Using Quasi-Experiments to Evaluate Firearm Laws: Comment on Britt et al.'s Reassessment of the D.C. Gun Law

David McDowall Colin Loftin
Brian Wiersema

The strength of research designs is relative. Compared with true experiments, quasi-experimental designs are weak. Compared with cross-sectional designs, they are strong. One can further strengthen inferences from quasi-experiments by examining a broader pattern of data. We agree with Britt, Kleck, and Bordua when they recommend that researchers expand the range of inquiry. We disagree with them when they recommend that researchers restrict it. The District of Columbia study is largely consistent with the available evidence, but it does not prove that restrictive handgun licensing will always reduce firearm deaths.

Britt, Kleck, and Bordua (1996) raise four issues about the interrupted time series design and its use to study firearm licensing in the District of Columbia. These include the selection of comparison groups, the timing of interventions, the form of intervention models, and the stability of estimates. We broadly agree with some of their observations, but we see little merit in others.

We separately examine each of Britt et al.'s major points. First, however, we consider time series studies within the larger context of research design.

The Strength of Research Designs as a Relative Matter

All research designs are subject to error, but some are less fallible than others. Most evaluations of designs use true experiments as the comparison standard (see, for example, Cook & Campbell 1979; Berk 1988:163). True experiments randomly assign cases to an experimental and a control group and then apply a study factor only to the experimental cases. By using probability to equate the two groups at the start of the research,

Address correspondence to Colin Loftin, School of Criminal Justice, University at Albany, State University of New York, Albany, NY 12222.

Law & Society Review, Volume 30, Number 2 (1996)

experiments eliminate many noncausal explanations for a difference at the end.

Quasi-experiments are much less desirable than are true experiments. Quasi-experiments also use an experimental and a control group, but they lack random assignment. The groups are more likely to differ at the start of the research, leaving more explanations for a difference between them at the end.

Still, some quasi-experiments rule out more rival explanations than do others, and a time series study is among the strongest in this respect. All major quasi-experiments also are stronger than are cross-sectional designs, which use statistical methods to equate the groups.[1]

One must judge any research design by comparing it with its alternatives. In this context, and as a general matter, one would prefer interrupted time series studies to most other approaches. We urge readers who doubt these points to consult standard texts on research design, especially Campbell & Stanley (1966) or Cook & Campbell (1979).

Comparison Groups in Interrupted Time Series Studies

Britt et al.'s first point correctly stresses the importance of comparison groups in time series analysis. By concentrating on external control series, however, they miss the essential logic of the time series design. The primary comparison in a time series study is between the average levels of a series before and after an intervention. The pre-intervention values take the role of the control group, and the post-intervention values are the experimental group. *All* the studies in Britt et al.'s Table 1 thus use control groups.

In the District of Columbia study, homicides with firearms fell by more than a chance amount after restrictive handgun licensing began (Loftin, McDowall, Wiersema, & Cottey 1991). Suicides with firearms also decreased, a central finding that Britt et al. do not mention. Together, these results suggest that the law reduced fatal firearm violence in the city (see Figures 1 and 2).

Time series designs are quasi-experiments, however, and it is conceivable that the pre-intervention patterns do not reflect what would have happened without the law. Inferences would be stronger if one examined a larger pattern of evidence. The District of Columbia study thus also analyzed homicides and suicides without guns in the District, and homicides and suicides with and without guns in the adjacent suburbs. None of these six control series decreased by more than chance after the law began.

[1] For a clear discussion of some of the problems facing cross-sectional studies of firearm laws, see Alba & Messner 1995.

https://doi.org/10.2307/3053964 Published online by Cambridge University Press

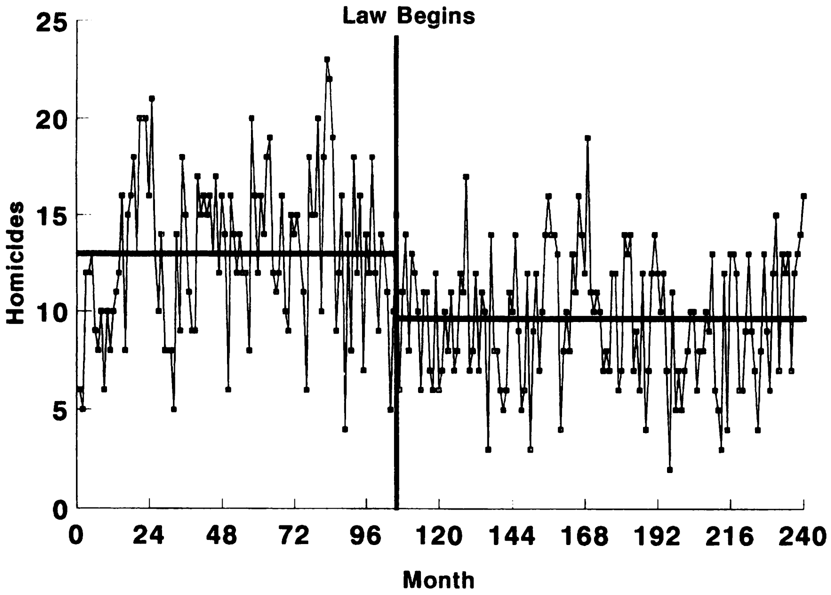

Figure 1. Firearm homicides by month, District of Columbia, 1968–1987

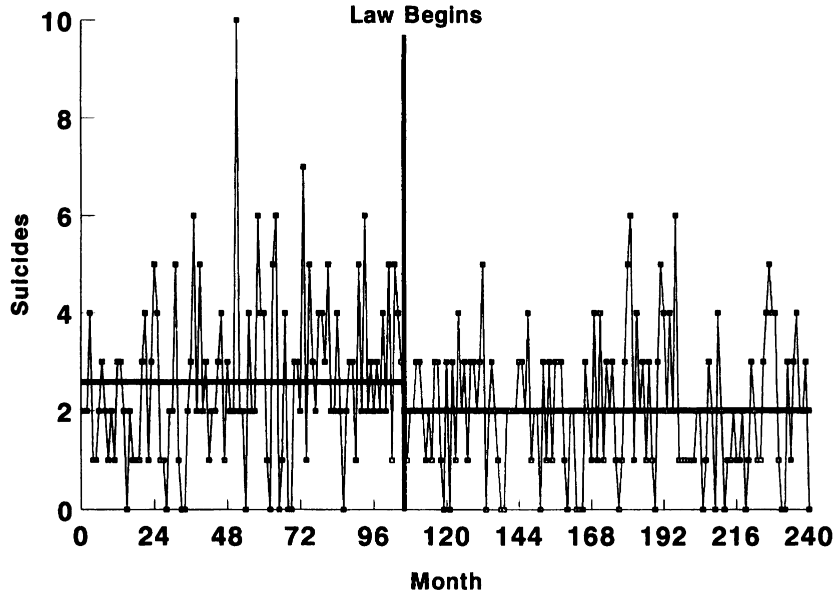

Figure 2. Firearm suicides by month, District of Columbia, 1968–1987

Additional supplementary tests are possible. Most usefully, one might replicate a study in other areas. Campbell & Stanley (1966) note that the natural sciences use interrupted time series designs extensively, supporting conclusions with multiple replications. We have used replication methods elsewhere (McDowall, Loftin, & Wiersema 1992; McDowall, Loftin, & Wiersema 1995),

https://doi.org/10.2307/3053964 Published online by Cambridge University Press

and we believe that they are extremely useful. If the same outcome occurs in different areas after they adopt a policy, rival explanations become much less plausible. Unfortunately, there is no suitable duplicate of the District's licensing law.

One also might examine violence in similar areas that did not change their laws. Britt et al. apply this approach to homicides in Baltimore. Baltimore is a reasonable choice for a control site, and the fact that firearm homicides fell there is relevant evidence.

After one considers the complete set of results, however, Baltimore's experience is not highly compelling. Firearm homicides decreased in Baltimore, but so did homicides and suicides without guns (Table 1, panel A).[2] Equally important, firearm suicides remained stable. Unlike the District, the Baltimore estimates do not show a pattern of decrease that is specific to gun-related deaths.

As a further test of whether the District findings reflected changes that occurred widely elsewhere, we analyzed data from two other areas. These were Boston and Memphis, the cities ranking immediately above and below the District in their 1990 populations. No series in either city decreased by more than a chance amount after the District's law began (Table 1, panel B). We therefore see no reason that the Baltimore results should alter conclusions about the District.

At a more basic level, we strongly disagree with Britt et al.'s implication that only a *single* comparison series is worth analyzing. By definition, quasi-experiments lack strictly comparable control groups. Baltimore is not the District of Columbia, just as gun homicides are not homicides without guns. Using common sense to choose several comparisons will be more convincing than will be conclusions that depend on the appropriateness of a single external control.

External comparisons are valuable supplements to an interrupted time series analysis. Yet no quasi-experimental control group will be fully satisfactory. Britt et al.'s insistence on one *perfect* control invites endless unproductive debates.[3] Ultimately, it is a counsel of despair.

[2] Due to space limitations, we present only the essential results of analyses. A detailed appendix is available on request. All data are from the U.S. National Center for Health Statistics (1993). Britt et al. are mistaken in saying that our nongun homicide series included legal interventions.

[3] Kleck (1991:254), for example, faults studies for ignoring nongun violence and for using cross-sectional matching to select comparison areas (pp. 383–87). Researchers thus cannot satisfy the standards both of Britt et al. and of Kleck.

https://doi.org/10.2307/3053964 Published online by Cambridge University Press

Table 1. Supplementary Analyses of District of Columbia Restrictive Handgun Licensing Law

Panel A. Change in Mean Numbers of Homicides and Suicides per Month after District of Columbia Law, Baltimore, 1968–1987

	Change after Law	*t*-Statistic
Gun-related homicides	−3.01	−3.29
Other homicides	−1.41	−3.47
Gun-related suicides	0.17	0.53
Other suicides	−0.62	−2.00

Panel B. Change in Mean Numbers of Homicides and Suicides per Month after District of Columbia Law, Memphis and Boston, 1968–1987

	Change after Law	*t*-Statistic
Memphis		
Gun-related homicides	0.74	1.06
Other homicides	0.37	1.60
Gun-related suicides	0.65	2.01
Other suicides	0.30	1.55
Boston		
Gun-related homicides	−0.80	−1.12
Other homicides	−0.31	−0.55
Gun-related suicides	0.10	0.69
Other suicides	−0.26	−0.76

Panel C. Change in Mean Numbers of Homicides and Suicides per Month after District of Columbia Law, District of Columbia, 1968–1990

	Change after Law	*t*-Statistic
Gun-related homicides	2.08	0.66
Other homicides	0.61	1.38
Gun-related suicides	−0.47	−2.23
Other suicides	−0.33	−0.94

Selecting Intervention Times in Interrupted Time Series Studies

The second issue that Britt et al. consider is the choice of an intervention point. They note that researchers will rarely know when an intervention began to influence behavior. As a remedy, they advise analysts to estimate effects using several different dates.

Ignoring the matter of whether everyone in a population responds to a legal change at a single date, we believe that this suggestion is problematic. If one conducts multiple tests on the same set of data, the notion of statistical significance rapidly loses its meaning. The conventional .05 alpha level implies a Type I error rate of 1 in 20. A researcher who tries 20 intervention

https://doi.org/10.2307/3053964 Published online by Cambridge University Press

points should expect at least one "significant" finding by chance.[4]

The most general version of Britt et al.'s suggestion would be to estimate all possible interventions and select the largest estimate as the intervention date. This procedure has low power, and it is heavily subject to chance events. Nevertheless, we applied it to the District of Columbia data. The largest estimate for firearm homicides was four months from the effective date of the law, and the largest estimate for firearm suicides was one month from it.

In any event, if the level of a series *did* change after a single intervention, a mistaken choice of the date should yield a conservative analysis. To see this, suppose that one studied the following set of data:

15, 15, 15, 15, 15, 15 (Intervention) 5, 5, 5, 5, 5, 5.

This series decreases by 10 units from a pre-intervention mean of 15.

Suppose one incorrectly placed the intervention after observations 4 or 8, two periods away from the change in the series. Here one would estimate a decrease of only 7.5 units. If one placed the intervention after observations 2 or 10, four periods away from the change, the decrease would be only 6 units. The further one moves from the true intervention point, the smaller the estimate of the effect. An incorrect location of the intervention thus underestimates the magnitude of the change.[5]

This exercise also suggests that Britt et al.'s method can find nontrivial effects far from the intervention date. Although the estimates will be smaller away from the correct date, they still may be statistically significant. Users of their method can therefore wrongly decide that a series changed long before or long after the intervention began.

Much of Britt et al.'s article reflects a concern about mistakenly concluding that a new policy affected a series. We share this concern, and we do not believe that a conservative bias is undesirable. The District of Columbia study used the effective date of the licensing law as the intervention point. We think that this is a reasonable choice, and it is easy to define. Other researchers might plausibly select other points, but we are skeptical of post hoc attempts to "fish" for an intervention.

[4] For additional criticism of this practice, see Kleck 1991:387–88.

[5] Matters are more complicated if one allows for random variation, but the same principle holds. Compare the District of Columbia estimates in Britt et al.'s Table 2 with those in their Table 4.

https://doi.org/10.2307/3053964 Published online by Cambridge University Press

Intervention Models in Interrupted Time Series Studies

Britt et al. discuss the form of intervention models in their third point, and they argue here that researchers should usually consider only gradual and permanent impacts. Much of this material is technically unsound, and we believe that it is the weakest part of their article.

Abrupt and permanent impact models are a special case of gradual and permanent ones. The abrupt permanent model has a single coefficient, ω, which measures the change in the series mean after the intervention. Besides ω, the gradual permanent model has a second coefficient, δ, which measures how rapidly the series reaches its final level. If δ equals zero, the series reaches this level immediately. The gradual model then becomes identical to the abrupt one.

Unlike the case of choosing an intervention point, statistical theory provides a solid foundation for selecting the model that best fits the data. The abrupt permanent model is more parsimonious than is the gradual permanent one. Unless a gradual impact fits better, one would select the simpler abrupt impact. As Britt et al. note, abrupt models provided the best fit to the data in the District of Columbia study.

A gradual permanent model implies a theory of how the intervention influenced a series. This theory is incorrect in the District of Columbia, where firearm homicides and suicides both abruptly decreased. Britt et al.'s advice to ignore this finding is a plea to blind oneself to obvious patterns of change. If one carried this practice to its logical extreme, one need not bother with empirical tests at all.

Perhaps more important, gradual and permanent intervention models make heavy demands on the data. Because of high correlations between the ω and δ coefficients, computer programs often cannot accurately estimate them, even if the gradual model is appropriate. One will then obtain insignificant nonsense results like those that Britt et al. report. Here the single coefficient abrupt model will still reasonably approximate the intervention's effect.

To show this, we used the SCA computer program (Lu & Hudak 1986) to simulate a gradual permanent intervention. We created a series with a pre-intervention level of 50, and a gradual decrease to a post-intervention level of 35 ($\omega = -1.5$, $\delta = .90$). To this we added a random (white noise) error term (see Fig. 3).

We then used SCA to estimate a gradual permanent model on the data that it had generated. The program returned grossly incorrect and statistically insignificant estimates of $\omega = -2.8$ and $\delta = .72$. Even with a large and known intervention, we could not obtain accurate results.

https://doi.org/10.2307/3053964 Published online by Cambridge University Press

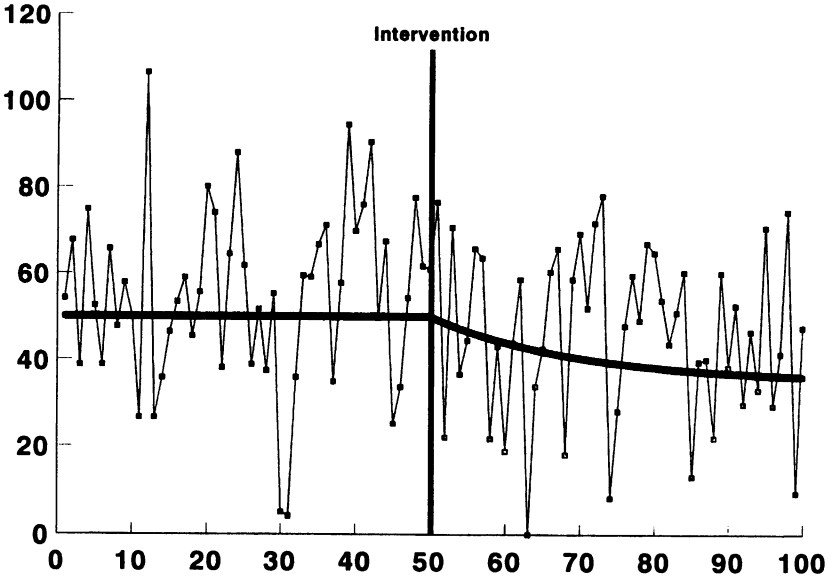

Figure 3. Simulated data with gradual permanent intervention after observation 50

Again, outcomes like this are due to practical problems in separating the two model coefficients. For the simulated data, the program reported a correlation of .98 between its estimates of ω and δ. As far as the program is concerned, the two coefficients amount to the same quantity, and it cannot estimate either after controlling for the other.

The same situation applies to the District of Columbia data. For both firearm homicides and firearm suicides, the correlation between ω and δ is an almost perfect .99. We suspect that this will occur whenever a change is small relative to the total series variation.

In cases like this, the single coefficient in the abrupt model can answer basic questions about whether a series increased or decreased after an intervention. For the simulated data, the abrupt model estimate was a statistically significant 9.1-unit decrease.

Researchers should routinely examine *both* gradual and abrupt models, using statistical theory to make their final choice. Because of the numerical problems in estimating gradual models, one should not entertain them alone. Britt et al.'s proposal to consider only these models often will lead to serious errors, and it can force one to ignore large and visually striking impacts. We strongly caution readers to avoid it.

https://doi.org/10.2307/3053964 Published online by Cambridge University Press

Model Stability in Interrupted Time Series Studies

Britt et al.'s final point considers the stability of model estimates. They recommend that one test model stability by varying the length of the series under study.

We agree with the spirit of this advice, but not for the reasons that Britt et al. offer. Time series data are not inherently unstable. A series is a set of realizations (a sample) from an underlying stochastic process. If the process operates in the future as it has in the past, the series level will be constant after one controls for the intervention.

In theory, there is no more reason to drop points from a time series than there is to drop respondents from a sample survey. Because one will obtain more precise estimates from larger samples, one would ordinarily use the entire series for the major analysis. The District of Columbia study examined homicides and suicides through 1987, the latest data then available.

Yet time series studies are vulnerable to historical threats, and other interventions might alter the course of a process. Ideally, researchers will examine the historical record and allow for the effects of other major changes on the analysis. Unknown or not clearly defined interventions always lurk in the background, however, and post hoc tests may help reveal them.

By the time the District of Columbia study appeared in print, firearm homicides were reaching record heights. The study speculated that this was due to another intervention, the violence accompanying the beginning of crack cocaine trafficking in the city.

If the drug hypothesis is correct, only the results for gun homicides should change if one considers the period after 1987. To test the hypothesis, we analyzed monthly data from the District through 1990 (Table 1, panel C). In support of the drug explanation, the estimates for all series except firearm homicides were largely identical to those in the original study.

Of course, post hoc tests for historical threats are prone to the same problems that beset searches for intervention dates. Ultimately, both methods make it easier to conclude that the study policy influenced a series. One can easily allow for other *known* interventions at the beginning of an analysis, but findings that depend on post hoc explorations will always rest on shaky ground.

Unmeasured historical events complicate time series studies, and they become more likely as the length of the series increases. We suggest two methods to supplement Britt et al.'s strategy of deleting observations.

First, and most convincingly, one might replicate the analysis in other areas. Independent tests in additional settings will pro-

https://doi.org/10.2307/3053964 Published online by Cambridge University Press

vide a more detailed set of findings, allowing one to better assess the threats posed by local history.

Second, one should use common sense in examining the data and interpreting the results. Firearm homicides in the District were visibly lower for more than 10 years after the licensing law began. The later rise in killings shows that this drop was not unalterable. Still, the duration of the change is stronger evidence of the law's effect than would be a decrease that lasted for only a few months.

Conclusions

One must evaluate time series and other quasi-experiments against some standard. Compared to true experiments, quasi-experiments are weak. Compared to cross-sectional designs, they are relatively strong. This deserves emphasis, because some firearm researchers mistakenly assert that cross-sectional methods are the most desirable (see Kleck 1995).

Although inferences from quasi-experiments are difficult, one can strengthen them by viewing a larger pattern of evidence. One might test multiple independent hypotheses, analyze external control series, or (most important) replicate studies in other areas. These additions cannot prove that an intervention influenced a series, but they can greatly narrow the set of possibilities.

We agree with Britt et al. when they advise researchers to apply valid methods to a wide range of data and to consider rival explanations for the results. We disagree when they insist on arbitrary tests and on a narrow range of models and control series.

Most emphatically, we disagree that one should suppress research that does not meet Britt et al.'s standards. They imply that one, perfect, study can resolve the issues; a study that does not fully satisfy their questionable standards for perfection (and this includes all existing studies) contains no useful information.

In contrast, we believe that trustworthy knowledge is most likely to come from multiple and redundant evaluations using a wide range of designs and sites. Worthwhile findings will appear repeatedly, while incorrect ones will fall by the wayside. We believe that researchers should examine *more* evidence, not *less*.

While the District of Columbia study holds up well under scrutiny, we do not claim—and have never implied—that restrictive licensing must always reduce firearm violence. The study reports the findings of a single quasi-experiment. It is a beginning, and that is all.

https://doi.org/10.2307/3053964 Published online by Cambridge University Press

References

Alba, Richard D., & Steven F. Messner (1995) "*Point Blank* against Itself: Evidence and Inference about Guns, Crime, and Gun Control," 11 *J. of Quantitative Criminology* 391.

Berk, Richard A. (1988) "Causal Inference for Sociological Data," in N. J. Smelser, ed., *Handbook of Sociology*. Newbury Park, CA: Sage Publications.

Britt, Chester L., Gary Kleck, & David J. Bordua (1996) "A Reassessment of the D.C. Gun Law: Some Cautionary Notes on the Use of Interrupted Time Series Designs for Policy Impact Assessment," 30 *Law & Society Rev.* 361.

Campbell, Donald T., & Julian C. Stanley (1966) *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Cook, Thomas D., & Donald T. Campbell (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

Kleck, Gary (1991) *Point Blank: Guns and Violence in America*. New York: Aldine de Gruyter.

——— (1995) "Using Speculation to Meet Evidence: Reply to Alba and Messner," 11 *J. of Quantitative Criminology* 411.

Loftin, Colin, David McDowall, Brian Wiersema, & Talbert J. Cottey (1991) "Effects of Restrictive Licensing of Handguns on Homicide and Suicide in the District of Columbia," 325 *New England J. of Medicine* 1615.

Lu, Lon-Mu, & Gregory B. Hudak (1986) *The SCA Statistical System, Version III for MSDOS*. DeKalb, IL: Scientific Computing Associates [computer software].

McDowall, David, Colin Loftin, & Brian Wiersema (1992) "A Comparative Study of Mandatory Sentencing Laws for Gun Crimes," 83 *J. of Criminal Law & Criminology* 378.

——— (1995) "Easing Concealed Firearm Laws: Effects on Homicide in Three States," 86 *J. of Criminal Law & Criminology* 193.

U.S. National Center for Health Statistics (1993) *Mortality Detail Files, 1968 to 1990*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [computer files].

https://doi.org/10.2307/3053964 Published online by Cambridge University Press