ANNALS, *AAPSS*, **578**, November 2001

# Does Research Design Affect Study Outcomes in Criminal Justice?

*By* DAVID WEISBURD, CYNTHIA M. LUM, and ANTHONY PETROSINO

ABSTRACT: Does the type of research design used in a crime and justice study influence its conclusions? Scholars agree in theory that randomized experimental studies have higher internal validity than do nonrandomized studies. But there is not consensus regarding the costs of using nonrandomized studies in coming to conclusions regarding criminal justice interventions. To examine these issues, the authors look at the relationship between research design and study outcomes in a broad review of research evidence on crime and justice commissioned by the National Institute of Justice. Their findings suggest that design does have a systematic effect on outcomes in criminal justice studies. The weaker a design, indicated by internal validity, the more likely a study is to report a result in favor of treatment and the less likely it is to report a harmful effect of treatment. Even when comparing randomized studies with strong quasi-experimental research designs, systematic and statistically significant differences are observed.

---

*David Weisburd is a senior research fellow in the Department of Criminology and Criminal Justice at the University of Maryland and a professor of criminology at the Hebrew University Law School in Jerusalem.*

*Cynthia M. Lum is a doctoral student in the Department of Criminology and Criminal Justice at the University of Maryland.*

*Anthony Petrosino is a research fellow at the Center for Evaluation, Initiative for Children Program at the American Academy of Arts and Sciences and a research associate at Harvard University. He is also the coordinator of the Campbell Crime and Justice Coordinating Group.*

NOTE: We are indebted to a number of colleagues for helpful comments in preparing this article. We especially want to thank Iain Chalmers, John Eck, David Farrington, Denise Gottfredson, Doris MacKenzie, Joan McCord, Lawrence Sherman, Brandon Welsh, Charles Wellford, and David Wilson.

THERE is a growing consensus among scholars, practitioners, and policy makers that crime control practices and policies should be rooted as much as possible in scientific research (Cullen and Gendreau 2000; MacKenzie 2000; Sherman 1998). This is reflected in the steady growth in interest in evaluation of criminal justice programs and practices in the United States and the United Kingdom over the past decade and by large increases in criminal justice funding for research during this period (Visher and Weisburd 1998). Increasing support for research and evaluation in criminal justice may be seen as part of a more general trend toward utilization of scientific research for establishing rational and effective practices and policies. This trend is perhaps most prominent in the health professions, where the idea of evidence-based medicine has gained strong government and professional support (Millenson 1997; Zuger 1997), though the evidence-based paradigm is also developing in other fields (see Nutley and Davies 1999; Davies, Nutley, and Smith 2000).

A central component of the movement toward evidence-based practice and policy is reliance on systematic review of prior research and evaluation (Davies 1999). Such review allows policy makers and practitioners to identify what programs and practices are most effective and in which contexts. The Cochrane Collaboration, for example, seeks to prepare, maintain, and make accessible systematic reviews of research on the effects of health care interventions (see Chalmers and Altman 1995; www.cochrane.org.) The *Cochrane Library* is now widely recognized as the single best source of evidence on the effectiveness of health care and medical treatments and has played an important part in the advancement of evidence-based medicine (Egger and Smith 1998). More recently, social scientists following the Cochrane model established the Campbell Collaboration for developing systematic reviews of research evidence in the area of social and educational interventions (see Boruch, Petrosino, and Chalmers 1999). In recognition of the growing importance of evidence-based policies in criminal justice, the Campbell Collaboration commissioned a coordinating group to deal with crime and justice issues. This group began with the goal of providing the best evidence on "what works in crime and justice" through the development of "systematic reviews of research" on the effects of crime and justice interventions (Farrington and Petrosino 2001 [this issue]).

In the Cochrane Collaboration, and in medical research in general, clinical trials that randomize participants to treatment and control or comparison groups are considered more reliable than studies that do not employ randomization. And the recognition that experimental designs form the gold standard for drawing conclusions about the effects of treatments or programs is not restricted to medicine. There is broad agreement among social and behavioral scientists that randomized experiments provide the best method for drawing causal inferences between treatments and

programs and their outcomes (for example, see Boruch, Snyder, and DeMoya 2000; Campbell and Boruch 1975; Farrington 1983; Feder, Jolin, and Feyerherm 2000). Indeed, a task force convened by the Board of Scientific Affairs of the American Psychological Association to look into statistical methods concluded that "for research involving causal inferences, the assignments of units to levels of the causal variable is critical. Random assignment (not to be confused with random selection) allows for the strongest possible causal inferences free of extraneous assumptions" (Wilkinson and Task Force on Statistical Inference 1999).

While reliance on experimental studies in drawing conclusions about treatment outcomes has become common in the development of evidence-based medicine, the Campbell Collaboration Crime and Justice Coordinating Group has concluded that it is unrealistic at this time to restrict systematic reviews on the effects of interventions relevant to crime and justice to experimental studies. In developing its *Standards for Inclusion of Studies in Systematic Reviews* (Farrington 2000), the group notes that it does not require that reviewers select only randomized experiments:

> This might possibly be the case for an intervention where there are many randomized experiments (e.g. cognitive-behavioral skills training). However, randomized experiments to evaluate criminological interventions are relatively uncommon. If reviews were restricted to randomized experiments, they would be relevant to only a small fraction of the key questions for policy and practice in criminology. Where there are few randomized experiments, it is expected that reviewers will select both randomized and non-randomized studies for inclusion in detailed reviews. (3)

In this article we examine a central question relevant both to the Campbell Collaboration crime and justice effort and to the more general emphasis on developing evidence-based practice in criminal justice: Does the type of research design used in a crime and justice study influence the conclusions that are reached? Assuming that experimental designs are the gold standard for evaluating practices and policies, it is important to ask what price we pay in including other types of studies in our reviews of what works in crime and justice. Are we likely to overestimate or underestimate the positive effects of treatment? Or conversely, might we expect that the use of well-designed nonrandomized studies will lead to about the same conclusions as we would gain from randomized experimental evaluations?

To examine these issues, we look at the relationship between research design and study outcomes in a broad review of research evidence on crime and justice commissioned by the National Institute of Justice. Generally referred to as the Maryland Report because it was developed in the Department of Criminology and Criminal Justice at the University of Maryland at College Park, the study was published under the title *Preventing Crime: What Works, What*

*Doesn't, What's Promising* (Sherman et al. 1997). The Maryland Report provides an unusual opportunity for assessing the impact of study design on study outcomes in crime and justice both because it sought to be comprehensive in identifying available research and because the principal investigators of the study devoted specific attention to the nature of the research designs of the studies included. Below we detail the methods we used to examine how study design affects study outcomes in crime and justice research and report on our main findings. We turn first, however, to a discussion of why randomized experiments as contrasted with quasi-experimental and nonexperimental research designs are generally considered a gold standard for making causal inferences. We also examine what prior research suggests regarding the questions we raise.

## WHY ARE RANDOMIZED EXPERIMENTS CONSIDERED THE GOLD STANDARD?

The key to understanding the strength of experimental research designs is found in what scholars refer to as the internal validity of a study. A research design in which the effects of treatment or intervention can be clearly distinguished from other effects has high internal validity. A research design in which the effects of treatment are confounded with other factors is one in which there is low internal validity. For example, suppose a researcher seeks to assess the effects of a specific drug treatment program on recidivism. If at the end of the evaluation the researcher can present study results and confidently assert that the effects of treatment have been isolated from other confounding causes, the internal validity of the study is high. But if the researcher has been unable to ensure that other factors such as the seriousness of prior records or the social status of offenders have been disentangled from the influence of treatment, he or she must note that the effects observed for treatment may be due to such confounding causes. In this case internal validity is low.

In randomized experimental studies, internal validity is developed through the process of random allocation of the units of treatment or intervention to experimental and control or comparison groups. This means that the researcher has randomized other factors besides treatment itself, since there is no systematic bias that brings one type of subject into the treatment group and another into the control or comparison group. Although the groups are not necessarily the same on every characteristic—indeed, simply by chance, there are likely to be differences—such differences can be assumed to be distributed randomly and are part and parcel of the stochastic processes taken into account in statistical tests. Random allocation thus allows the researcher to assume that the only systematic differences between the treatment and comparison groups are found in the treatments or interventions that are applied. When the study is complete,

the researcher can argue with confidence that if a difference has been observed between treatment and comparison groups, it is likely the result of the treatment itself (since randomization has isolated the treatment effect from other possible causes).

In nonrandomized studies, two methods may be used for isolating treatment or program effects. Quasi-experiments, like randomized experiments, rely on the design of a research study to isolate the effects of treatment. Using matching or other methods in an attempt to establish equivalence between groups, quasi-experiments mimic experimental designs in that they attempt to rule out competing causes by identifying groups that are similar except in the nature of the treatment that they receive in the study. Importantly, however, quasi-experiments do not randomize out the effects of other causes as is the case in randomized experimental designs; rather they seek to maximize the equivalence between the units studied through matching or other methods. Threats to internal validity in quasi-experimental studies derive from the fact that it is seldom possible to find or to create treatment and control groups that are not systematically different in one respect or another.

Nonexperimental studies rely primarily on statistical techniques to distinguish the effects of the intervention or treatment from other confounding causes. In practice, quasi-experimental studies often rely as well on statistical approaches to increase the equivalence of the comparisons made.[1] However, in nonexperimental studies, statistical controls are the primary method applied in attempts to increase the level of a study's internal validity. In this case, multivariate statistical methods are used to isolate the effects of treatment from that of other causes. This demands of course that the researcher clearly identify and measure all other factors that may threaten the internal validity of the study outcomes. Only if all such factors are included in the multivariate models estimated can the researcher be confident that the effects of treatment that have been reported are not confounded with other causes.

In theory, the three methods described here are equally valid for solving the problem of isolating treatment or program effects. Each can ensure high internal validity when applied correctly. In practice, however, as Feder and Boruch (2000) note, "there is little disagreement that experiments provide a superior method for assessing the effectiveness of a given intervention" (292). Randomization, according to Kunz and Oxman (1998), "is the only means of controlling for unknown and unmeasured differences between comparison groups as well as those that are known and measured" (1185). While random allocation itself ensures high internal validity in experimental research, for quasi-experimental and nonexperimental research designs, unknown and unmeasured causes are generally seen as representing significant potential threats to the internal validity of the comparisons made.[2]

## INTERNAL VALIDITY AND STUDY OUTCOMES IN PRIOR REVIEWS

While there is general agreement that experimental studies are more likely to ensure high internal validity than are quasi-experimental or nonexperimental studies, it is difficult to specify at the outset the effects that this will have on study outcomes. On one hand, it can be assumed that weaker internal validity is likely to lead to biases in assessment of the effects of treatments or interventions. However, the direction of that bias in any particular study is likely to depend on factors related to the specific character of the research that is conducted. For example, if nonrandomized studies do not account for important confounding causes that are positively related to treatment, they may on average overestimate program outcomes. However, if such unmeasured causes are negatively related to treatment, nonrandomized studies would be expected to underestimate program outcomes. Heinsman and Shadish (1996) suggested that whatever the differences in research design, if nonrandomized and randomized studies are equally well designed and implemented (and thus internal validity is maximized in each), there should be little difference in the estimates gained. Much of what is known empirically about these questions is drawn from reviews in such fields as medicine, psychology, economics, and education (for example, see Burtless 1995; Hedges 2000; Kunz and Oxman 1998; Lipsey and Wilson 1993). Following, what one would expect in theory, a general conclusion that can be reached from the literature is that there is not a consistent bias that results from use of nonrandomized research designs. At the same time, a few studies suggest that differences, in whatever direction, will be smallest when nonrandomized studies are well designed and implemented.

Kunz and Oxman (1998), for example, using studies drawn from the Cochrane database, found varying results when analyzing 18 meta-analyses (incorporating 1211 clinical trials) in the field of health care. Of these 18 systematic reviews, 4 found randomized and higher-quality studies[3] to give higher estimates of effects than nonrandomized and lower-quality studies, and 8 reviews found randomized or high-quality studies to produce lower estimates of effect sizes than nonrandomized or lower-quality studies. Five other reviews found little or inconclusive differences between different types of research designs, and in one review, low-quality studies were found to be more likely to report findings of harmful effects of treatments.

Mixed results are also found in systematic reviews in the social sciences. Some reviews suggest that nonrandomized studies will on average underestimate program effects. For example, Heinsman and Shadish (1996) looked at four meta-analyses that focused on interventions in four different areas: drug use, effects of coaching on Scholastic Aptitude Test performance, ability grouping of pupils in secondary schools, and psychosocial interventions for postsurgery outcomes. Included in their analysis were 98 published and unpublished studies. As a whole,

randomized experiments were found to yield larger effect sizes than studies where randomization was not used. In contrast, Friedlander and Robins (2001), in a review of social welfare programs, found that nonexperimental statistical approaches often yielded estimates larger than those gained in randomized studies (see also Cox, Davidson, and Bynum 1995; LaLonde 1986).

In a large-scale meta-analysis examining the efficacy of psychological, educational, and behavioral treatment, Lipsey and Wilson (1993) suggested that conclusions reached on the basis of nonrandomized studies are not likely to strongly bias conclusions regarding treatment or program effects. Although studies varied greatly in both directions as to whether nonrandomized designs overestimated or underestimated effects as compared with randomized designs, no consistent bias in either direction was detected. Lipsey and Wilson, however, did find a notable difference between studies that employed a control/comparison design and those that used one-group pre and post designs. The latter studies produced consistently higher estimates of treatment effects.

Support for the view that stronger nonrandomized studies are likely to provide results similar to randomized experimental designs is provided by Shadish and Ragsdale (1996). In a review of 100 studies of marital or family psychotherapy, they found overall that randomized experiments yielded significantly larger weighted average effect sizes than nonequivalent control group designs. Nonetheless, the difference between randomized and nonrandomized studies decreased when confounding variables related to the quality of the design of the study were included.

Works that specifically address the relationship between study design and study outcomes are scarce in criminal justice. In turn, assessment of this relationship is most often not a central focus of the reviews developed, and reviewers generally examine a specific criminal justice area, most often corrections (for example, see Bailey 1966; MacKenzie and Hickman 1998; Whitehead and Lab 1989). Results of these studies provide little guidance for specifying a general relationship between study design and study outcomes for criminal justice research. In an early review of 100 reports of correctional treatment between 1940 and 1960, for example, Bailey (1966) found that research design had little effect on the claimed success of treatment, though he noted a slight positive relationship between the "rigor" of the design and study outcome. Logan (1972), who also reviewed correctional treatment programs, found a slight negative correlation between study design and claimed success.

Recent studies are no more conclusive. Wilson, Gallagher, and MacKenzie (2000), in a meta-analysis of corrections-based education, vocation, and work programs, found that run-of-the-mill quasi-experimental studies produced larger effects than did randomized experiments. However, such studies also produced larger effects than did low-quality designs that clearly lacked comparability among groups. In a review of

165 school-based prevention programs, Whitehead and Lab (1989) found little difference in the size of effects in randomized and nonrandomized studies. Interestingly however, they reported that nonrandomized studies were much less likely to report a backfire effect whereby treatment was found to exacerbate rather than ameliorate the problem examined. In contrast, a more recent review by Wilson, Gottfredson, and Najaka (in press) found overall that nonrandomized studies yielded results on average significantly lower than randomized experiments' results, even accounting for a series of other design characteristics (including the overall quality of the implementation of the study). However, it should be noted that many of these studies did not include delinquency measures, and schools rather than individuals were often the unit of random allocation.[4]

## THE STUDY

We sought to define the influence of research design on study outcomes across a large group of studies representing the different types of research design as well as a broad array of criminal justice areas. The most comprehensive source we could identify for this purpose has come to be known as the Maryland Report (Sherman et al. 1997). The Maryland Report was commissioned by the National Institute of Justice to identify "what works, what doesn't, and what's promising" in preventing crime. It was conducted at the University of Maryland's Department of Criminology and Criminal Justice over a yearlong period between 1996 and 1997. The report attempted to identify all available research relevant to crime prevention in seven broad areas: communities, families, schools, labor markets, places, policing, and criminal justice (corrections). Studies chosen for inclusion in the Maryland Report met minimal methodological requirements.[5]

Though the Maryland Report did not examine the relationship between study design and study outcomes, it did define the quality of the methods used to evaluate the strength of the evidence provided through a scientific methods scale (SMS). This SMS was coded with numbers 1 through 5, with "5 being the strongest scientific evidence" (Sherman et al. 1997, 2.18). Overall, studies higher on the scale have higher internal validity, and studies with lower scores have lower internal validity. The 5-point scale was broadly defined in the Maryland Report (Sherman et al. 1997) as follows:

1: Correlation between a crime prevention program and a measure of crime or crime risk factors.
2: Temporal sequence between the program and the crime or risk outcome clearly observed, or a comparison group present without the demonstrated comparability to the treatment group.
3: A comparison between two or more units of analysis, one with and one without the program.
4: Comparison between multiple units with and without the program, controlling for other factors, or a non-equivalent com-

parison group has only minor differences evident.
5: Random assignment and analysis of comparable units to program and comparison groups. (2.18-2.19)

A score of 5 on this scale suggests a randomized experimental design, and a score of 1 a nonexperimental approach. Scores of 3 and 4 may be associated with quasi-experimental designs, with 4 distinguished from 3 by a greater concern with control for threats to internal validity. A score of 2 represents a stronger nonexperimental design or a weaker quasi-experimental approach. However, the overall rating given to a study could be affected by other design criteria such as response rate, attrition, use of statistical tests, and statistical power. It is impossible to tell from the Maryland Report how much influence such factors had on each study's rating. However, correspondence with four of the main study investigators suggests that adjustments based on these other factors were uncommon and generally would result in an SMS decrease or increase of only one level.

Although the Maryland Report included a measure of study design, it did not contain a standardized measure of study outcome. Most prior reviews have relied on standardized effect measures as a criterion for studying the relationship between design type and study findings. Although in some of the area reviews in the Maryland Report, standardized effect sizes were calculated for specific studies, this was not the case for the bulk of the studies reviewed in the report. Importantly, in many cases it was not possible to code such information because the original study authors did not provide the specific details necessary for calculating standardized effect coefficients. But the approach used by the Maryland investigators also reflected a broader philosophical decision that emphasized the bottom line of what was known about the effects of crime and justice interventions. In criminal justice, the outcome of a study is often considered more important than the effect size noted. This is the case in good part because there are often only a very small number of studies that examine a specific type of treatment or intervention. In addition, policy decisions are made not on the basis of a review of the effect sizes that are reported but rather on whether one or a small group of studies suggests that the treatment or intervention works.

From the data available in the Maryland Report, we developed an overall measure of study outcomes that we call the investigator reported result (IRR). The IRR was created as an ordinal scale with three values: 1, 0, and –1, reflecting whether a study concluded that the treatment or intervention worked, had no detected effect, or led to a backfire effect. It is defined by what is reported in the tables of the Maryland Report and is coded as follows:[6]

1: The program or treatment is reported to have had an intended positive effect for the criminal justice system or society. Outcomes in this case supported

the position that interventions or treatments lead to reductions in crime, recidivism, or related measures. [7]

0: The program treatment was reported to have no detected effect, or the effect was reported as not statistically significant.

–1: The program or treatment had an unintended backfire effect for the criminal justice system or society. Outcomes in this case supported the position that interventions or treatments were harmful and lead to increases in crime, recidivism, or related measures.[8]

This scale provides an overall measure of the conclusions reached by investigators in the studies that were reviewed in the Maryland Report. However, we think it is important to note at the outset some specific features of the methodology used that may affect the findings we gain using this approach. Perhaps most significant is the fact that Maryland reviewers generally relied on the reported conclusions of investigators unless there was obvious evidence to the contrary.[9] This approach led us to term the scale the *investigator reported result* and reinforces the fact that we examine the impacts of study design on what investigators report rather than on the actual outcomes of the studies examined.

While the Maryland reviewers examined tests of statistical significance in coming to conclusions about which programs or treatments work,[10] they did not require that statistical tests be reported by investigators to support the specific conclusions reached in each study. In turn, the tables in the Maryland Report often do not note whether specific studies employed statistical tests of significance. Accordingly, in reviewing the Maryland Report studies, we cannot assess whether the presence or absence of such tests influences our conclusions. Later in our article we reexamine our results, taking into account statistical significance in the context of a more recent review in the corrections area that was modeled on the Maryland Report.

Finally, as we noted earlier, most systematic reviews of study outcomes have come to use standardized effect size as a criterion. While we think that the IRR scale is useful for gaining an understanding of the relationship between research design and reported study conclusions, we recognize that a different set of conclusions might have been reached had we focused on standardized effect sizes. Again, we use the corrections review referred to above to assess how our conclusions might have differed if we had focused on standardized effect sizes rather than the IRR scale.

We coded the Scientific Methods Scale and the IRR directly from the tables reported in *Preventing Crime: What Works, What Doesn't, What's Promising* (Sherman et al. 1997). We do not include all of the studies in the Maryland Report in our review. First, given our interest in the area of criminal justice, we excluded studies that did not have a crime or delinquency outcome measure. Second, we excluded studies that did not provide an SMS score (a feature of some

TABLE 1
**STUDIES CATEGORIZED BY SMS**

| | Studies | |
|---|---|---|
| SMS | *n* | Percentage |
| 1 | 10 | 3 |
| 2 | 94 | 31 |
| 3 | 130 | 42 |
| 4 | 28 | 9 |
| 5 | 46 | 15 |
| Total | 308 | 100 |

TABLE 2
**STUDIES CATEGORIZED BY THE IRR**

| | Studies | |
|---|---|---|
| IRR | *n* | Percentage |
| –1 | 34 | 11 |
| 0 | 76 | 25 |
| 1 | 198 | 64 |
| Total | 308 | 100 |

tables in the community and family sections of the report). Finally, we excluded the school-based area from review because only selected studies were reported in tables.[11] All other studies reviewed in the Maryland Report were included, which resulted in a sample of 308 studies. Tables 1 and 2 display the breakdown of these studies by SMS and IRR.

As is apparent from Table 1, there is wide variability in the nature of the research methods used in the studies that are reviewed. About 15 percent were coded in the highest SMS category, which demands a randomized experimental design. Only 10 studies included were coded in the lowest SMS category, though almost a third fall in category 2. The largest category is score 3, which required simply a comparison between two units of analysis, one with and one without treatment. About 1 in 10 cases were coded as 4, suggesting a quasi-experimental study with strong attention to creating equivalence between the groups studied.

The most striking observation that is drawn from Table 2 is that almost two-thirds of the crime and justice studies reviewed in the Maryland Report produced a reported result in the direction of success for the treatment or intervention examined. This result is very much at odds with reviews conducted in earlier decades that suggested that most interventions had little effect on crime or related problems (for example, see Lipton, Martinson, and Wilks 1975; Logan 1972; Martinson 1974).[12] At the same time, a number of the studies examined, about 1 in 10, reported a backfire effect for treatment or intervention.

## RELATING STUDY DESIGN AND STUDY OUTCOMES

In Tables 3 and 4 we present our basic findings regarding the relationship between study design and study outcomes in the Maryland Report sample. Table 3 provides mean IRR outcome scores across the five SMS design categories. While the mean IRR scores in this case present a simple method for examining the results, we also provide an overall statistical measure of correlation, Tau-c (and the associated significance level), which is more appropriate for data of this type. In Table 4 we provide the

TABLE 3
**MEAN IRR SCORES ACROSS SMS CATEGORIES**

| SMS | Mean | *n* | Standard Deviation |
|---|---|---|---|
| 1 | .80 | 10 | .42 |
| 2 | .66 | 94 | .63 |
| 3 | .56 | 130 | .67 |
| 4 | .39 | 28 | .83 |
| 5 | .22 | 46 | .70 |
| Total | .53 | 308 | .69 |

NOTE: Tau-c = –.181. $p < .001$.

cross-tabulation of IRR and SMS scores. This presentation of the results allows us to examine more carefully the nature of the relationship both in terms of outcomes in the expected treatment direction and outcomes that may be classified as backfire effects.

Overall Tables 3 and 4 suggest that there is a linear inverse relationship between the SMS and the IRR. The mean IRR score decreases with each increase in step in the SMS score (see Table 3). While fully nonexperimental designs have a mean IRR score of .80, randomized experiments have a mean of only .22. The run of the mill quasi-experimental designs represented in category 3 have a mean IRR score of .56, while the strongest quasi experiments (category 4) have a mean of .39. The overall correlation between study design and study outcomes is moderate and negative (–.18), and the relationship is statistically significant at the .001 level.

Looking at the cross-tabulation of SMS and IRR scores, our findings are reinforced. The stronger the method in terms of internal validity as measured by the SMS, the less likely is a study to conclude that the intervention or treatment worked. The weaker the method, the less likely the study is to conclude that the intervention or treatment backfired.

While 8 of the 10 studies in the lowest SMS category and 74 percent of those in category 2 show a treatment impact in the desired direction, this was true for only 37 percent of the randomized experiments in category 5. Only in the case of backfire outcomes in categories 4 and 5 does the table not follow our basic findings, and this departure is small. Overall the relationship observed in the table is statistically significant at the .005 level.

### *Comparing the highest-quality nonrandomized studies with randomized experiments*

As noted earlier, some scholars argue that higher-quality nonrandomized studies are likely to have outcomes similar to outcomes of randomized evaluations. This hypothesis is not supported by our data. In Table 5 we combine quasi-experimental studies in SMS categories 3 and 4 and compare them with randomized experimental studies placed in SMS category 5. Again we find a statistically significant negative relationship ($p < .01$). While 37 percent of the level 5 experimental studies show a treatment effect in the desired direction, this was true for 65 percent of the quasi-experimental studies.

Even if we examine only the highest-quality quasi-experimental studies as represented by category 4 and

TABLE 4
**CROSS-TABULATION OF SMS AND IRR**

| | SMS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | | 4 | | 5 | |
| IRR | *n* | Percentage | *n* | Percentage | *n* | Percentage | *n* | Percentage | *n* | Percentage |
| –1 | 0 | 0 | 8 | 9 | 13 | 10 | 6 | 21 | 7 | 15 |
| 0 | 2 | 20 | 16 | 17 | 31 | 24 | 5 | 18 | 22 | 48 |
| 1 | 8 | 80 | 70 | 74 | 86 | 66 | 17 | 61 | 17 | 37 |
| Total | 10 | 100 | 94 | 100 | 130 | 100 | 28 | 100 | 46 | 100 |

NOTE: Chi-square = 25.487 with 8 *df* ($p < .005$).

TABLE 5
**COMPARING QUASI-EXPERIMENTAL STUDIES (SMS = 3 OR 4) WITH RANDOMIZED EXPERIMENTS (SMS = 5)**

| | SMS | | | |
|---|---|---|---|---|
| | 3 or 4 | | 5 | |
| IRR | *n* | Percentage | *n* | Percentage |
| –1 | 19 | 12 | 7 | 15 |
| 0 | 36 | 23 | 22 | 48 |
| 1 | 103 | 65 | 17 | 37 |
| Total | 158 | 100 | 46 | 100 |

NOTE: Chi-square = 12.971 with 2 *df* ($p < .01$).

TABLE 6
**COMPARING HIGH-QUALITY QUASI-EXPERIMENTAL DESIGNS (SMS = 4) WITH RANDOMIZED DESIGNS (SMS = 5)**

| | SMS | | | |
|---|---|---|---|---|
| | 4 | | 5 | |
| IRR | *n* | Percentage | *n* | Percentage |
| –1 | 6 | 21 | 7 | 15 |
| 0 | 5 | 18 | 22 | 48 |
| 1 | 17 | 61 | 17 | 37 |
| Total | 28 | 100 | 46 | 100 |

NOTE: Chi-square = 6.805 with 2 *df* ($p < .05$).

compare these to the randomized studies included in category 5, the relationship between study outcomes and study design remains statistically significant at the .05 level (see Table 6). There is little difference between the two groups in the proportion of backfire outcomes reported; however, there remains a very large gap between the proportion of SMS category 4 and SMS category 5 studies that report an outcome in the direction of treatment effectiveness. While 61 percent of the category 4 SMS studies reported a positive treatment or intervention effect, this was true for only 37 percent of the randomized studies in category 5. Accordingly, even when comparing those nonrandomized studies with the highest internal validity with randomized experiments, we find significant differences in terms of reported study outcomes.

### *Taking into account tests of statistical significance*

It might be argued that had we used a criterion of statistical significance, the overall findings would not have been consistent with the analyses reported above. While we cannot

examine this question in the context of the Maryland Report, since statistical significance is generally not reported in the tables or the text of the report, we can review this concern in the context of a more recent review conducted in the corrections area by one of the Maryland investigators, which uses a similar methodology and reports Maryland SMS (see MacKenzie and Hickman 1998). MacKenzie and Hickman (1998) examined 101 studies in their 1998 review of what works in corrections, of which 68 are reported to have included tests of statistical significance.

Developing the IRR score for each of MacKenzie and Hickman's (1998) studies proved more complex than the coding done for the Maryland Report. MacKenzie and Hickman reported all of the studies' results, sometimes breaking up results by gender, employment, treatment mix, or criminal history, to list a few examples. Rather than count each result as a separate study, we developed two different methods that followed different assumptions for coding the IRR index.

The first simply notes whether any significant findings were found supporting a treatment effect and codes a backfire effect when there are statistically significant negative findings with no positive treatment effects (scale A).[13] The second (scale B) is more complex and gives weight to each result in each study.[14]

Taking this approach, our findings analyzing the MacKenzie and Hickman (1998) data follow those reported when analyzing the Maryland Report. The correlation between study design and study outcomes is negative and statistically significant ($p < .005$) irrespective of the approach we used to define the IRR outcome scale (see Table 7). Using scale A, the correlation observed is –.29, while using scale B, the observed correlation is –.31.

TABLE 7

**RELATING SMS AND IRR ONLY FOR STUDIES IN MACKENZIE AND HICKMAN (1998) THAT INCLUDE TESTS OF STATISTICAL SIGNIFICANCE**

| | Scale A | | Scale B | |
|---|---|---|---|---|
| SMS | Mean | *n* | Mean | *n* |
| 1 | | 0 | | 0 |
| 2 | 0.83 | 24 | 1.46 | 24 |
| 3 | 0.62 | 26 | 1.04 | 26 |
| 4 | 0.36 | 11 | 0.64 | 11 |
| 5 | 0.00 | 7 | 0.14 | 7 |
| Total | .59 | 68 | 1.03 | 68 |

NOTE: Tau-c for scale A = –.285 ($p < .005$). Tau-c for scale B = –.311 ($p < .005$).

*Comparing effect size and IRR score results*

It might be argued that our overall findings are related to specific characteristics of the IRR scale rather than the underlying relationship between study design and study outcomes. We could not test this question directly using the Maryland Report data because, as noted earlier, standardized effect sizes were not consistently recorded in the report. However, MacKenzie and Hickman (1998) did report standardized effect size coefficients, and thus we are able to reexamine this question in the context of corrections-based criminal justice studies.

Using the average standardized effect size reported for each study reviewed by MacKenzie and Hickman (1998) for the entire sample (including studies where statistical significance is not reported), the results follow those gained from relating IRR and SMS scores using the Maryland Report sample (see Table 8). Again the correlation between SMS and study outcomes is negative; in this case the correlation is about –.30. The observed relationship is also statistically significant at the .005 level. Accordingly, these findings suggest that our observation of a negative relationship between study design and study outcomes in the Maryland Report sample is not an artifact of the particular codings of the IRR scale.

TABLE 8
**RELATING AVERAGE EFFECT SIZE AND SMS FOR STUDIES IN MacKENZIE AND HICKMAN (1988)**

| | Effect Size Available from the Entire Sample | |
|---|---|---|
| SMS | Mean | *n* |
| 1 | | 0 |
| 2 | .29 | 39 |
| 3 | .23 | 30 |
| 4 | .19 | 13 |
| 5 | .00 | 7 |
| Total | .23 | 89 |
| Missing values | | 12 |

NOTE: Correlation ($r$) = –.296 ($p < .005$).

## DISCUSSION

Our review of the Maryland Report Studies suggests that in criminal justice, there is a moderate inverse relationship between the quality of a research design, defined in terms of internal validity, and the outcomes reported in a study. This relationship continues to be observed even when comparing the highest-quality nonrandomized studies with randomized experiments. Using a related database concentrating only on the corrections area, we also found that our findings are consistent when taking into account only studies that employed statistical tests of significance. Finally, using the same database, we were able to examine whether our results would have differed had we used standardized effect size measures rather than the IRR index that was drawn from the Maryland Report. We found our results to be consistent using both methods. Studies that were defined as including designs with higher internal validity were likely to report smaller effect sizes than studies with designs associated with lower internal validity.

Prior reviews of the relationship between study design and study outcomes do not predict our findings. Indeed, as we noted earlier, the main lesson that can be drawn from prior research is that the impact of study design is very much dependant on the characteristics of the particular area or studies that are reviewed. In theory as well, there is no reason to assume that there will be a systematic type of bias in studies with lower internal validity. What can be said simply is that such studies, all else being equal, are likely to provide biased findings as compared with results drawn from randomized experimental designs. Why then do we find in reviewing a broad group of

crime and justice studies what appears to be a systematic relationship between study design and study outcomes?

One possible explanation for our findings is that they are simply an artifact of combining a large number of studies drawn from many different areas of criminal justice. Indeed, there are generally very few studies that examine a very specific type of treatment or intervention in the Maryland Report. And it may be that were we able to explore the impacts of study design on study outcomes for specific types of treatments or interventions, we would find patterns different from the aggregate ones reported here. We think it is likely that for specific areas of treatment or specific types of studies in criminal justice, the relationship between study design and study outcomes will differ from those we observe. Nonetheless, review of this question in the context of one specific type of treatment examined by the Campbell Collaboration (where there was a substantial enough number of randomized and nonrandomized studies for comparison) points to the salience of our overall conclusions even within specific treatment areas (see Petrosino, Petrosino, and Buehler 2001). We think this example is particularly important because it suggests the potential confusion that might result from drawing conclusions from nonrandomized studies.

Relying on a systematic review conducted by Petrosino, Petrosino, and Buehler (2001) on Scared Straight and other kids-visit programs, we identified 20 programs that included crime-related outcome measures. Of these, 9 were randomized experiments, 4 were quasi-experimental trials, and 7 were fully nonexperimental studies. Petrosino, Petrosino, and Buehler reported on the randomized experimental trials in their Campbell Collaboration review. They concluded that Scared Straight and related programs do not evidence any benefit in terms of recidivism and actually increase subsequent delinquency. However, a very different picture of the effectiveness of these programs is drawn from our review of the quasi-experimental and nonexperimental studies. Overall, these studies, in contrast to the experimental evaluations, suggest that Scared Straight programs not only are not harmful but are more likely than not to produce a crime prevention benefit.

We believe that our findings, however preliminary, point to the possibility of an overall positive bias in nonrandomized criminal justice studies. This bias may in part reflect a number of other factors that we could not control for in our data, for example, publication bias or differential attrition rates across designs (see Shadish and Ragsdale 1996). However, we think that a more general explanation for our findings is likely to be found in the norms of criminal justice research and practice.

Such norms are particularly important in the development of nonrandomized studies. Randomized experiments provide little freedom to the researcher in defining equivalence between treatment and comparison groups. Equivalence in randomized experiments is defined

simply through the process of randomization. However, nonrandomized studies demand much insight and knowledge in the development of comparable groups of subjects. Not only must the researcher understand the factors that influence treatment so that he or she can prevent confounding in the study results, but such factors must be measured and then controlled for through some statistical or practical procedure.

It may be that such manipulation is particularly difficult in criminal justice study. Criminal justice practitioners may not be as strongly socialized to the idea of experimentation as are practitioners in other fields like medicine. And in this context, it may be that a subtle form of creaming in which the cases considered most amenable to intervention are placed in the intervention group is common. In specific areas of criminal justice, such creaming may be exacerbated by self-selection of subjects who are motivated toward rehabilitation. Nonrandomized designs, even in relatively rigorous quasi-experimental studies, may be unable to compensate or control for why a person is considered amenable and placed in the intervention group. Matching on traditional control variables like age and race, in turn, might not identify the subtle components that make individuals amenable to treatment and thus more likely to be placed in intervention or treatment categories.

Of course, we have so far assumed that nonrandomized studies are biased in their overestimation of program effects. Some scholars might argue just the opposite. The inflexibility of randomized experimental designs has sometimes been seen as a barrier to development of effective theory and practice in criminology (for example, see Clarke and Cornish 1972; Eck 2001; Pawson and Tilley, 1997). Here it is argued that in a field in which we still know little about the root causes and processes that underlie phenomena we seek to influence, randomized studies may not allow investigators the freedom to carefully explore how treatments or programs influence their intended subjects. While this argument has merit in specific circumstances, especially in exploratory analyses of problems and treatments, we think our data suggest that it can lead in more developed areas of our field to significant misinterpretation and confusion.

## CONCLUSION

We asked at the outset of our article whether the type of research design used in criminal justice influences the conclusions that are reached. Our findings, based on the Maryland Report, suggest that design does matter and that its effect in criminal justice study is systematic. The weaker a design, as indicated by internal validity, the more likely was a study to report a result in favor of treatment and the less likely it was to report a harmful effect of treatment. Even when comparing studies defined as randomized designs in the Maryland Report with strong quasi-experimental research designs, systematic and statistically

significant differences were observed. Though our study should bscores e seen only as a preliminary step in understanding how research design affects study outcomes in criminal justice, it suggests that systematic reviews of what works in criminal justice may be strongly biased when including nonrandomized studies. In efforts such as those being developed by the Campbell Collaboration, such potential biases should be taken into account in coming to conclusions about the effects of interventions.

## Notes

1. Statistical adjustments for random group differences are sometimes employed in experimental studies as well.

2. We should note that we have assumed so far that external validity (the degree to which it can be inferred that outcomes apply to the populations that are the focus of treatment) is held constant in these comparisons. Some scholars argue that experimental studies are likely to have lower external validity because it is often difficult to identify institutions that are willing to randomize participants. Clearly, where randomized designs have lower external validity, the assumption that they are to be preferred to nonrandomized studies is challenged.

3. Kunz and Oxman (1998) not only compared randomized and nonrandomized studies but also adequately and inadequately concealed randomized trials and high-quality versus low-quality studies. Generally, high-quality randomized studies included adequately concealed allocation, while lower-quality randomized trails were inadequately concealed. In addition, the general terms *high-quality trials* and *low-quality trials* indicate a difference where "the specific effect of randomization or allocation concealment could not be separated from the effect of other methodological manoeuvres such as double blinding" (Kunz and Oxman 1998, 1185).

4. Moreover, it may be that the finding of higher standardized effects sizes for randomized studies in this review was due to school-level as opposed to individual-level assignment. When only those studies that include a delinquency outcome are examined, a larger effect is found when school rather than student is the unit of analysis (Denise Gottfredson, personal communication, 2001).

5. As the following Scientific Methods Scale illustrates, the lowest acceptable type of evaluation for inclusion in the Maryland Report is a simple correlation between a crime prevention program and a measure of crime or crime risk factors. Thus studies that were descriptive or contained only process measures were excluded.

6. There were also (although rarely) studies in the Maryland Report that reported two findings in opposite directions. For instance, in Sherman and colleagues' (1997) section on specific deterrence (8.18-8.19), studies of arrest for domestic violence had positive results for employed offenders and backfire results for nonemployed offenders. In these isolated cases, the study was coded twice with the same scientific methods scores and each of the investigator-reported result scores (of 1 and –1) separately.

7. For studies examining the absence of a program (such as a police strike) where social conditions worsened or crime increased, this would be coded as 1.

8. For studies examining the absence of a program (such as a police strike) where social conditions improved or crime decreased, this would be coded as –1.

9. Only in the school-based area was there a specific criterion for assessing the investigator's conclusions. As noted below, however, the school-based studies are excluded from our review for other reasons.

10. For example, the authors of the Maryland Report noted in discussing criteria for deciding which programs work, "These are programs that we are reasonably certain of preventing crime or reducing risk factors for crime in the kinds of social contexts in which they have been evaluated, and for which the findings should be generalizable to similar settings in other places and times. Programs coded as 'working' by this definition must have at least two level 3 evaluations with statistical

significance tests showing effectiveness and the preponderance of all available evidence supporting the same conclusion" (Sherman et al. 1997, 2-20).

11. It is the case that many of the studies in this area would have been excluded anyway since they often did not have a crime or delinquency outcome measure (but rather examined early risk factors for crime and delinquency).

12. While the Maryland Report is consistent with other recent reviews that also point to greater success in criminal justice interventions during the past 20 years (for example, see Poyner 1993; Visher and Weisburd 1998; Weisburd 1997), we think the very high percentage of studies showing a treatment impact is likely influenced by publication bias. The high rate of positive findings is also likely influenced by the general weaknesses of the study designs employed. This is suggested by our findings reported later: that the weaker a research design in terms of internal validity, the more likely is the study to report a positive treatment outcome.

13. The coding scheme for scale A was as follows. A value of 1 indicates that the study had any statistically significant findings supporting a positive treatment effect, even if findings included results that were not significant or had negative or backfire findings. A value of 0 indicates that the study had only nonsignificant findings. A value of –1 indicates that the study had only statistically significant negative or backfire findings or statistically significant negative findings with other nonsignificant results.

14. Scale B was created according to the following rules. A value of 2 indicates that the study had only or mostly statistically significant findings supporting a treatment effect (more than 50 percent) when including all results, even nonsignificant ones. A value of 1 indicates that the study had some statistically significant findings supporting a treatment effect (50 percent or less, counting both positive significant and nonsignificant results) even if the nonsignificant results outnumbered the positive statistically significant results. A value of 0 indicates that no statistically significant findings were reported. A value of –1 indicates that the study evidenced statistically significant backfire effects (even if nonsignificant results were present) but no statistically significant results supporting the effectiveness of treatment.

## References

Bailey, Walter C. 1966. Correctional Outcome: An Evaluation of 100 Reports. *Journal of Criminal Law, Criminology and Police Science* 57:153-60.

Boruch, Robert F., Anthony Petrosino, and Iain Chalmers. 1999. The Campbell Collaboration: A Proposal for Systematic, Multi-National, and Continuous Reviews of Evidence. Background paper for the meeting at University College–London, School of Public Policy, July.

Boruch, Robert F., Brook Snyder, and Dorothy DeMoya. 2000. The Importance of Randomized Field Trials. *Crime & Delinquency* 46:156-80.

Burtless, Gary. 1995. The Case for Randomized Field Trials in Economic and Policy Research. *Journal of Economic Perspectives* 9:63-84.

Campbell, Donald P. and Robert F. Boruch. 1975. Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six Ways in Which Quasi-Experimental Evaluations in Compensatory Education Tend to Underestimate Effects. In *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs*, ed. Carl Bennett and Arthur Lumsdaine. New York: Academic Press.

Chalmers, Iain and Douglas G. Altman. 1995. *Systematic Reviews*. London: British Medical Journal Press.

Clarke, Ronald V. and Derek B. Cornish. 1972. *The Control Trial in Institutional Research: Paradigm or Pitfall for Penal Evaluators?* London: HMSO.

Cox, Stephen M., William S. Davidson, and Timothy S. Bynum. 1995. A Meta-Analytic Assessment of Delinquency-

Related Outcomes of Alternative Education Programs. *Crime & Delinquency* 41:219-34.

Cullen, Francis T. and Paul Gendreau. 2000. Assessing Correctional Rehabilitation: Policy, Practice, and Prospects. In *Policies, Processes, and Decisions of the Criminal Justice System: Criminal Justice 3*, ed. Julie Horney. Washington, DC: U.S. Department of Justice, National Institute of Justice.

Davies, Huw T. O., Sandra Nutley, and Peter C. Smith. 2000. *What Works: Evidence-Based Policy and Practice in Public Services.* London: Policy Press.

Davies, Philip. 1999. What Is Evidence-Based Education? *British Journal of Educational Studies* 47:108-21.

Eck, John. 2001. Learning from Experience in Problem Oriented Policing and Crime Prevention: The Positive Functions of Weak Evaluations and the Negative Functions of Strong Ones. Unpublished manuscript.

Egger, Matthias and G. Davey Smith. 1998. Bias in Location and Selection of Studies. *British Medical Journal* 316:61-66.

Farrington, David P. 1983. Randomized Experiments in Crime and Justice. In *Crime and Justice: An Annual Review of Research*, ed. Norval Morris and Michael Tonry. Chicago: University of Chicago Press.

———. 2000. Standards for Inclusion of Studies in Systematic Reviews. Discussion paper for the Campbell Collaboration Crime and Justice Coordinating Group.

Farrington, David P. and Anthony Petrosino. 2001. The Campbell Collaboration Crime and Justice Group. *Annals of the American Academy of Political and Social Science* 578:35-49.

Feder, Lynette and Robert F. Boruch. 2000. The Need for Experiments in Criminal Justice Settings. *Crime & Delinquency* 46:291-94.

Feder, Lynette, Annette Jolin, and William Feyerherm. 2000. Lessons from Two Randomized Experiments in Criminal Justice Settings. *Crime & Delinquency* 46:380-400.

Friedlander, Daniel and Philip K. Robins. 2001. Evaluating Program Evaluations: New Evidence on Commonly Used Non-Experimental Methods. *American Economic Review* 85:923-37.

Hedges, Larry V. 2000. Using Converging Evidence in Policy Formation: The Case of Class Size Research. *Evaluation and Research in Education* 14:193-205.

Heinsman, Donna T. and William R. Shadish. 1996. Assignment Methods in Experimentation: When Do Nonrandomized Experiments Approximate Answers from Randomized Experiments? *Psychological Methods* 1:154-69.

Kunz, Regina and Andy Oxman. 1998. The Unpredictability Paradox: Review of Empirical Comparisons of Randomized and Non-Randomized Clinical Trials. *British Medical Journal* 317:1185-90.

LaLonde, Robert J. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review* 76:604-20.

Lipsey, Mark W. and David B. Wilson. 1993. The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation from Meta-Analysis. *American Psychologist* 48:1181-209.

Lipton, Douglas S., Robert M. Martinson, and Judith Wilks. 1975. *The Effectiveness of Correctional Treatment: A Survey of Treatment Evaluation Studies.* New York: Praeger.

Logan, Charles H. 1972. Evaluation Research in Crime and Delinquency—A Reappraisal. *Journal of Criminal Law, Criminology and Police Science* 63:378-87.

MacKenzie, Doris L. 2000. Evidence-based Corrections: Identifying What Works. *Crime & Delinquency* 46:457-71.

MacKenzie, Doris L. and Laura J. Hickman. 1998. *What Works in Corrections* (Report submitted to the State of Washington Legislature Joint Audit and Review Committee). College Park: University of Maryland.

Martinson, Robert. 1974. What Works? Questions and Answers About Prison Reform. *Public Interest* 35:22-54.

Millenson, Michael L. 1997. *Demanding Medical Excellence: Doctors and Accountability in the Information Age*. Chicago: University of Chicago Press.

Nutley, Sandra and Huw T. O. Davies. 1999. The Fall and Rise of Evidence in Criminal Justice. *Public Money & Management* 19:47-54.

Pawson, Ray and Nick Tilley. 1997. *Realistic Evaluation*. London: Sage.

Petrosino, Anthony, Carolyn Petrosino, and John Buehler. 2001. Pilot Test: The Effects of Scared Straight and Other Juvenile Awareness Programs on Delinquency. Unpublished manuscript.

Poyner, Barry. 1993. What Works in Crime Prevention: An Overview of Evaluations. In *Crime Prevention Studies*. Vol. 1, ed. Ronald V. Clarke. Monsey, NY: Criminal Justice Press.

Shadish, William R. and Kevin Ragsdale. 1996. Random Versus Nonrandom Assignment in Controlled Experiments: Do You Get the Same Answer? *Journal of Consulting and Clinical Psychology* 64:1290-305.

Sherman, Lawrence W. 1998. *Evidence-Based Policing*. In *Ideas in American Policing*. Washington, DC: Police Foundation.

Sherman, Lawrence W., Denise C. Gottfredson, Doris Layton MacKenzie, John E. Eck, Peter Reuter, and Shawn D. Bushway. 1997. *Preventing Crime: What Works, What Doesn't, What's Promising*. Washington, DC: U.S. Department of Justice, National Institute of Justice.

Visher, Christy A. and David Weisburd. 1998. Identifying What Works: Recent Trends in Crime Prevention. *Crime, Law and Social Change* 28:223-42.

Weisburd, David. 1997. *Reorienting Crime Prevention Research and Policy: From the Causes of Criminality to the Context of Crime* (Research Report NIJ 16504). Washington, DC: U.S. Department of Justice, National Institute of Justice.

Whitehead, John T. and Steven P. Lab. 1989. A Meta-Analysis of Juvenile Correctional Treatment. *Journal of Research in Crime and Delinquency* 26:276-95.

Wilkinson, Leland and Task Force on Statistical Inference. 1999. Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist* 54:594-604.

Wilson, David B., Catherine A. Gallagher, Doris L. MacKenzie. 2000. A Meta-Analysis of Corrections-Based Education, Vocation, and Work Programs for Adult Offenders. *Journal of Research in Crime and Delinquency* 37:347-68.

Wilson, David B., Denise C. Gottfredson, and Stacy S. Najaka. In Press. School-Based Prevention of Problem Behaviors: A Meta-Analysis. *Journal of Quantitative Criminology*.

Zuger, Abigail. 1997. New Way of Doctoring: By the Book. *New York Times*, 16 Dec.