Juvenile Justice: Advancing Research, Policy, and Practice
Edited by Francine T. Sherman and Francine H. Jacobs

CHAPTER 24

Better Research for Better Policies

Jeffrey A. Butts and John K. Roman

To do their jobs effectively, policy makers, professionals, and community partners must be able to access high-quality information about the impact of policies and programs for youth. Recent years have seen an increasing, and appropriate, focus on evidence-based policy. In setting priorities for funding and support, intervention programs demonstrated to be effective and efficient are preferred over programs that are well intentioned but untested by rigorous evaluation. An evidence-based approach is undeniably better than an approach based on faith or anecdotes, but the findings of existing evaluations are not sufficient by themselves as a basis for effective policy making. Translating research into practice requires more than a review of existing studies. It requires knowledge of the research process and its limitations. How do researchers generate evidence? What choices are involved in designing evaluation studies? Who sponsors research and how do they select one study over another? How do researchers and their funding bodies shape and interpret the results of research? Who disseminates research findings and how does the manner of presentation color the impact of information? A clear-eyed investigation of the entire evidence-generating process is an invaluable part of evidence-based policy.

There is no such thing as a perfect study. Criticizing social science research takes very little effort. Human behavior is enormously complex and the study of human behavior is fraught with challenges. Assessing the effect of these challenges is even more difficult due to the natural tendency of researchers to prefer positive findings to negative findings. Researchers typically work harder on fact-checking negative results—especially those contrary to their beliefs—than they do examining positive results that may support those same beliefs. Negative findings are scoured for errors while positive findings are accepted more readily. This apparent imbalance might lead one to conclude that the deck should be stacked against making changes to policy and practice simply to avoid risk. Until researchers produce solid evidence of impact, according to this argument, policy makers should stick to convention. We think such an approach is a mistake. Almost no one with detailed knowledge of the juvenile justice system would argue that its current policies and programs offer the best possible strategy for community safety and youth well-being. The take-no-risks approach would create a bias toward the status quo, which we know does not do an adequate job of protecting youth, families, and communities. Policy makers must continue to invest in new solutions and they must do so using the best available evidence, but the absence of perfect evidence should never be an excuse for inaction.

This chapter describes the strengths and weaknesses of the evidence-generating process as applied to juvenile justice. We begin by

describing how research fits into the decision-making process. We then describe different types of research evidence and how to evaluate the validity of conclusions developed using various strategies and techniques, followed by a description of some of the common obstacles facing efforts to translate evidence into practice. We take up the complex challenges of evaluating changes in policy as opposed to programs and offer ideas for improving evidence-based decision making. Finally, we suggest some practical strategies for reporting research results more clearly and we consider how to support the role of evaluation in designing future policies and programs.

THE RESEARCH MARKETPLACE

There is little doubt that research plays a role in the design of social policies and programs, including the selection of one juvenile crime prevention approach over another. There could be a debate, however, about the role that research *should* play in social policy. Most opinions on the matter fall between two extremes. At one end of the spectrum, it is possible to argue that the design of all policies and programs to address youth crime should be governed by research and that only "evidence-based" solutions merit the support of policy makers (Greenwood, 2006; see also Greenwood & Turner, Chapter 23, this volume). A lesser standard, after all, risks the possibility of failed policies and wasted resources. At the opposite end of the spectrum, one might argue that research is imperfect and it should never be allowed to exert complete control over which policies and programs are selected for implementation (Schorr, 2009). Research funding and research designs are vulnerable to contamination by bias and institutional self-interest. An extreme adherent of such a perspective might assert that the most appropriate sources of inspiration for social policy are the values and preferences of the people and their elected representatives. The role of research is to test whatever policies are pursued by experts and by public officials, not to control the selection of those policies.

Both extremes in such a debate, of course, would be impractical. Certainly, research and evaluation will always play some role in youth justice policy; just as certainly, that role will never be absolute or controlling. The design and implementation of policies and programs involves inevitable tension between values, beliefs, material incentives, and evidence. Managing this tension is a never-ending struggle that depends on the particular policy arena in which the debate occurs and on the motivations, abilities, and power of those engaged in the struggle. Public officials are always in favor of making decisions based on evidence. Like most human beings, however, they are usually interested in testing other people's ideas. They are less interested in scrutinizing their own ideas, believing them to be in need of confirmation rather than investigation.

Some policy ideas are never really tested. Ironically, it often seems that the largest questions of policy receive the least empirical examination. During the early years of the second Bush administration in the United States, for example, federal agencies were required to comply with a program of accountability using a procedure called PART (Program Assessment Rating Tool). Administered by the President's Office of Management and Budget (OMB), the PART program directed federal departments to conduct research on their programs and initiatives with the understanding that funding would be cut or reduced for programs that did not meet certain effectiveness thresholds (Frisco & Stalebrink, 2008). The PART program was an ambitious

effort to bring research-based accountability to the implementation and funding of federal programs. The standards were not always well conceived, however, and not all federal efforts were subjected to the standards in the same way. Social programs were far more likely to be targeted by PART than were military expenditures and foreign policy. The biggest policies are rarely subjected to the same level of evidentiary scrutiny imposed on social programs.

Some decisions, in fact, must be made without evidence. No matter how creative and how fair a program of accountability research may be, there will never be sufficient resources for researchers to test all possible beliefs and all possible theories behind a particular policy. Given that evaluation research will never be sweeping and comprehensive, policy makers are forced (and sometimes may prefer) to make some choices without solid evidence. The very fact that research cannot address all policy questions suggests that the funding of research itself is an important part of the policy process. Policy makers engage in the creation of evidence when they decide where, when, and how to deploy the tools of research and evaluation. Which programs and policies should be tested through rigorous research, and which can be assessed using less rigorous methods? Who should be charged with conducting such research? What standards of evidence should be considered minimally acceptable, and who should judge whether specific types of evidence are good enough or persuasive enough to inform policy and practice? Who should be responsible for communicating the findings of research and disseminating research products? In particular, who is responsible for translating the findings of research to audiences outside the self-referential confines of the research community? If scrutinizing the effectiveness of social programs is to be more than an insider's game, the findings of research have to be accessible to nontechnical audiences.

These issues are resolved through the competition of ideas and influence. It would be naïve, however, to assume that the programs and policies at the top of the evidentiary hierarchy got there solely on their own merits and due to the strength of their outcomes. Policy makers and funding authorities select how and where to invest their limited resources for research. Ideally, their investments would be focused entirely on improving the quality of evidence for the formulation of policy and practice. At times, however, they are subject to broader social dynamics related to politics and economics. A rigorous research program on the impact of juvenile incarceration, for example, would likely reveal basic conflicts over the purposes of juvenile justice. The official policy is that we incarcerate juveniles to facilitate rehabilitation and to ensure public safety; the reality, of course, is more complex. Other factors, such as ideology, fear, anger, and fiscal incentives influence the use of incarceration. Confronting these influences would be disquieting to many policy makers. Thus, we invest more in research to compare the relative effects of various drug treatment approaches than we do measuring the impact of the massive and costly social program called incarceration.

As policy makers decide how to invest in research, the cost of evaluation is a prominent concern but feasibility and salience are weighed heavily as well. A low-cost, highly feasible evaluation of a prominent policy issue is much more likely to attract funding and the support of decision makers than a high-cost, risky study of an issue that is less prominent or less understood by the public and by elected officials. Some research projects are more likely to be funded because they conform

more easily to conventional ways of thinking about social problems. If simply asking a research question threatens powerful interests or institutions, the authorities charged with funding research projects are less likely to risk asking the question. Research helps to shape juvenile justice policy and delinquency prevention programs, but it does so through a complex marketplace of resources, ideas, values, and power.

TYPES OF RESEARCH

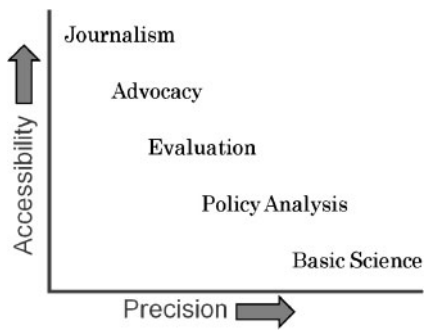
All forms of research on social policy, including youth justice policies and programs, could be divided along two dimensions: accessibility and precision (see Figure 24.1).

Investigative journalists, for example, conduct research and use data to tell a story. This form of research is not very precise, but it can be highly accessible with immediate impact. A small group of lawyers and journalists associated with the Innocence Project at the Benjamin N. Cardozo School of Law in the United States has been working for nearly 20 years to investigate the inner workings of the criminal justice system in death penalty cases with the goal of freeing prisoners wrongly convicted. Their investigations require extensive data collection and analysis, but their methods and results are not technical. Almost anyone can read the work and appreciate its importance. This "research" has resulted in a number of high-profile exonerations and the related stories have documented the origins and impact of wrongful convictions in the U.S. justice system. The stories have prevented several executions and may have helped to alter public opinion about capital punishment.

At the opposite end of the continuum, basic science research is often highly precise and targeted on clearly stated, empirical questions. These studies, however, often involve methodologies that require advanced training simply to understand, and they may address research questions that are not salient outside of a small, expert audience. Furthermore, many researchers in the social policy field write poorly. Even high-quality studies are sometimes presented using ponderous, dense language and complex scientific notation that is incompatible with popular consumption. The impact of their findings is often slow to develop at best. They must be translated into simpler language before they can be widely disseminated, and even then they may answer questions that few people care about.

Basic science, however, is a critical part of the process that leads to effective policies and

Figure 24.1 Various Types of Research Involve a Trade-off Between Precision and Accessibility



programs. Within the broader field of social research, it is often true that the people with the most advanced technical skills are found in academic institutions, where researchers are rewarded for pursuing the highly technical studies valued by the social sciences. Their work may be impenetrable to anyone outside their small peer group, and they may be incapable of describing their work to people who do not also share their vocabulary and reading list. Yet, when they are assisted by others who interpret their findings and translate them for a broader audience, basic science investigations can make profound contributions to the development of social policies and programs (see Table 24.1).

One recent example is the rapidly growing knowledge of brain functioning that has been facilitated by basic science involving scanning technology (Casey, Tottenham, Liston, & Durston, 2005). The ability to create detailed images of the working brain once seemed rather esoteric, but by the end of the 20th century, applied researchers were using the technique to understand how the brains of adolescents are different from those of adults. Science revealed that the brain of a 16-year-old adolescent, while more advanced than the brain of a 10-year-old child, is not as developed as the brain of a 25-year-old adult. Using scanning technology in applied research made it possible for the first time to show non-technical audiences that cognitive development is a continuous process that does not end suddenly at age 14. The science had a powerful impact on behavioral researchers and on legal thinking in the field of adolescent development, even affecting the reasoning of the U.S. Supreme Court when it banned the use of capital punishment for offenders who commit their crimes before age 18 (Scott & Steinberg, 2008).

A second example of the effective application of basic science for policy and program development is related to the sociological concept of "collective efficacy" (e.g., Sampson, Morenoff, & Earls, 1999). The notion that neighborhood social organization and the structure of interlocking relationships can have an independent effect on social problems that are normally measured at the individual level arose from basic social science on

Table 24.1. Researchers, Audiences, and Methods

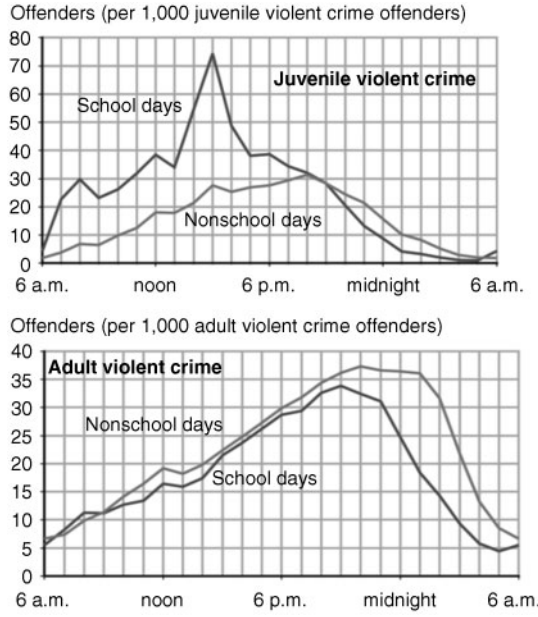
Professional Group	Primary Audience	Self-Perceived Mission	Preferred Methods	Dissemination Outlets
Basic science researchers	Other researchers	Basic science, hypothesis testing, theory development	Primary data collection, complex analysis, technical writing	Peer-reviewed journals, academic conferences
Statistical policy analysts	Public officials	Evaluation, policy simulation, program demonstration, implementation science	Primary data collection, secondary data analysis, complex analysis, technical writing	Peer-reviewed journals, government publications, professional conferences
Program evaluators	Public officials and practitioners	Evaluation, program demonstration, effectiveness testing, problem solving	Primary data collection, secondary analysis of agency data, basic analysis, accessible writing	Government publications, client reports, professional conferences
Policy advocates	Public officials, practitioners, general public	Reform, policy change, program development	Minimal data analysis, story telling, highly accessible writing	Trade publications, news media
Journalists	General public	Public education, reform	Minimal data analysis, story telling, highly accessible writing	News media

the formation of social capital and community capacity (Coleman, 1990; see also, Hawkins, Vashchenko, & Davis, Chapter 12, this volume). These findings helped to spark an entirely new direction in crime prevention policy (Browning, Feinberg, & Dietz, 2004; Mazerolle, Wickes, & McBroom 2010; Roman & Chalfin, 2008). Even nonscientists began to understand that criminal and delinquent behavior originates in social and community forces as well as in the individual characteristics of offenders.

In between the obvious accessibility of journalism and the admirable precision of basic science, there are several forms of research that play important roles in the formulation of social policies and the design of social programs (see Table 24.1). Policy analysts focus their efforts on specific social interventions, and their work can be just as sophisticated and technically complex as the work of researchers pursuing basic science. The mission of policy analysts is to assess the implementation of policies and to conduct quantitative simulations of their likely impact. They may publish their work in academic journals, but they value dissemination in government publications and professional conferences just as much. Like basic science researchers, policy analysts can be very mathematical and their work can be technically complex. Their presentation style is not always accessible to mainstream audiences, but when the findings of policy analysts are translated for broad consumption they can have significant impact.

One prominent instance of effective policy analysis involved an examination of the time of day when juvenile crime happens (see Figure 24.2). Beginning in the 1990s, a new type of law enforcement data made it possible

Figure 24.2 Unlike Crime by Adults, Juvenile Violent Crime Peaks in the Afternoon Hours



Source: Snyder and Sickmund, 2006.

for the U.S. Department of Justice to sponsor studies of juvenile crime according to the time of day when offenses tend to occur. Researchers found that, unlike adults, violent crime by juveniles peaked in the afternoon rather than in the evening hours. After a graphic portrayal of this finding appeared in national newspapers, including the *New York Times*, it began to change the awareness of the public and to draw the support of policy makers for afterschool programming for youth.

The work of another group of researchers, evaluators, is similar to that of policy analysts. Both types of research focus on the effects of policies and programs, but while policy analysis typically forecast future effects, evaluators often measure effects retrospectively (Brewer & deLeon, 1983). Evaluation research can be accessible to nontechnical audiences, although it requires advanced training. Evaluation researchers design their work to affect the understanding and actions of public officials and practitioners. Their work may involve primary data collection, but it often relies on the secondary analysis of agency data. Evaluations appear in government publications and client reports, but they may appear in the popular media as well, especially when the findings contradict conventional wisdom or when they address the impact of controversial or topical policies.

One recent example of the power of evaluation on policy formation concerned the anti-drugs program, DARE that was started by law enforcement and spread throughout the school systems of America. The first evaluation studies showed that DARE not only failed to reduce the use of drugs among students, it actually seemed to increase their interest in drugs and perhaps lessen their fear of using them (Ennett, Tobler, Ringwalt, & Flewelling, 1994). The ensuing controversy received plenty of media attention, and after an initial period of conflict, the founders of DARE changed their program model in an attempt to avoid future negative outcomes (Boyle, 2001).

Another example of effective evaluation can be found in studies of the Nurse Home Visitation program (now the Nurse–Family Partnership). Established in the 1970s, the program works with first-time single mothers, mainly teenagers, to improve outcomes for those mothers and their children. Beginning during pregnancy and continuing into the baby's second year, nurses regularly visit and provide guidance on nutrition and other health issues, as well as advice on parenting. By the 1980s, researchers were able to show that the program generated large, positive effects for both mothers and children (Olds, Henderson, Tatelbaum, & Chamberlin, 1988). In the ensuing years, the evidence had become so convincing that President Obama's first federal budget included nearly $10 billion for states to implement or expand similar programs (Eckenrode et al., 2010; Olds et al., 2004; Olds et al., 2007).

Evaluation research will always be an essential component in the development of social programs. While there is an academic field that focuses on the techniques of evaluation, and while evaluation researchers sometimes publish their work in peer-reviewed journals, evaluators are not always academics. Many professional evaluators work for government agencies, private research organizations, and consulting firms. They consider their audience to be the clients of their work (i.e., the agencies and programs they evaluate) as well as the general public.

JUDGING EVIDENCE

At this point, it may be helpful to identity exactly what we mean by research and evidence. In our view, *research* refers to any systematic investigation that uses a predefined

standard of quality to determine its methods and to govern the interpretation of evidence. The use of a predefined standard of quality is essential for distinguishing research from argument and speculation (or even fabrication). Ideally, a researcher will apply the highest possible standard in collecting and analyzing evidence to reach a conclusion. The appropriate standard of quality, however, may vary according to the type of research being conducted and the ways in which the findings are to be applied. Within the wide range of investigations that legitimately could be called research, quality standards might vary from *sufficient*, to *persuasive*, to *exacting*.

Journalists conduct research, as do literary scholars and lawyers. In these investigations, an appropriate standard of proof may be mere sufficiency, or "a preponderance of the evidence," or even "most experts agree." A lawyer may research the content and history of prior court rulings, for instance, to identify an underlying legal logic that was never articulated in the original body of opinions. Legal investigations should not be excluded from the phenomenon we call *research* simply because they rely on text rather than numerical data, or because they use logic rather than statistics to reach their conclusions. An investigation of legal precedent, however, can also be similar to literary criticism. Some legal investigations may merit the label *research*, and some may not.

Research evidence can take different forms. Some evidence originates from qualitative methods where the data are maintained in the form of stories or narratives and often begin with direct observation. Researchers may conduct interviews with elected officials and program staff and weave together a narrative about a program that identifies its critical components and draws inferences about its success. Qualitative studies have a clear role to play in the formation and assessment of delinquency prevention programs, but they rarely achieve the same policy impact as do quantitative, statistical studies.

The most influential type of research in delinquency prevention involves quantitative data and statistical procedures. Here also, however, standards of evidence vary. Some quantitative research focuses on the discovery of basic empirical facts and the testing of hypotheses. Does family poverty accelerate school failure, for example, or is the prevalence of vacant buildings in a community associated with its violent crime rate? Other research is designed to inform immediate policy decisions. For instance, should governments fund juvenile drug treatment courts? Do family violence interventions work? Are the effects of preventive health education worth the money it takes to deliver the training?

Studies of basic, empirical questions usually rely on statistical significance as their principal metric. Stated in terms of probability, or p values, a researcher might report that, indeed, family poverty *is* associated with school failure, and that the connection between the two concepts is so strong that there is less than a 1% probability ($p < 0.01$) that the statistical association could have occurred by chance alone. In other words, policy makers can be confident that such a finding points to something real and genuine. For example, say a particular study finds that the average annual income of youth who dropped out of school is 20% lower than the income of youth who stayed in school through graduation. If the statistical significance of this difference was described using a probability (or, p value) of less than one percent (i.e., $p < .01$), it would mean that the finding was very unlikely to be a fluke or coincidence. Statistical significance levels are traditionally stated in terms of the probability that a finding could be due to

chance alone. Researchers choose an appropriate level of probability depending on relevant theory and the findings of previous studies that suggest the extent of statistical variation likely to be found. Traditionally, studies use probability values of 1%, 5%, or 10% to determine the significance of their results, but the particular threshold is set in advance. Research findings are characterized as either significant or not significant; they are not compared, with one finding being termed more or less significant than another.

Statistical significance is not the only metric used by researchers, especially in evaluation studies. Significance is often not even the best way to assess the importance of a finding. Significance levels can be misconstrued. The level of significance associated with a particular finding is a function of several factors, but it is mainly derived from the absolute size of a difference in combination with the number of cases or observations (the N) used to establish that difference. Even a seemingly large difference (e.g., 80% of the youth in Group A were rearrested, but only 50% in Group B were) may fail to reach the level of statistical significance because the finding was generated with a very small sample. For example, the study may have collected data on just ten youth in each group.

In contrast, even a small difference (e.g., 50% versus 53% recidivism) could be statistically significant if the data set being analyzed was sufficiently large. Few public officials, however, would want to risk their resources or their reputations on a difference of 3 percentage points, unless they could describe the difference in some other way, perhaps economically or in terms of individual public safety (e.g., number of crimes averted). Researchers who focus on statistical significance alone can fail to appreciate the substantive importance of a finding. It is not uncommon to hear investigators at academic conferences draw profound conclusions or policy implications from relatively minor differences that are "significant" mainly because they were found using very large data sets. Outside of academic discussions, however, such minor differences are seen as relatively unimportant.

In the policy arena, "effect size" increasingly serves as an alternative to statistical significance. Effect size is often defined as the change in an outcome divided by its standard deviation, a traditional measure of statistical variation. Measures of effect size can be constructed in other ways as well, but all of the measures have a common function, which is to estimate the magnitude of a treatment effect given a specified level of intervention. A study might estimate the change in the prevalence of recent drug use among a sample of individuals following their participation in a new type of treatment, or researchers might measure change in the frequency of antisocial behavior in an entire community following the implementation of new juvenile curfew laws. Effect size, rather than statistical significance, is the primary language used to describe the benefits of social interventions.

The "effect" of a delinquency program could be the change observed in an important indicator of client behavior (e.g., recidivism). One common way to gauge whether an effect is large or small is to compare the scale of change in recidivism to the mean or average recidivism rate among a population of interest. A program that reduces recidivism by 50% will generally have a larger effect size than a program that lowers recidivism by just 10%. However, percent change is very sensitive to the mean level of the outcome. When mean levels are small, such as when only 10% of a sample is expected to be rearrested in the first place, a change of just 3 percentage points

(from 10% to 7%) would produce a relative change of 30%. Another way to gauge the size of an effect is to compare the scale of change with the natural variation of the key variable. For example, if recidivism for a particular group of youth was known to fluctuate between 5 and 90%, a change of 3 percentage points would seem trivial. On the other hand, if recidivism rarely varied outside a 5 percentage point range, for example from 45 to 50%, a program able to produce a consistent decline of 3 percentage points would likely have a very strong effect size.

In practice, effect sizes for delinquency prevention programs usually fall between −.30 (i.e., decreases the likelihood of delinquency) and +.30 (i.e., increases the likelihood of delinquency). The most successful interventions usually have effect sizes between –0.10 and –0.30. In their summary of program evaluations, Aos, Phipps, Barnoski, and Lieb (2001) reported that Multidimensional Treatment Foster Care (MTFC) reduced crime on average by 22%, which translated to an effect size of –0.37 given the other figures involved. Another well-known program, Multisystemic Therapy (MST), had an effect size of –0.31. Other programs with strong results have included nurse home visitation programs (–0.29), Functional Family Therapy (FFT) (–0.25), and the Seattle Social Development approach (–0.13) (Greenwood, 2006, p. 150; see also Greenwood & Turner, Chapter 23, this volume; Schiraldi, Schindler, & Goliday, Chapter 20, this volume).

Judging the evidence of delinquency interventions according to effect size alone, however, could also be inappropriate in some cases. Programs with lesser effect sizes may still be valuable. Some programs merit the label "evidence-based" because they generate a positive return on investment. A program that costs very little can be a worthwhile investment even if it has a relatively small effect size. For example, the Perry Preschool Project is considered highly successful by researchers despite its smaller effect size of −0.10 (Greenwood, 2006). Programs such as the Perry Preschool Project, of course, are the exception. For the most part, effect sizes for delinquency interventions need to fall in the range of –0.15 to –0.30 in order to have a real and lasting impact on policy and practice.

Cost is another, increasingly important component of research and evaluation on delinquency prevention programs. Government agencies and even private funders are beginning to require researchers to compare the costs and benefits of interventions. While studies of cost and benefits can take many forms, the usual approach is to first count all of the costs of program or policy inputs and then to develop a measure of program effectiveness. The final step is to translate effectiveness into a standardized measure of effect, usually by converting outcomes to dollars. The advantage of the approach is that comparisons can be made across different types of interventions. Critics argue that human behavior—particularly outcomes for youth—cannot be expressed in dollars and that a strictly economic approach ignores more basic values of justice and equity. Furthermore, estimating net benefits (benefits minus costs) or the ratio of costs to benefits creates an illusion of precision that belies the imprecise nature of evaluating human behavior. Nevertheless, evaluations that enumerate costs and benefits are increasingly common in crime prevention research.

EVALUATING POLICIES VERSUS PROGRAMS

Most evaluations are concerned with programs rather than policies, although the distinction can be elusive. By policy, we mean

any change in law, regulation or procedure that affects members of a group, community, or society. Lowering the upper age of juvenile court jurisdiction from 17 to 16, for example, would affect all 17- and 16-year-olds in a geographic area. A change in such a law would be considered a policy. Wholesale organizational reforms and system change efforts also generally fall within the domain of policy research. By contrast, programs target smaller or more defined populations. A decision to implement a juvenile drug court or to modify drug court procedures would affect just those youth admitted to drug court. It is often easier to evaluate programs than policies. Since the consequences of policy change are potentially more important than programmatic change, however, it is worth discussing the issues that surround policy evaluation.

Policy effects can be more difficult to evaluate for a variety of reasons. Policy questions are more likely to be politicized and value based and less amenable to facts and evidence. For example, the debate over determining the proper age for juvenile court jurisdiction has many factual elements, including cost and the relative effectiveness of the adult versus juvenile sanctions in reducing delinquency and promoting prosocial behavior. During an intense policy debate, however, beliefs can be more influential than facts. Those who believe that it is immoral and inhumane to treat juveniles as if they were adults are unlikely to be convinced by statistical evidence. Those who believe that youth offenders are no less culpable than their adult counterparts are similarly difficult to move. It is difficult to imagine an equally volatile debate over some aspect of juvenile probation practice.

Even when policy debates are more welcoming to research, the use of evaluation evidence can be difficult to conceptualize and implement. A thought experiment shows the difficulty. Suppose that a jurisdiction wants to implement a program that assigns youth to community-based treatment based on their risk factors and need for intervention. At the broadest level, all juveniles entering the system could be triaged with a screening tool that identifies substance disorders, mental health disorders, and other problems associated with delinquency. Youth with evidence of problems could be referred for more complete diagnostic evaluations and then referred to community-based interventions as appropriate. Deep-end services, such as residential treatment and other out-of-home placements would be reserved for youth who did not succeed in the less-restrictive environment of the community-based program.

Now, imagine that 1 year after the launch of the new process, local policy makers wanted to test the hypothesis that the juvenile justice system had become more efficient at assigning youth to appropriate intervention programs. How would researchers conduct such a test? Where would they find the needed data? One natural place to look would be in the rearrest data for all system-involved youth. The problem is that a greater rate of arrests overall could be either evidence of success (because a more efficient system would catch a higher percentage of youth violating the terms of their supervision) or evidence of failure (because the new process failed to produce better outcomes). The same problem would exist if an evaluation measured changes in the volume of treatment referrals, since they would be expected to increase initially but then decline as those youth expected to use the most services would be treated more effectively. Researchers could examine changing crime rates overall, but even a very strong program effect for some youth would probably not produce a perceptible change on the community's crime rate.

In addition to difficulties in conceptualizing appropriate outcomes, there is the problem of separating causation and correlation. What if an evaluation of a policy or systems change initiative found that crime declined? Was the decline due to the initiative or simply part of a broader trend? Such problems can be solved with a thoughtful, prospective evaluation that articulates the goals and objectives of the change effort, identifies the data necessary to measure the hypothesized change, and observes those data before and after implementation using an experimental design or a credible alternative. Too often, however, evaluations are started only after policy changes have been implemented, when some evidence of success has already appeared or evidence is needed urgently to support continued funding. In such a circumstance, it is nearly impossible to carry out a research effort that will produce credible results or generate real evidence of impact.

Some nontraditional evaluation approaches have appeared in recent years that avoid the pitfalls of retrospective analysis. One approach that has received a great deal of attention is the use *of instrumental variables*. Technically, these models solve the problem of endogeneity, where the outcome is at least a partial cause of the intervention. A classic example is the problem of measuring whether deploying more police lowers the crime rate. Neighborhoods with more crime are going to require more police protection, and thus it is difficult to determine the effect of more officers. The level of criminal offending may predict how many police officers are on patrol, not the other way around. Similar problems emerge on a host of important policy issues where the question is whether a policy change, such as lowering the age of juvenile jurisdiction or increasing the number and length of juvenile commitments, causes or reflects a change in juvenile offending.

Recently, economists have conducted several studies that try to solve these problems using instrumental variables. Researchers identify a policy that is highly correlated with a particular outcome, but that only affects that outcome through its relationship with another factor. Levitt published a series of articles using this approach to test the effects of prison size on crime rates (1996) and policing (1997; but see McCrary, 2002) on crime, and found strong effects for both. In the prison case, Levitt used prison overcrowding orders (i.e., court injunctions in response to excess populations) to approximate prison population size on the assumption that prison overcrowding orders affect prison populations, and thus crimes rates, but are not directly affected by the crime rate. While the approach is lauded for its creativity, other researchers cautioned against accepting the study's conclusion that each year of prison time served prevents 15 serious crimes. Useem, Piehl, and Liedka (2001) noted that the study's conclusion, "depends on a hypothesized symmetry in causal processes between an unusual set of circumstances and a usual set of circumstances" (p. 6). We discuss these issues not to criticize a particular study, but rather to point out the seriousness of the challenges researchers face in disaggregating the effects of policies on outcomes in the very common situation where the same outcomes are at least partly responsible for changes in policy.

A second common problem in estimating the effects of crime policies is estimating whether increased enforcement leads to less crime if greater enforcement increases the probability of arrest for crimes that were previously not often reported to police. Determining the effects of added police is confounded by the reality that police not only respond to calls for service; they initiate investigations. For certain kinds of crime, such

as prostitution and drug offending, more police personnel should lead to more crimes being reported. In the context of juvenile justice, a researcher measuring the effects of added enforcement of juvenile curfew laws would face a similar problem. More arrests would not necessarily indicate an increase in violations, but could indicate an increase in enforcement. In fact, since the true number of juvenile offenders can never be known (because the age of offenders who are not caught cannot be determined) it is generally impossible to determine whether an increase in the number of juvenile arrests is an indicator of an increase in juvenile crime, or merely a by-product of a change in enforcement practice. As Blumstein, Cohen, Roth, and Visher (1986) noted, "arrest records can be used to infer the volume of unobserved crimes committed" (p. 99), but a wealth of additional information is needed to make that inference. Rates of arrest are potentially important indicators of change in juvenile offending, but the relationship is only a correlation and often of unknown strength. Increases in juvenile arrests rates in the early 1990s caused many communities to implement draconian policies that increased the likelihood that juvenile arrestees would be prosecuted in the adult criminal justice system (Snyder & Sickmund, 1999). However, it is far from clear that the deleterious effects of those changes were outweighed by decreases in crime. In fact, the policies may have helped to spur additional crimes among those youth affected (McGowan et al., 2007).

A rather infamous example of the serious consequences that may arise from confusing the causes and correlates of crime is the prediction in the 1990s that a new population of youthful "superpredators" was about to savage American communities (Dilulio, 1995). The prediction was based on two trends that were emerging in the early 1990s and that together predicted an explosion in crime. John Dilulio warned Americans that they were sitting atop a "demographic crime bomb" with far more boys in the population under ten years of age than in previous decades. He argued that youth in general were becoming more violent, referring to evidence of increasing juvenile arrests for violent crime. Today, as the youth from that cohort of potential "superpredators" enter their mid-20s, youth crime rates are not only lower than they were in 1995, but lower than they have been at any time since the 1970s (Puzzanchera, 2009).

One final challenge exists in thinking about evaluating systems changes. The goal of youth serving agencies is to improve the life conditions and well-being of youth and communities. These agencies pay the costs of intervention, but the benefits of their efforts are more diffuse. The main beneficiaries are youth and families in the larger community. Problems emerge when agencies have implemented successful, systems-level changes and then seek funding to sustain those changes. While agencies can point to cost savings from efficiencies gained in delivering services, many of the benefits from large-scale change accrue outside the system and may be of little or no interest to those who determine the size of agency budgets. The increasing use of business models for budgeting social programs means that governments are increasingly interested in funding programs based on their "return on investment." The conceptualization of investment returns, however, does not extend very far beyond the agency's administrative costs.

It is reasonable at this point to inquire about the causes of what appears to be the sad state of research on juvenile justice policies and programs (see also Jacobs, Miranda-Julian, & Kaplan, Chapter 10, this volume). The simple explanation is that data collection and analysis are low priorities in most justice systems.

Data are rarely available to be examined at the individual or case level, and thus researchers are forced to work with data that simply are not up to the task, which leads to incorrect inferences (such as the "superpredator" forecast). More insidiously, the lack of data collection also suggests that these systems are not expending much energy investigating whether what they are doing is actually working, or even if they are doing the things they claim to be doing. A broad effort to collect better data would improve our capacity to test whether policy changes had the intended effect by providing a ready comparison that describes systems before they change. The best data systems would integrate information from a variety of ancillary service providers (mentor agencies, job placement assistance, health-care providers, etc.), other actors in the adult or justice system (Family Court) as well as the school system (see Schneider & Simpson, Chapter 22, this volume). Typically, the youth-serving agencies in any one community are concerned with the issue they are responsible for, but no one in the system is responsible for the whole child or the whole family. Data systems are similarly stove-piped and a complete picture of a juvenile's interaction with the larger service system is difficult, if not impossible to obtain. Thus, neither system insiders nor any outside researchers will be able to determine precisely which interventions are most effective.

We would recommend, at minimum, that agencies seeking to implement large-scale systems change think about the types of activities in which they expect to engage—before implementation—and then create a data system that can allow the agency to determine if the program is delivering the services it promises. Jurisdictions seeking to implement such a model need to measure not only whether juveniles are screened, assessed, and directed to the appropriate level of care, but also that accurate information about case processing is shared among all relevant decision makers. Each treatment agency in the larger system should measure not only how many youth assessments they conduct, for instance, but also how many assessments lead to referrals, how many referrals lead to successful treatment, how much treatment is actually delivered, and which individuals and agencies are involved in each point of service. More complete data would allow agencies to convince themselves and their stakeholders that they are achieving their stated objectives.

MAKING RESEARCH ACCESSIBLE

Communication is essential for good research. Research has little effect unless it reaches an audience that is motivated and prepared to receive it, and unless the members of that audience are in a position to use the lessons of research to improve the well-being of youth. There are always multiple mechanisms for sharing the findings of research, depending on the type of research involved and the specific audience (see Table 24.2). Unfortunately, some researchers view their primary audience as other researchers. Some of the best research on social policies and programs, in fact, is designed merely to advance the understanding of researchers rather than to change social institutions, policy, or practice. Especially among academic researchers, study findings are often disseminated using the insider's language of theoretical reasoning and incremental advances in understanding. Researchers can be very poor communicators, not always for lack of interest but because their training and professional culture encourages them to restrict the manner in which they communicate.

Science, including social science, requires objective scholarship. Throughout their

Table 24.2. Writing for Impact

Regardless of who it is that translates the technical language of researchers into accessible, nontechnical forms, someone must do it. It is simply unacceptable when the information and materials used to shape public policy remain inaccessible to nontechnical audiences, including the general public. We offer a few guiding principles for pursuing this task:

■ **There is no excuse for bad writing.**
By the time they have finished their professional training, researchers have consumed so much bad writing that they are actually impressed by bad writing, mistaking complexity for importance. Technical information does not have to be presented badly. It may be too late for many researchers to learn how to write plainly, but they should hire others to do it for them.

■ **Writing is teaching.**
The act of writing about research findings is akin to teaching, or at least it should be. Especially when findings are derived from complex research designs, and when they require elaborate statistical analysis, the results need to be presented in a way that actually communicates the findings to the audience. Every former student can probably remember a math or statistics teacher who stood in front of the room, facing away from the class, writing proofs on the board, as if that were teaching. It is not, and that style of teaching cannot be used to create research reports and articles. Again, if a researcher is incapable of making results accessible, someone who can do that must be added to the research team.

■ **Graphics deliver the message.**
People often learn more efficiently from visual information. Data graphics are essential and should be designed very carefully. They should be created with both accuracy and ease of interpretation in mind. Ideally, the complete message of a research report should be discernible by perusing the data graphics alone. Data tables are also crucial, but they should communicate effectively to the audience. Tables are not there simply to impress the nontechnical reader with small font sizes and Greek letters.

■ **If a picture is worth a thousand words, you don't need a thousand pictures.**
Data graphics are an essential part of communicating research findings, but more is not always better. Graphics should be reserved for conveying the central points of the research and for analytical messages that are complex and multidimensional. They should not be wasted on simple, descriptive tasks.

professional training, researchers are encouraged to view their work as the disinterested pursuit of knowledge. They are trained to gather evidence, analyze it, and report it as if they have no stake in its application. The production of knowledge is supposed to be their only goal. The application of knowledge is either secondary or entirely beyond their concern. One of the worst charges one researcher can make against another is to call that person an "advocate." Any evidence that a researcher is promoting a particular policy or program suggests that her/his research may not be credible, that he/she may be attempting to tip the scale toward one conclusion versus another. Researchers very much want their efforts to have an impact on social policy, but their occupational culture prohibits them from pursuing such an impact directly. This is especially true when research must be described to general audiences using nontechnical, straightforward language. Researchers worry that efforts to simplify their findings will gloss over important details and their carefully nuanced conclusions. Researchers seeking to avoid efforts to clarify their communication style often defend their own intransigence by saying that they refuse to "dumb down" their materials or their writing.

Certainly, most investigators do crave the recognition and sense of relevance that comes from wide dissemination and application of their study findings. At academic conferences and meetings, researchers often discuss the intersection of their work with the worlds of policy and politics. They may complain their work is misunderstood or disregarded by decision makers and elected officials. They may even lament the sloppy formulation of government policies that fail to apply sound knowledge in solving social problems. But, ultimately, their fear of losing reputational

capital among other researchers prevents them from investing their time and talents in translating the findings of their studies for a general audience.

Making a personal effort to bridge science and policy would be distasteful for many researchers. Within the community of researchers, a person who uses research evidence in an explicit attempt to influence public thinking and public policy is not a researcher, but a politician or an activist, even if the person was trained in research methods. Achieving influence and impact cannot be a researcher's first priority. Of course, every researcher (if only secretly) dreams of seeing his or her work have real impact on public debate and public policy. It is a question of method and intent. Within the culture of academia, it is acceptable and even laudable to be invited to inform a policy debate by describing the findings of research to key officials and other elite audiences. It is not culturally acceptable, however, to set a course on influencing debate and to seek out opportunities to exercise such influence, whether invited or not. All researchers enjoy being asked for their opinions on important matters of policy, but to design one's research explicitly to affect policy is an affront to academic culture. For this reason, there must be intermediaries between the spheres of research, policy formulation, and program implementation.

PROTECTING RESEARCH FROM SPONSORS AND CONSUMERS

Social programs and social policies would be improved by making research accessible outside of the technically oriented audience of researchers. Translating the findings of research for consumption by policy makers and practitioners would likely help them to incorporate the best knowledge about program effectiveness into the conceptualization and design of social interventions. What about the reverse? Are there reasons to be concerned about the possible impact of a stronger user orientation on research itself? Unfortunately, the answer is "yes."

Some researchers, especially those from the social sciences, are naturally resistant to viewing the goal of their efforts as the application and use of research findings. In some disciplines, good research practice requires a clear separation between the development and application of knowledge. Being too "user minded," some might worry, could lead researchers to change the questions they ask and the methods they use to find answers to questions. Especially in a competitive funding environment, it is instinctive to shape one's research to meet the needs and expectations of funding agencies. Modifying the methods and tactics of research for competitive reasons could easily undermine the quality of the effort and mislead policy makers and practitioners into following evidence that is less than sound.

A recent example can be found in the evaluation field. Researchers in the United States have been working for more than 10 years to assess the efficacy of juvenile drug courts, or juvenile drug treatment courts (e.g., Barnes, Miller, Miller, & Gibson, 2008; Belenko & Dembo, 2003; Butts & Roman, 2004; Hiller et al., 2010). Juvenile drug courts use a potentially persuasive combination of judicial authority and interorganizational coordination to motivate drug-involved offenders to stay in treatment and change their behavior. The courts use case management to coordinate services, drug tests to monitor offender compliance, and frequent court hearings to review case progress and establish effective social bonds between offenders, judges, and other court staff. Adult drug courts

started 20 years ago and are becoming a permanent part of the American justice system. In part, this is the work of advocacy organizations that promote drug courts, train drug court officials, publish reports extolling the virtues of drug courts, and work with news media to increase public awareness of drug courts. In their enthusiasm, drug court advocates have not always been sufficiently cautious in interpreting research evidence about drug courts, but this is as it should be. New programs need the support of advocates whose passions are impervious to empirical scrutiny.

To design effective programs, however, practitioners must put aside their passions and rely on evidence, especially evidence about the effectiveness of specific components of drug courts. Yet, many juvenile drug court evaluations in recent years have not been designed to discover effective program components. They have been designed to confirm the beliefs of practitioners and policy makers, especially those of the judges who operate drug courts and the federal agencies that support drug court programs. In the late 1990s, several practitioner groups made assertions about what they believed to be the "key components" of drug courts (see Table 24.3). These components were published in a list, even though the information was not based on any evaluation evidence or even on an articulated theory of program impact (Drug Courts Program Office, 1997). The list was simply the compiled judgment of drug court practitioners.

Scores of formative evaluation studies were funded during the next 10 years to hold drug court programs up against the standards set by the list of key components. Researchers were directed to examine the operations of drug courts and to determine whether they did or did not exhibit fidelity to the 10 key components. Because the list of components was not derived from or consistent with any research literature, however, there was no way to judge the importance of these findings. Significant amounts of state and federal resources were expended on what amounted to an extensive series of program audits that did not advance program development or even an empirical understanding of program effects.

Finally, another recent development demonstrates a different risk of mandating the connection between research, policy, and practice. In nearly all facets of social policy,

Table 24.3. 10 Key Components of Drug Courts as Identified by the U.S. Department of Justice

1.	The drug court integrates alcohol and other drug treatment services with justice system case processing.
2.	Using a nonadversarial approach, prosecution and defense counsel promote public safety while protecting participants' due process rights.
3.	Eligible participants are identified early and promptly placed in the drug court program.
4.	The drug court provides access to a continuum of alcohol, drug, and other related treatment and rehabilitative services.
5.	Abstinence is monitored by frequent alcohol and other drug testing.
6.	A coordinated strategy governs drug court responses to participants' compliance.
7.	Ongoing judicial interaction with each drug court participant is essential.
8.	Monitoring and evaluation measure the achievement of program goals and gauge effectiveness.
9.	Continuing interdisciplinary education promotes effective drug court planning, implementation, and operations.
10.	Forging partnerships among drug courts, public agencies, and community-based organizations generates local support and enhances drug court program effectiveness.

Source: Drug Courts Program Office, 1997.

policy makers and practitioners are becoming more enthusiastic about evidence-based policy and practice (e.g., see the work of the Coalition for Evidence-Based Policy in the United States and the Centre for Evidence Based Policy in the United Kingdom). The concepts of evidence-based policy and practice (EBPP) suggest that research findings should be used to weigh the desirability of social interventions and that research evidence should inform decisions to support one intervention model over another. In the best of all possible worlds, where research investment would follow innovation naturally and without bias or prejudice, following a strict EBPP regimen would be a sensible idea. Using research evidence to shape the design and implementation of social programs would make social policies more effective and result in improved social conditions. In our less-than-perfect world, however, research funding is intensely competitive and policy decisions are subject to political wrangling and the self-interests of policy makers. In our environment, a restrictive EBPP approach could stifle innovation and maintain unwanted sectarian control over policies and programs (Schorr, 2009).

For example, state and federal agencies in the United States are beginning to require that services for adolescent offenders be evidence based. In the field of crime prevention, however, very few interventions can make such a claim. While there are several early childhood programs that might survive the evidence test, including nurse home visitation programs and educational support programs such as Head Start, few programs for older youth have come close to reaching the status of "proven." Two such programs are Functional Family Therapy (FFT) and Multisystemic Therapy (MST). Is this because MST and FFT are the best possible interventions to prevent and reduce delinquency? For a small minority of offenders, this could indeed be true. For the vast majority of youth, however, it is certainly not true. Other interventions, even less expensive interventions, could be effective for many youth, but in a strict EBPP environment, untested programs are less likely to attract the significant investments required to generate high-quality evidence. The resources necessary to identify and disseminate high-quality evidence are limited, and untested programs are not likely to attract the funding necessary to prove their effectiveness if government agencies and private providers are locked into an evidence regime based on pre-existing research. In this way, the simple-minded adherence to evidence-based policy could actually be detrimental to the quality of programs and policies.

RESEARCH VERSUS QUALITY IMPROVEMENT

After decades of research on hundreds of programs and policy initiatives for justice-involved youth, it seems that we have relatively few proven approaches. One reason for this rather slow, haphazard rate of progress is that we spend much of the total pool of research resources on small, more limited studies. Smaller studies can be helpful in measuring the delivery of services or the implementation of new procedures, but they do not often involve random assignment and they cannot produce the experimental results necessary to earn that coveted spot on some future list of evidence-based interventions.

Rather than continuing to expend our resources on a multitude of small, inadequate, and potentially redundant evaluations, a new research paradigm may be required. Policy makers could derive greater benefit from a few high-quality impact evaluations that could

be used to establish operational principles for program implementation and a program of research that monitored quality and client outcomes in a more coordinated fashion.

The research funds currently used for numerous small studies could be combined and concentrated on a few high-profile evaluations targeted on very specific research questions.

There is a limit to the number of quality studies that can be funded and completed. The number of competent researchers available to do the work is finite, and given shortages in time, expertise, and funding, even the best studies can answer only a few key questions at a time. It is not enough simply to generate an ever-larger number of studies. This strategy has resulted in a glut of poorly conceived and underfunded evaluations, none of them capable of resolving key disputes about program design and policy reform. Rather than continuing to divide the pool of evaluation resources across an uncoordinated array of small studies, policy makers may want to concentrate research funds on a few, well-designed, theoretically oriented investigations of program impact and cost effectiveness.

The juvenile justice field could look beyond individual evaluations and start to build a system of research-based program accreditation. An effective accreditation process could relieve individual programs of the burdens imposed by constant evaluation. Standards could be set at the state or national level, and individual programs could demonstrate whether they meet the standards. Programs meeting their standards would then receive a credential. If the standards were carefully developed, accredited youth justice programs could use their credentials to demonstrate their competence and effectiveness to local stakeholders. Such a system would free resources for client services and program management that are currently diverted to inadequate and redundant evaluation studies. Policy and practice could be informed by a smaller number of carefully designed studies as long as those studies were used to inform a process of accreditation.

An accreditation process for youth justice would have to be sophisticated. It could not be simply an extension of "best practices," in which program design principles are extrapolated from the opinions and beliefs of practitioners. A rigorous system of accreditation would have to be empirically oriented and based on a foundation of solid evaluation findings. It would likely be years before the justice system was capable of implementing an effective accreditation regime, but the design and development work could begin immediately. Once established, this system of accreditation could profoundly affect juvenile justice programs and the policy environment in which they operate. The strength of the evaluation literature underlying accreditation would put pressure on programs to forego anecdotal evidence and personal preference in designing their procedures. If evaluators found, for example, that consistency in school attendance was associated with decreased probation failures, the accreditation process would encourage juvenile justice officials to work more closely with the schools. Accreditation would help to insulate individual programs against political attack. Currently, a state or local official who wishes to redirect funds away from a particular program merely has to challenge that program to produce evidence of its effectiveness. Unless the program is fortunate enough to have findings from a recent and high-quality study, its funding is vulnerable. A respected accreditation process would allow programs to demonstrate their value using outcome data from other jurisdictions without constantly funding their own small and insufficient evaluations.

An accreditation-focused evaluation agenda would also free researchers to investigate critical questions about program effectiveness. Currently, a significant amount of research is undertaken not to illuminate the elements of program effectiveness, but to fulfill funding requirements and provide support for future funding requests. Studies conducted under these circumstances are rarely carried out with a true spirit of discovery. Investigators cannot afford the luxury of building an evidence base with precise measurements of a few key program components at a time. They must develop broad indicators of effectiveness, and they must do it as quickly as possible. An accreditation process would allow evaluation researchers to focus on measuring the fundamental components of program effectiveness. An accreditation system could standardize the scheduling and delivery of services and allow practitioners to focus on quality. They could work to develop expanded prevention and intervention programs for juvenile offenders. The accreditation agency could help in answering questions about the effectiveness of program components. It would no longer be necessary for every program manager and every researcher to reinvent the wheel each time a question of effectiveness arose. The accreditation entity would identify, synthesize, and disseminate authoritative research that would be trusted for policy and program development. Practitioners could focus on building programs knowing their models and frameworks were sound.

CONCLUSION

Most textbooks for training social science researchers describe a similar set of ingredients needed to cook up effective policy research. According to most recipes, good research starts with complete data on all individuals in all justice agencies and any relevant nonjustice agencies, often including social services, drug treatment, housing and employment. An effective research design maximizes the size of the population to which the study results can be generalized but also minimizes the chances of an erroneous conclusion. Randomized, controlled trials are preferred; matched controls are acceptable. Natural experiments are generally frowned upon. Prospective designs are preferred over retrospective designs. Long follow-up periods with multiple observations are optimal. Finally, to be effective in shaping new policy, study results should be disseminated in a way that conveys rigor to the expert but provides clarity to policy makers and to members of the public.

Unfortunately, few studies in the youth justice field achieve many of those goals in practice. If all research studies were required to meet each of these standards before being approved for funding, many—if not most—important research questions would go unanswered. The data available to researchers are usually incomplete. Youth justice agencies rarely collect just the right kind of information needed for a particular study. Most researchers have to work with agency officials to generate new data. Next, they have to choose between selecting a research sample that will allow their study to observe a true answer, versus a sample where the study results can be generalized to a broader population. Randomized controlled trials produce rigorous and generalizable results, but they are not feasible in most circumstances. Even when they are feasible, by definition they answer very narrow questions. Natural experiments are the best way to look at broad policy changes, but they fail to control for competing explanations of change or impact. Prospective studies take too long for all but the most patient stakeholders, and thus retrospective studies are usually a more practical alternative. Finally, few writers trained in

the social sciences are capable of preparing a research report that is accessible enough for the average citizen yet precise enough for an expert audience.

Our recommendations for designing and implementing real-world policy research and program evaluation are quite different from those that would be favored in controlled, laboratory settings. In policy research, it may be more important to answer a useful question with less precision than it is to test a narrow hypothesis with great accuracy. Researchers often have to take whatever data they can get, including information from other locations and from other time frames, and even less than rigorous sources, such as interviews, focus groups, and direct observation. Researchers should focus on the comparability of treatment and control groups, which means caring equally about the quality of a study design *and* how well the study withstands real world inconveniences of lost data points and skewed information. Studies should be designed to create equivalent samples, but also to produce generalizable results knowing that compromises will still have to be made on both objectives. For broader policy questions, natural experiments with pre-post designs are often more informative than well-designed studies that address a more limited range of policy concerns. Study results must be communicated clearly and concisely, respecting clarity as much as precision, and favoring transparency above all else. And finally, all researchers should avoid the missteps that inevitably follow from the presumption that whatever situation they are studying is completely different from anything that others have studied. Every research project is special, but few are truly unique. Social science research will never reach the levels of precision associated with laboratory studies, but such precision is not always necessary to improve youth-serving systems. Thoughtful research that acknowledges but does not succumb to real world obstacles can lead to substantial improvements in policy and practice.

REFERENCES

Aos, S., Phipps, P., Barnoski, R. & Lieb, R. (2001). *The comparative costs and benefits of programs to reduce crime* (Version 4.0). Olympia: Washington State Institute for Public Policy.

Barnes, J. C., Miller, J. M., Miller, H. V., & Gibson, C. (2008). Juvenile drug court program admission, demeanor and cherry-picking: A research note. *American Journal of Criminal Justice*, *33*(2), 166–176.

Belenko, S., & Dembo, R. (2003). Treating adolescent substance abuse problems in the juvenile drug court. *International Journal of Law and Psychiatry*, *26*(1), 87–110.

Blumstein, A., Cohen, J., Roth, J. A., & Visher, C. A. (Eds.). (1986). *Criminal careers and career criminals* (National Research Council, Panel on Research on Criminal Careers, Volume 1). Washington, DC: National Academy Press.

Boyle, P. (2001). A DAREing rescue: How intervention by critics and federal officials brought the youth anti-drug program into rehab. *Youth today*. Washington, DC: American Youth Work Center.

Brewer, G. D., & deLeon, P. (1983). *The foundations of policy analysis*. Homewood, IL: Dorsey Press.

Browning, C. R., Feinberg, S. L., & Dietz, R. D. (2004). The paradox of social organization: Networks, collective efficacy, and violent crime in urban neighborhoods. *Social Forces*, *83*(2), 503–534.

Butts, J. A., & Roman, J. (Eds.). (2004). *Juvenile drug courts and teen substance abuse*. Washington, DC: Urban Institute Press.

Casey, B. J., Tottenham, N., Liston, C., & Durston, S. (2005). Imaging the developing brain: What have we learned about cognitive development? *Trends in Cognitive Sciences*, *9*(3), 104–110.

Coleman, J. (1990). *Foundations of social theory*. Cambridge, MA: Harvard University Press.

Dilulio, J. J. (1995). The coming of the super-predators. *Weekly Standard*, *1*(11), 23.

Drug Courts Program Office. (1997). *Defining drug courts: The key components*. Washington, DC: U.S. Department of Justice, Office of Justice Programs, Drug Courts Program Office.

Eckenrode, J., Campa, M., Luckey, D. W., Henderson C. R., Jr., Cole, R., Kitzman, H., . . . Olds, D. L.

(2010). Long-term effects of prenatal and infancy nurse home visitation on the life course of youths: 19-year follow-up of a randomized trial. *Archives of Pediatric Adolescent Medicine*, *164*, 9–15.

Ennett, S. T., Tobler, N. S., Ringwalt, C. L., & Flewelling, R. L. (1994). How effective is Drug Abuse Resistance Education? A meta-analysis of Project DARE outcome evaluations. *American Journal of Public Health*, *84*(9), 1394–1401.

Frisco, V., & Stalebrink, O. J. (2008). Congressional use of the Program Assessment Rating Tool. *Public Budgeting & Finance*, *28*(2), 1–19.

Greenwood, P. W. (2006). *Changing lives—delinquency prevention as crime-control policy*. Chicago, IL: University of Chicago Press.

Hiller, M. L., Malluche, D., Bryan, V., DuPont, M. L., Martin, B., Abensur, R., . . . Payne, C. (2010). A multisite description of juvenile drug courts. *International Journal of Offender Therapy and Comparative Criminology*, *54*(2), 213–235.

Levitt, S. D. (1996). The effect of prison population size on crime rates: Evidence from prison overcrowding litigation. *Quarterly Journal of Economics*, *111*(2), 319–351.

Levitt, S. D. (1997). Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review*, *87*(3), 270–290.

Mazerolle, L., Wickes, R., & McBroom, J. (2010). Community variations in violence: The role of social ties and collective efficacy in comparative context. *Journal of Research in Crime and Delinquency*, *47*(1), 3–30.

McCrary, J. (2002). Using electoral cycles in police hiring to estimate the effect of police on crime: Comment. *American Economic Review*, *92*(4), 1236–1243.

McGowan, A., Hahn, R., Liberman, A., Crosby, A., Fullilove, M., Johnson, R., . . . Stone, G. (2007). Effects on violence of laws and policies facilitating the transfer of juveniles from the juvenile justice system to the adult justice system: A systematic review. *American Journal of Preventive Medicine*, *32* (4S), S7–S28.

Olds, D. L., Henderson, C. R., Jr., Tatelbaum, R., & Chamberlin, R. (1988). Improving the life-course development of socially disadvantaged mothers: A randomized trial of nurse home visitation. *American Journal of Public Health*, *78*(11), 1436–1445.

Olds, D. L., Robinson, J., Pettitt, L. M., Luckey, D. W., Holmberg, J., Ng, R. K., . . . Henderson, C. R. Jr. (2004). Effects of home visits by paraprofessionals and by nurses: Age four follow-up results of a randomized trial. *Pediatrics*, *114*(6), 1560–1568.

Olds, D. L., Kitzman, H., Hanks, C., Cole, R., Anson, E., Sidora-Arcoleo, K., . . . Bondy, J. (2007). Effects of nurse home visiting on maternal and child functioning: Age nine follow-up of a randomized trial. *Pediatrics*, *120*, e832–e845.

Puzzanchera, C. (2009). Juvenile arrests 2007. *Juvenile justice bulletin*. Washington, DC: U.S. Department of Justice, Office of Juvenile Justice and Delinquency Prevention [NCJ 225344].

Roman, C., & Chalfin, A. J. (2008). Fear of walking outdoors: A multilevel ecologic analysis of crime and disorder. *American Journal of Preventive Medicine*, *34*, 306–312.

Sampson, R. J., Morenoff, J. D., & Earls, F. (1999). Beyond social capital: Spatial dynamics of collective efficacy for children. *American Sociological Review*, *64* (5), 633–660.

Schorr, L. B. (2009, August 20). To judge what will best help society's neediest, let's use a broad array of evaluation techniques. *Chronicle of Philanthropy*.

Scott, E. S., & Steinberg, L. (2008). *Rethinking juvenile justice*. Cambridge, MA: Harvard University Press.

Snyder, H., & Sickmund, M. (1999). *Juvenile offenders and victims: 1999 national report*. Washington, DC: Office of Juvenile Justice and Delinquency Prevention.

Useem, B., Piehl, A. M., & Liedka, R. V. (2001). *Crime-control effect of incarceration: Reconsidering the evidence*. Washington, DC: National Institute of Justice. Document No. 188265.