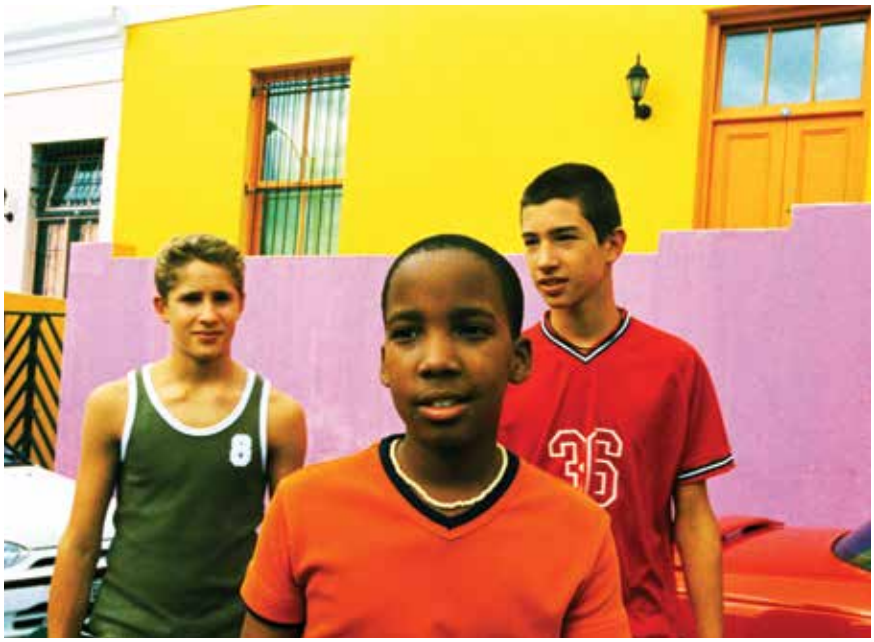
U.S. Department of Justice
Office of Justice Programs
National Institute of Justice

U.S. Department of Health and Human Services
Centers for Disease Control and Prevention
National Center for Injury Prevention and Control

CHANGING COURSE

Preventing Gang Membership



Chapter 11. Program Evaluation: How Do We Know If We Are Preventing Gang Membership?

NCJ 243475

Program Evaluation: How Do We Know If We Are Preventing Gang Membership?

Finn-Aage Esbensen and Kristy N. Matsuda

- **A well-designed program evaluation can determine the effectiveness of a program; the purpose of an evaluation is to determine whether a program (and not some other factor) caused the intended outcomes.**
- **Because the veracity of statements regarding a program's effectiveness depends on the quality of the program evaluation, practitioners and policymakers should understand basic research design concepts (such as *comparison group* and *pretest*) and sampling concepts (such as *representativeness* and *bias*).**
- **An *outcome evaluation* assesses whether a program or strategy achieved the desired outcome or result.**
- **A *process evaluation* assesses the extent to which a program or strategy was implemented as designed; confidence in the success of a program is diminished if implementation is flawed, and a process evaluation can help identify why and where program improvements should be made.**
- **Cost-effectiveness or cost-benefit analyses (conducted during or after a program evaluation) can help policymakers, practitioners and communities determine the most economically efficient gang-joining prevention strategies to implement.**

In Brief

Program evaluations are essential to determining if a prevention program is effective, whether the program is focused on gangs or on other issues. Although evaluations require time and money, they are central to making well-informed decisions about resource allocation and support of prevention programs. This chapter reviews key components of program evaluations and highlights the importance of adhering to these components. Because rigorously designed process and outcome evaluations are the best way to determine program effectiveness — and calls for "evidence-based" programs and policies have become more frequent — the components of the term *rigorous* are also discussed. In addition, the value of cost-effectiveness and cost-benefit analyses in program assessment is introduced. To use the knowledge produced by program evaluations effectively, policymakers and practitioners must be able to interpret the quality of the research.

Most strategies and programs with the potential to prevent kids from joining gangs have not been adequately evaluated to assess their impact on gang membership. This is unfortunate. Significant human and financial resources are allocated to solving social problems, such as gangs, but not enough attention is paid to whether or not these efforts are successful. Because resources are finite, there is a need to determine which policies and programs actually achieve the intended results and which do not.

One often hears stories, anecdotes or testimonials that are presented as "evidence" of a program's or policy's effectiveness. Many policymakers and practitioners are motivated by the most salient or unique "success stories." But, although they may be inspirational — and even capture some of a program's impact — such stories, absent other data, are not sufficient measures of effectiveness. Judgments about a program's effectiveness should not be based on the salient or the unique; they should be based on science. The primary goal of evaluation research is to assess whether a program *causes* a desired outcome to everyone — or at least to a substantial proportion of those who have been exposed to the program — and not just the few exceptional cases. How do we know that certain prevention programs, policies or initiatives "work" and others do not? First, we must understand the evidence that exists or why it is important to "know what we know." Would it not be easier to support all programs and policies intending to reduce gangs (or anything else, for that matter)? Isn't something always better than nothing? There are three primary reasons the answer to both of these questions is "no."

First, resources are limited. It is a matter of necessity to prioritize programs when allocating financial and other support. There are a number of ways to accomplish this — for example, by looking at measures of cost-effectiveness or the number of people affected. One excellent way to determine which program to fund is to determine whether the program achieves its goals or, as this book discusses, whether it actually prevents gang membership.

Second, it is often impractical to implement multiple programs with the same intended outcome in the same population or community. This is simply the way the "real world" works: Loyalty to an ineffective or less effective program can make it impossible for a more effective program to be implemented. For example, the D.A.R.E. (Drug Abuse Resistance Education) program is one of the most widely implemented programs aimed at reducing drug use among adolescents. It has been in operation for more than 25 years and has been implemented in 43 countries.[1] D.A.R.E. is taught in 75 percent of school districts in the United States.[2] Evaluations of the program, however, have consistently failed to conclude that D.A.R.E. reduces drug use among youth.[3] The wide implementation of the program — and continued allocations of scarce resources to support it — decreases the likelihood that another drug-prevention program will be implemented in D.A.R.E. districts.

Finally, it is misguided to believe that, even if a program is not effective, it at least does no harm. An evaluation of a gang-intervention program in Los Angeles showed that participation in the program actually *increased* delinquency, solidarity and resistance among the gang members who participated.[4] In short, we need to know what works and why, and this chapter provides the groundwork for how this should be done.

To date, the Gang Resistance Education and Training (G.R.E.A.T.) program is the only program specifically focused on reducing gang membership that has been rigorously evaluated. The evaluation of the original G.R.E.A.T. program showed modest positive results with respect to several risk factors that are associated with delinquency and gang membership: peer group association, attitudes about gangs and law enforcement, and risk-seeking behaviors. However, the original evaluation showed no significant effects on delinquency or gang membership itself.

In response to those results — revealed through a rigorous evaluation — G.R.E.A.T. underwent a thorough review that resulted in a revised curriculum that emphasizes a skills-building approach. This revised program is currently being evaluated; the results after one year are promising: for example, there was a 39-percent reduction in the odds of gang-joining among the G.R.E.A.T. students compared with students in the control group.[5]

Although other gang-membership prevention programs exist, they either have not been evaluated or have not been evaluated using a research design that was rigorous enough to assess program effectiveness. For example, one common evaluation approach is to simply collect data on participants before and after participating in a program and then attribute any change in the post-test to the program. Such a conclusion, however, is not supported by evidence from a simple "pre-post" design because any number of factors could actually have caused the observed change — if, for example, another prevention strategy had been implemented in the community at the same time. Increasing the number of rigorous program evaluations is certainly the first step in determining what programs are effective. But spending resources on poorly designed program evaluations is almost as problematic as conducting no evaluation at all. This is particularly salient with a topic such as gang-membership prevention, where the will to implement programs is high but the knowledge of what works is limited. It is tempting to implement programs without, or with limited, evaluations; but a poor evaluation can lead to erroneous conclusions in either direction: that a truly effective program did not work or that an ineffective program had a benefit. Erroneous conclusions can lead to wasted resources if an ineffective program is continued; erroneous conclusions can also lead to missed opportunities if an effective program is stopped.

A well-designed program evaluation is critical to making the most strategically sound policy decisions. In the following discussion, we explain the characteristics of a well-designed evaluation, including factors that commonly impact the quality of a program evaluation.

The Basics of a Well-Designed Evaluation

There are three primary steps in evaluating a program or policy. First, program evaluators must identify the program's goals or intended outcome. This may seem basic, but it is fundamental. Although it is tempting — and often appropriate — to look for other outcomes, programs should be judged on the outcome that practitioners are *intending* to achieve.

Second, one must determine if the program was implemented as intended and designed; researchers call this program fidelity. Programs can have considerable variety in terms of content, duration, frequency and general delivery style. For a program to be deemed effective, however, it must bring about the intended changes as designed. For instance, the G.R.E.A.T. program is designed as 13 lessons to be taught by a law enforcement officer to middle-school students; if, however, the G.R.E.A.T. program were taught by current gang members (instead of police officers) and an evaluation found that it *increased* gang-joining, it would be inaccurate to say that the program does not work. Effectiveness — or ineffectiveness — cannot be assessed if a program has not been implemented with sufficient fidelity to its design.

Once a program's intended outcome is identified and it is determined that the program was implemented as designed, it can be evaluated to determine whether it had the intended result. A determination of effectiveness must also be based on (1) valid and reliable measures, and (2) appropriate sampling and research design.[6] Without appropriate measures and research design, it is possible, at best, to establish only an *association* between the program and its effect; but a determination that the program *caused* the change cannot be established.

An association between a program and its outcome (that youth in the program are less likely to join a gang, for example) is only the first element necessary to establish causality. Does the fact that gang membership is found to be lower after the introduction of a gang-membership prevention program mean that the program caused the reduction? Maybe yes, but possibly no. It could be that a decline in gang membership was part of a trend that started well before the program was implemented. It could be that a police crackdown on gangs occurred during the same period the program was implemented. It could be that people who participated in the program were the type of people who were less likely to join a gang. To conclude that participation in a program actually *caused* the decline in gang membership, researchers must use a strong evaluation design.

Only after an appropriate and well-designed program evaluation is in place can issues related to cost-effectiveness of the program be

assessed. From a policymaker perspective, cost-effectiveness is paramount. Can an initial investment in a prevention program lead to a reduction in costs associated with youth joining gangs? The answer depends on the effectiveness of the program. It must first be determined how much of a reduction in the outcome — in this case, fewer kids joining gangs — can be attributed to the program.

Program Design: Identifying Goals

Identifying a program's goals or intended outcomes may seem straightforward but, in reality, it is the foundation of an evaluation and should not be taken lightly. It is a task that must come early in the process because a strong evaluation hinges on this determination. Program administrators need to work with evaluators and make explicit what the programmatic goals are. Program administrators also need to clearly inform evaluators about the program components and the goals they are expected or designed to achieve. Absent a clear statement of program goals in the program design and a specific statement about how the program components are intended to achieve the desired outcomes, it is virtually impossible to assess a program's effectiveness.

In a recent evaluation, for example, my (Finn-Aage Esbensen) colleagues and I sifted through countless descriptions of a delinquency-prevention program to identify its goals — and found more than a dozen different stated goals.[7, 8] Through a review of the program design materials and discussions with program staff, we were able to specify three main program goals, including the reduction of victimization. Only after we had a clear understanding of the program's intended goals were we able to determine which program components were designed to achieve which goals. From this information, we were able to develop an evaluation (including the research design and instruments) that would measure the program goals.

Process Evaluation: Assessing Implementation Fidelity

It is important that policymakers and practitioners understand the importance of conducting a process evaluation of a program, policy or initiative in conjunction with an outcome evaluation.[6, 9, 10, 11, 12] A process evaluation assesses the quality of program implementation, or what researchers call program fidelity. A process evaluation determines if what is actually being delivered is consistent with what was intended. As Peter Rossi and his colleagues state in their work on the importance of evaluations, "A precondition for impact on the social conditions a program addresses is that the program actually be implemented in a manner that could plausibly affect those conditions."[13]

One of the key predictors of a program's effectiveness is the quality of its implementation.[14] Unfortunately, program implementation failure is common. Gary Gottfredson and his colleagues concluded, in the *National Study of Delinquency Prevention in Schools,* that "about half of school-based prevention activities are of such poor quality that they cannot reasonably be expected to make a difference in levels of problem behavior."[10] Process evaluations are necessary to determine which programs fail because of poor design as opposed to poor implementation. One of the reasons researchers often cannot tell policymakers and practitioners precisely which programs do — and do not — "work" is that programs fail to provide services consistent with the program's design, reducing their chances of effectiveness or success. Too often, there are fundamental changes in who delivers the program, content is skipped, or the program is modified in ways that were not part of the original program design.

There are a number of ways to determine whether a program was implemented with fidelity. It is important for policymakers and practitioners to be aware that there are pros and cons associated with each method. One of the simpler strategies is to ask program providers to indicate the extent to which they comply with the program design. This method can be a cheap and easy way to obtain useful information, but the disadvantage is that self-reported representations of fidelity to a program's design would not be independently verified.

In an evaluation of a school-based delinquency program, self-reports from program staff revealed that most were not engaged in one core feature of the program — using outside experts to supplement the program content.[7, 8] This was a major indicator of the lack of program fidelity. A downside to this approach is that individuals involved in program delivery may have a vested interest in showing their own compliance or

effectiveness or in concealing other variations from the design that should have been exposed. Given these sources of potential bias, results from the self-reported process evaluations should be interpreted with caution.

Having a third-party evaluator observe program delivery is a more objective approach than self-evaluation in assessing program fidelity. With the cooperation of program personnel, many types of prevention programs can be observed for consistency in implementation. Observations are well-suited for lesson-based programs like those common in schools. In the G.R.E.A.T. program evaluation, for example, my (Finn-Age Esbensen) fellow researchers and I developed a detailed instrument for each lesson taught by officers in the classroom.[15] A trained observer provided an overall assessment of the fidelity (quality, dosage and adherence) of the officer's program delivery. Each classroom included in the study was observed multiple times to assess whether problems with program delivery were a one-time occurrence or a common event. This observational process evaluation strategy is far more costly and time-intensive than the self-evaluation approach, but it offers much more reliable information about the quality of program implementation.

In our "In the Spotlight" interview (below), Lieutenant Raj Ramnarace, with the LaCrosse (WI) Police Department, describes his experience with

IN THE SPOTLIGHT: THE G.R.E.A.T. PROGRAM

INTERVIEW WITH RAJ RAMNARACE

The Gang Resistance Education and Training (G.R.E.A.T.) program is a national, school-based gang and violence prevention program taught by trained law enforcement officers. The program has three primary goals:

1. Teach youth to avoid gang membership.

2. Prevent violence and criminal activity.

3. Assist youth to develop positive relationships with law enforcement.

We interviewed Lieutenant Raj Ramnarace, M.Ed., the former regional administrator of G.R.E.A.T.'s Midwest Atlantic Region. Ramnarace currently serves with the LaCrosse (WI) Police Department.

How did you become involved in G.R.E.A.T.?

I became a certified G.R.E.A.T. instructor in 1993. A year later, I joined the National Training Team and began training other officers to become G.R.E.A.T. officers.

How did your work influence the actual development of G.R.E.A.T.?

For 10 years (1998–2008), I was the Midwest Regional Administrator and, as a member of the National Training Team, I served on the committee that produced the new curriculum. Because we knew that program implementation fidelity was the best way to ensure success, we supported the use of monitoring program implementation and effectiveness.

How was this actually manifested?

The U.S. Department of Justice's Bureau of Justice Assistance, the agency responsible for operational control of G.R.E.A.T., trained a select group of senior G.R.E.A.T. officers to conduct program audits during site visits to observe program delivery. As another way of enhancing program fidelity, a number of cities began pre- and post-testing of students.

What other approaches does G.R.E.A.T. use to improve the program's effectiveness?

We use evaluation forms to survey teachers, and we conduct periodic reviews with school districts to assess how their G.R.E.A.T. program fits into other curricula the school district is providing.

What challenges have you faced in ensuring implementation fidelity?

Once officers are in the classroom, they are on their own. Many teachers are not familiar enough with G.R.E.A.T. curricula to know whether the curriculum is being delivered appropriately. We have to rely on G.R.E.A.T. instructors and their agencies to provide information on the curriculum to the classroom instructors.

How do you address this?

Over the years, I've seen officers who represent two different points on the fidelity-effectiveness continuum. The first are those who lack either the experience or the motivation to deliver the curricula in a way that engages students, regardless of whether they are following the curriculum outlines. The second are those who are really experienced and are able to add examples from their own experiences to the program's curriculum when they see a teachable moment. The key, of course, is knowing how much an instructor can add without violating program fidelity. For that reason, we stress program fidelity and interpersonal skills to our instructors in training. We consistently stress the benefits of adherence to the curriculum, while also trying to impress upon our instructors that building positive relationships with students will improve the curriculum's effectiveness. Nationally, G.R.E.A.T. has offered training for G.R.E.A.T. supervisors in local agencies so that they know what is expected of the G.R.E.A.T. instructors they supervise. And, as I mentioned before, we do conduct classroom audits, which are very valuable.

strategies for determining whether a program is being implemented with fidelity to its original design and goals.

Outcome Evaluation: Assessing Program Effectiveness

A program's effectiveness cannot be determined without a rigorous research design that is able to causally link its components to outcomes and excludes other potential explanations for the outcomes. A rigorous outcome evaluation can establish that a change in behavior — for example, preventing youth from joining gangs — is due to the program and not to other external factors such as maturation (aging of the participants), selection of program participants who were not at high risk of becoming gang-involved (what researchers call *sample selection bias,* or "creaming the sample"), or some other shared experiences by program participants. It is important that policymakers and practitioners understand the components of the most rigorous evaluations and, most important, be able to articulate to their constituents the real-world occurrences that sometimes make an outcome evaluation difficult to execute.

The research community has been proactive in determining what kind of evaluation leads to the determination of an effective program. Although there are different criteria that can define program effectiveness, the most rigorous classification is the one used by the University of Colorado's Blueprints program.[16] To be classified as a "promising" Blueprints program, the program evaluation must have used an experimental design (which involves random assignment to the treatment or to the control group, along with pre- and post-test measures) or a quasi-experimental design (with treatment and control groups matched on key variables) and have found evidence of significant effects.

To earn the "Blueprints model" label, the program must meet two additional criteria: The significant effect must be sustained for at least one year post-treatment, and the program must have at least one high-quality replication that also shows effectiveness.[11]

As we know, youth gangs are found all over the U.S., yet a gang-membership prevention program that is effective in Minot, ND, may not be as effective in an urban area with greater population diversity. To have confidence that a program works, it must be replicated in multiple locales with different characteristics.

When all is said and done, program evaluators may make definitive statements regarding the effectiveness or ineffectiveness of a program, but informed policymakers and program administrators are responsible for deciding what such findings mean for the future of the program. For example, policymakers and practitioners are likely to face the situation where a program may be found to be effective in one group of individuals but not another. Should the program be eliminated or refocused on a narrower client base? Perhaps additional resources should be allocated to evaluate the program in a different area, with a different population or with a commitment to better methods. Perhaps the program could be revised? Practitioners and policymakers are best equipped to determine the most appropriate course of action only if the quality of the program implementation was high and a rigorous outcome evaluation design was used.

Control or comparison groups, and pretests. To determine whether a program reduces gang membership, two things must occur:

- Treatment and control or comparison groups must be employed.
- Answers to gang affiliation questions after program completion must be compared with answers before implementation of the program.

Without both the comparison group and pretests, the important questions, "Compared with what?" and "Compared with whom?" cannot be answered. The following example illustrates the importance of these comparisons.

Gang-joining starts relatively early, around 12 or 13 years old, and escalates through the mid-teen years. A prevention program that targets 12-year-olds may seem ineffective when gang involvement has increased from the pre- to the post-test.

Such a conclusion, however, may not be the entire story; such an increase could be due to aging or maturation of the students. If, however, one also compared the rate of gang membership of the treatment group with a group of students not exposed to the program, the results may show that the increase in gang membership for program participants is less than that observed among nonprogram participants — that is, the program does have a preventive effect.

Realities of random assignment. To ensure that no other factor is the actual cause of a change in the outcome, the comparison group needs to be as similar as possible to the treated group before exposure to the program. The surest way for an evaluation to accomplish equality between treatment and control groups is to follow the true experimental method and randomly assign individuals to each group. Of course, it is important that policymakers and practitioners understand that this is not always practical. For example, with school-based prevention programs, it would be impractical (from the school's perspective) for researchers or program administrators to randomly assign students to receive, or not receive, the intervention because this would disrupt intact classrooms. Does this mean that equitable groups cannot be created? No. It is possible to randomly assign *classrooms* to receive or not receive the program. In this way, treatment and control groups are still available, educators can deliver the program with ease, and the school is minimally disrupted.

It is important that policymakers and practitioners understand the ethical considerations of random assignment. A common concern is the ethics of withholding services from individuals (those in the control group) who might benefit from the program. For example, assigning individuals who qualify for treatment as part of standard practice to the "no treatment" control group would be unethical despite the use of the evaluation "gold standard" of a randomized control trial (RCT). There are a number of ways to address this issue. At a minimum, control subjects should receive whatever intervention or treatment they would otherwise receive. This can be considered the "usual care" condition. When there is no usual intervention, it is often still desirable to provide a minimal version or different form of intervention, or the control subjects could be put on a "wait list," going without treatment only until a spot in the treatment group becomes available. However, when assessing a program evaluation, it is important to keep the treatment and control groups distinct. Undermining the randomized assignment, for any reason, can jeopardize the validity of study results and lead to over- or underestimating the true effect of the program.

Attrition. The Blueprints standard requires that the outcome show at least some sign of stability: evidence of a program effect one year after participants complete the program. This requires that study participants be tracked across time. Whenever such follow-up is required, evaluators must deal with the issue of sample loss or attrition. People move, die or otherwise disappear — and, therefore, it is important that policymakers and practitioners understand the potential consequences of attrition.

First, attrition can lead to a loss of statistical power, which means that if too many subjects are lost, it might not be possible to detect meaningful differences across groups. Second, individuals "lost" over time may represent higher-risk youth, and this may bias the results. Especially in gang-related research, gang members may be more likely to leave a study, and the loss of the most extreme members may compromise the findings. Evaluators need to make every effort to retain study participants, but attrition is a reality in longitudinal research — and it is important that policymakers and practitioners understand the challenges that attrition may pose for interpreting evaluation results.

Bias due to attrition may be indicated when the final treatment and control samples are notably different from the original samples. Assume, for example, that there were 50 high-risk kids in both the original treatment group and the control group but, at the end of the study, there were 45 high-risk kids in the control group and only 20 in the treatment group. In this case, if the final treatment group is at low risk for gang membership, it would be difficult to know whether this was because the program worked or because many high-risk kids could not be assessed.

Generalizability and selection bias. In addition to being essential in assessing the effectiveness of a program, proper sampling can also increase what researchers call the *generalizability* of the outcome results. A robust sample size drawn from a diverse study population is crucial to making well-informed inferences about the potential effectiveness of a program or policy in a variety of settings. Subsequent replication with other populations should also be conducted. As researchers, practitioners and policymakers know, however, program evaluations are time-consuming and costly. In addition, local evaluations can tell us things that a national evaluation cannot — and vice versa. That is why it is important that everyone understands this reality: an evaluation concluding that a program is effective for 14 participants in town X in year Y with counselor P is certainly informative, but it is not as informative as being able to conclude that a program is effective in a diverse population in several cities over multiple years and across many program administrators. Evaluations can get "more value for their dollar" with careful planning and effort in the design phase.

Generalizability is possible only if the sample that is included for treatment, and therefore evaluation, is not biased. Bias can occur in a number of ways during the recruitment of study participants. This is particularly important to keep in mind when considering research on gangs and gang-membership prevention programs. Sometimes clients can be selected or deselected based on certain criteria. A *gang-membership prevention* program may be reluctant to accept *current gang members,* which would limit the generalizability of the findings only to individuals with no prior history of gang membership.

Another type of selection bias may occur if, for example, youth at greatest risk for gang-joining are excluded due to staff concerns about meeting program goals or expectations. This type of selection bias, restricting a program to youth at low risk of gang membership, is referred to as "creaming the sample" and it could increase the probability of finding benefits, such as might occur if these youth are inherently easier to work with or more likely to participate fully in the program.

Selection Bias: What Is the Role of Informed Consent?

Another issue related to selection bias is informed consent. When governmental agencies and research institutions (universities and private research firms) conduct research involving human subjects, they must detail plans and strategies to guarantee the protection of the participants' rights.

When this research involves minors — as it inevitably will with programs or strategies to prevent gang-joining — this protection generally requires obtaining parental consent for the child's participation.

Two types of parental consent exist — *passive* and *active:*

- Passive parental consent requires a form to be returned only if the parents *do not* want their child to participate.
- Active parental consent mandates that the researcher obtain permission for every child to be included in the study.

Active consent is more difficult to obtain and increases the risk of the selective loss of higher risk subjects.[17, 18] If the parents of only the high-risk youth were to refuse to sign the form, for example, this would reduce the generalizability of the results.

Under certain circumstances — if, for example, the study involves a sample for which parental consent is not a reasonable requirement to protect a child — an organization's Internal Review Board (IRB) can grant a waiver of parental consent.

For example, imagine a program in which gang members *volunteer* to participate in an evaluation. The evaluation design randomly assigns some to the program and others to the control group. The program is eventually shown to reduce gang involvement. Is the program an effective gang-reduction program? Yes and no. Yes, it is effective in reducing the involvement of the participants, but who are the participants? As two highly respected researchers, Malcolm Klein and Cheryl Maxson, have observed, it is important to understand this potential bias in gang research: Gang members who are willing or interested in participating in research "are likely to be atypical of the general gang membership."[19] Gang members who volunteer to participate in gang programs may be looking for a way out of the gang or may be less serious gang members to start with and, therefore, would not likely be representative of the gang members who did not volunteer.

Without due attention, selection bias can confound program evaluations and lead to results that only apply to specific subgroups. Practitioners and policymakers using program evaluations to inform their decisions should pay particular attention to possible selection bias in any program and evaluation — and be prepared to question inflated claims. If, for example, representations are made that a program "works" for *all* gang members, but it was administered to and evaluated only for male gang members, the veracity of the claim should be questioned. (See sidebar, "Selection Bias: What Is the Role of Informed Consent?")

Cost-Effectiveness

One important, but often overlooked, aspect of program evaluation is assessing the relative cost associated with achieving a desired outcome. Once evaluators have determined that a program reduces gang membership, the question then becomes one of cost. Two strategies for addressing this issue are cost-benefit and cost-effectiveness analyses. Cost-benefit analyses are more difficult to implement because program benefits must be expressed in some form of monetary terms. For instance, in reducing gang membership, a cost could be placed on the savings of crimes not committed, arrests not having to be made, and savings associated with lower incarceration rates.

A cost-benefit analysis of a variety of crime and delinquency prevention programs, including the Blueprints programs, has been conducted.[20, 21] The cost-benefit analysis of Multisystemic Therapy, for example, revealed a benefit of $4.36 for every $1.00 of program costs. The cost-benefit analysis of Functional Family Therapy revealed a savings of $10.42 for every $1.00 of program costs.[20, 21] Most of the savings are from reduced crime. Although neither of these programs has gang-membership prevention or gang-crime reduction as a primary programmatic outcome, both include principles that may help keep kids from joining a gang.

On the other hand, a cost-effectiveness analysis calculates only the cost to run a particular program — such as personnel, supplies, space, transportation — but assesses program *effectiveness* in terms of behavioral or performance outcomes. For gang-membership prevention, for example, we could examine the percentage reduction in gang membership relative to the actual cost of delivering the program. Of importance is determining the extent to which a given program, relative to other programs or to program costs, is cost-effective and therefore worthy of implementation or continuation.

Policy Issues

Policymakers are key players in ensuring accountability in gang-membership prevention programs and policies. The first step is to understand that program evaluations are crucial. Unfortunately, many policymakers fail to require program evaluations when they award funding to new prevention programs or when a program has been substantially modified or is being used with a new population. This practice should change. Evaluation of both the implementation (delivery process) of the program and the outcome (effectiveness) should become the norm. Policymakers should encourage evaluators to conduct the most rigorous evaluations, based on the criteria described above, if possible. This will help ensure that evaluations are both fruitful and economical. Policymakers should also encourage practitioners to develop partnerships with researchers (from local universities, for example) to facilitate objective, rigorous evaluation of their programs.

A solid, scientifically dependable "evidence-based" gang-membership prevention program demands the collaboration of practitioners, policymakers and researchers. To achieve the best results, it is critical that a program evaluation be funded and designed *prior* to implementation of the program. Program personnel must cooperate with evaluators to reduce problems of selection bias and sample attrition. If program goals are not clearly articulated to allow evaluators to develop appropriate measures of key outcomes, the evaluators are forced to design their evaluation based on previous decisions made by the program's practitioners, which may lead to sacrifices in the rigor of the evaluation that could have been avoided. For example, if a program is designed to be administered to every child in a state prior to "pretesting," program evaluators will have no comparison group within that state. They will be forced to select a control group from a different, but hopefully similar, state. Or — if a program is delivered with minimal design detail — it could be implemented without consistency (fidelity) across all sites, rendering a determination of effectiveness impossible or unreliable.

To really be confident about what works to prevent gang-joining, we must have rigorous evaluations. Funding, commitment and a shared belief in the importance of evaluations are what it takes to get the job done. To improve the prospects for accurately concluding that a program works, there must be sufficient funds allocated to conduct a rigorous evaluation. Without rigorous process and outcome evaluations, it is unlikely that scarce community resources — both monetary and personnel — will be used as effectively as possible.

About the Authors

Finn-Aage Esbensen

Finn-Aage Esbensen has conducted research in the areas of youth violence and gangs for more than three decades. For the past 18 years, Dr. Esbensen has focused on the evaluation of school-based prevention strategies, most notably the multisite evaluations of the Gang Resistance Education and Training (G.R.E.A.T.) program. He received his Ph.D. from the University of Colorado and is currently a professor in the Department of Criminology and Criminal Justice at the University of Missouri-St. Louis.

Kristy N. Matsuda

Kristy N. Matsuda studies the causes of and responses to juvenile delinquency and gang membership. In addition to evaluating a national gang-membership prevention program, Dr. Matsuda has extensive experience in evaluating programs, policies and the causes of violence for the California Department of Corrections and Rehabilitation. She received her Ph.D. from the University of California, Irvine, and is currently an Assistant Research Professor at the University of Missouri-St. Louis.

Endnotes

1. D.A.R.E. America. 2*5 years D.A.R.E. America.* Los Angeles, CA: D.A.R.E. America, 2008.

2. Berman G, Fox A. *Lessons From the Battle Over D.A.R.E.: The Complicated Relationship Between Research and Practice.* Washington, DC: U.S. Department of Justice, Bureau of Justice Assistance, 2009.

3. Ennett ST, Tobler NS, Ringwalt CL, Flewelling RL. How effective is drug abuse resistance education? A meta-analysis of project D.A.R.E. outcome evaluations. *Am J Public Health* 1994; 84:1394-1401.

4. Klein M. *The American Street Gang.* New York, NY: Oxford University Press, 1995.

5. Esbensen F-A, Peterson D, Taylor TJ, Osgood DW. Results from a multi-site evaluation of the G.R.E.A.T. program. *Justice Q.* 2012; 29:125-151.

6. Summerfelt WT. Program strength and fidelity in evaluation. *Appl Dev Sci.* 2003; 7:55-61.

7. Esbensen F-A. *Evaluation of the Teens, Crime, and the Community and Community Works Program.* Washington, DC: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice, 2005.

8. Melde C, Esbensen F-A, Tusinki K. Addressing program fidelity using onsite observations and program provider descriptions of program delivery. *Eval Rev.* 2006; 30:1-27.

9. Fagan AA, Mihalic S. Strategies for enhancing the adoption of school-based prevention programs: Lessons learned from the Blueprints for Violence Prevention replications of the Life Skills Training program. *J Community Psychol.* 2003; 31:235-253.

10. Gottfredson GD, Gottfredson DC, Czeh ER. *National Study of Delinquency Prevention in Schools, Final Report.* Ellicott City, MD: Gottfredson Associates, Inc., 2000.

11. Mihalic S, Fagan A, Irwin K, Ballard D, Elliott D. *Blueprints for Violence Prevention Replications: Factors for Implementation Success.* Boulder, CO: University of Colorado, Center for the Study and Prevention of Violence, 2002.

12. Mihalic SF, Irwin K. Blueprints for violence prevention: From research to real-world settings — factors influencing the successful replication of model programs. *Youth Violence and Juv Justice* 2003; 1:307-329.

13. Rossi PH, Freeman HE, Lipsey MW. *Evaluation: A Systematic Approach.* 6th ed. Thousand Oaks, CA: Sage Publications. 2004.

14. Lipsey MW. The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Vict Offender* 2009; 4:127-149.

15. Esbensen F-A, Peterson D, Taylor TJ. *2009 Report to Schools and Communities: Program Implementation Quality and Preliminary Outcome Results.* St. Louis, MO: University of Missouri-St. Louis, 2009.

16. Center for the Study and Prevention of Violence. Blueprints for Healthy Youth Development. Boulder, CO: University of Colorado at Boulder, 2010. Available at http://www.blueprintsprograms.com. Accessed on May 30, 2012.

17. Esbensen F-A, Miller MH, Taylor TJ, He N, Freng A. Differential attrition rates and active parental consent. *Eval Rev.* 1999; 23:316-335.

18. Esbensen F-A, Melde C, Taylor TJ, Peterson D. Active parental consent in school-based research: How much is enough and how do we get it? *Eval Rev.* 2008; 32:335-362.

19. Klein M, Maxson CL. *Street Gang Patterns and Policies.* New York, NY: Oxford University Press, 2006:164.

20. Aos S, Lieb R, Mayfield J, Miller M, Pennucci A. *Benefits and Costs of Prevention and Early Intervention Programs for Youth.* Olympia, WA: Washington State Institute for Public Policy, 2004.

21. Lee S, Aos S, Drake E, Pennucci A, Miller M, Anderson L. Return on investment: Evidence-based options to improve statewide outcomes (Document No. 12-04-1201). Olympia, WA: Washington State Institute for Public Policy, April 2012.