# RECRUITING “HIDDEN” POPULATIONS: METHODOLOGICAL CONSIDERATIONS FOR ADAPTING RESPONDENT-DRIVEN SAMPLING TO PRESERVE PARTICIPANT ANONYMITY IN RESEARCH IN NATIVE COMMUNITIES

by

Katrina L. Ramsey

A THESIS

Presented to the Department of Public Health & Preventive Medicine

and the Oregon Health & Science University School of Medicine

in partial fulfillment of

the requirements for the degree of

Master of Public Health

June 2007

School of Medicine
Oregon Health & Science University

# CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Katrina L. Ramsey

has been approved

Jodi A. Lapidus, PhD

Victoria Warren-Mears, PhD, RD

Dongseok Choi, PhD

# TABLE OF CONTENTS

## List of Tables

## List of Figures

## Acknowledgments

For their contributions, I wish particularly to thank the following:

- Jodi Lapidus, my advisor, whose skill and patience and commitment to community-based research made the entire project possible;
- Dongseok Choi and Victoria Warren-Mears, for reading and improving many drafts;
- Jeanne Bertolli and Patrick Sullivan of the National Center for HIV, STD, and TB Prevention, Centers for Disease Control and Prevention, for their useful suggestions;
- Project Red Talon and the rest of The Northwest Portland Area Indian Health Board for their support of this project and many other health-related efforts on the part of tribal communities;
- Many other urban American Indian and Alaska Native community members who contributed their thoughts, time, and experience to the original HITS project;
- My parents and my husband, Aaron, who were sounding boards and coaches, and my son, Luke, whose “help” sometimes made progress seem impossible but also supplied the joy to keep me going.

# Abstract

### Context

Respondent-driven sampling (RDS) is useful for accessing hard-to-reach populations but requires tracking respondents in a way that may not be feasible in anonymous surveys or in some communities.

### Purpose

This study explores modifications to RDS that preserve respondent anonymity and provides recommendations for implementing RDS.

### Methods

Results were simulated for a hypothetical study that tracked recruitment in groups of respondents rather than in person-to-person recruitment networks. In regular RDS, recruitment network data are used to generate specific population estimates. In anonymous RDS, network data are lacking; instead, group data are used to generate ranges of possible scenarios. These possible scenarios were compared to the results that would have been obtained under regular RDS.

### Findings

By simulating possible recruitment scenarios, it is possible to generate point and interval estimates for the distribution of characteristics in a sample that has reached equilibrium. Mean, median, and probability-weighted estimates produce intervals that vary in precision.

## Conclusions

Modifying RDS to preserve respondent anonymity requires sacrificing some precision in analyzing the sample and generating population estimates. Anonymous methods most closely approximate regular RDS methods when the sampling scheme has been successful; however, the degree to which the sampling scheme has been successful is unknown in anonymous RDS. Further study should provide results that better reflect the overall population. Sampling methods may also be developed to provide more data about recruitment patterns without identifying individual respondents. In spite of these weaknesses, anonymous RDS is a systematic method that, in contrast to the convenience samples that are more commonly used in hard-to-reach populations, provides tools to evaluate a study’s external validity.

# Introduction

## Motivating example

A 2001 survey of urban American Indians and Alaska Natives (AI/ANs) posed a particular challenge in recruiting participants. The survey was a slightly modified version of the Center for Disease Control and Prevention's (CDC) HIV Testing Survey (HITS) designed for urban AI/ANs who either had HIV or who were at risk for contracting the disease and who lived in the metropolitan area of Portland, Oregon.[1] The target population included AI/AN intravenous drug users, men who have sex with men, and high-risk heterosexuals. During the formative phase of the study, approximately 25 one-on-one interviews and 2 focus groups were conducted to find out, among other things, whether the sampling methods usually used with this survey would be successful in recruiting a representative sample of high-risk AI/ANs in the Portland area.

Most CDC-conducted HITS use systematic venue-based sampling. Research staff systematically recruit individuals from venues where intravenous drug users (street corners or needle exchange programs), men who have sex with men (gay bars), or high-risk heterosexuals (STD clinics) congregate.[2-5] For the AI/AN study the formative interviews identified only one location where high-risk AI/ANs might congregate. This location was a Native treatment facility, and though it might have been possible to generate a large sample there, the client population excluded not-in-treatment drug users and many individuals with high-risk sexual behaviors. The treatment facility attracted clients from areas outside of Portland, whereas the goal of the study was to assess risk behaviors of Portland AI/AN. People in treatment were also likely to have recently changed high-risk behaviors, which would have led to bias in reporting current risk behaviors. Systematic venue-based sampling was not feasible for this study.

Without the possibility of using standard CDC sampling methods and lacking a sampling frame for AI/ANs in Portland, organizers turned to a non-random sampling method. Respondent-driven sampling (RDS) is a chain-referral sampling method[6] that has been used to recruit “hard to reach” or “hidden” populations—groups that are difficult to find because they do not appear on lists or gather in centralized locations, or because they are defined by behaviors or characteristics they keep private. RDS has been implemented successfully in studies of jazz musicians,[7] intravenous drug users,[8, 9] stimulant drug users in rural areas,[10] Latino gay men,[11] transgender persons,[12] and sex workers in Vietnam.[13] The first public health-related study to use this method was a long-term HIV/AIDS intervention, detailed in a 1997 and later publications by Heckathorn et al.[8, 14, 15]

RDS is closely related to another type of chain-referral sampling known as snowball sampling.[6, 16] Both begin with a core group of respondents, called “seeds.” Additional respondents are recruited from among the seeds’ social contacts, then successively from the peers of each “wave” of respondents. In snowball sampling, researchers ask for the names and contact information of others with similar risk factors and then sample randomly from those contact lists. Because it requires asking people to reveal potentially harmful information about others without their consent, this method is difficult to justify. In RDS, rather than asking for names and contact information, researchers enlist respondents to recruit their peers, typically by providing respondents with a number of recruitment coupons to give to their social contacts. Those contacted by their peers decide on their own whether they wish to participate, and contact the research staff directly.

Because chain-referral samples lack the sampling frames of traditional probability samples, it is difficult to judge how representative they are of the hidden population they are designed to tap into. As a rule, people tend to associate with others who are similar (in some respect) to them, and if those preferences influence who is recruited into the study, then the final sample may reflect the characteristics of the seed respondents rather than the population of interest. RDS was developed to address this type of bias in chain-referral samples. One of its distinguishing features is the careful tracking of who recruits whom to the study. While tracking does not eliminate preferences for members of one’s own group, it makes it possible to examine those preferences analytically.

Accurate recruitment tracking can be accomplished using recruitment coupons. After each respondent has participated in the study, he or she receives a set number of coupons to give to his or her social contacts. If a new person chooses to participate, he or she brings the coupon to the first appointment. The coupons are marked with a unique identifier that connects the new respondent to her recruiter. Other methods for tracking recruitment could be devised.

The 2001 AI/AN HITS study was proposed to its two institutional review boards (IRBs)[*] with RDS in the protocol. Some elements of RDS were rejected, including the requirement to track recruitment from person to person. RDS was developed in the context of long-term intervention studies, where participants could be promised

[*] The Oregon Health & Science University IRB and the Portland Area Indian Health Service IRB

confidentiality but not real anonymity. Recruitment coupons in those studies were printed with identification numbers so that researchers knew who recruited whom to the study. The one-time AI/AN survey, however, was intended to be truly anonymous. The IRBs argued that even if the identification numbers on the coupons were not linked to any personal identifiers, they might give the appearance of identifiers and discourage participation or damage the credibility of the research and the institutions that supported it.

In the protocol eventually approved by the two IRBs, the coupon tracking system was modified. The coupons did not carry identification numbers, but instead were printed on colored paper. Coupon colors were changed weekly. Respondents were asked to recruit up to 5 individuals in their social network by handing out these coupons. Interviewers recorded the color of the coupon each person had received from a friend and the color that person took to potentially recruit more respondents. Whereas in standard RDS, the data show who recruited whom, this modified system pointed to a *group* of possible recruiters. The lack of individual-level detail made calculations from traditional RDS impossible. Nonetheless, this anonymous adaptation of RDS was an efficient method for finding AI/ANs with risk behaviors for HIV, a population that would otherwise be difficult to recruit into such a study.

Starting with 17 seed respondents, the AI/AN HITS survey interviewed 222 AI/ANs age 18 years and older who were residing in the Portland metropolitan area. The process included 6 waves over approximately 14 weeks. After basic eligibility was confirmed, the interviewers explained the confidentiality statement and obtained consent. A handmade Medicine Bag was given to honor each person and their willingness to participate. Then the trained interviewers administered surveys in a standard fashion, and, at the completion of the 35-40 minute survey, a quality assurance review was conducted on the survey responses. Finally, interviewers provided a customized HIV prevention session based on responses to the survey and gave respondents $25. The surveys were both anonymous and confidential; no respondent identifiers were collected.

# Background: Respondent-driven sampling

## Review of studies using respondent-driven sampling

Respondent-driven sampling (RDS) first appeared as a published method for public health studies in 1997.[8] Since then, RDS has been used in a number of studies, most of them related to illicit drug use or risky sexual behavior. Some studies have incorporated RDS as one of several recruitment methods or have modified the original version (as in the example above) to meet specific study needs. While RDS has most often been used to study illicit or stigmatized behaviors, it also lends itself to sampling non-stigmatized groups, such as musicians or other professionals.

In the original 1997 study, 277 active intravenous drug users (IDUs) were recruited for surveys related to a peer-driven intervention at two sites in Connecticut.[8] Each participant initially received three recruitment coupons to distribute to acquaintances but could return for more coupons if that quota was filled. Participants received cash as a primary incentive for participating in the intervention or survey and were also eligible for secondary incentives: $10 for each person successfully recruited, with a $5 bonus for recruiting a woman when women proved more difficult to attract to the study. Accumulating the sample took one year at each of the two sites.

That study directly influenced a second peer-driven intervention for IDUs, this time in Yaroslavl, Russia. Over twenty-four months in 1996-1998, the Russian project recruited 484 injection drug users, mostly young male ethnic Russians. Each respondent received three recruitment coupons to distribute to peers. These coupons were used to track recruitment at the individual level and to disburse secondary incentives for successfully recruiting others to the study. Participants received both primary and secondary incentives and provided peer education as well as participating in a series of interviews and voluntary testing for HIV, Hepatitis B and C, and sexually transmitted infections.[17]

RDS proved effective for studies of HIV risk in urban populations elsewhere. Recruitment was rapid for a cross-sectional study of IDUs in US-Mexico border cities. Between February and April 2005, two groups of about 200 IDUs were recruited from 15 and 9 seeds, using three recruitment coupons per person and both primary and secondary monetary incentives. Respondents provided an anonymous interview and a venous blood sample for serologic testing of HIV, HCV, HBV, and syphilis antibody.[18]

In Tirana, Albania, and St. Petersburg, Russia, researchers recruited 200 IDUs at each site within 8 weeks, starting with 15 and 13 seeds, respectively. Methods differed slightly between sites on the recommendations of local stakeholders: In Tirana, participants received cash incentives of 10 euros for participating and secondary cash incentives of 5 euros for successfully recruiting others; in St. Petersburg, stakeholders advised against the use of cash incentives and gifts of chocolate, personal hygiene products, and coffee were given as primary and secondary incentives. Participants received three coupons each for recruiting additional participants. Biometric measures were used in Tirana to minimize duplicate respondents, and Heckathorn’s free software, RDS Analysis Tool version 5.4 (RDSAT),[19] was used to generate population estimates and confidence intervals.[20]

A study of HIV seroprevalence among IDUs in New York City in 2004 used RDS to recruit 118 more participants than the 500 originally proposed, and in one quarter of the time originally allotted.[21] Participants received cash incentives of $20 for an interview and blood draw and were given three coupons each, coded with unique numbers, to recruit their friends and acquaintances. Participants were assigned unique codes and physical traits were tracked using IRIS Plus,[19] a database designed for RDS, to prevent duplicate responses. The research team found that RDS was efficient with respect to time and funding, and that it produced a diverse sample of IDUs. They suggest that recruitment may have been particularly easy in this population, which was large, familiar with research, and very willing to participate.

In two urban centers in Papua, New Guinea, RDS quickly generated large samples of female sex workers and men who have sex with men for surveys of risk behaviors. Primary incentives in the form of cash and gift bags were given for participation, and secondary cash incentives were given for successful recruitment. Recruitment was tracked using coupons and a project-specific database, which included biometric measurements to reduce duplication of study subjects. As in the study described above, RDSAT[19] was used to generate population estimates and confidence intervals; the paper presents RDSAT results contrasted with non-adjusted estimates.[22]

The success of RDS as a sampling method was more equivocal in other studies in urban areas of Eastern Europe and Russia. Three studies in Serbia, Montenegro, and Russia failed to recruit sufficient numbers of sex workers and IDUs.[13] A number of factors were identified as possible contributors to the lack of success: The cash incentives offered were less than a typical sex worker's hourly rate and may have been perceived as inadequate; the social networks of sex workers were small, consisting mainly of small clusters; and sex workers distrusted officials, which may have extended to researchers.

On the other hand, RDS has successfully been applied to rural populations, where the same problems of network density might be anticipated. Several articles provide detailed discussion of the methods used in an Ohio study of rural drug users.[23-26] The natural history study included respondents aged 18-30 who lived in Ohio, were not in any formal drug abuse treatment program within 30 days prior to entering the study, and who reported using MDMA (3,4-methelyenedioxymethamphet-amine), or ecstasy, in the previous 6 months. Sampling began with 28 seed respondents who were identified through ethnographic fieldwork and by recruitment consultants who were familiar with networks of people who had experience with MDMA. Each respondent was asked to refer others "like themselves" who had used MDMA recently and received three recruitment coupons with serial numbers to link that person to any participants he or she successfully recruited. When the recruitment process slowed, additional coupons were given to those who requested them. A primary incentive of $50 was given to each respondent for time spent completing the 2-3 hour baseline interview. Initially, $10 was given to a participant for successfully recruiting a peer to the project, but this amount was increased to $15 to speed up the recruitment process about 2 months after the start of the process.

Some studies implement RDS as a recruitment mechanism alone, ignoring one of the primary advantages of RDS over other methods – that it provides a framework for adjusting sample estimates to approximate the overall population. A review of nine studies of IDUs across Russia and Estonia compared RDS to a similar chain-referral method using "indigenous field workers" (IFS), evaluating the two methods with respect to recruitment.[27] In IFS, field workers are trained in the research protocols and identify individuals known to them to participate in the study. The field workers then ask their initial respondents to introduce them to other members of their social networks. The review of the two methods found significant differences between the samples recruited with RDS and with IFS, but neither was clearly more successful in recruiting a more marginalized or more diverse sample. The cost of RDS was slightly higher, averaging $43 per respondent compared to $37 using IFS, due mainly to higher incentives offered to RDS participants.

Two other studies state that they use the methods outlined by Heckathorn but seem to have used RDS mainly for recruitment and not for analyzing the resulting sample. A team in Nairobi, Kenya, used RDS to recruit sexually active men older than 18 for a detailed survey of knowledge and behaviors related to genital hygiene. The pace of recruitment was quite brisk: 463 men were recruited between April and October, 2002.[28] RDS was one of several recruitment methods used to interview and provide HIV testing to 392 male-to-female and 123 female-to-male transgender persons,[12, 29] along with targeted sampling and agency referrals, without a formal framework or attempt at statistical inference based on the sampling method. Participants received $40 for completing the interview, $10 for a follow-up appointment, and $5 for each of up to 5 successful referrals. Recruitment from July through December 1997, but the relative importance of respondent-driven sampling among other methods is unclear.

RDS is not limited to studies of stigmatized groups. A recent study of jazz musicians in New York and San Francisco employed RDS.[7] The study looked at affiliations such as union membership, preferred style of jazz, and primary instrument played. The sampling method was ideally suited to the population, since jazz musicians are networked by performing together or learning from each other, but no list exists from which a representative sample could be drawn and musicians represent a small fragment of the overall population. RDS also succeeded in bringing a wide range of musicians to the study, not just those who were better known or who played in the center of the city.

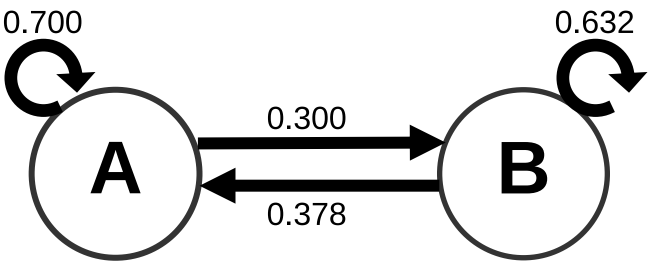
In practice, RDS is becoming established as an efficient and effective method for sampling participants from a wide variety of groups. RDS is most useful for targeting groups for whom no membership list exists, who tend to prefer to keep private, who have social connections to each other, and who represent a relatively small part of the overall population. In Indian Country, RDS could be very effective in studies of stigmatized behaviors, such as illicit drug use, but also for locating Native health professionals, artists, or traditional practitioners for strength-based wellness promotion studies.

### Respondent-driven sampling as a Markov process

The usefulness of respondent-driven sampling (RDS) hinges on the independence of the final sample composition from the characteristics of the initial seed respondents. This independence occurs if the sampling occurs as a regular Markov process, a type of process that passes from one "state" to another according to transition probabilities.[8, 30] As applied to RDS, a "state" can be understood as a characteristic, such as gender or city of residence. The sampling process moves from one state to another when one person recruits another; when a female respondent recruits a male respondent, the process moves from a female state to a male state. The transition probabilities describe recruiters' choices of recruit: The event that a person with Characteristic A recruits another person with Characteristic A occurs with one transition probability; the event that a person with Characteristic A recruits someone with Characteristic B occurs with a different transition probability, and so on.

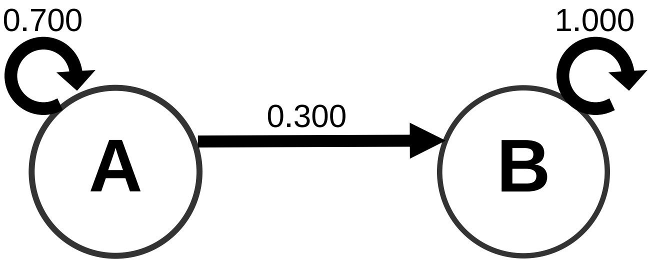
Analysis of RDS is successful only if the sampling process fits the definition of a *regular* Markov process. A regular Markov process is one that is “ergodic,” which means that any state can be reached from any other state – a person can recruit any other person in the population, regardless of their characteristics. Figure 1A illustrates a regular, ergodic Markov process with two groups, or “states,” A and B. Arrows represent transitions with probabilities given by the adjacent numbers. The probability that an A will recruit another A is 0.700. Figure 1B illustrates a non-ergodic process, where one respondent group recruits only from within itself. For group B, no arrow leads to A, representing a probability of 0, and the probability of a B recruiting a B is 1. That group is an “absorbing state,” the process is non-ergodic, and RDS theory does not apply. A regular Markov process is also non-cyclic, meaning that any kind of respondent can be recruited at any point in time.

**Figure 1A.**



**A regular Markov process.** This Markov process is *ergodic*. At any time, an A can recruit an A or B, and a B can recruit an A or B.
The numbers represent transition probabilities (the probability that person with characteristic A or B) recruits someone with characteristic A or B).

**Figure 1B.**



In this process, B is an absorbing state. The transition probability that a B will recruit another B is 1.0. The process is *non-ergodic.*

Figures adapted from Heckathorn DD. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*. May 1997;44(2):174-199.

Strictly speaking, Markov processes are linear; RDS is not. If RDS were linear, each respondent would recruit exactly one other person. In fact, respondents can recruit multiple people, or they can recruit none, which creates a branching, rather than a linear, pattern. According to Heckathorn,[16] the violation of the linear assumption does not invalidate the model. He argues that a branching referral structure can be analyzed as a set of linear structures, with each line tracing the path from the seeds through all the individuals in later waves, and that “it seems reasonable to suppose” that the analysis of a single linear chain would also be valid for a set of linear chains. He also suggests that whether the Markov model fits the data should be determined empirically, by comparing the theoretically-computed equilibrium with the actual sample distribution. Where the discrepancy is very small (e.g. ≈2%), the Markov model can be assumed to be a good fit.

Some practical steps can improve the chances of obtaining a sample that approximates the Markov model. When the sampling process extends over a larger number of waves, and when each respondent recruits a limited number of new participants, the sampling process is closer to linear in shape than a sampling process with only a few waves and large recruitment clusters. The closer to linear, the better the Markov model fits the data.

Over a “large” number of waves, a Markov process reaches an equilibrium state that is independent of its starting state.[30, 31] (This “large” number can actually be quite small; Magnani et al. estimate that six recruitment waves or less should suffice.[6]) In the equilibrium state, the composition of the sample ceases to change from wave to wave. The independence of the equilibrium composition from the starting state means that no matter who the seed respondents are, the sample will reach the same equilibrium distribution of characteristics. A sampling process could start with one man as a seed respondent and yield a final equilibrium sample that is 66% women and 34% men. In the same population, the sampling could begin with one woman and yield the same final distribution of 66% women and 34% men. Figures 2A and 2B illustrate this concept. In both scenarios, the transition probabilities are the same; only the starting points differ.

**Figure 2A. Sample composition through 10 waves, starting with one man, given the transition probabilities in 2C**

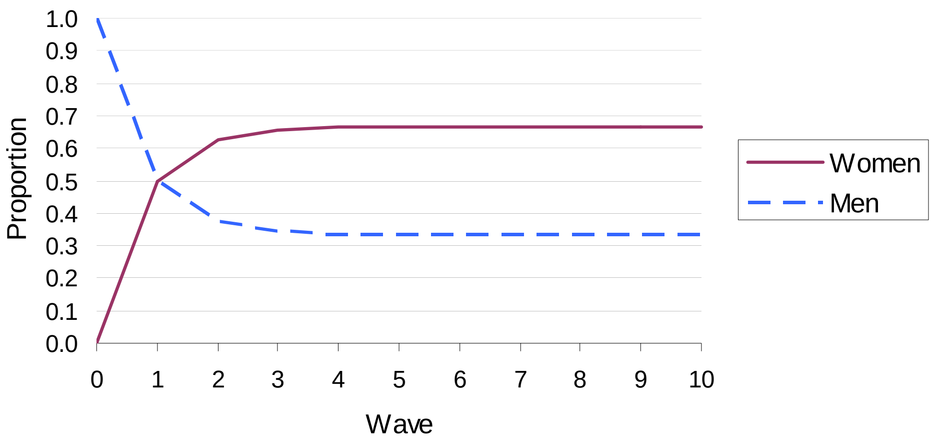


**Figure 2B. Sample composition through 10 waves, starting with one woman, given the transition probabilities in 2C**

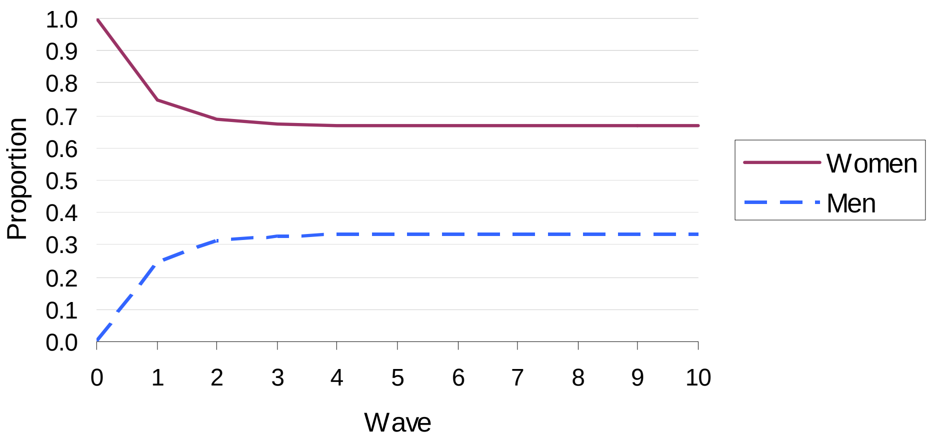


**Figure 2C. Transition probabilities for 2A and 2B**

| | **Recruits** | |
|---|---|---|
| **Recruiters** | **Women** | **Men** |
| **Women** | 0.75 | 0.25 |
| **Men** | 0.50 | 0.50 |

Figures adapted from Heckathorn DD. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*. May 1997;44(2):174-199.

A Markov process approaches equilibrium rapidly, which, as Heckathorn points out, makes RDS practical for research settings.[8] In the figures above, the sample gender distribution reached levels that were within 1% of the equilibrium distribution as early as the fourth wave. If more diverse seed respondents are chosen – 50% women and 50% men – the distribution approximates equilibrium in only three waves. (Heckathorn and colleagues have developed programs available for free download for estimating the number of waves required to reach equilibrium with varying starting assumptions.[19]) Starting with a good assortment of seed respondents speeds the progression to equilibrium.

### Calculating the equilibrium distribution

The equilibrium distribution can be calculated for traits that are objective – such as gender, age, or place of residence – and have mutually exclusive categories. Some recommend that these traits should also be "verifiable" by someone other than the respondent[8, 32]; such traits include gender and regular intravenous drug use, which is verifiable by checking for injection scars or "track marks." The most useful traits for equilibrium distribution calculations are those relevant to questions of the study's representativeness because of the close relationship between the equilibrium distribution and external validity.

The equilibrium distribution is calculated from transition probabilities. As discussed above, the sampling process "moves" from one state to another when a person recruits someone else; transition probabilities are the probabilities of those recruitment patterns. In a process with a two-level characteristic of interest, such as gender, there are four transition probabilities:

(1) The probability that a woman recruits a woman (or the probability of recruiting a woman, given that the recruiter is a woman) → P(W | W)

(2) The probability that a woman recruits a man → P(M | W)

(3) The probability that a man recruits a woman → P(W | M)

(4) The probability that a man recruits a man → P(M | M)

In Figure 3 below, gender is replaced by an unspecified characteristic with two levels, A and B. This hypothetical example started with an unspecified number of seed respondents, Wave 0, which is not shown in the figure. The next three waves, Waves 1-3, are shown on the left. Three people were recruited in Wave 1, two of them by As and one by a B. In Wave 2, a total of nine people were recruited, and in Wave 3, twenty-seven people joined the study. The three waves yield a total of 39 recruits, combined in the summary table. The cell counts in the summary table are obtained by summing the same cells over all three waves, i.e. cell A-A in the summary table is the sum of cell A-A in Wave 1, cell A-A in Wave 2, and cell A-A in Wave 3.

**Figure 3. Recruiter-recruit pairs through three waves and calculation of transition probabilities for a hypothetical RDS scenario**

| **Wave 1** | | |
|---|---|---|
| | **Recruits** | |
| **Recruiters** | **A** | **B** |
| **A** | 1 | 1 |
| **B** | 0 | 1 |

| **Wave 2** | | |
|---|---|---|
| | **Recruits** | |
| **Recruiters** | **A** | **B** |
| **A** | 2 | 1 |
| **B** | 3 | 3 |

| **Wave 3** | | |
|---|---|---|
| | **Recruits** | |
| **Recruiters** | **A** | **B** |
| **A** | 11 | 4 |
| **B** | 4 | 8 |

| **Summary Table, Waves 1-3** | | | |
|---|---|---|---|
| | **Recruits** | | |
| **Recruiters** | **A** | **B** | **Totals** |
| **A** | 1+2+11 = **14** | 1+1+4 = **6** | **20** |
| **B** | 0+3+4 = **7** | 1+3+8 = **12** | **19** |

| **Transition Probabilities, Waves 1-3** | | | |
|---|---|---|---|
| | **Recruits** | | |
| **Recruiters** | **A** | **B** | **Totals** |
| **A** | 14/20 = **0.70** | 6/20 = **0.30** | 1.00 |
| **B** | 7/19 = **0.37** | 12/19 = **0.63** | 1.00 |

The transition probabilities are the row proportions in the summary table. The row total for recruiters with characteristic A is 20; this is the denominator for the first two transition probabilities, P(A|A) = 14/20 = 0.70 and P(B|A) = 6/20 = 0.30. The remaining two transition probabilities are calculated the same way. Thus the transition probabilities are averaged over the entire recruitment process and are not specific to any one point in the recruitment process.

Once the transition probabilities are known, the equilibrium proportions can be calculated either by (1) applying them iteratively to a series of matrices or by (2) solving multiple equations simultaneously. The number of simultaneous equations is one more than the number of categories measured for the characteristic of interest. For the two-level gender variable, for example, three simultaneous equations are required. These equations are given below, both in general terms and using the data from the example given in Figure 3.

**Table 1. Transition probabilities, sample proportions, and equilibrium proportions in a hypothetical respondent-driven sampling scenario**

| **Notation** | | | **Example values** | | |
|---|---|---|---|---|---|
| Transition probabilities | | | | | |
| | **Recruits** | | | **Recruits** | |
| **Recruiters** | **A** | **B** | **Recruiters** | **A** | **B** |
| **A** | $S_{AA}$ | $S_{AB}$ | **A** | 0.70 | 0.30 |
| **B** | $S_{BA}$ | $S_{BB}$ | **B** | 0.37 | 0.63 |
| Actual sample proportions of recruits after recruitment is complete | | | | | |
| **Overall sample proportion** | $n_A / n_{TOTAL}$ | $n_B / n_{TOTAL}$ | **Overall sample proportion** | 0.538 | 0.462 |
| Equilibrium proportions calculated by solving simultaneous equations | | | | | |
| $1 = E_A + E_B$<br>$E_A = (s_{AA})(E_A) + (s_{BA})(E_B)$<br>$E_B = (s_{AB})(E_A) + (s_{BB})(E_B)$ | | | $1 = E_A + E_B$<br>$E_A = (0.70)(E_A) + (0.37)(E_B)$<br>$E_B = (0.30)(E_A) + (0.63)(E_B)$<br>$E_A = 0.552$<br>$E_B = 0.448$ | | |
| **Equilibrium proportion** | $E_A$ | $E_B$ | **Equilibrium proportion** | 0.552 | 0.448 |

Note that the transition probabilities are the only values used in solving for the equilibrium proportions. This has important implications for the bias (or lack thereof) in RDS. The probability that someone with a given characteristic will be recruited is affected both by the preferences of the recruiter and by the prevalence of the characteristic in the population, but the relative importance of preference and prevalence are not measured. (See the "Representativeness of the equilibrium proportion" section below for further discussion.) If each respondent's network size is known (that is, if the survey instrument includes questions about network size), then the equilibrium proportions can be weighted to yield unbiased estimates of population proportions.[16, 32]

In summary, RDS is a chain-referral sampling method useful for recruiting hidden populations for research. Sampling begins with an initial group of respondents who recruit others from among their social contacts, and those respondents in turn recruit additional participants. RDS fits the definition of a regular Markov process, which means that the characteristics of respondents in each recruitment wave depend only on the wave before and not on the initial respondents. As the sampling process progresses through a number of recruitment waves, the composition of the sample reaches equilibrium and remains the same from wave to wave. Typically, RDS requires tracking recruitment from individual to individual and limiting the number of people any one respondent can recruit, usually by using recruitment coupons. RDS produces a reliable sample regardless of the sampling starting point and can be implemented in a way that preserves the privacy of respondents and their social contacts to a greater extent than other chain-referral methods.

## Objectives

As described above, respondent-driven sampling is an attractive method for tapping into hard-to-reach populations. However, there are some scenarios in which keeping track of recruiters is challenging. It is possible to adapt respondent-driven sampling so that similar methods can be used in contexts such as anonymous surveys where, as in the motivating example presented above, the need to preserve the anonymity of participants takes precedence over the requirement to track recruiters.

The objectives for this paper include the following:

(1) Describe an adapted respondent-driven sampling technique that can be used when respondent anonymity is important (anonymous RDS);

(2) Develop a method for estimating the equilibrium distribution of two-level traits in anonymous RDS;

(3) Examine the properties of the equilibrium estimations produced;

(4) Suggest methods for producing estimates for traits with three or more levels; and

(5) Detail the method and theory for those who might use respondent-driven sampling to conduct community research.

## Methods: Anonymous respondent-driven sampling

As described in the section above, regular RDS uses information about the characteristics of participants and the characteristics of the individuals who recruited them to the study. Recruitment typically is tracked using recruitment coupons printed with serial numbers that link the recruit to the recruiter. The tracking system is illustrated below in Figure 4A, where every individual in the study has an identifiable recruiter. The paired recruiters and recruits give rise to a transition probability matrix that can then be used to calculate the equilibrium proportions, which gives the distribution of characteristics expected in the sample once the sampling process has continued long enough to become independent of its starting point.

**Figure 4A. Regular RDS**

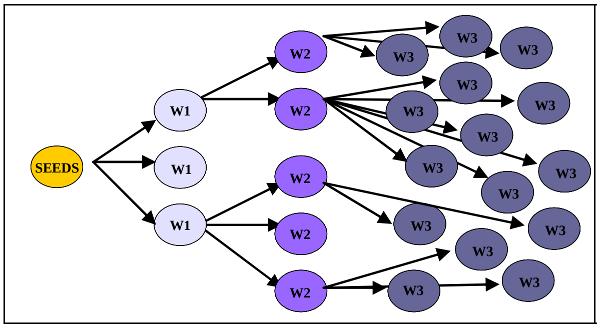
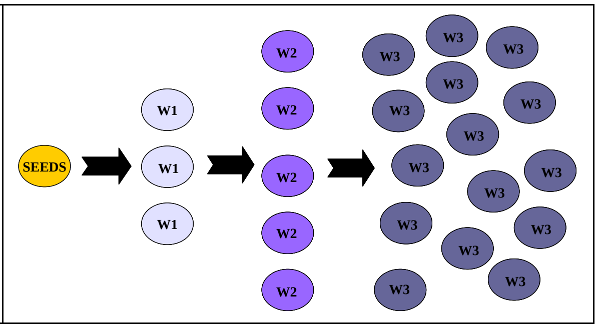


**Figure 4B. Anonymous RDS**



Anonymous RDS, in contrast, lacks the information about who recruited whom that is necessary for the equilibrium distribution calculations. Recruitment coupons would not be printed with serial numbers but would instead be coded in some way, such as different coupon colors, to distinguish recruitment waves. In this context, a recruitment wave is a group of people recruited in the same number of steps from the initial seed respondents. For example, seed respondents might receive gold-colored coupons to give to their peers to recruit them into the study. When those recruits arrive for interviews, they would bring the gold-colored coupons with them. They become Wave 1 and receive coupons in the next color, such as light blue. A recruit who brings a light blue coupon to the interview is identified as Wave 2 and receives the next color, and so on. Because some people may respond more quickly than others, several recruitment waves might occur simultaneously. Recruitment waves are defined by sequence rather than by time frame.*

Tracking recruitment waves provides limited information about recruitment patterns. As illustrated in Figure 4B, it is unclear who recruited any given individual. On the other hand, a participant in Wave 2 could only have been recruited by someone in Wave 1. Because the relationships are not specific, anonymous RDS gives rise to multiple, *potential* transition probability matrices rather than the single, *actual* transition probability matrix of regular RDS. The potential probability matrices of anonymous RDS yield multiple equilibrium proportions. Summary statistics for the set of equilibrium proportions – mean, median, range – can serve as point estimates for the directly calculated equilibrium proportions under regular RDS.

### Equilibrium distributions for 2x2 tables in anonymous respondent-driven sampling

The steps to constructing 2x2 tables for anonymous RDS begin with setting the marginal totals. The overall total n is known – this is the total number of recruits, or everyone in the dataset excluding the seed respondents. Column totals are also known – these are

* In fact, in the motivating example, researchers used time frames rather than recruitment sequences to define recruitment waves. Coupon colors were changed periodically and were not determined by the color of the coupon the recruit had received. There is some utility in tracking recruitment this way, but the sequential method offers more detail for analysis.

recruits who have either trait A or trait B. Row totals are unknown but can be estimated. Note that the row totals differ in meaning from the column totals: In the columns, every individual recruited into the study is counted once; the rows, on the other hand, reflect the instances of recruitment that brought those individuals into the study. Many of the individuals represented in the column totals as recruits may also have acted as recruiters, and if they recruited several people, they may be counted in the row totals several times. If they did not recruit anyone, they are not counted in the row totals at all. Like the column totals, the row totals sum to n.

**Figure 5. Range of values for 2x2 tables in anonymous RDS**

| | **Recruits** (all but seeds) | | |
|---|---|---|---|
| **Recruiters**<br>(all but last wave) | **A** | **B** | **Total** |
| **A** | Unknown, but all possible permutations can be generated once the row totals are set | | Unknown, but falls in known range |
| **B** | | | Unknown, but falls in known range |
| **Total** | **Known** | **Known** | **Known** |

Under anonymous RDS, row totals must be made of potential – rather than actual – recruiters. Tracking waves of recruitment narrows the possibilities. The people in the final recruitment wave could not have recruited anyone else, so they can be excluded from the row counts. The initial seed respondents must be counted among potential recruiters, even though they are excluded from the column totals. In other words,

$$\text{Potential recruiters} = n - (\text{final wave}) + (\text{seed respondents})$$

Once the potential recruiters have been identified, the next step is to determine which group each potential recruiter belongs to in order to find the relative distribution of characteristics in that subset. Additional steps are required to set row totals. The number of potential recruiters does not equal n (unless the final recruitment wave happens to contain the same number of individuals as the group of seed respondents) which means that the counts of potential recruiters must be manipulated before they can serve as row totals. A number of options are outlined below.

## Point estimates for equilibrium distributions under anonymous RDS

One possible starting assumption is that members of both groups A and B were equally likely to recruit additional study participants. If this is the case, then the row totals will be distributed between A and B in the same proportions as the potential recruiters. To set the total for row A, multiply n by the proportion of potential recruiters who have characteristic A and round to the nearest whole number. The total for row B is the remainder.

Once set, the row and column totals define a “family” of possible tables. The number of tables in each table family depends on the size of the sample and the marginal totals. Other than the requirement that the sums of cells match the row and columns, there is only one restriction. Because the tables model a regular, ergodic Markov process, each cell must contain at least one individual to yield transition probabilities greater than 0.

The simplest way to simulate the entire family of possible tables is to use statistical software. (Sample SAS code is included in the appendix.) After all possible tables are initialized, the cell values determine the transition probabilities, as described above, which in turn determine one set of equilibrium proportions per table. The mean and median equilibrium proportions taken over all tables in a family can serve as point estimates for the actual equilibrium distribution. Neither estimate is obviously better than the other, and in most simulations the mean and median are quite close.

In summary, the steps to calculate this estimate for the equilibrium distribution, where all respondents are considered equally likely to recruit others to the study, are as follows:

(1) Calculate the total number of recruits, *n*, by summing over the entire dataset and excluding the initial seed respondents.

(2) Calculate column totals by dividing *n* between those with characteristic A and those with characteristic B.

(3) Determine the numbers of potential recruiters in the whole dataset, including seed respondents but excluding those recruited in the final recruitment wave.

(4) Set row totals by multiplying *n* by the proportions of potential recruiters with characteristics A and B and rounding to the nearest whole number.

(5) Generate all possible tables for the given column and row totals with at least one person in each table cell using statistical software.

(6) Calculate transition probabilities (row proportions) for each possible table.

(7) Calculate equilibrium proportions for each possible table using the transition probabilities.

(8) Generate summary statistics (mean, median, etc.) over all possible tables.

## Example: Gender of intravenous drug users

In this section, the equilibrium distribution of an actual sample is calculated twice, first using regular RDS methods, then as if the study had been conducted anonymously. The example is taken from a recruitment diagram in a study of intravenous drug users in

Connecticut published by Heckathorn[§]; this example was chosen because it represents an actual sampling experience – rather than a hypothetical one – and because the diagram provided sufficient information to run the analysis for both regular and anonymous RDS. The diagram shows only one cluster and not the entire study sample, however. The actual study began with multiple seed respondents who produced a number of such recruitment chains and this example does not reflect the results of the study.

The recruitment chain depicted in the diagram from which these data were taken started with one male seed and proceeded through ten additional waves. Overall, 33% of the 109 total respondents in this subset were women and 67% were men. The transition probabilities, shown in Table 2b, were used to calculate the equilibrium proportions with the simultaneous equations given in Table 1 above. The overall sample distribution approximated the equilibrium distribution with only 3.5% discrepancy between the sample and equilibrium proportions.

**Table 2a. Recruitment by gender with regular RDS (n)**

| | Gender of recruit | | |
|---|---|---|---|
| Gender of recruiter | Female | Male | Total |
| Female | 11 | 11 | 22 |
| Male | 25 | 62 | 87 |
| Total | 36 | 73 | 109 |

**Table 2b. Transition probabilities with regular RDS**

| | Gender of recruit | |
|---|---|---|
| Gender of recruiter | Female | Male |
| Female | 0.500 | 0.500 |
| Male | 0.287 | 0.713 |
| Overall distribution | 0.330 | 0.670 |
| *Equilibrium distribution* | *0.365* | *0.635* |

[§] Figure 1 of Heckathorn DD. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*. May 1997;44(2);178. The final wave contained only three respondents and was excluded from the example.

The same study could have been conducted tracking only waves, with color-coded coupons to reflect the order in which people were recruited. To simulate the difference, the existing data are stripped of recruiter information, leaving a variable to show which wave the recruit belonged to. The known quantities in Table 2a are now limited to the column and overall totals. We would know how many people were recruited of each gender, but not who recruited them:

**Table 3. Known gender of recruits (all respondents except seeds), anonymous RDS**

| | Gender of recruit | | |
|---|---|---|---|
| Gender of recruiter | Female | Male | Total (n) |
| Female | ? | ? | ? |
| Male | ? | ? | ? |
| Total | 36 | 73 | **109** |

We would, however, be able to figure out how many potential recruiters of each gender were in the dataset by starting with the total n, 109, subtracting the 10 people in the final wave and adding the 1 seed respondent:

Potential recruiters = 109 – (10 people in the final wave) + (1 seed respondent) = 100

Examination of the data shows that of these potential recruiters, 34 are female and 66 are male:

**Table 4. Potential recruiters, anonymous RDS**

| | Gender of recruit | | |
|---|---|---|---|
| Gender of recruiter | Female | Male | Total (n) |
| Female | ? | ? | 34 |
| Male | ? | ? | 66 |
| Total | ? | ? | **100** |

The number of actual recruits does not match the number of potential recruiters because each potential recruiter might have recruited several new people, or none, and the recruitment waves are of varying size.

In the final table, though, row totals represent instances of recruitment rather than individuals and the number of recruits (n=109) should match the instances in which they were recruited. Some assumptions need to be made to make this match between potential recruiters and actual recruits. As a starting assumption, we can treat the male and female groups as though both groups recruited new participants at the same rate and set the row totals to match the proportion of potential recruiters of each gender. This is accomplished by multiplying the total number of recruits by the proportion of potential recruiters in each row and rounding the result to the nearest whole number. For these data, the calculations yield 37 females and 72 males:

**Table 5. Row totals set to match proportions of potential recruiters**

| | Gender of recruit | | | |
|---|---|---|---|---|
| Gender of recruiter | Female | Male | Total (n) | |
| Female | ? | ? | (34/100)*109 = | 37 |
| Male | ? | ? | (66/100)*109 = | 72 |
| Total | 36 | 73 | | **109** |

The next step is to simulate all possible tables that would give rise to those row and column totals, a task accomplished quickly using statistical software (SAS[33]; see appendix for sample code). The possible tables are subject to one limitation: Because RDS is a regular Markov process and must be ergodic, a person with characteristic A must be able to recruit others from either group A or group B, and the same must be true for group B. This means that all transition probabilities must be greater than 0, and for this to be true, each cell must contain at least one individual. In this instance, there are 35 possible combinations, given in Table 6 below. Using the cell values, we can then calculate the transition probabilities for each table and use those transition probabilities to calculate the equilibrium distribution. The equilibrium proportion of women in this “family” of tables ranges from 0.2059 to 0.3331, with an average value of 0.3178 and a median of 0.3272. Figure 6 plots the distribution of the possible equilibrium proportions (found in the final two columns of Table 6) and their medians.

**Table 6. Values in all possible tables for gender example when rows are set to the same distribution as potential recruiters, anonymous RDS**

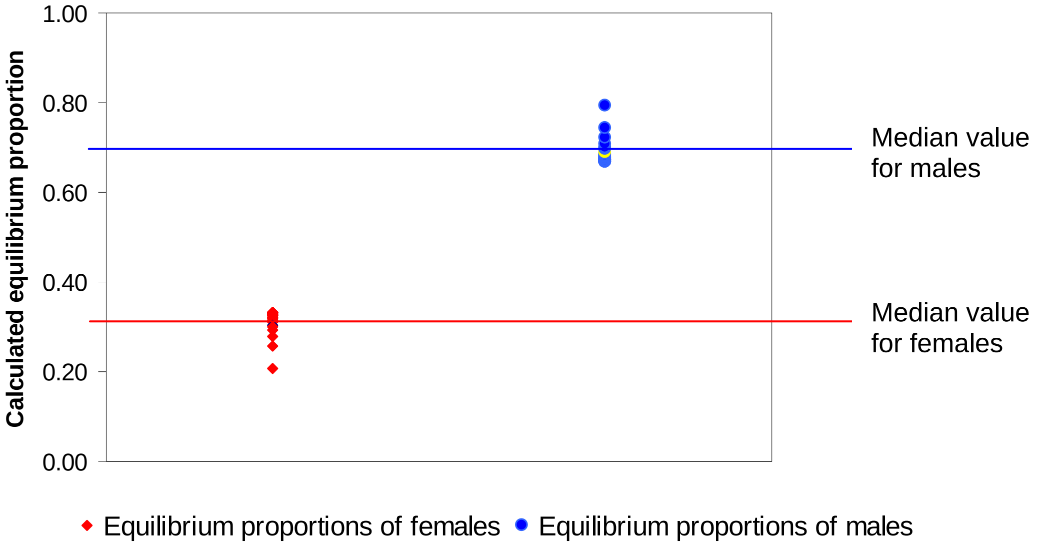
| n in cell FF | n in cell FM | n in cell MF | n in cell MM | $S_{FF}$ | $S_{FM}$ | $S_{MF}$ | $S_{MM}$ | Equilibrium proportion for females | Equilibrium proportion for males |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 36 | 35 | 37 | 0.027 | 0.973 | 0.486 | 0.514 | 0.3331 | 0.6669 |
| 2 | 35 | 34 | 38 | 0.054 | 0.946 | 0.472 | 0.528 | 0.3329 | 0.6671 |
| 3 | 34 | 33 | 39 | 0.081 | 0.919 | 0.458 | 0.542 | 0.3326 | 0.6674 |
| 4 | 33 | 32 | 40 | 0.108 | 0.892 | 0.444 | 0.556 | 0.3323 | 0.6677 |
| 5 | 32 | 31 | 41 | 0.135 | 0.865 | 0.431 | 0.569 | 0.3326 | 0.6674 |
| 6 | 31 | 30 | 42 | 0.162 | 0.838 | 0.417 | 0.583 | 0.3323 | 0.6677 |
| 7 | 30 | 29 | 43 | 0.189 | 0.811 | 0.403 | 0.597 | 0.3320 | 0.6680 |
| 8 | 29 | 28 | 44 | 0.216 | 0.784 | 0.389 | 0.611 | 0.3316 | 0.6684 |
| 9 | 28 | 27 | 45 | 0.243 | 0.757 | 0.375 | 0.625 | 0.3313 | 0.6687 |
| 10 | 27 | 26 | 46 | 0.270 | 0.730 | 0.361 | 0.639 | 0.3309 | 0.6691 |
| 11 | 26 | 25 | 47 | 0.297 | 0.703 | 0.347 | 0.653 | 0.3305 | 0.6695 |
| 12 | 25 | 24 | 48 | 0.324 | 0.676 | 0.333 | 0.667 | 0.3300 | 0.6700 |
| 13 | 24 | 23 | 49 | 0.351 | 0.649 | 0.319 | 0.681 | 0.3295 | 0.6705 |
| 14 | 23 | 22 | 50 | 0.378 | 0.622 | 0.306 | 0.694 | 0.3297 | 0.6703 |
| 15 | 22 | 21 | 51 | 0.405 | 0.595 | 0.292 | 0.708 | 0.3292 | 0.6708 |
| 16 | 21 | 20 | 52 | 0.432 | 0.568 | 0.278 | 0.722 | 0.3286 | 0.6714 |
| 17 | 20 | 19 | 53 | 0.459 | 0.541 | 0.264 | 0.736 | 0.3280 | 0.6720 |
| 18 | 19 | 18 | 54 | 0.486 | 0.514 | 0.250 | 0.750 | 0.3272 | 0.6728 |
| 19 | 18 | 17 | 55 | 0.514 | 0.486 | 0.236 | 0.764 | 0.3269 | 0.6731 |
| 20 | 17 | 16 | 56 | 0.541 | 0.459 | 0.222 | 0.778 | 0.3260 | 0.6740 |
| 21 | 16 | 15 | 57 | 0.568 | 0.432 | 0.208 | 0.792 | 0.3250 | 0.6750 |
| 22 | 15 | 14 | 58 | 0.595 | 0.405 | 0.194 | 0.806 | 0.3239 | 0.6761 |
| 23 | 14 | 13 | 59 | 0.622 | 0.378 | 0.181 | 0.819 | 0.3238 | 0.6762 |
| 24 | 13 | 12 | 60 | 0.649 | 0.351 | 0.167 | 0.833 | 0.3224 | 0.6776 |
| 25 | 12 | 11 | 61 | 0.676 | 0.324 | 0.153 | 0.847 | 0.3208 | 0.6792 |
| 26 | 11 | 10 | 62 | 0.703 | 0.297 | 0.139 | 0.861 | 0.3188 | 0.6812 |
| 27 | 10 | 9 | 63 | 0.730 | 0.270 | 0.125 | 0.875 | 0.3165 | 0.6835 |
| 28 | 9 | 8 | 64 | 0.757 | 0.243 | 0.111 | 0.889 | 0.3136 | 0.6864 |
| 29 | 8 | 7 | 65 | 0.784 | 0.216 | 0.097 | 0.903 | 0.3099 | 0.6901 |
| 30 | 7 | 6 | 66 | 0.811 | 0.189 | 0.083 | 0.917 | 0.3051 | 0.6949 |
| 31 | 6 | 5 | 67 | 0.838 | 0.162 | 0.069 | 0.931 | 0.2987 | 0.7013 |
| 32 | 5 | 4 | 68 | 0.865 | 0.135 | 0.056 | 0.944 | 0.2932 | 0.7068 |
| 33 | 4 | 3 | 69 | 0.892 | 0.108 | 0.042 | 0.958 | 0.2800 | 0.7200 |
| 34 | 3 | 2 | 70 | 0.919 | 0.081 | 0.028 | 0.972 | 0.2569 | 0.7431 |
| 35 | 2 | 1 | 71 | 0.946 | 0.054 | 0.014 | 0.986 | 0.2059 | 0.7941 |
| Mean equilibrium proportion | | | | | | | | 0.3178 | 0.6822 |
| Median equilibrium proportion | | | | | | | | 0.3272 | 0.6728 |
| Equilibrium proportion under regular RDS | | | | | | | | 0.3647 | 0.6353 |

**n in cell**: The number of respondents in each cell for the given possible table. Cell FF is the first cell, which gives the number of female recruiters of female recruits; cell FM is the number of female recruiters of male recruits.

***s*FF**: Each *s* term represents a transition probability. $s_{FF}$ is the transition probability of a female recruiter for a female recruit; $s_{FM}$ is the transition probability of a female recruiter for a male recruit (see Figure 4).

**Equilibrium proportion**: The distribution of characteristics once the sampling process has reached equilibrium, calculated by solving the simultaneous equations in Figure 4 above.

**Figure 6. All possible equilibrium proportions for tables with rows set to the same distribution as potential recruiters**



Treating this example first as a regular RDS and then as an anonymous RDS study provides an opportunity to compare the results from the different methods. Under regular RDS, the equilibrium proportions were 0.365 for women and 0.635 for men. Under anonymous RDS and the assumption that men and women recruited at the same rate, the point estimates and the range of equilibrium proportions were just outside the regular RDS proportions. The mean anonymous RDS equilibrium proportions differed from the regular RDS equilibrium proportions by about 4.7%. The median equilibrium proportions under anonymous RDS differed from regular RDS by only 3.7%.

The assumption that female and male respondents recruited new participants at the same rate yielded a relatively small discrepancy between regular RDS and anonymous RDS. In an actual anonymous RDS study there would be no way to compare the results to regular RDS and no way to judge the appropriateness of that starting assumption.

Instead, we propose looking at the entire range of possibilities by repeating the above calculations for the full range of possible row totals.

## Interval estimates for equilibrium distributions under anonymous RDS

The same process used to generate a single family of possible tables and its mean and median equilibrium proportions can be applied to all families of possible tables, yielding a range of mean and median equilibrium values. This requires extra steps, first to determine the actual range of potential row totals, then to generate a dataset containing those row totals with the column totals, which remain constant.

The first limitation on possible row totals relates to the assumptions of the RDS model. As before, the assumption that RDS is a regular Markov process and must be ergodic requires that all transition probabilities must be greater than 0, and for this to be true, each cell must contain at least one individual. For a 2x2 table, therefore, the minimum row total for any characteristic is 2. The maximum row total is n-2, where n is the total number of participants excluding the initial seed respondents.

When coupons are rationed to limit recruitment, they can place further limits on the row totals. If each participant receives only 3 recruitment coupons, then the maximum possible row total is (3)*(potential recruiters with characteristic), which may be less than n-2. Potential recruiters are counted the same as above, i.e. (potential recruiters) = n – (final wave) + (seed respondents). If the number of potential recruiters with the characteristic multiplied by the number of coupons is less than n-2, then the product is the maximum row total. To complete the table, the second row total is the remainder; together the rows should sum to n.

The two sets of row totals, one with row A set to its maximum, the other with row B set to maximum, define the range of table families possible with the anonymous RDS dataset. Using these two sets of row totals as endpoints, statistical software can generate all possible row totals in between. Column totals, being a known quantity, remain constant for all scenarios.

Once the row and column totals have been defined, the same steps as above can be applied: Generate all possible tables for each family; calculate transition probabilities; use the transition probabilities to calculate equilibrium proportions; take the mean and median of all possible tables in each family. This process results in a set of mean and median equilibrium proportions, rather than the single estimates described above. The middle 95% of these estimates can substitute for a 95% confidence interval.

## Example revisited: Gender of intravenous drug users

This section returns to the previous example of a subset of 110 intravenous drug users from a study in Connecticut. The total dataset of 110 individuals includes 1 seed respondent and 109 recruits. Of the 109 recruits, 36 were female and 73 were male. In the actual study, recruitment coupons were not rationed; if the study had been conducted as anonymous RDS, they would have been. For the purposes of the example, assume that each person received three coupons.

The row total for either group is at most n-2. In this example, n=109 recruits and n-2=107. The maximum row total for women may be lower, depending on the number of female potential recruiters and coupons distributed. From the example above, we know that there were 34 women among potential recruiters. If each woman used all three of her coupons, then a maximum of (34*3), or 102, participants would have been recruited by women. Because 102<107, 102 is the maximum row total for women. The corresponding minimum row total for men is n-102=7.

**Table 7a. Row totals set to match the maximum possible for female recruiters**

| | Gender of recruit | | | |
|---|---|---|---|---|
| Gender of recruiter | Female | Male | Total | |
| Female | ? | ? | 34*3 = | 102 |
| Male | ? | ? | 109 – 102 = | 7 |
| Total (n) | 36 | 73 | | **109** |

The maximum row total for men can be approached the same way. If each man recruited the maximum 3 people to the study, then 66*3=198 people would have been recruited by men. This is far outside the limit of 107 determined by the size of the dataset, so the actual maximum row total for men is 107 and the corresponding minimum row total for women is 2.

**Table 7b. Row totals set to match the maximum possible for male recruiters**

| | Gender of recruit | | | |
|---|---|---|---|---|
| Gender of recruiter | Female | Male | Total | |
| Female | ? | ? | | 2 |
| Male | ? | ? | 109 – 2 = | 107 |
| Total (n) | 36 | 73 | | **109** |

These two sets of row totals constitute the endpoints of a range of row totals possible for this dataset. Column totals are held constant throughout because the number of recruits in each group is known and does not need to be estimated.

**Table 8. Range of possible row totals for table “families”**

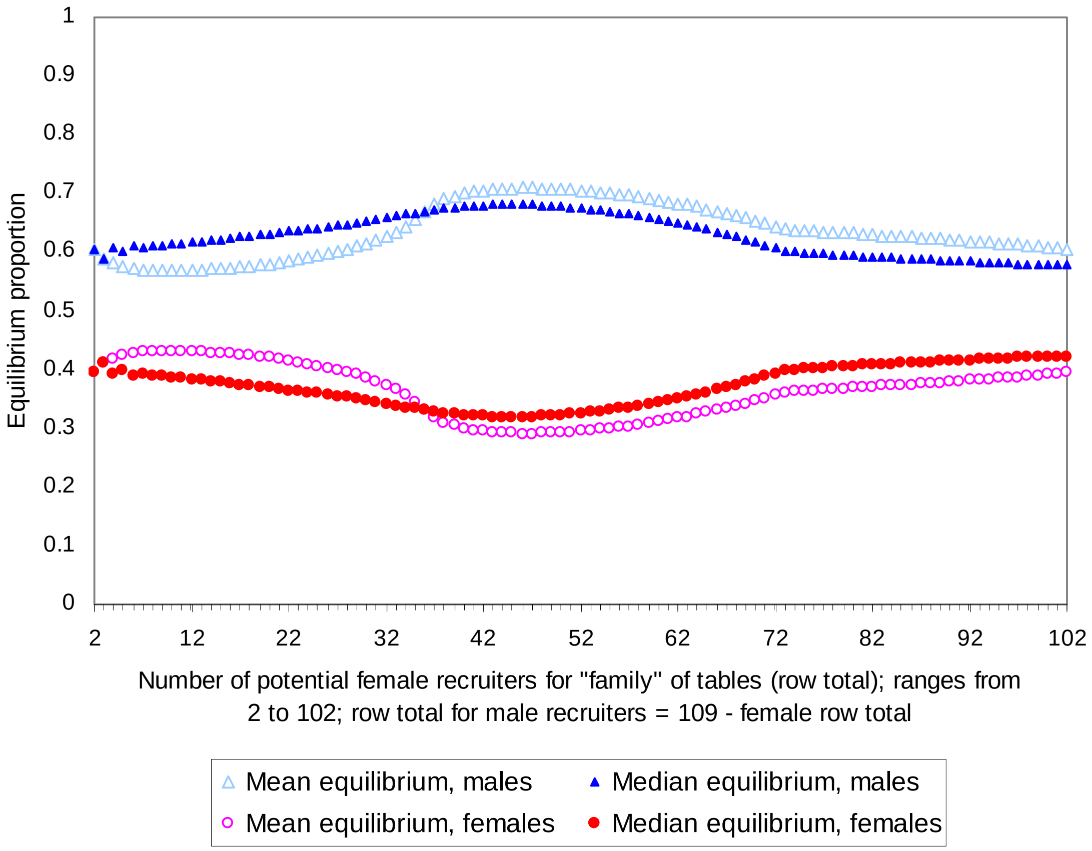
| Gender of recruiter | Gender of recruit | | Total | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Female | Male | Female maximum … | | | | | | | | | | … Male maximum |
| Female | ? | ? | 102 | 101 | 100 | 99 | 98 | 97 | 96 | … | 4 | 3 | 2 |
| Male | ? | ? | 7 | 8 | 9 | 10 | 11 | 12 | 13 | … | 105 | 106 | 107 |
| Total (n) | 36 | 73 | 109 | 109 | 109 | 109 | 109 | 109 | 109 | 109 | 109 | 109 | 109 |

Each set of row and column totals in the range produces a “family” of possible tables from which a set of equilibrium proportions can be calculated and summarized. Statistical software can generate all possible tables quickly and calculate their equilibrium distributions as described above.

The mean equilibrium distribution calculated over each family of tables is fairly stable (see Figure 7). The mean values for women range from 0.291 to 0.431; because of rounding, the middle 95% of mean equilibrium distributions falls in the same range, 0.291 to 0.431. For male respondents, the interval is 0.569 to 0.709. Those intervals correspond to a margin of error of ±7%, which is wider than the 2% discrepancy standard used by Heckathorn and others to determine whether sampling has reached equilibrium, but still gives a relatively precise estimate of the expected equilibrium distribution.

Median equilibrium estimates vary less than the means (see Figure 7). For female respondents, the middle 95% of median values range from 0.320 to 0.421, and for males, from 0.579 to 0.680. The margin of error is just over 5%. Another interesting feature of the mean and median estimates is that they are most similar at the point that corresponds to the earlier example, the family of tables where male and female recruiters were assumed to have recruited at the same rate. In Figure 7, the mean crosses the median at the family of tables where the row columns are distributed in the same proportions as the potential recruiters in the dataset.

**Figure 7. Mean and median equilibrium proportions for all "families" of possible tables, anonymous respondent-driven sampling**



When RDS is conducted anonymously, then, either the mean or median equilibrium estimates could substitute for the actual equilibrium distribution. Those estimates are averaged over all possible tables in a given family, meaning that the tables are derived from the same set of row and column totals. While column totals are known and held constant, row totals fall within a wide range of potential values and determining appropriate row totals is a difficult step in the process. Rather than choosing one set of row totals that might represent the reality of recruitment patterns, it is possible to repeat the calculations for every family. The resulting means and medians fall within a fairly narrow range which could serve as an interval estimate for the equilibrium distribution.

## Weighted equilibrium estimates

Instead of averaging equilibrium estimates over a family of tables, or taking the median, it is possible to weight each table's equilibrium proportions by the table's probability under the hypergeometric distribution. This technique borrows from Fisher's Exact Test for 2x2 contingency tables[34] which provides a definition for the probability of each table under the hypergeometric distribution (see Figure 8). Under this distribution, tables where the cells are most evenly distributed carry the highest probability. For a single set of row and column totals, the individual probabilities of possible tables sum to a total probability of 1.

**Figure 8. Probabilities of tables under the hypergeometric distribution**

|  | 1 | 2 | Total |
|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | $r_1$ |
| 2 | $n_{21}$ | $n_{22}$ | $r_2$ |
| Total | $c_1$ | $c_2$ | $n$ |

$$\text{Table probability} = \frac{r_1!\, r_2!\, c_1! c_2!}{n_{11}! n_{12}! n_{21}! n_{22}!\, n!}$$

Table probabilities under RDS require a slight adjustment to the formula used for Fisher's Exact Test. Tables with cells equal to zero are included among tables possible under Fisher's Exact Test, whereas zero cells violate the assumptions of RDS and are excluded from the set of possible tables generated for RDS. Because the table probabilities sum to 1, the Fisher's Exact Test probabilities must be rescaled for anonymous RDS. Rescaling the probabilities for RDS requires an extra step of calculating and then summing the probabilities of tables containing zero cells. The probability of each RDS table under the hypergeometric distribution is divided by that sum.

Calculating the weighted equilibrium proportions for a family of tables starts with the same steps as before – generating all possible tables and calculating their equilibrium proportions. The probability of each of those possible tables is also calculated. Each equilibrium proportion is multiplied by its table's probability. The products are summed over all tables; the sum is the weighted equilibrium estimate.

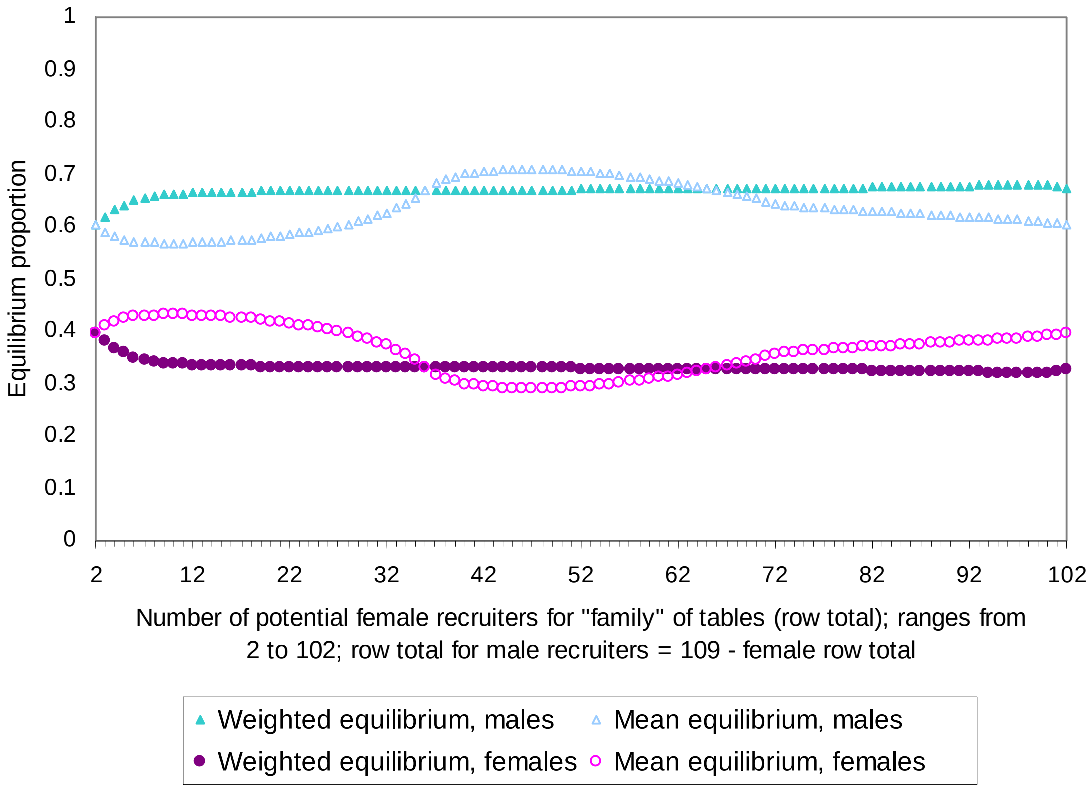
Table 9 shows each of these steps for the first anonymous RDS example, in which the row totals for the gender table were set to the same distribution as all potential recruiters (see Table 6). The equilibrium distributions in the first two columns are multiplied in turn by the middle column. The last two columns are meaningful only in that their totals are the weighted equilibrium proportions, 0.330 females and 0.670 males.

**Table 9. Individual table probabilities and weighted equilibrium calculations for the family of possible tables where rows are set to the same distribution as potential recruiters, gender example, anonymous RDS**

| Equilibrium proportions of females | Equilibrium proportions of males | Probability of this table under the hypergeometric distribution, rescaled | Partial weighted equilibrium, females | Partial weighted equillibrium, males |
|---|---|---|---|---|
| 0.3331 | 0.6669 | 1.83467 E-07 | 6.1113 E-08 | 1.2235 E-07 |
| 0.3329 | 0.6671 | 3.04168 E-06 | 1.0124 E-06 | 2.0292 E-06 |
| 0.3326 | 0.6674 | 3.09368 E-05 | 1.0289 E-05 | 2.0647 E-05 |
| 0.3323 | 0.6677 | 0.00022 | 7.2098 E-05 | 0.0001 |
| 0.3326 | 0.6674 | 0.00112 | 0.0004 | 0.0007 |
| 0.3323 | 0.6677 | 0.00440 | 0.0015 | 0.0029 |
| 0.3320 | 0.6680 | 0.01359 | 0.0045 | 0.0091 |
| 0.3316 | 0.6684 | 0.03359 | 0.0111 | 0.0225 |
| 0.3313 | 0.6687 | 0.06735 | 0.0223 | 0.0450 |
| 0.3309 | 0.6691 | 0.11069 | 0.0366 | 0.0741 |
| 0.3305 | 0.6695 | 0.15031 | 0.0497 | 0.1006 |
| 0.3300 | 0.6700 | 0.16962 | 0.0560 | 0.1136 |
| 0.3295 | 0.6705 | 0.15976 | 0.0526 | 0.1071 |
| 0.3297 | 0.6703 | 0.12598 | 0.0415 | 0.0844 |
| 0.3292 | 0.6708 | 0.08333 | 0.0274 | 0.0559 |
| 0.3286 | 0.6714 | 0.04627 | 0.0152 | 0.0311 |
| 0.3280 | 0.6720 | 0.02157 | 0.0071 | 0.0145 |
| 0.3272 | 0.6728 | 0.00843 | 0.0028 | 0.0057 |
| 0.3269 | 0.6731 | 0.00276 | 0.0009 | 0.0019 |
| 0.3260 | 0.6740 | 0.00075 | 0.0002 | 0.0005 |
| 0.3250 | 0.6750 | 0.00017 | 0.0001 | 0.0001 |
| 0.3239 | 0.6761 | 3.22267E-05 | 1.0437 E-05 | 2.1789 E-05 |
| 0.3238 | 0.6762 | 4.98718E-06 | 1.6148 E-06 | 3.3723 E-06 |
| 0.3224 | 0.6776 | 6.30324E-07 | 2.0321 E-07 | 4.2711 E-07 |
| 0.3208 | 0.6792 | 6.44791E-08 | 2.0682 E-08 | 4.3797 E-08 |
| 0.3188 | 0.6812 | 5.27992E-09 | 1.6832 E-09 | 3.5966 E-09 |
| 0.3165 | 0.6835 | 3.41441E-10 | 1.0805 E-10 | 2.3339 E-10 |
| 0.3136 | 0.6864 | 1.71483E-11 | 5.3770 E-12 | 1.1771 E-11 |
| 0.3099 | 0.6901 | 6.55000E-13 | 2.0298 E-13 | 4.5201 E-13 |
| 0.3051 | 0.6949 | 1.85253E-14 | 5.6529 E-15 | 1.2872 E-14 |
| 0.2987 | 0.7013 | 3.74608E-16 | 1.1189 E-16 | 2.6271 E-16 |
| 0.2932 | 0.7068 | 5.16463E-18 | 1.5142 E-18 | 3.6504 E-18 |
| 0.2800 | 0.7200 | 4.53635E-20 | 1.2701 E-20 | 3.2661 E-20 |
| 0.2569 | 0.7431 | 2.28724E-22 | 5.8754 E-23 | 1.6996 E-22 |
| 0.2059 | 0.7941 | 5.52250E-25 | 1.1369 E-25 | 4.3855 E-25 |
|  | **Sum** | **1.00000** | **0.330** | **0.670** |

As before, these steps can be repeated for the entire range of table families in the data, starting with the assumption of the minimum total number of female recruiters (row total = 2) and shifting one person to the female category with each new round until the maximum possible is reached (row total = 102). Figure 9 shows the full range weighted equilibrium estimates plotted with the original mean equilibrium estimates. The weighted estimates are consistent. The middle 95% of weighted estimates range from 0.320 to 0.369 for females, a range of 4.9%. If tables with fewer than 5 female recruiters are excluded, the weighted estimates cover a range only 3.8% wide.

**Figure 9. Weighted and unweighted equilibrium proportions for all "families" of possible tables, weighted by the probability of the table under the hypergeometric distribution, anonymous respondent-driven sampling**



While the stability of the weighted estimates recommends this technique, the way that probabilities are assigned may not reflect the reality of RDS. The calculations for table probabilities are borrowed from a statistical test that assumes no association between traits, which translates to higher probabilities being assigned to tables where the cell counts are more or less evenly distributed. This means that possible tables where recruiters show no preference for recruiting from one group or another are counted as more likely than tables where recruiters bring more recruits from their own group, whereas experience shows that people tend to associate with others who are like them (women may have more female contacts than males) and may recruit preferentially from among those who share their characteristics. If participants recruit preferentially from their own groups, the weighted equilibrium is weighted toward the tables that least reflect reality.

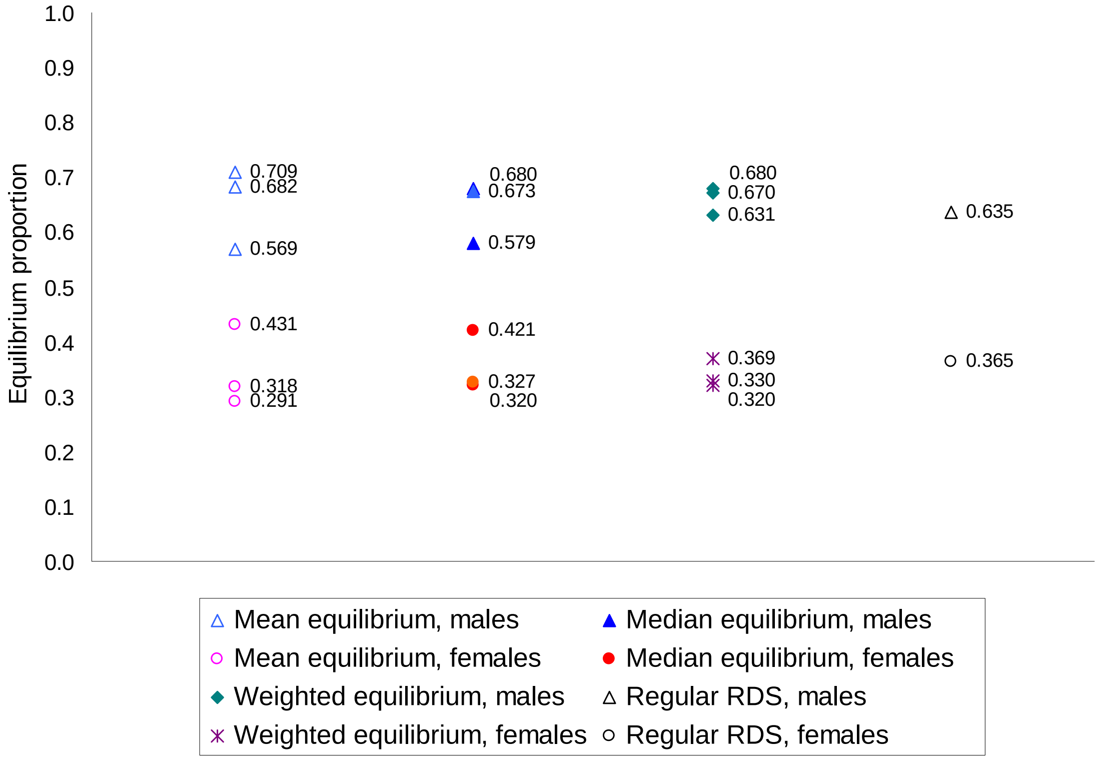
The effect of this limitation is unclear. In the case of the gender example, the probability of the actual recruitment table under regular RDS (Table 2a) is only $p = 0.035$, yet the weighted equilibrium distribution of 0.330 females and 0.670 males is reasonably close to the regular RDS equilibrium distribution of 0.365 females and 0.635 males.

## **Summary: Measures for 2x2 tables**

A variety of measures exist for estimating the equilibrium distribution for 2x2 tables when individual-level data about recruitment are lacking, including both point and interval estimates. Point estimates start with an assumption about how each group recruited new participants to the study; one such assumption is that both groups recruited at the same rate. Row totals are set based on the chosen assumption and all possible tables are generated. The mean and median of the resulting equilibrium estimates can both serve as point estimates, as well as an estimate weighted by the probability of the constituent tables under the hypergeometric distribution. If the same process is repeated for all possible sets of row totals, the middle 95% of estimates make a reasonable interval estimate.

Figure 10 provides a visual comparison of the ranges covered by the middle 95% of mean, median, and weighted equilibrium estimates, along with the point estimates under the assumption that males and females recruited equally from the previous section. The median estimates cover a slightly narrower range than the mean estimates, though the point estimates are very similar for medians and means. The median point estimates are

**Figure 10. Point estimates under the equal recruitment rate assumption and bounds of 95% intervals for mean, median, and weighted equilibrium proportions, anonymous RDS, compared to regular RDS**



very close to the outside of the 95% range for all median estimates; this occurs because most of the calculated median equilibrium proportions fall within a very narrow (±2%) range. The median and weighted estimates are almost indistinguishable. In this example, comparisons to regular RDS were possible. All ranges included the actual equilibrium distribution from regular RDS.

### Methods for 3x3 tables

The gender and other examples above have been variables with only two levels (male and female, A and B), but variables with three or more levels may also be important. It may be useful, for example, to examine recruitment by city of residence or race. Adding a third category does not substantially change the methods presented here but is more challenging conceptually and computationally. The number of potential tables is exponentially larger than the number of possible 2x2 tables, as Table 11 shows. For a hypothetical example with only 30 study subjects, one set of row and column totals in a 2x2 table might give rise to as many as 14 possible tables, but for the same sample, a 3x3 table family might contain 666 possible tables. For larger sample sizes – such as those typical in public health research – the computations quickly become cumbersome, even with a relatively fast computer. The data from the example above, when only n=105 subjects were included, generated over 87 million possible tables and required an inordinate amount of computer processing time.

**Table 10. Comparison of computations for 2x2 versus 3x3 tables, when generating all possible tables – hypothetical example**

| | **2x2 Table** | **3x3 Table** |
|---|---|---|
| Total sample size | n = 30 | n = 30 |
| Column totals | Column 1 = 15<br>Column 2 = 15 | Column 1 = 10<br>Column 2 = 10<br>Column 3 = 10 |
| Number of table "families" | 27 | 253 |
| Maximum tables in a family | 14 | 666 |
| Total tables possible for all families combined | 196 | 46,656 |

Generating a random sample of possible tables, rather than the whole set, takes less time and yields almost identical results. The steps are simple: Begin by initializing the column totals, which are known. Generate the first row of cells by multiplying the maximum possible value – the column total minus two, one for each of the remaining cells – by a random number and rounding up. Generate the next row of cells by multiplying those maximum values – the remainder minus one – by a random number and rounding up. The third row of cells is the remainder, and the rows totals are summed across the rows. Statistical software can repeat the process a large number of times (in our simulations, we used 50,000 iterations).

For 3x3 tables, calculating three unknown equilibrium proportions requires solving four simultaneous equations rather than the three required for a 2x2 table (see Figure 11). Otherwise the process remains the same.

**Figure 11. Transition probability matrix and equilibrium proportions for 3x3 tables**

| | **Recruits** | | |
|---|---|---|---|
| **Recruiters** | **A** | **B** | **C** |
| **A** | $S_{AA}$ | $S_{AB}$ | $S_{AC}$ |
| **B** | $S_{BA}$ | $S_{BB}$ | $S_{BC}$ |
| **C** | $S_{CA}$ | $S_{CB}$ | $S_{CC}$ |

$$1 = E_A + E_B + E_C$$
$$E_A = (s_{AA})(E_A) + (s_{BA})(E_B) + (s_{CA})(E_C)$$
$$E_B = (s_{AB})(E_A) + (s_{BB})(E_B) + (s_{CB})(E_C)$$
$$E_C = (s_{AC})(E_A) + (s_{BC})(E_B) + (s_{CC})(E_C)$$

Working with estimates for 3x3 tables can also be conceptually more challenging than for 2x2 tables because of the extra dimension added by the third level. When a row total is set to its maximum in a 2x2 table, for example, the remaining row total is at its minimum, by definition. When a row total in a 3x3 table is set to its maximum, the remaining two rows could take on a range of values. Is it necessary to perform the analysis twice – once with the second row total set to an absolute minimum, then with the third set to minimum? If so, what is the utility of multiple estimates for the same value? These questions remain to be answered.

## Example: Race of intravenous drug users

The diagram that produced the gender example above also provided data on the race of study participants, coded as white, black, Hispanic, other, and unknown. To get the three-level variable needed for demonstration here, the “Hispanic” and “other” categories were combined and the “unknown” observations excluded from analysis, yielding a total n = 105 (see Table 11a). The equilibrium distribution calculated under regular RDS is 0.612 white, 0.210 Hispanic, and 0.178 black (see Table 11b.). There is a rather large discrepancy between the sample distribution and equilibrium, but because this is a subset consisting of the network resulting from just one seed respondent, that result is unsurprising.

**Table 11a. Recruitment by race with regular RDS**

| | Race of recruit | | | |
|---|---|---|---|---|
| Race of recruiter | White | Hispanic | Black | Total |
| White | 41 | 3 | 7 | 51 |
| Hispanic | 8 | 29 | 6 | 43 |
| Black | 5 | 2 | 4 | 11 |
| Total | 54 | 34 | 17 | 105 |

**Table 11b. Transition probabilities with regular RDS**

| | Race of recruit | | |
|---|---|---|---|
| Race of recruiter | White | Hispanic | Black |
| White | 0.804 | 0.059 | 0.137 |
| Hispanic | 0.186 | 0.674 | 0.140 |
| Black | 0.455 | 0.182 | 0.364 |
| Overall distribution | 0.514 | 0.324 | 0.162 |
| *Equilibrium distribution* | *0.612* | *0.210* | *0.178* |

When calculating the point estimates for this sample under anonymous RDS, we would lack records of who recruited whom. Instead, we would know which recruitment wave each participant belonged to. Potential recruiters are found in all but the final recruitment wave and include the initial seed respondent(s). Row totals can be set to reflect characteristics in the same proportions as among potential recruiters by first examining the relative frequencies by race, as in Table 12 below, then adjusting those numbers to match the total sample size. The total number of potential recruiters is less than the number of actual recruits, so the row totals must be rescaled (Table 13).

**Table 12. Potential recruiters with anonymous RDS**

|  | Race of recruit |  |  |  |
|---|---|---|---|---|
| Race of recruiter | White | Hispanic | Black | Total (n) |
| White | ? | ? | ? | 48 |
| Hispanic | ? | ? | ? | 33 |
| Black | ? | ? | ? | 15 |
| Total | ? | ? | ? | **96** |

**Table 13. Row totals set to match proportions of potential recruiters**

|  | Race of recruit |  |  |  |  |
|---|---|---|---|---|---|
| Race of recruiter | White | Hispanic | Black | Total (n) |  |
| White | ? | ? | ? | (48/96)*105 = | 53 |
| Hispanic | ? | ? | ? | (33/96)*105 = | 36 |
| Black | ? | ? | ? | (15/96)*105 = | 16 |
| Total | 54 | 34 | 17 |  | **105** |

This set of proportional row totals, together with the actual column totals, produces 29,015 possible tables. Averaged over all of those tables, the mean equilibrium proportions are 0.519 for whites, 0.317 for Hispanics, and 0.163 for blacks. The median equilibrium proportions vary only slightly from the means, at 0.515 for whites, 0.322 for Hispanics, and 0.163 for blacks (see Table 14).

Even for a small dataset, the number of possible 3x3 tables can overwhelm some software packages. In this example, if the same calculations are performed on a random sample of 10,000 possible tables (see appendix for sample code) rather than the entire set of 29,015 tables, the estimates are identical to three decimal points – the random sample produces a very close approximation of all possible tables.

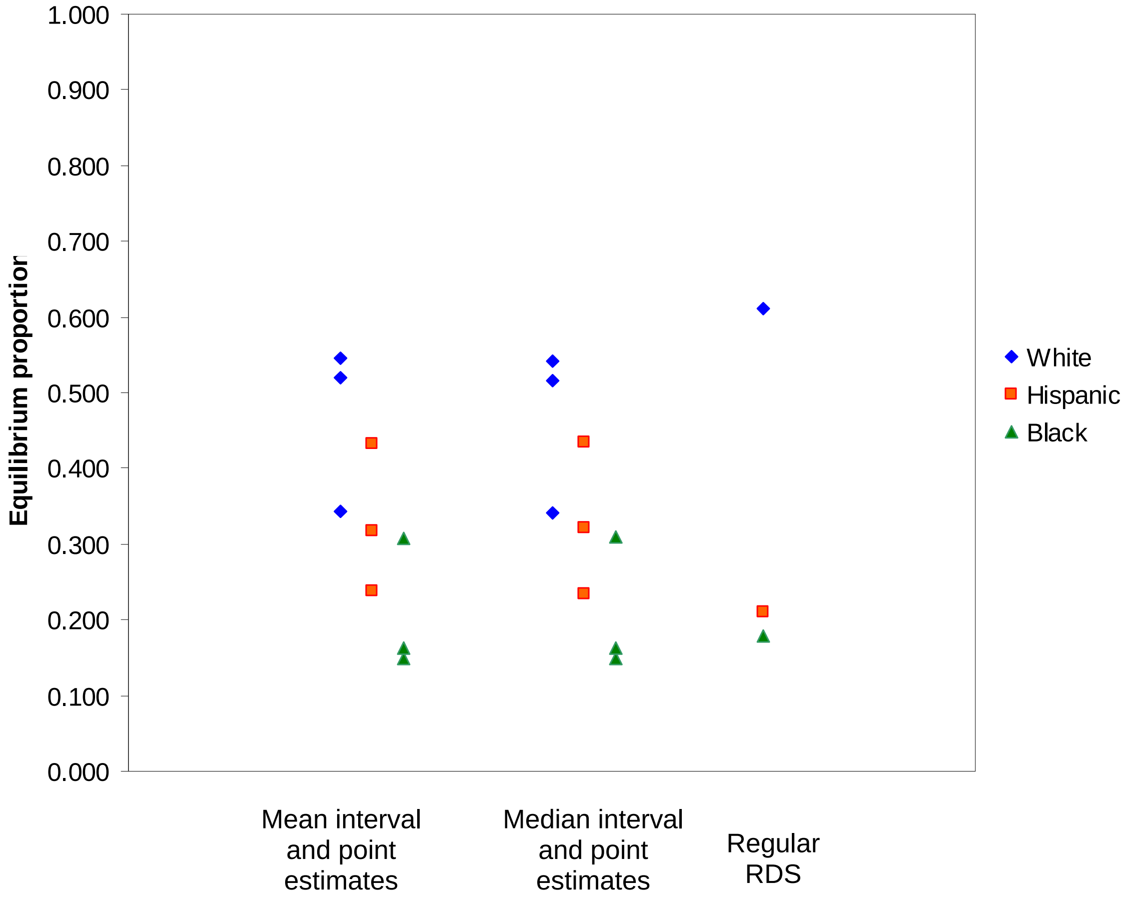
**Table 14. Point estimates for equilibrium proportions, anonymous and regular RDS**

| | White | Hispanic | Black |
|---|---|---|---|
| **Anonymous RDS, all possible cell combinations for Table 13 (n=29,015 tables)** | | | |
| Mean equilibrium | 0.519 | 0.317 | 0.163 |
| Median equilibrium | 0.515 | 0.322 | 0.163 |
| **Anonymous RDS, random sample of cell combinations for Table 13 (n=10,000 tables)** | | | |
| Mean equilibrium | 0.519 | 0.317 | 0.163 |
| Median equilibrium | 0.515 | 0.322 | 0.163 |
| **Regular RDS, Table 11b (n=1 table)** | | | |
| Equilibrium | 0.612 | 0.210 | 0.178 |
| Overall sample distribution | 0.514 | 0.324 | 0.162 |

Discrepancies between the anonymous RDS and regular RDS point estimates vary. The difference between the estimates for white and Hispanic participants is about 10% in both groups, but only 1.5% for black participants. The anonymous RDS estimates closely approximate the overall sample distribution, however.

Intervals for 3x3 tables are created on the same general principles as for 2x2 tables, except that a random sample of possible tables substitutes for the entire range of over 87 million possibilities. Once the random sample of tables has been generated and grouped into table families with the same row and column totals, the equilibrium distributions are calculated and the means and medians taken for each family. For a 95% interval, the table families between the 0.025 and 0.975 percentiles serve as endpoints. In Figure 12 below, the 95% intervals for mean and median equilibrium distributions in this example are plotted next to the point estimates calculated above and the regular RDS equilibrium proportions. Unlike the 2x2 table examples, which had non-overlapping intervals and consistent point estimates within those bounds, here the intervals overlap and the regular RDS estimates lie outside the intervals for two of three categories.

**Figure 12. Upper and lower bounds of 95% intervals (n=50,000 tables), point estimates under equal recruitment assumption (n=10,000 tables), and regular RDS equilibrium estimates (n=1 table)**



While conceptually similar to 2x2 tables, 3x3 tables in anonymous RDS pose greater challenges for calculation and interpretation of equilibrium estimates. These calculations seem less likely to yield good approximations of regular RDS than those for 2x2 tables. However, these generalizations are based on a small number of experimental samples. Further simulation and exploration are required to understand the important implications of adding a third response category.

### Representativeness of the equilibrium proportion

The equilibrium distribution is a reliable estimate: If several samples were taken from the same population, starting with different sets of seed respondents, the samples would have approximately the same equilibrium distributions. The equilibrium distribution does not necessarily reflect population prevalence, however, because of the role that personal preference plays in a recruiter's choice of recruit. The only numbers used to calculate the equilibrium proportions are the transition probabilities (see Table 2 and Figure 11), which reflect the preferences of the participants in addition to the prevalence of traits in the population of interest.[8, 32]

If people chose new recruits randomly from among their peers, the result would mimic a traditional probability sample. The prevalence of a trait in the sample would reflect the prevalence of the trait in the reference population. Most likely, though, people do not choose new recruits randomly. People often have more ties to those who share certain traits—women often have more female than male friends—which manifests as a preference in recruitment, or "homophily."[32] Occasionally the opposite may occur and members of a population may "mask" or protect a subgroup by not recruiting them, or members of a subgroup may be more cautious and therefore more difficult to recruit to the study.

With a traditional sampling frame, it is possible to evaluate bias in a sample by comparing the demographics of the sample to known demographic measures in the reference population. If a sample of state residents had a very different age distribution than the census results for that state, the discrepancy would call into question the representativeness of the sample. With an RDS sample, however, known demographic measures most likely do not exist because the method is used for hidden populations. This requires one to (a) trust (or not trust) that the RDS equilibrium distribution is representative of the population as a whole, or (b) take extra steps to avoid or evaluate sampling bias.

There are reasons to trust the representativeness of the equilibrium sample. Each transition probability can be conceptualized as two components. First, the probability that a person with a given characteristic will be recruited depends on how prevalent that characteristic is in the population. Second, the transition probability is also influenced by the homophily (or other preference) of the recruiter. In other words, each transition probability is the sum of the random probability of selection (which reflects the prevalence of a trait) together with homophily. Heckathorn demonstrates that when homophily is equally strong in all subgroups, the transition probabilities approximate random probabilities.[8]

Additional steps may also be taken to compensate for or eliminate the bias resulting from participant preferences. One extension of RDS methods requires asking participants how many people they know and using the “network size” as weights to produce unbiased estimates.[15, 32] Even without plans to perform this type of analysis, it seems prudent to include a similar question for any trait where the knowing the population prevalence is important or useful as a check on the quality of survey data. If it was important to know how many intravenous drug users were female, for example, it would be worthwhile to ask each participant how many women they knew who injected drugs. If few women showed up in the sample but most respondents reported knowing a large number of women using intravenous drugs, it would suggest the need for targeted incentives to recruit women to the study.

Alternatively, a hybrid RDS method has been developed to minimize the problem of homophily. Instead of simply asking participants to recruit others from among their social contacts, researchers ask participants to list all of their social contacts – using only first names or pseudonyms – and then sample randomly from the list, asking participants to recruit those individuals to the study.[35, 36] This reduces biased selection from within respondents' social networks, but it does not eliminate the homophily bias entirely, as the social networks themselves may be biased. Asking respondents to list their peers may also lead to "masking," where respondents would not mention certain friends and acquaintances, even by pseudonym. On the other hand, this method has the benefit of producing a good estimate of network size that could be used to weight estimates.

Knowing the equilibrium distribution serves several useful functions: (1) When the equilibrium distribution and the sample distribution are very similar, the sampling process has gone through sufficient waves to reach an equilibrium state; (2) when the equilibrium distribution and the sample distribution are different, it could also indicate that the sampling process had run into a closed group (absorbing state), which a close examination of the data would reveal; (3) if both groups share equal in-group preferences, then the equilibrium proportions are unbiased population estimates; and (4) if the two groups have unequal in-group preferences but each respondent's network size is known, the equilibrium distribution can be weighted by network size to produce unbiased population estimates.[15]

In summary, the equilibrium distribution can be an unbiased estimate of population prevalence, but it is only unbiased when all subgroups are equally strongly inclined to recruit from within their own groups. Asking respondents how many people they know provides an additional variable that can be used to weight the equilibrium sample to better reflect population characteristics.

## Implementing respondent-driven sampling: Some questions to ask when planning a study

### Is respondent-driven sampling the best method in this situation?

Respondent-driven sampling (RDS) can be a very efficient and effective method for recruiting “hidden” populations, groups with members that are hard to locate by traditional methods or are defined by illicit or stigmatized activities. But even though methods have been developed to allow statistical inference under RDS, they require additional effort and may be more vulnerable to bias than traditional methods, depending on how the study is implemented. In studies of AI/AN, sampling frames are often available for random sampling, such as tribal enrollment lists or clinic user populations. If the population of interest can be recruited from venues, such as clinics, powwows, or needle exchanges, then systematic or random venue-based sampling may be easier to implement than RDS.

If neither a good sampling frame nor a venue can be identified – as was the case in the motivating example in the introduction to this paper – RDS may be the best choice, but only if the members of the population of interest have social connections to each other. RDS was effective for the HIV testing survey because people who engage in high-risk sexual behaviors or drug use cannot do so without partners or suppliers. Social behaviors, most risk factors for infectious diseases, immigration status, and homelessness all create connections between individuals. On the other hand, RDS would likely fail in a study of people suffering from depression because people who are depressed often do not know of anyone else with the same condition.

### How much effort should we devote to formative research?

RDS has been promoted as requiring less intensive formative research than other non-random sampling methods.[37] Formative research can potentially predispose the research team to look for (and find) certain results that do not truly reflect reality. This is a particular danger if the formative research misses an important element of the research question or fails to uncover important stakeholders.

On the other hand, RDS is designed for hard-to-reach and hidden populations about whom little is known. Formative research can provide good answers to most design questions, such as the use of incentives and tracking systems. In fact, one research team credits the success of their research project to extensive formative research.[24]

### How do we want to track recruitment?

Tracking recruitment and network size at the individual level can yield great population estimates, but this level of detail was not appropriate for the 2001 anonymous survey. Benefits include knowing the distribution of important characteristics, such as gender, minor vs. adult, town of residence, or certain risk behaviors.

Tracking this information requires time and effort and funding to design and maintain an effective tracking system (though at least one tracking database is available for free download[19]). The tracking system might also look suspicious to potential participants, who may decide not to take the risk of participation. Particularly for populations who do not want to be found, any system that looks like tracking might discourage participation and bias results toward individuals whose risk perception is lower.

Alternatives to the recruitment coupons with serial numbers include color-coded coupons corresponding to recruitment waves, or coupons printed with wave numbers rather than serial numbers. Another option is to decide a priori which characteristics are most important to analyze with respect to external validity and to ask each respondent about the person who referred them to the study, e.g. “Was this person a woman or a man?” Such questions would provide all necessary data for constructing the transition probability matrix.

### What kinds of incentives should we offer?

In the 2001 AI/AN HITS study, only primary incentives for participation were given. Participants received medicine bags and $25 each. Social incentives – the possibility of helping one’s friends and community – were also in effect.

Other RDS studies have also used secondary incentives to encourage recruitment. Secondary incentives are given to the recruiter who successfully encourages someone else to participate. For example, a person might receive $25 for answering a survey, then an additional $15 for each of two friends referred who also took the survey. Secondary incentives require a tracking system, outside of study results, and are not possible in a truly anonymous study.

### How should we handle “volunteers”?

Volunteers – people who show up to participate without having been formally sampled – are common in Indian Country, where inclusiveness tends to be valued. Ideally, if volunteers are interviewed or otherwise included, their records should be marked so that their data can be treated appropriated. These respondents may be excluded from equilibrium estimates. It may also be possible, in analysis, to treat volunteer respondents as seeds who start new recruitment chains.

Volunteers could be minimized by printing recruitment coupons on bold or textured papers that are hard for well-meaning individuals to photocopy. Potential participants might be required to have their actual coupon in hand in order to be eligible for participation.

**What subgroups or characteristics are most important to know about?**

Deciding in advance on a few subgroups or characteristics that are important to study has practical advantages. For one, the recruitment and incentive schemes can be adjusted to encourage recruitment of certain types of respondents if the preliminary sample lacks members of that group, as in the IDU study where “steering” incentives were given for recruiting younger respondents.[15]

It may also be possible to use an anonymous RDS scheme but ask about certain characteristics of the recruiter who brought the respondent into the study. While regular RDS requires serial numbers for tracking recruitment from person to person, those serial numbers are used to identify recruits and recruiters as members of subgroups. If those subgroups can be identified without the serial numbers, the actual transition probabilities can be calculated and the all-possible-tables methods are unnecessary. One might, for example, ask for the respondent’s gender and the gender of the person who gave the respondent the recruitment coupon. These two variables, collected for every participant, would generate the same kind of transition probability matrix as would be generated in regular RDS.

### Should we ask about personal network size?

If your instrument includes at least one question about network size—the number of people the respondent knows within the target population—you can use your equilibrium calculations to generate population estimates, which will give you a much better idea of what is going on in the community. A free software package, the Respondent Driven Sampling Analysis Tool (RDSAT)[19], has been developed especially for generating RDS estimates of important population parameters such as population proportions, average personal network sizes, homophily, and measures of significance for population estimates.

At minimum, RDSAT requires that the data contain the following three data for each individual: (1) personal network size; (2) the respondent's serial number from the couon s/he was recruited with; and (3) the serial numbers from the coupons the respondent is given to recruit others.

Without the serial number data, it may be possible to use the methods applicable to anonymous RDS to make the same population inferences as under regular RDS, but with less precision in the estimates.

## Conclusion

Respondent-driven sampling (RDS) can be an efficient and cost-effective recruitment method applicable in a wide variety of populations. RDS has been used successfully in both urban and rural settings, with intravenous drug users, sex workers, and jazz musicians. The modifications that were made to preserve respondent anonymity in the 2001 HIV Testing Survey described in the introduction did not substantially change the recruitment process. The fact that this type of sampling functions as a Markov process means that the characteristics of recruits depend only on the characteristics of those who recruited them, and not on the initial “seed” respondents, so that the sample quickly becomes independent of the seeds and approaches a stable equilibrium.

Methods for analyzing an anonymous RDS sample, however, differ from regular RDS. While analysis of regular RDS involves generating a single table of transition probabilities and a single set of equilibrium proportions, anonymous RDS substitutes a range of possible tables and summaries of equilibrium proportions. Anonymous RDS estimates thus are less precise than regular RDS estimates.

Anonymous RDS estimates are also most successful when the sample is closest to equilibrium. This is a weakness in the method that deserves further study. In regular RDS, the equilibrium proportion estimates are less influenced by the sample composition and can be calculated at an intermediate point in the sampling process to determine whether the process has reached equilibrium.

In spite of these weaknesses, anonymous RDS does provide information about the external validity of study results, in contrast to most non-random methods available. The alternatives considered in the motivating example (see “Introduction”) would have produced convenience samples or systematic samples with significant biases. The modifications that led to anonymous RDS were made in deference to the values of the community being surveyed as well as the concerns of two institutional review boards. The resulting sample provided more information about the population than would have otherwise been possible.

## Wave Tables

The two tables in this section show the anticipated progression of recruitment under a variety of assumptions. The first, Table 15, shows the expected composition of each recruitment wave and the eventual equilibrium distributions for two-level variables. The proportions approach equilibrium quite rapidly; almost all of the scenarios given in the table arrive within two percent of the equilibrium distribution in fewer than 10 waves. Generally, more waves are required when the in-group recruiting preferences of both groups are very high (e.g. the transition probability of A recruiting A is 0.9 and of B recruiting B is 0.9).

Table 16 gives the anticipated cumulative sample size for each recruitment wave, given varying numbers of seed respondents and recruitment rates, given as the average number of people recruited by each participant.

Taken together, the two tables should give some idea of how many waves would be required to recruit for a study. Recruitment waves do not correspond to any specific time frames; in fact, recruitment waves overlap in time. In the study used as the motivating example for this paper, 6 recruitment waves were completed in 14 weeks and each person recruited, on average, 1.2 new participants to the study. Please see the review section for examples of other studies: Any such details published in peer-reviewed literature have been included there.

**Table 15. Expected composition of recruitment waves and equilibrium distributions for two-level variables, given the composition of seeds and overall transition probabilities**

Transition probabilities represent each group's tendency toward in-group versus out-group recruitment. Cells AA and BB represent in-group recruitment, and cells AB and BA represent out-group recruitment. Wave compositions in **bold** font are within ±2.0% of the equilibrium distribution.

| Seeds | | Transition probabilities | | Expected composition of recruitment wave … | | | | | | | | | | Equilibrium |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (%) | | A | B | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | distribution |
| 0.50 | A | 0.1 | 0.9 | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | 0.100 |
| 0.50 | B | 0.1 | 0.9 | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | 0.900 |
| 0.50 | A | 0.1 | 0.9 | 0.150 | **0.185** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | 0.182 |
| 0.50 | B | 0.2 | 0.8 | 0.850 | **0.815** | **0.819** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | 0.818 |
| 0.50 | A | 0.1 | 0.9 | 0.250 | **0.325** | **0.303** | **0.309** | **0.307** | **0.308** | **0.308** | **0.308** | **0.308** | **0.308** | 0.308 |
| 0.50 | B | 0.4 | 0.6 | 0.750 | **0.675** | **0.698** | **0.691** | **0.693** | **0.692** | **0.692** | **0.692** | **0.692** | **0.692** | 0.692 |
| 0.50 | A | 0.1 | 0.9 | 0.350 | 0.425 | **0.388** | **0.406** | **0.397** | **0.402** | **0.399** | **0.400** | **0.400** | **0.400** | 0.400 |
| 0.50 | B | 0.6 | 0.4 | 0.650 | 0.575 | **0.613** | **0.594** | **0.603** | **0.598** | **0.601** | **0.600** | **0.600** | **0.600** | 0.600 |
| 0.50 | A | 0.1 | 0.9 | 0.450 | **0.485** | **0.461** | **0.478** | **0.466** | **0.474** | **0.468** | **0.472** | **0.469** | **0.471** | 0.471 |
| 0.50 | B | 0.8 | 0.2 | 0.550 | **0.515** | **0.540** | **0.522** | **0.534** | **0.526** | **0.532** | **0.528** | **0.531** | **0.529** | 0.529 |
| 0.50 | A | 0.1 | 0.9 | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.50 | B | 0.9 | 0.1 | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.50 | A | 0.2 | 0.8 | 0.150 | **0.115** | **0.112** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | 0.111 |
| 0.50 | B | 0.1 | 0.9 | 0.850 | **0.885** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | 0.889 |
| 0.50 | A | 0.2 | 0.8 | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | 0.200 |
| 0.50 | B | 0.2 | 0.8 | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | 0.800 |
| 0.50 | A | 0.2 | 0.8 | 0.300 | **0.340** | **0.332** | **0.334** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | 0.333 |
| 0.50 | B | 0.4 | 0.6 | 0.700 | **0.660** | **0.668** | **0.666** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | 0.667 |
| 0.50 | A | 0.2 | 0.8 | 0.400 | **0.440** | **0.424** | **0.430** | **0.428** | **0.429** | **0.428** | **0.429** | **0.429** | **0.429** | 0.429 |
| 0.50 | B | 0.6 | 0.4 | 0.600 | **0.560** | **0.576** | **0.570** | **0.572** | **0.571** | **0.572** | **0.571** | **0.571** | **0.571** | 0.571 |
| 0.50 | A | 0.2 | 0.8 | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.50 | B | 0.8 | 0.2 | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.50 | A | 0.2 | 0.8 | 0.550 | **0.515** | **0.540** | **0.522** | **0.534** | **0.526** | **0.532** | **0.528** | **0.531** | **0.529** | 0.529 |
| 0.50 | B | 0.9 | 0.1 | 0.450 | **0.485** | **0.461** | **0.478** | **0.466** | **0.474** | **0.468** | **0.472** | **0.469** | **0.471** | 0.471 |

**Table 15. Expected composition of recruitment waves and equilibrium distributions for two-level variables, given the composition of seeds and overall transition probabilities**

Transition probabilities represent each group's tendency toward in-group versus out-group recruitment. Cells AA and BB represent in-group recruitment, and cells AB and BA represent out-group recruitment. Wave compositions in **bold** font are within ±2.0% of the equilibrium distribution.

| Seeds (%) | | Transition probabilities | | Expected composition of recruitment wave … | | | | | | | | | | Equilibrium distribution |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **A** | **B** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | |
| 0.50 | **A** | 0.4 | 0.6 | 0.250 | 0.175 | **0.153** | **0.146** | **0.144** | **0.143** | **0.143** | **0.143** | **0.143** | **0.143** | 0.143 |
| 0.50 | **B** | 0.1 | 0.9 | 0.750 | 0.825 | **0.848** | **0.854** | **0.856** | **0.857** | **0.857** | **0.857** | **0.857** | **0.857** | 0.857 |
| 0.50 | **A** | 0.4 | 0.6 | 0.300 | **0.260** | **0.252** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | 0.250 |
| 0.50 | **B** | 0.2 | 0.8 | 0.700 | **0.740** | **0.748** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | 0.750 |
| 0.50 | **A** | 0.4 | 0.6 | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | 0.400 |
| 0.50 | **B** | 0.4 | 0.6 | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | 0.600 |
| 0.50 | **A** | 0.4 | 0.6 | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.50 | **B** | 0.6 | 0.4 | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.50 | **A** | 0.4 | 0.6 | 0.600 | **0.560** | **0.576** | **0.570** | **0.572** | **0.571** | **0.572** | **0.571** | **0.571** | **0.571** | 0.571 |
| 0.50 | **B** | 0.8 | 0.2 | 0.400 | **0.440** | **0.424** | **0.430** | **0.428** | **0.429** | **0.428** | **0.429** | **0.429** | **0.429** | 0.429 |
| 0.50 | **A** | 0.4 | 0.6 | 0.650 | 0.575 | **0.613** | **0.594** | **0.603** | **0.598** | **0.601** | **0.600** | **0.600** | **0.600** | 0.600 |
| 0.50 | **B** | 0.9 | 0.1 | 0.350 | 0.425 | **0.388** | **0.406** | **0.397** | **0.402** | **0.399** | **0.400** | **0.400** | **0.400** | 0.400 |
| 0.50 | **A** | 0.6 | 0.4 | 0.350 | 0.275 | 0.238 | **0.219** | **0.209** | **0.205** | **0.202** | **0.201** | **0.201** | **0.200** | 0.200 |
| 0.50 | **B** | 0.1 | 0.9 | 0.650 | 0.725 | 0.763 | **0.781** | **0.791** | **0.795** | **0.798** | **0.799** | **0.799** | **0.800** | 0.800 |
| 0.50 | **A** | 0.6 | 0.4 | 0.400 | 0.360 | **0.344** | **0.338** | **0.335** | **0.334** | **0.334** | **0.333** | **0.333** | **0.333** | 0.333 |
| 0.50 | **B** | 0.2 | 0.8 | 0.600 | 0.640 | **0.656** | **0.662** | **0.665** | **0.666** | **0.666** | **0.667** | **0.667** | **0.667** | 0.667 |
| 0.50 | **A** | 0.6 | 0.4 | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.50 | **B** | 0.4 | 0.6 | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.50 | **A** | 0.6 | 0.4 | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | 0.600 |
| 0.50 | **B** | 0.6 | 0.4 | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | 0.400 |
| 0.50 | **A** | 0.6 | 0.4 | 0.700 | **0.660** | **0.668** | **0.666** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | 0.667 |
| 0.50 | **B** | 0.8 | 0.2 | 0.300 | **0.340** | **0.332** | **0.334** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | 0.333 |
| 0.50 | **A** | 0.6 | 0.4 | 0.750 | **0.675** | **0.698** | **0.691** | **0.693** | **0.692** | **0.692** | **0.692** | **0.692** | **0.692** | 0.692 |
| 0.50 | **B** | 0.9 | 0.1 | 0.250 | **0.325** | **0.303** | **0.309** | **0.307** | **0.308** | **0.308** | **0.308** | **0.308** | **0.308** | 0.308 |

**Table 15. Expected composition of recruitment waves and equilibrium distributions for two-level variables, given the composition of seeds and overall transition probabilities**

Transition probabilities represent each group's tendency toward in-group versus out-group recruitment. Cells AA and BB represent in-group recruitment, and cells AB and BA represent out-group recruitment. Wave compositions in **bold** font are within ±2.0% of the equilibrium distribution.

| Seeds (%) | | Transition probabilities | | Expected composition of recruitment wave … | | | | | | | | | | Equilibrium distribution |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **A** | **B** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | |
| 0.50 | **A** | 0.8 | 0.2 | 0.450 | 0.415 | 0.391 | 0.373 | 0.361 | **0.353** | **0.347** | **0.343** | **0.340** | **0.338** | 0.333 |
| 0.50 | **B** | 0.1 | 0.9 | 0.550 | 0.585 | 0.610 | 0.627 | 0.639 | **0.647** | **0.653** | **0.657** | **0.660** | **0.662** | 0.667 |
| 0.50 | **A** | 0.8 | 0.2 | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.50 | **B** | 0.2 | 0.8 | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.50 | **A** | 0.8 | 0.2 | 0.600 | 0.640 | **0.656** | **0.662** | **0.665** | **0.666** | **0.666** | **0.667** | **0.667** | **0.667** | 0.667 |
| 0.50 | **B** | 0.4 | 0.6 | 0.400 | 0.360 | **0.344** | **0.338** | **0.335** | **0.334** | **0.334** | **0.333** | **0.333** | **0.333** | 0.333 |
| 0.50 | **A** | 0.8 | 0.2 | 0.700 | **0.740** | **0.748** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | 0.750 |
| 0.50 | **B** | 0.6 | 0.4 | 0.300 | **0.260** | **0.252** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | 0.250 |
| 0.50 | **A** | 0.8 | 0.2 | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | 0.800 |
| 0.50 | **B** | 0.8 | 0.2 | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | 0.200 |
| 0.50 | **A** | 0.8 | 0.2 | 0.850 | **0.815** | **0.819** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | 0.818 |
| 0.50 | **B** | 0.9 | 0.1 | 0.150 | **0.185** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | 0.182 |
| 0.50 | **A** | 0.9 | 0.1 | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.50 | **B** | 0.1 | 0.9 | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.50 | **A** | 0.9 | 0.1 | 0.550 | 0.585 | 0.610 | 0.627 | 0.639 | **0.647** | **0.653** | **0.657** | **0.660** | **0.662** | 0.667 |
| 0.50 | **B** | 0.2 | 0.8 | 0.450 | 0.415 | 0.391 | 0.373 | 0.361 | **0.353** | **0.347** | **0.343** | **0.340** | **0.338** | 0.333 |
| 0.50 | **A** | 0.9 | 0.1 | 0.650 | 0.725 | 0.763 | **0.781** | **0.791** | **0.795** | **0.798** | **0.799** | **0.799** | **0.800** | 0.800 |
| 0.50 | **B** | 0.4 | 0.6 | 0.350 | 0.275 | 0.238 | **0.219** | **0.209** | **0.205** | **0.202** | **0.201** | **0.201** | **0.200** | 0.200 |
| 0.50 | **A** | 0.9 | 0.1 | 0.750 | 0.825 | **0.848** | **0.854** | **0.856** | **0.857** | **0.857** | **0.857** | **0.857** | **0.857** | 0.857 |
| 0.50 | **B** | 0.6 | 0.4 | 0.250 | 0.175 | **0.153** | **0.146** | **0.144** | **0.143** | **0.143** | **0.143** | **0.143** | **0.143** | 0.143 |
| 0.50 | **A** | 0.9 | 0.1 | 0.850 | **0.885** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | 0.889 |
| 0.50 | **B** | 0.8 | 0.2 | 0.150 | **0.115** | **0.112** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | 0.111 |
| 0.50 | **A** | 0.9 | 0.1 | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | 0.900 |
| 0.50 | **B** | 0.9 | 0.1 | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | 0.100 |

**Table 15. Expected composition of recruitment waves and equilibrium distributions for two-level variables, given the composition of seeds and overall transition probabilities**

Transition probabilities represent each group's tendency toward in-group versus out-group recruitment. Cells AA and BB represent in-group recruitment, and cells AB and BA represent out-group recruitment. Wave compositions in **bold** font are within ±2.0% of the equilibrium distribution.

| Seeds | | Transition probabilities | | Expected composition of recruitment wave … | | | | | | | | | | Equilibrium |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (%) | | **A** | **B** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | distribution |
| 0.66 | **A** | 0.1 | 0.9 | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | 0.100 |
| 0.34 | **B** | 0.1 | 0.9 | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | 0.900 |
| 0.66 | **A** | 0.1 | 0.9 | 0.134 | **0.187** | **0.181** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | 0.182 |
| 0.34 | **B** | 0.2 | 0.8 | 0.866 | **0.813** | **0.819** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | 0.818 |
| 0.66 | **A** | 0.1 | 0.9 | 0.202 | 0.339 | **0.298** | **0.311** | **0.307** | **0.308** | **0.308** | **0.308** | **0.308** | **0.308** | 0.308 |
| 0.34 | **B** | 0.4 | 0.6 | 0.798 | 0.661 | **0.702** | **0.689** | **0.693** | **0.692** | **0.692** | **0.692** | **0.692** | **0.692** | 0.692 |
| 0.66 | **A** | 0.1 | 0.9 | 0.270 | 0.465 | 0.368 | **0.416** | **0.392** | **0.404** | **0.398** | **0.401** | **0.399** | **0.400** | 0.400 |
| 0.34 | **B** | 0.6 | 0.4 | 0.730 | 0.535 | 0.633 | **0.584** | **0.608** | **0.596** | **0.602** | **0.599** | **0.601** | **0.600** | 0.600 |
| 0.66 | **A** | 0.1 | 0.9 | 0.338 | 0.563 | 0.406 | 0.516 | 0.439 | 0.493 | **0.455** | **0.482** | **0.463** | **0.476** | 0.471 |
| 0.34 | **B** | 0.8 | 0.2 | 0.662 | 0.437 | 0.594 | 0.484 | 0.561 | 0.507 | **0.545** | **0.518** | **0.537** | **0.524** | 0.529 |
| 0.66 | **A** | 0.1 | 0.9 | 0.372 | 0.602 | 0.418 | 0.566 | 0.448 | 0.542 | 0.466 | 0.527 | 0.479 | **0.517** | 0.500 |
| 0.34 | **B** | 0.9 | 0.1 | 0.628 | 0.398 | 0.582 | 0.434 | 0.552 | 0.458 | 0.534 | 0.473 | 0.521 | **0.483** | 0.500 |
| 0.66 | **A** | 0.2 | 0.8 | 0.166 | **0.117** | **0.112** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | 0.111 |
| 0.34 | **B** | 0.1 | 0.9 | 0.834 | **0.883** | **0.888** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | 0.889 |
| 0.66 | **A** | 0.2 | 0.8 | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | 0.200 |
| 0.34 | **B** | 0.2 | 0.8 | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | 0.800 |
| 0.66 | **A** | 0.2 | 0.8 | 0.268 | **0.346** | **0.331** | **0.334** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | 0.333 |
| 0.34 | **B** | 0.4 | 0.6 | 0.732 | **0.654** | **0.669** | **0.666** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | 0.667 |
| 0.66 | **A** | 0.2 | 0.8 | 0.336 | 0.466 | **0.414** | **0.434** | **0.426** | **0.430** | **0.428** | **0.429** | **0.429** | **0.429** | 0.429 |
| 0.34 | **B** | 0.6 | 0.4 | 0.664 | 0.534 | **0.586** | **0.566** | **0.574** | **0.570** | **0.572** | **0.571** | **0.571** | **0.571** | 0.571 |
| 0.66 | **A** | 0.2 | 0.8 | 0.404 | 0.558 | 0.465 | 0.521 | **0.488** | **0.507** | **0.496** | **0.503** | **0.498** | **0.501** | 0.500 |
| 0.34 | **B** | 0.8 | 0.2 | 0.596 | 0.442 | 0.535 | 0.479 | **0.512** | **0.493** | **0.504** | **0.497** | **0.502** | **0.499** | 0.500 |
| 0.66 | **A** | 0.2 | 0.8 | 0.438 | 0.593 | 0.485 | 0.561 | 0.507 | **0.545** | **0.519** | **0.537** | **0.524** | **0.533** | 0.529 |
| 0.34 | **B** | 0.9 | 0.1 | 0.562 | 0.407 | 0.515 | 0.439 | 0.493 | **0.455** | **0.481** | **0.463** | **0.476** | **0.467** | 0.471 |

**Table 15. Expected composition of recruitment waves and equilibrium distributions for two-level variables, given the composition of seeds and overall transition probabilities**

Transition probabilities represent each group's tendency toward in-group versus out-group recruitment. Cells AA and BB represent in-group recruitment, and cells AB and BA represent out-group recruitment. Wave compositions in **bold** font are within ±2.0% of the equilibrium distribution.

| Seeds (%) | | Transition probabilities | | Expected composition of recruitment wave … | | | | | | | | | | Equilibrium distribution |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **A** | **B** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | |
| 0.66 | **A** | 0.4 | 0.6 | 0.298 | 0.189 | **0.157** | **0.147** | **0.144** | **0.143** | **0.143** | **0.143** | **0.143** | **0.143** | 0.143 |
| 0.34 | **B** | 0.1 | 0.9 | 0.702 | 0.811 | **0.843** | **0.853** | **0.856** | **0.857** | **0.857** | **0.857** | **0.857** | **0.857** | 0.857 |
| 0.66 | **A** | 0.4 | 0.6 | 0.332 | **0.266** | **0.253** | **0.251** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | 0.250 |
| 0.34 | **B** | 0.2 | 0.8 | 0.668 | **0.734** | **0.747** | **0.749** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | 0.750 |
| 0.66 | **A** | 0.4 | 0.6 | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | 0.400 |
| 0.34 | **B** | 0.4 | 0.6 | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | 0.600 |
| 0.66 | **A** | 0.4 | 0.6 | 0.468 | **0.506** | **0.499** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.34 | **B** | 0.6 | 0.4 | 0.532 | **0.494** | **0.501** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.66 | **A** | 0.4 | 0.6 | 0.536 | **0.586** | **0.566** | **0.574** | **0.571** | **0.572** | **0.571** | **0.571** | **0.571** | **0.571** | 0.571 |
| 0.34 | **B** | 0.8 | 0.2 | 0.464 | **0.414** | **0.434** | **0.426** | **0.429** | **0.428** | **0.429** | **0.429** | **0.429** | **0.429** | 0.429 |
| 0.66 | **A** | 0.4 | 0.6 | 0.570 | **0.615** | **0.593** | **0.604** | **0.598** | **0.601** | **0.600** | **0.600** | **0.600** | **0.600** | 0.600 |
| 0.34 | **B** | 0.9 | 0.1 | 0.430 | **0.385** | **0.408** | **0.396** | **0.402** | **0.399** | **0.400** | **0.400** | **0.400** | **0.400** | 0.400 |
| 0.66 | **A** | 0.6 | 0.4 | 0.430 | 0.315 | 0.258 | 0.229 | **0.214** | **0.207** | **0.204** | **0.202** | **0.201** | **0.200** | 0.200 |
| 0.34 | **B** | 0.1 | 0.9 | 0.570 | 0.685 | 0.743 | 0.771 | **0.786** | **0.793** | **0.796** | **0.798** | **0.799** | **0.800** | 0.800 |
| 0.66 | **A** | 0.6 | 0.4 | 0.464 | 0.386 | 0.354 | **0.342** | **0.337** | **0.335** | **0.334** | **0.334** | **0.333** | **0.333** | 0.333 |
| 0.34 | **B** | 0.2 | 0.8 | 0.536 | 0.614 | 0.646 | **0.658** | **0.663** | **0.665** | **0.666** | **0.666** | **0.667** | **0.667** | 0.667 |
| 0.66 | **A** | 0.6 | 0.4 | 0.532 | **0.506** | **0.501** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.34 | **B** | 0.4 | 0.6 | 0.468 | **0.494** | **0.499** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.66 | **A** | 0.6 | 0.4 | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | 0.600 |
| 0.34 | **B** | 0.6 | 0.4 | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | 0.400 |
| 0.66 | **A** | 0.6 | 0.4 | **0.668** | **0.666** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | 0.667 |
| 0.34 | **B** | 0.8 | 0.2 | **0.332** | **0.334** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | 0.333 |
| 0.66 | **A** | 0.6 | 0.4 | **0.702** | **0.689** | **0.693** | **0.692** | **0.692** | **0.692** | **0.692** | **0.692** | **0.692** | **0.692** | 0.692 |
| 0.34 | **B** | 0.9 | 0.1 | **0.298** | **0.311** | **0.307** | **0.308** | **0.308** | **0.308** | **0.308** | **0.308** | **0.308** | **0.308** | 0.308 |

**Table 15. Expected composition of recruitment waves and equilibrium distributions for two-level variables, given the composition of seeds and overall transition probabilities**

Transition probabilities represent each group's tendency toward in-group versus out-group recruitment. Cells AA and BB represent in-group recruitment, and cells AB and BA represent out-group recruitment. Wave compositions in **bold** font are within ±2.0% of the equilibrium distribution.

| Seeds (%) | | Transition probabilities | | Expected composition of recruitment wave … | | | | | | | | | | Equilibrium distribution |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A | B | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| 0.66 | A | 0.8 | 0.2 | 0.562 | 0.493 | 0.445 | 0.412 | 0.388 | 0.372 | 0.360 | **0.352** | **0.347** | **0.343** | 0.333 |
| 0.34 | B | 0.1 | 0.9 | 0.438 | 0.507 | 0.555 | 0.588 | 0.612 | 0.628 | 0.640 | **0.648** | **0.653** | **0.657** | 0.667 |
| 0.66 | A | 0.8 | 0.2 | 0.596 | 0.558 | 0.535 | 0.521 | **0.512** | **0.507** | **0.504** | **0.503** | **0.502** | **0.501** | 0.500 |
| 0.34 | B | 0.2 | 0.8 | 0.404 | 0.442 | 0.465 | 0.479 | **0.488** | **0.493** | **0.496** | **0.497** | **0.498** | **0.499** | 0.500 |
| 0.66 | A | 0.8 | 0.2 | **0.664** | **0.666** | **0.666** | **0.666** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | 0.667 |
| 0.34 | B | 0.4 | 0.6 | **0.336** | **0.334** | **0.334** | **0.334** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | 0.333 |
| 0.66 | A | 0.8 | 0.2 | **0.732** | **0.746** | **0.749** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | 0.750 |
| 0.34 | B | 0.6 | 0.4 | **0.268** | **0.254** | **0.251** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | 0.250 |
| 0.66 | A | 0.8 | 0.2 | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | 0.800 |
| 0.34 | B | 0.8 | 0.2 | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | 0.200 |
| 0.66 | A | 0.8 | 0.2 | **0.834** | **0.817** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | 0.818 |
| 0.34 | B | 0.9 | 0.1 | **0.166** | **0.183** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | 0.182 |
| 0.66 | A | 0.9 | 0.1 | 0.628 | 0.602 | 0.582 | 0.566 | 0.552 | 0.542 | 0.534 | 0.527 | 0.521 | **0.517** | 0.500 |
| 0.34 | B | 0.1 | 0.9 | 0.372 | 0.398 | 0.418 | 0.434 | 0.448 | 0.458 | 0.466 | 0.473 | 0.479 | **0.483** | 0.500 |
| 0.66 | A | 0.9 | 0.1 | **0.662** | **0.663** | **0.664** | **0.665** | **0.666** | **0.666** | **0.666** | **0.666** | **0.666** | **0.666** | 0.667 |
| 0.34 | B | 0.2 | 0.8 | **0.338** | **0.337** | **0.336** | **0.335** | **0.334** | **0.334** | **0.334** | **0.334** | **0.334** | **0.334** | 0.333 |
| 0.66 | A | 0.9 | 0.1 | 0.730 | 0.765 | **0.783** | **0.791** | **0.796** | **0.798** | **0.799** | **0.799** | **0.800** | **0.800** | 0.800 |
| 0.34 | B | 0.4 | 0.6 | 0.270 | 0.235 | **0.218** | **0.209** | **0.204** | **0.202** | **0.201** | **0.201** | **0.200** | **0.200** | 0.200 |
| 0.66 | A | 0.9 | 0.1 | 0.798 | **0.839** | **0.852** | **0.856** | **0.857** | **0.857** | **0.857** | **0.857** | **0.857** | **0.857** | 0.857 |
| 0.34 | B | 0.6 | 0.4 | 0.202 | **0.161** | **0.148** | **0.144** | **0.143** | **0.143** | **0.143** | **0.143** | **0.143** | **0.143** | 0.143 |
| 0.66 | A | 0.9 | 0.1 | 0.866 | **0.887** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | 0.889 |
| 0.34 | B | 0.8 | 0.2 | 0.134 | **0.113** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | 0.111 |
| 0.66 | A | 0.9 | 0.1 | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | 0.900 |
| 0.34 | B | 0.9 | 0.1 | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | 0.100 |

**Table 15. Expected composition of recruitment waves and equilibrium distributions for two-level variables, given the composition of seeds and overall transition probabilities**

Transition probabilities represent each group's tendency toward in-group versus out-group recruitment. Cells AA and BB represent in-group recruitment, and cells AB and BA represent out-group recruitment. Wave compositions in **bold** font are within ±2.0% of the equilibrium distribution.

| Seeds (%) | | Transition probabilities | | Expected composition of recruitment wave … | | | | | | | | | | Equilibrium distribution |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **A** | **B** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | |
| 0.75 | **A** | 0.1 | 0.9 | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | 0.100 |
| 0.25 | **B** | 0.1 | 0.9 | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | 0.900 |
| 0.75 | **A** | 0.1 | 0.9 | 0.125 | **0.188** | **0.181** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | 0.182 |
| 0.25 | **B** | 0.2 | 0.8 | 0.875 | **0.813** | **0.819** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | 0.818 |
| 0.75 | **A** | 0.1 | 0.9 | 0.175 | 0.348 | **0.296** | **0.311** | **0.307** | **0.308** | **0.308** | **0.308** | **0.308** | **0.308** | 0.308 |
| 0.25 | **B** | 0.4 | 0.6 | 0.825 | 0.653 | **0.704** | **0.689** | **0.693** | **0.692** | **0.692** | **0.692** | **0.692** | **0.692** | 0.692 |
| 0.75 | **A** | 0.1 | 0.9 | 0.225 | 0.488 | 0.356 | 0.422 | **0.389** | **0.405** | **0.397** | **0.401** | **0.399** | **0.400** | 0.400 |
| 0.25 | **B** | 0.6 | 0.4 | 0.775 | 0.513 | 0.644 | 0.578 | **0.611** | **0.595** | **0.603** | **0.599** | **0.601** | **0.600** | 0.600 |
| 0.75 | **A** | 0.1 | 0.9 | 0.275 | 0.608 | 0.375 | 0.538 | 0.424 | 0.503 | 0.448 | **0.487** | **0.459** | **0.478** | 0.471 |
| 0.25 | **B** | 0.8 | 0.2 | 0.725 | 0.393 | 0.625 | 0.462 | 0.576 | 0.497 | 0.552 | **0.513** | **0.541** | **0.522** | 0.529 |
| 0.75 | **A** | 0.1 | 0.9 | 0.300 | 0.660 | 0.372 | 0.602 | 0.418 | 0.566 | 0.448 | 0.542 | 0.466 | 0.527 | 0.500 |
| 0.25 | **B** | 0.9 | 0.1 | 0.700 | 0.340 | 0.628 | 0.398 | 0.582 | 0.434 | 0.552 | 0.458 | 0.534 | 0.473 | 0.500 |
| 0.75 | **A** | 0.2 | 0.8 | 0.175 | **0.118** | **0.112** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | 0.111 |
| 0.25 | **B** | 0.1 | 0.9 | 0.825 | **0.883** | **0.888** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | 0.889 |
| 0.75 | **A** | 0.2 | 0.8 | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | 0.200 |
| 0.25 | **B** | 0.2 | 0.8 | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | 0.800 |
| 0.75 | **A** | 0.2 | 0.8 | 0.250 | **0.350** | **0.330** | **0.334** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | 0.333 |
| 0.25 | **B** | 0.4 | 0.6 | 0.750 | **0.650** | **0.670** | **0.666** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | 0.667 |
| 0.75 | **A** | 0.2 | 0.8 | 0.300 | 0.480 | 0.408 | **0.437** | **0.425** | **0.430** | **0.428** | **0.429** | **0.428** | **0.429** | 0.429 |
| 0.25 | **B** | 0.6 | 0.4 | 0.700 | 0.520 | 0.592 | **0.563** | **0.575** | **0.570** | **0.572** | **0.571** | **0.572** | **0.571** | 0.571 |
| 0.75 | **A** | 0.2 | 0.8 | 0.350 | 0.590 | 0.446 | 0.532 | **0.481** | **0.512** | **0.493** | **0.504** | **0.497** | **0.502** | 0.500 |
| 0.25 | **B** | 0.8 | 0.2 | 0.650 | 0.410 | 0.554 | 0.468 | **0.519** | **0.488** | **0.507** | **0.496** | **0.503** | **0.498** | 0.500 |
| 0.75 | **A** | 0.2 | 0.8 | 0.375 | 0.638 | 0.454 | 0.582 | 0.492 | 0.555 | **0.511** | **0.542** | **0.521** | **0.536** | 0.529 |
| 0.25 | **B** | 0.9 | 0.1 | 0.625 | 0.363 | 0.546 | 0.418 | 0.508 | 0.445 | **0.489** | **0.458** | **0.479** | **0.464** | 0.471 |

**Table 15. Expected composition of recruitment waves and equilibrium distributions for two-level variables, given the composition of seeds and overall transition probabilities**

Transition probabilities represent each group's tendency toward in-group versus out-group recruitment. Cells AA and BB represent in-group recruitment, and cells AB and BA represent out-group recruitment. Wave compositions in **bold** font are within ±2.0% of the equilibrium distribution.

| Seeds (%) | | Transition probabilities | | Expected composition of recruitment wave … | | | | | | | | | | Equilibrium distribution |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **A** | **B** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | |
| 0.75 | **A** | 0.4 | 0.6 | 0.325 | 0.198 | **0.159** | **0.148** | **0.144** | **0.143** | **0.143** | **0.143** | **0.143** | **0.143** | 0.143 |
| 0.25 | **B** | 0.1 | 0.9 | 0.675 | 0.803 | **0.841** | **0.852** | **0.856** | **0.857** | **0.857** | **0.857** | **0.857** | **0.857** | 0.857 |
| 0.75 | **A** | 0.4 | 0.6 | 0.350 | **0.270** | **0.254** | **0.251** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | 0.250 |
| 0.25 | **B** | 0.2 | 0.8 | 0.650 | **0.730** | **0.746** | **0.749** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | 0.750 |
| 0.75 | **A** | 0.4 | 0.6 | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | 0.400 |
| 0.25 | **B** | 0.4 | 0.6 | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | 0.600 |
| 0.75 | **A** | 0.4 | 0.6 | 0.450 | **0.510** | **0.498** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.25 | **B** | 0.6 | 0.4 | 0.550 | **0.490** | **0.502** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.75 | **A** | 0.4 | 0.6 | 0.500 | 0.600 | **0.560** | **0.576** | **0.570** | **0.572** | **0.571** | **0.572** | **0.571** | **0.571** | 0.571 |
| 0.25 | **B** | 0.8 | 0.2 | 0.500 | 0.400 | **0.440** | **0.424** | **0.430** | **0.428** | **0.429** | **0.428** | **0.429** | **0.429** | 0.429 |
| 0.75 | **A** | 0.4 | 0.6 | 0.525 | 0.638 | **0.581** | **0.609** | **0.595** | **0.602** | **0.599** | **0.601** | **0.600** | **0.600** | 0.600 |
| 0.25 | **B** | 0.9 | 0.1 | 0.475 | 0.363 | **0.419** | **0.391** | **0.405** | **0.398** | **0.401** | **0.399** | **0.400** | **0.400** | 0.400 |
| 0.75 | **A** | 0.6 | 0.4 | 0.475 | 0.338 | 0.269 | 0.234 | **0.217** | **0.209** | **0.204** | **0.202** | **0.201** | **0.201** | 0.200 |
| 0.25 | **B** | 0.1 | 0.9 | 0.525 | 0.663 | 0.731 | 0.766 | **0.783** | **0.791** | **0.796** | **0.798** | **0.799** | **0.799** | 0.800 |
| 0.75 | **A** | 0.6 | 0.4 | 0.500 | 0.400 | 0.360 | **0.344** | **0.338** | **0.335** | **0.334** | **0.334** | **0.333** | **0.333** | 0.333 |
| 0.25 | **B** | 0.2 | 0.8 | 0.500 | 0.600 | 0.640 | **0.656** | **0.662** | **0.665** | **0.666** | **0.666** | **0.667** | **0.667** | 0.667 |
| 0.75 | **A** | 0.6 | 0.4 | 0.550 | **0.510** | **0.502** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.25 | **B** | 0.4 | 0.6 | 0.450 | **0.490** | **0.498** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | **0.500** | 0.500 |
| 0.75 | **A** | 0.6 | 0.4 | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | **0.600** | 0.600 |
| 0.25 | **B** | 0.6 | 0.4 | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | **0.400** | 0.400 |
| 0.75 | **A** | 0.6 | 0.4 | **0.650** | **0.670** | **0.666** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | 0.667 |
| 0.25 | **B** | 0.8 | 0.2 | **0.350** | **0.330** | **0.334** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | 0.333 |
| 0.75 | **A** | 0.6 | 0.4 | **0.675** | **0.698** | **0.691** | **0.693** | **0.692** | **0.692** | **0.692** | **0.692** | **0.692** | **0.692** | 0.692 |
| 0.25 | **B** | 0.9 | 0.1 | **0.325** | **0.303** | **0.309** | **0.307** | **0.308** | **0.308** | **0.308** | **0.308** | **0.308** | **0.308** | 0.308 |

**Table 15. Expected composition of recruitment waves and equilibrium distributions for two-level variables, given the composition of seeds and overall transition probabilities**

Transition probabilities represent each group's tendency toward in-group versus out-group recruitment. Cells AA and BB represent in-group recruitment, and cells AB and BA represent out-group recruitment. Wave compositions in **bold** font are within ±2.0% of the equilibrium distribution.

| Seeds (%) | | Transition probabilities | | Expected composition of recruitment wave … | | | | | | | | | | Equilibrium distribution |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **A** | **B** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | |
| 0.75 | **A** | 0.8 | 0.2 | 0.625 | 0.538 | 0.476 | 0.433 | 0.403 | 0.382 | 0.368 | 0.357 | **0.350** | **0.345** | 0.333 |
| 0.25 | **B** | 0.1 | 0.9 | 0.375 | 0.463 | 0.524 | 0.567 | 0.597 | 0.618 | 0.632 | 0.643 | **0.650** | **0.655** | 0.667 |
| 0.75 | **A** | 0.8 | 0.2 | 0.650 | 0.590 | 0.554 | 0.532 | **0.519** | **0.512** | **0.507** | **0.504** | **0.503** | **0.502** | 0.500 |
| 0.25 | **B** | 0.2 | 0.8 | 0.350 | 0.410 | 0.446 | 0.468 | **0.481** | **0.488** | **0.493** | **0.496** | **0.497** | **0.498** | 0.500 |
| 0.75 | **A** | 0.8 | 0.2 | 0.700 | **0.680** | **0.672** | **0.669** | **0.668** | **0.667** | **0.667** | **0.667** | **0.667** | **0.667** | 0.667 |
| 0.25 | **B** | 0.4 | 0.6 | 0.300 | **0.320** | **0.328** | **0.331** | **0.332** | **0.333** | **0.333** | **0.333** | **0.333** | **0.333** | 0.333 |
| 0.75 | **A** | 0.8 | 0.2 | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | 0.750 |
| 0.25 | **B** | 0.6 | 0.4 | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | **0.250** | 0.250 |
| 0.75 | **A** | 0.8 | 0.2 | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | **0.800** | 0.800 |
| 0.25 | **B** | 0.8 | 0.2 | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | **0.200** | 0.200 |
| 0.75 | **A** | 0.8 | 0.2 | **0.825** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | **0.818** | 0.818 |
| 0.25 | **B** | 0.9 | 0.1 | **0.175** | **0.183** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | **0.182** | 0.182 |
| 0.75 | **A** | 0.9 | 0.1 | 0.700 | 0.660 | 0.628 | 0.602 | 0.582 | 0.566 | 0.552 | 0.542 | 0.534 | 0.527 | 0.500 |
| 0.25 | **B** | 0.1 | 0.9 | 0.300 | 0.340 | 0.372 | 0.398 | 0.418 | 0.434 | 0.448 | 0.458 | 0.466 | 0.473 | 0.500 |
| 0.75 | **A** | 0.9 | 0.1 | 0.725 | 0.708 | 0.695 | 0.687 | **0.681** | **0.676** | **0.674** | **0.671** | **0.670** | **0.669** | 0.667 |
| 0.25 | **B** | 0.2 | 0.8 | 0.275 | 0.293 | 0.305 | 0.313 | **0.319** | **0.324** | **0.326** | **0.329** | **0.330** | **0.331** | 0.333 |
| 0.75 | **A** | 0.9 | 0.1 | 0.775 | **0.788** | **0.794** | **0.797** | **0.798** | **0.799** | **0.800** | **0.800** | **0.800** | **0.800** | 0.800 |
| 0.25 | **B** | 0.4 | 0.6 | 0.225 | **0.213** | **0.206** | **0.203** | **0.202** | **0.201** | **0.200** | **0.200** | **0.200** | **0.200** | 0.200 |
| 0.75 | **A** | 0.9 | 0.1 | 0.825 | **0.848** | **0.854** | **0.856** | **0.857** | **0.857** | **0.857** | **0.857** | **0.857** | **0.857** | 0.857 |
| 0.25 | **B** | 0.6 | 0.4 | 0.175 | **0.153** | **0.146** | **0.144** | **0.143** | **0.143** | **0.143** | **0.143** | **0.143** | **0.143** | 0.143 |
| 0.75 | **A** | 0.9 | 0.1 | **0.875** | **0.888** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | **0.889** | 0.889 |
| 0.25 | **B** | 0.8 | 0.2 | **0.125** | **0.113** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | **0.111** | 0.111 |
| 0.75 | **A** | 0.9 | 0.1 | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | **0.900** | 0.900 |
| 0.25 | **B** | 0.9 | 0.1 | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | **0.100** | 0.100 |

**Table 16. Cumulative sample size by number of seeds and average number of new recruits per participant**

| Seeds (n) | Average new recruits per participant | Total sample size as of wave … | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 5 | 0.66 | 3 | 4 | 6 | 9 | 14 | 23 | 38 | 63 | 104 | 172 |
| 8 | 0.66 | 5 | 8 | 13 | 21 | 34 | 56 | 92 | 152 | 252 | 418 |
| 12 | 0.66 | 7 | 11 | 18 | 29 | 48 | 79 | 131 | 217 | 360 | 597 |
| 15 | 0.66 | 9 | 14 | 23 | 38 | 63 | 104 | 172 | 285 | 473 | 785 |
| 20 | 0.66 | 13 | 21 | 34 | 56 | 92 | 152 | 252 | 418 | 693 | 1150 |
| 3 | 0.75 | 2 | 3 | 5 | 8 | 14 | 24 | 42 | 73 | 127 | 222 |
| 5 | 0.75 | 3 | 5 | 8 | 14 | 24 | 42 | 73 | 127 | 222 | 388 |
| 8 | 0.75 | 6 | 10 | 17 | 29 | 50 | 87 | 152 | 266 | 465 | 813 |
| 12 | 0.75 | 9 | 15 | 26 | 45 | 78 | 136 | 238 | 416 | 728 | 1274 |
| 15 | 0.75 | 11 | 19 | 33 | 57 | 99 | 173 | 302 | 528 | 924 | 1617 |
| 20 | 0.75 | 15 | 26 | 45 | 78 | 136 | 238 | 416 | 728 | 1274 | 2229 |
| 3 | 0.9 | 2 | 3 | 5 | 9 | 17 | 32 | 60 | 114 | 216 | 410 |
| 5 | 0.9 | 4 | 7 | 13 | 24 | 45 | 85 | 161 | 305 | 579 | 1100 |
| 8 | 0.9 | 7 | 13 | 24 | 45 | 85 | 161 | 305 | 579 | 1100 | 2090 |
| 12 | 0.9 | 10 | 19 | 36 | 68 | 129 | 245 | 465 | 883 | 1677 | 3186 |
| 15 | 0.9 | 13 | 24 | 45 | 85 | 161 | 305 | 579 | 1100 | 2090 | 3971 |
| 20 | 0.9 | 18 | 34 | 64 | 121 | 229 | 435 | 826 | 1569 | 2981 | 5663 |
| 3 | 1.0 | 3 | 6 | 12 | 24 | 48 | 96 | 192 | 384 | 768 | 1536 |
| 5 | 1.0 | 5 | 10 | 20 | 40 | 80 | 160 | 320 | 640 | 1280 | 2560 |
| 8 | 1.0 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 | 2048 | 4096 |
| 12 | 1.0 | 12 | 24 | 48 | 96 | 192 | 384 | 768 | 1536 | 3072 | 6144 |
| 15 | 1.0 | 15 | 30 | 60 | 120 | 240 | 480 | 960 | 1920 | 3840 | 7680 |
| 20 | 1.0 | 20 | 40 | 80 | 160 | 320 | 640 | 1280 | 2560 | 5120 | 10240 |
| 3 | 1.1 | 3 | 6 | 12 | 25 | 52 | 109 | 228 | 478 | 1003 | 2106 |
| 5 | 1.1 | 5 | 10 | 21 | 44 | 92 | 193 | 405 | 850 | 1785 | 3748 |
| 8 | 1.1 | 8 | 16 | 33 | 69 | 144 | 302 | 634 | 1331 | 2795 | 5869 |
| 12 | 1.1 | 13 | 27 | 56 | 117 | 245 | 514 | 1079 | 2265 | 4756 | 9987 |
| 15 | 1.1 | 16 | 33 | 69 | 144 | 302 | 634 | 1331 | 2795 | 5869 | 12324 |
| 20 | 1.1 | 22 | 46 | 96 | 201 | 422 | 886 | 1860 | 3906 | 8202 | 17224 |
| 3 | 1.2 | 3 | 6 | 13 | 28 | 61 | 134 | 294 | 646 | 1421 | 3126 |
| 5 | 1.2 | 6 | 13 | 28 | 61 | 134 | 294 | 646 | 1421 | 3126 | 6877 |
| 8 | 1.2 | 9 | 19 | 41 | 90 | 198 | 435 | 957 | 2105 | 4631 | 10188 |
| 12 | 1.2 | 14 | 30 | 66 | 145 | 319 | 701 | 1542 | 3392 | 7462 | 16416 |
| 15 | 1.2 | 18 | 39 | 85 | 187 | 411 | 904 | 1988 | 4373 | 9620 | 21164 |
| 20 | 1.2 | 24 | 52 | 114 | 250 | 550 | 1210 | 2662 | 5856 | 12883 | 28342 |

| Seeds (n) | Average new recruits per participant | Total sample size as of wave … | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 3 | 1.3 | 3 | 6 | 13 | 29 | 66 | 151 | 347 | 798 | 1835 | 4220 |
| 5 | 1.3 | 6 | 13 | 29 | 66 | 151 | 347 | 798 | 1835 | 4220 | 9706 |
| 8 | 1.3 | 10 | 23 | 52 | 119 | 273 | 627 | 1442 | 3316 | 7626 | 17539 |
| 12 | 1.3 | 15 | 34 | 78 | 179 | 411 | 945 | 2173 | 4997 | 11493 | 26433 |
| 15 | 1.3 | 19 | 43 | 98 | 225 | 517 | 1189 | 2734 | 6288 | 14462 | 33262 |
| 20 | 1.3 | 26 | 59 | 135 | 310 | 713 | 1639 | 3769 | 8668 | 19936 | 45852 |
| 3 | 1.4 | 4 | 9 | 21 | 50 | 120 | 288 | 691 | 1658 | 3979 | 9549 |
| 5 | 1.4 | 7 | 16 | 38 | 91 | 218 | 523 | 1255 | 3012 | 7228 | 17347 |
| 8 | 1.4 | 11 | 26 | 62 | 148 | 355 | 852 | 2044 | 4905 | 11772 | 28252 |
| 12 | 1.4 | 16 | 38 | 91 | 218 | 523 | 1255 | 3012 | 7228 | 17347 | 41632 |
| 15 | 1.4 | 21 | 50 | 120 | 288 | 691 | 1658 | 3979 | 9549 | 22917 | 55000 |
| 20 | 1.4 | 28 | 67 | 160 | 384 | 921 | 2210 | 5304 | 12729 | 30549 | 73317 |
| 3 | 1.5 | 4 | 10 | 25 | 62 | 155 | 387 | 967 | 2417 | 6042 | 15105 |
| 5 | 1.5 | 7 | 17 | 42 | 105 | 262 | 655 | 1637 | 4092 | 10230 | 25575 |
| 8 | 1.5 | 12 | 30 | 75 | 187 | 467 | 1167 | 2917 | 7292 | 18230 | 45575 |
| 12 | 1.5 | 18 | 45 | 112 | 280 | 700 | 1750 | 4375 | 10937 | 27342 | 68355 |
| 15 | 1.5 | 22 | 55 | 137 | 342 | 855 | 2137 | 5342 | 13355 | 33387 | 83467 |
| 20 | 1.5 | 30 | 75 | 187 | 467 | 1167 | 2917 | 7292 | 18230 | 45575 | 113937 |
| 3 | 1.6 | 4 | 10 | 26 | 67 | 174 | 452 | 1175 | 3055 | 7943 | 20651 |
| 5 | 1.6 | 8 | 20 | 52 | 135 | 351 | 912 | 2371 | 6164 | 16026 | 41667 |
| 8 | 1.6 | 12 | 31 | 80 | 208 | 540 | 1404 | 3650 | 9490 | 24674 | 64152 |
| 12 | 1.6 | 19 | 49 | 127 | 330 | 858 | 2230 | 5798 | 15074 | 39192 | 101899 |
| 15 | 1.6 | 24 | 62 | 161 | 418 | 1086 | 2823 | 7339 | 19081 | 49610 | 128986 |
| 20 | 1.6 | 32 | 83 | 215 | 559 | 1453 | 3777 | 9820 | 25532 | 66383 | 172595 |
| 3 | 1.7 | 5 | 13 | 35 | 94 | 253 | 683 | 1844 | 4978 | 13440 | 36288 |
| 5 | 1.7 | 8 | 21 | 56 | 151 | 407 | 1098 | 2964 | 8002 | 21605 | 58333 |
| 8 | 1.7 | 13 | 35 | 94 | 253 | 683 | 1844 | 4978 | 13440 | 36288 | 97977 |
| 12 | 1.7 | 20 | 54 | 145 | 391 | 1055 | 2848 | 7689 | 20760 | 56052 | 151340 |
| 15 | 1.7 | 25 | 67 | 180 | 486 | 1312 | 3542 | 9563 | 25820 | 69714 | 188227 |
| 20 | 1.7 | 34 | 91 | 245 | 661 | 1784 | 4816 | 13003 | 35108 | 94791 | 255935 |
| 3 | 1.8 | 5 | 14 | 39 | 109 | 305 | 854 | 2391 | 6694 | 18743 | 52480 |
| 5 | 1.8 | 9 | 25 | 70 | 196 | 548 | 1534 | 4295 | 12026 | 33672 | 94281 |
| 8 | 1.8 | 14 | 39 | 109 | 305 | 854 | 2391 | 6694 | 18743 | 52480 | 146944 |
| 12 | 1.8 | 21 | 58 | 162 | 453 | 1268 | 3550 | 9940 | 27832 | 77929 | 218201 |
| 15 | 1.8 | 27 | 75 | 210 | 588 | 1646 | 4608 | 12902 | 36125 | 101150 | 283220 |
| 20 | 1.8 | 36 | 100 | 280 | 784 | 2195 | 6146 | 17208 | 48182 | 134909 | 377745 |
| 3 | 1.9 | 5 | 14 | 40 | 116 | 336 | 974 | 2824 | 8189 | 23748 | 68869 |
| 5 | 1.9 | 9 | 26 | 75 | 217 | 629 | 1824 | 5289 | 15338 | 44480 | 128992 |
| 8 | 1.9 | 15 | 43 | 124 | 359 | 1041 | 3018 | 8752 | 25380 | 73602 | 213445 |
| 12 | 1.9 | 22 | 63 | 182 | 527 | 1528 | 4431 | 12849 | 37262 | 108059 | 313371 |
| 15 | 1.9 | 28 | 81 | 234 | 678 | 1966 | 5701 | 16532 | 47942 | 139031 | 403189 |
| 20 | 1.9 | 38 | 110 | 319 | 925 | 2682 | 7777 | 22553 | 65403 | 189668 | 550037 |

| Seeds (n) | Average new recruits per participant | Total sample size as of wave … | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 3 | 2.0 | 6 | 18 | 54 | 162 | 486 | 1458 | 4374 | 13122 | 39366 | 118098 |
| 5 | 2.0 | 10 | 30 | 90 | 270 | 810 | 2430 | 7290 | 21870 | 65610 | 196830 |
| 8 | 2.0 | 16 | 48 | 144 | 432 | 1296 | 3888 | 11664 | 34992 | 104976 | 314928 |
| 12 | 2.0 | 24 | 72 | 216 | 648 | 1944 | 5832 | 17496 | 52488 | 157464 | 472392 |
| 15 | 2.0 | 30 | 90 | 270 | 810 | 2430 | 7290 | 21870 | 65610 | 196830 | 590490 |
| 20 | 2.0 | 40 | 120 | 360 | 1080 | 3240 | 9720 | 29160 | 87480 | 262440 | 787320 |

## Glossary

**Absorbing state:** In RDS, this term describes a subgroup whose members only recruit other members of the same subgroup.

**Ergodic:** A term to describe a process that can move from any state to any other state. In RDS, an ergodic system means that any respondent could potentially recruit anyone else, regardless of personal characteristics.

**Non-ergodic:** A term to describe a system with an absorbing state (see above). In RDS, a non-ergodic system would be one in which one subgroup never recruits from another, e.g. women recruit other women but never men.

**Equilibrium sample**: The sample achieved when the composition of characteristics changes very little from wave to wave; occurs after a “large” number of waves.

**Equilibrium proportion**: The proportion of people with a given characteristic in the equilibrium sample (see above).

**Transition probability**: The probability that a respondent in a given group was recruited by a member of the same (or of another) group.

**Probability matrix**: All transition probabilities for a trait, derived from the 2x2 or 3x3 table matching recruiters and recruits.

## References

**1.** Lapidus JA, Bertolli J, McGowan K, Sullivan P. HIV-related risk behaviors, perceptions of risk, HIV testing, and exposure to prevention messages and methods among urban American Indians and Alaska Natives. *AIDS Educ Prev.* Dec 2006;18(6):546-559.

**2.** Centers for Disease Control and Prevention. *HIV Testing Survey, 2000.* Atlanta, GA: Department of Health and Human Services, Centers for Disease Control and Prevention; 2003. HIV/AIDS Special Surveillance Report, Vol. 1, No. 1.

**3.** Centers for Disease Control and Prevention. *HIV Testing Survey, 2001.* Atlanta, GA: Department of Health and Human Services, Centers for Disease Control and Prevention; 2004. Special Surveillance Report Number 1.

**4.** Centers for Disease Control and Prevention. *HIV Testing Survey, 2002.* Atlanta, GA: Department of Health and Human Services, Centers for Disease Control and Prevention; 2004. Special Surveillance Report Number 5.

**5.** Hecht FM, Chesney MA, Lehman JS, et al. Does HIV reporting by name deter testing? MESH Study Group. *Aids.* Aug 18 2000;14(12):1801-1808.

**6.** Magnani R, Sabin K, Saidel T, Heckathorn D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. *Aids.* May 2005;19 Suppl 2:S67-72.

**7.** Heckathorn D, Jeffri J. Social networks of jazz musicians. *Changing the beat : a study of the worklife of jazz musicians.* Vol 3. Washington, D.C.: National Endowment for the Arts; 2003:48-61.

**8.** Heckathorn DD. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems.* May 1997;44(2):174-199.

**9.** Frost SDW, Brouwer KC, Firestone Cruz MA, et al. Respondent-driven sampling of injection drug users in two U.S.-Mexico border cities: recruitment dynamics and impact on estimates of HIV and syphilis prevalence. *Journal Of Urban Health: Bulletin Of The New York Academy Of Medicine.* 2006;83(6 Suppl):i83-97.

**10.** Falck RS, Siegal HA, Wang J, Carlson RG, Draus PJ. Nonmedical drug use among stimulant-using adults in small towns in rural Ohio. *Journal of Substance Abuse Treatment.* 2005;28(4):341-349.

**11.** Ramirez-Valles J, Heckathorn DD, VÃ¡zquez R, Diaz RM, Campbell RT. From networks to populations: the development and application of respondent-driven sampling among IDUs and Latino gay men. *AIDS And Behavior.* 2005;9(4):387-402.

**12.** Clements-Nolle K, Marx R, Katz M. Attempted Suicide Among Transgender Persons: The Influence of Gender-Based Discrimination and Victimization. *Journal of Homosexuality.* 2006;51(3):53-69.

**13.** Simic M, Johnston LG, Platt L, et al. Exploring barriers to 'respondent driven sampling' in sex worker and drug-injecting sex worker populations in Eastern Europe. *Journal Of Urban Health: Bulletin Of The New York Academy Of Medicine.* 2006;83(6 Suppl):i6-15.

**14.** Heckathorn D. Development of a theory of collective action: from the emergence of norms to AIDS prevention and the analysis of social structure. In: Berger J, Zelditch M, eds. *New directions in contemporary sociological theory*. Lanham, Md.: Rowman & Littlefield Publishers; 2002:79-108.

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

**16.** Heckathorn DD. Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems.* 2002;49(1):11.

**17.** Servegev B, Oparina T, Rumyantseva TP, et al. HIV prevention in Yaroslavl, Russia: A peer-driven intervention and needle exchange. *Journal of Drug Issues.* Fal 1999;29(4):777-804.

**18.** Frost SD, Brouwer KC, Firestone Cruz MA, et al. Respondent-driven sampling of injection drug users in two U.S.-Mexico border cities: recruitment dynamics and impact on estimates of HIV and syphilis prevalence. *J Urban Health.* Nov 2006;83(6 Suppl):i83-97.

**19.** Heckathorn D. Respondent-driven sampling downloads. 4/21/2006; www.respondentdrivensampling.org. Accessed April 2007.

**20.** Stormer A, Tun W, Guli L, et al. An analysis of respondent driven sampling with Injection Drug Users (IDU) in Albania and the Russian Federation. *Journal Of Urban Health: Bulletin Of The New York Academy Of Medicine.* 2006;83(6 Suppl):i73-82.

**21.** McKnight C, Des Jarlais D, Bramson H, et al. Respondent-driven sampling in a study of drug users in New York City: notes from the field. *J Urban Health.* Nov 2006;83(6 Suppl):i54-59.

**22.** Yeka W, Maibani-Michie G, Prybylski D, Colby D. Application of respondent driven sampling to collect baseline data on FSWs and MSM for HIV risk reduction interventions in two urban centres in Papua New Guinea. *Journal Of Urban Health: Bulletin Of The New York Academy Of Medicine.* 2006;83(6 Suppl):i60-72.

**23.** Carlson RG, Wang J, Falck RS, Siegal HA. Drug use practices among MDMA/ecstasy users in Ohio: a latent class analysis. *Drug & Alcohol Dependence.* 2005;79(2):167-179.

**24.** Draus PJ, Siegal HA, Carlson RG, Falck RS, Wang J. Cracking the cornfields: Recruiting illicit stimulant drug users in rural Ohio. *Sociological Quarterly.* Winter 2005;46(1):165-189.

**25.** Wang J, Carlson RG, Falck RS, Siegal HA, Rahman A, Li L. Respondent-driven sampling to recruit MDMA users: a methodological assessment. *Drug & Alcohol Dependence.* 2005;78(2):147-157.

**26.** Wang J, Falck RS, Li L, Rahman A, Carlson RG. Respondent-driven sampling in the recruitment of illicit stimulant drug users in a rural setting: Findings and technical issues. *Addictive Behaviors.* 2007;32(5):924-937.

**27.** Platt L, Wall M, Rhodes T, et al. Methods to recruit hard-to-reach groups: comparing two chain referral sampling methods of recruiting injecting drug users across nine studies in Russia and Estonia. *Journal Of Urban Health: Bulletin Of The New York Academy Of Medicine.* 2006;83(6 Suppl):i39-53.

**28.** Steele MS, Bukusi E, Cohen CR, Shell-Duncan BA, Holmes KK. Male genital hygiene beliefs and practices in Nairobi, Kenya. *Sexually Transmitted Infections.* 2004;80(6):471-476.

**29.** Clements-Nolle K, Marx R, Guzman R, Katz M. HIV prevalence, risk behaviors, health care use, and mental health status of transgender persons: implications for public health intervention. *American Journal Of Public Health.* 2001;91(6):915-921.

**30.** Grinstead CM, Snell JL, Snell JL. *Introduction to probability*. 2nd rev. ed. Providence, RI: American Mathematical Society; 1997.

**31.** Kemeny JG, Snell JL. *Finite Markov chains*. Princeton, N.J.: Van Nostrand; 1960. Cited in Heckathorn, D. (1997) "Respondent-driven sampling: a new approach to the study of hidden populations." Social Problems 44(2):174-199.

**32.** Salganik MJ, Heckathorn DD. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology.* 2004;34(1):193-239.

**33.** *SAS* [computer program]. Version 8.02. Cary, NC, USA; 1999-2001.

**34.** Agresti A. A survey of exact inference for contingency tables. *Statistical Science.* Feb 1992;7(1):131-153.

**35.** Evans-Campbell T, Lindhorst T, Huang B, Walters KL. Interpersonal violence in the lives of urban American Indian and Alaska Native women: implications for health, mental health, and help-seeking. *Am J Public Health.* Aug 2006;96(8):1416-1422.

**36.** Walters KL, Evans-Campbell T, Simoni J, Old Person R, Stately A. Historical trauma, microaggressions, and colonial trauma response: A decolonization framework for HIV prevention efforts among indigenous communities. *Embracing Our Traditions, Values, and Teachings [conference]*. Anchorage, Alaska; 2006.

**37.** Semaan S, Lauby J, Liebman J. Street and network sampling in evaluation studies of HIV risk-reduction interventions. *AIDS Rev.* Oct-Dec 2002;4(4):213-223.

## Appendix: Sample SAS code

```
***********************************************************************
***********************************************************************
**                                                                   **
**  This code creates a “family” of all possible 3x3 tables with     **
**  fixed row and column totals (set in the first step).             **
**                                                                   **
***********************************************************************
***********************************************************************

** Start with the assumption of total n = 109;
** The range of possible row totals will be set using the frequencies
** of a and b in all but the last wave (potential recruiters); n in
** group a = 34, b = 66;
** Recruits in columns c1 = 36, c2 = 73;
** Coupons = 0 to 3 per person;

** Under these assumptions, the most "extreme" scenarios are:
** r1 = 2, r2 = 107 ;
** r1 = 102, r2 = 7 ;

** initialize dataset to contain all row permutations for 2x2 tables in
this range;

data data1;
      do i = 2 to 102;
            r1 = i;
            n = 109;
            r2 = n - r2;
            c1 = 36;
            c2 = 73;
      output;
      end;
run;


** Create all possible tables by setting the first cell to 0;
** Some of the tables will be non-ergodic but are needed for rescaling
** table probabilities;
** If table probabilities are not required, set the first cell to 1;
** Cycle the first cell from minimum value to the maximum value - which
is one less than either the row total or the column total (whichever is
smallest);
** This will generate some excess tables with negative cells – these
will be deleted in the next step;

data data2;
      set data1;

   * the first cell is n11;
   * the cell in the first row, second column, is n12, etc;
```

```
      do i = 0 to min(r1,c1) by 1;

            n11=i;
            n12=r1-n11;
            n21=c1-n11;
            n22=r2-n21;

     *output observation to dataset;
        output;
        end;

      keep r1 r2 c1 c2 n n11 n12 n21 n22;
run;

data data3;
      set data2;
      IF (n11<0 OR n12<0 OR n21<0 OR n22<0) then delete;
run;


** Number the tables sequentially;

data data4;
  set data3;
  table + 1;
  by r2;
  if first.r2 then table = 1;
run;


data data5;
      set data4;

      * Calculate the hypergeometric probability for each table;
            f1=fact(r1)/fact(n11);
            f2=fact(r2)/fact(n12);
            f3=fact(c1)/fact(n21);
            f4=fact(c2)/fact(n22);
            fn=fact(n);

            * probability under hypergeometric distribution;
            probtable=(f1*f2*f3*f4)/fn;

      * transition probability based on row marginals;
            tp11 = round(n11/r1, .001);
            tp12 = round(n12/r1, .001);
            tp21 = round(n21/r2, .001);
            tp22 = round(n22/r2, .001);
      keep r2 table n11 n12 n21 n22 probtable tp11 tp12 tp21 tp22;
run;


* flag tables where Markov process is non-ergodic, i.e. transition
probability = 0;

data data6;
      set data5;
      ergodic = 1;
      if (tp11=0 or tp12=0 or tp21=0 or tp22=0) then ergodic = 0;
run;

* create file containing the sum of the probabilities of non-ergodic
tables for adjusting later;

PROC MEANS DATA=data6 NWAY ;
  WHERE (ergodic = 0);
  CLASS r2 ;
  VAR probtable ;
  OUTPUT OUT=data7a SUM=kprob N=cnt ;
RUN;

* sort that output file and the full data file and merge by r2;

DATA data7b;
      set data7a;
      keep r2 kprob;
RUN;

PROC SORT DATA=data7b;
      BY r2;
RUN;

PROC SORT DATA=data6;
      BY r2;
RUN;

DATA data8;
      MERGE data6 data7b;
      BY r2;
RUN;

* create variable for rescaled table probabilities,
      p(j)' = p(j) / ( 1 - p(1) - p(k))
      where p(1) and p(k) are the probabililites summed in the earlier
step;

DATA data9;
      set data8;
      jprobtab = probtable / (1-kprob);
RUN;


* Calculate equilibrium proportions E1 and E2 directly;
* Create variables partE1 and partE2 that will contribute to weighted
mean in next step;

data data10;
      set data9;
      E1 = tp21/(1-tp11+tp21);
      E2 = tp12/(1+tp12-tp22);
      parte1 = jprobtab*E1;
      parte2 = jprobtab*E2;
run;
```

```
PROC MEANS DATA=data10 NWAY ;
  WHERE (ergodic = 1);
  CLASS r2 ;
  VAR e1 ;
  OUTPUT OUT=data11 MEAN=meane1 MEDIAN=mediane1 N=cnt ;
RUN;

PROC MEANS DATA=data10 NWAY ;
  WHERE (ergodic = 1);
  CLASS r2 ;
  VAR e2 ;
  OUTPUT OUT=data12 MEAN=meane2 MEDIAN=mediane2 N=cnt ;
RUN;

PROC MEANS DATA=data10 NWAY ;
  WHERE (ergodic = 1);
  CLASS r2 ;
  VAR parte1 ;
  OUTPUT OUT=data13 SUM=weighte1 N=cnt ;
RUN;
PROC MEANS DATA=data10 NWAY ;
  WHERE (ergodic = 1);
  CLASS r2 ;
  VAR parte2 ;
  OUTPUT OUT=data14 SUM=weighte2 N=cnt ;
RUN;

data data11;
      set data11;
      keep r2 meane1 mediane1 cnt;
run;

data data12;
      set data12;
      keep r2 meane2 mediane2;
run;

data data13;
      set data13;
      keep r2 weighte1;
run;

data data14;
      set data14;
      keep r2 weighte2;
run;

proc sort data=data11; by r2; run;
proc sort data=data12; by r2; run;
proc sort data=data13; by r2; run;
proc sort data=data14; by r2; run;

data data15;
      merge data11 data12 data13 data14;
      by r2;
run;
```

```
PROC PRINT data=data15;
      var r2 cnt meane1 meane2 mediane1 mediane2 weighte1 weighte2;
run;



** print the range of values in the summary dataset;

PROC MEANS DATA=data15 n min max;
  VAR meane1 meane2 mediane1 mediane2 weighte1 weighte2 cnt;
RUN;

proc univariate data=data15 noprint;
   var meane1 meane2 mediane1 mediane2 weighte1 weighte2;
   output out=percentiles pctlpts=2.5 5 95 97.5
            pctlpre=meane1 meane2 mediane1 mediane2 weighte1 weighte2;
run;




***********************************************************************
***********************************************************************
**                                                                   **
**  This code creates a “family” of all possible 3x3 tables with     **
**  fixed row and column totals (set in the first step).             **
**                                                                   **
***********************************************************************
***********************************************************************

* In this example, row totals are set to the same proportions as
potential recruiters in the data;

* Data steps are ordered to minimize file size;

data data1;

  * initialize row and column totals - proportional rows;

  r1 = 53;
  r2 = 36;
  r3 = 16;
  c1 = 54; c2 = 34; c3 = 17;

run;


data data2;
      set data1;

  * n13 is the cell in the first row, third column;

      do i = 1 to (c3-2) by 1;
            n13=i;
      output;
      end;
      drop i;
run;


data data3;
      set data2;
      do i = 1 to (c2-2) by 1;
            n12 = i;
      output;
      end;
      drop i;
run;

data data4;
      set data3;
      do i = 1 to (c1-2) by 1;
            n11 = i;
            row1 = n11 + n12 + n13;
      output;
      end;
      drop i;
run;

data data4a;
      set data4;
      where row1 = r1;
run;

data data5;
      set data4a;
      do i = 1 to (c3-n13-1) by 1;
            n23=i;
      output;
      end;
      drop i row1;
run;

data data6;
      set data5;
      do i = 1 to (c2-n12-1) by 1;
            n22 = i;
      output;
      end;
      drop i;
run;

data data7;
      set data6;
      do i = 1 to (c1-n11-1) by 1;
            n21 = i;
            row2 = n21 + n22 + n23;
      output;
      end;
      drop i;
run;

data data8;
```

```
      set data7;
      where row2 = r2;
      n31=c1-n11-n21;
      n32=c2-n12-n22;
      n33=c3-n13-n23;
      n=r1+r2+r3;
      drop row2;
run;


* At this point, all tables have been generated;
* The proc means step is used to examine the data and make sure the
simulation was successful;

proc means n min max;
      var n11 n12 n13 n21 n22 n23 n31 n32 n33 r1 r2 r3 n;
run;


* The final step is to calculate the transition probabilities
(represented by tpxx variables) and equilibrium proportions
(represented by Ex variables);

data data9;
      set data8;

   * transition probabilities - based on row marginals;

   tp11 = round(n11/r1, .001);
   tp12 = round(n12/r1, .001);
   tp13 = round(n13/r1, .001);
   tp21 = round(n21/r2, .001);
   tp22 = round(n22/r2, .001);
   tp23 = round(n23/r2, .001);
   tp31 = round(n31/r3, .001);
   tp32 = round(n32/r3, .001);
   tp33 = round(n33/r3, .001);

   * equilibrium proportions – algebraic rearrangement of simultaneous
equations;

   E3 = (tp13-(tp13*tp22)+(tp12*tp23))/(1+tp12-tp22+(tp13*tp32)+tp13-
(tp13*tp22)+(tp12*tp23)-(tp23*tp32)-tp33-(tp12*tp33)+(tp22*tp33));
   E2 = (tp12-(tp12*E3)+(tp32*E3))/(1+tp12-tp22);
   E1 = 1 - E2 - E3;

   * optional – delete columns and cells;

   drop c1 c2 c3 n11 n12 n13 n21 n22 n23 n31 n32 n33 n;
run;

proc means n mean median;
   var e1 e2 e3;
run;
```

```
***********************************************************************
***********************************************************************
**                                                                   **
** This code creates a random sample of all possible 3x3 tables with **
** varying row totals, but with fixed column totals (set in the      **
** first step).                                                      **
**                                                                   **
***********************************************************************
***********************************************************************

* initialize column totals;

data step1;
      c1 = 54;
      c2 = 34;
      c3 = 17;
run;

data step2;
      set step1;

      * designate how many random tables to simulate;

      DO tablenum = 1 TO 1000 BY 1;

      * simulate all cells of the table;

      n13 = ceil((c3-2)*ranuni(0));
      n12 = ceil((c2-2)*ranuni(0));
      n11 = ceil((c1-2)*ranuni(0));
      n23 = ceil((c3-n13-1)*ranuni(0));
      n22 = ceil((c2-n12-1)*ranuni(0));
      n21 = ceil((c1-n11-1)*ranuni(0));
      n31=c1-n11-n21;
      n32=c2-n12-n22;
      n33=c3-n13-n23;
      r1=n11+n12+n13;
      r2=n21+n22+n23;
      r3=n31+n32+n33;
      n=r1+r2+r3;

  OUTPUT;
  END;

run;

* confirm that all variables within appropriate ranges, i.e.;
* n11, n21, and n31 should be between 1 and c1-3;
* n12, n22, and n32 should be between 1 and c2-3;
* n13, n23, and n33 should be between 1 and c3-3;
* r1, r2, and r3 should be between 3 and n-6;
* n should only equal n;

proc means n min max;
      var n11 n12 n13 n21 n22 n23 n31 n32 n33 r1 r2 r3 n;
run;
```