Improving Local Evaluation Utility

Within Multisite Prevention Intervention Programs

Norman A. Constantine

School and Community Health Research Group, WestEd, Oakland, CA

(now at the Public Health Institute, Berkeley, CA)

Helen H. Cagampang

Institute for Health Policy Studies, University of California at San Francisco

Paper presented at the 106th Annual Conference of the American Psychological Association,

San Francisco, CA, August, 1998

(revised April, 2000)

The authors thank Wendy Constantine, Marycruz Diaz, BethAnn Berliner, and Roy Gabriel for helpful comments. This paper also was presented at the European Evaluation Society International Conference on Evaluation, Rome, Italy, October, 1998. Norm Constantine's e-mail address is nconstantine@phi.org.

Abstract

Multisite prevention programs typically involve a set of programmatically similar, locally-implemented behavior-change interventions. Evaluations of these programs often are characterized by a precarious balance between research rigor and program-wide summative evaluation priorities on the one hand, and adaptability to disparate local community conditions and needs on the other. In this paper we consider the detrimental effects on local staff motivation that occur when this balance is tilted in favor of program-wide summative priorities. We argue that the resulting motivational drift not only hurts the local program implementation, but at the same time reduces the validity of the cross-site summative evaluation. We propose that the most effective means of abating motivational drift is to enrich the evaluation with an infusion of local-site wisdom and perspective, best accomplished through inclusive evaluation strategies. Several principles and specific strategies are presented for enhancing local evaluation utility within a multisite program context.

Introduction

Behaviors that put humans at risk for a variety of health-related problems, including risky sexual activity, alcohol, tobacco, and other drug use, violence, poor nutrition, and inadequate physical activity, are increasingly the focus of school- or community-based prevention intervention programs. Interventions might focus on the individual, the family, the school, or the community, or some combination of these, but all strive to achieve lasting changes in individual risk behaviors as at least part of the ultimate outcomes. It is rare to find prevention programs that have conclusively demonstrated effectiveness in real world applications across multiple diverse sites. Many programs have been developed outside of a research and evaluation context; some have not been evaluated at all. Others have demonstrated efficacy only in an intensively supervised, high fidelity, single-site implementation. Still others have failed to demonstrate efficacy under any conditions, but have been skillfully packaged and promoted.

To address this critical void, multisite demonstration programs are increasingly employed to test the effectiveness of prevention interventions under diverse conditions of real-world program operations. Demonstration programs typically involve locally-implemented prevention interventions, usually within the bounds of common cross-site theoretical frameworks, program specifications, and goals, and often are tied to uniform evaluation requirements. Programs vary in terms of the fidelity to a common intervention program model expected of the participating sites, the rigor of the multisite evaluation design, and the relative allocation of resources between the program-wide and local aspects of the evaluation. One feature that many of these programs share is a generalized cross-site evaluation plan, combined with some support for a local evaluation component at each site. Evaluation efforts are often characterized by a precarious balance between research rigor and program-wide summative evaluation priorities on the one hand, and attention to disparate local community conditions and needs on the other. When local

needs are insufficiently addressed, motivation is likely to suffer among critical local participants, simultaneously attenuating the effectiveness of the local program implementation, the utility of the local evaluation, and ultimately the validity of the cross-site summative evaluation findings.

In this paper, we start by reviewing some of the common challenges encountered and strategies employed in balancing cross-site requirements with local needs and priorities. We then discuss several potential factors contributing to motivational drift, illustrating this discussion with specific examples. Next we discuss the potential of inclusive evaluation approaches in helping to ameliorate this drift. Finally we recommend several specific strategies for remediation.

Challenges in Balancing Local and Multisite Perspectives

In noting the growing emphasis on scientific rigor in demonstration programs funded by the Center for Substance Abuse Prevention (CSAP), Resnicow & Kirby (1997) explore several issues in finding the common ground between evaluation and demonstration. A key issue is the difference between formal research studies, where the intervention is generally designed by scientists, and demonstration programs, where public or private service agency staff and administrators have a more substantial role in designing and implementing the local intervention. With the demonstration program model it is often difficult to know whether "the evaluation drives the intervention or the intervention drives the evaluation" (p. 11). Specific challenges include (a) determining the appropriate level of interpersonal distance, (b) bridging the cultural gap between evaluators, clients, and program staff, (c) bridging the terminology gap between evaluators and program staff, (d) distrust of research within the community served by the program, (e) the multi-component nature of most interventions, (f) the belief among program staff that evaluation is a poor use of resources, and (g) fear among program staff and community members that evaluation overemphasizes the negative. According to Resnicow and Kirby (1997),

several strategies are available to meet these challenges: making evaluation collaborative rather than hierarchical, making evaluation into a clinical tool, increasing communication between program and evaluation staff, including positive behaviors as outcomes, and involving the community in the evaluation.

Several CSAP national cross-site demonstration programs provide cogent illustrations of local versus cross-site balance-related challenges. The CSAP High Risk Youth Demonstration Program (Sambrano, Springer, & Hermann, 1997) is one example. This program to prevent substance abuse among youth involves more than 400 local sites across several funding cohort cycles. Each site employs locally-relevant prevention strategies targeted at youth and families in high-risk settings. During the first cycle of grants, the need for explicit local evaluation requirements was identified and later grantees were required to set aside at least 15 percent of program funds for local evaluations. Across all sites a common theoretical framework (a risk and resilience model) was mandated. In addition, the need for sufficient implementation time to allow the program to stabilize before beginning evaluation was recognized, leading to the allocation of additional years of funding for the later grantees. In spite of these exemplary and unusual efforts to enhance the utility of local evaluation efforts for the second cycle of funding, major challenges remained unmet. First, it was impossible to achieve consensus among the local sites for a set of core outcome measures that all would report. Even a compromise resolution, whereby a list of appropriate measures of commonly applicable program outcomes would be provided as a menu of choices for local selection, failed: “None of the available standardized questionnaires were adopted by more than a handful of programs, and outcome measurement remained highly idiosyncratic across programs”(p. 381). Second, the quality of the data provided varied substantially by site as a result of disparate local efforts and skill, and competing priorities. The results of attempting to apply these data to the cross-site evaluation were “mixed

at best" (p. 381). Based on these disappointing experiences with the first two cohorts of grantees, CSAP developed a mandatory questionnaire for the third cohort of funded sites, and took over responsibility for all aspects of the cross-site data collection. At the same time, additional strategies were employed to build local support for the cross-site evaluation, including convening national workshops, providing regular communication, and returning locally generated data files to the sites.

Another CSAP experience worth considering is the national cross-site evaluation of the Community Partnership Program (Cook, Roehl, Oros, & Trudeau, 1994; Hansen & Kaftarian, 1994; Yin, Kaftarian, Yu, & Jansen, 1997). This program supports substance abuse prevention efforts involving coordinated partnerships of key community organizations. To classify local strategies across 251 diverse community grantees, a partnership taxonomy was developed. This taxonomy was used to categorize sites by composition of the partnership, age of the partnership, and community density. In addition, local evaluation plans were content-analyzed, yielding a total of 60 process, outcome, and impact categories of measurement available for cross-site analysis. Because of the "obvious burdens and constraints" placed on the evaluation by the large number of sites (Cook et al., 1994, p. 159), a two-tiered data collection system was developed. Minimal process data were collected from all sites, while more in-depth process data were collected from a sample of 36 sites, stratified across taxonomy categories. Impact data were collected from 24 sampled sites, and from 24 comparison communities that were selected by the national evaluators. For impact measures, investigators looked for identical measures where possible, such as some of the drug-use self-report items, and viewed small variations in wording or administration methods as potentially enhancing the robustness of any positive findings. For psychological constructs, similar but non-identical measures tapping the same construct were sought, and interpreted through meta-analytic methods. Because of the large number of sites, the

evaluators looked for commonalties across sites' measures and then conducted subset analyses in an exploratory manner. They concluded that:

> The presence of process, outcome, and impact measurement types with high frequency in a sample of partnership evaluation plans supports the idea that cross-site comparison studies are possible. Such studies will be non-experimental in character and will have strict limitations with regard to demonstrating effectiveness of partnerships as a means of changing the prevalence of substance use and abuse in communities. Nonetheless, such comparisons will yield important information that will be useful for the formulation of hypotheses that may subsequently be tested in experimental or quasi experimental field trials. Further, the examination of such data will profoundly influence the development of methodologies that will be crucial for examining community-based intervention issues (Hansen & Kaftarian, 1994, p. 186).

In a similar vein, the challenges of "thinking globally while acting locally" are discussed by Saxe et al. (1997) from the perspective of their evaluation of the Fighting Back program. This is a national multisite demonstration of community substance abuse prevention, funded by the Robert Wood Johnson Foundation. Within this program, exceptional latitude was given to each local site to design and implement an intervention uniquely suited to the needs and resources of their local community. This flexibility presented additional challenges to the evaluators, summed up as an elusive requirement for "…understanding a global concept while measuring its various manifestations over time in very different locations" (p. 364). Most helpful in successfully meeting these challenges was the use of management information systems to aid in the systematic collection of both qualitative and quantitative process and outcome data, together with providing quarterly reports from these data for local-site use in understanding and improving program implementations. The evaluators concluded that these reports were useful

not only at the local-site level, but also enriched the cross-site evaluation by providing for assessment of variations of local program context and implementation to be integrated with the quantitative outcome data.

Similar challenges have been identified in multisite teen pregnancy prevention programs, as illustrated by two statewide programs in California: the California Department of Education's Teen Pregnancy Prevention Grant Program (Constantine & Curry, 1997), and the California Department of Health Services' Community Challenge Grant Program (Brindis, 1999). Together these two statewide programs comprise 149 local school and community program sites involving more than 1,000 collaborating agencies. Major evaluation challenges documented during the initial years of implementation included: (a) delay in the design of the cross-site evaluation until local-site implementation was well underway; (b) discrepancies between legislated outcome requirements and local program goals, objectives, and strategies; (c) variable levels of success across sites in program implementation; (d) lack of program implementation support to the local sites; (e) extensive data collection and reporting requirements for the statewide component; and (f) weak and limited local evaluation design options. These challenges were addressed and partly ameliorated through modular student survey options for outcome data collection and reporting, multiple quasi-experimental evaluation design strategies for each site to chose from, support for site-specific local evaluation components, telephone and onsite technical support, and biannual meetings of site staff and evaluators to build community, share experiences, and provide feedback (Brindis, 1999; Constantine & Curry, 1997).

The randomized clinical trial is sometimes considered the gold standard of research rigor (e.g., Meinert, 1986). One such trial, the Robert Wood Johnson Foundation's Infant Health and Development Program, provided a standardized intervention for low birth weight infants and their families at eight diverse sites around the country (Infant Health and Development

Program,1991). As is typical with clinical trials, program implementation and data collection were highly standardized across sites, and locally designed site-specific evaluation activities were limited. Many of the same challenges described above were encountered, requiring continuous attention together with a substantial investment of resources from the national study office. Strategies employed to help motivate local-site data collection staff included: (a) providing comprehensive written instruction manuals for each of the 68 data-collection forms employed; (b) providing regular, immediate, constructive, and public data quality feedback to the local-site staff; (c) conducting regular on-site visits and all-site training meetings for local-site staff; and (d) providing close communication and regular support between the national study office and the local-site staff (Constantine, Shing, & Pechler, 1987; Constantine, Constantine, & Wrona, 1997; Constantine, Haynes, & Spiker, 1997). The researchers concluded that "Developing a good rapport with the site staff is probably the single most important step that the National Study Office took in producing higher staff morale, which in turn strengthened the staff's commitment to collecting standardized high quality data" (Constantine et al., 1997, p. 393).

The above examples illustrate the commonalties across program areas, evaluation models, and funding mechanisms of challenges and strategies in balancing local and multisite perspectives. Next we will examine these challenges within a formal organizational framework. This will set the stage for a more in-depth discussion of key principles and practical strategies to address these challenges.

The Etiology of Motivational Drift

The cross-site evaluator typically has control over the design of the program-wide evaluation as well as the specifications for local-site data collection, sometimes with input, albeit often of a token nature, from representatives of the local sites. At the same time, the local-site

staff are primarily responsible for program implementation and data collection. The intrinsic motivation and commitment of the local-site data collection staff are a key determinant of data quality (consisting of data integrity, data completeness, and data representativeness) (Constantine et al., 1997; Lutz, 1977). This motivation, while typically strong in the beginning of a program, often fades over time. Through our experiences working with a wide variety of local sites participating in national and statewide programs, we have repeatedly observed a phenomenon that we label motivational drift. This drift is illustrated by the evolving nature of the questions and themes that often surface among the local-site participants over the course of a program:

- **Evaluation month 1 (naiveté, enthusiastic support) :** How can we help make this evaluation as useful and as valid as possible?
- **Evaluation month 6 (practicality, grudging cooperation):** How much of our program resources is this evaluation going to take?
- **Evaluation month 12 (increasing concern, superficial compliance):** What is the least we need to do to meet this part of the evaluation mandate?
- **Evaluation month 18 and beyond (survivalism):** Can we get away with ignoring this part of the evaluation mandate?

Of course these precise questions are not universally experienced according to the above timeline, yet our experiences suggest that they might well illustrate a common phenomenon – motivational drift among local site staff in multisite evaluations. To the extent that this does occur, it is useful to ask what could explain such a drift from enthusiastic support, to grudging cooperation, to superficial compliance, to outright survivalism?

We believe that there are three interrelated causal factors: detachment, irrelevance, and outcome pessimism. Singularly and collectively, such factors can be substantial obstacles to a

useful and valid evaluation. Figure 1 illustrates how they might interact within a mutually reinforcing network to negatively affect the quality, and therefore the validity, of the data. Once these three factors are better understood, however, they can be potentially ameliorated by implementing specific strategies within an inclusive evaluation approach.

→ PLEASE INSERT **FIGURE ONE** HERE

Detachment: Lack of Ownership of the Externally Imposed Multisite Evaluation Design

The cross-site component of the evaluation plan is often imposed upon the local sites with little or no meaningful local participation in its design. Sometimes input is sought; however, it is often at a token level involving an infrequently convened advisory group or other means of representing general concerns of the local sites to the evaluation design authorities. Timing problems often exacerbate this situation in one of two ways. First, an evaluation plan might be designed prior to sufficient development of local implementations, thereby further restricting the opportunity for local input. Conversely, the evaluation plan might not be ready in time for the sites to collect uncontaminated baseline data, thereby further compromising a site's ability to demonstrate positive outcome results.

One special condition that can severely inhibit local ownership and buy-in of the evaluation plan is that of legislated outcome requirements for the program. Some of the mandated outcomes to be measured might have little to do with the specific local-site context and program goals. It is not uncommon to find legislation that is more precisely prescriptive about outcome requirements than about program strategies. For example, the authorizing legislation for the California Teen Pregnancy Prevention Grant Program, SB-1170 (see http://www.leginfo.ca.gov), mandates ten specific outcome requirements to be reported by each site (birthrates, delayed sexual activity,

school attendance, academic performance, dropout rates, pupil grades, birth weights, self-esteem, child protective services referrals, and family functioning). Program strategies, however, are much more loosely prescribed. It is not hard to imagine how this local program flexibility would lead to some local programs with goals and intended results that do not span the full scope of the inflexible legislated outcome requirements, as indeed was documented with the SB-1170 program (Constantine & Curry, 1997).

Irrelevance: Insufficient Local Utility of the Cross-Site Design

Cross-site evaluations typically focus on future-oriented and highly generalized summative outcomes. These are intended to provide the evidence needed to demonstrate that the programs are worthy of political support and continued funding. This type of focus, however, often conflicts with the more immediate priorities and specific needs of the local sites. Local evaluation needs and uses are too often considered as supplemental to the primary cross-site evaluation focus. But to focus on cross-site outcomes at the expense of local program needs is usually an ineffective strategy, because many of the local needs are at the same time critical prerequisites for valid and complete outcome data.

One local-site need that often is fully neglected or insufficiently addressed in cross-site summative evaluations is the need for program implementation support. While there is a growing consensus around the general principles of effective intervention programs (e.g., American Psychological Association, 1993; Kirby, 1997; Lynch & Bonnie, 1994; National Institutes of Health 1997; National Institute on Drug Abuse, 1997), the implementation of these principles in specific environmental and social contexts remains challenging with many unknowns. The assumption that a winning program proposal guarantees expertise in implementing the program effectively has been frequently invalidated by experience. Professional grant writers might be employed, who have little to do with implementing the program. Further, staff turnover and

reassignment sometimes lead to replacement of the program coordinator or other key program staff between the time the proposal is submitted and the time the program is operational. Finally, changing population, local politics, resource availability, and other conditions between proposal submission and program startup often lead the local program implementation staff into unfamiliar territory. The need for responsive and sufficiently intense implementation support and technical assistance is critical both to the local program's success, and to the ability of the cross-site evaluation to document positive results.

As a local program implementation matures, local issues, community needs, service delivery, and organizational stability often become increasingly salient, while the early promise of the program and the evaluation decreases. Demands on management in a maturing program, including the need to sustain funding, become increasingly intense. Local managers often regard evaluation as an area for economizing. Clients, the program recipients whose needs often appear most vivid, might seem more deserving of available resources. Cross-site evaluations rarely contribute to a manger's ability to deal with these immediate challenges. Given all the competing demands, it is not surprising that generalized outcome evaluations completed at the time that the program ends, or later, can be of little relevance to local program managers.

Outcome Pessimism: Barriers to Demonstrating Positive Program Outcomes

As illustrated in Figure 1, outcome pessimism -- the belief (often realistic) of program staff and local evaluators that positive results will not be found by the evaluation -- is both a result of and a contributor to motivational drift. Because of the previously discussed lack of ownership of the evaluation design (detachment), combined with the insufficient local relevance of the design (irrelevance), the local staff commitment to proper data collection is likely to be low. And without quality data collection, of course, the likelihood of demonstrating positive outcomes will be compromised. As site staff sense this reality, they become susceptible to developing outcome

pessimism, thereby further entrenching their motivational drift. As a result, data collection efforts will be further compromised. And the cycle goes on.

Even with committed and motivated data collection, there might be little reason to expect positive evaluation outcome results. Sites are often resistant to randomly assigning program participation between intervention and control groups, and naturally-occurring comparison groups that are sufficiently comparable to the intervention group are difficult if not impossible to find (Constantine & Curry, 1997; Sambrano, et al., 1997). Baseline data are often not truly baseline due to a late start in data specification and collection, and outcome data are often contaminated by the unknown effects of concurrent alternative interventions. Without adequate implementation support and sufficient formative evaluation, the program's implementation might have been severely compromised. Basch et al. (1985) label this problem Type III error, which they define as obtaining negative evaluation findings from a program that has not been adequately implemented. Because the program has not been given a fair test, the negative results are meaningless. The potential for Type III error can be even greater within a multisite program evaluation than for a single-site evaluation. To the extent that local-site staff recognize this, or any of the other potential detractors from positive outcome results, motivational drift is reinforced.

Inclusive Evaluation Approaches

We believe that the most effective strategies for abating motivational drift and improving local evaluation utility are based upon enriching the evaluation with an infusion of local-site perspective. This approach, often referred to as inclusive evaluation, provides an explicit operational framework for the evaluation participants. And while it holds great potential for abating motivational drift, it does so in large part by making the whole evaluation a more comprehensive, more useable, and more valid enterprise. Just as we have shown above how

some of the causes of motivational drift can interact together into a self-perpetuating and debilitating cycle, the benefits of inclusivity can evolve into an increasingly relevant and powerful self-reinforcing venture.

Practitioners of inclusive evaluation work to carefully identify and recruit the appropriate stakeholders as genuine partners in the evaluation effort. Going beyond merely "asking for input," these participants are integrated into all aspects of the evaluation process. Regular and meaningful on-site meetings of an evaluation work group representing various interests and potential uses of the evaluation, both local and cross-site, are key to local utility. Selling this level of commitment to the participants (as well as to the funder) is one of the critical tasks of the inclusive evaluator. Meaningfully involving this group during all levels of discourse is another.

Inclusive approaches are not magical solutions. Many challenges can be expected, even in applying them to single-site evaluations (Greene, 1988; Ryan, Greene, Lincoln, Mathison, & Mertens, 1998), and these challenges are typically magnified when dealing with multisite evaluations having separate local and cross-site components. A primary challenge to be addressed is that busy local-site program staff often do not recognize that the benefits of an inclusive approach will justify the additional investment in their time that genuine participation requires. For this approach to work, the evaluator must hold a strong enough belief in and commitment to its value to sell reluctant site staff on its likely benefits. The evaluator also needs to structure the process so that meaningful intermediate benefits occur frequently and regularly. Not every stakeholder under any and all circumstances can be convinced, nor are all circumstances even appropriate for inclusive approaches. It has been our experience, however, that most stakeholders can be convinced and most circumstances are indeed appropriate. Recognizing those that truly are not (e.g., due to severe internal site conflict or political chaos in the community, extremely limited time or resources available, insurmountable opposition from a

funding source or key administrative staff, etc.) is another skill the inclusive evaluator must develop. In spite of these and other challenges and required investments, we maintain that the promise will be worth the effort in many situations.

A comprehensive presentation of inclusive evaluation philosophy, principles, and strategies has been provided by Patton (1997). Informed by his background in organizational development, Patton's utilization-focused evaluation [UFI] subsumes participatory, collaborative, and empowerment approaches into a tool chest of situation-specific strategies and methods. At the foundation of UFI is the intensive involvement across most evaluation activities of a wide range of interested parties having potential uses for the evaluation results. Within this context, Patton briefly discusses cross-site evaluation issues, offering several lines of evidence to counter the common fallacy that a multisite program implementation must be fixed and unchangeable across sites and time. Instead, Patton argues, a critical aspect of successful multisite programs is mutual adaptation between the local and cross-site methods and goals (1997, p. 204, citing McLaughlin, 1976).

Also having its roots in the field of organizational development, the total quality management [TQM] (Deming, 1986; Kerridge, 1991) approach to evaluation (also know as continuous process improvement) focuses on achieving full involvement of local stakeholders in establishing questions of interest, devising ways to measure them, and finding ways to incorporate the data into program improvement strategies. Every member of the local-site program staff is involved and every stakeholder is expected to contribute to program evaluation and improvement. There are many variants of TQM-based evaluation; all include collaborative problem solving and regular collection and review of process and outcome data, with a primary goal of making program staff more responsive to clients, and more effective in achieving program goals. (e.g., Cagampang, 1996; Cagampang, Barth, Williams, & Neufeld-Wall, 1998;

Huber, 1997). Kaluzny, McLaughlin, and Simpson (1992) describe six dimensions on which TQM differs from the performance and quality assurance standards approach to evaluation: (a) TQM relies on collective or managerial responsibility rather than legal or professional authority; (b) accountability rests not with specialists but with everyone engaged in the TQM process; (c) interdisciplinary teams work together to monitor quality, rather than referring to externally set quality standards; (d) locally-developed rather than external process and performance expectations help staff take responsibility for their work and the team's performance; (e) shorter, rather than longer, planning horizons support continuous improvement; and (f) continuous improvement is the responsibility of all personnel, not just managers. Building on these emphases, multisite evaluators and program managers can make local evaluation activities more useful and more meaningful.

Within the context of multisite community mental health center evaluations, Faulkner and colleagues (1982) propose the participatory planning model [PPM]to satisfy state accountability requirements while at the same time preserving local flexibility and relevance. This model has seven key characteristics: (a) involvement of a broad base of participants in development of statewide monitoring standards, (b) control of standards development and site review by those involved in the process, (c) acceptance of democratic principles in the standards development and monitoring process, (d) integration of external monitoring and internal evaluation systems, (e) back-up by staff and consultants in a manner that is educational and supportive but not controlling, (f) production of standards that are flexible guidelines and objectives rather than specific procedures, and (g) development of strategies to influence the political system to accept the accountability of this type of evaluation process.

Considered together, Patton's utilization-focused evaluation, Deming's total quality management, and Faulkner's participatory planning model yield three common key principles: a

full range of stakeholders must be meaningfully involved in all aspects of the evaluation; an evaluation must be flexible and responsive to local conditions and needs; and data, a valuable evaluation product, must be regularly and meaningfully shared. These principles are fully consistent with the program experience examples provided earlier for meeting the evaluation challenges of balancing local and multisite perspectives. They provided a solid foundation for the development and discussion of specific practical strategies for abating motivational drift and enhancing local utility within multisite program evaluations.

Practical Strategies

Six strategies have been especially useful in our work across a variety of multisite programs, and are discussed below. These strategies are to (a) support implementation, (b) evaluate for program development, (c) facilitate community building, (d) provide regular data feedback, (e) match measured outcomes to program specifics, and (f) identify realistic goals.

Support Implementation

While program implementation is often taken for granted, we have argued above that this can be a dangerous oversight. Especially in multisite programs, implementation is a critical factor to be assessed as part of the evaluation. More than assessment, however, quality program implementation must be sufficiently supported throughout the multisite program.

Local site staff should understand the research literature relevant to the program being implemented. Morrisey et al. (1997) document three reasons for local-site staff not staying current with the research literature -- insufficient time, the overly technical tone of the literature, and limited access to literature sources. Several strategies to support knowledge transfer to the local-site level are recommended, including providing summarized literature reviews to site staff, training in how to conduct literature searches, training on use of on-line prevention databases, and increasing practitioner access to other centralized databases.

Another approach to supporting implementation is to provide a locally-based field consultant to serve each site or geographic clusters of sites. This was employed as part of an evaluation of a multisite HIV prevention education grant program funded by the California Department of Education (Constantine, Curry, & Berry, 1999). In this program, a planned multisite outcome evaluation was largely redirected into program implementation support and developmental evaluation, with a small sample of more stabilized programs selected for more rigorous outcome evaluation. A key component of this program is the use of a cadre of locally-based field consultants to build relationships with up to four local sites each, focusing heavily on program implementation support. Other parts of the support system include facilitated conference calls on topics selected by the sites, electronic communication groups, biannual statewide meetings of all grantees and local evaluation consultants, and ongoing needs assessment activities to identify future services and priorities. Feedback from the sites about this system has been positive, and in the process the local programs are increasing capacity to engage in more formal and rigorous outcome evaluation (Constantine et al., 1999).

Evaluate for Program Development

Formative and/or developmental evaluation are likely to be among the most critical needs of the local sites in multisite program evaluations. Formative evaluation involves helping "programs get ready for summative evaluation by improving program processes and providing feedback about strengths and weaknesses that appear to affect goal attainment" (Patton, 1994, p. 312). Developmental evaluation goes a step further, involving "long-term partnering relationships with clients who are, themselves, engaged in ongoing program development" (p. 312). Multisite prevention intervention programs provide an ideal environment and obvious need for formative and developmental evaluation. But the value of these types of evaluation is not

often well recognized and understood, and frequently neglected or completely omitted (Moskowitz, 1993).

Peterson and colleagues (1994) also discuss the dangers of premature focus on outcome evaluation. Before outcome evaluation, they argue, several earlier stages of program assessment should be completed. First are program and model building activities such as working with program staff to clearly state the problem they are addressing, focusing program goals and objectives, adequately defining the intervention, and developing the program model of expected relationships among intervention components, mediating variables, and outcomes. Next are the preliminary program measurement activities, including accounting of clients, measuring the intervention, and monitoring participation. Only if and when these earlier stages are adequately addressed should a program commit to formal outcome evaluation. When outcome evaluation is attempted prematurely, "poor evaluations of questionable validity often result" (p. 116).

Multisite prevention programs tend to be large investments, and funders are understandably interested in obtaining outcome results to justify the investments. To the extent that funders and policy makers can be educated about the dangers of premature outcome evaluation, and the benefits of formative and developmental evaluation approaches to the local site and ultimately to the cross-site outcome evaluation, the perceived legitimacy and funding support of these types of evaluation will grow. And as funders recognize their value, multisite evaluators will have even more motivation, and means, to attend to evaluation for program development.

Facilitate Community Building

Bringing together cross-site program managers and staff for regular training and review meetings serves not only as a means to communicate knowledge and monitor progress, but also as an effective method of building community across sites. Local-site staff and managers participating as part of a multisite evaluation are in a unique position of mutual dependence, with

overlapping needs as well as wisdom to share. Mistakes made and lessons learned at one site can facilitate the learning process at another. Furthermore, the bonding and community building that typically occur at these meetings can be a strong protective factor against motivational drift. For the two-day meetings convened twice a year in the California HIV Prevention Education Project described above, participant evaluations indicated high levels of perceived usefulness and satisfaction across all meeting components, with peer and consultant networking most highly rated of all (Constantine et al., 1999). At each meeting time was allowed for cross-site break-out sessions, with staff and managers from the 32 participating sites organizing themselves around geographic or programmatic similarities. Other strategies to reinforce community building include facilitated conference calls on topics selected by the sites, and e-mail discussion groups.

Provide Regular Data Feedback

Data collection responsibilities can be burdensome to site staff. Multisite programs always need to be vigilant about mandating only those data elements that are essential to the cross-site study. Even so, a substantial investment in data collection is often required at the local level. To further motivate local-site staff, it is important to provide regular and timely data-based feedback in locally useable form.

There are two primary types of data feedback systems that are relevant in a multisite study. First are systems to provide feedback about the quality of the data collected. Constantine and colleagues (1987, 1997) describe a comprehensive data collection feedback system for a multisite national study. This system incorporated regular and immediate monthly feedback reports to study sites around the country. Reports included listings of missing forms for specific cases, questionable data flags and requests for verification or correction, and graphic displays of month to month percentages of forms received within the appropriate assessment window of time. Each site received its own report as well as a comparative report that showed how a its

performance compared to the other sites. This system provided a means of quickly remediating specific problem data, positive reinforcement to sites for quality and timely data reporting, and an timely and specific indication of need for further training and support at specific sites.

The second type of data feedback systems are those that return locally provided process or outcome data to each site in a format intended to increase understanding of the local program implementation and to improve program functioning. An excellent example is provided by Fawcett et al. (1995), who describe a monthly feedback system based on log forms completed by local-site staff. This system provides community coalitions with cumulative graphs of coalition activities, community changes, services provided, resources generated, and other aspects of program implementation and effect. Because head counts are frequently required program-wide (e.g., of parent participation, community partner participation, activities completed, etc.), the additional investment of developing a management information system and either sending the graphs back to the sites or, better yet, installing the systems on site, can be well worth the extra effort. The data then become immediately useful at the local level, and in turn can motivate those involved locally to sustain interest (Fawcett et al.,1995). While Fawcett's system is largely based on program activities and community-level outcomes, aggregate individual-level outcome data also can be provided, particularly with reference to national or cross-site totals to inform local sites how they compare with other sites.

Match Measured Outcomes to Program Specifics

Large-scale, multisite program evaluations are inherently complex and usually evolve over time. Early identification of potential intermediate and ultimate outcomes in an evolving intervention can challenge even the most skilled evaluator (Kubisch, Weiss, Schorr, & Connell, 1995). Identifying change agents that are appropriate evaluation targets within multiple collaborative relationships can be very difficult; identifying the paths through which change is

likely to occur even more so. These paths can be effectively traced only when local agencies and staff are closely involved with designing and implementing the evaluation, especially in thinking through the intended short and long term outcomes (Brown, 1995).

Local stakeholders can help evaluators balance generalized outcomes with local goals and objectives, and identify the nuances of the complex relationships in their programs that are likely to influence behavior (Brown, 1995). To the extent that local stakeholders are meaningfully involved in the cross-site evaluation design, therefore, this design will be more relevant and powerful. Sufficient attention to the integration of local and program-wide models, goals, and indicators is needed. Local sites can be expected to vary across important program and context dimensions, however, and local involvement itself might not achieve sufficient match between cross-site outcomes and local program specifics.

To capture the wide range of potential outcomes from more than 100 locally-designed pregnancy prevention programs, California's Community Challenge Grant Program employed a modular survey (Brindis, 1999). The instrument was composed of a core survey module administered at all sites together with supplemental modules enabling each participating project to customize the survey to fit its own intervention strategies. While this approach made it possible for agencies to collect information related to their range of locally-designed interventions, it also proved cumbersome and too long for the majority of participants to complete in the available time. In response to feedback from program managers and staff, the instrument was substantially shortened and simplified for the second funding cycle. The consensus of all involved was that a shorter survey would provide more reliable data for the selected outcomes, and that locally-designed instruments could be used to capture more specific and relevant data about each of the local interventions. Similar challenges were encountered with the California Healthy Kids Survey (www.wested.org/hks), a new survey system implemented in

school districts throughout California. Its modular approach, consisting of a core module and six supplemental modules, provided flexibility to customize at the district level according to local needs, however, the challenge of fitting an 85-item core and one or more supplemental modules into a typical 50 minute class period has limited the actual benefits achieved.

A practical yet under-utilized alternative strategy is to design the cross-site evaluation as a meta-analysis, rather than as an analysis of pooled data elements across sites. This approach allows sites to select locally relevant data elements within a set of common program-wide general constructs, instead of requiring identical and potentially misfit data element specifications for all. This also removes the often dubious assumptions required for pooling of data across demographically, culturally, and otherwise variable sites (Constantine & Shing, 1988; Kraemer, 1978). Meta-analysis methodologies have been well developed over the last 20 years and are commonly applied to a sample or population of unconnected studies on a common topic (Durlack & Lipsey, 1991; Fleiss, 1993; Glass, 1976). They have been underused, however, as evaluation design options for multisite studies.

Identify Realistic Program Goals

Fishbein (1997) warns against the perils of unrealistic expectations for a behavioral change intervention. Not only do funders and policy makers sometimes hold unrealistically high expectations, but local-site grant applicants can be the most unrealistic of all. This problem is exacerbated within multisite programs when unrealistic cross-site expectations encourage overly ambitious expectations at the local-site level. Two strategies can be helpful in promoting more realistic program goals and thereby eliminating a primary source of motivation drift. First, funders must become sufficiently aware of what is realistic for different types of interventions. Intensive programs targeting a few individuals are more likely to achieve noticeable change in a shorter time than are more diffuse programs with a wider target group. And sustained behavioral

effects over time are more difficult to generate than immediate knowledge improvements. To the extent that local programs focus on unrealistic goals, their likelihood of achieving even the more modest successes might be diminished. The need for realistic goals also must be communicated in request for applications. Goals and objectives in a proposal should not be valued solely on how ambitious they are, but also on the research or evidence-based justification from the applicant that they are realistic and feasible.

In addition, one of the first tasks of an evaluator should be to work with local-site staff to critically examine and revise as necessary their goals and objectives. This task is readily built into the other early activities of creating a working program model and developing local evaluation questions.

Conclusions

We have proposed a comprehensive explanation for the motivational drift that frequently occurs among local-site staff responsible for cross-site evaluation activities and requirements. Strategies have been described to ameliorate this drift, all within the framework of inclusive evaluation. When evaluators are able to nurture the strong motivation that local-site program staff typically bring to the early stages of the program, not only will local evaluation utility improve but simultaneously the generalized validity of the multisite evaluation outcome findings will be enhanced – via the specification and collection of local data that are more relevant, of better quality and integrity, and more representative and complete. These solutions are not always easy to implement, but with committed participants, sufficient resources, and some creativity, we have found that it is possible to successfully evaluate multisite prevention programs while at the same time increasingly local evaluation utility and enhancing the motivation of local-site staff.

References

American Psychological Association. (1993). Violence and youth: Psychology's response. Washington, D.C.: Author.

Anderson, R. B. (1977). The effectiveness of Follow-Through: What have we learned. Presented at the Annual Meeting of the American Educational Research Association, New York.

Basch, C. E., Sliepcevich, E. M., Gold, R. S., Duncan, F., & Kolbe. L. (1985). Avoiding Type III errors in health education program evaluations: A case study. Health Education Quarterly, 12, 315-331.

Brindis, C. (1999). The Community Challenge Grant Program: Preliminary assessment. San Francisco, CA: Institute for Health Policy Studies, Center for Reproductive Health Research, University of California.

Brown, P. (1995). The role of the evaluator in comprehensive community initiatives. In J. P. Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), New approaches to evaluating community initiatives: Concepts, methods, and contexts. (pp. 201-225). Washington, DC: The Aspen Institute.

Cagampang, H. H. (1996). Focus on performance: A continuous process improvement system for information and education programs funded by the Office of Family Planning. Berkeley, CA:, University of California, School of Social Welfare.

Cagampang, H. H., Barth, R. P., Williams, L., & Neufeld-Wall, S. (1998). Responsible fatherhood and male involvement projects: Continuous program improvement handbook. Berkeley, CA: University of California, School of Social Welfare.

Constantine, N. A. & Curry, K. (1997, November). The local site perspective in statewide teen pregnancy prevention program evaluations. Paper presented at the American Evaluation Association Annual Conference, San Diego, CA.

Constantine, N. A., Curry, K., & Berry, C. (1999, August). The California HIV Prevention Education Technical Assistance Network: Supporting school-linked HIV prevention education. Paper presented at the National Leadership Conference to Strengthen HIV/AIDS Education and Coordinated School Health Programs. Atlanta, GA.

Constantine, N. A., Constantine, W. L., & Wrona, M. R. (1997). Data management systems. In R. T. Gross, D. Spiker, & C. Haynes (Eds.), Helping low birth weight, premature babies: The Infant Health and Development Program (pp. 408-424). Stanford, CA: Stanford University Press.

Constantine, N. A. & Shing, M. S. (1988). PIGCOR: A SAS macro program for computing pooled intra-group correlation coefficients and testing for group homogeneity. Applied Psychological Measurement, 12 (2), 214

Constantine, N. A., Shing, M. S., & Pechler, H. (1987). Receipt control and assessment monitoring systems for a multisite clinical trial. SAS Users Group International 12th Annual Conference Proceedings, 12, 669-673. Cary, NC: SAS Institute.

Constantine, W., Haynes, C., and Spiker, D. (1997). Operational issues in implementing the evaluation. In R. T. Gross, D. Spiker, & C. Haynes (Eds.), Helping low birth weight, premature babies: The Infant Health and Development Program (pp. 381-393). Stanford, CA: Stanford University Press.

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.

Deming, S. E. (1986). Out of the crisis. Cambridge, MA: Massachusetts Institute of Technology.

Durlak, J. A. & Lipsey, M. W. (1991). A practitioner's guide to meta-analysis. American Journal of Community Psychology, 19, 291-332

Faulkner, L. R., Cutler, D. L., & Middleton, M. H. (1982). A participatory planning process for the evaluation of community mental health centers. Community Mental Health Journal, 18, 169-179.

Fawcett, S. B. (1995). Workgroup evaluation handbook: Evaluating and supporting community initiatives for health and development. Lawrence, KS: University of Kansas.

Fishbein, M. (1997). Great expectations, or do we ask too much from community-level interventions? American Journal of Public Health, 86, 1075-1076.

Fleiss, J. L. (1994). The statistical basis of meta-analysis. Statistical Methods in Medical Research, 2, 121-145.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Greene, J. G. (1988). Stakeholder participation and utilization in program evaluation. Evaluation Review,12, 91-116.

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.

Huber, S. C. (1997). Using evaluation as a management tool: The family planning manager. Newton, MA: Management Sciences for Health.

Infant Health and Development Program. (1990). Enhancing the outcomes of low birth weight, premature infants: A multisite, randomized trial. Journal of the American Medical Association, 263(22), 3035-3042.

Kaluzny, A. D., McLaughlin, C. P., and Simpson, K. (1992). Applying total quality management concepts to public health organizations. Public Health Reports, 107, 257-264.

Kerridge, D. (1991). Deming's ideas change the world. American Statistical News, 173, 1-4.

Kirby, D. (1997). No easy answers: Research findings on programs to reduce teen pregnancy. Washington, D.C.: National Campaign to Prevent Teen Pregnancy.

Kraemer, H. C. (1978). Individual and ecological correlation in a general context. Behavioral Science, 23, 67-72.

Kubisch, A. C., Weiss, C. H., Schorr, L. B., & Connell, J. P. (1995). Introduction. In J. P. Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), New approaches to evaluating community initiatives: Concepts, methods, and contexts. (pp. 1-21).Washington, DC: The Aspen Institute.

Lutz, T. D. (1977). Managing people responsible for managing data. Proceedings of the statistical computing section of the American Statistical Association, 69-72. Washington, DC.

Lynch, B. S. & Bonnie, R. J. (Eds.) (1994). Growing up tobacco free: Preventing nicotine addiction in children and youths. Washington, D.C.: National Academy Press.

McLaughlin, M. (1976). Implementation as mutual adaptation. In W. Williams & R. F. Elmore (Eds.), Social program implementation (pp. 167-180). New York: Academic Press.

Meinert, C. L. (1986). Clinical trials: Design, conduct, and analysis. New York: Oxford University Press.

Morrisey, E., Wandersman, A., Seybolt, D., Nation, M., Crusto, C., & Davino, K. (1997). Toward a framework for bridging the gap between science and prevention: A focus on evaluator and practitioner perspectives. Evaluation and Program Planning, 20, 367-377.

Moskowitz, J. (1993). Why reports of outcome evaluations are often biased or uninterpretable. Evaluation and Program Planning, 16,1-9.

National Institutes of Health. (1997). Interventions to prevent HIV risk behaviors. NIH Consensus Statement, 15(2), 1-41.

National Institute on Drug Abuse. (1997). Preventing drug use among children and adolescents. (NIH Publication No. 97-4212). Washington, D.C.: Author.

Patton, M. Q. (1994). Developmental evaluation. Evaluation Practice, 15, 311-20.

Patton, M. Q. (1996). A world larger than formative and summative. Evaluation Practice, 17, 131-144.

Patton, M. Q. (1997). Utilization-focused evaluation: The new century text. Thousand Oaks, CA: Sage.

Patton, M. Q. (1986). Utilization-focused evaluation (2nd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (1978). Utilization- focused evaluation. Thousand Oaks, CA: Sage.

Peterson, J. L., Card, J. J., Eisen, M. B., & Sherman-Williams, B. (1994). Evaluating teenage pregnancy prevention and other social programs: Ten stages of program assessment. Family Planning Perspectives, 26, 116-120, 132.

Resnicow, K. & Kirby, J. (1997). Finding common ground: On the marriage of evaluation and demonstration programs. Journal of Early Adolescence, 17, 10-20.

Ryan, K., Greene, J., Lincoln, Y., Mathison, S., & Mertens, D. (1998). Advantages and challenges of using inclusive evaluation approaches in evaluation practice. American Journal of Evaluation, 19, 101-122.

Sambrano, S, Springer, J. F., & Hermann, J. (1997). Informing the next generation of prevention programs: CSAP's cross-site evaluation of the 1994-95 high-risk youth grantees. Journal of Community Psychology, 25(5), 375-395

Saxe, L., Reber, E., Hallfors, D., Kadushin, C., Jones, D., Rindskopf, D., & Beveridge, A. (1997). Think globally, act locally: Assessing the impact of community-based substance abuse prevention. Evaluation and Program Planning, 20, 357-366.

Yin, R. K. & Kaftarian, S. J. (1997). Challenges of community based program outcome evaluations. Evaluation and Program Planning, 20, 293-297.

Yin, R. K., Kaftarian, S. J., Yu, P., & Jansen, M.A. (1997). Outcomes from CSAP's community partnership program: Findings from the national cross-site evaluation. Evaluation and Program Planning, 20, 345-356.

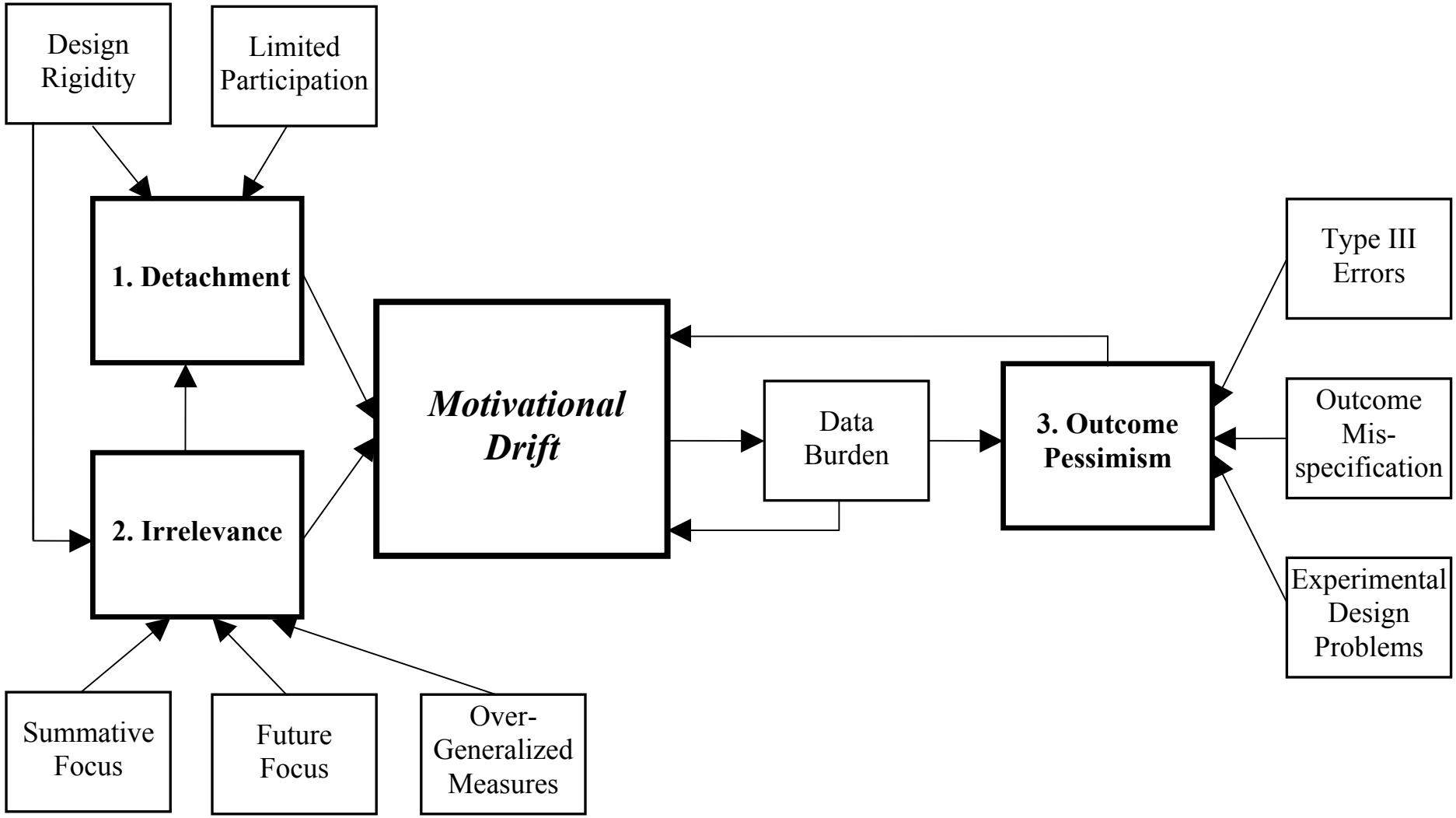

Figure 1. The Etiology of Motivational Drift

From: Constantine, N. A. & Cagampang, H. H. (1998, August). *Improving local evaluation utility within multi-site risk behavior prevention programs.* Paper presented at the 106th Annual Conference of the American Psychological Association, San Francisco, CA.