When Can We Conclude That Treatments or Programs "Don't Work"?

By
DAVID WEISBURD,
CYNTHIA M. LUM,
and
SUE-MING YANG

In this article, the authors examine common practices of reporting statistically nonsignificant findings in criminal justice evaluation studies. They find that criminal justice evaluators often make formal errors in the reporting of statistically nonsignificant results. Instead of simply concluding that the results were not statistically significant, or that there is not enough evidence to support an effect of treatment, they often mistakenly accept the null hypothesis and state that the intervention had no impact or did not work. The authors propose that researchers define a second null hypothesis that sets a minimal threshold for program effectiveness. In an illustration of this approach, they find that more than half of the studies that had no statistically significant finding for a traditional, no difference null hypothesis evidenced a statistically significant result in the case of a minimal worthwhile treatment effect null hypothesis.

Keywords: what works; null hypothesis significance testing; effect sizes; statistical significance; statistical power

Criminal justice researchers and policy makers have begun to focus greater attention on the question of what works in preventing and controlling crime (Sherman et al. 1998). Following a growing interest in evidence-based practice in medicine and other fields (Davies, Nutley, and Smith 2000; Millenson 1997; Nutley and Davies 1999; Zuger 1997), there is now wide recognition of the need to answer core questions

David Weisburd is a professor in the Department of Criminology and Criminal Justice at the University of Maryland and in the Institute of Criminology at the Hebrew University Law School in Jerusalem.

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

Sue-Ming Yang is a doctoral student in the Department of Criminology and Criminal Justice at the University of Maryland.

NOTE: We are indebted to a number of colleagues for helpful comments in preparing this article. We especially want to thank Thomas Cook, Ian Chalmers, Mark Lipsey, Joseph Naus, Anthony Petrosino, and David Wilson.

DOI: 10.1177/0002716202250782

about what types of policies, programs, and treatments are useful in developing effective crime prevention and control practices (MacKenzie 2000; Sherman et al. 2002). But in trying to define programs that work in criminal justice, researchers have also been called on to decide which programs do not work. And here, researchers are faced with a difficult dilemma that is often not understood by the policy makers who seek answers to core policy questions or even many researchers who must provide to them interpretable, policy-relevant conclusions.

In everyday logic, answering one of our questions naturally leads us to a conclusion about the other. If we conclude that something does work, then by inference, we reject the conclusion that it does not work. If we cannot conclude that a program or treatment works, we assume that it does not. But in the scientific logic that underlies most evaluation research in the social sciences and medicine, our two questions do not reflect a simple dichotomy. This logic, which is often defined under the general term *null hypothesis statistical testing*, draws us to a very specific way of thinking about the outcomes of studies and the conclusions we can draw from them. It begins with a very simple assumption that the researcher then tries to test. This assumption, defined as the null hypothesis, is ordinarily that the program or treatment has no effect. The task of the researcher is to see whether the empirical findings gained in a study provide enough reliable evidence to reject this null hypothesis. When such evidence is brought to bear, we say that the results are statistically significant. When a criminal justice researcher finds, for example, that the group receiving a sanction or treatment has significantly fewer arrests than a group that did not (the control group in a randomized experiment), this is ordinarily seen as solid evidence that the practice evaluated has worked in preventing crime.[1]

Our dilemma using the logic of null hypothesis statistical testing is that a finding that there is not sufficient empirical evidence to conclude that a program works cannot by implication lead us to the conclusion that a program does not work. Indeed, this approach says very little directly about this latter question either statistically or substantively. A finding that there is not enough evidence to state that the treatment has a statistically significant outcome does not mean that we have enough evidence to decide that the treatment does not work. When we fail to reject the null hypothesis and come to the conclusion that our results are not statistically significant, we have not made a scientific statement about whether the null hypothesis is true. We have simply stated that based on the evidence available to us we cannot conclude that it is false.

This fact has important implications for what evaluation researchers can say about whether practices do not work based on null hypothesis statistical testing logic. It is formally incorrect based on this logic to conclude that the null hypothesis is true. By implication, the researcher is incorrect when using language such as there is "no effect," there is "no difference," there is "no impact," or the treatment "did not work" based on results that are not statistically significant. As Jacob Cohen (1988), a distinguished psychological statistician, wrote,

> Research reports in the literature are frequently flawed by conclusions that state or imply that the null hypothesis is true. For example, following the finding that the difference

> between two sample means is not statistically significant, instead of properly concluding from this failure to reject the null hypothesis that the data do not warrant the conclusion that the population means differ, the writer concludes, at least implicitly, that there is no difference. The latter conclusion is always strictly invalid. (P. 16)

Formally then, a finding of no statistically significant difference should not lead the criminal justice researcher to accept the null hypothesis that there is no difference or no treatment effect (see also Alderson and Chalmers in press; Finch, Cumming, and Thomason 2001). But in the real world, not only are we concerned with whether programs or treatments work; we are also concerned with whether we have enough evidence to conclude that they are not useful. Accordingly, as Cook and Campbell (1979) noted, "While we cannot prove the null hypothesis, in

---

[C]riminal justice researchers often make formal errors in reporting study results...[that] cannot be justified by the design or outcomes observed in the studies reviewed.

---

many practical contexts we have to make decisions and act as though the null hypothesis were true" (p. 45; see also Cook et al. 1979). For example, if we could state confidently based on our empirical study that the effects of the treatment were very small, then it might be warranted in practical if not statistical terms to state that the program does not work. Similarly, if we could be confident that our research was adequately designed to identify a meaningful treatment effect, if none were found, it would seem reasonable to draw a conclusion that such an effect did not exist.

In this article, we examine common practices of reporting statistically nonsignificant findings in criminal justice evaluation studies. Do criminal justice researchers follow closely the norms of good statistical practice? Or do they often make the mistake of claiming that their findings support the null hypothesis? In practical terms, does the design of criminal justice studies or the outcomes observed suggest that we can act as though the null hypothesis were true in describing research results even if we cannot formally state this in statistical terms? Below, we describe the sample that forms the basis for our study and the methods we used to examine reporting practices. We then describe our main findings. Our study shows that criminal justice researchers often make formal errors in reporting study results and that such formal reporting errors cannot be justified by the design or outcomes observed in the studies reviewed. In concluding, we suggest and illus-

trate an alternative null hypothesis statistical testing method that allows conclusions to be reached about whether programs or practices failed to meet a minimal threshold of success.

The Study

We sought to examine criminal justice practices of reporting statistically nonsignificant findings across a large group of studies representing 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).[2] 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, Department of Criminology and Criminal Justice, during a year 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 were chosen for inclusion in the Maryland Report that met minimal methodological requirements.[3]

While the Maryland Report includes the most comprehensive listing of relevant criminal justice studies presently available and describes the findings of each study briefly, the report does not consistently identify whether a test of statistical significance was conducted in a study. Because of this, we conservatively selected every study in the Maryland Report in which the reported findings could be interpreted as indicating a statistically nonsignificant result. We excluded from our initial sample studies that reported beneficial or harmful effects for interventions, as well as those whose outcomes were specifically reported in the Maryland Report as statistically significant. We found that seventy-three of the studies reported on in the Maryland Report met this initial criterion.

We were able to locate reports or articles on sixty-five (89 percent) of the seventy-three studies we initially identified.[4] After careful examination of reports or articles associated with each of these sixty-five studies, thirteen were determined ineligible for our study either because evaluators did not indicate that any tests of significance were conducted or because the actual reported outcome in the Maryland Report was inconsistent with the author's conclusions. In four cases, we decided that a single Maryland study should be coded for our purposes as separate studies. For example, two independent neighborhood foot patrol programs conducted in two different cities were treated as a single study in the Maryland Report but are distinguished in our investigation (Police Foundation 1981). Overall, fifty-eight independent studies were included in our final sample.

To examine reporting practices for findings that were not statistically significant, we recorded the wording used in describing the results of the evaluation in each of three sections of research reports or articles: the abstract (or when unavailable, the introduction, the initial summary, or in some cases, the executive summary given at the beginning of lengthy reports), the results (often labeled the findings of a study),

TABLE 1
EXAMPLES OF FORMALLY INCORRECT AND FORMALLY CORRECT WORDING

Formally Incorrect Wording (Accepting the Null Hypothesis)	Formally Correct Wording (Failing to Reject the Null Hypothesis)
• no difference/change • did not work/failed • no effect/influence/impact/benefit • as likely • no relation between • finding not statistically significant and therefore there is no difference between groups	• no statistically significant differences or findings not statistically significant (without accepting the null) • failed to affect significantly • not significant or no significance (omitted the word "statistically") • no evidence to support claim • little difference, no substantial difference, virtually identical

and the conclusion/discussion, which was found toward the end of the reports or articles. A dichotomized coding system was used in which reporting was classified as formally incorrect or formally correct. We used an earlier article examining reporting practices in psychology as a guide in categorizing the studies (Finch, Cumming, and Thomason 2001). We coded reporting practices as "formally incorrect" that accepted the null hypothesis of "no difference" or "no impact" or that reported a finding as statistically nonsignificant and then directly interpreted that finding as accepting the null hypothesis. We classified the description of outcomes as "formally correct" when a statistically nonsignificant result was reported without stating that the null hypothesis was accepted. Authors who used wording such as "little difference," "virtually no difference," or "not a substantial difference" were given the benefit of the doubt, and these were coded conservatively as using formally correct wording. Table 1 provides some examples of formally incorrect and formally correct wording.

While the interpretation of statistically nonsignificant findings forms the main focus of our study, we also wanted to assess whether the nature of the outcomes observed or the design used warranted a practical decision to "act as though the null hypothesis were true" (Cook and Campbell 1979, 45). This would be the case, for example, if the effects observed in the studies examined were very close to zero or if the study designs were very sensitive and thus able to detect with a high degree of certainty even very modest outcomes.

In recent years, statisticians have developed a number of commonly used standardized measures of effect size for comparing studies using different types of measurement (Lipsey and Wilson 2001; Rosenthal 1991). While only a handful of the studies reviewed here reported standardized effect size (ES) measures, we tried to calculate effect size on the basis of other statistics that were reported in the articles we reviewed. Using this method, we were able to gain enough information to calculate effect sizes in forty-three of the fifty-eight independent studies. Findings both within and across studies were often reported using different statistics.

TABLE 2
HOW STATISTICALLY NONSIGNIFICANT FINDINGS ARE REPORTED (BY SECTION OF THE ARTICLE)

	Abstract		Results		Conclusions	
	n	%	*n*	%	*n*	%
Formally incorrect wording	17	49	17	29	30	64
Formally correct wording	18	51	41	71	17	36
Total	35	100	58	100	47	100
Specific finding not reported in section	23		0		11	

Because of this variability and our need to compare effect sizes across statistically nonsignificant findings, we calculated and then converted all effect sizes into the standardized mean difference effect size, often denoted as ES_{sm} (Lipsey and Wilson 2001). This is also known as Cohen's d (Cohen 1988), which is the difference between the means of two comparison groups divided by their pooled standard deviation.[5] If an independent study had more than one reported statistically nonsignificant finding (31 percent of the independent studies), the smallest, average, and largest effect sizes were recorded.[6]

The sensitivity of a research design is generally assessed by measuring the statistical power of a study. Statistical power tells us how likely it would be to observe an effect of a certain size in a specific study if that effect is found in the population of interest (Lipsey 1998; Weisburd 1993; Weisburd and Britt 2002). In formal terms, statistical power represents the probability that a study will lead to rejection of the null hypothesis under a specific set of assumptions. For our purposes, statistical power provides a method for examining whether a study would be likely to identify a specific type of effect in a sample if it existed in the population of interest. If a study had a high level of statistical power to identify an effect that was defined as meaningful, then a statistically nonsignificant finding could be seen practically as a good indication that a meaningful effect did not exist—even if it would be formally incorrect to conclude that the null hypothesis was true. To calculate statistical power for the studies reviewed, it was necessary to collect information on the sample sizes associated with each individual finding reported.[7] This was possible for fifty-one of the fifty-eight independent studies we examined.[8]

How Are Statistically Nonsignificant Findings Reported?

Our main research concern is with the nature of the reporting of statistically nonsignificant results. Do researchers generally state their findings in a formally correct manner, or do they often use language that violates the formal logic of null

hypothesis statistical testing? Table 2 suggests that it is fairly common in criminal justice evaluations to describe statistically nonsignificant results as leading to a finding of no difference or no impact or as evidence that the intervention or practice does not work. Our findings also show that criminal justice evaluators are much more likely to make such a formal error in reporting results in the abstract or conclusions of their article.

Of the thirty-five studies that reported statistically nonsignificant results in their abstracts, almost half did so in a way that formally violates the statistical logic that underlies the tests used. In the case of the conclusions, this was true for almost two-thirds of the forty-seven studies in which statistically nonsignificant results were described. Not surprisingly, researchers were much more likely to report study findings using formally correct terms such as "not statistically significant" or "no evidence to support the claim" in the results section.[9] More generally, this is the section of the report or article where statistical details of studies are provided. Nonetheless, 29 percent of these studies use terms such as the treatment "doesn't work" or has "no impact" or interpret statistically nonsignificant results as accepting the null hypothesis.

A finding that there is not enough evidence to state that the treatment has a statistically significant outcome does not mean that we have enough evidence to decide that the treatment does not work.

We might question whether reporting practices have improved over time. In this case, we might speculate that as criminal justice researchers have become more methodologically sophisticated, so too have they become more careful in the reporting of statistically nonsignificant research results. While this proposition seems reasonable, it is not supported by our study. We do not find a statistically significant relationship between the year of publication and the use of formally correct or incorrect reporting of results.[10]

Our first main conclusion, accordingly, is that criminal justice researchers often report statistically nonsignificant results in ways that formally violate good statistical reporting practices.[11] They commonly use language such as there is "no impact" or that the "program doesn't work," which suggests acceptance of the null hypothesis. This despite the fact that a statistically nonsignificant result cannot formally

lead to acceptance of the null hypothesis. But are criminal justice researchers unique in their failure to formally follow good reporting practices in regard to statistically nonsignificant results?

We could find only two comparable studies that examined reporting practices. Finch, Cumming, and Thomason (2001) took a sample of articles from the *Journal of Applied Psychology* from the years 1940, 1955, 1970, 1985, and 1999. They also took a sample from the year 1999 from the *British Journal of Psychology*. Their results suggest that "the proportion of [statistically nonsignificant] articles in which a null hypothesis is accepted averages 38%" (p. 195).[12] Using a similar threshold, in which a statistically nonsignificant result was incorrectly reported at least once in a study, the result for criminal justice evaluations would be much higher, about 66 percent. Alderson and Chalmers (in press) examined reporting practices in medicine using the *Cochrane Database of Systematic Reviews* for 2001 and the first half of 2002. They found that 22.5 percent of the studies reviewed in 2001 and 13.3 percent in 2002 used incorrect reporting in the Main Results or Reviewers' Conclusions sections. Accordingly, at least as compared with psychology and medicine, criminal justice evaluators seem especially likely to incorrectly report statistically nonsignificant findings.

Do the Effect Sizes Evident in These Studies or Their Statistical Power Imply That It Is Possible to Act as Though the Null Hypothesis Were True?

While criminal justice evaluators often make formal errors in reporting statistically nonsignificant results, it may be that such errors are not serious in practice. As Mark Lipsey (2000) observed, the "proper goal of intervention research is to detect any effects of meaningful magnitude relative to the nature of the intervention and the conditions it addresses" (p. 108). If, for example, the observed effect size in a study is very close to zero, it might be argued that a statement that there is no impact or no difference, while formally incorrect, does not provide a misleading view of study results.

Moreover, it is nearly impossible for the null hypothesis to be formally true in any study (Anderson, Burnham, and Thompson 2000; Chow 1998; Cohen 1994; Johnson 1995; Rosenthal 1995). As David Bakan (1966) observed in the *American Psychologist* almost forty years ago,

> The fact of the matter is that there is really no good reason to expect the null hypothesis to be true in any population. . . . Why should any correlation coefficient be exactly .00 in the population? . . . A glance at any set of statistics on total populations will quickly confirm the rarity of the null hypothesis in nature. (P. 426)

Does the fact that the null hypothesis is almost never true mean that we are wrong in ever coming to a conclusion that a program or treatment does not work? Clearly,

when the effect of a treatment or program is very small, it would seem reasonable, as Cook and Campbell (1979) noted, to "act as if the null hypothesis were true" (p. 45).

Is this the case in criminal justice studies that report statistically nonsignificant results? To assess this question, we use the ES estimates described earlier. Cohen (1988) suggested that a value of .20 on this scale may be seen as a small effect, an ES of .50 a moderate effect, and an ES of .80 a large effect. Lipsey (2000) suggested that an effect size of .10 could "easily be of practical significance" (p. 109). A generally interpretable sense of the magnitude of such differences can be gained by interpreting effect sizes according to differences in the proportion of treatment success. For example, using a base rate of 40 percent, a small effect size of .20 could be interpreted as the difference between 50 percent and 40 percent, a medium effect size of .50 as the difference between 65 percent and 40 percent, and a large effect size of .80 as the difference between 78 percent and 40 percent.

Our study shows that criminal justice evaluators often make formal errors in the reporting of statistically nonsignificant findings.

In Table 3, we report the distribution of ES scores for the forty-three studies that provided sufficient information to calculate ES measures. Because many independent studies had multiple findings, two separate distributions are presented: one for the average and one for the largest effect sizes reported. In cases with only one statistically nonsignificant finding, the result is the same on each of these measures. While there is considerable variability in these distributions, many of the studies include results that are not trivial and in some cases suggest relatively larger effect sizes. This is clearly the case for the largest effect size index. Here, two-thirds of the studies have effect sizes greater than .10, Lipsey's criterion for a meaningful effect. Almost a third of the studies have effect sizes larger than .20, Cohen's definition of a small effect. Five of these studies have effect sizes greater than .40, and one greater than .50. While the percentage of studies with larger effect sizes is slightly smaller when we look at the average ES, it remains the case that only 40 percent of the studies examined have an ES between 0 and .10.

We also examined the relationship between the use of formally incorrect wording in reporting statistically nonsignificant results and effect size. It might be, for example, that researchers were unlikely to use incorrect wording when effect sizes were relatively larger. As is apparent from Table 4, this position is not supported in

TABLE 3
DISTRIBUTION OF THE ABSOLUTE VALUE OF EFFECT SIZES FOR STATISTICALLY NONSIGNIFICANT STUDIES

	Average ES		Largest ES	
	n	%	*n*	%
0 ≤ ES ≤ .1	17	40	15	35
.1 < ES ≤ .2	13	30	14	33
.2 < ES ≤ .4	12	28	9	21
.4 < ES ≤ .5	1	2	4	9
.5 < ES	0	0	1	2
Total	43	100	43	100

NOTE: ES = standardized effect size.

TABLE 4
PERCENTAGE OF STUDIES THAT HAD ANY INCORRECT WORDING BY EFFECT SIZES

	0 ≤ ES ≤ .10	.10 < ES ≤ .20	ES > .20
Study had incorrect wording	76.5	77	69
Study did not have incorrect wording	23.5	23	31
Total	100	100	100

our data. The relationship between effect size and reporting practices is relatively small and not statistically significant (tau-c = –.061, p = .68). About 77 percent of studies with effect sizes of .10 or smaller used formally incorrect wording in one of the three sections we examined in each study. This was true for the same proportion of studies with effect sizes between .10 and .20, though for a somewhat smaller percentage of studies with effect sizes greater than .20.

From these analyses, it appears that investigators were often not justified in acting as if the null hypothesis were true. Not only were effect sizes of study differences not trivial, but in many instances, authors reported these nontrivial differences as representing no difference. But it still might be argued that the studies were designed in ways that generally allowed them to detect meaningful program effects. As noted earlier, statistical power analysis provides a method for assessing how sensitive a study is in identifying program effects of a given size. If these studies evidenced high levels of statistical power for identifying modest program impacts, it would follow that the practical error of using language accepting the null hypothesis would not be a serious one.

To assess this possibility, we estimated statistical power levels for each of the findings reported in our study sample. These estimates were developed based on a .05 threshold of statistical significance and followed the authors' recommendations regarding the directionality of the test employed.[13] Sample size was drawn directly from the studies, as noted earlier. Power levels were calculated based on the type of

TABLE 5
STATISTICAL POWER USING A HYPOTHESIZED SMALL EFFECT SIZE (d = .20)

	n	%
0 ≤ power ≤ .3	26	51
.3 < power ≤ .6	11	22
.6 < power ≤ .8	5	100
.8 < power ≤ 1.0	9	18
Total	51	100

test utilized in the study and were averaged when more than one statistically nonsignificant finding was reported. We provide an estimate of statistical power based on Cohen's (1988) small estimate of ES (.20). In practice, what we ask is how sensitive these studies were to detecting a small effect if such an effect existed in the population under study. If these studies were generally very likely to detect a small effect, we might conclude that they were in a strong position to act as if the null hypothesis were true.

Table 5 reports the results of our analysis. For a standard small effect size (d = .20), we identify the number of studies that fell within four power-level groupings. It is generally accepted that the most powerful studies seek a power level of .80 or greater (e.g., see Cohen 1973; Gelber and Zelen 1985). Such studies are highly likely to evidence a significant finding based on sample statistics if the hypothesized effect examined exists in the population to which the researcher seeks to infer. When the power level of a study is very low, it means that even if there is an effect in the population of the magnitude hypothesized, the researcher is unlikely to detect it (by identifying a statistically significant result in the sample under study).

Table 5 shows that criminal justice studies that report statistically nonsignificant results seldom reach the threshold required for a statistically powerful study for identifying small program effects. Only 18 percent of the studies examined have power levels of greater than .80 when a small effect is considered. In fully half of these studies, the power level for detecting a small effect is .30 or less. In such studies, the researcher is very likely to fail to reject the null hypothesis of no difference even when the program has a standardized effect of .20 in the study population. This means that these studies are designed with little sensitivity for detecting what Cohen described as a small program impact. It is not reasonable in this context for researchers to act as if the null hypothesis were true.

Constructing an Alternative Method for Concluding That Programs or Treatments Do Not Work

Our study so far illustrates that reporting practices for statistically nonsignificant results are often formally in error and suggests that formal reporting errors often

cannot be justified using the argument that it is possible to act as if the null hypothesis were true. But we remain with the problem of defining when it is possible for a researcher evaluating a single study to conclude with confidence that the program or intervention evaluated does not work.[14] Is this conclusion possible to reach using the logic of null hypothesis statistical testing that underlies most evaluation research?

In practice, there is one scenario in which researchers commonly and correctly conclude that a program or treatment does not work and do so on the basis of a specific test of the null hypothesis (see Cook and Campbell 1979).[15] When a researcher evaluates a program or intervention and finds that there is a statistically significant backfire effect, he or she can conclude with confidence that the program did not work. A backfire effect reflects an outcome that is the opposite of that expected or desired by the criminal justice system. For example, for some of the measures of program effectiveness in the Cambridge-Somerville Youth Study, there was a statistically significant effect of treatment (see McCord 2003 [this issue]; Powers and Witmer 1951). Importantly, however, the treated participants were less likely to evidence successful outcomes than the control group participants. Similarly, in a series of experimental evaluations of Scared Straight programs, statistically significant differences were found between treatment and control conditions (Petrosino, Turpin-Petrosino, and Buehler forthcoming). Again, however, those receiving the intervention were found to have higher rates of reported deviance after the intervention than were the control group participants.

The general form of null hypothesis statistical testing presently employed by criminal justice evaluation researchers does not allow a clear method for coming to a conclusion that an intervention "does not work."

In such cases, the researcher is on solid ground in coming to a conclusion that the program did not work. Here, the researcher can reject the null hypothesis of no difference or no effect and conclude that the treatment or intervention had a harmful effect. In the Maryland Report, 11 percent of the studies reported a backfire effect (Weisburd, Lum, and Petrosino 2001). When such conclusions were based on a statistically significant finding, it would have been formally correct for the

researcher to conclude that there was statistical evidence that the program did not work.

While the case of a backfire effect illustrates how the logic of null hypothesis statistical testing can be used to conclude that programs or interventions do not work, setting a standard of a statistically significant backfire effect to come to this conclusion sets an unrealistically high threshold. If we required that a program be shown to be harmful to claim that it is not effective, we would of course be drawing this conclusion about very few studies. We believe, however, that the same logic can be used more generally and more realistically as a tool for formally deciding whether programs fail to meet a basic criterion of success.

When a statistically significant backfire effect is reported, it is based on a contrast with a null hypothesis of no difference. But what if we set the null hypothesis at a different threshold, one that represented a minimal level of program effectiveness rather than the traditional threshold of no difference? There is no statistical barrier to this readjustment. The null hypothesis can be set at any value. Our suggestion is that researchers define at the outset of a study the effect size that would be required to conclude that an intervention is worthwhile. Such an approach is suggested already in "cost effectiveness" studies (see Welsh and Farrington 2000). In such studies, the effectiveness of an intervention is calculated in terms of its specific costs to the criminal justice system or society. While this approach is often considered in terms of calculating the outcomes of an intervention, for example, the benefit gained for each police officer hired (e.g., see Levitt 1997), there is no reason policy makers, practitioners, and researchers could not define a specific level of effect at the outset that would be required for the program to be seen as a worthwhile practice (see Jacobson and Truax 1991; Lipsey 2000).

Having chosen such a threshold, the researcher could then set that outcome as the null hypothesis for defining whether the study allowed a conclusion that the treatment was ineffective or perhaps more correctly that it did not reach a minimal level of effectiveness. If the observed outcome of a test did not achieve statistical significance using the traditional no difference null hypothesis, the researcher could then test this second null hypothesis of a minimal worthwhile treatment effect. Below, we illustrate this approach using our study sample of statistically nonsignificant findings.

In practice, our proposal would require study investigators to define a minimal level of effect for each study in terms of the specific nature of the treatments and outcomes observed. Depending on the costs involved and the potential impacts on offenders, the community, or society more generally, different thresholds would likely be chosen in each specific case. However, for the sake of simplifying analysis and because we cannot make such judgments for the studies examined, we choose a standard level of effect for all the studies in our sample. We think a small effect size (ES = .20) as defined by Cohen provides a reasonable threshold. In this case, we are arguing that studies must have at least a small effect size for them to be considered cost effective.

TABLE 6
PERCENTAGE OF STUDIES SIGNIFICANTLY DIFFERENT FROM A COST-EFFECTIVE THRESHOLD (STANDARDIZED EFFECT SIZE = .20)

	n	%
Nonsignificant	19	44
Significant	24	56
Total	43	100

For each study, we adjusted the null hypothesis test employed, replacing a null hypothesis of no difference with that of a difference of .20 in favor of treatment. We use a one-tailed test of statistical significance because we are concerned only with whether treatments or programs have a smaller effect than this standard. We also use the standard .05 level of statistical significance.[16] A statistically significant finding in this case allows us to conclude that the interventions examined did not reach a minimal level of effectiveness. We were able to calculate this statistic for forty-three of the fifty-eight studies examined.

Table 6 reports our findings. In more than half of these cases, we gained a statistically significant outcome. That is, in 56 percent of the cases where researchers failed to reject the null hypothesis of no difference, we were able to reject the null hypothesis of a standardized small treatment effect. In these studies, the investigator could come to a much stronger conclusion than that allowed by simply conducting a no difference null hypothesis statistical test. A conclusion could now be reached that there is statistical evidence that the program failed to meet the minimum threshold of success that we have defined.

Conclusions

Our study shows that criminal justice evaluators often make formal errors in the reporting of statistically nonsignificant findings. Instead of simply concluding that the results were not statistically significant, or that there is not enough evidence to support an effect of treatment, they often mistakenly accept the null hypothesis and state that the intervention had no impact or did not work. We also examined whether such reporting errors might be viewed more as a formal rather than substantive error on the part of researchers. If effect sizes were trivial and the studies involved were statistically sensitive and thus likely to detect meaningful program impacts, it may have been reasonable for investigators to "act as though the null hypothesis were true" (Cook and Campbell 1979, 45). Our results show that there is little basis for this argument. The effect sizes we found in criminal justice studies reporting statistically nonsignificant results were often not trivial, and such investigations seldom met accepted thresholds for statistically powerful studies.

Our study suggests that criminal justice researchers need to be more cautious in the conclusions that they reach on the basis of statistically nonsignificant findings. One way to do this would be for criminal justice researchers to establish better defined and more rigorous standards for reporting practices, such as those developed by the American Psychological Association (Wilkinson and the Task Force on Statistical Inference 1999).[17] David Farrington (2003 [this issue]) has also suggested that criminologists pay more attention to reporting practices citing the usefulness of Lösel and Koferl's (1989) "descriptive validity," which would require clear and precise description of study results.

But irrespective of the quality of reporting practices, the general form of null hypothesis statistical testing does not allow a clear method for coming to a conclusion that an intervention does not work. This is true despite the fact that researchers are often called on by practitioners to provide insight not only about what works but also about what does not work. Drawing on conclusions that can be reached from studies that show statistically significant backfire effects, we propose that researchers define a second null hypothesis that sets a minimal threshold for program effectiveness. Such a threshold would take into account the potential costs

Our study suggests that criminal justice researchers need to be more cautious in the conclusions that they reach on the basis of statistically nonsignificant findings.

and benefits of a program and be specific to the particular intervention examined. In an illustration of this approach, we found that more than half of the studies that had no statistically significant finding for a no difference null hypothesis evidenced a statistically significant result in the case of a minimal worthwhile treatment effect null hypothesis. In such cases, there is statistical evidence that the programs failed to meet a minimal threshold of success.

Notes

1. Whether this effect is sufficient to make the intervention cost effective and thus useful in criminal justice practice is another matter. For example, relatively small effects might be found to be statistically significant in a very large sample. See Welsh and Farrington (2000) for a detailed discussion of cost-effectiveness research in criminal justice.

2. At the time the analysis for this investigation was undertaken, the updated Maryland Report, titled "Evidence Based Crime Prevention," edited by Sherman, Farrington, Welsh, and MacKenzie (2002) was not yet available.

3. To be included in the Maryland Report, studies had to be methodologically rigorous enough to be rated at least a 1 on a scientific methods scale, a 1 being a correlation between a crime prevention program and a measure of crime or crime risk factors.

4. The Maryland Report authors did not develop a central archive of the studies that were used. Therefore, we searched for reports and articles (some unpublished) using the University of Maryland Library System and the National Criminal Justice Reference Service Library in Rockville, Maryland. We also consulted with the original authors of the Maryland Report and attempted to contact the actual authors of studies whom we could not locate.

5. These conversions were carried out with the assistance of formulas provided by Lipsey and Wilson (2001) as well as David Wilson's Effect Size Calculator, located at http://mason.gmu.edu/~dwilsonb/ma.html.

6. In eighteen of the studies, the Maryland Report findings pointed to multiple yet related statistically nonsignificant results in the actual research. For example, a finding reported in the Maryland Report of "no appreciable effect on crime" may be reported in the actual research as nonsignificant results of four types of crimes.

7. Following Cohen (1988), we use adjusted sample size (n′) for calculating power estimates:

$$n' = \frac{2n_A n_B}{n_A + n_B}$$

8. Statistical power estimates were gained by defining a specific threshold of effect, the type of test employed, the significance criterion used, and the sample size examined (see Cohen 1988). We discuss statistical power and the methods used to gain power estimates later in our article.

9. The difference in the proportions of incorrect wording between the results and the conclusions sections is statistically significant at $p < .01$, and that between the results section and the abstract has a p value of .05.

10. The relationship between year of publication and reporting practices was very small in each of the three sections of the report, with p values much greater than conventional thresholds.

11. We also looked at the relationship between reporting of statistically nonsignificant findings and the Maryland Scientific Methods Score (see Sherman et al. 1997, 2.18-2.19). In both the abstract and the results section, the relationship between the scientific methods scale score and the wording used to report statistically nonsignificant results were not statistically significant (tau-c = –.091, $p = .624$, and tau-c = .109, $p = .354$, respectively). Yet in the conclusions section, there is a negative and statistically significant relationship between the scientific methods scale score and the wording used (tau-c = –.288, $p < .05$), suggesting that studies with increased methodological rigor are more likely to use formally incorrect wording.

12. Interestingly, Finch, Cumming, and Thomason (2001) also did not find a statistically significant relationship between the year of publication of results and reporting practices.

13. In many cases, authors did not describe whether a one- or two-tailed test was used. When possible, we relied on other components of the analysis to make this decision. When information was not available, we assumed a two-tailed test.

14. Importantly, we are concerned here with statistical generalizations regarding a specific program. Most scholars agree that multiple studies in multiple sites are required for reaching more general policy conclusions (e.g., see Cook et al. 1992; Greenberg, Meyer, and Wiseman 1994; MacKenzie and Souryal 1994; McShane, Williams, Wagoner 1992; Weisburd and Taxman 2000).

15. Cook and Campbell (1979) suggested such an approach in their book on quasi-experimentation. They noted that

> when an explicit directional hypothesis guides the research, it is sometimes possible to conclude with considerable confidence that the derived effect was not obtained under the conditions in which the testing occurred. This conclusion is easiest to draw when the results are statistically significant and in the opposite direction to that specified in the hypothesis. (P. 45)

16. First, we defined any value of average standardized effect size of .20 or greater as not meeting our significance threshold since these values are all greater than the criterion for our test. We then subtracted .20 from the nonabsolute value of the effect size for each remaining study. We then compared this result to D_c, the significance criterion that Cohen (1988) provided. If our result was larger than D_c, then we coded that study as having an outcome that is significantly different from the requirement of a minimal worthwhile treatment effect under a small effect size expectation.

17. For example, many scholars have suggested that standardized effect sizes and confidence intervals be required in the reporting of study findings (Cohen 1994; Finch, Cumming, and Thomason 2001; Lipsey 2000; Schmidt and Hunter 2002; Rosenthal 1995).

References

Alderson, Phil, and Iain Chalmers. In press. Claims in abstracts of Cochrane reviews that health care interventions have "no effect." *British Medical Journal*.

Anderson, David, Kenneth Burnham, and William Thompson. 2000. Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management* 64:912-23.

Bakan, David. 1966. The test of significance in psychological research. *Psychological Bulletin* 66:423-37.

Chow, Siu L. 1998. Précis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences* 21:169-239.

Cohen, Jacob. 1973. Statistical power analysis and research results. *American Educational Research Journal* 10:225-30.

———. 1988. *Statistical power analysis for the behavioral sciences*. 2d ed. Hillsdale, NJ: Lawrence Erlbaum.

———. 1994. The earth is round ($p < .05$). *American Psychologist* 49:997-1003.

Cook, Thomas, and Donald Campbell. 1979. *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cook, Thomas, Harris Cooper, David Cordray, Heidi Hartmann, Lawrence Hedges, Richard Light, Thomas Louis, and Frederick Mosteller. 1992. *Meta-analysis for explanation: A casebook*. New York: Russell Sage.

Cook, Thomas, Charles Gruder, Karen Hennigan, and Brian Flay. 1979. The history of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. *Psychological Bulletin* 86:662-79.

Davies, Huw T. O., Sandra Nutley, and Peter Smith. 2000. *What works: Evidence-based policy and practice in public services*. London: Policy Press.

Farrington, David P. 2003. Methodological quality standards for evaluation research. *Annals of the American Academy of Political and Social Science* 587:49-68.

Finch, Sue, Geoff Cumming, and Neil Thomason. 2001. Reporting of statistical inference in the "Journal of Applied Psychology": Little evidence of reform. *Educational and Psychological Measurement* 61:181-210.

Gelber, R., and M. Zelen. 1985. Planning and reporting clinical trials. In *Medical oncology: Basic principals and clinical management of cancer*, edited by Paul Calabrese, Philip Schein, and S. Rosenberg. New York: Macmillan.

Greenberg, David, Robert Meyer, and Michael Wiseman. 1994. Multisite employment and training program evaluations: A tale of three studies. *Industrial Labor Relations Review* 47:679-91.

Jacobson, Neil, and Paula Truax. 1991. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology* 59:12-19.

Johnson, Douglas. 1995. Statistical sirens: The allure of nonparametrics. *Ecology* 76:1998-2000.

Levitt, Steven. 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review* 87:270-90.

Lipsey, Mark. 1998. Design sensitivity: Statistical power for applied experimental research. In *Handbook of applied social research methods*, edited by Leonard Bickman and Debra Rog. Thousand Oaks, CA: Sage.

———. 2000. Statistical conclusion validity for intervention research: A significant ($p < .05$) problem. In *Validity and social experimentation: Donald Campbell's legacy*, edited by Leonard Bickman. Thousand Oaks, CA: Sage.

Lipsey, Mark, and David Wilson. 2001. *Practical meta-analysis*. Applied social research methods series 49. Thousand Oaks, CA: Sage.

Lösel, Friedrich, and Peter Koferl. 1989. Evaluation research on correctional treatment in West Germany: A meta-analysis. In *Criminal behavior and the justice system: Psychological perspectives*, edited by Hermann Wegener, Friedrich Lösel, and Jochen Haisch. New York: Springer-Verlag.

MacKenzie, Doris. 2000. Evidence-based corrections: Identifying what works. *Crime and Delinquency* 46:457-71.

MacKenzie, Doris, and Claire Souryal. 1994. *Multisite evaluation of shock incarceration: Evaluation report*. Washington, DC: National Institute of Justice.

McCord, Joan. 2003. Cures that harm: Unanticipated outcomes of crime prevention programs. *Annals of the American Academy of Political and Social Science* 587:16-30.

McShane, Marilyn, Frank Williams, and Carl Wagoner. 1992. Prison impact studies: Some comments on methodological rigor. *Crime & Delinquency* 38:105-20.

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 and Management* 19:47-54.

Petrosino, Anthony, Carolyn Turpin-Petrosino, and John Buehler. Forthcoming. The effects of Scared Straight and other juvenile awareness programs on delinquency: A systematic review of randomized controlled trials. *Annals of the American Academy of Political and Social Science*.

Police Foundation. 1981. *The Newark foot patrol experiment*. Washington, DC: Police Foundation.

Powers, Edwin, and Helen Witmer. 1951. *An experiment in the prevention of delinquency: The Cambridge-Somerville Youth Study*. New York: Columbia University Press.

Rosenthal, Robert. 1991. *Meta-analytic procedures for social research*. Applied social research methods series 6. Newbury Park, CA: Sage.

———. 1995. Progress in clinical psychology: Is there any? *Clinical Psychology: Science and Practice* 2:133-50.

Schmidt, Frank, and John Hunter. 2002. Are there benefits from NHST? *American Psychologist* 57:65-71.

Sherman, Lawrence, David Farrington, Brandon Welsh, and Doris MacKenzie, eds. 2002. *Evidence based crime prevention*. New York: Routledge.

Sherman, Lawrence, Denise Gottfredson, Doris MacKenzie, John Eck, Peter Reuter, and Shawn Bushway. 1997. *Preventing crime: What works, what doesn't, what's promising: A report to the United States Congress*. Washington, DC: National Institute of Justice.

———. 1998. *Preventing crime: What works, what doesn't, what's promising. Research in brief*. Washington, DC: National Institute of Justice.

Weisburd, David, with Anthony Petrosino and Gail Mason. 1993. Design sensitivity in criminal justice experiments. In *Crime and justice: A review of research 17*, edited by Michael Tonry. Chicago: University of Chicago Press.

Weisburd, David, and Chester Britt. 2002. *Statistics in criminal justice*. 2d ed. Belmont, CA: Wadsworth.

Weisburd, David, Cynthia M. Lum, and Anthony Petrosino. 2001. Does research design affect study outcomes in criminal justice? *Annals of the American Academy of Political and Social Science* 578:50-70.

Weisburd, David, and Faye Taxman. 2000. Developing a multicenter randomized trial in criminology: The case of HIDTA. *Journal of Quantitative Criminology* 16:315-40.

Welsh, Brandon, and David Farrington. 2000. Monetary costs and benefits of crime prevention programs. In *Crime and justice: A review of research 27*, edited by Michael Tonry. Chicago: University of Chicago Press.

Wilkinson, Leland, and the Task Force on Statistical Inference. 1999. *American Psychologist* 54:594-604.

Zuger, Abigail. 1997. New way of doctoring: By the book. *The New York Times*, 16 December.