

Schonlau, Matthias; Liebau, Elisabeth

Article — Published Version

## Respondent-Driven Sampling

Stata Journal

**Provided in Cooperation with:**

German Institute for Economic Research (DIW Berlin)

Suggested Citation: Schonlau, Matthias; Liebau, Elisabeth (2012) : Respondent-Driven Sampling, Stata Journal, Stata Press, College Station, Texas, Vol. 12, Iss. 1, pp. 72-93, <http://www.stata-journal.com/article.html?article=st0247>

This Version is available at:

<http://hdl.handle.net/10419/67489>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# THE STATA JOURNAL

## Editor

H. Joseph Newton  
Department of Statistics  
Texas A&M University  
College Station, Texas 77843  
979-845-8817; fax 979-845-6077  
jnewton@stata-journal.com

## Editor

Nicholas J. Cox  
Department of Geography  
Durham University  
South Road  
Durham DH1 3LE UK  
n.j.cox@stata-journal.com

## Associate Editors

Christopher F. Baum  
Boston College

Nathaniel Beck  
New York University

Rino Bellocco  
Karolinska Institutet, Sweden, and  
University of Milano-Bicocca, Italy

Maarten L. Buis  
Tübingen University, Germany

A. Colin Cameron  
University of California–Davis

Mario A. Cleves  
Univ. of Arkansas for Medical Sciences

William D. Dupont  
Vanderbilt University

David Epstein  
Columbia University

Allan Gregory  
Queen's University

James Hardin  
University of South Carolina

Ben Jann  
University of Bern, Switzerland

Stephen Jenkins  
London School of Economics and  
Political Science

Ulrich Kohler  
WZB, Berlin

Frauke Kreuter  
University of Maryland–College Park

Peter A. Lachenbruch  
Oregon State University

Jens Lauritsen  
Odense University Hospital

Stanley Lemeshow  
Ohio State University

J. Scott Long  
Indiana University

Roger Newson  
Imperial College, London

Austin Nichols  
Urban Institute, Washington DC

Marcello Pagano  
Harvard School of Public Health

Sophia Rabe-Hesketh  
University of California–Berkeley

J. Patrick Royston  
MRC Clinical Trials Unit, London

Philip Ryan  
University of Adelaide

Mark E. Schaffer  
Heriot-Watt University, Edinburgh

Jeroen Weesie  
Utrecht University

Nicholas J. G. Winter  
University of Virginia

Jeffrey Wooldridge  
Michigan State University

**Stata Press Editorial Manager**  
**Stata Press Copy Editors**

Lisa Gilmore  
Deirdre Skaggs

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go “beyond the Stata manual” in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the *Stata Journal*, including information for authors, see the webpage

<http://www.stata-journal.com>

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index<sup>®</sup>
- Current Contents/Social and Behavioral Sciences<sup>®</sup>
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch<sup>®</sup>)
- Scopus<sup>™</sup>
- Social Sciences Citation Index<sup>®</sup>

**Copyright Statement:** The *Stata Journal* and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

The articles appearing in the *Stata Journal* may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the *Stata Journal*, in whole or in part, on publicly accessible websites, file servers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the *Stata Journal* or the supporting files understand that such use is made without warranty of any kind, by either the *Stata Journal*, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the *Stata Journal* is to promote free communication among Stata users.

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata, Mata, NetCourse, and Stata Press are registered trademarks of StataCorp LP.

# Respondent-driven sampling

Matthias Schonlau  
University of Waterloo  
Waterloo, Canada  
schonlau@uwaterloo.ca

Elisabeth Liebau  
DIW Berlin  
(German Institute for Economic Research)  
Berlin, Germany  
eliebau@diw.de

**Abstract.** Respondent-driven sampling is a network sampling technique typically employed for hard-to-reach populations (for example, drug users, men who have sex with men, people with HIV). Similarly to snowball sampling, initial seed respondents recruit additional respondents from their network of friends. The recruiting process repeats iteratively, thereby forming long referral chains. Unlike in snowball sampling, it is crucial to obtain estimates of respondents' personal network sizes (that is, number of acquaintances in the target population) and information about who recruited whom. Markov chain theory makes it possible to derive population estimates and sampling weights. We introduce a new Stata command for respondent-driven sampling and illustrate its use.

**Keywords:** st0247, rds, rds\_network, respondent-driven sampling

## 1 Introduction

Some populations are difficult to sample. Consider the homeless: It is not possible to construct a sampling frame because there are no registries or other reasonably complete lists of the homeless. Random-digit dialing does not work because most homeless are not known to carry phones. Address-based sampling procedures also do not work well because the homeless do not have addresses. Invented by Heckathorn in the mid-1990s, respondent-driven sampling (RDS) (Heckathorn 1997, 2002; Salganik and Heckathorn 2004) offers an alternative method that allows inference in populations for which traditional sampling methods are unfeasible or impractical. RDS has proven particularly popular for behavioral surveillance of HIV and has been employed by the Centers for Disease Control and Prevention (Abdul-Quader et al. 2006).

RDS works as follows: seed respondents recruit a fixed number of additional respondents from their network of friends. At each wave, recruits continue to recruit from among their friends. When the desired sample size is reached, the process is terminated. Although this sounds like snowball sampling, RDS differs from snowball sampling in that each respondent must be able to give an estimate of his or her network size (number of people known in the target population; also called “degree”), and it is important to trace who recruited whom. Unlike in snowball sampling, it is also important that recruiting chains are sufficiently long to converge to a sampling equilibrium.

RDS has two additional requirements that do not affect sampling theory but are nonetheless an integral component of the method because they facilitate recruiting. First, there is a double-incentive system. A respondent receives one incentive for par-

ticipating in the survey and receives another incentive for each successfully recruited respondent. Second, recruiting is driven by respondents rather than by interviewers. This feature lends RDS its name. The idea is that respondents are more likely to participate when motivated by their friends, especially when dealing with a sensitive topic such as AIDS or illegal drugs.

RDS is designed for univariate population inference in situations where traditional sampling strategies are not possible. For a categorical variable, the primary purpose of RDS is to obtain unbiased estimates of population proportions. Accordingly, for a categorical variable, the primary goal of RDS software is to compute the population proportions (or, equivalently, the weights that lead to the population proportions). For a continuous variable, the primary purpose of RDS software is to compute the individualized weights, which can then be used to estimate the distribution of that variable. RDS methodology at present has not developed weights for multivariate analyses.

Currently, the only implementations of RDS to our knowledge are the standalone software package RDSAT (Volz et al. 2010)<sup>1</sup> and an independent implementation in the software R that will be made available in the future (Gile 2011). In this article, we introduce an implementation in Stata consisting of two commands: `rds_network` and `rds`. The purpose of `rds_network` is to compute information about respondents' recruiters that is required as input for `rds`. The purpose of `rds` is to compute estimates of population proportions, weights, and other statistics.

The remainder of this article is organized as follows: Section 2 outlines some of the RDS theory including required assumptions. Section 3 contains information about the Stata implementation. Section 4 illustrates RDS by means of a toy example, and section 5 presents a larger example, the SATHCAP study, for the analysis of a categorical variable. Section 6 describes the analysis of continuous variables. Section 7 concludes with a discussion.

## 2 Respondent-driven sampling

Suppose we are interested in the population proportions of a categorical variable such as race and ethnicity or the prevalence of AIDS (yes/no). We will call this variable an analysis variable, and we will call each category (for example, Hispanics) a group. Because we know who recruited whom, it is possible to compute a transition matrix of the analysis variable. RDS makes a Markov assumption: the value of the analysis variable of the recruited (for example, Hispanic ethnicity) depends on the value of the analysis variable of the recruiter but not on that of the recruiter's recruiter.

For Markov chains, the transition matrix converges to a sample equilibrium, and this equilibrium is independent of the seed (Heckathorn 2002, theorem 1). Therefore, it does not matter who the seed respondents are. In practice, so-called social stars (respondents who will be able to recruit respondents easily) are chosen as seed respondents. The proportions in the sample equilibrium do not equal the population proportions, however,

---

1. RDSAT is available for download at <http://www.respondentdrivensampling.org>.

because respondents' inclusion probability is proportional to their degree (the number of people they know in the target population); therefore, people who know more people in the target population are more likely to be recruited into the sample. Likewise, groups with larger average network size will be overrepresented in the equilibrium.

## 2.1 Estimating population proportions

To derive population proportions, reciprocity (or bidirectional) recruiting relations are assumed. This means that if respondent A recruited respondent B, then the reverse could have also occurred. Denote  $k$  as the number of groups for which to compute population proportions. Denote  $N_i$ ,  $i = 1, \dots, k$ , as the sample sizes of group  $i$ , and denote  $D_i$  as the average degree in group  $i$ . Furthermore, denote  $S_{ij}$  as the transition matrix between groups  $i$  and  $j$ ; group  $i$  is the group of the recruiter, and  $j$  is the group of the recruit.

The number of ties originating from members of group 1 is  $N_1 D_1$ , which is the number of respondents in group 1 times the average number of ties of group 1 respondents. The number of ties between groups 1 and 2 can be computed as the number of ties in group 1 times the proportion of ties that go from group 1 to group 2:  $N_1 D_1 S_{12}$ . Because of reciprocity, the number of ties from group 2 to group 1 ( $N_2 D_2 S_{21}$ ) is equally large. Dividing by  $N$  turns the number of ties into population proportions ( $P_1$  and  $P_2$ ), and the following equality is obtained (Heckathorn 2002, eq. 8; Salganik and Heckathorn 2004, eq. 6):

$$P_1 D_1 S_{12} = P_2 D_2 S_{21} \tag{1}$$

The constraint that proportions sum to 1 gives a second equation. If there are only two groups (for example, one group for HIV-positive people and one for everyone else), then we can solve the two equations for the two unknown proportions. If there are more than two groups, then equations analogous to (1) can be constructed for all pairs of groups. For  $m$  groups, that yields  $m \times (m - 1)/2$  equations (plus the constraint that proportions have to sum to 1) for only  $m$  parameters. The problem is over-determined.

This dilemma can be solved, for example, by estimating the unknown parameters using least squares, as in linear regression. Heckathorn's preferred solution, however, is a form of data smoothing (Heckathorn 2002, 24–25). The underlying idea is as follows: if groups recruit with equal effectiveness, then the number of people recruiting out of a group and into a group should be equal. The resulting demographically adjusted recruiting matrix  $\mathbf{R}^*$  can be computed as follows (Heckathorn 2007, sec. 3.2)

$$\mathbf{R}^* = \begin{pmatrix} S_{11}E_1N_r & S_{12}E_1N_r & \cdots & S_{1m}E_1N_r \\ S_{21}E_2N_r & S_{22}E_2N_r & \cdots & S_{2m}E_2N_r \\ \vdots & \vdots & \ddots & \vdots \\ S_{m1}E_mN_r & S_{m2}E_mN_r & \cdots & S_{mm}E_mN_r \end{pmatrix}$$

$N_r$  is the number of recruits and  $E_i, i = 1, \dots, m$ , is the proportion of group  $i$  in the equilibrium. Because each row of the transition matrix is multiplied with a constant,  $E_1 \times N_r$ , the transition probabilities are not affected. The smoothed, demographically adjusted recruiting matrix  $\mathbf{R}^{**}$  is a symmetric matrix where the smoothing consists of averaging:

$$\mathbf{R}^{**} = \begin{pmatrix} R_{11}^* & \frac{R_{12}^* + R_{21}^*}{2} & \cdots & \frac{R_{m1}^* + R_{1m}^*}{2} \\ \frac{R_{12}^* + R_{21}^*}{2} & R_{22}^* & \cdots & \frac{R_{m2}^* + R_{2m}^*}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{R_{m1}^* + R_{1m}^*}{2} & \frac{R_{m2}^* + R_{2m}^*}{2} & \cdots & R_{mm}^* \end{pmatrix}$$

Using the demographically adjusted recruiting matrix  $\mathbf{R}^{**}$ , the transition matrix  $\mathbf{S}^{**}$  can now be computed. Finally, proportion estimates can be obtained by solving the following system of  $m$  equations:

$$\begin{aligned} 1 &= P_1^{**} + P_2^{**} + \cdots + P_m^{**} \\ P_1^{**} D_1 S_{12}^{**} &= P_2^{**} D_2 S_{21}^{**} \\ P_1^{**} D_1 S_{13}^{**} &= P_3^{**} D_3 S_{31}^{**} \\ &\vdots \\ P_1^{**} D_1 S_{1m}^{**} &= P_m^{**} D_m S_{m1}^{**} \end{aligned}$$

The smoothing renders additional equations redundant (Heckathorn 2007, 172). If there are only two groups, the smoothing adjustment does not affect the estimates of the proportions.

## 2.2 Estimating average group degree

The degree is the network size of an individual respondent. The average group degree is the average network size of a group. The average sample degree of a group is an overestimate of average group degree because respondents with a larger network are overrepresented in a sample. The multiplicity estimate of average degree (Rothbart, Fine, and Sudman 1982; Heckathorn 2007, sec. 2.1) for group  $a$  corrects for this:

$$D_a = N_a / \sum_{i=1}^{N_a} (1/D_i)$$

where  $N_a$  is the sample size of group  $a$  and  $D_i$  is the degree of respondent  $i$ . (Seeds are excluded in the calculations of average group degree because seeds were not recruited by peers [Salganik and Heckathorn 2004, 215; Heckathorn 2007, 197].)

### 2.3 Sampling weights

The population weights are computed by dividing the estimated population proportion for a given group equally among all sample members of that group:

$$W_a = P_a/C_a$$

where  $C_a$  refers to the sample proportion of group  $a$  (Heckathorn et al. 2002; Salganik and Heckathorn 2004). All members of group  $a$  receive the same population weight.

The population weight can be separated into a degree component,  $DC_a$ , and a recruitment component,  $RC_a$  (Heckathorn 2007, eq. 26):

$$W_a = (P_a/E_a) \times (E_a/C_a) = DC_a \times RC_a$$

The degree component represents a correction for differential average group degree. If the average group degrees are equal, then  $P_a = E_a$  and the degree component  $DC_a = 1$ . The recruitment component represents differences in recruiting. When the sample proportion equals the equilibrium proportion (when  $C_a = E_a$ ), then the recruitment proportion  $RC_a = 1$ .

This partition leads to the introduction of individualized weights (Heckathorn 2007) or dual-component weights,  $DW_i$ :

$$DW_i = c \times RC_i/D_i \tag{2}$$

where  $c$  is a normalizing constant chosen such that the average individualized weight equals 1. Individualized weights contain two components: degree and recruitment. For the degree component, person-specific estimates exist; for the recruitment component, estimates do not vary within category. Individualized weights are proportional to the inverse of a respondent's degree  $D_i$ , making them robust to large outliers in individual degree. Individualized weights are more commonly used than population weights.

### Convergence

From theoretical work, it is known that convergence to an equilibrium is reached quickly (Heckathorn 2002, theorem 2). Starting with an extreme distribution (100% of respondents in one group and 0% in all other groups), we can simulate how many recruitment waves are required for a given transition matrix to reach equilibrium. Convergence is achieved when two successive simulated recruitment waves do not differ by more than a prespecified convergence tolerance for any group. The Stata implementation of RDS requires that convergence is achieved from all  $m$  extreme distributions. Whether convergence is reached should be recomputed for each variable of interest. However, in



practice, variables with the same number of categories tend to reach convergence at about the same number of iterations.

## Homophily

Homophily measures to what extent respondents prefer to recruit from their own group rather than at random. The probability of selecting from the same group is the probability that selection is controlled by homophily plus the probability of random selection (Heckathorn 2002, 20):

$$S_{aa} = H_a + (1 - H_a)P_a \quad (3)$$

for group  $a$ . Solving for  $H_a$  yields the equation for homophily. Homophily values range from  $-1$  to  $+1$ . The value  $0$  corresponds to random recruitment; the value  $1$  corresponds to always recruiting from one's own group; the value  $-1$  corresponds to never recruiting from one's own group. Moderate homophily is not problematic. If homophily is very large (for example,  $0.9$ ), however, the transition matrix may take a long time to converge, which may be a sign that the groups are not networked.

## Assumptions

The theory underlying RDS is based on a set of five assumptions:

- *Assumption 1: Reciprocity.* The reciprocity assumption implies that if respondent A recruited respondent B, then in principle, B could have recruited A. In practice, this assumption is tested by including a survey question about the relationship between the respondent and the recruiter. The assumption is violated if many of the recruited persons are strangers.
- *Assumption 2: Networked population.* All respondents are interconnected. This assumption would be violated, for example, if the target population consisted of rivaling gangs who do not communicate with one another. The solution in this case would be to conduct separate RDS samples for each of the noncommunicating groups. If the number of waves required to reach equilibrium for any variable is large, one may suspect a problem.
- *Assumption 3: Sampling with replacement.* Sampling with replacement means that in principle, a respondent could be contacted again and the respondent would participate a second time. In practice, a respondent would probably refuse to fill out the questionnaire a second time. In addition, duplicate respondents are usually actively screened out to prevent fraud related to obtaining multiple incentives. However, assuming that the sample is only a small fraction of the total population, this assumption can be ignored.
- *Assumption 4: Network size.* Respondents can accurately report their personal network size. Biased estimates (for example, consistent underestimation or overestimation of network size) are unproblematic because respondents uniformly un-

derestimate or overestimate their network size (Wejnert 2009, sec. “Degree estimation”). There is ongoing concern that self-reported network sizes may be problematic (Wejnert and Heckathorn 2008, 119), though there is also evidence that different ways of assessing network size lead to essentially the same result (Wejnert 2009).

- *Assumption 5: Random recruitment.* Respondents recruit from their network at random. To verify this assumption, one might ask about attributes (for example, gender and race) of respondents’ networks and compare expected characteristics with actual sample composition. This assumption does not hold in practice and is one reason why respondent-driven sampling should be used only when traditional sampling methods are unfeasible.

### Nonresponse

Nonresponse deserves a mention because it plays a large role in traditional sampling methodology. Nonresponse matters in RDS but is not talked about much. Because respondents (rather than interviewers) recruit respondents, it is not possible to estimate nonresponse unless the respondents are interviewed a second time.

## 3 Stata implementation

RDS data look different from regular data because they embed the recruiting network structure. Table 1 gives an example of minimum data requirements: ID (coupon number), network size, referral coupon numbers (this example has six of them), and an analysis variable (this example uses race/ethnicity). The respondents need not be ordered in any way. Missing referral coupons indicate that the respondents were not given a full set of referral coupons. In table 1, no respondent was given more than four coupons. Whether a referral coupon was handed out without leading to a new respondent or whether no coupon was handed out because sampling was terminated does not affect estimation.

Table 1. Example data for RDS. The seed id is 172.

id	netsize	numcpn1	numcpn2	numcpn3	numcpn4	numcpn5	numcpn6	ethnic
40282	1	40307	40306	30306	30305	.	.	other
40361	3	40374	40375	30375	30376	.	.	white
172	18	40274	40275	.	.	.	.	other
40360	289	40383	40458	.	.	.	.	white
40383	12	30453	30454	40446	40447	.	.	black
40274	7	40335	40278	.	.	.	.	other
40275	4	40282	40283	.	.	.	.	other
40283	2	40361	40360	30359	30360	.	.	white
40278	6	40308	40309	.	.	.	.	white

The analysis is split into two Stata commands: `rds_network` and `rds`. The command `rds_network` determines the longest chain length (needed to assess convergence to the equilibrium), and it collects information about the recruiter of a respondent (variables `recruiter_id()` and `recruiter_var()`).

### 3.1 Syntax

The syntax is as follows:

```
rds_network varname, id(varname) coupon(str) ncoupon(#) degree(varname)
  [ancessor(newvar) depth(newvar) recruiter_id(newvar)
  recruiter_var(newvar) ]
```

The options `id()`, `coupon()`, `ncoupon()`, and `degree()` specify the unique coupon code of respondents, the stems of the variable names for their referral coupons, the number of referral coupons, and the network size, respectively. The `rds_network` command should always be called with the full RDS network for a given site. If a respondent is removed, the recruitment chain is broken into subchains before and after the deleted respondent. `rds_network` intentionally does not support `if` or `in` qualifiers.

The four optional options of the command generate four additional variables, as shown in the syntax. `ancestor()` contains the ID of the seed through which the respondent was recruited. `depth()` contains the depth of the recruiting tree for a given recruit. Seeds have depth 0, their recruits have depth 1, and so forth. `recruiter_id()` contains the ID of the respondent's recruiter. `recruiter_var()` contains the value of `varname` of the respondent's recruiter.

`rds` is the main estimation command. The recruiter variables `recruiter_id()` and `recruiter_var()`, computed by `rds_network`, are now required as input variables. The syntax is as follows:

```
rds varname [if] [in], id(varname) degree(varname) recruiter_id(varname)
  recruiter_var(varname) [wgt(newvar) wgt_pop(newvar) detail convtol(#)
  network_size_method(str) ]
```

Degree refers to the estimate of network size (number of friends in the target population). Optionally, `wgt()` generates a variable with individualized sampling weights. Additional options (related to convergence to the equilibrium and the algorithm used to compute average network size) are also available.

### 3.2 Input validation and potential errors

The `rds_network` command verifies that the respondent ID and all referral coupons are unique. `rds_network` also verifies that there is no self-referral (a respondent's coupon

points to himself or herself). Furthermore, `rds` will give an error if the estimated equilibrium proportion for a group is 0. Missing values for network size (degree) are allowed; missing values for the analysis variable specified in *varname* are not allowed. All network sizes (degrees) must be positive.

### 3.3 Standard errors and the bootstrap

Standard errors and confidence intervals can be estimated via Taylor linearization (the `svy` routines in Stata) or bootstrapping. The bootstrapping approach is preferred because of concern that the other approach does not adequately reflect variability in the sampling process. The bootstrap method is also implemented in RDSAT (Volz et al. 2010). Even so, recent simulations suggest that confidence intervals are typically too narrow (Goel and Salganik 2010). In Stata, `svy` routines can be applied as follows:

```
svyset [pweight=myweight]
svy: proportion myvar
```

Standard errors of the proportions using a traditional nonparametric bootstrap of the ties between recruiter and recruited are computed as follows:

```
bootstrap _b, reps(1000): rds varname, id() recruiter_id() [...]
```

This results in a bootstrap sample of the observed transitions.

The software RDSAT uses a slightly different bootstrapping procedure (Heckathorn 2002, 27–29; Salganik 2006). Roughly, RDSAT simulates a new recruiting chain using the estimated transition matrix. The first simulated recruit is chosen arbitrarily. Each following simulated recruit is selected at random based on the probabilities specified in the transition matrix. In RDSAT, the bootstrapping procedure is applied to the least-squares algorithm, not to the smoothing algorithm (RDS 2006, 30).

## 4 Small toy example

We present a toy example from Heckathorn (2007, app. A). The purpose of this section is to illustrate the concepts introduced earlier but also to validate the Stata command with calculations by hand. This example has 20 respondents (table 2). The outcome variable is colored with the levels “red” (indexed as group 1) and “blue” (indexed as group 2). In figure 1, circles correspond to red respondents and squares to blue respondents.

The `id` variable has 20 unique values; they need not be consecutive or ordered as in table 2. Respondent order does not affect calculations. In this example, respondents received three coupons labeled `ref1`, `ref2`, and `ref3`. Table 2 shows only those coupons that led to a new recruit—where a coupon corresponds to a respondent ID in another row. Often, an interviewer gives a respondent a coupon but the respondent is unable to recruit someone with that coupon. In those cases, whether missing values or coupon numbers are listed does not affect the calculations.

Table 2. Toy dataset from Heckathorn (2007, app. A)

	id	ref1	ref2	ref3	degree	color
1.	1	2	3	4	8	red
2.	2	5	6	.	8	red
3.	3	7	8	9	.	red
4.	4	10	.	.	10	blue
5.	5	11	12	13	5	red
6.	6	.	.	.	7	blue
7.	7	14	15	16	4	blue
8.	8	17	.	.	7	blue
9.	9	18	19	20	5	red
10.	10	.	.	.	2	blue
11.	11	.	.	.	4	red
12.	12	.	.	.	.	blue
13.	13	.	.	.	3	red
14.	14	.	.	.	2	blue
15.	15	.	.	.	3	blue
16.	16	.	.	.	3	red
17.	17	.	.	.	7	blue
18.	18	.	.	.	3	blue
19.	19	.	.	.	5	red
20.	20	.	.	.	8	blue

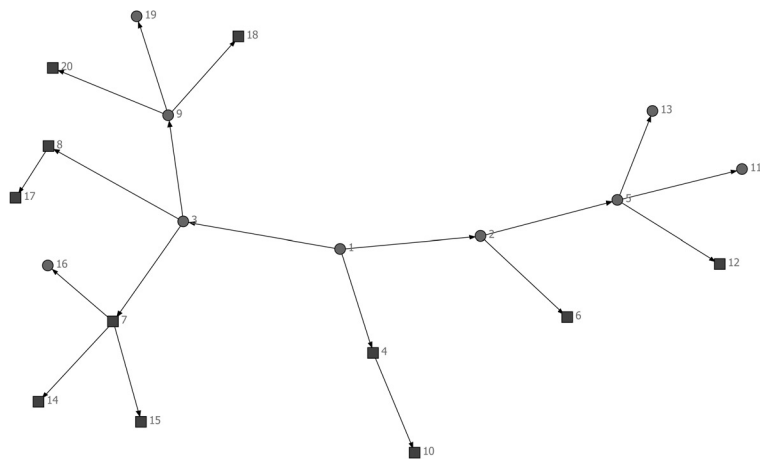


Figure 1. Network graph for the toy example. Each respondent belongs to one of two categories: “red” (circle) or “blue” (square).

Table 3 contains some information about the structure and various outcomes that are computed. The total count in the observation matrix is 19, corresponding to 19 transitions (20 respondents minus one seed). Red recruiters recruit both red and blue recruits seven times each. Conditional on being a red recruiter, the probability of recruiting a red recruit is 50%. Correspondingly, the observed transition matrix (table 3) from red to red is 0.5.

Table 3. Intermediate output from the rds command for the toy example

```

Number of categories of (key): 2
Required referral length until convergence: 4
Method to compute Av. Network Size Method = multiplicity

Observation matrix
      red blue
red    7   7
blue   1   4

Transition Matrix (Before Smoothing)
      red blue
red   .5   .5
blue  .2   .8

Demographically adjusted matrix
      red   blue
red  2.7142857  2.7142857
blue 2.7142857 10.857143

Data-Smoothed Recruitments
      red   blue
red  2.7142857  2.7142857
blue 2.7142857 10.857143

Transition Matrix
      red blue
red   .5   .5
blue  .2   .8

```

The demographic adjustment and data-smoothing steps are needed to address the over-determination that arises when there are more than two categories. In this example, there are only two categories, and these additional steps do not affect the estimate of transition matrix. Therefore, table 3 shows the same transition matrix before and after smoothing.

The multiplicity estimate for degree is  $D_i = n_i / \{\sum_j (1/d_{ij})\}$ . For red respondents, this estimate gives  $D_1 = 7 / (1/8 + 1/5 + 1/4 + 1/3 + 1/5 + 1/5 + 1/3) = 4.26$ . For comparison, the average estimate for degree is biased and larger,  $(8 + 5 + 4 + 3 + 5 + 5 + 3) / 7 = 4.71$ . The seed is typically not included in the calculation of degree. Similarly,  $D_2 = 3.89$ . Average and multiplicity estimates of degree are reported in table 4.

Table 4. Output from the `rds` command for the toy example

	red	blue
Categories	0	1
SampleSize	9	11
Recruits	8	11
Seeds	1	0
SampleProportion	.45	.55
Equilibrium	.28571429	.71428571
AverageDegree	4.7142859	5.3000002
MultiplicityDegree	4.2639594	3.8906901
Homophily	.31750811	.25203045
Weight	.59420125	1.3320172
RecruitmentComponent	.63492063	1.2987013
DegreeComponent	.93586697	1.0256532
PopulationProportion	.26739056	.73260944

To estimate population proportions,  $p_1$  and  $p_2$ , the following equation system needs to be solved:  $p_1 D_1 S_{12} = p_2 D_2 S_{21}$  and  $1 = p_1 + p_2$ . Because there are only two categories, there is a closed-form solution. Two equations with two unknowns can be solved and yield  $p_1 = D_2 S_{21} / (D_1 S_{12} + D_2 S_{21})$ . Substituting gives  $p_1 = 3.8907 \times 0.2 / (4.2639 \times 0.5 + 3.8907 \times 0.2) = 0.2674$ . It follows that  $p_2 = 1 - 0.2674 = 0.7326$ . Population proportions are reported in table 4.

The population weight is computed by dividing the population estimate by the sampling fraction. For red respondents, the weight is  $w_1 = 0.2673 / (9/20) = 0.5942$ . For blue respondents, the weight is  $w_2 = 0.7326 / (11/20) = 1.332$ . Population weights are also reported in table 4. All individuals in the same category have the same population weight.

In addition, there are individual weights that take into account estimates for individual degree. Because each individual has a different weight, individualized weights cannot be displayed in table 4. However, both components of population weights, the recruiting and degree components, are displayed. For example, multiplying the two components for the red category yields the population weight  $0.6349 \times 0.9359 = 0.5942$ . The individualized weights (also called dual-component weights) are computed by dividing the recruitment component by individual degree:  $DW_i = 0.6349 / D_i \times c$  where  $i$  enumerates all respondents and  $c$  is a normalizing constant. For example, a respondent with an individual degree of 8 (that is, 8 people known in the target population) has individual weight  $0.6349 / 8 \times c = 0.0793 \times c$ . The exact value of  $c$  does not matter. From sampling theory, we know that multiplying sampling weights with a constant does not affect weighted analyses.

Table 4 gives the estimated homophily for group 1 as 0.3175. Using  $S_{11} = 0.5$  from the final transition matrix in table 3, (3) holds:  $0.5 = 0.3175 + (1 - 0.3175) \times 0.2674$ . The calculation for homophily for group 2 is analogous.

## 5 SATHCAP study example: Categorical variable

The Sexual Acquisition and Transmission of HIV Cooperative Agreement Program, also known as SATHCAP, applied RDS to sample men who had sex with men (MSM) and drug users (DU) in four metropolitan areas (Chicago, IL; Los Angeles, CA; Raleigh-Durham, NC; and St. Petersburg, Russia) (Iguchi et al. 2009). In addition, sex partners of this target population were sampled but were not part of the official RDS sample. The SATHCAP study used an innovative dual-recruitment method with different coupon colors for different segments of the target population to ensure both MSM and DU were sampled. Public release data are available through a website.<sup>2</sup> The data used here to illustrate RDS correspond to phase II at the Los Angeles site. We first analyze the network:

```
rds_network ethnic, id(id) coupon(numcpn) ncoupon(6) degree(netsize) ///
  recruiter_id(p_id) recruiter_var(p_key) depth(depth) ancestor(ancestor)
```

`rds` output (not shown) notes that there are 117 seed respondents.<sup>3</sup> This is an unusually large number. The maximum chain or referral length is 18 (not counting the seed). The output also lists the length of the maximal referral chains for each individual seed (table 5 gives an excerpt). Most seeds in table 5 do not recruit anyone. Figure 2 shows the sample size by referral depth (using the variable `depth` specified above). Seeds have depth 0. The sample size decreases while the referral depth increases. Based on calculations with the variable specified in option `ancestor()`, it turns out that 13 of the 117 seeds produce 71% of the sample. It is common that only a small percentage of seeds are highly productive (Malekinejad et al. 2008).

---

2. You can access the data at <http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/29181> (registration is required for data access).

3. The number of seeds reported by Iguchi et al. (2009) is somewhat lower. During field work, referral IDs of some respondents were lost. Rather than reporting the number of intended seeds, the command reports the number of actual seeds (namely, respondents without a recruiter).



Table 5. Excerpt of output from `rds_network` identifying seeds and the length of each seed's recruiting chain. Most seeds shown fail to recruit anyone.

Seed	MaxDepth
<i>(output omitted)</i>	
2309	0
2378	0
2389	0
2395	0
2421	0
2462	2
2480	18
2499	1
2503	0
2602	0
<i>(output omitted)</i>	

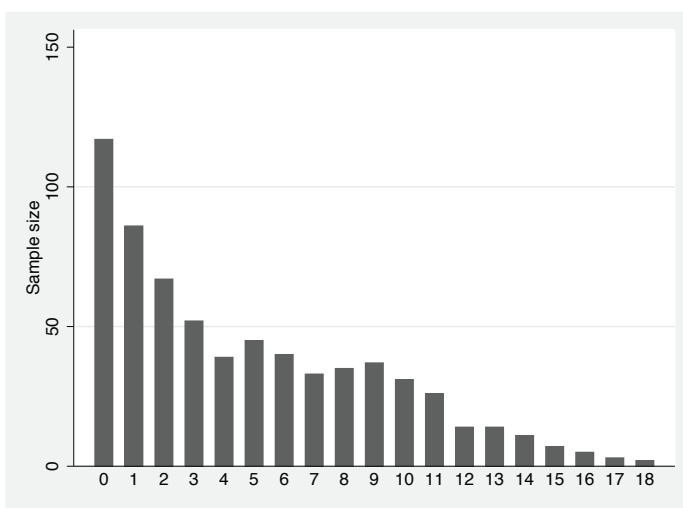


Figure 2. Sample size (excluding sex partners) by depth of the referral chain. Depth 0 corresponds to seed respondents.

Having computed the recruiter information, we can now proceed with assessing convergence and estimation:

```
rds ethnic, id(id) degree(netsize) recruiter_id(p_id) recruiter_var(p_key) ///
  wgt(wgt) wgt_pop(wgt2)
```

Originally, the variable `netsize` was calculated from three different questions corresponding to the number of MSM, DU, and their overlap. Inconsistent answers could result in negative values and zeros. We set those values to missing.<sup>4</sup>

### Convergence

The `rds` output (not shown) states that the required minimum referral length until convergence is 5. From the `rds_network` output, we know that the longest chain in our data has length 18. Therefore, convergence for the variable `ethnic` is achieved. The required referral length needed to achieve convergence is simulated based on the transition matrix. It is also interesting to see how the sampling proportions converge. Figure 3 shows the cumulative sampling proportion of racial/ethnic groups calculated for all data up to a maximal depth or chain length. Indeed, we find that the proportions converge while the maximal wave increases, although in practice, the convergence may have taken a little longer.

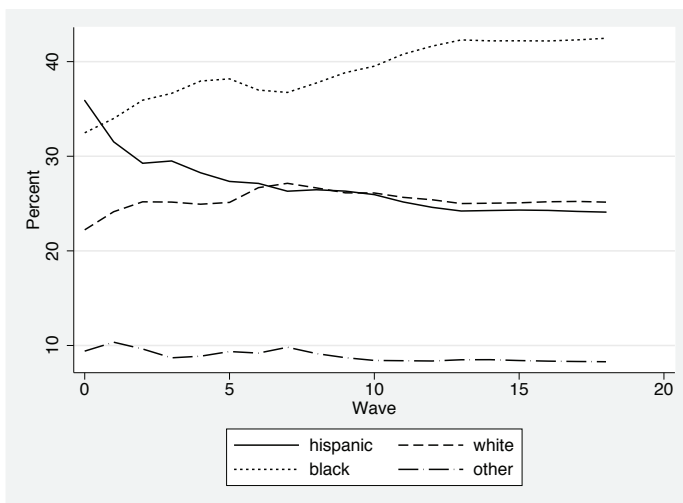


Figure 3. Percentage of four racial/ethnic groups for increasing length of the recruitment chain. Percentages are based on cumulative samples up to a given chain length.

### Estimation

The final transition matrix is shown in table 6. `rds` output (not shown) contains intermediate matrices ( $S$ ,  $R^*$ ,  $R^{**}$ , and  $S^{**}$ ) for the calculation of this transition matrix, and output contains the matrix of observed counts. If there are only two groups, the

4. Setting zeros to one is less attractive because it would give those individuals very high weight. RDSAT routinely treats zeros as missing.

estimates of the initial and the final transition matrices are identical. In the transition matrix, we notice that black respondents recruit other black respondents 67.5% of the time. We will return to this in the context of homophily below.

Table 6. Estimated final transition matrix

	hispanic	white	black	other
hispanic	0.421	0.243	0.252	0.084
white	0.246	0.508	0.200	0.046
black	0.111	0.127	0.675	0.087
other	0.224	0.293	0.362	0.121

Table 7 displays estimation results. The sample size is the sum of the number of seeds and the number of recruits. There were seeds in all four racial/ethnic categories. There are three different proportion estimates: sample proportion, proportion in the equilibrium, and population proportion. Proportion in the equilibrium refers to the theoretical sampling proportion if the transition matrix has reached its equilibrium. If network size (degree) is constant, population proportions equal the equilibrium proportions. In practice, the network size varies, and recruits who have a larger network are more likely to be sampled. The population proportion is an average-network-size-adjusted equilibrium proportion.

Table 7. Estimation results

	hispanic	white	black	other
Categories	1	2	3	4
SampleSize	160	167	282	55
Recruits	118	141	244	44
Seeds	42	26	38	11
Sample_Proportion	0.241	0.252	0.425	0.083
Equilibrium	0.226	0.268	0.427	0.078
AverageDegree	15.939	19.978	17.731	13.488
MultiplicityDegree	4.432	5.491	5.309	5.021
Homophily	0.217	0.344	0.448	0.045
Weight	1.081	0.992	0.967	0.959
RecruitmentComponent	0.939	1.067	1.006	0.943
DegreeComponent	1.151	0.929	0.961	1.016
PopulationProportion	0.26	0.249	0.411	0.079

There are two measures of average network size in table 7: “average” and “multiplicity”. The naïve estimate “average” does not take into account that respondents with a larger network are more likely to be recruited into the sample. Therefore, the sample average for a group (for example, Hispanics) overestimates the population average. The “multiplicity” estimate corrects for this. If the network sizes were constant, then the two estimates would give identical results.

The population sampling weights are designed to reproduce the estimated population proportion. The commands

```
svyset [pweight=wgt2]
svy: proportion ethnic
```

(where the variable `wgt2` was specified as an option in `rds`) reproduce the population proportions exactly. The variable `weight` contains only four distinct values corresponding to the four racial/ethnic categories.

Table 8 shows a comparison of estimated standard errors using Taylor linearization, bootstrap using RDS (estimates based on the smoothing algorithm introduced in section 2,  $n = 2500$ ), and the bootstrap from RDSAT (estimates based on the least-squares algorithm,  $n = 2500$ ). The standard errors based on Taylor linearization are much smaller than the two bootstrap estimates. The two bootstrap standard errors are similar to one another.

Table 8. Three estimates of the standard error of the population proportions of race/ethnicity: 1) Standard error based on Taylor approximation (using `svyset`), 2) bootstrap standard error ( $n = 2500$ ) using `rds` in Stata, and 3) bootstrap standard error ( $n = 2500$ ) using the RDSAT software

	Taylor linearized standard error	Bootstrap standard error	RDSAT bootstrap standard error
hispanic	0.018	0.033	0.036
white	0.017	0.033	0.033
black	0.019	0.041	0.042
other	0.010	0.017	0.019

### Homophily

Homophily is a diagnostic statistic that estimates to what extent respondents tend to recruit within group rather than at random. For example, table 7 shows that black respondents recruit other black respondents 44.8% of the time, and they recruit at random from any of the four groups 55.2% of the time. Only very large homophily values (for example, 0.9) would raise a concern.

### Reciprocity

The SATHCAP questionnaire contained a question about the relationship between the respondent and his or her recruiter. Only 4.5% of the recruited respondents described their recruiter as a stranger. This percentage is small and does not raise concerns. There are no guidelines of what percentage is considered too large.

### Networked population

The number of iterations required to achieve convergence did not raise a red flag for any variable we looked at. Likewise, we found no anomalies in the corresponding transition matrices.

### Random recruitment

Iguchi et al. (2009) argued it may not always be obvious to respondents how their friends self-identify in terms of race/ethnicity. Therefore, they looked at other variables including gender to verify the random recruitment assumption. Recruits are male 88.7% of the time (excluding a small number of transsexuals and excluding sex partners). Recruits reported that 71.4% of their network is male. The difference is significant:  $X^2(1) = 74.0, p < 0.001$ . Therefore, the random-recruitment assumption is violated with respect to gender. Iguchi et al. (2009) argued that it is not clear whether the differences are due to measurement error in the self-reported characteristics of their network or whether they are due to nonrandom recruitment.

## 6 SATHCAP study example: Continuous variable

Continuous variables must first be converted into categorical variables. Once converted, the analysis of continuous variables is identical to that of categorical variables leading to (2) for individualized weights. At that point, individualized weights can be applied to the continuous variable itself (rather than its categorized version).

Of course, the question arises regarding how many categories to use. Few categories may result in a loss of information because the recruitment component of individualized weights does not vary within category; too many categories will result in a sparse transition matrix and numerical instabilities. Therefore, the number of categories will also depend on the sample size. One option is to consider multiple choices for the number of categories and to find a sweet spot where estimates appear to converge. If the number of categories is too large, estimates may diverge again while the recruitment component becomes unstable (Heckathorn 2007, 178). A continuous variable is typically split up into tertiles for three categories, quartiles for four categories, and so forth.

We illustrate this with the number of HIV tests as reported by individuals in the SATHCAP study. The number of HIV tests reported range from 0 to 555 visits; however, 90% of the respondents estimate between 0 and 10 visits. Figure 4 shows estimates and their confidence intervals. The two estimates shown at the left are the unweighted estimate and an estimate using inverse degree as a weight. The inverse degree estimate ignores the recruitment component of individualized weights in (2). The remaining weighted estimates are based on categorizing the number of HIV tests into 2 to 10 groups. Although the means are not significantly different from one another, the largest shift in mean occurs between the unweighted estimate and the estimate using inverse degree as a weight. Therefore in this example, the recruitment component did not affect estimates much. Based on figure 4, the estimated average number of HIV tests is about 6.5.

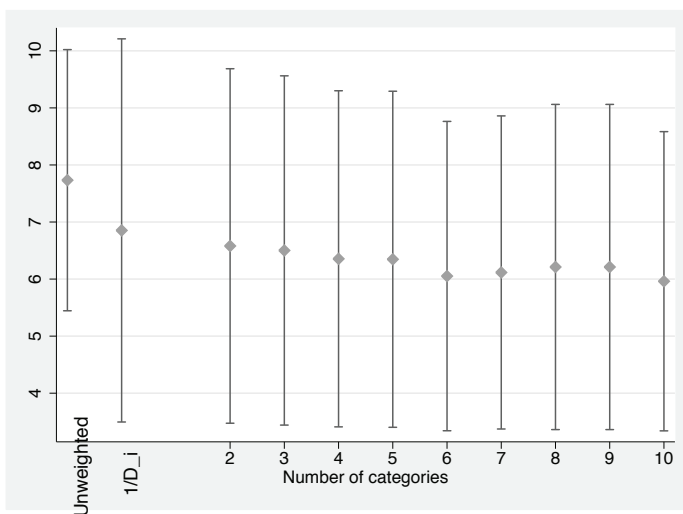


Figure 4. Estimated mean number of HIV visits and 95% confidence intervals. The two estimates to the left are the unweighted estimate and the estimate using the inverse of degree as a weight. The remaining estimates are based on splitting the number of HIV visits into 2–10 categories or groups.

## 7 Discussion

The integration of RDS within the Stata programming environment easily accommodates additional programming needs that require special-purpose programming in a standalone package. For example, the bootstrap routine can be used with `rds` as explained earlier. Unusually large outliers of network size can be “pulled in” by setting large values to a user-defined maximum. Furthermore, some researchers might want to analyze data only after reaching equilibrium. If the equilibrium is reached after five referral waves, this can be accomplished as follows:

```
rds_network varname, depth(mydepth) [...]
rds varname if mydepth>=5, [...]
```

Weights can be poststratified to known totals by using the `poststrata` option in `svyset`, or equivalently, a new adjusted weight variable can be computed using `svygen poststratify`.

There is currently no consensus of how to conduct regression with RDS data. Sampling weights are calculated based on a single analysis variable, such as race/ethnicity. In multivariable analyses such as regression, it is unclear what to do. The best practice currently is to conduct a sensitivity analysis (Johnston et al. 2010) using the weight constructed for the dependent variable.

RDS is an area of active research and the literature is expanding. In practice, there are numerous implementation challenges such as defining eligibility criteria (Johnston 2008; Johnston et al. 2008). Relative to a simple random sample, the RDS sample size should be at least twice as large to account for design effects, and it may need to be even larger (Salganik 2006; Goel and Salganik 2010). RDS has also been conducted through a web survey (Wejnert and Heckathorn 2008).

More recently, the so-called RDS II estimator (Volz and Heckathorn 2008) has been derived. This estimator corresponds to using the inverse individual degree as a sampling weight. However, the estimate of variance is more complex because of the network dependencies. The RDS II estimator is particularly appealing for continuous variables because continuous variables need not be split into categorical variables as described in section 6. Building on the RDS II estimator, an estimator that does not require the assumption “sampling with replacement” has been derived (Gile 2011). We expect many more exciting developments on RDS.

## 8 Acknowledgments

We are grateful to the SATHCAP project for allowing us to use some of their data in the example. Data from this project will be made publicly available at a later date. We are grateful to Juergen Schupp, whose intriguing and challenging questions inspired this project. This work was conducted while the first author was affiliated with DIW Berlin.

## 9 References

- Abdul-Quader, A. S., D. D. Heckathorn, C. McKnight, H. Bramson, C. Nemeth, K. Sabin, K. Gallagher, and D. C. Des Jarlais. 2006. Effectiveness of respondent-driven sampling for recruiting drug users in New York City: Findings from a pilot study. *Journal of Urban Health* 83: 459–476.
- Gile, K. J. 2011. Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association* 106: 135–146.

- Goel, S., and M. J. Salganik. 2010. Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences* 107: 6743–6747.
- Heckathorn, D. D. 1997. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems* 44: 174–199.
- . 2002. Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems* 49: 11–34.
- . 2007. Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology* 37: 151–207.
- Heckathorn, D. D., S. Semaan, R. S. Broadhead, and J. 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: 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 (Suppl. 1): 5–31.
- Johnston, L., H. O’Bra, M. Chopra, C. Mathews, L. Townsend, K. Sabin, M. Tomlinson, and C. Kendall. 2010. The associations of voluntary counseling and testing acceptance and the perceived likelihood of being HIV-infected among men with multiple sex partners in a South African township. *AIDS and Behavior* 14: 922–931.
- Johnston, L. G. 2008. Behavioural surveillance: Introduction to respondent driven sampling. Participant manual, Global AIDS Program (GAP) Surveillance Team, Centers for Disease Control and Prevention.  
<http://globalhealthsciences.ucsf.edu/sites/default/files/content/pphg/surveillance/modules/global-trainings/respondent-driven-sampling-2008.pdf>.
- Johnston, L. G., M. Malekinejad, C. Kendall, I. M. Iuppa, and G. W. Rutherford. 2008. Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: Field experiences in international settings. *AIDS and Behavior* 12 (Suppl. 1): 131–141.
- Malekinejad, M., L. G. Johnston, C. Kendall, L. R. Kerr, M. R. Rifkin, and G. W. Rutherford. 2008. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: A systematic review. *AIDS and Behavior* 12 (Suppl. 1): 105–130.
- RDS. 2006. *RDS Analysis Tool v5.6 User Manual*. Ithaca, NY: RDS.
- Rothbart, G. S., M. Fine, and S. Sudman. 1982. On finding and interviewing the needles in the haystack: The use of multiplicity sampling. *Public Opinion Quarterly* 46: 408–421.



- Salganik, M. J. 2006. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health* 83 (Suppl. 1): 98–112.
- Salganik, M. J., and D. D. Heckathorn. 2004. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology* 34: 193–239.
- Volz, E., and D. D. Heckathorn. 2008. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics* 24: 79–97.
- Volz, E., C. Wejnert, I. Degani, and D. D. Heckathorn. 2010. *Respondent-Driven Sampling Analysis Tool (RDSAT) Version 6.0 Beta*. Ithaca, NY: Cornell University.
- Wejnert, C. 2009. An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. *Sociological Methodology* 39: 73–116.
- Wejnert, C., and D. D. Heckathorn. 2008. Web-based network sampling: Efficiency and efficacy of respondent-driven sampling for online research. *Sociological Methods and Research* 37: 105–134.

#### **About the authors**

Matthias Schonlau is a professor of statistics at the University of Waterloo in Waterloo, Canada. Previously, he spent one year at the German Institute for Economic Research (DIW Berlin) and many years at the RAND Corporation in Santa Monica, CA, and Pittsburgh, PA. His research interests include survey sampling and methodology.

Dr. Elisabeth Liebau is a survey manager with the SOEP group at the German Institute for Economic Research (DIW Berlin). She is especially interested in various techniques of migrant sampling and migration/integration analysis.