

3. Evaluation and its Discontents

In the previous chapter we saw the inherent difficulties in defining the multi-layered phenomenon of HIV transmission through sexual contact. We also saw the unique structures which have been created to carry out the challenging task of prevention. Our particular focus was on the contribution of community-based organizations with their emphasis on structural causes and social change. As could be expected, evaluating prevention efforts in this context is trying, at best. At worst, it is a Sisyphean task in which the immensity of doubt threatens to crush the most fearless of researchers. Here we will examine what unites and what divides those who are trying to measure the outcomes of prevention efforts conducted by community organizations.

3.1 What is Evaluation?

There is no universally accepted definition of the term “evaluation.” Most authors would agree, however, with the following statement proposed by UNAIDS (UNAIDS 2000, p. 8):

Evaluation is a collection of activities designed to determine the value or worth of a specific program, intervention or project.

Although a strong focus on *outcome* evaluation is evident in the literature, there is wide agreement that evaluation activity needs to accompany all phases of project development, from conception to implementation. Typically, the evaluation spectrum is depicted as having three parts (cf. Coyle et al. 1991):

- Formative Evaluation: This includes all activities related to assessing the need of any given target population and the resources available to address that need. The result is a clarification of how HIV is spread, the factors contributing to this spread, and defining potentially effective interventions to contain the epidemic. Reference to previously tried theories and practical approaches is useful.
- Process Evaluation: This includes all activities which accompany the implementation of the intervention. All aspects relevant to how the intervention is delivered by project staff and how it is received by the target population are considered. This can include, for example, consis-

tency among workers in terms of service delivery; satisfaction or acceptance regarding the intervention on the part of the target population; problems encountered when delivering the intervention; etc.

- Outcome Evaluation: This includes all activities which examine what the intervention accomplished. Outcome is always considered in terms of specific goals, either external and general (e.g. a lower incidence of HIV) or specific to a particular setting (e.g. achieving a certain response from a given segment of the target population in a particular place).

There are variations to this schema. For example, UNAIDS (2000) does not include the formative stages as part of the evaluation process. For UNAIDS there are *process*, *outcome* and *impact* evaluations. The process evaluation is identical to that described above. The outcome evaluation focuses on the specific goals of changing behavior within the target group. Impact evaluations measure the effect of a program, or several programs, on the overall epidemiological development (HIV incidence and prevalence).

Regardless of the specific typology, however, it is clear that evaluation is concerned with providing data to support the development and maintenance of programs, projects, and interventions so they can have a lasting effect in preventing the spread of HIV. The three levels named above are clearly interdependent. Only if a need has been described (formative evaluation) can a program be developed (process evaluation) with the goal of producing the necessary change to address that need (outcome evaluation).

3.2 What is the Purpose of Evaluation?

Differences arise among the various authors in regard to two points: (1) the focus of evaluation and (2) the type of information considered to be adequate proof that the program, intervention or project is worthwhile. These differences are both ideological and conceptual in origin. The ideological aspects will be discussed below under 3.3.4 Design and Methods. Here the conceptual aspects can be clarified by considering a description of the various forms of evaluation research with regard to their purpose, as adapted from Jean-Claude Manderscheid (1996). Manderscheid like Van de Ven and Aggleton (1999) argue that the “how” and “what” of evaluation is largely determined by the “why”—that is, by the reason driving the evaluation and, in turn,

by those whose interests are being served. According to these authors, there is no evaluation form which is “right” or “better,” but rather the evaluation needs to fit the questions being asked and the reason for conducting the evaluation in the first place (Figure 1).

Upon looking at Figure 1 we see an overriding interest among all four stakeholders in an adequate and appropriate response to stemming the epidemic. The focus of evaluation activity varies, however, depending on each stakeholder’s particular interests. For example, questions of theory-building and causality are central to researchers, but of no immediate relevance to the group being served. Decision makers want definable problems with practical solutions and measurable outputs; whereas, the community-based organization is interested in how well it is serving the target group. There is also an implicit qualitative difference in the types of information required by the various stakeholders. Researchers emphasize methodological issues related to how data is gathered so as to establish causal links between intervention and outcome. Decision makers look rather for politically acceptable and generally plausible connections between the problem and its solution in order to assure accountability, regardless of formal definitions of causality. The community-based organizations themselves are most concerned with the daily problems encountered in the practice of prevention and thus seek direct input on their work. And finally, the target group is most interested in the degree to which service structures give them what they need in order to confront the reality of the disease in their everyday lives.

As Manderscheid, Van de Ven, and Aggleton argue, all of these interests and their associated information preferences reflect the social and political reality of evaluation research, and thus are equally valid and necessary. At the same time, one needs to recognize the unique advantages and disadvantages of the associated approaches. For example, Van de Ven & Aggleton (1999) see the strength of methodologically rigorous research as defining rules for the development of new programs; the risk, however, is artificiality and a lack of generalizability because of the restrictions imposed on intervention design and the selection of subjects. They argue further that research geared to decision makers has the benefit of setting up communication between funders and programs, while often overvaluing efficiency and undervaluing a program’s stated goals. Finally, the sort of self-study implied by participative re-

search with and for target groups increases the personal responsibility of those involved, but can overburden members of the group and exclude the views of outsiders.

Although there is general agreement in the literature regarding a need for a variety of approaches to evaluation, there are differing opinions as to the relative weight of the type of information produced. For example, there are those who hold the *randomized clinical trial* (RCT) as being the ultimate test of whether an HIV prevention intervention is effective. This conflict will be discussed below.

3.3 Current Themes in the Evaluation of HIV Prevention

In this section the current theory and practice of HIV prevention will be discussed further in terms of important themes in the literature.

3.3.1 Regional Differences: North America vs. Europe

It is important to recognize that the international literature on the evaluation of HIV prevention largely originates from North America, particularly the United States. For example, of the 68 outcome evaluations found in an extensive search of the literature by Oakley et al. (1995), 50 (74%) were conducted in North America (3 UK, 6 Other European, 9 Other). Of the 110 studies on sexual health interventions for young people reviewed by Peersman & Levy (1998), 92 (84%) were from North America (13 Europe, 2 Developing Countries, 1 Other). In preparation for this dissertation, an extensive literature search was conducted of studies in English, German and French, using both electronic data banks and contact to key researchers (Figure 2). The goal was to identify process and outcome evaluations of community-based HIV prevention focusing on the sexual transmission of HIV (see exclusion criteria Figure 3). Of the identified 191 articles 139 (72.8%) were from the United States (52 or 27.2% from other countries⁴). The literature was also searched in the three languages for theoretical and review articles. Of the 58 identified publications, 36 (62.1%) originated from the US. If we exclude the UK (15 articles), Australia (3) and Canada (2) from the latter category, we have only 5 review articles or theoretical treatments from non-English-speaking countries or international organizations.

⁴ 6 each from the UK and Thailand; 4 each from Mexico and Canada; 3 from Kenya; 2 each from Singapore, Nigeria, Indonesia, Australia, Zaire, Malawi, Zimbabwe, and India; 1 each from Ghana, Sweden, Honduras, Bolivia, France, Uganda, Sri Lanka, Greece, Nicaragua, New Zealand, Germany, South Africa, and Columbia

The international discourse on the evaluation of HIV programs is, therefore, influenced for the most part by the experience of American researchers and practitioners.

An obvious explanation for the disproportionate representation of the US in the literature as compared with other industrialized countries would be the larger scope of the American epidemic as compared with Europe. Not only was the US the first country in which the virus was identified, but it continues to exhibit the most serious epidemic among industrialized nations. For example, the estimated annual incidence of HIV in the US is 40,000 (CDC 2001a) as compared to 2,000 in Germany (RKI 2001), which when taking into account the relative differences in population represents six times the magnitude of the German epidemic. And unlike Germany, large-scale sub-epidemics continue to emerge among previously unaffected groups. HIV/AIDS thus constitutes a health problem in the US of a far larger scope than in Europe. One would thus expect a larger investment in related research.

Epidemiological differences do not, however, explain the virtual non-existence of process and outcome evaluations on community interventions in German and French. Here, cultural differences may be at work, as suggested by Maja Heiner (1992) and Jean-Claude Manderscheid (1996). These authors describe a particular emphasis in English-speaking countries on evaluation as an integral part of planning. Both point to a different research tradition in Germany and France which is not driven by planning imperatives, but rather by questions related to the social determinants and social impact of a given problem. That is, evaluation research in both countries operates less in the interest of immediate political utility and is thus less concerned with the results of specific interventions and more with longer term social change. The German social welfare state has traditionally not required specific needs assessments of particular target groups in order to act, but has designed policy based on larger societal trends as they have been documented at the population level over time through research of various forms. The resulting interventions have been primarily at the macro level in terms of legislation regarding such issues as job security, housing availability, etc. The individual social service projects are thus viewed as being supplementary to legislative and structural interventions which seek to address such issues as poverty, class disparities, unemployment, etc. The under-representation of French and German process and outcome studies is likely attributable to this different

tradition in terms of the relationship between social policy and research. This tradition is a reflection of the welfare state principles which have shaped social policy in Europe, and which have had less of an influence in the US. State supports and structural interventions have formed the core of European social policy; whereas, the United States has a unique tradition which emphasizes the non-profit and private sector in addressing social problems (cf. Axinn and Levin 1982, Reinhardt 1996).

In an editorial by Peter Aggleton (1998), this contrast between the US vs. Europe and Australia is formulated more strongly, making reference to other differences which will be discussed later (p. 1):

There was no search [in Europe and Australia], as in the USA, a country whose uncontrolled HIV epidemic should act as a lesson to us all, for 'social magic bullets'—non-existent interventions which, regardless of context, produce desired changes in behavior. There was no demand to build a 'science of prevention' through the evaluation of behavioral interventions. And there was no claims that there exists only one way of knowing what works in HIV prevention—the randomized clinical trial (RCT).

Instead, these countries made headway against the epidemic by developing their programs according to the basic health promotion principle enshrined in the Ottawa Charter for Health Promotion. They sought to use social science to illuminate contemporary responses to the epidemic; raise questions about the ways in which beliefs of the seriousness of HIV and AIDS varied and changed; and chart changes in behaviors among groups of communities at special risk.

A primary challenge for researchers and practitioners in Germany is thus to examine critically the North American discourse, transferring ideas and experiences into the German context in which such realities as universal healthcare, the social welfare state (*Sozialstaat*), and a structural understanding of social problems have formed the basis for social and health policy.⁵

⁵ Although the social welfare state is characteristic of Europe as whole, it is important to recognize that some countries such as the Netherlands have a tradition of evaluation in the social and health care system which operates within these structures. Thus, the argument of this section should not be understood to mean there is no intervention research in the North American sense within Europe; however, the emphasis in the European context is on social causes and state-funded interventions.

3.3.2 Quality of Reported Studies, Publishing Bias, and Gray Literature

Not all evaluations which are conducted are published, and not all which are published are of comparable quality.

Kay Dickersin (1997) summarizes the work of several researchers in public health and medicine who have attempted to ascertain the scope of the problem of unpublished studies. A range of 7-51% of studies which were commissioned by funding bodies and/or described in abstract form were never published. In addition, Dickersin cites the considerable evidence for the tendency of journals to publish studies which show that an intervention works, which means that studies showing failure or a neutral outcome are under-represented.

There is no such review of the literature to determine specifically to what extent evaluation research on community-based HIV prevention is never published. Two observations suggest, however, that the rate of non-reporting is high. This would mean that the vast majority of evaluation activities conducted on community-based HIV programs is never disseminated beyond an immediate circle of stakeholders. The first observation is the tendency for evaluation research to be either commissioned by a funding body (for the purposes of monitoring or decision-making) or to be organized by a community-based organization (as a form of self-assessment). Evaluation research is thus often conducted primarily to serve specific organizational purposes and not to contribute to a larger body of knowledge. Second, the International AIDS Conference has, since its conception, featured the integration of community experience and scientific papers. This has included a large number of reports on the activities of community-based organizations around the world, many of which have been evaluated in some form. The published literature itself, however, hardly reflects the diversity of countries and programs represented at the conferences.

Of the studies which have been published, the majority have to do with interventions which involve the community but which have not necessarily been developed by community-based organizations themselves. Of the 191 research-based articles mentioned above which were identified in the preliminary work for this dissertation, only 34 (17.8%) are actual evaluations of existing activities within a community-based organization. If we only include outcome evaluations (156 studies), the percentage drops to 8.3% (13 articles). The others are interventions which were initiated by a researcher or funding body to be implemented at the community level. In many

cases the intervention is developed and implemented in cooperation with community members; however, the intervention design and protocol is under the direction of a research team. Two articles were found which do not focus on the effects of specific interventions, but rather on the practice and impact of community-based organizations in a particular geographical area using KAB surveys in Zimbabwe (Mercer et al. 1996) and a combination of methods in New York State (Maslanka et al. 1995).

Another problem in the published evaluation research is the uneven quality of the reporting which has been observed by several reviewers. Irrespective of the question of which design is most appropriate for determining program success (to be discussed below) there is often a lack of fundamental information about what was actually done and how it was assessed. For example, the following are not clearly described: the aims of an intervention or program; exactly what was done in terms of intervention; the qualities of the target population; the theory underlying the intervention; and the extent to which the target population was included in planning and implementation (cf. Stanton et al. 1996; Oakley et al. 1995).

The result of these problems—unpublished studies, publishing bias, and uneven quality in reporting—is that we can not be sure exactly what the published literature represents, which makes comparing studies a difficult task. Clearly, the results of the vast majority of prevention activities being carried out in the world at the community-level have never been reported in the literature.

3.3.3 Selecting Outcome Measures

Selecting outcome measures is an extremely difficult task when it comes to measuring the success of HIV prevention. Indisputably, the general goal of prevention is to decrease (or at least stabilize) incidence so as to decrease (or stabilize) prevalence over the longer term.

Incidence is, however, a poor measure of the development of an epidemic when the rate of infection in a population is low and/or the proportion of those being tested is small and not representative of the target group. Given a quantitatively adequate number of people being tested who are representative of the target population, and given a consistent rate of testing, prevalence can be useful in tracking an epidemic over time. However, due to the extended incubation of HIV and other delays in becoming tested, incidence and prevalence need to be interpreted with caution:

Both measures reflect infections taking place at an often unknown earlier point in time, thus not being able to reflect immediately the effects of particular prevention programs. An additional problem with incidence and prevalence measures is that they need to be interpreted within the diffusion pattern of the particular epidemic. For example, saturation will be reached in many populations, regardless of outside intervention. In this case incidence and also, in the presence of mortality or migration, prevalence will peak and decline without any changes in the factors influencing risk. Finally, because of the drastic decrease in mortality for people with HIV over the last several years due to combination therapies, there will be a rising prevalence even with a steady or lowered incidence in a country such as Germany where treatment is widely available.

It is for these and related reasons that KAB (knowledge, attitudes, and behavior) measures have been the preferred means of describing the effects of prevention interventions, both in the short and long term. Originating in the Health Belief Model (Fishbein and Middlestadt 1989) the KAB approach seeks to describe at the individual level the risk behavior itself and the primary psychological factors with which it is associated. The vast majority of evaluation studies thus include one or more psychological or behavioral measures as the primary outcome indicator. Common outcome variables include knowledge of the means of transmission, frequency of condom use, partner selection, behavioral intention, risk perception, negotiating skills, etc. The implicit or explicit goal of interventions is thus changing knowledge, attitudes and behavior which are most associated with a risk for HIV transmission within the target group. The direction of change (that is, how the group should think or what the group should do to protect itself) is determined in advance by the prevention program.

A more advanced but less practiced form of using such measures is the construction of composite risk scores. As O'Leary et al. (1997) observe (p. 11): "Conceivably in response to the intervention, individuals may engage in alternative composite behavioral strategies to reduce risk, and the construction of composite risk scores may be the best approach to the measurement of outcomes." Indeed, we know that individuals respond to the epidemic with risk-reduction strategies that may include several components (Wright 2000c, Paicheler 2000). Thus, by using composite risk scores, the reduction of risk is not isolated to specific, pre-determined responses,

but takes into account a host of possible reactions at the individual level which would decrease the total risk at the collective level of the target group.

The problem with the focus on psychological and behavioral measures at the individual level is that other levels of causal factors promoting the spread of HIV in a population are not considered. As described in Chapter 2, community-level and structural factors also play an important role. Kaye Wellings (1994) thus concludes in her summary of lessons learned from HIV evaluation research in Europe (p. 208): “KAB surveys are not appropriate instruments for the evaluation of changes in the social context, and ideally their findings should be synthesized with those from other studies using other approaches, such as work exposing barriers to preventive action.” Wellings is concerned primarily with measurement at the population-level in each country, and thus recommends the use of condom sales figures, calls to AIDS helplines, and media analysis. In regard to this problem, UNAIDS (2000) also recommends, for example, measures of discrimination against people with AIDS in order to assess the social climate.

This problem of excluding the social context when measuring outcomes is particularly apparent when considering the impact of community-based organizations. As discussed in Chapter 2, it is precisely the social-structural factors which such organizations seek to address. And it is this political role which the WHO Ottawa Charter defines as the unique contribution of community groups. Indeed, there is much support in the evaluation literature for this special role. For example, Jeffrey A. Kelly (1999) writes (p. 300):

Community-level HIV prevention interventions that prove successful and sustainable cannot be imposed on a population. They must grow from—and be owned by—the population segments one hopes to reach. It is now standard practice when developing an HIV prevention intervention to solicit input, recommendations, and advice from members of the community population toward which the intervention will be directed. This is essential. However, it is possible to push our prevention paradigms further and view members of at-risk communities not just as the recipients of an intervention but also as partners in the intervention’s delivery.

Steven E. Hobfoll (1998) also recommends placing the community context in the center of HIV prevention efforts, but comments (p. 142):

If this recommendation seems obvious, I end with the opinion that it is not what is practiced in most AIDS prevention research and it is often not practiced in community psychology research. Rather the issues of ecology and community needs tend to emerge more in the discussion sections than the introduction and hypotheses. Ecology is used to explain findings but not to create methodologies and avenues of exploration.

Thus, in spite of the tremendous support for community-based approaches in the literature, the prevention research is so conducted that these organizations are seen more as transporters of behavioral interventions, rather than as sources for new theory and practice in preventing the spread of HIV.

To illustrate the latter point, two important American studies can be cited which have attempted to illustrate the role of community members in HIV prevention. The first is the Mpowerment Project conducted by the Center for AIDS Prevention Studies at the University of California, San Francisco (Kegeles et al. 1996). In this intervention, HIV prevention activities for young gay men were integrated into a range of social activities in a mid-sized American city with the goal of reducing the amount of risk behavior (unprotected anal intercourse) in this group. The post-intervention behaviors were statistically compared to those in a similar city where the project was not operating. The approach was based on the theory of the diffusion of innovations and sought behavioral change based on communication of norms through peer groups. Several mediating and sociodemographic variables are considered in the outcome analysis, which demonstrated a significant fall in risk behavior in the intervention community.

The second study was conducted by Jeffrey A. Kelly and colleagues (1997) in eight cities, four serving as controls. The intervention, also based on diffusion theory, recruited popular members of the community who serve as opinion leaders in the gay bar scene. These peer leaders were trained to endorse systematically risk reduction behavior among gay men in the bars which they frequented. As above, the goal was re-defining social norms in such a way as to promote less unprotected anal intercourse. Self-reports were corroborated with the number of condoms taken from dispensers pre- and post-intervention. The cities in which the intervention was conducted showed a significant drop in risk behavior.

These two studies are so important in terms of community-level HIV prevention because they demonstrate under experimental conditions (randomized clinical trial) that interventions at this level uniquely cause changes in risk behavior. However, even in these two examples of unusually well-designed and well-reported research, we find two limitations which exist in the literature as a whole: (1) The focus on behavioral outcomes at the individual level and (2) The collaboration with community groups as a vehicle for transporting theory and methods initiated by the researchers. These limitations mean that the structural and political focus of community-based work as discussed in Chapter 2 are not addressed. Also, the autonomy and innovation of the groups themselves is not examined, but rather, the way in which communities can transport messages and norms. Although Kegeles et al. explicitly sought to empower young gay men in the intervention community, no measures of empowerment are reported. It is implied that behavioral change will result if the target group is “mobilized and empowered,” the latter apparently resulting from the men designing and running the intervention themselves.

The continuing problem, therefore, in defining outcome measures for community-based prevention is the lack of fit between the goals and structures of community-based work and the criteria for evaluation research. Which brings us back to the interests of the various stakeholders, as discussed under 3.2.

What sort of outcome measures would reflect the goals of community-based organizations while addressing epidemiologically relevant indicators of changing risk? Keith Tones (1999) suggests adopting outcome indicators which reflect the social change components themselves. This can include, for example, social policy measures or target-group specific characteristics, such as empowerment. Taken together, the indicators should reflect key aspects of “a supportive environment in which the ‘healthy choice is the easy choice.’” Such an approach to evaluation research implies, however, a multi-level analyses of HIV risk as proposed by Van de Ven & Aggleton (1999), as opposed to individualistic psychological models. Here we seen the intricate connection between the selection of outcome indicators, theoretical considerations, and research design. These will be discussed in the next section.

3.3.4 Design and Methods: RCT vs. Triangulation

No researcher would dispute the need for high quality data collected systematically and in such a way as to demonstrate both internal and external validity. However, just what constitutes “valid” is at the core of scientific debate, distinguishing the various schools of thought from one another. The larger theoretical dimensions of this question will be considered in Chapter 4. Here we will examine at the practical level the current debate in the literature regarding designs and methods for conducting evaluation research.

On the one side we have those who promote the *randomized clinical trial* (RCT) or *true experiment* as the “gold standard” for testing the effect of health interventions, regardless of their form. According to this view, not only medications but also, for example, surgical and behavioral interventions should be subjected to this same standard of proof before being recognized as effective. Ann Oakley and colleagues (1995) describe the advantages of the experimental approach in HIV prevention research (p. 480):

In answering questions about what works in terms of affecting health outcomes, RCT provide a remedy to the inferential uncertainties of non-experimental designs by securing an equivalence between the social characteristics of experimental and control groups, distributing unknown factors capable of influencing outcome equally between study groups, and reducing the possibility of research bias.

In reference to the inherent difficulty in determining causal links between interventions and health outcomes, the same research team writes (Fullerton et al. 1995, p. 92; cf. Peterman & Aral 1993):

It is the very complexity and multiplicity of factors influencing health attitudes and behaviors that strengthens the case for properly designed randomized control trials (RCT) with sufficiently large numbers as the only reliable way of establishing the effectiveness of different types of intervention.

But as Keith Tones (2000) comments, the ability of an experimental design to produce convincing evidence is limited when it comes to health promotion precisely because of this complexity (p. 228):

Prior to questioning this view, it is important to state at the outset that the RCT is a peculiarly elegant, powerful and emotionally satisfying tool. It

cannot be reasonably challenged when, under conditions of uncertainty, we wish to know whether intervention “A” has a greater effect than intervention “B”—provided only that the interventions in question are relatively simple and clearly definable. Unfortunately, health promotion interventions differ substantially from the administration of a new drug or a well-defined surgical procedure; as we will note below, they are usually intrinsically complex and multi-factorial.

Tones and others (e.g., Speller et al. 1997; MacNeil & Hogle 1998; IUHPE 1999; Heiner 1992; Manderscheid 1996; Kippax & Van de Ven 1998; Van de Ven & Aggleton 1999) argue for a variety of approaches, without privileging the RCT as being the best source of information on the effectiveness of an intervention. They support a *triangulation* of methods to produce the most reliable and meaningful data for evaluating health promotion activities, including HIV prevention. In addition to experimental designs this can include, for example: quasi-experimental designs, epidemiological surveillance data, qualitative data from the target group regarding interventions, and mathematical models or simulations to estimate impact at the population level. Ideally, researchers, practitioners, and policy makers will have access to several sources of systemically gathered information regarding an intervention in order to form an opinion of its appropriateness.

Irrespective of differing epistemological assumptions, to be discussed in Chapter 4, we find in these two positions different points of view concerning the nature of health promotion and the purpose of evaluation research. These will be summarized presently.

Mechanism of the Intervention: Proponents of the experimental approach assume that there is some discrete core of each intervention which constitutes the “active ingredient” and is thus responsible for the intended change in the target group. It is believed that, through carefully designed experiments, we can over time find out exactly which aspects of an intervention have an effect (for example, what parts of a video presentation or what messages transported during a counseling session). These aspects can then be precisely defined and packaged in appropriate ways to be used with other groups in other settings. Those critical of the RCT remark that the assumption of an “active ingredient” is the result of a faulty analogy between medications and health promotion interventions. Whereas, one may be able to isolate and re-

produce the specific molecules of a substance which have a therapeutic effect; interventions aimed at changing social or behavioral dimensions of a problem are not so easily defined. For example: The word “aspirin” contains little ambiguity, either from a scientific or a lay perspective. However, “information brochure” is another matter. Depending on culture, language, setting, historical period, etc. just what constitutes a useful printed description of a problem can vary widely. Even basic facts can be presented in myriad ways. There is, therefore, no clearly identifiable core to such an intervention.

Generalizability (external validity): It follows from the differing assumptions concerning the mechanism of the intervention that there are different underlying views concerning the generalizability of a particular approach outside of the experimental population. If there is, in fact, an active ingredient to an intervention, then showing a significant effect, for example, in a group of gay men implies that the intervention will likely be effective with other gay men. One can even assume that the intervention may apply to other target groups, as well, depending on whether the groups are deemed to be similar in significant ways. Detractors of the RCT, on the other hand, argue that the experimental findings on health promotion have no clear external validity. Given, for example, that there is a sexual education video which has been shown under experimental conditions to reduce risk-taking behavior in inner city youth at a clinic for sexually transmitted infections in New York, what does that tell us about videos and other populations? We cannot conclude that videos in themselves work; clearly, sexual education videos can differ in countless ways in both quality and content. It is also difficult to conclude that videos “of this type” are effective; what would we mean? How would we recognize a similar video which could also be effective without also testing it experimentally? Another problem is the possibility that context played a role in how the video was received, representing a culmination of other influences outside of the researcher’s purview (relationship between the target group and the organization showing the video; social and historical factors defining the target group’s relationship to HIV/AIDS; etc.), thus limiting applicability to other settings or to similar populations in other cities.

Comparability of Target Groups: Implicit in experimental studies is that the subjects, which have been divided into control and treatment groups, are representative of a larger population. In fact, in the ideal scenario, the subjects would be a ran-

dom subset of this population, so as to eliminate selection bias. Typically, the larger population to which researchers generalize their findings is a group with a common socio-demographic or socio-psychological characteristic, for example: gay-identified men, youth, inner-city women, African-American women, or college students. It is assumed that this characteristic is stable and salient to such a degree, that any unique features of the subjects which are related to their particular locality are not relevant. Thus, for example, an experimentally proven four-session workshop with gay men in San Francisco will likely be effective for gay men in other cities in the US, and perhaps in other industrialized countries, as well. Critics of the experimental approach argue that although target groups do share similarities from place to place, the effectiveness of prevention interventions have much to do with the particular history and social context of any given locality. For example: How homosexuality is lived is dependent on culture of origin, the size of the place in which one lives, the amount of state-sanctioned oppression, etc. Thus, even gay identified men can exhibit varying degrees of self-esteem, community support, sexual negotiating skills, information level, etc. based on differences in the larger social environment. Also, different geographical areas are confronted with very different epidemiological situations; the objective risk of infection in epicenters is dramatically higher than in more rural locations. Finally, outside of epicenters the contact with people who are HIV positive tends to be less as well as the amount of time in which a target group has been living with the risk of infection. Such differences based on historical and social context thus make comparison between target groups from place to place very problematic, which calls into question the degree to which experimental subjects can represent more than their own localities.

Illumination vs. Causal Proof: The goal of randomized clinical trials is to provide definitive evidence for the causal link between an intervention and a specific outcome. Through careful design and study management, subjects are thus exposed to an intervention (or a package of interventions) whose effects are then measured. It can thus be concluded whether or not a particular intervention has a unique impact and, if so, the degree of this impact. Those who question the primacy of the RCT have, however, observed several problems with the attempt to establish causality in this way. Firstly, there is the problem of secular trends. Commonly, a target group is made aware of factors affecting their risk situation from sources other than those providing

viding an intervention. Thus, contamination of experiments is an ongoing problem, ultimately masking potential effects. Secondly, in the everyday practice of prevention, target groups are typically subjected to multiple interventions simultaneously. This is not only another source of contamination, but also reflects the real-world context in which interventions are provided. Thus, experimental evidence that one intervention has a significant effect does not answer the question whether the palette of preventions services offered by a community-based organization is appropriate. In addition, the synergistic (or antagonistic) effects between interventions are not taken into account. Thirdly, the tailoring of interventions which is necessary so that they can be tested within the context of an experimental protocol can often mean reducing prevention work to a very limited number of activities delivered according to strictly standardized methods, all of which bears little resemblance to the daily practice of prevention in the field. For these reasons, the researchers promoting the triangulation of data place their emphasis on the *illumination* of prevention practice as opposed to strict proofs of causality. Concretely, this means an ongoing evaluation of the relationship between the needs of the target group, the process of intervention development and delivery, and the epidemiological situation. (This will be described in more detail in Chapter 6).

Cost and Logistics: Proponents of the RCT argue that experimental designs are the most efficient way to test the effects of interventions, thus justifying the considerable costs and logistical planning which such studies require. Detractors call attention to two problems in this regard. The first is the limited amount of research resources as compared to the dynamic landscape of the myriad interventions being developed. There will never be enough resources to test all interventions experimentally. And even those tested in this way do not necessarily reflect the actual practice of prevention in community-based organizations which continue to adapt approaches based on experience. Secondly, many interventions which are developed by community groups cannot logistically be subjected to experimental designs. For example, programs already in operation which deal with hard-to-reach populations cannot be asked to change their work so that it can be experimentally tested. Rather, researchers need to develop forms of evaluation which capture the everyday realities of prevention practice. If the latter does not happen, then evaluation research will be biased in

its focus on narrowly measurable interventions which will, in turn, be construed as being the only forms of intervention which are effective.

To sum: No one disputes the strength of experimental designs in clarifying causality when the link between an action and an outcome is straightforward. The argument against such designs is that they are grossly inadequate when applied to health promotion interventions which, by their very nature, need to reflect particular social and historical contexts of target groups in specific localities. Because of the multi-faceted nature of HIV prevention at the community level, answering the question “what works” is always a matter of interpretation and judgment, based on theory and the accumulation of evidence. Assumptions need to be made, regardless of approach, as outlined above. Those who promote the triangulation of data from several sources argue that, given such uncertainty, it is better to use information provided by a host of indicators measured in different ways, rather than by relying on one particular research methodology.

3.3.5 Evaluation and Community Development

Even the best and greatest evaluations only minimally affect how decisions get made

Carol H. Weiss (as quoted in Heiner 1992, p. 357)

It is a truism of social science that research findings all too seldom affect decisions concerning policy and practice (cf. King et al. 1998). The field of HIV prevention is no exception. As mentioned above, descriptions of interventions rarely mention theories found in the literature. And the results of evaluation research appear to be seldom disseminated beyond an immediate circle of stakeholders.

Maja Heiner (1992) provides a useful overview of this issue. In summarizing the opinion of experts in the field of prevention research, she begins by identifying four characteristics of evaluations which are most likely to be used by decision-makers:

- **Relevance**—the findings can be used directly by stakeholders to solve a current problem.
- **Credibility**—the trustworthiness of the evaluator has been established by the stakeholder. This has often less to do with the particular evaluation method

employed than with a demonstrated expertise on the part of the evaluator in understanding the concrete problems faced by the stakeholders.

- **Readability**—the evaluation report is short and to the point, avoiding jargon and methodological discussions which are only of interest to other researchers.
- **Timeliness**—the evaluation report is turned in within the agreed upon time-frame.

Heiner goes on to describe that the process of evaluation is part of the complex process of decision-making within organizations and, as such, needs to be viewed as a modest intervention in systems with multiple levels of influence. There is thus no guaranteed way of producing institutional change in the interest of improving services. One can, however, maximize the probability of impact—not only by observing the above four guidelines, but also by assuming a “client and serve-oriented” approach (p. 352) :

[Michael Quinn] Patton characterized his concept as “client and service-oriented” and the opposite model as “academic.” In his opinion “academic” evaluators emphasize research purposes and the traditional standards of methodological rigor. They conduct summative outcome studies and want to contribute mainly to social science theory. “Service-oriented” evaluators concentrate on the needs of their clients. They want to contribute to program improvement using qualitative methods and assisting program decision makers in utilizing their findings.

Of course, defining who one’s client is—that is, whose interests are to be served by an evaluation—is not always easy and can involve ethical considerations, for example, if the interests of an organization and the target population are at odds. To assign methodological rigor and theory-building exclusively to the realm of the “academics” is also an oversimplification. However, the main point is that the evaluator sees him or herself as an active player in the change process, and not simply as a bearer of “facts.” Heiner describes the evaluator as “an orchestrator of negotiation processes” who works closely with those who have contracted him or her to conduct the evaluation so as to produce information which will most likely lead to an improvement in services. This implies a certain epistemological stance (354-355):

This stakeholder and participation-oriented approach stands in the tradition of Stake’s “responsive” evaluation. [. . .] In Germany this paradigm is better

known (with different shades of meaning) as “hermeneutic” or “interpretive” and has also been gaining ground in evaluation research lately. It would not sufficiently be earmarked by calling it “qualitative” as opposed to “quantitative,” although it does rely rather on qualitative methods. It is characterized by different approaches to social reality and to the role of social scientists. This leads to a strong emphasis on the necessity of stakeholder participation, not only to give different interest groups a chance. Participation is also needed to understand the different perceptions of relevant actors and to interpret their behavior.

Heiner mentions the empowering nature of this approach as it enables stakeholders to identify and make the institutional changes necessary to improve the quality of their work. This theme is developed more thoroughly in the work of Mary Secret et al. (1999) who describe the use of *empowerment evaluation* for community-based HIV programs. Citing previous work in the field, they define empowerment as (p. 120) “an increase in the actual power of the client or community so that action can be taken to change and prevent the problems clients face.” And in terms of evaluation (p. 121):

Empowerment evaluation fosters in these agencies [that serve oppressed peoples] the capability of using research findings to improve their service delivery systems and to shape their own programmatic destinies. [. . .] Empowerment evaluation is also “illuminating” and “liberating.” [. . .] Ideally, the process results in stakeholders being better able to document program effectiveness to clients and appropriate policy makers, solve their own management problems, secure their own resources, or promote social change on behalf of the population they serve.

Echoing Heiner, Secret et al. see the evaluator as playing an active role in a process of knowledge transfer. They identify the *teaching* of research methods, *facilitation* toward defining evaluation goals and objectives, and *advocacy* for social change as being the three important aspects of this role. Such an approach, they argue, makes the relevance of research findings for health promotion agencies apparent and thus addresses the long-standing problem of the rift between research and practice.

The growing interest in community-based research of this kind was clearly evident at the XIII International AIDS Conference in Durban, South Africa (Wright

2001). However, as Gary Dowsett (2000) of Australia commented, the resources and commitment for such work are often lacking:

The commitment to community-based HIV social research is much stronger in principle than in practice. HIV social research is dominated by quantitative behavioural monitoring studies and experimental intervention trials. Rarely do the research needs of the HIV community sector gain the resources needed or attract sufficient academic interest to produce really useful knowledge to assist practitioners in the community sector to undertake their work more effectively with a more secure knowledge base.

An important step toward strengthening this type of research internationally was taken by the International Network for Community-Based Research on HIV/AIDS (INCBR) which made its international debut at the conference in Durban. This collaboration of scientists and community groups from various countries presented several principles which could well serve as a foundation for future activity in this area (INCBR 2000). Here they have been translated by the author into German to reflect the unique aspects of the situation of self-help and prevention in this country:

Gewinn für die Selbst-Hilfe-Gruppen

Wissenschaftliche Arbeit im Sinne der Selbst-Hilfe sind Untersuchungen, die für und von den Betroffenengruppen und deren Organisationen durchgeführt werden. Dieses Vorgehen hat die Entwicklung der Kapazitäten und Kompetenzen der Gruppen und Organisationen zum Ziel, um organisationspezifische Erkenntnisse zu gewinnen, die dann zur Optimierung der eigenen Arbeit dienen können.

Aufbau von Kapazität

Wissenschaftliche Arbeit im Sinne der Selbst-Hilfe fördert die Entwicklung der Fähigkeit zur systematischen Erhebung und Auswertung von Informationen. Ziel ist der Aufbau einer langfristigen Kompetenz zur Selbstreflexion sowie zur selbstgesteuerten Weiterentwicklung.

Zusammenarbeit

Die Erfahrungen eines Selbst-Hilfe-Projekts sind das Eigentum des Projekts. Aus diesem Grund sollen Selbst-Hilfe-Projekte so früh wie möglich im Forschungsprozess zur Teilnahme an der Konzipierung und Gestaltung von Untersuchungen über ihre Arbeit einbezogen werden. Dadurch kann eine Zusammenarbeit zwischen Wissen-

schaftler/inne/n und Selbst-Hilfe vereinbart werden, die ethische Themen, den Umgang mit Daten sowie die Veröffentlichung der Ergebnisse berücksichtigt.

Gleichstellung

Wissenschaftliche Arbeit im Sinne der Selbst-Hilfe beinhaltet ein besonderes Verhältnis zwischen dem Beobachter/der Beobachterin und den Beobachteten, da die „Probanden“ auch an der Analyse ihres kollektiven Zustandes teilnehmen. Wissenschaftler/innen und Selbst-Hilfe müssen sich deshalb als gleichberechtigte Partner verstehen, damit die Zusammenarbeit so ausgehandelt werden kann, dass Respekt, Würde und Empowerment realisiert werden können.

Einbeziehung

Wissenschaftliche Arbeit im Sinne der Selbst-Hilfe bezieht alle Betroffenen in allen Phasen der Arbeit ein: die Überprüfung von Forschungsanträgen, das Schreiben von Anträgen, das Sammeln von Daten, und die Interpretation sowie die Veröffentlichung der Ergebnisse.

Zugang

Die Teilnahme von Selbst-Hilfe an der Wissenschaft ist eine Art Weiterbildung, die den Aufbau von Kapazitäten für wissenschaftliche Zusammenarbeit fördert. Aus diesem Grund sollen Sprache und Methode aller Untersuchungen allgemein verständlich sein, um den weitesten Zugang für Vertreter/innen der Selbst-Hilfe zu erreichen, da letztendlich die Fähigkeit der Selbst-Hilfe-Projekte zur „Selbstforschung“ die Basis ihrer Weiterentwicklung ist.

Empowerment

Der Forschungsprozess selbst sowie die dadurch gewonnenen Daten und Erkenntnisse sind das Eigentum und die Instrumente der Selbst-Hilfe-Projekte und müssen als solche von Wissenschaftler/innen anerkannt und respektiert werden. Vertreter/innen der Selbst-Hilfe sollen dabei unterstützt werden, ihre eigenen Untersuchungen durchzuführen und deren Ergebnisse zu veröffentlichen und damit anderen Selbst-Hilfe-Projekten nutzbar zu machen.

3.3.6 Best Practice and Evaluation Research

To close this chapter it is useful to discuss briefly the relationship between what has become known as “best practice” and evaluation research. UNAIDS (2002) defines best practice as follows:

Best Practice, the continuous process of learning, feedback, reflection and analysis of what works (or does not work) and why, is the basis from which UNAIDS, its Cosponsors and partners identify, exchange and document important lessons learned. Best Practice has been shared through exchange forums, networks, Best Practice Collection publications, and technical assistance.

As is clear from this definition, what constitutes “best practice” is essentially a consensus of expert opinion, in this case, the opinion of UNAIDS and its member agencies. This consensus is usually based on a combination of evaluation research findings, the opinion of those having implemented or received program services, and/or an assessment of the degree to which certain health promotion principles (for example, participation of the target group) have been applied. Commonly, examples of model programs are presented, the salient aspects of which are summarized under a rubric “lessons learned,” which is usually a summary of the experience gained by practitioners while conducting the program or intervention. Best practice thus goes beyond what evaluation research can offer, providing a cohesive picture of the state-of-the-art in HIV prevention based on several sources of information. In effect, best practice collections set standards by which existing programs can measure their services and present models for new organizations which are seeking to become involved in HIV prevention for the first time.