Sampling for Evaluation: Issues and Strategies for Community-Based HIV Prevention Programs

Ann Aileen O'Connell

Eval Health Prof 2000; 23; 212
DOI: 10.1177/016327870002300206

SAMPLING FOR EVALUATION
Issues and Strategies for Community-Based HIV Prevention Programs

ANN AILEEN O'CONNELL
University of Connecticut

Sampling methods are an important issue in the evaluation of community-based HIV prevention initiatives because it is through responsible sampling procedures that a valid model of the population is produced and reliable estimates of behavior change determined. This article provides an overview on sampling with particular focus on the needs of community-based organizations (CBOs). As these organizations continue to improve their capacity for sampling and program evaluation activities, comparisons across CBOs can become more rigorous, resulting in valuable information collectively regarding the effectiveness of particular HIV prevention initiatives. The author reviews several probability and nonprobability sampling designs; discusses bias, cost, and feasibility factors in design selection; and presents six guidelines designed to encourage community organizations to consider these important sampling issues as they plan their program evaluations.

EVALUATION & THE HEALTH PROFESSIONS, Vol. 23 No. 2, June 2000 212-234

Community-based organizations (CBOs) play a vital role in HIV prevention primarily due to their success in providing individually or community-tailored programs and services and their ability to reach marginalized or disadvantaged groups within a community. With the emerging results of several promising behavioral interventions for HIV prevention (Centers for Disease Control [CDC], 1996a; CDC AIDS Community Demonstration Projects Research Group, 1999; DiClemente & Wingood, 1995; Kalichman, Carey, & Johnson, 1996; Kelly et al., 1992; Lauby, Smith, Stark, Person, & Adams, 1999), many CBOs are in a unique position to begin to replicate these interventions, adapting program components or delivery as necessary to fit their organizational or community needs. Although this is clearly a direction that community-based HIV prevention needs to move in, funding or staffing limitations often interfere with the ability of CBOs to apply sound evaluation principles and methods for assessing the progress or outcomes of their intervention activities. As a result, little is currently known about the successes or challenges that CBOs experience during evaluation of newly developed or replicated HIV prevention programs. However, efforts to understand the process of evaluation for community-based initiatives are evolving, particularly in the area of substance abuse prevention (Cook, Roehl, Oros, & Trudeau, 1994; Goodman & Wandersman, 1994; Hansen & Kaftarian, 1994).

One of the important methodological issues confronting community organizations and behavioral or community researchers in the field of HIV prevention is that of sampling (National Research Council [NRC], 1991; Oakley, Fullerton, & Holland, 1995; Ostrow, Kessler, Stover, & Pequegnat, 1993). A good sampling strategy is necessary for making informed program evaluation decisions and unbiased assessments of change or effectiveness (Burstein & Freeman, 1985; Rossi & Freeman, 1993). Given the necessity for quick and accurate identification of effective programs for diverse populations and communities, the quality of the sample of individuals on which the assessment of *effectiveness* is based becomes of paramount importance. Poorly constructed samples can lead to erroneous decisions about program effectiveness, which in turn may result in less than optimal use of competitive and limited funds. Decisions based on poor samples interfere with efforts to clarify intervention implementation issues and understand the replication process within diverse

communities—particularly regarding community factors and their relationships to intervention success or failure.

The purpose of this article is to provide an overview on sampling that may be particularly relevant to CBOs planning and conducting evaluations of their HIV prevention activities. Several critical concerns regarding sampling and sample design choices are reviewed, and a set of guiding principles on sampling for community organizations is developed. This information should prove useful to agencies trying to improve their capacity for and understanding of evaluation skills and approaches. A review of two promising alternative sampling strategies is also provided. In-depth technical details for any particular sampling strategy may be found in reference texts on sampling (e.g. Henry, 1990, 1998; Kalton, 1983; Kish, 1995). Organizations or agencies that perceive a need for further assistance in developing their sampling schemes may also want to consult a sampling or evaluation specialist from within their community.

COMMUNITY-BASED ORGANIZATINS

Community-based organizations are typically nonprofit, nongovernmental, local, and often grassroots organizations. By their very nature, they are considerably diverse in terms of their philosophical emphases, the programs and services they offer, and their staff size, budget, and location throughout the communities they serve. CBOs generally have close ties to members of their community and often target their interventions or programs to disadvantaged or hard-to-reach groups. Regarding HIV/AIDS, those groups disproportionately affected continue to be minority populations, women, and youth (CDC, 1996b, 1997).

Due to staffing or financial limitations, many CBOs maintain a limited capacity for sampling and evaluation design concerns. Often, CBOs rely on process evaluations for either program monitoring or fiscal accountability needs instead of attempting to conduct outcome or impact evaluations of program effects (O'Connell, Bol, & Langley, 1997; Thomas & Morgan, 1991). Clearly, it becomes unrealistic to expect many of these small local agencies to conduct methodologically rigorous experimental, randomized, treatment/control studies for their interventions of the sort that are expected from well-

funded primary HIV prevention research initiatives. Yet, it remains important to ascertain how well a highly regarded, replicated intervention might be achieving its goals in different communities or cities. Given recent trends of HIV infection, particularly among disadvantaged or minority groups, one clearly cannot ignore the information that could be available collectively from local CBO efforts; hence, everyone benefits in a push to move from a reliance on useful but limited process evaluations to evaluations that include both process and outcome or impact evaluations.

Evaluators need to work toward breaking down some of the methodological barriers that community organizations might experience in their evaluations of programs and activities and offer their assistance to CBOs as they begin to strengthen their competence for sample selection and program evaluation. An earlier article by the author presented evaluation strategies for CBOs (O'Connell et al., 1997); the current article is intended to be a companion piece to that work, focusing on the importance of sampling within the context of evaluation. With improved rigor in both sampling strategies and overall evaluation procedures, comparisons across CBOs should become much more valuable for informing researchers and practitioners as to what works best and under what community conditions.

BEHAVIOR AND HIV PREVENTION

Ultimately, it is one's behavior that impacts on risk of contracting HIV (Curran, 1996; Fishbein & Guinan, 1996). Preventing HIV depends in large measure on peoples' commitment to behavior change and their continued engagement in safe and protective behaviors. Community-based prevention initiatives will therefore be most effective when they contain a behavioral intervention component. These behavioral interventions are designed to influence the behaviors or practices of a particular target population—resulting in healthier behaviors through reduction of risk or adoption of safer activities. Intervention activities have been well-informed and strengthened by behavioral theories, and the methodology for evaluation of these programs or program components needs to keep the pace as well. Typically, several different levels of intervention components are possible: individual, group, street/community outreach, and/or community level

(Academy for Educational Development [AED], 1995). These are outlined briefly in Table 1. For each type of program to be evaluated, sampling poses several unique situations and challenges.

PURPOSE OF SAMPLING

In a nutshell, the purpose of sampling is to create a model of the targeted population (Henry 1990, 1998). Occasionally, it may be possible to survey or interview all participants in a particular program or workshop activity, but this is rarely the norm. Due to time and cost constraints, data collected from a sample instead of the entire target population is used to assess the effectiveness of a program or activity. The target population is, of course, situational: Each of the levels of interventions described in Table 1 could ostensibly have a different target population as its focus.

The level of congruence between members of the sample and members of the target of the intervention directly affects the quality of the information to be derived during the evaluation.

If the sample does not reflect or model the targeted population adequately, then the information resulting from that sample will be biased no matter how well thought out the intervention itself is. More often than not, these biases are hidden, and it is only through thoughtful understanding of the context of the evaluation, the program or intervention being evaluated, and the methods used for sample selection that credible and informative evaluation results will be achieved. Concerns regarding the balance between sample and population characteristics is one of the primary considerations in choosing among alternative sampling procedures because the evaluator or researcher would want to choose the most efficient modeling approach possible given the particular circumstances of the CBO.

Related to the efficiency of modeling the population of interest and concerns about potential sources of bias is the issue of sample size because the precision of estimates regarding program effect will rely in large part on the number of individuals included in the sample. For a CBO beginning the process of developing a sampling strategy, the most salient question is: What quantifiable effect do I wish to observe to say with confidence that the intervention has practical significance? For example, if an intervention is designed to encourage the consistent

TABLE 1
Characteristics of Four Types of Behavioral Interventions

Intervention Type	*Description*
Individual level	Peer- or nonpeer-led programs or services providing one-on-one interventions to support development and maintenance of an individual's safer behavior. These kinds of programs might also have a referral component such as referrals to drug treatment programs or other needed services.
Group level	Peer- or nonpeer-mediated risk reduction training sessions administered to groups such as classrooms, workplace groups, and so on.
Street/community outreach level	Typically focused on individuals of hard-to-reach populations such as the homeless, runaways, or injection drug users (IDUs). Prevention materials or messages are delivered to individuals in locations where they are found throughout the community, often by trained members of the targeted group.
Community level	Designed to reach the greatest number of people in a defined community. The community can be geographic or can be focused on particular subgroups of individuals such as gay males or IDUs. Communities might also be institutions such as schools, bars, or congregations. These interventions are designed to foster individual change and build support for that change through a change in community norms and/or the community environment.

use of condoms among sexually active youth, what proportion reporting consistent condom use would one like to see to claim that the program was successful? Suppose one feels that a definite increase of 20% using condoms consistently, either from baseline or in reference to a comparison group, is a reasonable and important increase. Would the program be regarded as successful if the sample results revealed a change of only 19% reporting consistent condom use? What about 15%? What is the range above and below the stated value of 20% that one would allow oneself before claiming practical significance of the results? Is the use of a percentage even the best measure to represent effectiveness or improvement? These kinds of questions are often difficult to answer but they are extremely important for establishing sample size because the precision with which program effectiveness is measured is completely dependent on this initial range, or *tolerable error*, as well as on how one decides to quantify success or change.

In general, the larger the sample, the more precise any estimates might be; however, in the real world, it is not quite so easy, and many other factors come into play. Both Henry (1990, 1998) and Lipsey (1990) gave excellent reviews of many of these factors and should be

consulted when sampling decisions are being made. Kellow (1998) clarified many of the issues involved in quantifying a treatment or intervention effect because some commonly used measures of program effectiveness can be easily influenced by sample size, and it is important to be aware of their potential to be misleading. Because of this influence, as Kellow (1998) pointed out, "many programs are continued (or terminated) based on the size of the sample used to evaluate the program and not the actual effectiveness of the program" (p. 125). It is always important to remember that the usefulness of a sample depends on how well that sample models the population of interest. A large sample can be just as biased as a small sample—sample size does not guarantee integrity—and attaining statistical significance does not guarantee a meaningful outcome.

The complexity involved in sample size decisions is derived from the need for an efficient sample size given external constraints such as cost, feasibility, and the nature of the program being evaluated. For example, sampling individuals to evaluate a street outreach intervention targeted at injection drug users (IDUs) will clearly involve a different strategy and different cost requirements than the sampling methods employed to evaluate the effectiveness of small groups of youth participating in HIV prevention workshops.

PROBABILITY VERSUS NONPROBABILITY SAMPLES

Evaluators have a choice between probability and nonprobability samples. This choice can affect how well a sample models the intended population. Probability samples are based on principles of probability theory and, relative to nonprobability samples, generally have greater external validity and credibility. In terms of statistical theory, they are preferred over nonprobability samples; however, they are often difficult and sometimes impossible to construct. Common probability sampling techniques include simple random sampling (SRS) and systematic, stratified, cluster, or multistage sampling. Due to the rigor required in the construction of strict probability samples, many CBOs may choose to rely on carefully selected nonprobability or observational samples. Before turning to those methods, elements of probability sampling are reviewed next.

Briefly, simple random sampling implies that every member of the intended population has a known and equal chance of being selected for the sample. Imagine putting all the names for a targeted population in a hat and choosing at random a set of names meeting the desired sample size. The use of this method implies that a list, or sampling frame, of names for the targeted population is available, and for some kinds of interventions, such as workshops or group activities where the attendees are known and could be selected for a sample completely at random, the SRS approach might be feasible. However, for interventions where the targeted population cannot be exhaustively listed or selected from at random, the SRS approach is clearly unrealistic. Consider a street-level intervention aimed at populations such as the homeless, injection drug users, sex workers, teenage runaways, and so on. Although desirable, lists for these kinds of transient and often hard-to-reach populations are extremely difficult to create (if possible at all), and the composition of these groups may change too rapidly for a list to be considered complete at any given point in time.

Other methods of probability sampling, such as systematic, stratified, and cluster sampling, are all variations on SRS and are designed to help model a population with as little bias as possible. Depending on the situation, these modifications of SRS may be easier to apply than SRS. Extensive information on the application of these sampling techniques is widely available (e.g., Henry, 1990, 1998; Kalton, 1983; Kish, 1995), and therefore only a few are mentioned here.

Systematic sampling involves selecting every kth person from a population list, such as every 10th person or every 100th person, until the desired sample size is reached. Stratified sampling allows the evaluator to construct strata (such as age groups or education levels) that might be differentially related to the outcome of interest (such as consistent condom use), and then SRS is employed within each strata. In cluster sampling, the evaluator randomly selects particular clusters, such as apartment blocks in a city, classrooms in a school, or even schools in a district, and every element in the cluster becomes part of the overall sample. If random samples are taken from the randomly chosen clusters, the result is referred to as a two-stage sample. With the addition of other stages of random sampling (e.g., students randomly sampled from classrooms that were randomly sampled from schools that were randomly sampled from districts), the result is a multistage sample.

Probability sampling techniques require that a list or frame of the intended population be available at some stage during the sampling process, which may pose a logistical problem for many smaller community organizations. Nonprobability samples, when properly constructed, may be a better way for CBOs to begin to understand their interventions' progress and effectiveness. Although not considered as rigorous as probability samples, some nonprobability sample designs are clearly stronger than others in terms of how well they model the population of interest. The most common kinds of nonprobability sampling designs are convenience sampling, purposive sampling, and snowball sampling.

Among these approaches, convenience sampling procedures yield the greatest potential for bias because the sample is formed from people who tend to be conveniently available to the evaluator or researcher and may not reflect the general characteristics of the population at all. The evaluator should always take a step back and reflect on just how well the resulting sample models the intended population and consider the impact that subsequent evaluation decisions regarding program success or failure might have on the HIV epidemic in the community and particularly for the targeted group. A reliance on convenience sampling should only occur when other more methodologically sound approaches—including stronger nonprobability designs—have been deemed unfeasible.

Purposive sampling is a step above simple convenience sampling. As Weiss (1998) explained, "purposive sampling is useful in evaluation when the evaluator is interested in data not just on average participants but on participants at the extremes" (p. 164). The evaluator purposively chooses participants for the sample, based on specific a priori criteria. For example, if an intervention for HIV prevention among the homeless is to be evaluated, the evaluator may choose to select a sample from within local homeless shelters as well as from homeless people encountered on the street because the intervention may work differentially among those homeless people willing to seek out shelter and those who cannot or otherwise choose not to go to a shelter. Although it would clearly be more convenient to sample solely from the shelter, purposively including street dwellers allows the CBO to assess the effect of its intervention among different subgroups of the homeless, thus improving the understanding of the intervention.

Snowball sampling can be thought of as a special case of purposive sampling. In snowball sampling, initial members of the sample are asked to identify others for possible inclusion in the sample. This approach is often successful when unobservable networks might exist among members of a targeted population. With a guarantee of confidentiality for their participation, crack users, sex workers, runaways, and other hard-to-reach groups within a community may be able to recruit other group members into the sample.

None of the earlier mentioned sampling designs is free from the potential for bias. Regardless of whether the design involves probability or nonprobability sampling, one of the most important issues is selection bias—that is, the existence of important differences between the sample and the target population. Ellenberg (1994) pointed out that selection bias is never completely avoidable and that researchers cannot accurately assess for selection bias in an observed sample in any quantifiable way. They can, however, review the quality of the sampling process because this will inform them about the potential for important differences between the sample and the population they would hope to generalize to (Ellenberg, 1994). Community organizations are responsible for making careful choices regarding the desired composition of the final sample, which needs to be as defensible as possible regarding its representation of the intended population. Only then will one begin to see evaluation results based on samples that have the credibility desired for determining program success or failure.

Attention to other details of the evaluation, such as the use of randomization in a comparative study, can strengthen the validity of the evaluation results even if nonprobability samples have been used (NRC, 1991). *Randomization* refers to the process of randomly assigning individuals into a treatment group or a control group. Strictly speaking, experiments using this process, such as randomized clinical trials (RCTs), are uniquely suited for obtaining valid and reliable estimates of a treatment or a program's effectiveness. However, participants recruited for RCTs are not randomly selected from some population of interest; enrollment is based on eligibility in terms of similar medical conditions, medical history, medication history, age, and so on.

In very specific situations, randomization is an important and useful strategy for ensuring the internal validity of a study; but for community-based organizations, the use of randomization of participants to two different groups may not be desirable or feasible. Weiss (1998) detailed the common criticisms associated with the use of randomization and outlined some evaluation suggestions for organizations studying the effectiveness of their programs. Through sampling, and when random assignment to treatment and comparison groups is not acceptable, threats to internal validity may be limited by ensuring that people in the treatment group and the comparison group are as alike as possible.

The process of constructing samples to be used for making unbiased estimates of treatment versus comparison group differences across comparable groups of people implies that the samples were recruited using similar approaches. Sampling procedures, whether probability or nonprobability, should be replicable across different communities implementing and evaluating the same intervention as well as across intervention and comparison groups for a comparative evaluation. In constructing a sample, the emphasis should always be on limiting bias; that is, the focus should be on understanding—and accounting for—aspects of the sample that might impact on the evaluation findings.

To avoid bias, sampling designs for evaluation purposes need to include those members of the population targeted by the intervention. Sampling schemes may need to be modified somewhat when the population of interest is hidden or hard to reach within a community, such as crack users, teenage runaways, the homeless, or people involved in the sex trades. *Targeted sampling* (Watters & Biernacki, 1989) and *respondent driven sampling* (Heckathorn, 1997) are two less familiar but extremely useful alternatives to the nonprobability designs discussed earlier. Although the two approaches focus on adaptations of different methodologies, these sampling strategies are appropriate for sampling hard-to-reach populations within a community, particularly for evaluating street/outreach-level or community-level interventions. Both methods emphasize consistency between the sample and the population and attempt to account for and eliminate bias as much as possible. Furthermore, these procedures could be replicated in different communities or different areas of a city for

comparison purposes, thus making these sampling designs ideal for a comparative type evaluation.

TARGETED SAMPLING

Targeted sampling (Watters & Biernacki, 1989) is a "purposeful, systematic method by which controlled lists of specified populations within geographical districts are developed and detailed plans are designed to recruit adequate numbers of cases within each of the targets" (p. 420). Respondents are selected for specific attributes important to the subject under study. Targeted sampling was developed by Watters and Biernacki through their work on HIV and injection drug use in high risk areas of San Francisco. It consists of a combination of methodologies, including ethnographic mapping and other qualitative techniques to locate and identify the target population, and a mixture of techniques from probability and nonprobability sampling to recruit a comprehensive cross section of the targeted group into the sample. Primarily, these sampling procedures include stratified sampling and snowball sampling, both discussed in the sections earlier, and quota sampling, as described by Kalton (1983), where a given number of people with very specific characteristics are selected for the sample.

Because the targeted sample is not a true probability sample, there is some concern regarding bias—mostly resulting from constraints on the time and location of the actual sampling or from the recruitment strategy employed. Watters and Biernacki (1989) reported that their final sample "was dominated by underclass individuals and is not representative of middle-class injecting drug users" (p. 426) because that group would have required different recruitment strategies than the ones employed in the study.

The strength of the targeted approach lies in the fact that although it is a nonprobability sample, a targeted sample is not a convenience sample either. With a strong initial effort at locating and describing the target population through the ethnographic mapping, this approach is a powerful method for constructing samples of hard-to-reach populations. The method is also useful for establishing comparable samples for evaluation designs requiring a comparison group. Targeted sampling procedures employed across communities or geographic areas with similar characteristics (size, availability of treatment programs,

etc.) may yield samples with similar profiles of members of these hard-to-reach populations. From this viewpoint, targeted sampling is an extremely useful methodology for constructing replicable samples, namely, those that maintain a strong congruence to the targeted population and are comparable across other cities or communities as well.

RESPONDENT-DRIVEN SAMPLING

Respondent-driven sampling (Heckathorn, 1997) is another approach useful for locating and sampling hidden or hard-to-reach populations. This method was developed during Heckathorn's work on HIV prevention and injection drug use in eastern Connecticut. It is a sophisticated modification of chain referral or snowball sampling. The respondent-driven sample is composed through systems of social networks, assuming that "those best able to access members of hidden populations are their own peers" (Heckathorn, 1997, p. 178). The methodology for creating a respondent driven sample is systematic and could be replicated in different communities in an attempt to obtain comparable samples.

An important characteristic of respondent driven sampling is that it is based on a system of primary and secondary incentives. First, incentives are provided to participants as a reward for their participation in the study, and secondly, incentives are also provided for participants who recruit others with similar or specific characteristics into the sample (e.g., a female IDU as opposed to any IDU). For confidentiality purposes, recruiters were not asked to name possible network members; instead, they are provided with recruitment coupons and are paid the recruitment incentive when their coupon is used by a new recruit into the study.

As with all nonprobability designs, some bias also exists in this approach. For respondent-driven sampling, the bias typically results from network characteristics as well as concern over coverage of an area. To reduce the tendency for oversampling from recruiters with large personal networks, recruitment quotas were established. In addition, "special incentives can be employed to increase recruitment of subjects with traits associated with a small personal network" (Heckathorn, 1997, p. 179). Heckathorn (1997) included a discussion of bias assessment and results of several sensitivity analyses. His empirical results indicate that when accurately applied, the respondent-driven

sampling approach overcomes many of the problems associated with chain referral samples and can result in a reliable sample that aptly models a hidden or hard-to-reach population.

FACTORS INVOLVED IN SELECTING A SAMPLE DESIGN

One of the challenges in designing effective evaluation research is that there are no hard and fast rules for what kinds of sampling designs might be applicable for every possible situation. Each research endeavor deserves an exploration of the choice of the many different sample designs available and the impact these choices have on the composition of the final sample. Other interrelated factors that need to be considered as a choice among strategies is being made include: cost and feasibility, the safety of data collectors, whether to use a cohort or cross-sectional approach to data collection, the need to limit and reduce the potential for bias, and internal and external validity concerns.

Cost and feasibility. In terms of cost, CBOs should begin with a careful comparison and contrast of the possible sampling designs. Trade-offs can and will exist at many different levels. Be cautioned, however, in that what may appear to be the least expensive sampling approach may turn out to be more costly later in terms of arriving at an accurate assessment of the intervention's effects. Such may be the case if a convenience sample is settled on. Feasibility of a design involves the ability to actually carry out the sampling technique as intended. Examples of cost and feasibility considerations include: staffing and staff expertise, training requirements, development of survey or interview protocols, mode of data collection desired (face-to-face interviews, direct mailing, phone calls, etc.), recruiting interviewers or data collectors (most often these will be from the community or the target group), incentives for participation, required sample size, and the time and effort needed to reach the targeted population.

Safety factors. If particular street corners, housing projects, or gatherings of individuals are deemed too dangerous for data collectors to work with or near, or if the hour of data collection is an issue, some

sample designs simply may not be feasible. For example, sending a data collector out into the field to locate and interview IDUs in high-drug use areas would most likely compromise the safety of the interviewer, even at the best time of day. Heckathorn's (1997) respondent-driven sampling approach to recruiting and interviewing IDUs was clearly more appropriate and safe for their interviewers because many drug users would be physically able to come to an interview area at specified times, and the interviewer's safety would then be far less likely to be compromised. Safety may also be a concern in evaluations of HIV-prevention programs for residents of public housing. Gwiasda, Taluc, and Popkin (1997) evaluated the Chicago Housing Authority's Anti-Drug Initiative and reported on the lessons they learned in that endeavor regarding the safety of their interviewers. The interviewers' safety was their highest priority, "even if it meant turning down an interview" (p. 91). Safety concerns prompted the modification of some of the sampling strategies employed across the 4 years of their study, with the end result being a very structured data collection procedure, uncompromised safety of their interviewers, sound data collected from areas of severe poverty and high crime rates, and a protocol for data collection that could be replicated in other dangerous neighborhoods.

Cohort or cross-sectional samples. For assessment of intervention effectiveness in terms of desired behavior change over time, the most meaningful program evaluations will be comparative in nature, either through use of recruited subjects as their own controls (following a cohort of people over time) or through collection of repeated cross-sectional samples (obtaining data from a fresh cross-section of people for comparison over time). Both kinds of samples have their own benefits and limitations, and both can be used to address the important questions of whether change occurred in the outcomes of interest (Koepsell et al., 1992).

There are many complex issues involved in the optimal choice between cohort or cross-sectional designs for estimating behavior change (Diehr et al., 1995). Many community-based programs want to change risky or unsafe individual behavior as well as influence community and social norms, so cross-sectional samples may be more feasible and appropriate for these kinds of programs. Given cost or other constraints for the evaluation, cross-sectional samples may be

nonprobability samples at each wave of data collection, which implies that an emphasis in sampling will be on consistency of procedures across time or across different communities so that replicable samples are maintained.

An often neglected concern regarding cross-sectional samples involves duplicity, or the repeated inclusion of individuals in the sample either at the same wave or over time. All effort should be made to eliminate the possibility of duplicity—which may be a severe problem when monetary incentives are offered to potential sample subjects for their participation. One approach is to carefully ask enough demographic questions (birthday, age, place of birth, etc.) in an attempt to recognize, match, and eliminate duplicates.

Nonresponse in cross-sectional designs and loss to follow-up in cohort or longitudinal studies also pose serious difficulties in the interpretation of data. Suggestions for understanding and limiting nonresponse and loss to follow-up, such as writing/asking clear questions, offering attractive incentives, emphasizing confidentiality, and so on can be found in Fowler (1993) and Braverman and Slater (1996).

Understanding bias. Although there are many different kinds of bias that could result in any research endeavor, in terms of sample selection and evaluation, the two most important types of bias to be aware of are selection bias and comparison bias. Selection bias occurs when some individuals have a greater chance of being included in a sample over others in that population. As a result, the effect of the intervention could be confounded by differences between those selected for the sample and those individuals who are not selected. It is clearly quite important to strive for the best model of the target population as possible—and eliminate or at least identify and attempt to control for any probable biases in selection. Ellenberg (1994) suggested that the most appropriate way to avoid the insidious effects of bias in selection is through responsible and proper sample design and constantly reflecting on systematic differences that might exist between sample and population.

Comparison bias occurs when there are differences in samples taken from the intervention group and a comparison group who does not receive the intervention. This kind of bias leads to difficulty in assessing the true net effect of the intervention in terms of behavior change. Severe preexisting differences between the two groups to be

compared can be avoided through attention to details in how the two samples were selected. Although statistical adjustments may be made in an attempt to control for some of these preexisting differences, these approaches cannot adjust for every possible way in which two groups might differ (Rosenbaum, 1995; Rosenbaum & Rubin, 1984; Weiss, 1998). Again, the best protection against bias is responsible sampling.

Internal and external validity. No sample will yield estimates that will be entirely free from bias. There are a host of variables and factors that could bias the findings from a study and threaten the internal and external validity of an evaluation. *Internal validity* refers to the ability to claim that an observed change in behavior, for example—whether at the individual, group, or community level—is clearly due to the intervention and not the result of confounding factors or interactions. In a comparative evaluation, this would mean being confident that outcome differences between the intervention and comparison samples are the result of solely the intervention. To address problems of internal validity, particularly for nonprobability samples, one approach is to randomize who receives the intervention and who does not. Randomization tends to balance out effects of extraneous factors evenly across groups and offers protection against threats to internal validity. However, randomization may be a luxury that just isn't an option for many CBOs either for ethical, political, or other reasons important to the community or the stakeholders for the intervention. In such situations, CBOs would need to be very careful in how their sample is chosen to counteract the potential of bias from selection.

Internal validity does not guarantee that a study will have external validity. *External validity* refers to generalizability, that is, that the findings can be faithfully extended or generalized to other members of the target population that were not included in the sample. If the sampling scheme missed an entire collection of members of the population so that the sample did not model the population well, attempts to generalize the findings could be extremely misleading even if the study was determined to have internal validity. For example, if a community-level HIV prevention program for encouraging women to use condoms failed to include Hispanic women in the community sample, it would be erroneous to conclude that findings from that

evaluation, no matter how internally valid, would have been the same for Hispanic women in this underrepresented group.

CONCLUSIONS AND GUIDELINES

Clearly, an important public health goal is that HIV-prevention activities and interventions be successfully identified and replicated. CBOs need to employ responsible sample collection procedures to contribute to a larger understanding of which activities and interventions are most effective for all members of a target population. In essence, there needs to be a push toward a systematic way to compile CBO activities, particularly regarding implementation and evaluation of their prevention efforts.

Several important questions begin to arise. For example: What needs to be known about the characteristics of a CBO, or of a given community, to be able to say that an intervention is likely to be successful? What are the important features of a community that could help one understand the success or weakness of intervention replications?

Given that there is a clear need to understand what works and what does not in diverse communities, particularly regarding replication of promising behavioral interventions from rigorously evaluated large-scale demonstration projects, this is where the need for high quality evaluations of CBO interventions and activities becomes extremely valuable. Collectively, the evaluations of these multiple locally focused CBO programs could be used to take a multipluralist approach (Shadish, 1993) to further evaluating an overall behavioral intervention. When similar effects are observed in multiple communities where programs have been adapted for local needs and priorities, then there is clearly evidence of effective technology transfer and greater knowledge about the impact of specific behavioral interventions can be gained.

Encouraging responsible sampling designs for program evaluations among diverse CBOs would be a first step toward better understanding how community-based behavioral interventions reduce individual and/or community-level risk for HIV. Next, a brief list of sampling guidelines that might be useful for CBOs as they consider their sampling needs is provided. Much remains to be learned about the effectiveness of HIV prevention initiatives, particularly regarding

the impact of community or CBO-level factors on implementation as well as outcome. Hopefully, these guidelines will provide some assistance to CBOs as they continue to work toward improved capacity for program evaluation, both as individual organizations and collectively.

1. Do not neglect the process. Put as much attention up front to sampling who/what you need for program implementation or monitoring purposes as you would to assessing attainment of your objectives or comparative measures of effect. This emphasis on process is particularly important for multisite interventions within single communities and for those interventions containing several different components such as peer outreach and individual counseling. Sites experiencing difficulties in implementing an intervention or specific components of an intervention would clearly not observe the expected intervention effect. Identification of common obstacles to implementation across different sites offers a valuable invitation toward improvement.

2. Assess community level variables as well. To begin to understand the collective efforts of CBOs in HIV prevention, community characteristics or variables should be considered when describing the nature of a community-based sample. The role of these variables would be to provide context for a community program or intervention in an attempt to clarify findings or anticipate possible threats to the internal validity of an evaluation. Recently, researchers have suggested that structural and environmental factors may have an impact on the effectiveness of community-level interventions (CDC AIDS Community Demonstration Projects Research Group, 1999; Tawil, Verster, & O'Reilly, 1995).

Some of these structural or environmental factors are local or national laws governing crime or drug policies, the availability of employment or job training opportunities within communities, and the general social composition of the community. Other potential variables to consider when describing a community include crime rates, economic measures such as the median income or the percentage of schoolchildren receiving free/reduced lunch, and social indicators such as the number of single family dwellings, the percentage of female heads of household, the number of community service agencies, AIDS case rates, the availability of HIV/AIDS initiatives within the community, and so on. These kinds of variables provide important

contextual information regarding the population from which a sample is drawn and, as such, may help clarify the kinds of interventions that are most successful in different kinds of communities.

3. Know your sample. As Henry (1998) explained, representativeness is a subjective judgment. Researchers need to remove subjectivity as much as possible from the samples selected and ensure that it is not elements of the sample, either through systematic inclusion or exclusion of particular groups of people from the desired population, that are influencing decisions for or against the effectiveness of particular interventions. When constructing a sample, it is the researcher's responsibility to ensure that the sample accurately models the intended population. A useful framework for incorporating sampling choices and concerns into the overall research and design process is provided by Henry (1990, 1998). Researchers adopting these guidelines for sample design and selection will be in a strong position to adequately assess the quality of the consistency between the sample and the desired population.

4. Know your comparison group. The purpose of having a comparison group is to be able to infer what would have happened to individuals in the intervention group if they had not received the intervention. Differences between the intervention and comparison group tends to mask or even exaggerate the true intervention effect. Statistical adjustments for preexisting differences may be useful, but the availability of these after-the-fact approaches should never be an excuse not to work toward establishing comparability before the statistical analysis stage. The best comparison groups are those that are comparable before the intervention program begins. In the absence of randomization, consistent and replicable sampling strategies can help to limit the effect of comparison bias.

5. Randomize where possible. Selection bias is the norm, not the exception, and randomization is the best protection against it. Selection bias can occur when the sample is selected nonrandomly so that members of the population may systematically be excluded from the sample. Another form of selection bias can occur if people can elect whether to participate in a program, such as the case with volunteers. Where randomization is simply not possible, evaluators should temper

their findings and attempt to clarify in some way what the influence of selection may be for their evaluation. Some insight into the degree of possible selection bias may be possible through a study or identification of characteristics of nonparticipants or an investigation of variables affecting the choice for program participation.

6. Employ cost comparisons so that you are ensured of collecting the best possible sample given all financial constraints. Be sure you are able to identify and compare the expected precision of your outcome estimates as well as the potential for selection or comparison bias that might occur as a result of different sample designs. Unfortunately, HIV prevention programs are hindered by limitations of federal, state, local, or philanthropic funding. Careful consideration of the costs involved among the different sampling options, while being aware of the trade-offs that may result, should help guide the CBO to a responsible choice of the best sampling strategy for their needs.

REFERENCES

Academy for Educational Development. (1995). *Overview of HIV/AIDS prevention interventions: An approach to examining their effectiveness*. Washington, DC: Author.

Braverman, M. T., & Slater, J. K. (Eds.). (1996). *Advances in survey research*. San Francisco: Jossey-Bass.

Burstein, L.. & Freeman, H. E. (1985). Perspectives on data collection in evaluations. In L. Burstein, H. E. Freeman, & P. H. Rossi (Eds.), *Collecting evaluation data* (pp. 15-32). Beverly Hills, CA: Sage.

Centers for Disease Control. (1996a). Community-level prevention of human immunodeficiency virus infection among high-risk populations: The AIDS Community Demonstration Projects. *Morbidity and Mortality Weekly Reports*, *45*, 1-24.

Centers for Disease Control. (1996b). *HIV/AIDS Surveillance Report*, 9(1), 1-37.

Centers for Disease Control. (1997). Update: Mortality attributable to HIV infection among persons aged 24-44—United States, 1994. *Morbidity and Mortality Weekly Report*, *45*, 121-125.

Centers for Disease Control AIDS Community Demonstration Projects Research Group. (1999). Community-level HIV intervention in 5 cities: Final outcome data from the CDC AIDS Community Demonstration Projects. *American Journal of Public Health*, *89*, 336-345.

Cook, R., Roehl, J., Oros, C., & Trudeau, J. (1994). Conceptual and methodological issues in the evaluation of community-based substance abuse prevention coalitions: Lessons learned from the national evaluation of the Community Partnership Program. *Journal of Community Psychology*, CSAP special issue, 155-169.

Curran, J. W. (1996). Bridging the gap between behavioral science and public health practice in HIV prevention. *Public Health Reports*, *111*(Suppl. 1), 3-4.

DiClemente, R. J., & Wingood, G. M. (1995). A randomized controlled trial of an HIV sexual risk-reduction intervention for young African-American women. *Journal of the American Medical Association*, *274*, 1271-1276.

Diehr, P., Martin, D. C., Koepsell, T., Cheadle, A., Psaty, B. M., & Wagner, E. H. (1995). Optimal survey design for community intervention evaluations: Cohort or cross-sectional? *Journal of Clinical Epidemiology*, *48*, 1461-1472.

Ellenberg, J. H. (1994). Selection bias in observational and experimental studies. *Statistics in Medicine*, *13*, 557-567.

Fishbein, M., & Guinan, M. (1996). Behavioral science and public health: A necessary partnership for HIV prevention. *Public Health Reports*, *111*(Suppl. 1), 5-10.

Fowler, F. J. (1993). *Survey research methods* (2nd ed.). Newbury Park, CA: Sage.

Goodman, R. M., & Wandersman, A. (1994). FORECAST: A formative approach to evaluating community coalitions and community-based initiatives. *Journal of Community Psychology*, CSAP special issue, 6-25.

Gwiasda, V., Taluc, N., & Popkin, S. J. (1997). Data collection in dangerous neighborhoods: Lessons from a survey of public housing residents in Chicago. *Evaluation Review*, *21*(1), 77-93.

Hansen, W. B., & Kaftarian, S. J. (1994). Strategies for comparing multiple-site evaluations under non-equivalent design conditions. *Journal of Community Psychology*, CSAP special issue, 170-187.

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

Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.

Henry, G. T. (1998). Practical sampling. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 101-126). Thousand Oaks, CA: Sage.

Kalichman, S. C., Carey, M. P., & Johnson, B. T. (1996). Prevention of sexually transmitted HIV infection: A meta-analytic review of the behavioral outcome literature. *Annals of Behavioral Medicine*, *18*, 6-15.

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

Kelly, J. A., St. Lawrence, J. S., Stevenson, L. Y., Hauth, A. C., Kalichman, S. C., Diaz Y. E., Brasfield, T. L., Koob, J. J., & Morgan, M. G. (1992). Community HIV/AIDS risk reduction: The effects of endorsements by popular people in three cities. *American Journal of Public Health*, *82*, 1483-1489.

Kellow, J. T. (1998). Beyond statistical significant tests: The importance of using other estimates of treatment effects to interpret evaluation. *American Journal of Evaluation*, *19*, 123-124.

Kish, L. (1995). *Survey sampling*. New York: John Wiley. (Original work published 1965)

Koepsell, T. D., Wagner, E. H., Cheadle, A. C., Patrick, D. L., Martin, D. C., Diehr, P. H., Perrin, E. B., Kristal, A. R., Allan-andrilla, C. H., & Dey, L. J. (1992). Selected methodological issues in evaluating community-based health promotion and disease prevention programs. *Annual Review Public Health*, *13*, 31-57.

Lauby, J. L., Smith, P. J., Stark, M., Person, B., & Adams, J. (1999). *A community-level HIV prevention intervention for inner city women: Results of the women and infants demonstration trial*. Manuscript submitted for publication.

Lipsey, M. (1990). *Design sensitivity: Statistical power for experimental research*. Thousand Oaks, CA: Sage.

National Research Council. (1991). *Evaluating AIDS prevention programs: Expanded edition*. Washington, DC: National Academy Press.

Oakley, A., Fullerton, D., & Holland, J. (1995). Behavioral interventions for HIV/AIDS prevention. *AIDS*, *9*, 479-486.

O'Connell, A. A., Bol, L., & Langley, S. C. (1997). Evaluation issues and strategies for community-based organizations developing women's HIV prevention programs. *Evaluation & the Health Professions, 20*, 428-454.

Ostrow, D. G., Kessler, R. C., Stover, E., & Pequegnat, W. (1993). Design, measurement and analysis issues in AIDS mental health research. In D. G. Ostrow & R. C. Kessler (Eds.), *Methodological issues in AIDS behavioral research* (pp. 1-16). New York: Plenum.

Rosenbaum, P. (1995). *Observational studies*. New York: Springer.

Rosenbaum, P., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.

Rossi, P. H., & Freeman, H. E. (1993). *Evaluation: A systematic approach* (5th ed.). Newbury Park, CA: Sage.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. *New Directions for Program Evaluation, 60*, 13-57.

Tawil, O., Verster, A., & O'Reilly, K. R. (1995). Enabling approaches for HIV/AIDS prevention: Can we modify the environment and minimize the risk? *AIDS, 9*, 1299-1306.

Thomas, S. B., & Morgan, C. H. (1991). Evaluation of community-based AIDS education and risk reduction projects in ethnic and racial minority communities: A survey of projects funded by the U.S. Department of Public Health Service. *Evaluation and Program Planning, 14*, 247-255.

Watters, J. K., & Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems, 36*, 416-430.

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.