AIDS and Behavior, Vol. 9, No. 4, December 2005 (© 2005)
DOI: 10.1007/s10461-005-9031-0

Response to Commentary

The Fit Between Theory and Data in Respondent-Driven Sampling: Response to Heimer

Jesus Ramirez-Valles,[1] Douglas D. Heckathorn,[2] Raquel Vázquez,[1] Rafael M. Diaz,[3] and Richard T. Campbell[4]

Published online: Dec. 2, 2005

We thank Dr. Robert Heimer for his thoughtful commentary on our paper (Ramirez-Valles *et al*., this issue). As Heimer states, Respondent-Driven Sampling (RDS) is "an innovative and powerful methodology," thus, we welcome this opportunity to further clarify the theoretical basis and applications of RDS.

Like any other probability sampling method, RDS is based on a statistical theory of the sampling process. Such theory provides the means to calculate population estimators of minimal or zero bias, and estimates of the variability of those estimators in the form of confidence intervals or standard errors. This theory also constrains the contexts in which sampling can validly take place, because necessary information required by the statistical theory must be attainable, and data structures must have the form presumed by the statistical theory. Consequently, evaluating an application of a sampling method must assess the fit between the requirements of the sampling method's statistical theory and the data from which research conclusions are derived. To the extent that the assumptions of the statistical theory are violated, confidence on the validity and reliability of estimators is correspondingly reduced. This is the focus of Heimer's comments. He repeatedly expresses concerns that we failed to address what he sees as discrepancies between the data reported in our article and the requirements of the RDS method. These concerns are based on an overly restrictive characterization of the requirements of RDS.

Many of his concerns focus directly or indirectly on the concept of equilibrium, so a brief background on this issue is useful. Prior to the development of RDS, the conventional wisdom regarding snowball-type samples of hidden populations, as expressed in Erickson's (1979) classic article, was that they began with an initial bias due to nonrandom selection of seeds, for if the seeds could be selected randomly, the population would not be hidden. This initial bias, then, was assumed to be compounded in unknown ways as the sample expands wave by wave. The implication was that snowball methods could only serve as convenience samples, because bias could only be judged by what Kalton (1983) termed "subject evaluation."

The original presentation of RDS (Heckathorn, 1997) showed that this conventional wisdom was incorrect. Modeling peer recruitment as a Markov chain showed that bias introduced from the nonrandom choice of seeds was not compounded wave by wave, but instead was progressively eliminated. This occurs because the sample attains a stable composition (i.e., an equilibrium, that is independent of the seeds from which it began). The bias introduced from the necessarily nonrandom selection of seeds, thus, would not forever contaminate the sample; instead, this bias is overcome if sampling attains a sufficient number of waves. This is the first theorem to which Heimer refers.

A second theorem (Heckathorn, 1997) states that as the sample expands wave by wave, it approaches equilibrium at a geometric rate. Therefore, equilibrium is attained quickly, so only a modest

[1] School of Public Health, University of Illinois at Chicago, Chicago, Illinois.

[2] Cornell University, IthacaNew York.

[3] César E. Chávez Institute, College of Ethnic Studies, San Francisco State University, San Francisco.

[4] Department of Sociology, University of Illinois at Chicago, Chicago, Illinois.

1090-7165/05/1200-0409/0 © 2005 Springer Science+Business Media, Inc.

number of waves (e.g., no more than four or five) are required for equilibrium to be approximated. Heimer incorrectly interprets this theorem as requiring that the number of respondents must increase geometrically in each wave. However, this theorem says nothing about the *number* of respondents in each wave, but about the *rate at which equilibrium is approximated*. For example, starting with a U.S.-born seed in Chicago, and using a convergence radius (i.e., error tolerance) of ±2% (see Fig. 2A, Ramirez-Valles *et al.*, in this issue), approximating equilibrium requires four waves. (For computational procedures, see Heckathorn *et al.*, 2002, pp. 66–67.) If the convergence radius is reduced in size by one half, to 1%, one might expect it to require twice as many waves for equilibrium to be approximated. Yet because equilibrium is approached at a geometric rate, only one additional wave, for a total of five, is required. Furthermore, if the convergence radius is reduced 20-fold, to 0.1%, only seven waves (not $4 \times 20 = 80$) are required. Thus, even when defined by a highly stringent standard, approximating equilibrium requires a modest number of waves.

Heimer's mischaracterization of the geometric convergence theorem leads him to attribute excessively stringent requirements to RDS. First, unproductive seeds may present practical problems if they are numerous because they consume scarce resources, but they do not present theoretical problems for RDS. They are dropped out from the analysis. The unit of analysis for RDS is the peer recruitment, which is the unit enumerated in the recruitment matrices from which the RDS population estimator is calculated (e.g., see Table II, Ramirez-Valles *et al.*, in this issue). Unproductive seeds do not appear on such tables because they were not recruited by a peer and did not recruit. Consequently, and contrary to Heimer's argument, they do not increase the number of respondents in the early waves. Furthermore, short recruitment chains do not present a theoretical problem because in RDS analyses, all peer recruitments are treated equally. Whether a recruiter/recruit relationship lies within a small or a large chain, or whether it lies at the beginning or end of a chain, it is treated equivalently. All have the same status, as unitary and equivalent elements of the recruitment matrix, from which transition probabilities are calculated. Therefore, one also need not worry about censoring when recruitment is ended after the sample size has been reached.

There are, however, two tests of the equilibrium requirement that are useful, both of which were reported in our article. First, it is useful to verify that the sample composition approximates equilibrium. As reported in our article, the equilibrium level of HIV seroprevalence was closely approximated in both the Chicago and San Francisco sites. The value of attention to this requirement was illustrated in a RDS study of Middletown and Meriden IDUs (see Heckathorn, 1997). The equilibrium and sample compositions were highly divergent, and examination of the data indicated that this resulted from extremely tenuous network connections between the two towns. The problem was resolved by dividing the dataset into two samples, and then confirming that the close correspondence between equilibrium and the sample composition was restored.

It is also useful to ensure that the number of waves in the sample exceeds that required for equilibrium to be approximated. This helps ensure that the sample will have sufficient sociometric depth (i.e., the number of links from the terminus of the longest recruitment chain to its seed), to guarantee that all members of the target population had a non-zero probability of inclusion in the sample. The number of waves attained at each site, 10 in Chicago and 7 in San Francisco, are sufficient for this purpose.

A second set of Heimer's concerns focus on the representativeness of RDS population estimates. In evaluating these concerns, it is important to clarify an issue that his comments obscure, between bias in the sample composition, and bias in the RDS population estimate. If a sample is self-weighting, the sample composition is unbiased; otherwise weighting is required to produce unbiased estimates. In the case of RDS, the sample is self-weighting when the strength of clustering (i.e., homophily) is equal across groups. This is the third theorem (see Heckathorn, 1997) to which Heimer refers. According to a fourth theorem (Heckathorn, 2002), self-weighting also occurs when network sizes are equivalent across groups. Of course, it is generally not the case that groups have identical levels of clustering or network sizes. As a result, RDS samples are generally not self-weighting. In these cases, the RDS population estimator corrects for biases in the sample (Heckathorn, 2002). Specifically, it has been shown (Salganik and Heckathorn, 2004) that when the assumptions of the statistical theory are satisfied, RDS produces estimates that are asymptotically unbiased, which means that bias is on the order of 1/[sample size], so bias is trivial in samples of meaningful size.

Heimer confuses bias in the sample composition and bias in the RDS estimator when he states that we claim to have obtained an unbiased estimate of HIV seroprevalence only in San Francisco. The sample proved (nearly) self-weighting in San Francisco due to similar homophily (i.e., −.021 and −.026 for HIV-positives and HIV-negatives, respectively) and similar network sizes (i.e., 5.876 and 5.835 for HIV-positives and HIV-negatives, respectively). In Chicago, however, both terms were highly divergent (i.e., homophily of .227 and −.082, and network sizes of 20.517 and 11.287 for HIV-positives and HIV-negatives, respectively), so the estimate of seroprevalence diverged substantially from the sample proportion (i.e., though there were 24.7% HIV-positives in the sample, the estimated seroprevalence is 16.8%).

Our choice of HIV status to illustrate homophily and the calculation of population estimates is a valid one for this population. Our field experience and own research indicate the existence of social grouping along HIV status. HIV-positive individuals come to know each other and create their own networks through the social services and other groups (including the internet) they attend because of their status. HIV status is a discernible attribute among gay male populations, and frequently does not require an explicit or overt (e.g., verbal) disclosure to be made known. Furthermore, this variable is interesting methodologically because sampling was self-weighting in one site, but not in the other. Yet, his observation that the recruitment relationships upon which RDS is based should not be confused with risk networks is correct. Respondents frequently recruit friends or acquaintances that are not part of their sex or drug-associated risk networks. Thus, in RDS studies, as in studies employing other sampling methods, information on risk networks must come directly, not indirectly, via recruitment relationships.

A further critique offered by Heimer focuses on the population coverage. He objects to our Venn diagram, which depicts the set of respondents accessible through venues and the set of subjects accessible through institutions, as the subset of the set of respondents accessible through networks. His objection is that some respondents will prove impossible to recruit. Our intent in offering the diagram was not to suggest that problems of non-response can be ignored, for they are important in any sampling method. The extent of this problem varies greatly depending on factors such as the sensitivity of the information gathered; the extent to which researchers have established a trusting relationship with the community from which respondents are drawn; the level and form of incentives; and potential participants' prior experiences with researchers. All these factors have little connection to the sampling method. However, given the dual incentive system upon which RDS is based, in which material rewards for participating are reinforced by social influence of peer recruiters, RDS may reduce non-response bias. Yet, non-response bias always must remain a concern in any study.

Regardless of the ever-present potential for non-response bias, the essential message of the Venn diagram remains valid. Institutional sampling is limited by its inability to reach those lacking institutional affiliations and venue-based sampling is limited by its inability to reach those who avoid public venues. Network-based sampling methods, such as RDS, can potentially reach both respondents who inhabit these settings. Social networks tend to form within institutions and venues, as well as beyond both settings.

In his discussion of the Venn diagram, Heimer expresses special concerns about the assumption that respondents recruit randomly from their personal networks. He also suggests ethnographic investigation to determine in detail how recruitment decisions are made. Yet, the assumption is *not* that each respondent must recruit randomly from his or her network. Clearly, individual respondents may recruit for many reasons. For example, some may recruit the first eligible acquaintance they see and others may seek out friends who need money. The relevant assumption, however, is that in aggregate recruitment patterns will reflect personal network composition, so respondents "recruit *as though* they were selecting randomly from their personal networks" [emphasis added] (Heckathorn, 2002, p. 19). Ethnographic investigations documenting the bases for individual recruitment decisions would have limited value in verifying aggregate recruitment patterns. A suitable approach, employed in previous RDS studies, is to ask respondent about the composition of the personal networks with respect to visible attributes such as gender and ethnicity. This approach has provided support for the random-recruitment assumption (Heckathorn *et al.*, 2002; Wang *et al.*, 2005).

The network-comparison approach was not well suited to the current study, because the sample was homogeneous with respect to the most visible characteristics (e.g., gender and ethnicity). Other variables on which a comparison could be made, such as national origin, language spoken, or HIV serostatus, are frequently not matters of public knowledge. An alternative means for evaluating the random-recruitment assumption can be proposed that determines whether the data structure fits the assumption. Such a method is termed "reciprocity index."

The reciprocity index is based on two features of the sampling process. First, respondents recruit acquaintances, friends, and those closer than friends, and these are ties that tend to be reciprocal (e.g., my friends and acquaintances tend to consider me to be one of their friends or acquaintances). The implication is that the number of ties linking any pair of groups must be equal in each direction (i.e., for any groups X and Y, the ties from X to Y equals those from Y to X). Second, when the sampling process reaches equilibrium, and when groups recruit with equal effectiveness, ties are randomly selected from the target population's network (Salganik and Heckathorn, 2004). Therefore, if ties are randomly sampled, and the numbers of ties linking groups are equal in both directions, cross-recruitment counts will differ due only to stochastic variation. Consequently, if differences in cross-recruitment counts are small enough to have been produced by chance, support for the random-recruitment assumption is provided. Alternatively, if recruitment is consistently biased toward a particular group, cross-recruitment counts would differ. For example, if some process biased recruitment in favor of HIV-positives, the number of recruitments from HIV-negatives to HIV-positives would exceed that from HIV-positives to HIV-negatives by an amount reflecting the strength of that bias. In contrast, approximately equal cross-counts suggest that there is no consistent bias toward either group.

To evaluate the significance of differences in cross-counts, it is useful to compare them with that which would be expected if recruitments were randomly allocated among categories. In that case, cross-recruitment counts would be no more likely to be equal than would within-category recruitment counts. To see how this can be implemented, consider the analysis of HIV in Chicago, expanded to include the considerable number of respondents for which HIV serostatus is unknown (see Table I). In this system, there are three types of cross-category recruitments, between HIV-positives and HIV-negatives (i.e., 9 and 13), between HIV-positives and HIV-status unknown (i.e., 4 and 1), and between HIV-negatives and HIV-status unknown (i.e., 8 and 7). The mean observed discrepancy for this system, D_o, is $[(13-9)+(4-1)+(8-7)]/3 = 2.67$.

This observed discrepancy, D_o, can be compared to the discrepancy expected if recruitment were independent of network structure. Assume that recruitment is not a random selection from respondent's networks, but instead is strictly random such that each recruiter has an equal probability of recruiting a HIV-positive, a HIV-negative, or a HIV status unknown. The mean number of expected cross recruitments can then be derived by simulation. Consistent with recruitment counts in Table I, assume that there are 18 recruitments by HIV-positives, and that these are randomly allocated among the three categories, as are the 63 recruitments by HIV-negatives and 11 recruitments by respondents of unknown serostatus. The expected discrepancy in cross-category recruitments, D_e, can be calculated by constructing recruitment matrices consistent with these assumptions and calculating the discrepancy for each system. Based on 10,000 simulations, this yields an expected discrepancy of $D_e = 10.87$, a figure larger by a factor of 4 than the observed discrepancy of $D_o = 2.67$. The relationship between these two numbers can be expressed as a proportional reduction in error (PRE) statistic, termed the reciprocity index, R, where expected error is the mean expected discrepancy, D_o, as defined above, and the observed error is the observed discrepancy, D_e. The reduction in error is the difference between the expected and observed discrepancy, $D_e - D_o$, and the proportional reduction in error, the reciprocity index, R, is:

$$R = \frac{D_e - D_o}{D_e} \tag{1}$$

For example, for the Chicago HIV analysis, $R = (10.87-2.67)/\ 10.87 = 0.75$, so 75% of error is reduced. Similarly, for the San Francisco analysis (Table I) the observed discrepancy $D_o = 3$, the expected discrepancy is 9.46, which yields a reciprocity index of $R = 0.68$.[5]

[5] The reciprocity index can also be calculated using not the original recruitment matrix, but the demographically adjusted recruitment matrix (Heckathorn, 2002). This corrects for differences in cross-count recruitments that are due, not to non-random recruitment from personal networks, but from differential recruitment, such as HIV-positive recruiting more than HIV-negatives. However, because demographic adjustment tends to reduce cross-counts differentials, this approach yields a less conservative form of the index.

Table I. Recruitment by HIV Status Among Latino Gay and Bisexual Men and Transgender Persons in Chicago and San Francisco

	Serostatus of recruit			
Serostatus of recruiter	Negative	Positive	Unknown	Total
A. Recruitment by HIV status, Chicago				
Negative	42	13	8	63
Positive	9	5	4	18
Unknown	7	1	3	11
Mean observed cross-category discrepancy in recruitment, D_o		2.67		
Mean expected cross-category discrepancy in recruitment, D_e (based on 10,000 simulations)		10.86		
Reciprocity Index, $R(=(D_e-D_o)/D_e)$		.75		
Significance, p		.0094		
B. Recruitment by HIV status, San Francisco				
Negative	15	15	7	37
Positive	22	20	3	45
Unknown	6	2	1	9
Mean observed cross-category discrepancy in recruitment, D_o		3		
Mean expected cross-category discrepancy in recruitment, D_e (Based on 10,000 simulations)		9.46		
Reciprocity Index, $R(=(D_e-D_o)/D_e)$		.68		
Significance, p		.0277		

The statistical significance of the observed discrepancies can be assessed by determining the frequency with which the expected discrepancy (i.e., the discrepancy in a system consistent with the null hypothesis) is equal to or less than the observed one. This procedure was used to compute the statistical significance of the observed discrepancies, again based on 10,000 simulations, yielding $p = .0094$ for Chicago (i.e., the mean discrepancy was equal to or less than the observed discrepancy, $D_o = 2.67$, in 94 of the 10,000 simulations, for an estimated significance of 94/10,000 = .0094). The corresponding figure for San Francisco is $p = .028$. Thus, the fit between the structure of the recruitment matrices and the theoretic requirements of the RDS statistical theory appears to be good.

A further concern expressed by Heimer is the validity of the comparison of RDS with a simulated time–location sampling (TLS). He suggests that the only valid comparison between methods would be a parallel empirical application of both. This is incorrect, for it is acknowledged in every comprehensive discussion of TLS that this method is suitable only for populations that are geographically concentrated (Amon *et al.*, 2000). Any study that reveals limited geographic concentration also shows the applicability of TLS to be limited. This is a type study for which TLS is ill suited, because it does not provide information on those who shun public settings. The degree of geographic concentration is best be revealed by sampling methods which coverage extends beyond public settings, because they reveal who would be missed by an exclusive focus on these settings. Our finding that many Latino gay and bisexual men and transgender person (GBT) avoid the public venues from which TLS draws its samples are directly germane to evaluations of the applicability of TLS. In addition, the finding that the respondents who would be missed have equivalent levels of HIV risk behavior suggest the importance of supplementing TLS with sampling methods that reach beyond public venues.

Heimer also asserts that a study by Marin *et al.* (2003), in which a random household sample was compared to a snowball sample in San Francisco, is superior to our study. Heimer, however, does not explain why this is the case. Indeed, he acknowledges that Marin and colleagues' study is limited because the snowball sample only included two recruiting waves. Marin and colleagues make a significant contribution to our understanding of network-based sampling, but their snowball sample hardly reflects a recruiting network. It only included two recruitment waves of young (ages 18–29 years) gay men. Moreover, unlike Marin and colleagues, we compared RDS with hypothetical recruitment from eight different venues in Chicago and San Francisco. We agreed with Heimer that further research is needed comparing RDS with other sampling methods, yet to the best of our knowledge, our study provides the best evidence available.

Finally, we welcome Heimer's call for additional research evaluating RDS. As a method that is less than a decade old, and used only by a few dozen researchers, it is in its infancy compared to standard probabilities sampling methods that have been refined over the course of more than half a century. Further extensions of the statistical theory upon which RDS is based may increase the efficiency of indicators. Additional research will also be required to identify optimal applications in varying contexts that take advantage of RDS considerable flexibility in operational procedures. Moreover, extensions of RDS to additional populations provide useful information. Our article reported on the first application of RDS to a GBT population. It thereby answered fundamental questions such as that GBT individuals can be motivated to recruit peers; and that the resulting sample reflects the diversity of the GBT population with respect to HIV risk and status, and demographics such as national original and educational level. The study also confirmed that the sample could be drawn efficiently with respect both to time and resources.

REFERENCES

Amon, J., Brown, T., Hogle, J., MacNeil, J., Magnani, R., Mills, S., Pisani, E., Reble, T., Saidel, T., and Sow, C. K. (2000). *Behavioral Surveillance Surveys: Guidelines for repeated behavioral surveys in populations at risk of HIV*. Arlington, VA: Family Health International.

Erickson, B. H. (1979). Some problems of inference from chain data. *Sociological Methodology*, *10*, 276–302.

Heckathorn, D. D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, *44*, 174–199.

Heckathorn, D. D. (2002). Respondent driven sampling II: Deriving statistically valid population estimates from chain-referral samples of hidden populations. *Social Problems*, *39*, 11–34.

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

Kalton, G. (1983). *Introduction to survey sampling*. Newbury Park, CA: Sage Publications.

Marin, J. L., Wiley, J., and Osmond, D. (2003). Social networks and unobserved heterogeniety in risk for AIDS. *Population Research Policy Review*, *22*, 65–90.

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

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