J Exp Criminol (2011) 7:183–198
DOI 10.1007/s11292-010-9117-1

Research design influence on study outcomes in crime and justice: a partial replication with public area surveillance

Brandon C. Welsh · Meghan E. Peel · David P. Farrington · Henk Elffers · Anthony A. Braga

Published online: 12 October 2010

Abstract Does the quality of research design have an influence on study outcomes in crime and justice? This was the subject of an important study by Weisburd et al. (2001). They found a moderate and significant inverse relationship between research design and study outcomes: weaker designs, as indicated by internal validity, produced stronger effect sizes. Using a database of evaluations (n=136) from systematic reviews that investigated the effects of public area surveillance on crime, this paper carried out a partial replication of Weisburd et al.'s study. We view it as a partial replication because it included only area- or place-based studies (i.e., there were no individual-level studies) and these studies used designs at the lower end of the evaluation hierarchy (i.e., not one of the studies used a randomized experimental design). In the present study, we report findings that are highly concordant with the earlier study. The overall correlation between research design and study outcomes is moderate but negative and significant (Tau-b= –.175, p=.029). This suggests that stronger research designs are less likely to report desirable effects or, conversely, weaker research designs

B. C. Welsh (✉) · M. E. Peel
School of Criminology and Criminal Justice, Northeastern University, Churchill Hall, 360 Huntington Avenue, Boston, MA 02115, USA
e-mail: b.welsh@neu.edu

D. P. Farrington
Cambridge University, Cambridge, UK

H. Elffers
Netherlands Institute for the Study of Crime and Law Enforcement, Amsterdam, The Netherlands

A. A. Braga
Rutgers University, Newark, NJ, USA

A. A. Braga
Harvard University, Cambridge, MA, USA

may be biased upward. We explore possible explanations for this finding. Implications for policy and research are discussed.

Keywords Evaluation design · Evidence-based crime policy · Public area surveillance · Systematic review

Does the quality of research design have an influence on study outcomes in crime and justice? This was the subject of an important study by Weisburd et al. (2001). On the heels of the launching of the Campbell Collaboration (in February 2000) and the beginning of expanded use of systematic reviews to assess the effectiveness of criminological interventions and inform public policy (see Farrington and Welsh 2001; Sherman 2003; Weisburd et al. 2003), the authors were chiefly motivated by two central methodological issues that concerned an evidence-based approach to crime and justice. The first was that research design matters. Higher-quality evaluations, as indicated by internal validity, provide more confidence in observed effects, and randomized experiments have the highest internal validity (Farrington and Welsh 2005, 2006; Shadish et al. 2002). The second issue was that criminal justice, with few exceptions (e.g., Braga 2005; Sherman et al. 2005), does not have the luxury (unlike medicine) to base policies and practices solely on evidence from randomized experiments. Instead, it must consider a wider range of evaluation designs (Mears 2007, 2010). On this point, the words of the authors are instructive: "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" (Weisburd et al. 2001: 52).

Weisburd et al. (2001) found a moderate and significant inverse relationship between research design and study outcomes. This meant that weaker research designs, as indicated by internal validity, produced stronger effect sizes. For example, the weakest research design, a correlation between a program and a measure of crime, had a mean score of .80 on the authors' investigator reported result scale (the highest and lowest possible scores were 1 and –1, respectively) compared to the strongest design of a randomized experiment with a mean score of .22. In fact, the inverse relationship was linear, with each successive type of research design (from weak to strong) showing a smaller effect. In addition to weaker research designs being more likely to report positive or desirable effects, these designs were less likely to report harmful or backfire effects. The study was based on 308 evaluations that were included in the now-famous Maryland report on what works and what does not in preventing crime (Sherman et al. 1997).

The authors concluded that their results "point to the possibility of an overall bias in nonrandomized criminal justice studies" (Weisburd et al. 2001: 65). While they cautioned that this bias may be due to a number of factors that they were not able to investigate (e.g., publication bias, differential attrition), they were more confident that an explanation for their results could be found in the "norms of criminal justice research and practice" (p. 65). By this the authors meant that compared to the routinized and prescribed nature of randomized experiments, nonrandomized studies can more easily fall prey to selection bias and other confounding factors, which criminal justice practitioners find more difficult to address than their counterparts in medicine, for example.

A large body of evaluation research on surveillance methods designed to prevent crime in public space exists. Closed-circuit television (CCTV) and improved street lighting are the most well developed of these measures that are in current use. Other widely used surveillance measures that perform a crime-prevention function in public space and that have been evaluated include security guards, place managers, and defensible space.[1] Recently completed systematic reviews of these forms of public area surveillance, which incorporated only the highest-quality evaluation designs (Farrington and Welsh 2007; Welsh and Farrington 2009a, b; Welsh et al. 2010), presented an opportunity to investigate the relationship between research design and study outcomes.

A brief summary of the results of these reviews seems warranted. It was found that CCTV is effective in car parks, improved street lighting is effective in city and town centers and residential/public housing communities, and the defensible space practice of street closures or barricades is effective in inner-city neighborhoods. Also of importance is evidence showing that CCTV and improved street lighting are more effective in reducing property (and especially vehicle) crimes than in reducing violent crimes. Street closure or barricade schemes are effective in reducing both property and violent crimes. For security guards, the weight of the evidence suggests that they are promising when implemented in car parks and targeted at vehicle crimes. In contrast, place managers appear to be of unknown effectiveness in preventing crime. These less-than-conclusive statements about the effectiveness of security guards and place managers have everything to do with the small number of high-quality evaluations that have been carried out on these measures.[2]

Drawing upon the evaluations collected for these systematic reviews, we set out to conduct a replication of Weisburd et al.'s (2001) study. Two factors suggest that our sample could serve as the basis of a suitable replication. First, public area surveillance represents a well-known group of criminological interventions, within the more general area of situational crime prevention (Cornish and Clarke 2003). Second, the group of studies brought together here represents the full body of evaluation research (published and unpublished) on public area surveillance (see the Methodology section for more details). The present study also goes some way toward addressing Lipsey's (2003) call for more uniform bodies of research to be used in investigating the relationship between research design and study outcomes.

[1] Briefly, place managers (Eck 1995) are persons such as bus drivers, parking lot attendants, train conductors, and others who perform a surveillance function by virtue of their position of employment. Unlike security personnel, however, the task of surveillance for these employees is secondary to their other job duties. Defensible space (Newman 1972) involves design changes to the built environment to maximize the natural surveillance of open spaces (e.g., streets and parks) provided by people going about their day-to-day activities. Examples of design changes include the construction of street barricades or closures, re-design of walkways, and installation of windows. They can also include more mundane techniques such as the removal of objects from shelves or windows of convenience stores that obscure lines of sight in the store and the removal or pruning of bushes in front of homes so that residents may have a clear view of the outside world (Cornish and Clarke 2003).

[2] For more details on the results, interested readers should consult the separate reviews or the larger study that includes the full body of this work (see Welsh and Farrington 2009a).

Literature review

An evaluation of an intervention is considered to be high quality if it possesses a high degree of internal, construct, and statistical conclusion validity. Put another way, we can have a great deal of confidence in the observed effects of an intervention if it has been evaluated using a design that controls for the major threats to these three forms of validity.[3] Internal validity, which refers to how well the study unambiguously demonstrates that an intervention (e.g., parent training) had an effect on an outcome (e.g., delinquency), has generally been regarded as the most important type of validity (Shadish et al. 2002: 97).[4] Here, some kind of control condition is necessary to estimate what would have happened to the experimental units (e.g., people or areas) if the intervention had not been applied to them—termed the "counterfactual inference." Internal validity is central to the discussion on the quality or strength of research design.

The main threats to internal validity include selection, history, maturation, instrumentation, testing, regression to the mean, differential attrition, and causal order (see Farrington 2003: 53). In addition, there may be interactive effects of threats. For example, a selection-maturation effect may occur if the experimental and control conditions have different preexisting trends, or a selection-history effect may occur if the experimental and control conditions experience different historical events (e.g., where they are located in different settings).

In principle, a randomized experiment has the highest possible internal validity because it can rule out all these threats, although in practice, differential attrition may still be problematic. Randomization is the only method of assignment that controls for unknown and unmeasured confounders as well as those that are known and measured. The next best method is to use a nonrandomized experimental or quasi-experimental design. In the former, experimental and control units are matched or statistically equated (e.g., using a prediction or propensity score) prior to intervention. One example of a quasi-experimental design involves before-and-after measures in experimental and comparable control conditions. According to Cook and Campbell (1979) and Shadish et al. (2002), this is the minimum design that is interpretable or understandable. Control conditions are needed to counter threats to internal validity.

Weisburd et al. (2001) reviewed the key studies on the relationship between research design and study outcomes and drew a number of important conclusions.

[3] External validity, which refers to how well the effect of an intervention on an outcome is generalizable or replicable in different conditions, is difficult to investigate within one evaluation study. External validity can be established more convincingly in systematic reviews and meta-analyses of a number of evaluation studies.

[4] Statistical conclusion validity is concerned with whether the presumed cause (the intervention) and the presumed effect (the outcome) are related. The main threats to this form of validity are insufficient statistical power to detect the effect (e.g., because of small sample size) and the use of inappropriate statistical techniques. Construct validity refers to the adequacy of the operational definition and measurement of the theoretical constructs that underlie the intervention and the outcome. The main threats to this form of validity rest on the extent to which the intervention succeeded in changing what it was intended to change (e.g., to what extent was there treatment fidelity or implementation failure) and on the validity and reliability of outcome measures (e.g., how adequately police-recorded crime rates reflect true crime rates).

First, despite broad consensus that experimental studies are stronger than quasi-experimental or non-experimental studies with respect to internal validity, researchers face difficulties in specifying the effect of research design on study outcomes. It is assumed that weaker research designs will lead to biases in assessment of program effects; however, the directionality and magnitude of the bias is contingent on a number of factors related to the nature of the research. Second, in comparisons of randomized versus nonrandomized (quasi-experimental and non-experimental studies) designs, there is "not a consistent bias" that results from the use of the latter, and any differences in effects between the two research designs will be reduced when studies with nonrandomized designs are "well designed and implemented" (p. 55). The finding that there is not a consistent bias but rather mixed results between these designs is found across disciplines, including medical and social sciences. Third, efforts to investigate the relationship between research design and study outcome are scarce in crime and justice, and the studies that have been done are limited. In the words of Weisburd et al.: "Results of these studies provide little guidance for specifying a general relationship between study design and study outcomes for criminal justice research" (p. 56).

We set out to update Weisburd et al.'s (2001) literature review. The last 10 years of criminology and criminal justice journals were searched using the major electronic bibliographic databases (e.g., Criminal Justice Abstracts, National Criminal Justice Reference Service Abstracts) to locate studies that examined the effect of research design on crime and offending outcomes. Forward searches were carried out to identify studies that cited Weisburd et al. and related articles. We also contacted each of the three study authors and other key scholars in the field in an effort to identify any works (published or unpublished) on the topic.

Interestingly, not one new study was found that explicitly investigated the relationship between research design and study outcomes in crime and justice. Ten years later, we are left to concur with Weisburd and his colleagues that there is a real scarcity of works on this topic in crime and justice. Two studies are noteworthy.

Cook et al. (2008) compared estimates from randomized experiments with estimates from observational studies covering a wide range of areas (e.g., education, job training, school dropout prevention). An important feature of this research is that it involved within-study comparisons where the randomized experiments and observational studies shared treatment groups. The aim was to "test whether different causal estimates result when a counterfactual group is formed, either with or without random assignment, and when statistical adjustments for selection are made in the group from which random assignment is absent" (p. 724). The authors found that the two designs produced comparable results when the observational studies are well designed and analyzed, but not when the observational studies are poorly designed and analyzed.

Lipsey (2003) explored the difficulties of investigating and interpreting moderator variables in meta-analyses, with a particular focus on research design. He argued that, because moderator variables are generally related to each other and to the effect of the intervention (or effect size), it can be difficult to determine the influence of a single moderator on effect size. Lipsey referred to this as the confounding effect of moderator variables.

Drawing upon his delinquency intervention database, which includes almost 400 studies, Lipsey carried out a number of analyses to investigate the relationship between effect size and research design (i.e., random or nonrandom) and other key moderator variables (e.g., research demonstration or routine practice, age of sample, and intervention intensity). While analyses showed that effect sizes were significantly larger for weaker designs, he cautioned against concluding that "nonrandomized designs are biased upward" (p. 74). This is because effect-size differences were also significantly related to other important moderator variables, in some cases more strongly than to research design. In the context of meta-analysis, Lipsey called for a "very careful multivariate analysis with the full set of correlated moderator variables (if we could identify and measure all such variables)" to test if in fact nonrandomized designs are biased upward (p. 77).

Methodology

As noted above, five recently completed systematic reviews of public area surveillance measures served as the basis for the population of studies used here. These reviews resulted in the collection of a total of 150 unique evaluation studies (i.e., each representing a different project). A small number of these evaluations did not measure crime, some of the experimental-control evaluations only reported measures of crime for the target area, and a couple were carried out in private settings. Exclusion of these reduced the sample to 136 evaluations.

One of the differences between our study and Weisburd et al.'s is that we collected, coded, and analyzed all of the evaluations in our sample. Of course, this is a function of carrying out a systematic review. Weisburd and his colleagues had to rely on the results reported by the chapter authors in the Maryland report (Sherman et al. 1997). They termed this the investigator reported result (IRR). To reflect this difference, we titled our measure of results the systematic reviewer result or SRR. The SRR was also created as an ordinal scale with three numerical values corresponding to the main study outcomes: positive or desirable effect (1), null effect (0), and negative or backfire effect (–1). The conventional $p < .05$ was used as the significance level cut-off. We followed the same rules for coding studies as used by Weisburd et al. (2001: 58–59):

1: The program had an intended positive effect in the study sample/population. Outcomes in this case supported the position that interventions lead to reductions in crime.
0: The program was reported to have no detected effect, or the effect was reported as not statistically significant.
–1: The program had an unintended backfire effect in the study sample/population. Outcomes in this case supported the position that interventions were harmful and lead to increases in crime.

In some of our systematic reviews (on CCTV and street lighting), it was possible to carry out a meta-analysis. The starting point for a meta-analysis is the calculation of a comparable measure of effect size and an estimate of its variance for each

evaluation (Lipsey and Wilson 2001). For our purposes here, this information served as the SRR for each evaluation.

In the case of public area surveillance evaluations, the measure of effect size had to be based on the number of crimes in the experimental and control areas before and after the intervention. This is because this was the only information that was regularly provided in these evaluations. Here, the "relative effect size" or RES (which can be interpreted as an incident rate ratio) is used to measure effect size. The RES is calculated from the following table:

	Before	After
Experimental	a	b
Control	c	d

where a, b, c, d are numbers of crimes

$$\mathrm{RES} = a*d/b*c$$

The RES is intuitively meaningful because it indicates the relative change in crimes in the control area compared with the experimental area. RES=2 indicates that d/c (control after/control before) is twice as great as b/a (experimental after/experimental before). This value could be obtained, for example, if crimes doubled in the control area and stayed constant in the experimental area, or if crimes decreased by half in the experimental area and stayed constant in the control area, or in numerous other ways. A RES greater than 1.0 indicates a desirable effect of the intervention, and a RES less than 1.0 indicates an undesirable effect. The significance of RES, and its confidence intervals, are calculated by calculating the natural logarithm of RES, LRES, which has a known variance:

$$\mathrm{VAR(LRES)} = 1/a + 1/b + 1/c + 1/d$$

Not all of the included evaluations used a control condition. For these uncontrolled studies, a slightly modified procedure was used to calculate effect size and variance. Here, the RES was calculated by dividing crimes in the before period by crimes in the after period, or a/b. This assumes that crimes did not change in the control condition (the missing c and d). The variance of RES is calculated as follows:

$$\mathrm{VAR(LRES)} = [1/a + 1/b]*2$$

In a handful of cases, it was not possible to calculate an effect size and its variance using the above procedures. In these cases, the SRR had to be based on the effect size (e.g., relative difference in percentages) that was reported in the study.

Another important component of this study is the type of research designs under investigation. Nearly all of the studies used one of three designs: no control, non-comparable control, or comparable control. These correspond approximately to levels 1, 2, and 3, respectively, on the Scientific Methods

Scale (SMS) (Sherman et al. 1997; see also Farrington et al. 2006).[5] Not one of the studies used a randomized experimental design. Weisburd et al. were able to use the full range of evaluation designs, from level 1 (no control or correlational) to level 5 (randomized experiment) on the SMS.

One other important component of this study—another that distinguishes it from Weisburd et al.'s study—is that it includes only area- or place-based evaluations. Most evaluations of situational crime-prevention programs are what are called area-based studies, where the unit of analysis is the area. In these studies, the effect of crime on the area or place (e.g., city center, public transportation facility, school) is measured, rather than the effect of crime on the individual, which is assessed in commonly used evaluation studies. In area-based studies, the best and most feasible design usually involves before-and-after measures of crime in experimental and comparable control conditions, together with statistical control of extraneous variables. This is an example of a quasi-experimental evaluation design. Few area-based studies have used experimental designs (Weisburd 2005). It is for these reasons that we view the present study as a partial replication of the one by Weisburd et al. (2001).

Results

Descriptive statistics of the studies are presented in Tables 1 and 2. As might be expected, there was a fair degree of variability in the research designs of the 136 studies (see Table 1). A little more than four out of ten (44%) studies used a comparable control design (the most rigorous one represented here) to evaluate the effects of the intervention on crime. Just over one-third (37%) of the studies employed a no control design, and approximately one-fifth (19%) of the studies used a non-comparable control design. This distribution of research designs differs from that reported in Weisburd et al. (2001). While they found that an almost identical percentage of studies used a comparable control design (42%), a larger share used a non-comparable control design (31%) and a considerably smaller share used the weakest design (3%). A more accurate comparison requires the removal of the two strongest research designs used in their study, represented by levels 4 and 5 on the SMS. This had the effect of making the distribution of research designs even less similar between our study and Weisburd et al.'s; for example, comparable control designs now accounted for more than half (56%) of all their studies.

Study outcomes, as categorized by the SRR, are shown in Table 2. Nearly half (49%) of the studies reported a positive or desirable effect on crime. A slightly smaller percentage of studies reported a null effect (43%), and only 7% reported a negative or backfire effect. Weisburd et al. (2001) also found that the greatest share of studies reported a positive effect (64%) and the smallest share reported a negative effect (11%).

[5] Two street lighting studies that used a comparable control design (Painter and Farrington 1997, 1999) also controlled for extraneous variables. Eck (2006) rated them as level 4 on the SMS. A level 4 study involves measures of crime before and after the program in experimental and comparable control conditions, together with statistical control of extraneous variables.

Table 1 Studies categorized by research design

Research design	Studies	
	n	Percentage
No control	50	37
Non-comparable control	26	19
Comparable control	60	44
Total	136	100

Basic findings on the relationship between research design and study outcomes are presented in Tables 3 and 4 (these results are presented in a similar manner to those reported in Weisburd et al.). Table 3 reports on the mean SRR scores across the three types of research design. The mean SRR scores and an overall statistical measure of correlation, Tau-b, allow for an examination of the results. Table 4 reports on a cross-tabulation of research design and SRR scores.

Both tables indicate that there is a significant inverse relationship between research design and the SRR. The most internally valid studies, as represented by comparable control designs, have a mean SRR score of .28, while the weakest studies, as represented by no control designs, have a mean SRR score of .50 (see Table 3). There is little difference between the mean SRR scores for the no-control and non-comparable control studies (.50 vs. .58). The overall correlation between research design and study outcomes is moderate and negative (−.175), and the relationship is statistically significant at the .05 level (p=.029). This suggests that stronger research designs with respect to internal validity are less likely to report desirable effects or, conversely, weaker research designs may be biased upward.

Table 4 provides a cross-tabulation of research design type and the SRR. Stronger research designs with respect to internal validity are less likely to conclude that the intervention had a desirable effect. Negative or backfire effects were found in five studies (10%) with a no control design and five studies (8%) with a comparable control design; none of the non-comparable control design studies reported a negative effect. Desirable effects were found in 30 studies (60%) with a no control design, 15 studies (58%) with a non-comparable control design, and 22 studies (37%) with a comparable control design. Going from weakest to strongest research design (left to right in Table 4) there is a linear increase in the percentage of studies with null effects, while the reverse is true for studies with

Table 2 Studies categorized by the SRR

SRR	Studies	
	n	Percentage
–1	10	7
0	59	43
1	67	49
Total	136	100*

*Does not add due to rounding

Table 3 Mean SRR scores across research design categories

Research design	Mean	*n*	Standard deviation
No control	.50	50	0.678
Non-comparable control	.58	26	0.504
Comparable control	.28	60	0.613
Total	.42	136	0.627

Tau-b=−.175, $p<.05$

desirable effects. The overall relationship observed in Table 4 is statistically significant at the .05 level ($p=.042$).

In Table 5, we combine the two weakest research designs (no control and non-comparable control) for a comparison with the strongest or most valid in the comparable control design. Our reasoning is that because the latter is considered the minimum interpretable design for concluding that an intervention caused a change in the outcome, any research design below this threshold will produce results that are inconclusive. The combined no-control and non-comparable control studies have a mean SRR score of .53 and the comparable control studies have a mean SRR score of .28 (data drawn from Table 3). As shown in Table 5, an almost identical percentage of studies in the two groups reported a negative effect on crime, while there were large differences between the two groups in the percentage of studies with null and desirable effects. For example, desirable effects on crime were found in just over one-third (37%) of comparable control studies and almost three-fifths (59%) of the combined no control and non-comparable control studies. Once again, we find a statistically significant negative relationship between research design and the SRR ($p=.031$).

Discussion and conclusions

In the first direct, not to mention the most rigorous and largest, test of the influence of research design on study outcomes in crime and justice, Weisburd et al. (2001) found a moderate and significant inverse relationship between research design and

Table 4 Cross-tabulation of research design and SRR

	Research design					
	No control		Non-comparable control		Comparable control	
SRR	*n*	Percentage	*n*	Percentage	*n*	Percentage
–1	5	10	0	0	5	8
0	15	30	11	42	33	55
1	30	60	15	58	22	37
Total	50	100	26	100	60	100

Chi-square = 9.882 with 4 *df* ($p<.05$)

Table 5 Comparing no-control and non-comparable control studies with comparable control studies

	Research design			
	No control/Non-comparable control		Comparable control	
SRR	*n*	Percentage	*n*	Percentage
–1	5	7	5	8
0	26	34	33	55
1	45	59	22	37
Total	76	100	60	100

Chi-square = 6.940 with 2 *df* ($p<.05$)

study outcomes. Weaker research designs, as indicated by internal validity, were more likely to report desirable effects and were less likely to report backfire or harmful effects. In the present study, we report findings that are highly concordant with Weisburd et al.'s study. The overall correlation between research design and study outcomes is moderate but negative and significant.

The main aim of this paper was to carry out a replication of Weisburd et al.'s study. We view it as a partial replication because it included only area- or place-based studies (i.e., there were no individual-level studies) and these studies used designs at the lower end of the evaluation hierarchy (i.e., not one of the studies used a randomized experimental design). Two other major differences exist between the present study and Weisburd et al.'s. First, we focused on a specific area of criminological intervention in the form of public area surveillance and, second, we used a database of evaluations from a number of systematic reviews that we have carried out on this topic. This means that the group of studies brought together here represents the full body of evaluation research (published and unpublished) on public area surveillance, and we collected, coded, and analyzed all of the evaluations in our sample. By contrast, Weisburd et al. focused on a broad spectrum of criminological interventions and had to rely on the results reported in the Maryland report.

We certainly do not believe that research design alone—or any one other moderator variable for that matter—can explain our main finding. As reported above, Lipsey (2003), in the context of a meta-analysis on delinquency intervention, found that effect size differences were also significantly related to other important moderator variables, in some cases more strongly than research design. There are, of course, some key factors that may contribute to this research design effect that merit investigation. One of these is publication bias. Weisburd et al. acknowledged that they were unable to look into this. One of the advantages of systematic reviews is that they are meant to include both published and unpublished studies. At the very least, this allows one to investigate for publication bias. We found no evidence of publication bias in our systematic reviews of the effects of public area surveillance on crime.

Another key factor is program implementation. Poor implementation is considered the most common reason for a program's failure to demonstrate an

impact on crime or, more correctly, "for an impact evaluation to fail to put the 'theory' to a fair test" (Ekblom and Pease 1995: 594). Detailed information on implementation is sometimes lacking in evaluations, and our sample is no different, but we found no evidence to suggest that the programs with the highest quality evaluation design (i.e., comparable control), which had the lowest mean SRR score, were more likely to be poorly implemented than those with the lower-quality evaluation designs (i.e., no control and non-comparable control).

We are apt to concur with Weisburd and his colleagues (2001) that the upward bias of weaker evaluation designs is rooted in the "norms of criminal justice research and practice" (p. 65). As noted above, this has to do with the conduct of nonrandomized studies and their increased susceptibility to selection bias and other confounding factors, which criminal justice practitioners find more difficult to address than their counterparts in medicine.

Rather than being seen as a defeatist position, this may well be a marker of the still relatively new tradition of the application of the scientific method to evaluating interventions in crime and justice (compared to medicine). It may also go some way to highlighting the often messy nature of conducting evaluations in the field and with multiple stakeholders of sometimes differing interests. These points draw attention to a key question for future research: Is it that crime and justice researchers are coming from disciplines where training in methods is weaker or does the nature of our context influence the outcomes? Where few procedures or protocols exist to guide these evaluations, unlike the more routinized and prescribed randomized experiments, methodological quality control can fall by the wayside. Moreover, it may raise the need for increased practitioner-academic partnerships, which have made some important advances in criminal justice research in recent years (see Braga 2010).

A growing number of scholars suggest that academics should engage in ongoing collaborations with practitioners to create opportunities to test crime and justice interventions using more rigorous designs with higher degrees of internal validity (Braga 2010; Petersilia 2008). Past partnerships between academics and criminal justice practitioners were often characterized by role conflicts, such as researchers reporting the "bad news" that an evaluated program was not effective in preventing crime (Weisburd 1994). For academic researchers, success or failure matters less than their commitment to the development of knowledge on what does or does not work in preventing crime. For criminal justice practitioners, this news could be interpreted as their personal failure, and the skepticism of academics may be viewed as irritating. Traditional research and evaluation roles played by academics, often involving data collection and analysis *after* programs have been developed, can also be viewed by practitioners as not particularly helpful in their efforts to do a better job in preventing crime.

Current partnerships between academics and practitioners are more likely to be characterized by action research methodologies and a shared sense of responsibility for implementing locally relevant crime prevention programs and understanding whether these implemented programs actually work in preventing crime (Braga and Hinkle 2010; Klofas et al. 2010). For instance, in the realm of public area surveillance evaluations, researchers could partner with city officials to pinpoint the optimal locations for CCTV placement through strategic analyses of crime data.

Budget-conscious cities often lack the resources to adequately cover all existing high-crime locations. Researchers would then be well positioned to suggest a deployment scheme that could randomly allocate CCTV cameras across the identified high-crime locations (and thereby create equivalent treatment and control conditions). A case could be made that this more rigorous design would provide city officials with an equitable distribution of existing scarce resources and, if the evaluation reported evidence of crime prevention benefits, city officials would also be better positioned to acquire external funds to then cover the control locations that did not receive the cameras during the study time period.

The overall state of crime and justice research, which is generally methodologically weak (National Research Council 2008), could be improved through this new wave of collaborative relationships as more opportunities will arise to conduct rigorous evaluations of interventions. Research partnerships allow academics to get their feet in the door, develop trust with practitioners, and position themselves to make a stronger argument for using rigorous evaluation designs such as randomized controlled trials. Petersilia (2008) suggests that policy-makers and practitioners today are often willing to support true randomized experiments and are more likely to be influenced by experimental findings than in the past. Many higher-level managers have had research methods courses and most understand and are familiar with medical trials where new drugs are routinely tested with experimental designs (Petersilia 2008).

To some observers, close working relationships between practitioners and academics may violate the purported scientific necessity to separate program developer and evaluator roles (for discussion, see Petrosino and Soyden 2005). However, unless there is some convincing evidence of widespread evaluator bias associated with such arrangements, these collaborative arrangements seem necessary to put academics in the position of being able to conduct higher-quality evaluations of crime-prevention programs. As Olds (2009) suggests in his recent essay in support of "disciplined passion," balancing scientific integrity with the practical challenges associated with program evaluation in real-world settings needs to be addressed through higher standards for reporting trials, better peer review, improved investigator training, and rigorous collegial support of those who choose this line of work.

The field could also benefit from greater use of randomized experimental designs in place-based studies more generally. There are some promising signs that this is beginning to take hold (see Weisburd 2005). Where random assignment to treatment and control conditions is not feasible, the use of even more rigorous quasi-experimental designs should be explored (Henry 2009). Some of these designs include regression discontinuity, instrumental variable, propensity score matching, and interrupted time series with comparable control areas.

Notwithstanding the limitations of the present study, we believe that as a test of the influence of research design on study outcomes, it further demonstrates that research design does indeed matter, and this has everything to do with confidence in the observed effects. By and large, the highest-quality evaluation design used in our sample involved before-and-after measures of crime in experimental and comparable control conditions. Cook and Campbell (1979) and Shadish et al. (2002) considered this to be the minimum design that is interpretable. In short, control conditions are

needed to counter the many threats to internal validity. This is also the threshold that has been adopted by the Campbell Collaboration, not only for systematic reviews of criminological interventions, but also in the areas of education and social work and social welfare.

Our focus on a specific criminological intervention addresses Lipsey's (2003) call for more uniform bodies of research to be used to investigate the relationship between research design and study outcomes. This seems to be the most productive way to advance knowledge on this front. Also, evaluation designs at the lower end of the evaluation hierarchy should be included in future tests, if only because these weak designs continue to be the most widely used forms of evaluation in our field.

This study has gone some way toward contributing to the debate on the influence of research design on study outcomes in crime and justice. Its specific focus on area-based studies and those at the lower end of the evaluation design hierarchy has, we believe, dealt with two important but neglected areas of criminological research, but by no means has this debate been resolved. To this end, we join Weisburd et al. (2001), Lipsey (2003), and other scholars in calling for further research to advance knowledge on this important methodological and policy question.

Acknowledgments We are grateful to Chet Britt, Dan Mears, Chris Sullivan, the journal editor, and the anonymous reviewers for helpful comments.

References

Braga, A. A. (2005). Hot spots policing and crime prevention: a systematic review of randomized controlled trials. *Journal of Experimental Criminology, 1*, 317–342.

Braga, A. A. (2010). Setting a higher standard for the evaluation of problem-oriented policing initiatives. *Criminology and Public Policy, 9*, 173–182.

Braga, A. A., & Hinkle, M. (2010). The participation of academics in the criminal justice working group process. In J. M. Klofas, N. K. Hipple, & E. F. McGarrell (Eds.), *The new criminal justice: American communities and the changing world of crime control* (pp. 114–120). New York: Routledge.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*, 724–750.

Cornish, D. B., & Clarke, R. V. (2003). Opportunities, precipitators and criminal decisions: A reply to Wortley's critique of situational crime prevention. In M. J. Smith & D. B. Cornish (Eds.), *Theory for practice in situational crime prevention. Crime prevention studies, vol. 16* (pp. 41–96). Monsey: Criminal Justice Press.

Eck, J. E. (1995). A general model of the geography of illicit retail marketplaces. In J. E. Eck & D. Weisburd (Eds.), *Crime and place. Crime prevention studies, vol. 4* (pp. 67–94). Monsey: Criminal Justice Press.

Eck, J. E. (2006). Preventing crime at places. In L. W. Sherman, D. P. Farrington, B. C. Welsh, & D. L. MacKenzie (Eds.), *Evidence-based crime prevention, rev. ed* (pp. 241–294). New York: Routledge.

Ekblom, P., & Pease, K. (1995). Evaluating crime prevention. In M. Tonry & D. P. Farrington (Eds.), *Building a safer society: Strategic approaches to crime prevention. Crime and justice: A review of research, vol. 19* (pp. 585–662). Chicago: University of Chicago Press.

Farrington, D. P. (2003). Methodological quality standards for evaluation research. *The Annals of the American Academy of Political and Social Science, 587*, 49–68.

Farrington, D. P. & Welsh, B. C. (Eds.) (2001). What works in preventing crime? Systematic reviews of experimental and quasi-experimental research. *Annals of the American Academy of Political and Social Science*, 578 [full issue]

Farrington, D. P., & Welsh, B. C. (2005). Randomized experiments in criminology: what have we learned in the last two decades? *Journal of Experimental Criminology, 1*, 9–38.

Farrington, D. P., & Welsh, B. C. (2006). A half-century of randomized experiments on crime and justice. In M. Tonry (Ed.), *Crime and justice: A review of research, vol. 34* (pp. 55–132). Chicago: University of Chicago Press.

Farrington, D. P., & Welsh, B. C. (2007). *Improved street lighting and crime prevention: A systematic review*. Stockholm: National Council for Crime Prevention.

Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2006). The Maryland scientific methods scale. In L. W. Sherman, D. P. Farrington, B. C. Welsh, & D. L. MacKenzie (Eds.), *Evidence-based crime prevention, rev. ed* (pp. 13–21). New York: Routledge.

Henry, G. T. (2009). Estimating and extrapolating causal effects for crime prevention policy and program evaluation. In J. Knutsson & N. Tilley (Eds.), *Evaluating crime reduction. Crime prevention studies, vol. 24* (pp. 147–173). Monsey: Criminal Justice Press.

Klofas, J. M., Hipple, N. K., & McGarrell, E. F. (2010). The new criminal justice. In J. M. Klofas, N. K. Hipple, & E. F. McGarrell (Eds.), *The new criminal justice: American communities and the changing world of crime control* (pp. 3–17). New York: Routledge.

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *The Annals of the American Academy of Political and Social Science, 587*, 69–81.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks: Sage.

Mears, D. P. (2007). Towards rational and evidence-based crime policy. *Journal of Criminal Justice, 35*, 667–682.

Mears, D. P. (2010). *American criminal justice policy: An evaluation approach to increasing accountability and effectiveness*. New York: Cambridge University Press.

National Research Council. (2008). *Parole, desistence from crime, and community integration*. Washington, DC: National Academies Press.

Newman, O. (1972). *Defensible space: Crime prevention through urban design*. New York: Macmillan.

Olds, D. L. (2009). In support of disciplined passion. *Journal of Experimental Criminology, 5*, 201–214.

Painter, K. A., & Farrington, D. P. (1997). The crime reducing effect of improved street lighting: The Dudley project. In R. V. Clarke (Ed.), *Situational crime prevention: Successful case studies* (2nd ed., pp. 209–226). Guilderland: Harrow and Heston.

Painter, K. A., & Farrington, D. P. (1999). Street lighting and crime: Diffusion of benefits in the Stoke-on-Trent project. In K. A. Painter & N. Tilley (Eds.), *Surveillance of public space: CCTV, street lighting and crime prevention. Crime prevention studies, vol. 10* (pp. 77–122). Monsey: Criminal Justice Press.

Petersilia, J. (2008). Influencing public policy: An embedded criminologist reflects on California prison reform. *Journal of Experimental Criminology, 4*, 335–356.

Petrosino, A., & Soyden, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology, 1*, 435–450.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sherman, L. W. (Ed.) (2003). Misleading evidence and evidence-led policy: Making social science more experimental. *Annals of the American Academy of Political and Social Science,* 589 [full issue]

Sherman, L. W., Gottfredson, D. C., MacKenzie, D. L., Eck, J. E., Reuter, P., & Bushway, S. D. (1997). *Preventing crime: What works, what doesn't, what's promising*. Washington DC: National Institute of Justice, U.S. Department of Justice.

Sherman, L. W., Strang, H., Angel, C., Woods, D., Barnes, G. C., Bennett, S., et al. (2005). Effects of face-to-face restorative justice on victims of crime in four randomized, controlled trials. *Journal of Experimental Criminology, 1*, 367–395.

Weisburd, D. (1994). Evaluating community policing: Role tensions between practitioners and evaluators. In D. P. Rosenbaum (Ed.), *The challenge of community policing: Testing the promises* (pp. 274–277). Thousand Oaks: Sage.

Weisburd, D. (2005). Hot spots policing experiments and criminal justice research: Lessons from the field. *The Annals of the American Academy of Political and Social Science, 599*, 220–245.

Weisburd, D., Lum, C. M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? *The Annals of the American Academy of Political and Social Science, 578*, 50–70.

Weisburd, D., Lum, C. M., & Petrosino, A. (Eds.) (2003). Assessing Systematic Evidence in Crime and Justice: Methodological Concerns and Empirical Outcomes. *Annals of the American Academy of Political and Social Science*, 587 [full issue]

Welsh, B. C., & Farrington, D. P. (2009a). *Making public places safer: Surveillance and crime prevention*. New York: Oxford University Press.

Welsh, B. C., & Farrington, D. P. (2009b). Public area CCTV and crime prevention: An updated systematic review and meta-analysis. *Justice Quarterly, 26*, 716–745.

Welsh, B. C., Mudge, M. E., & Farrington, D. P. (2010). Reconceptualizing public area surveillance and crime prevention: Security guards, place managers, and defensible space. *Security Journal, 23,* 299–319.

Brandon C. Welsh, Ph.D. is an Associate Professor in the School of Criminology and Criminal Justice at Northeastern University and a Senior Research Fellow at the Netherlands Institute for the Study of Crime and Law Enforcement. He is an author or editor of eight books, including *Making Public Places Safer: Surveillance and Crime Prevention* (Oxford University Press, 2009) and *The Oxford Handbook on Crime Prevention* (Oxford University Press, forthcoming).

Meghan E. Peel, M.Sc. is a Doctoral Candidate in the School of Criminology and Criminal Justice at Northeastern University and a Research Associate at the Netherlands Institute for the Study of Crime and Law Enforcement. Her dissertation is titled "Defining Crime, Social Control, and the Enduring Influence of Neighborhood Context: A Mixed Methods Approach."

David P. Farrington, Ph.D. O.B.E., is Professor of Psychological Criminology at the Institute of Criminology at Cambridge University. His major research interest is in developmental criminology, and he is Director of the Cambridge Study in Delinquent Development, which is a prospective longitudinal survey of over 400 London males from ages 8 to 48. In addition to over 500 published journal articles and book chapters on criminological and psychological topics, he has published over 75 books, monographs, and government publications.

Henk Elffers, Ph.D. is Professor of Empirical Research into Criminal Law Enforcement at Vrije Universiteit Amsterdam and a Senior Research Fellow at the Netherlands Institute for the Study of Crime and Law Enforcement. His research interests encompass environmental criminology and guardianship, effects of punishment, and criminal evidence.

Anthony A. Braga, Ph.D. is a Professor in the School of Criminal Justice at Rutgers University and a Senior Research Fellow in the Program in Criminal Justice Policy and Management at Harvard University.