—when a respondent's contacts have unequal probabilities of selection—can result in a pool of recruits that is systematically different from the pool of respondents' contacts. Because existing inferential methods assume that recruits are a simple random sample from among contacts, these systematic differences may bias resulting estimates (Gile and Handcock, 2010; Tomas and Gile, 2011). To examine the effects of such biases on the sample composition of a specific trait, employment status, we introduced the following questions in the questionnaires for the DU (for a related approach, see Yamanis *et al.* (2013)):

'(J) How many of them [repeat number of contacts in question F] are currently working?
(K) (follow-up questionnaire): Do the persons to whom you gave the coupons have work? (asked separately for each of 1 to 3 persons).
(L) Are you actually working? (we consider responses given by recruits of each respondent)'.

Overall, then, these questions, in order, should measure the employment characteristics of the pool of potential recruits, the employment characteristics of those who were chosen for referral by the respondents and accepted coupons, and the employment characteristics of those who then chose to return the coupons and to enrol in the study. The difference between the characteristics reported in the first question (J) and second question (K) reflect the joint effects of the decisions to pass and accept coupons, whereas the difference in characteristics between the second question (K) and third question (L) reflect the effect of the decision to participate in the interview.

Fig. 10, which we call a *recruitment bias plot*, provides a summary of the responses to these questions. This plot compares the composition of comparable sets of respondents' social contacts, coupon recipients and recruits. To do this, we restrict analysis to the set S of recruiters with data available on all three levels, and then calculate the average percentage of contacts, coupon recipients and recruits who are employed as follows:

$$\frac{1}{|S|}\sum_{i\in S}\frac{J_i}{F_i},$$

$$\frac{1}{|S|}\sum_{i\in S}\frac{\sum_{j=1}^{n_i^{\mathrm{c}}} K_{ij}}{n_i^{\mathrm{c}}}$$

and

$$\frac{1}{|S|}\sum_{i\in S}\frac{\sum_{j\in \mathrm{Recruits}(i)} L_j}{\sum_{j\in \mathrm{Recruits}(i)} 1},$$

where F_i, J_i and L_i refer to respondent i's response to questions (F), (J) and (L), K_{ij} is a binary indicator of i's report of the employment status of the person receiving his or her jth coupon and n_i^{c} is the number of coupons that i reported distributing.

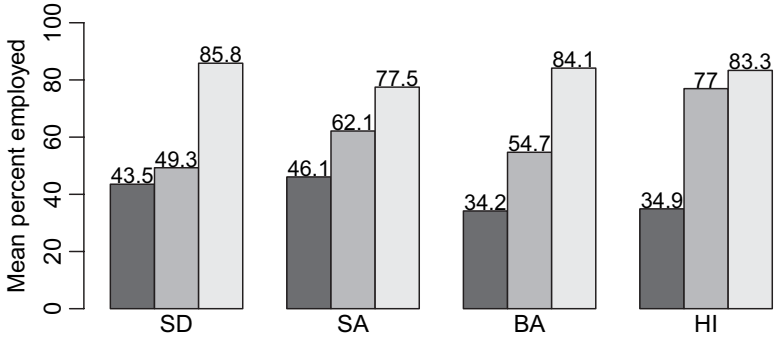


Fig. 10. Recruitment bias plot: percentage of drug users employed, by location and question (■, contacts; ■, coupons; ■, recruits)

In every site, there is a marked increase in the reported rate of employment for each stage in the referral process (Fig. 10). These data suggest that respondents distributing coupons are more likely to give them to those among their contacts who are employed, and that, among those receiving coupons, those who are employed are more likely to return them.

These results are a provocative suggestion of aberrant respondent behaviour and could belie a dramatic oversampling of employed DU. These particular results, however, should be seen in light of other possible explanations, in particular the possibility of survey response bias. The succession of questions, reflecting increased proportions of reported employment, also correspond to increasing social closeness to the respondent. Because it is possible that 'having work' is a desirable status, a response bias based on social desirability would also explain the results in this section.

Other researchers (e.g. Heckathorn *et al.* (2002), Wang *et al.* (2005), Wejnert and Heckathorn (2008), Iguchi *et al.* (2009), Rudolph *et al.* (2011), Liu *et al.* (2012) and Yamanis *et al.* (2013)) have introduced and used statistical tests assessing the assumption of random recruitment. We address these and introduce a new non-parametric test, in the on-line supporting information.

8.3. Non-response

Non-response, where intended respondents do not participate, is a problem in most surveys. If non-responders differ systematically from responders, estimates will suffer from non-response bias. Non-response and non-response bias are particularly challenging to measure in RDS studies because non-responders are contacted by other participants rather than by researchers and because non-response can arise in two ways—by refusing a coupon or by failing to return the coupon to participate in the study.

To understand non-response better, respondents were asked during the follow-up interview

> '(M) How many coupons did you distribute?
> (N) How many people did not accept a coupon you offered to them?'.

We estimate the *coupon refusal rate* R_{C} by comparing responses to question (M) and question (N), and we estimate the *non-return rate* R_{N} by comparing responses to question (M) with the number of survey participants presenting coupons from each respondent. Finally, comparing the number of respondents with the number of attempted eligible coupon distributions (refused and distributed) we estimate the *total non-response rate* R_{T}. Specifically, these rates are respectively computed as follows:

$$R_{\mathrm{C}} = \frac{\sum_{i \in S} n_i^{\mathrm{r}}}{\sum_{i \in \mathrm{S}} n_i^{\mathrm{r}} + \sum_{i \in \mathrm{S}} n_i^{\mathrm{c}}}, \tag{2}$$

$$R_{\mathrm{N}} = 1 - \frac{\sum_{i \in S} |\mathrm{Recruits}(i)|}{\sum_{i \in S} n_i^{\mathrm{c}}}, \tag{3}$$

$$R_{\mathrm{T}} = 1 - \frac{\sum_{i \in S} |\mathrm{Recruits}(i)|}{\sum_{i \in S} n_i^{\mathrm{r}} + \sum_{i \in S} n_i^{\mathrm{c}}} = 1 - (1 - R_{\mathrm{C}})(1 - R_{\mathrm{N}}), \tag{4}$$

where S is again restricted to those with data on all relevant questions, n_i^{c} is the number of coupons distributed by i, n_i^{r} is the number of refused coupons reported by i and $|\mathrm{Recruits}(i)|$

Table 4. RDS non-response rates†

Rate	*Results for the FSW*				*Results for the DU*				*Results for the MSM*			
	SD	*SA*	*BA*	*HI*	*SD*	*SA*	*BA*	*HI*	*SD*	*SA*	*BA*	*HI*
Coupon refusal, R_C	56.5	45.3	7.5	28.0	0.4	15.9	11.3	41.3	7.7	16.5	25.4	29.2
Non-return, R_N (%)	13.4	43.9	43.0	41.4	26.1	35.3	44.6	33.9	29.4	23.6	39.7	31.9
Total non-response, R_T (%)	62.3	69.3	47.2	57.8	26.3	45.6	50.9	61.2	34.8	36.3	55.0	51.8
Number of recruiters	123	136	141	151	126	105	164	141	153	128	152	102

†The coupon refusal rate is the total number of reported coupon refusals to eligible alters divided by that number plus the number of reported coupons distributed. Coupon non-return is the percentage of coupons that were not returned (among accepted coupons). The total non-response rate is the percentage of attempted recruitments of eligible alters not resulting in survey participation. All rates were computed on the basis of only recruits of respondents who completed the follow-up interview.

represents the number of successful recruits of i. In general, the problems of refusal, non-return and total non-response were slightly more serious among the FSW than DU or MSM (Table 4). Note that all these estimates may be underestimates of non-response, as respondents with none or fewer successful recruitments were less likely to complete the follow-up survey (see section S6 in the on-line supporting information).

To know whether this non-response could induce non-response bias, we would need to know whether the people who refused were different from those who participated. We could not collect information about non-responders directly so we asked recruiters why their non-respondents had refused coupons, as has been done in previous studies (Stormer *et al.*, 2006; Johnston, Khanam, Reza, Khan, Banu, Alam, Rahman and Azim, 2008; Iguchi *et al.*, 2009). For each of up to five refusals, the return survey asked

'(O) What is the principal reason why these persons did not accept a coupon?'.

Responses to question (O) are summarized in the *coupon refusal analysis* in Table 5. The most common reason given for refusal was aversion to being identified as a member of the target population (26.6%). Many refusers also reported fear of test results (especially HIV test results: 16.3%). Some were 'uninterested' (22.0%). Interestingly for study organizers, among the reasons for 'other', 5.2% of MSM refusers reportedly did not trust the study or did not believe that the incentive was true. For an alternative approach to collecting information about refusals, see Yamanis *et al.* (2013).

8.4. Decisions to accept coupon and to participate in study

In addition to exploring reasons for not participating in the study, we also asked about each respondent's reason for participating, as in Johnston, Khanam, Reza, Khan, Banu, Alam, Rahman and Azim (2008):

'(P) What is the principal reason why you decided to accept a coupon and participate in this study?'.

Responses are reported in Table 6. In every site, a substantial majority reported participating in the interest of receiving HIV test results. Further, we go beyond previous researchers and assess whether the motivation for participation is associated with important study outcomes. For example, we found that the odds of having HIV among those who expressed motivation based on the HIV test was from 0.43 (study MSM–HI) to 2.03 (study MSM–SA) times the

Table 5. Coupon refusal analysis: responses to the question 'What is the principal reason why these persons did not accept a coupon?'

Response	*Results for the FSW*				*Results for the DU*				*Results for the MSM*			
	SD	*SA*	*BA*	*HI*	*SD*	*SA*	*BA*	*HI*	*SD*	*SA*	*BA*	*HI*
Too busy	7.3	80.0	10.0	0.8	0.0	10.3	0.0	3.0	0.0	30.8	4.5	12.1
Fear being identified	31.4	0.0	63.3	21.3	100.0	17.9	22.4	20.5	4.2	30.8	30.3	31.3
Incentive low or location far	3.6	0.0	0.0	2.5	0.0	2.6	0.0	1.8	0.0	0.0	0.8	1.0
Not interested	26.3	0.0	10.0	38.5	0.0	2.6	10.2	15.1	0.0	30.8	30.3	19.2
Fear HIV or other results	15.3	0.0	6.7	28.7	0.0	15.4	20.4	10.2	75.0	7.7	16.7	4.0
Fear giving blood	0.7	0.0	0.0	0.8	0.0	33.3	0.0	22.9	0.0	0.0	0.0	7.1
Fail eligibility	0.0	0.0	10.0	0.8	0.0	0.0	4.1	6.0	0.0	0.0	3.8	2.0
Already got coupon	1.5	20.0	0.0	0.0	0.0	0.0	2.0	0.6	0.0	0.0	10.6	0.0
Other	13.9	0.0	0.0	6.6	0.0	17.9	40.8	19.9	20.8	0.0	3.0	23.2
Total reasons reported	137	5	30	122	1	39	49	166	24	13	132	99

Table 6. Responses to question (P), 'What is the principal reason why you decided to accept a coupon and participate in this study?'†

Response	*Results for the FSW*				*Results for the DU*				*Results for the MSM*			
	SD	*SA*	*BA*	*HI*	*SD*	*SA*	*BA*	*HI*	*SD*	*SA*	*BA*	*HI*
Incentive	2.9	5.0	3.3	1.3	11.6	16.5	6.0	5.0	5.7	5.8	1.8	5.9
For HIV test	88.5	77.7	90.1	86.4	51.3	63.2	65.1	70.1	71.5	71.9	81.5	83.3
Other or all test	1.0	2.3	0.4	2.6	18.4	0.6	4.7	3.0	1.4	1.2	5.0	1.9
Recruiter	1.7	5.0	1.6	2.0	3.9	10.3	6.3	17.6	7.7	5.8	4.6	4.8
Study interest	4.6	10.0	4.1	7.3	10.0	8.7	17.9	4.3	11.3	14.7	5.0	3.7
Other	1.2	0.0	0.4	0.3	4.8	0.6	0.0	0.0	2.4	0.6	2.1	0.4
Total	410	301	243	302	310	310	301	301	505	327	281	269

†The 'other' category includes 'I have free time', 'To stop using' (DU only) and 'Other'.

odds for those who did not. Similar relationships hold when analyses are restricted to those who have not had an HIV test in the last 3 months or last 6 months. We summarize these results in the *motivation–outcome plot* in Fig. 11. The unknown dependence structure does not allow for formal statistical testing; however, note that to the extent that the probability of participation is associated with participant motivation and participant motivation is associated with an outcome of interest, bias will be introduced in the estimates even if these associations are not statistically significant.

8.5. Current recommendations

The approaches in this section do not directly indicate the extent to which estimates may be impacted by the various forms of participation bias. Our approaches for measuring and monitoring potential sources of participation bias, instead, were developed in the interest of

(a) adjusting the sampling process,

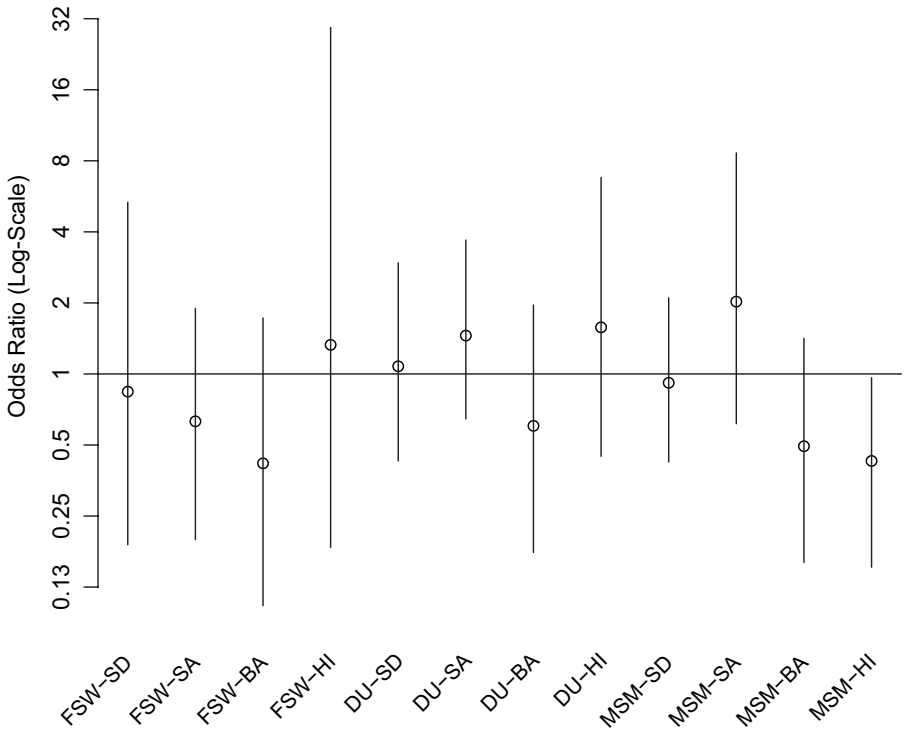


Fig. 11. Motivation–outcome plot: odds ratios of having HIV given HIV test motivation for study participation (ratios greater than 1 indicate those participating for HIV test results are more likely to have HIV; for reference, nominal 95% intervals are based on the inversion of Fisher's exact test (these would be confidence intervals if the data were independent identically distributed))

(b) informing the choice of an estimator or
(c) informing the development of new approaches to inference.

Ideally, the quantitative survey-based approaches that are presented here should be paired with qualitative evaluation of decision making associated with recruitment and participation (Scott, 2008; Broadhead, 2008; Ouellet, 2008; de Mello *et al.*, 2008; Kerr *et al.*, 2011; McCreesh *et al.*, 2012, 2013).

Fortunately, differential recruitment effectiveness is possible to evaluate by using data that are readily available in all RDS studies and is directly actionable in terms of estimators. Recruitment effectiveness plots should be made to study the relationship between recruitment and key study variables, both during and after data collection. Where differences are found, qualitative study or discussion with survey staff may reveal areas for improvement in the sampling process. Further, these findings may influence the choice of estimators. Tomas and Gile (2011) showed that the estimator in Salganik and Heckathorn (2004) is more robust to differential recruitment effectiveness than other estimators are. The newer estimator of Gile and Handcock (2011) allows researchers to adjust for differential recruitment effectiveness by outcome and wave of the sample.

Recruitment biases are more difficult to evaluate, in part because they require more specialized data collection. The particular characteristics of interest, such as employment status, will be study specific and require researchers to be very familiar with the population and sampling process. Any characteristic that may be associated with increased participation should

be measured for respondents, potential respondents (i.e. contacts of respondents) and coupon recipients so that researchers can create recruitment bias plots (Fig. 10). The collection of such data may also inspire further development of statistical inference for RDS data. The relationship between drug user employment and participation is a good example. If employed alters are indeed more likely to be sampled, and if this tendency can be measured (as in these data), methods may be developed to adjust inference for this tendency. The estimators in Gile and Handcock (2011) and Lu (2013) are particularly conducive to this kind of adjustment.

A thorough evaluation of non-response bias requires a follow-up study of non-responders. Despite the obvious logistical challenges (see de Mello *et al.* (2008), Kerr *et al.* (2011) and McCreesh *et al.* (2012)), we recommend such a study whenever researchers have special concerns about non-response. Absent such a study, computing RDS non-response rates could alert researchers to possible problems with non-response. A coupon refusal analysis could then help researchers to adjust their studies to remove barriers to participation. Further, the results of a coupon refusal analysis could suggest individual characteristics that might be related to non-response and which, therefore, should be measured. For example, if distance to the study site seems burdensome, researchers could introduce an additional study site, or, minimally, collect data on a measure of distance burden either to adjust estimators or to monitor recruitment bias.

Participant motivations should also be measured in all studies, as an indication of potential differential valuation of incentives by different subpopulations. Motivation–outcome relationships can be studied between any combinations of expressed motivations and relevant respondent characteristics. Mechanisms to adjust inference for biases that are introduced by measurable differential incentives to participation, such as those due to interest in HIV test results, are not yet developed. The precise quantification of these effects and their effects on inference are an important area for future research.

9. Discussion

RDS is designed to enact a near statistical miracle: beginning with a convenience sample, selecting subsequent samples dependent on previous samples, then treating the final sample as a probability sample with known (or estimable) inclusion probabilities. This is in contrast with traditional survey samples, where sampling is intended to be conducted from a well-defined sampling frame according to sampling procedures fully controlled by the researcher.

Miracles do not come for free and, where alternative workable strategies are available, RDS is often not advisable. Unfortunately, alternative approaches are unavailable for many populations of interest. Therefore, researchers need to be aware of two main costs of RDS: large variance of estimates (see, for example Goel and Salganik (2010), Szwarcwald *et al.* (2011), Wejnert *et al.* (2012), Mouw and Verdery (2012) and Johnston, Chen, Silva-Santisteban and Raymond (2013)) and many assumptions, including those considered in this paper. Unfortunately, these assumptions are difficult to assess with certainty in real hard-to-reach populations when the sampling is largely conducted by respondents outside the view of researchers. Rather than attempting to provide definitive critical values for statistical tests, which would themselves rest on numerous untested assumptions, we have provided a broad set of intuitive tools to allow RDS researchers to understand better the processes that generated their data.

For researchers planning future data collection, we briefly summarize our current recommendations in Table 7. We have made some of these methods available in the R (R Core Team,

Table 7. Summary of recommendations

Before data collection
Formative research
Add the following questions to the initial survey:
questions to assess finite population effects on sampling (e.g. question (B))
questions to assess validity in time window of degree questions (e.g. questions (H) and (I))
questions to asses reciprocity (e.g. question (C) and those in Heckathorn (2002))
questions to assess recruitment bias (e.g. questions (J) and (L))
questions for motivation–outcome plot (e.g. question (P))
Add the following questions to the follow-up survey
questions to assess finite population effects on sampling (e.g. question (A))
questions to assess recruitment bias (e.g. question (K))
questions to assess non-response (e.g. questions (M) and (N))
questions for coupon refusal analysis (e.g. question (O))
During data collection
For all traits of interest, create convergence plots, bottleneck plots and all points plots (Section 5)
Create recruitment effectiveness plots (Fig. 9)
Create recruitment bias plots (Fig. 10)
Calculate the reciprocation rate (Table 3)
Check the validity of the timeframe used in the degree question (Section 7.1)
Calculate the non-response rates (Table 4)
Conduct a coupon refusal analysis (Table 5)
Conduct motivation–outcome analysis (Fig. 11)
After data collection
Assess finite population effects in data collection and estimates (Section 4)
Calculate test–retest reliability of the degree question (Section 7.2)

2012) packages `RDS` (Handcock *et al.*, 2009) and `RDS Analyst` (Handcock *et al.*, 2013). We emphasize that these diagnostics should continue to be refined and improved as more is learned about RDS sampling and as new estimators are developed. In fact, we hope that this paper will stimulate just such research.

Acknowledgements

We are grateful to Tessie Caballero Vaillant, El Consejo Presidencial del SIDA, Dominican Republic, for allowing us to use these data. We also thank Maritza Molina Achécar, Juan Jose Polanco and Sonia Baez of El Centro de Estudios Sociales y Demograficos, Dominican Republic, for overseeing data collection, and Chang Chung, Sharad Goel, Mark Handcock, Doug Heckathorn, Martin Klein, Dhwani Shah and Cyprian Wejnert for helpful discussions. Finally, we thank all those who participated in these studies. The research reported in this publication was supported by grants from the National Institutes of Health–National Institute of Child Health and Development (R01-HD062366 and R24-HD047879), the National Institutes of Health (R21-A604273) and the National Science Foundation (CNS-0905086 and SES-1230081), including support from the National Agricultural Statistics Service. The content is solely the responsibility of the authors.

The authorship is alphabetical; all the authors contributed equally to the paper.

References

Barbosa Júnior, A., Pati Pascom, A. R., Szwarcwald, C. L., Kendall, C. and McFarland, W. (2011) Transfer of sampling methods for studies on most-at-risk populations (MARPs) in Brazil. *Cad. Sde Publ.*, **27**, suppl. S1, S36–S44.

Bengtsson, L., Lu, X., Nguyen, Q. C., Camitz, M., Hoang, N. L., Liljeros, F. and Thorson, A. (2012) Implementation of web-based respondent-driven sampling among men who have sex with men in Vietnam. *PLOS ONE*, **7**, no. 11, article e49417.

Bengtsson, L. and Thorson, A. (2010) Global HIV surveillance among MSM: is risk behavior seriously underestimated? *AIDS*, **24**, 2301–2303.

Bernard, H. R., Hallett, T., Iovita, A., Johnsen, E. C., Lyerla, R., McCarty, C., Mahy, M., Salganik, M. J., Saliuk, T., Scutelniciuc, O., Shelley, G. A., Sirinirund, P., Weir, S. and Stroup, D. F. (2010) Counting hard-to-count populations: the network scale-up method for public health. *Sexlly Transmttd Infect.*, **86**, suppl. 2, ii11–ii15.

Bernard, H. R., Killworth, P., Kronenfeld, D. and Sailer, L. (1984) The problem of informant accuracy: the validity of retrospective data. *A. Rev. Anthrop.*, **13**, 495–517.

Borgatti, S. P. (2002) NetDraw software for network visualization. *Technical Report.* Analytic Technologies, Lexington.

Brewer, D. D. (2000) Forgetting in the recall-based elicitation of personal and social networks. *Socl Netwrks*, **22**, 29–43.

Broadhead, R. S. (2008) Notes on a cautionary (tall) tale about respondent-driven sampling: a critique of Scott's ethnography. *Int. J. Drug Poly*, **19**, 235–237.

Burt, R. D. and Thiede, H. (2012) Evaluating consistency in repeat surveys of injection drug users recruited by respondent-driven sampling in the Seattle area: results from the NHBS-IDU1 and NHBS-IDU2 surveys. *Ann. Epidem.*, **22**, 354–363.

Frost, S. D., Brouwer, K. C., Firestone Cruz, M. A., Ramos, R., Ramos, M. E., Lozada, R. M., Magis-Rodriguez, C. and Strathdee, S. A. (2006) Respondent-driven sampling of injection drug users in two US–Mexico border cities: recruitment dynamics and impact on estimates of HIV and Syphilis prevalence. *J. Urb. Hlth*, **83**, 83–97.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–472.

Gile, K. J. (2011) Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J. Am. Statist. Ass.*, **106**, 135–146.

Gile, K. J. and Handcock, M. S. (2010) Respondent-driven sampling: an assessment of current methodology. *Sociol. Methodol.*, **40**, 285–327.

Gile, K. J. and Handcock, M. S. (2011) Network model-assisted inference from respondent-driven sampling data. *Preprint arXiv:1108.0298.*

Goel, S. and Salganik, M. J. (2009) Respondent-driven sampling as Markov chain Monte Carlo. *Statist. Med.*, **28**, 2202–2229.

Goel, S. and Salganik, M. J. (2010) Assessing respondent-driven sampling. *Proc. Natn. Acad. Sci. USA*, **107**, 6743–6747.

Handcock, M. S. (2011) size: estimating hidden population size using respondent driven sampling data. *R Package Version 0.20*.

Handcock, M. S., Fellows, I. E. and Gile, K. J. (2013) *RDS Analyst: Analysis of Respondent-driven Sampling Data*, version 1.0. Los Angeles.

Handcock, M. S., Gile, K. J. and Mar, C. M. (2012) Estimating hidden population size using respondent-driven sampling data. *Working Paper.*

Handcock, M. S., Gile, K. J. and Neely, W. W. (2009) RDS: R functions for respondent-driven sampling. *R Package Version 0.10*.

Heckathorn, D. D. (1997) Respondent-driven sampling: a new approach to the study of hidden populations. *Socl Prob.*, **44**, 174–199.

Heckathorn, D. D. (2002) Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Socl Prob.*, **49**, 11–34.

Heckathorn, D., Semaan, S., Broadhead, R. and Hughes, J. (2002) Extensions of respondent-driven sampling: a new approach to the study of injection drug users aged 18-25. *AIDS Behav.*, **6**, 55–67.

Heimer, R. (2005) Critical issues and further questions about respondent-driven sampling: comment on Ramirez-Valles, *et al*. (2005). *AIDS Behav.*, **9**, 403–408.

Iguchi, M. Y., Ober, A. J., Berry, S. H., Fain, T., Heckathorn, D. D., Gorbach, P. M., Heimer, R., Kozlov, A., Ouellet, L. J., Shoptaw, S. and Zule, W. A. (2009) Simultaneous recruitment of drug users and men who have sex with men in the United States and Russia using respondent-driven sampling: sampling methods and implications. *J. Urb. Hlth*, **86**, suppl. 1, 5–31.

Johnston, L. G. (2008) Introduction to respondent-driven sampling. *Technical Report.* (Available from `http://bit.ly/OHgGX5`.)

Johnston, L. G. (2013) Introduction to respondent-driven sampling. *Technical Report.* World Health Organization, Geneva.

Johnston, L. G., Caballero, T., Dolores, Y. and Values, H. M. (2013) HIV, Hepatitis B/C and Syphilis prevalence and risk behaviors among gay/trans/men who have sex with men, Dominican Republic. *Int. J. STD AIDS*, **24**, 313–321.

Johnston, L. G., Chen, Y.-H., Silva-Santisteban, A. and Raymond, H. F. (2013) An empirical examination of respondent driven sampling design effects among HIV risk groups from studies conducted around the world. *AIDS Behav.*, **17**, 2202–2210.

Johnston, L. G., Khanam, R., Reza, M., Khan, S. I., Banu, S., Alam, M. S., Rahman, M. and Azim, T. (2008) The effectiveness of respondent driven sampling for recruiting males who have sex with males in Dhaka, Bangladesh. *AIDS Behav.*, **12**, 294–304.

Johnston, L. G., Malekinejad, M., Kendall, C., Iuppa, I. M. and Rutherford, G. W. (2008) Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: field experiences in international settings. *AIDS Behav.*, **12**, suppl. 1, 131–141.

Johnston, L. G., Prybylski, D., Raymond, H. F., Mirzazadeh, A., Manopaiboon, C. and McFarland, W. (2013) Incorporating the service multiplier method in respondent-driven sampling surveys to estimate the size of hidden and hard-to-reach populations. *Sexlly Transmttd Dis.*, **40**, 304–310.

Kerr, L. R. F. S., Kendall, C., Pontes, M. K., Werneck, G. L., McFarland, W., Mello, M. B., Martins, T. A. and Macena, R. H. M. (2011) Selective participation in a RDS survey among MSM in Ceara, Brazil: a qualitative and quantitative assessment. *J. Bras. Doenc. Sexmnte Transmiss.*, **23**, 126–133.

Lansky, A., Abdul-Quader, A. S., Cribbin, M., Hall, T., Finlayson, T. J., Garfein, R. S., Lin, L. S. and Sullivan, P. S. (2007) Developing an HIV behavioral surveillance system for injecting drug users: the National HIV Behavioral Surveillance System. *Publ. Hlth Rep.*, **122**, suppl. 1, 48–55.

Lansky, A., Drake, A., Wejnert, C., Pham, H., Cribbin, M. and Heckathorn, D. D. (2012) Assessing the assumptions of respondent-driven sampling in the National HIV Behavioral Surveillance System among injecting drug users. *Open AIDS J.*, **6**, 77–82.

Liu, H., Li, J., Ha, T. and Li, J. (2012) Assessment of random recruitment assumption in respondent-driven sampling in egocentric network data. *Socl Netwrkng*, **1**, 13–21.

Lu, X. (2013) Linked ego networks: improving estimate reliability and validity with respondent-driven sampling. *Socl Netwrks*, **35**, 669–685.

Lu, X., Bengtsson, L., Britton, T., Camitz, M., Kim, B. J., Thorson, A. and Liljeros, F. (2012) The sensitivity of respondent-driven sampling. *J. R. Statist. Soc.* A, **175**, 191–216.

Lu, X., Malmros, J., Liljeros, F. and Britton, T. (2013) Respondent-driven sampling on directed networks. *Electron. J. Statist.*, **7**, 292–322.

Magnani, R., Sabin, K., Saidel, T. and Heckathorn, D. (2005) Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS*, **19**, suppl., S67–S72.

Malekinejad, M., Johnston, L. G., Kendall, C., Kerr, L. R., Rifkin, M. R. and Rutherford, G. W. (2008) Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: a systematic review. *AIDS Behav.*, **12**, 105–130.

Marsden, P. V. (1990) Network data and measurement. *A. Rev. Sociol.*, **16**, 435–463.

McCreesh, N., Frost, S. D. W., Seeley, J., Katongole, J., Tarsh, M. N., Ndunguse, R., Jichi, F., Lunel, N. L., Maher, D., Johnston, L. G., Sonnenberg, P., Copas, A. J., Hayes, R. J. and White, R. G. (2012) Evaluation of respondent-driven sampling. *Epidemiology*, **23**, 138–147.

McCreesh, N., Tarsh, M. N., Seeley, J., Katongole, J. and White, R. G. (2013) Community understanding of respondent-driven sampling in a medical research setting in Uganda: importance for the use of RDS for public health research. *Int. J. Socl Res. Methodol.*, **16**, 269–284.

de Mello, M., de Pinho, A. A., Chinaglia, M., Tun, W., Júnior, A. B., Ilário, M. C. F. J., Reis, P., Salles, R. C. S., Westman, S. and Díaz, J. (2008) Assessment of risk factors for HIV infection among men who have sex with men in the metropolitan area of Campinas City, Brazil, using respondent-driven sampling. *Technical Report.* Population Council, New York.

Mills, H., Colijn, C., Vickerman, P., Leslie, D., Hope, V. and Hickman, M. (2012) Respondent driven sampling and community structure in a population of injecting drug users, Bristol, UK. *Drug Alc. Depend.*, **126**, 324–332.

Montealegre, J. R., Johnston, L. G., Murrill, C. and Monterroso, E. (2013) Respondent driven sampling for HIV biological and behavioral surveillance in Latin America and the Caribbean. *AIDS Behav.*, **17**, 2313–2340.

Mouw, T. and Verdery, A. M. (2012) Network sampling with memory: a proposal for more efficient sampling from social networks. *Sociol. Methodol.*, **42**, 206–256.

Nesterko, S. and Blitzstein, J. (2014) Bias-variance and breadth-depth tradeoffs in respondent-driven sampling. *J. Statist. Computn Simuln*, to be published.

Ouellet, L. J. (2008) Cautionary comments on an ethnographic tale gone wrong. *Int. J. Drug Poly*, **19**, 238–240.

Poon, A. F. Y., Brouwer, K. C., Strathdee, S. A., Firestone-Cruz, M., Lozada, R. M., Kosakovsky Pond, S. L., Heckathorn, D. D. and Frost, S. D. W. (2009) Parsing social network survey data from hidden populations using stochastic context-free grammars. *PLOS ONE*, **4**, article e6777.

Raftery, A. E. and Lewis, S. (1992) How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 763–773. Oxford: Oxford University Press.

Ramirez-Valles, J., Heckathorn, D. D., Vázquez, R., Diaz, R. M. and Campbell, R. T. (2005a) From networks to populations: the development and application of respondent-driven sampling among IDUs and Latino gay men. *AIDS Behav.*, **9**, 387–402.

Ramirez-Valles, J., Heckathorn, D. D., Vázquez, R., Diaz, R. M. and Campbell, R. T. (2005b) The fit between theory and data in respondent-driven sampling: Response to Heimer. *AIDS Behav.*, **9**, 409–414.
R Core Team (2012) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rudolph, A. E., Crawford, N. D., Latkin, C., Heimer, R., Benjamin, E. O., Jones, K. C. and Fuller, C. M. (2011) Subpopulations of illicit drug users reached by targeted street outreach and respondent-driven sampling strategies: implications for research and public health practice. *Ann. Epidem.*, **21**, 280–289.
Rudolph, A. E., Fuller, C. M. and Latkin, C. (2013) The importance of measuring and accounting for potential biases in respondent-driven samples. *AIDS Behav.*, **17**, 2244–2252.
Salganik, M. J. (2012) Commentary: Respondent-driven sampling in the real world. *Epidemiology*, **23**, 148–150.
Salganik, M. J., Fazito, D., Bertoni, N., Abdo, A. H., Mello, M. B. and Bastos, F. I. (2011) Assessing network scale-up estimates for groups most at risk of HIV/AIDS: evidence from a multiple-method study of heavy drug users in Curitiba, Brazil. *Am. J. Epidem.*, **174**, 1190–1196.
Salganik, M. J. and Heckathorn, D. D. (2004) Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol. Methodol.*, **34**, 193–240.
Scott, G. (2008) "They got their program, and I got mine": a cautionary tale concerning the ethical implications of using respondent-driven sampling to study injection drug users. *Int. J. Drug Poly*, **19**, 42–51.
Stormer, A., Tun, W., Guli, L., Harxhi, A., Bodanovskaia, Z., Yakovleva, A., Rusakova, M., Levina, O., Bani, R., Rjepaj, K. and Bino, S. (2006) An analysis of respondent driven sampling with injection drug users (IDU) in Albania and the Russian Federation. *J. Urb. Hlth*, **83**, 73–82.
Szwarcwald, C. L., de Souza Júnior, P. R. B., Damacena, G. N., Junior, A. B. and Kendall, C. (2011) Analysis of data collected by RDS among sex workers in 10 Brazilian cities, 2009: estimation of the prevalence of HIV, variance, and design effect. *J. Acq. Immune Defic. Synd.*, **57**, suppl., S129–S135.
Tomas, A. and Gile, K. J. (2011) The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electron. J. Statist.*, **5**, 899–934.
UNAIDS (2010) *Guidelines on Estimating the Size of Populations Most at Risk to HIV.* Geneva: UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance.
Volz, E. and Heckathorn, D. D. (2008) Probability based estimation theory for respondent driven sampling. *J. Off. Statist.*, **24**, 79–97.
Wang, J., Carlson, R. G., Falck, R. S., Siegal, H. A., Rahman, A. and Li, L. (2005) Respondent-driven sampling to recruit MDMA users: a methodological assessment. *Drug Alc. Depend.*, **78**, 147–157.
Wejnert, C. (2009) An empirical test of respondent-driven sampling: point estimates, variance, degree measures, and out-of-equilibrium data. *Sociol. Methodol.*, **39**, 73–116.
Wejnert, C. and Heckathorn, D. D. (2008) Web-based network sampling efficiency and efficacy of respondent-driven sampling for online research. *Sociol. Meth. Res.*, **37**, 105–134.
Wejnert, C., Pham, H., Krishna, N., Le, B. and DiNenno, E. (2012) Estimating design effect and calculating sample size for respondent-driven sampling studies of injection drug users in the United States. *AIDS Behav.*, **16**, 797–806.
White, R. G., Lansky, A., Goel, S., Wilson, D., Hladik, W., Hakim, A. and Frost, S. D. (2012) Respondent driven sampling—where we are and where should we be going? *Sexlly Transmttd Infectns*, **88**, 397–399.
Yamanis, T. J., Merli, M. G., Neely, W. W., Tian, F. F., Moody, J., Tu, X. and Gao, E. (2013) An empirical analysis of the impact of recruitment patterns on RDS estimates among a socially ordered population of female sex workers in China. *Sociol. Meth. Res.*, **42**, 392–425.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supporting information: Diagnostics for respondent-driven sampling'.

Supporting Information: Diagnostics for Respondent-driven Sampling

Krista J. Gile
University of Massachusetts, Amherst, MA, USA
Lisa G. Johnston
Tulane University, New Orleans, LA, USA
and University of California, San Francisco, San Francisco, CA, USA
Matthew J. Salganik
Microsoft Research, New York, NY USA
and Princeton University, Princeton, NJ, USA
**Authorship alphabetical; all authors contributed equally to the paper.*

S1. With-replacement Sampling

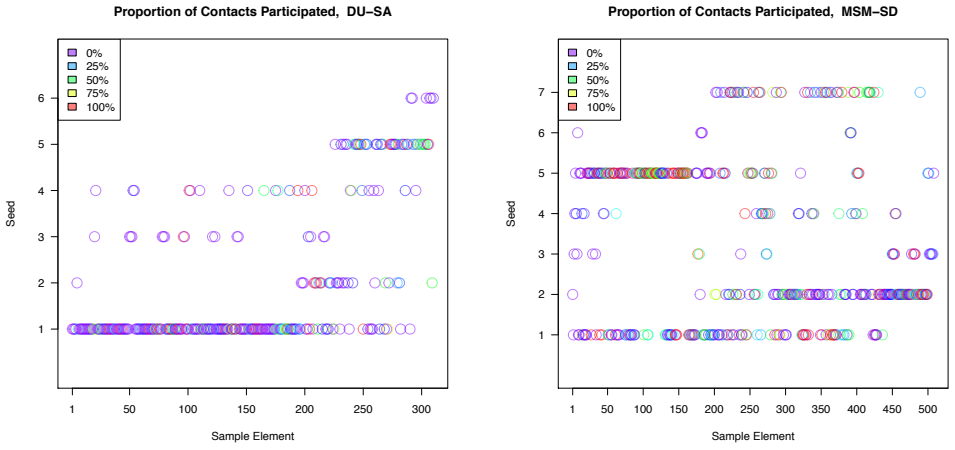
S1.1. Multiple Connections to Survey Participants

In Section 4.3 we presented results about the proportion of respondents' contacts who had already participated in the study. It may also be of interest to visualize these trends. Figs. S1(a) and S1(b) show the reported proportions that already participated for each respondent, by seed, over time. In Fig. S1(a), we can see that within seed, particularly seed 1, periods of low proportion already sampled are often followed by periods of higher proportion already sampled. This may be indicative of the exhaustion of local subgroups. Fig. S1(b) shows less evidence of a positive trend in proportion already sampled over time. Finally, Fig. S2 shows the fitted linear trends for all 12 sites.

S1.2. Decreasing Degree over Time in Sample

Under a broad range of assumptions, link-tracing samples result in higher draw-wise sampling probabilities for people with higher degrees (Gile, 2011). Thus, as the sample begins to deplete the target population, we would expect higher-degree nodes to be sampled earlier, followed by lower-degree nodes, suggesting that a decreasing trend in degree over time could be an indication of finite population effects on sampling. We compared several options for evaluating the trend of degree over time. We used time-order in the study to measure time in these analyses, although results were robust to using survey date. These approaches grouped roughly into two families: those sensitive to a small number of outliers (linear regression, Poisson regression), and those robust to a small number of outliers (regression on log degree, robust regression approaches such as least trimmed squares, M regression, and median regression, as well as rank-based Kendall's Tau and Spearman's Rho). Approaches within each family tended to produce similar results. Because of the unknown dependence in the data structure, we considered only the sign of the coefficient of time in each model. Surprisingly, we find little evidence of decreasing degree over time

Address for correspondence: Krista J. Gile, Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003-9305, U.S.A.
E-mail: gile@math.umass.edu



(a) (b)

Fig. S1. Proportion of alters already participated, by seed.

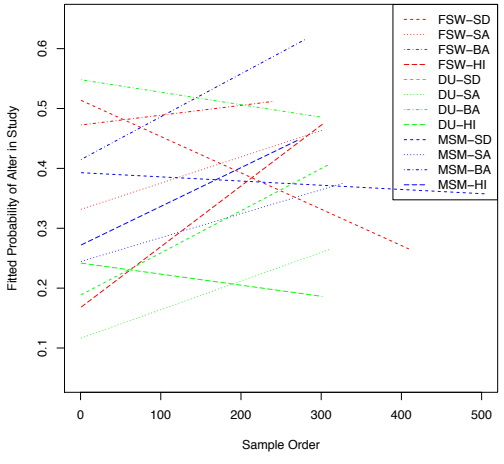
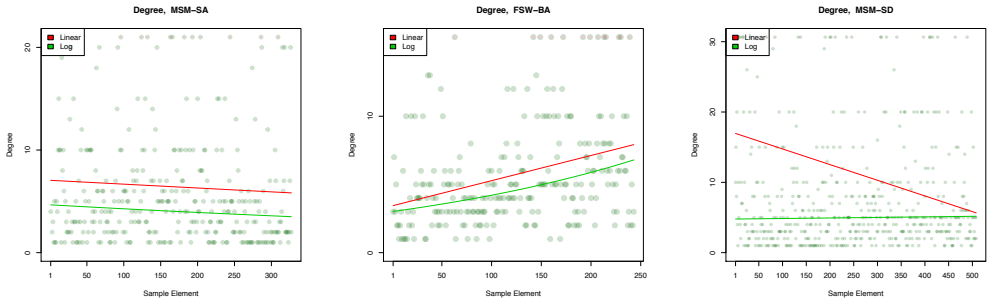


Fig. S2. Fitted linear trends for 12 sites for the proportion of respondents' contacts who had already participated in the study.

with either the non-robust (5 of 12 flagged for linear slope with the linear model) or robust methods (1-3 of 12 flagged).

Fig. S3 illustrates the fitted linear relationship between degree and sample order, as well as the linear relationship fitted to log degree for three sites. In Fig. S3(a) (MSM-SA), both approaches found a negative relationship between degree and sample order; in Fig. S3(b) (FSW-BA) both approaches found a positive relationship; and in Fig. S3(c) (MSM-SD) the two approaches found differing trends, likely driven by the few high responses early in the sample.

Because we have more confidence in the more robust methods, we conclude that this



(a) (b) (c)

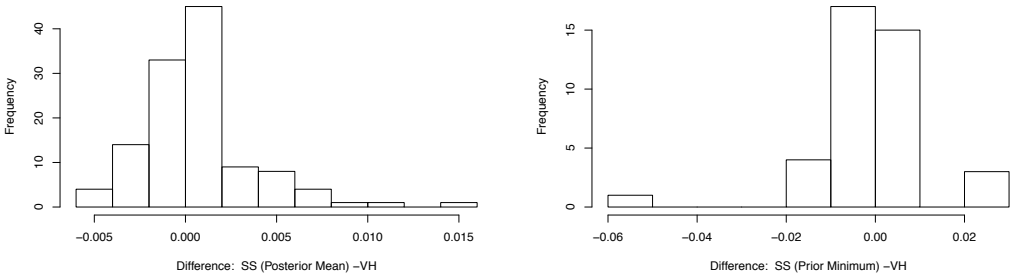
Fig. S3. Degree of respondents over time, with fitted linear model and linear model for log of degree. For visualization, the highest responses were truncated and represented in red at the tops of the plots.

indicator clearly suggests finite population effects for MSM-SA (flagged by all indicators) and perhaps MSM-SD (flagged by most robust indicators). It is surprising to us that all the other populations, including the three known to have not reached their target sample size (FSW-BA, MSM-BA, MSM-HI), suggested positive or null trends in sample degree over time. Because we have strong theoretical reasons (Gile, 2011) to expect negative trends in these cases, we hope future research, with other data sets, will help explain this phenomenon.

S1.3. Successive Sampling Estimation of Finite Population Bias

If researchers have an estimate of the size of the target population, they can compare the SS estimator (Gile, 2011) to the VH estimator (Volz and Heckathorn, 2008) in order to assess finite population effects on estimates. As is typically the case, however, there were no existing estimates of the sizes of our target populations. Therefore, we use the RDS data itself in order to estimate the sizes of our target populations using the approach introduced in Handcock et al. (2012) and implemented in the `R` (R Core Team, 2012) package `size` (Handcock, 2011).

The method of Handcock et al. (2012) requires specifying a prior distribution for the size of the population. To specify the prior distribution for populations of MSM, we drew on a meta-analysis of Caceres et al. (2006), which provides broad bounds on the proportion of men who have had sex with another man in the past year. The estimate for the Dominican Republic (and all of Latin America) is 1-8% of the sexually active adult male population, which we assume to constitute 15-64 year olds. Combining this information with information on the number of males between 15-64 in each city from the Dominican Republic's National Statistical Office (Oficina Nacional de Estadistica, 2009), we created a conservative upper and lower bound for the size of the MSM population in each city. These bounds are then used to define the lower and upper quartiles of a prior distribution. For DU and FSW, no comparable meta-analyses existed so we used broad ranges, consisting of 1-10% of the 15-64 year old total population (DU), or population of women (FSW). As with the MSM, we used these ranges, combined with information from the Dominican Republic's National Statistical Office (Oficina Nacional de Estadistica, 2009) to create prior distributions. When setting the priors in this manner, the method of Handcock et al. (2012) results in posterior mean MSM population size estimates within the original range for SD, SA, and HI, and just above the higher end of the range in Barahona. For DU and FSW, this procedure produced



(a) Posterior Mean (b) Prior Minimum (MSM only)

Fig. S4. Histograms of the differences between Volz-Heckathorn and Successive Sampling estimates over many traits of interest, with SS estimate based on (a) the posterior mean and (b) the prior minimum (MSM populations only).

6 estimates consistent with the ranges specified in the prior, one (FSW in BA) higher than the 10% number, and two (DU in SD and SA) lower than 1%.

When using the SS estimator, therefore, we used three plausible low population sizes:

- The posterior mean (best point estimate from the population size estimation)
- The lower bound of the posterior highest probability density (HPD) region (lowest plausible estimate from the population size estimation)
- For MSM populations, 1% of the 15-64 year old male population (lower bound of the plausible region from the meta-analysis of Caceres et al. (2006)).

Using each of these estimates of population size, we estimate prevalence of each of the characteristics described in Section S2 using the SS estimator, as well as using the VH estimator. A histogram of the differences based on the lower bound of the HPD region is given in the main text (Fig. 4). Corresponding plots for the other two population size estimates are given in Figure S4. For an single site, a plot like those in Figure S5 may be more useful. These *Population Size Sensitivity Plots* summarize the differences across several population size estimates for the traits of interest in an individual study. For completeness, all items with difference greater than .01 are summarized in Table S1.

S2. Seed dependence

We recommend visual inspection of Convergence Plots and Bottleneck Plots, but in cases where there are many study sites and many traits of interest, it may be difficult to monitor all of these plots. Therefore, we develop a set of procedures that enable researchers to automatically flag plots for further inspection.

For the Convergence Plots, further inspection is called for if the estimates seem to be changing at the end of the sample. That is, a trait should be flagged if

$$\text{there exists} \quad t < \tau \text{ such that } \mid \hat{p}_{(n-t)} - \hat{p}_{(n)} \mid > \epsilon \tag{1}$$

where τ is a parameter that sets how much of the trace will be examined and ϵ represents the maximum allowable difference between the estimate at time t and the final estimate. We suspect that the desired values of τ and ϵ will vary from study to study, but in this case we set $\tau = 50$ and $\epsilon = 0.02$. In other words, we ask whether there are any of the final 50 estimates that have a difference of more than 0.02 from the final estimate. We run

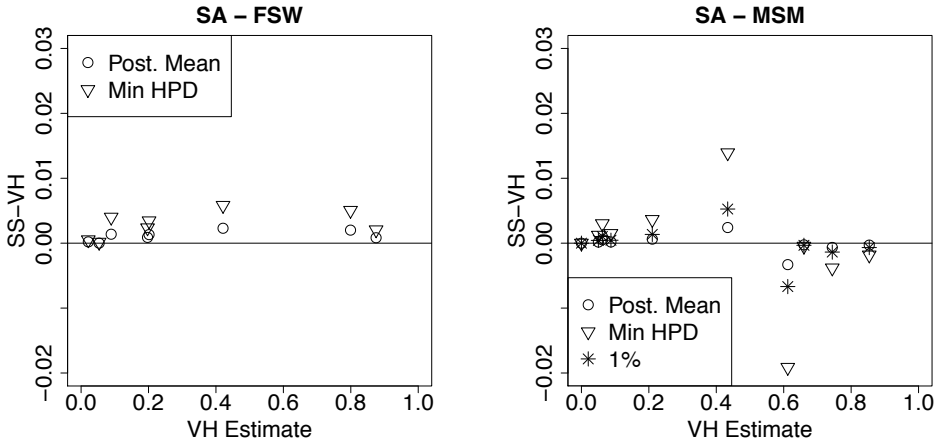


Fig. S5. *Population Size Sensitivity Plots* comparing Volz-Heckathorn estimates and Successive Sampling estimates of prevalence of many traits of interest in two target populations.

Table S1. Prevalence estimates based on Successive Sampling and Volz-Heckathorn estimators for each trait with maximum absolute difference greater than .01.

		Trait	VH	Max HPD	Post. Mean	Min HPD	1%
FSW	HI	Last Client Brothel	0.306	0.316	0.321	0.334	-
FSW	HI	Been In Program	0.345	0.349	0.351	0.356	-
DU	SD	Main Drug Crack	0.263	0.267	0.270	0.275	-
DU	SD	Use Drugs Every Day	0.378	0.385	0.388	0.397	-
DU	SA	Use Drugs Every Day	0.360	0.364	0.367	0.374	-
DU	SA	Been Imprisoned	0.370	0.374	0.376	0.382	-
DU	BA	Use Drugs Every Day	0.391	0.397	0.400	0.410	-
DU	HI	Main Drug Cocaine	0.422	0.418	0.416	0.410	-
DU	HI	Use Drugs Every Day	0.406	0.410	0.413	0.419	-
DU	HI	Been Imprisoned	0.259	0.263	0.265	0.271	-
MSM	SA	Had HIV Test	0.434	0.434	0.436	0.448	0.439
MSM	SA	Bisexual	0.612	0.612	0.609	0.593	0.605
MSM	BA	HIV+	0.087	0.087	0.086	0.085	0.074
MSM	BA	Had HIV Test	0.331	0.330	0.328	0.321	0.277
MSM	BA	Working	0.711	0.712	0.712	0.716	0.735
MSM	BA	Use Drugs	0.607	0.608	0.609	0.613	0.633
MSM	BA	Sex With Woman	0.858	0.859	0.859	0.863	0.884
MSM	HI	Had HIV Test	0.503	0.503	0.501	0.493	0.491
MSM	HI	Used Condom	0.790	0.790	0.787	0.776	0.773
MSM	HI	Sex With Woman	0.834	0.834	0.833	0.825	0.823

this procedure on 120 group $\times$ trait $\times$ city combinations shown in Fig. S6, and we find the most convergence flags in MSM data: 37.5% of traits were flagged, as compared with 25% of traits for DU and 22% for FSW. Increasing ϵ to 0.05 results in flagging only two traits, both in MSM populations: Bisexual in Santiago and Use Drugs in Higuey. The convergence problems that we detected could be caused by the network structure in the population, the method of seed selection, and the interaction between the two.

For the Bottleneck Plots, further inspection is called for if the estimates from each seed

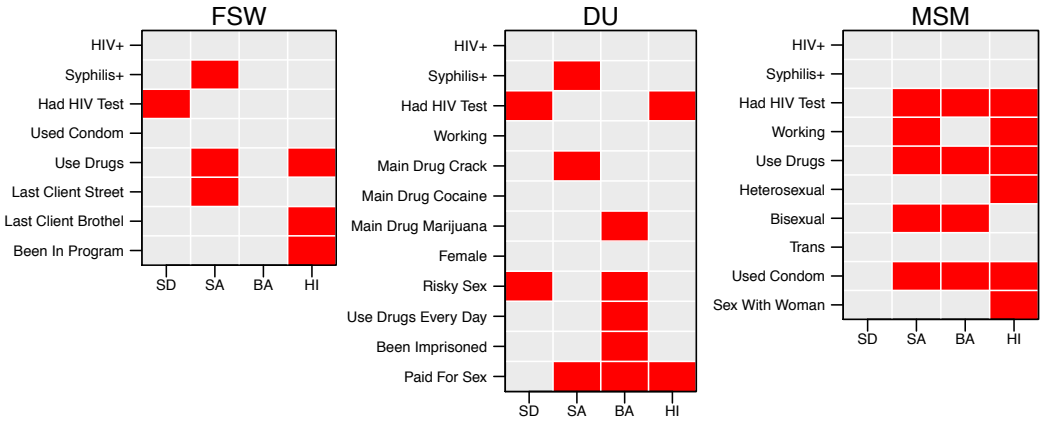


Fig. S6. Convergence test results for $\tau = 50$ and $\epsilon = 0.02$. Red cells represent traits flagged for possible lack of convergence.

deviate from the overall estimate. We formalize that intuition by calculating a weighted squared deviation:

$$WSD = \sum_s n_s \cdot (\widehat{p}_s - \widehat{p})^2 \tag{2}$$

where $\widehat{p}_s$ is the estimate using sample only from the tree originating at seed s and n_s is the corresponding sample size (not including the seed itself and not including cases with missing data on the trait of interest or degree). In order to assess whether this statistic is unusual, we perform a permutation procedure where the structure of trees are preserved (including weights) while the individual traits are permuted. We then calculate the WSD for the permuted data, and we repeat this procedure 10,000 times. We flag a trait for further investigation if the observed WSD is greater than 90% of the permuted WSD values; this threshold can be adjusted for desired sensitivity.

We ran this procedure on the same 120 group $\times$ trait $\times$ city combinations examined previously and found that the rates of flagging were highest among FSW (41%) followed by MSM (30%) and then DU (23%) (Fig. S7). Although no trait was flagged in all four cities, these results suggest that likely sources of bottlenecks for FSW are based on sources of clients (e.g., brothel vs. street), drug use, and disease status (HIV and Syphilis); for DU based on type of drug used (Marijuana), employment status, and gender; and for MSM based on self-identification (e.g., bi-sexual and transsexual). These results also suggest that bottlenecks can occur across traits that are not visible to respondents (e.g., disease status) possibly because these traits are correlated with other traits (e.g., age or risky behavior) that do affect social tie formation. Finally, it is important to note that some target populations (e.g., MSM in Santo Domingo) appear to have bottlenecks along many traits.

We wish to emphasize that our flagging procedure for Bottleneck Plots may not always match the intuition of experienced RDS researchers. While it does correctly flag obvious cases of bottlenecks (Fig. 8(a)) and it does not flag in cases where there do not appear to be bottlenecks (Fig. 8(b)), there are a small proportion of edge cases where our flagging procedure did not match our expectations. For example, our procedure flags HIV status for MSM in Barahona (Fig. 8(c)) although this is caused by two chains of length 1, and is therefore probably not cause for concern. On the other hand, our procedure does not flag

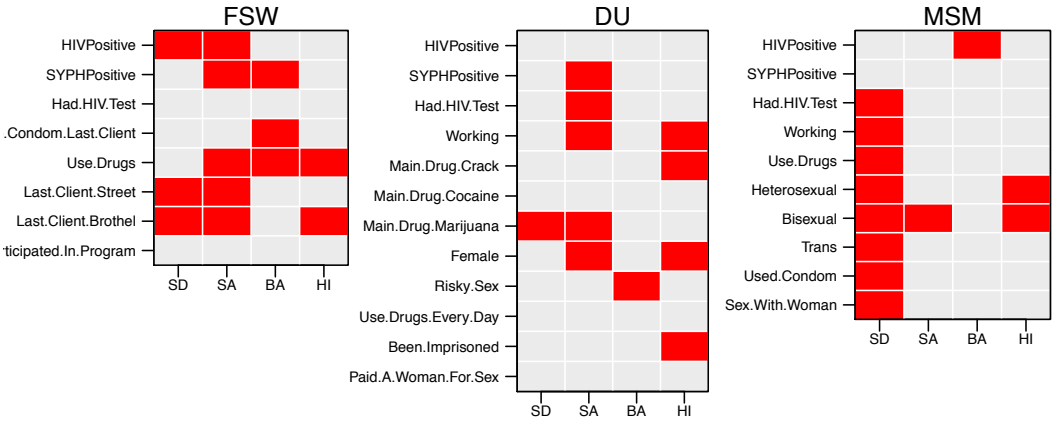


Fig. S7. Bottleneck test results. Red cells represent traits flagged for possible bottlenecks.

HIV status for MSM in Santiago (Fig. 8(d)) even though a review of the plot seems to call for further investigation into the difference between the long chain with approximately 15% estimated prevalence to the other chains with close to zero estimated prevalence. Thus, while we find this procedure a useful heuristic, we hope that future researchers will develop a more formal approach.

S3. Reciprocation

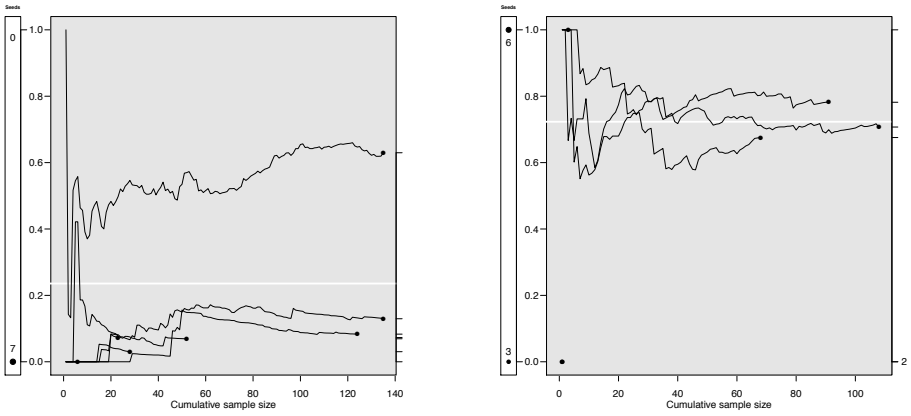
In this section, we introduce a measure of reciprocation of all network ties, rather than just the ties associated with coupon-passing. Although the recall task associated with reporting these data is more complicated than asking only about the recruiter, it is the reciprocation of all ties, not just those involved in coupon-passing that is necessary for estimating sampling probabilities. This is because the estimation of sampling probabilities in RDS relies on the self-reported number of network connections. If all relationships are reciprocated, then the number of network connections is related to a respondent's sampling probability. Otherwise, it is the respondent's in-degree, or number of incoming relations, that is related to sampling probability. Unfortunately, reporting numbers of incoming relations is very difficult. Current estimators for RDS data therefore require reciprocation for two reasons. First, out-ties are easier to self-report and therefore more often recorded, while in-ties are more directly related to sampling probabilities. If all ties are reciprocated, then self-reported out-degree is the same as in-degree. Furthermore, if all ties are reciprocated, the sampling process more closely approximates a random walk on an undirected graph, a common assumption of estimators used.

During the initial visit, participants were asked the following questions about their alters (MSM versions; other groups were analogous):

(Q) How many of them (repeat the number in F) know you well enough that they could give you a coupon within a week if they had been in this study?

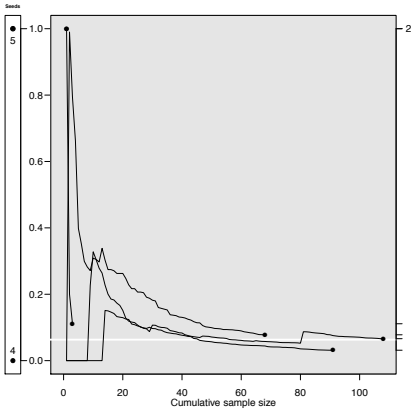
(R) If we were to give you as many coupons as you wanted, how many of them (repeat the number in F) could you give a coupon to?

(S) If we were to give you as many coupons as you wanted, how many of these MSM (repeat the number in R) do you think you could give a coupon to by this time next week?

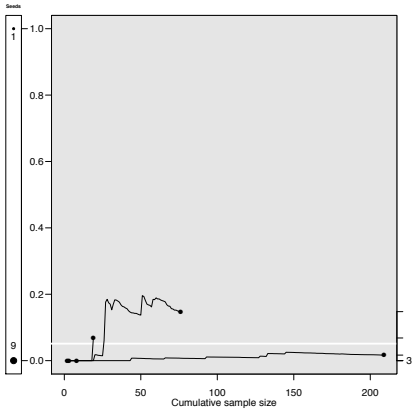


(a) MSM-SD, Use drugs everyday

(b) MSM-BA, Employed



(c) MSM-BA, HIV+



(d) MSM-SA, HIV+

Fig. S8. *Bottleneck Plots:* The left panel in each plot reports the composition of the seeds and the tick marks on the right axis show the final estimates. If there is more than one tree with the same final estimate, that number is also shown on the right axis (see (c) and (d)).

Among all 3,860 respondents who responded to all of these questions, 29.7% gave the same answer for both questions Q and S, 46.7% reported they could give more coupons than they might receive, and 23.6% reported the opposite. The median difference between responses to these questions was 0 and the mean difference of 1.5 more coupons that could be given out than received. Larger differences are positively associated with larger maximum response to either question. For this reason, we also consider normalized difference values, computed as follows:

$$\frac{|Q - S|}{max(Q, S)}.$$

Using these normalized values, the median difference is still 0, with mean 0.40 and third quartile 0.67. This approach is conceptually closer to the full requirement of the reciprocity assumption, but it is also subject to larger concerns of reporting accuracy. Therefore, we prefer the approaches described in Section 6.

S4. Measurement of Degree

S4.1. Time dynamics

We conducted three analyses to check whether the one week time frame in question G was reasonable (see Sec. 7.1). First, for each respondent, we calculated the proportion of his or her alters (based on question F) that could be reached in a specific time frame (based on questions H and I). Fig. S9 depicts the average proportion of alters reachable in each period, by site, with logically inconsistent results excluded.† With the exception of the DU in Santiago, almost all alters were reachable within seven days. The average rate across sites was 92% reachable within one week. Within one day, the across-site-average percent reachable was 62%.

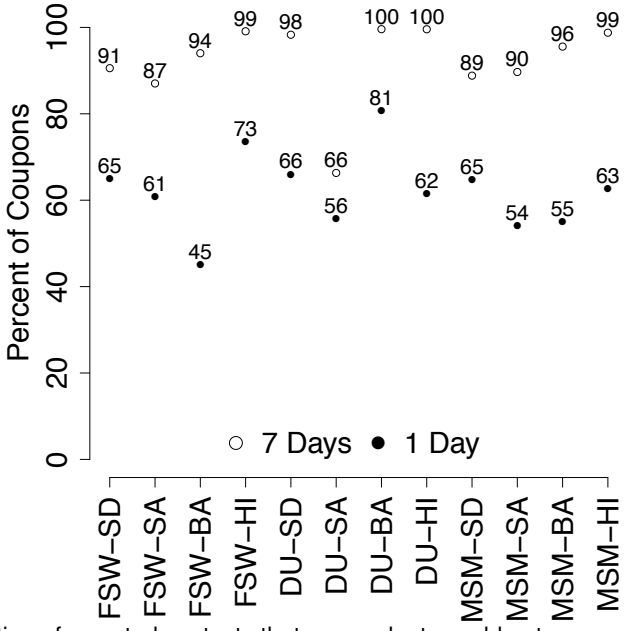


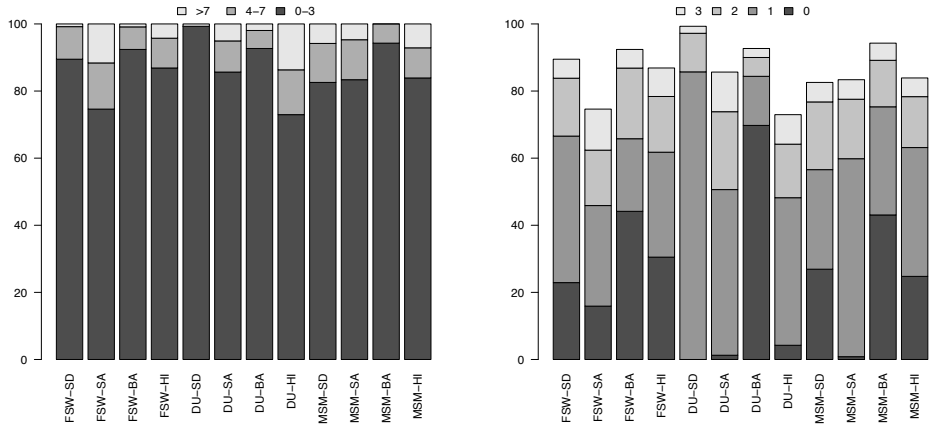
Fig. S9. Proportion of reported contacts that respondents could get a coupon to in 1 or 7 days.

Second, we considered the self-reported number of days each respondent took to distribute his or her coupons (asked at follow-up). Fig. S10 illustrates that across sites, over half (64%) of coupons were distributed in one day and almost all (95%) within seven days.

Finally, we examined the difference between the interview dates for each recruiter-recruit pair, a measure of time dynamics that does not rely on respondent's reports (but which may, unfortunately, be influenced by the capacity for study sites to process interviews during high demand.) Fig. S11 shows that in each site, a substantial majority (79% overall) of interviews occur within a week of the recruiter's interview.

Overall, these three results suggest that restricting social network recall to people a respondent has seen within the last week appears reasonable in this study. Nearly all coupons

†Responses were deemed logically inconsistent and therefore were excluded if a respondent reported being able to reach more contacts than (s)he knew (F). Three sites had high numbers of logically inconsistent responses: FSW-SD (56, 21) (7 days, 1 day), DU-SA (65, 39), MSM-SD (63, 40). A total of 42 responses were inconsistent across the remaining 7 sites.



(a) All days (b) Days 0 - 3

Fig. S10. Percent of coupons distributed by number of days, by site. Most coupons were distributed within 3 days, and nearly all within 7 (a). Among DU in Barahona most coupons were distributed in one day (b).

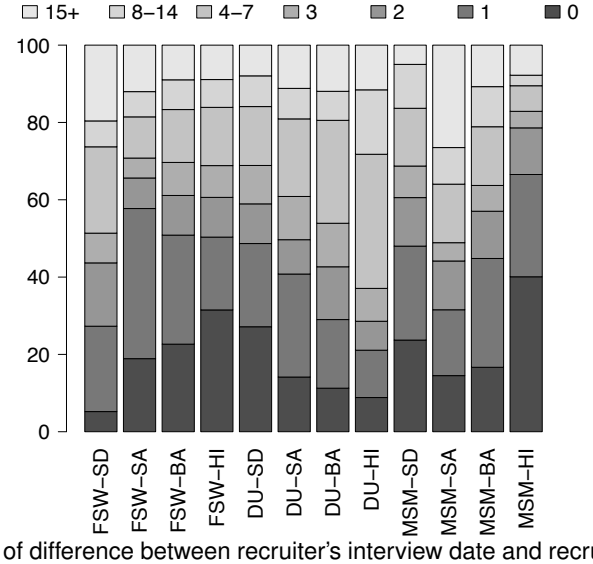


Fig. S11. Distribution of difference between recruiter's interview date and recruit's interview date, by site.

were distributed within a week, and aside from the DU in Santiago, most respondents reported being able to reach nearly all social contacts within a week. Because most coupons were distributed within a shorter period of a few days, it might even make sense to further restrict the recall period to two or three days. Note that the validity of this measure, however, relies on the assumption that coupons were distributed to people incidentally encountered, rather than sought out. Further study is necessary to determine whether respondents seek

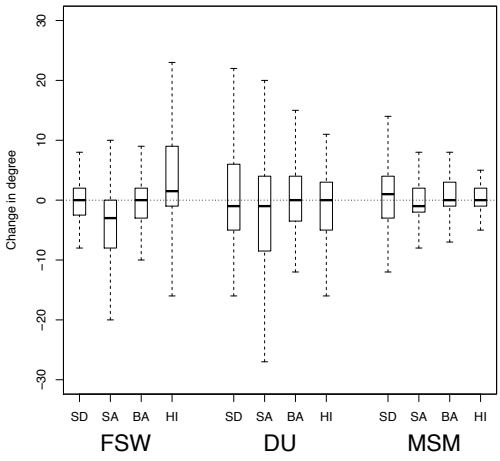
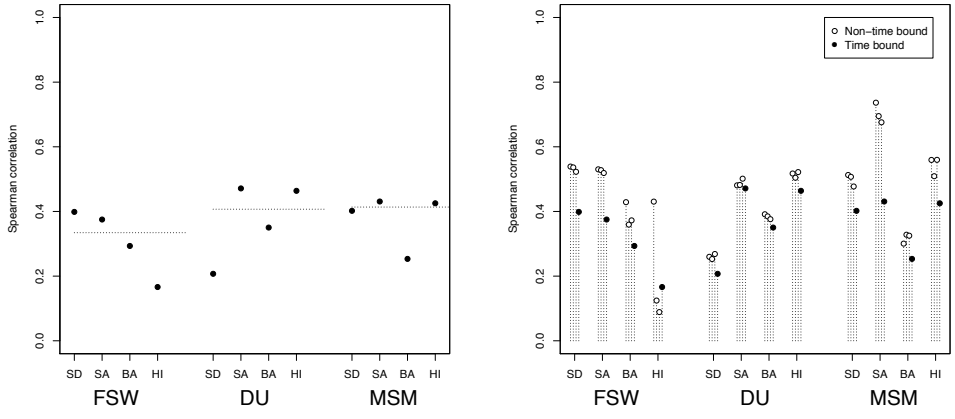


Fig. S12. Boxplots of the test-retest difference in reported degree, measured by Question (G), by group and city. There is no general pattern of increase or decrease. In order to show the median, 25th and 75th percentiles more clearly, this plot does not include points outside of the whiskers.



(a) (b)

Fig. S13. (a) Spearman rank correlation between test and retest measures for the main degree question (G) with the median value for each group marked by the horizontal line. (b) The measures that are not time-bounded (D, E, F) have higher correlation.

out their recruits, or select them from among incidentally encountered alters.

S4.2. Test-retest reliability

In order to assess the test-retest reliability of the degree questions, questions D-G were included in both the initial and follow-up interviews. The median difference in degree (G) at interview and follow-up is 0 (25^{th} percentile = -3, 75^{th} percentile = 3). Further, Fig. S12 shows that results were similar across study site and target population.

Fig. S13(a) shows the test-retest reliability for the degree question used for estimation (question G). One potential reason for the low test-retest reliability of question G is that it refers to a seven day time frame. Therefore, even if respondents are perfectly accurate in their responses there could be test-retest variation because of week-to-week variation. This issue of time-bounded questions has come up in other test-retest studies (e.g., van Groenou et al. (1990)), but is difficult to resolve because it is not reasonable to ask respondents at the follow-up interview about their experiences in the one week proceeding their initial interview as that is often about three weeks in the past. However, one way to roughly gauge how much extra variability is introduced by this time frame is by examining the first three network size questions which are not time-bounded. Fig. S13(b) shows that the test-retest reliability is higher for the non time-bounded questions, but only slightly so.

Finally, we note that when considering measures of test-retest reliability, it is critical to consider any potential sources of dependence between the measures. Here interviewers at the follow-up visit did not know respondent's answers from the initial visit. Further, since the time period between interviews was generally around three weeks, it is extremely unlikely that respondents remembered their original responses to the degree questions. One possible source of dependence that did exist in this study is that the respondents may have been interviewed by the same interviewer at the initial and follow-up visits, thus possibly increasing perceived test-retest reliability.

S5. Testing Recruitment Bias

Most RDS inference relies on the assumption that recruits are selected at random from among the contacts of each recruiter. Under this assumption, successful recruits should constitute a simple random sample of the personal networks of respondents. In most cases, reviewing a Recruitment Bias Plot (e.g., Fig. 10) should be sufficient to inform researchers' intuition about whether recruitment bias is a concern. In some cases, researchers may want to test whether the observed recruitment patterns are consistent with random recruitment. However researchers should note that statistical significance is not the same as estimator bias, so even a perfect test would not be a good judge of whether a recruitment bias is strong enough to impact estimates.

With no recruitment bias, the coupons should be passed to a simple random sample of the recruiter's contacts, and the coupons returned should be returned by a simple random sample of those receiving coupons. To test these assumptions non-parametrically, we compare the (unweighted) count of employed at each stage to a null distribution approximated by simulated simple random sampling from the reported composition of the relevant subset of each recruiter's eligible alters. To test for biased coupon passing, for example, we simulate the coupon-recipients of each recruiter by drawing n_i^c samples from among F_i units, including J_i employed, where n_i^c is the reported number of coupons distributed by i, F_i is i's reported number of contacts, and J_i is i's reported number of employed alters. Non-parametric null distributions for returning coupons and for overall recruitment were constructed similarly, with test statistics and reference distributions described in Table S2.

Our tests show very small p-values, suggesting the reported recruitment patterns are very unlikely absent recruitment bias (see Table S3). However, one reason for these extreme findings could be poor data quality, perhaps due to a desirability bias of employment status reporting. Many data points are logically inconsistent, with more employed alters receiving coupons than were originally reported known, or with more employed recruits than coupons given to employed people. In addition, there is no evidence that those with more reported employed contacts tend to recruit more employed people (the correlation between

Table S2. Test statistics and reference distributions for testing for Recruitment Bias at three levels: in the passing of coupons, in returning coupons, and overall.

Test	Test statistic	Reference Distribution
Coupon Passing	Count of employed coupon-recipients	SRS from contact composition of each recruiter
Returning Coupons	Count of employed recruits	SRS from composition of coupon recipients of each recruiter
Overall	Count of employed recruits	SRS from contact composition of each recruiter

these proportions is negative in many of the samples). Therefore, while we feel this test is mathematically appropriate, we suggest caution in its use, or the use of earlier tests relying on self-reported network compositions.

Finally, we note that other approaches have been used previously to compare reported network composition to actual sample recruits. These approaches can be roughly divided into two categories: those that make standard distributional assumptions, including independence, to perform chi-square and t-tests of recruitment patterns, and those that test for the impact on estimators, using standard RDS confidence intervals. While each of these other approaches certainly adds to information available to researchers, we prefer our non-parametric testing approach because it relies on neither parametric assumptions, nor the validity of standard RDS confidence intervals.

The earliest work in assessing non-random recruitment is in Heckathorn et al. (2002), which looks at the correlation between implied population proportions across several groups under random sampling and observed cross-group recruitment patterns. This approach is not ideal because we would like to test whether the compositions are the same, not just correlated. Wang et al. (2005) therefore extend this approach by using a t-test to compare the sample proportion of the observed data to the reported network compositions. This approach relies on a binomial approximation to the distribution of the estimated proportions. Wejnert and Heckathorn (2008) introduce a chi-square test to compare the expected referral matrix under random referral to the observed referral matrix. This approach also relies on distributional assumptions, in particular an assumption of independence of observations. It is unclear from the Wang et al. (2005) and Wejnert and Heckathorn (2008) papers which form of weighting is used to estimate the composite alter characteristics. Rudolph et al. (2011) also use chi-square tests and t-tests to test for non-random sampling, but they do not give details on their methods. Finally, Jenness et al. (2014) use a t-test to compare the geographical distance to sampled contacts to (a proxy for) the distance to contacts in general.

Iguchi et al. (2009) compare population proportion estimates directly, by substituting reported network composition for composition of referrals in RDS estimators. Their test is based on comparing the usual and network-composition-based population proportion estimates, using the RDS uncertainty estimates under the two conditions. This approach has the advantage of placing the comparison on the scale of the measure of interest, but is only applicable for non-random recruitment on the feature to be estimated, and can only be used with estimators relying on estimates of proportions of cross-group ties (e.g. Salganik and Heckathorn (2004)). Liu et al. (2012) also propose a test relying on standard RDS estimates for some (five) visible attributes and their confidence intervals, and also make use of information collected on the composition of the social network alters of the respondents. They compare the estimates to the value of the "population proportions of five visible

Table S3. P-values for non-parametric tests of recruitment bias based on employment status on three levels: Which contacts are given coupons, which coupon recipients return coupons to become recruits, and overall, which contacts become recruits. P-values suggest the reported recruitment patterns are very unlikely absent recruitment bias. The second section records the proportion of cases in each setting in which the number of employed persons selected was larger than the number available. This suggests the apparently strong recruitment bias may be due to data quality issues.

	P-value				Proportion Inconsistent			
	SD	SA	BA	HI	SD	SA	BA	HI
Coupon Passing	0.369	$< .0001$	$< .0001$	$< .0001$	0.094	0.137	0.201	0.216
Returning Coupons	$< .0001$	$< .0001$	$< .0001$	$< .0001$	0.519	0.286	0.352	0.243
Overall	$< .0001$	$< .0001$	$< .0001$	$< .0001$	0.202	0.143	0.192	0.206

variables among the total drug-using alters from which the RDS sample was drawn." It is not clear from their paper how they compute this proportion, or precisely how it relates to the population proportion.

The work of Yamanis et al. (2013), concurrent with this work, provides descriptive analyses most similar to our work. These authors also compare the inferred characteristics of invited recruits, successful recruits, and the full population of all alters, in a tabular representation similar to a *Recruitment Bias Plot*, and broken down by recruiter characteristics. Their descriptives are also slightly different from ours because their measurement strategy is different from ours, asking respondents only collectively about the features of their uninvited alters. Rather than testing for non-random recruitment directly, they suggest a test for the impact of non-random recruitment on inference, similar to the approach of Iguchi et al. (2009). They suggest a modified bootstrap procedure to compute theoretical proportion estimates under random sampling from invited recruits, successful recruits, and the full population of alters. They derive statistical significance based on the relative values of these estimates as compared to their bootstrap confidence intervals.

S6. Non-response to Follow-up

We use data collected from the follow-up interview to provide some evidence for some of our diagnostics, but only about half of participants completed a follow-up interview (Fig. 2). Therefore, we compare the participants who did and did not complete a follow-up questionnaire on several characteristics, and summarize those results in Table S4. Table S5 presents comparisons of each population type, aggregated across the four cities, as well as a grand aggregation (total) over all 12 sites. Chi-square and t-tests for statistical significance were conducted for each of the 12 sites, each of the 3 groups, and the study total, resulting in 16 tests for each trait; tests that were significant at the .05 level are listed in table S5.

We first compare the degrees of respondents and non-respondents across the 16 comparisons, the mean degrees of follow-up respondents and non-respondents were only significantly different in one site, Drug Users in Santo Domingo.

Many respondents return to the study center to receive secondary incentives for successful recruitments. For this reason, we might expect higher rates of recruitment among follow-up respondents. Indeed, this is the case in our data, with significantly higher average recruitments among follow-up respondents in all 16 comparisons. A comparison of the proportions with no recruits continues this pattern, with a significantly higher proportion with no recruits among follow-up non-respondents. Finally, we also compare the average number of recruits for follow-up respondents and non-respondents, excluding those with no recruits. Again, all sites have significantly higher average numbers of recruits among follow-up respondents.

Table S4. Comparison of respondents (Resp.) to follow-up and non-respondents (Non-Resp) to follow-up, based on average values of various variables for each population and for the full sample.

	FSW Resp	FSW Non-Resp	DU Resp	DU Non-Resp	MSM Resp	MSM Non-Resp	Total Resp	Total Non-Resp
N	558	698	557	665	562	826	1677	2189
Mean Degree	9.08	7.68	14.04	14.61	9.68	6.82	10.93	9.46
# Recruits	1.57	0.50	1.47	0.64	1.62	0.56	1.55	0.57
No Recruits	0.22	0.68	0.23	0.57	0.18	0.63	0.21	0.63
# Recruits/0	2.02	1.57	1.89	1.49	1.97	1.54	1.96	1.53
Study Day	24.04	35.53	15.23	19.71	28.43	34.33	22.59	30.27
Study Day/0	22.21	35.39	13.23	16.43	27.51	31.17	21.13	26.79
HIV Positive	0.04	0.05	0.05	0.08	0.09	0.06	0.06	0.06
For Incentive	0.03	0.03	0.07	0.12	0.14	0.14	0.08	0.10
For HIV Test	0.85	0.86	0.59	0.65	0.12	0.14	0.52	0.53
For Any Test	0.86	0.87	0.61	0.66	0.37	0.38	0.61	0.62

Table S5. Listing of populations showing significant differences between follow-up respondents and non-respondents for each variable considered in Table S4. Significant results for the full sample are indicated by "Total", and for aggregated data by population by "FSW," "DU," and "MSM." Average differences for these populations can be found from Table S4. For individual sites, numbers in parentheses are the difference in means (Mean_Resp-Mean_Non-Resp). Statistical significance based on t-test and chi-square tests at level 0.05.

Term	Significant Sites
Mean Degree	DU-SD (-7.83)
# Recruits	Total, FSW, DU, MSM, FSW-SD (1.98), FSW-SA (0.88), FSW-BA (0.92), FSW-HI (0.57), DU-SD (0.87), DU-SA (0.78), DU-BA (0.88), DU-HI (0.82), MSM-SD (0.97), MSM-SA (1.45), MSM-BA (0.86), MSM-HI (1.13)
No Recruits	Total, FSW, DU, MSM, FSW-SD (-0.75), FSW-SA (-0.35), FSW-BA (-0.41), FSW-HI (-0.26), DU-SD (-0.29), DU-SA (-0.29), DU-BA (-0.43), DU-HI (-0.38), MSM-SD (-0.39), MSM-SA (-0.56), MSM-BA (-0.45), MSM-HI (-0.47)
# Recruits/0	Total, FSW, DU, MSM, FSW-SD (0.66), FSW-SA (0.45), FSW-BA (0.54), FSW-HI (0.24), DU-SD (0.55), DU-SA (0.45), DU-BA (0.28), DU-HI (0.25), MSM-SD (0.37),, MSM-SA (0.89), MSM-BA (0.29), MSM-HI (0.5)
Study Day	Total, FSW, DU, MSM, FSW-SD (-10.7), FSW-SA (-19.77), FSW-BA (-9.33), FSW-HI (-4.27), DU-SD (-4.56), DU-BA (-7.82), MSM-SA (-12.72), MSM-BA (-5.28), MSM-HI (-11.73)
Study Day/0	Total, FSW, DU, MSM, FSW-SD (-12.63), FSW-SA (-21.37), FSW-BA (-10.8), DU-SD (-6.81), MSM-SA (-16.89), MSM-HI (-16.82)
HIV Positive	DU, FSW-BA (-0.08), MSM-HI (0.08)
For Incentive	DU
For HIV Test	DU, MSM-SA (-0.1)
For Any Test	none

It is also possible there is a censoring effect deterring response to follow-up among respondents completing their primary interviews closer to the end of the study. Indeed, we find that follow-up non-respondents tend to have their initial interviews significantly later in the study than follow-up respondents in all but 3 sites. Part of this effect may be attributable to the fact that respondents at the very end of the study are not given coupons and are therefore unable to make recruitments. Comparing the average study date for follow-up respondents and non-respondents excluding those with no recruits reveals remaining significant differences in all but 6 sites.

Finally, the second contact with the interview site also serves to deliver HIV and other test results. We therefore compare follow-up respondents and non-respondents on HIV status as well as based on their stated motivations for participating in the study. We find significant differences in HIV status in only 3 cases (drug users overall, female sex workers in Barahona, and men who have sex with men in Higuey), without a consistent pattern in the direction of the association. In terms of motivations for participation, drug users overall (but not in any particular site) are less likely to participate in the follow-up study if their primary stated motivation was for the financial incentive. Surprisingly, drug users whose primary stated motivation was HIV test results were less likely to participate in follow-up, while MSM in Higuey were more likely to participate in follow-up when motivated by HIV test results. There were no significant differences in other sites. We also compared the proportions who mentioned any disease test as a primary motivation, and found no significant differences.

Overall, then, our follow-up non-respondents seem to differ from follow-up respondents primarily based on their rates of recruitment (follow-up respondents recruited more often), and study date (earlier participants more likely to follow-up). Where relevant to our conclusions, the impacts of these results are noted in the text.

References

Caceres, C., K. Konda, M. Pecheny, A. Chatterjee, and R. Lyerla (2006). Estimating the number of men who have sex with men in low and middle income countries. *Sexually Transmitted Infections 82*(suppl 3), iii3–iii9.

Gile, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association 106*(493), 135–146.

Handcock, M. S. (2011). size: Estimating hidden population size using respondent driven sampling data. R package version 0.20.

Handcock, M. S., K. J. Gile, and C. M. Mar (2012). Estimating hidden population size using respondent-driven sampling data. *Working paper*.

Heckathorn, D., S. Semaan, R. Broadhead, and J. Hughes (2002). Extensions of respondent-driven sampling: A new approach to the study of injection drug users aged 18-25. *AIDS and Behavior 6*(1), 55–67.

Iguchi, M. Y., A. J. Ober, S. H. Berry, T. Fain, D. D. Heckathorn, P. M. Gorbach, R. Heimer, A. Kozlov, L. J. Ouellet, S. Shoptaw, and W. A. Zule (2009). Simultaneous recruitment of drug users and men who have sex with men in the United States and Russia using respondent-driven sampling: Sampling methods and implications. *Journal of Urban Health 86*(S1), 5–31.

Jenness, S. M., A. Neaigus, T. Wendel, C. Gelpi-Acosta, and H. Hagan (2014). Spatial recruitment bias in respondent-driven sampling: Implications for HIV prevalence estimation in urban heterosexuals. *AIDS and Behavior*, forthcoming.

Liu, H., J. Li, T. Ha, and J. Li (2012). Assessment of random recruitment assumption in respondent-driven sampling in egocentric network data. *Social Networking 1*(2), 13–21.

Oficina Nacional de Estadistica (2009). Poblacion estimada y proyectada region provincia y municipio 2000-2010. September 18, 2009; Accessed August 2, 2012.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org/.

Rudolph, A. E., N. D. Crawford, C. Latkin, R. Heimer, E. O. Benjamin, K. C. Jones, and C. M. Fuller (2011). Subpopulations of illicit drug users reached by targeted street outreach and respondent-driven sampling strategies: Implications for research and public health practice. *Annals of Epidemiology 21*(4), 280–289.

Salganik, M. J. and D. D. Heckathorn (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology 34*(1), 193–240.

van Groenou, M. B., E. v. Sonderen, and J. Ormel (1990). Test-retest reliability of personal network

delineation. In T. Antonucci and C. Knipscheer (Eds.), *Social Network Research: Substantive Issues and Methodological Questions*, pp. 121–136. Amsterdam: Swets and Zeitlinger.

Volz, E. and D. D. Heckathorn (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics 24*(1), 79–97.

Wang, J., R. G. Carlson, R. S. Falck, H. A. Siegal, A. Rahman, and L. Li (2005). Respondent-driven sampling to recruit MDMA users: A methodological assessment. *Drug and Alcohol Dependence 78*(2), 147–157.

Wejnert, C. and D. D. Heckathorn (2008, August). Web-based network sampling efficiency and efficacy of respondent-driven sampling for online research. *Sociological Methods & Research 37*(1), 105–134.

Yamanis, T. J., M. G. Merli, W. W. Neely, F. F. Tian, J. Moody, X. Tu, and E. Gao (2013, August). An empirical analysis of the impact of recruitment patterns on RDS estimates among a socially ordered population of female sex workers in China. *Sociological Methods & Research 42*(3), 392–425.