Accuracy of Common Statistical Methods for Detecting Effects of State Policies on Mass Shootings

Rosanna Smart[a]

Terry L. Schell[a]

Grant Duwe[b]

Denis Agniel[a]

[a]RAND Corporation, 1776 Main Street, Santa Monica, CA 90407

[b]Minnesota Department of Corrections, 1450 Energy Park Drive, Suite 200, St. Paul, MN 55108

Abstract

The enormous costs of mass shootings prompt public interest in identifying policies that effectively reduce their frequency and lethality. While numerous studies have used common quasi-experimental methods to identify such policies, most find statistically nonsignificant effects. It is unclear if those methods have adequate power to evaluate this question in the available data. Without sufficient power, both nonsignificant and significant findings are likely to be inaccurate. This study evaluates the statistical performance of common methods used to assess state policy effects on mass shootings. Using empirical simulations based on real mass shooting data and fictitious state policies, we compare the ability of standard models with different link functions, geographic controls, and time-varying covariates to accurately recover causal effects and accurately estimate standard errors across outcomes, effect magnitudes, and numbers of adopting states. The results highlight serious challenges. Unless effect sizes are very large (20-percent decrease) and many states adopt the policy, the best-performing models correctly identify significant effects less than 20% of the time. For mass public shootings (a rarer outcome than mass shootings generally) and few treated states, methods on average detect true effects less than 5% of the time, and one of three significant estimates is in the wrong direction. Findings highlight severe statistical limitations facing policy evaluations targeting rare but consequential forms of firearm violence.

Acknowledgments: The authors thank Dulani Woods for work on an earlier version of the simulation design, as well as participants of the National Research Conference for the Prevention of Firearm-Related Harms for helpful feedback. This work was funded by a grant from Arnold Ventures. The authors have no potential conflicts of interest to disclose.

Introduction

Policymakers often face questions about effective strategies to prevent rare but costly outcomes. For example, terrorist attacks, fatal child maltreatment, and mass shootings occur rarely in any given state or year, yet each carries social costs substantial enough to command legislative and public attention (Cook 2020; Clauset and Woodard 2013; Fang et al. 2012). The rarity of these outcomes, however, also makes it challenging to evaluate the policies that have been implemented to address them. If a state often experiences zero incidents in a year, observing zero incidents in the year following a law change offers little evidence for whether the new law is responsible. The literature on state firearm laws and mass shooting outcomes highlights this challenge. Despite a relatively large body of research using quasi-experimental methods to estimate the effects of state laws on mass shooting incidence and lethality, most studies have found statistically nonsignificant effects (Smart et al. 2026). Even in those studies that have found statistically significant effects, effect sizes are often estimated so imprecisely that they can neither rule out that the true effect is near-zero nor can they rule out implausibly large effects (Smart et al. 2026).

Despite the importance of understanding how policies affect rare but costly outcomes, the policy evaluation literature has paid little attention to these statistical problems. For decades, most methodological research in the policy evaluation field has focused on reducing bias in causal effect estimates without discussing the variance or the total error of the estimates (Olsen et al. 2013; Deaton and Cartwright 2018). A growing set of studies have noted statistical problems created by high variance causal effect estimates, and the low statistical power this causes in particular policy evaluation contexts, such as estimating the effects of state policies on firearm deaths (Schell et al. 2018) or fatal opioid overdoses (Griswold et al. 2026; Griffin et al. 2021). The field of applied economics in particular appears to have a large fraction of studies with inadequate statistical power (Ioannidis et al. 2017). A literature dominated by low statistical power creates significant inferential problems, beyond a low rate of statistical significance; even when effects from such studies are statistically significant the effect sizes may be systematically biased in magnitude (Type-M error), and an unacceptably high proportion of significant effects may be in the wrong direction (Type-S error; Gelman and Carlin (2014)) when power is low.

Some recent research suggests that when estimating state policy effects on an all-cause mortality outcome, several common methods have power to detect changes in mortality rates of 2 percent, and Type-S error is negligible for effects half that size (Black et al. 2022). However, we suspect that these encouraging findings do not generalize to evaluating state policies on rarer outcomes, such as mass shooting incidents, where you may have zero or one incident in a given state-year, rather than thousands. A rigorous assessment of the performance of commonly used quasi-experimental methods in the evaluation of such rare outcomes in panel data does not appear to be available. This study fills that gap using mass shooting outcomes as our case study. We conducted a series of statistical simulations based on data representing a census of US mass shooting incidents over a 45-year period to assess statistical performance of the types of models that have been used to estimate the effects of firearm laws on mass shooting outcomes (Smart et al., 2026). We compare model performance in terms of: accuracy of estimates, accuracy of

model-based standard errors, power to detect true effects, and Type-S error rate (i.e., the proportion of statistically significant effects that were in the wrong direction relative to the true effect). An additional contribution of this paper is establishing a uniform and valid approach to simulating causal effects in sparse data. It is not possible to create homogenous causal effects on a discrete count outcome that includes zeros (without creating negative or fractional incidents), and this paper extends prior simulation-based approaches (Black et al. 2022; Schell et al. 2018) to creating simulated effects that are nearly homogenous with these sparse discrete outcomes.

We document three novel findings. First, even under the best conditions and using the best among the tested models, we have reasonable power to detect only extremely large effect sizes (e.g., a 20 percent change in mortality rates) with such a sparse outcome. Second, in this setting, quasi-experimental methods that are standard in the existing literature produce Type I error rates of four to six times the nominal rate, indicating that standard errors are substantially underestimated even when using cluster-robust methods. Third, in contrast to the findings of Black et al. (2022) where Type-S error rates were negligible for effect sizes above 1%, we find that Type-S rates can remain substantial even at imposed effect sizes of 20%, an effect size outside the range considered plausible for any state gun law by expert assessments (Smart et al. 2021). It appears that estimates of policy effects using standard statistical methods are too inaccurate to provide useful guidance about which policies can effectively reduce mass shooting incidents. For these types of extremely rare but important outcomes, empirical estimates of policy effects may be inherently too imprecise to be useful for policy guidance, at least absent novel statistical methods that have not yet appeared in this literature.

The remainder of the paper proceeds as follows. We first describe our data, simulation design, and the estimation methods we evaluate. We then present results on estimation accuracy, Type I error, correct rejection rates, and sign errors. Finally, we conclude by discussing implications for the evaluation of policies targeting rare outcomes and for the reporting standards such evaluations should meet.

Methods

Our simulations compare the statistical performance of a range of quantitative methods for estimating state-level policy effects using outcome data that is based on actual mass shooting incidents across states and over time. Specifically, we are assessing performance in an ideal case where models' assumption about the functional form of the causal effects are exactly true, there are no actual omitted confounds that would necessitate more complex models, and the effects are as homogeneous as possible across states and over time. Given differences in how researchers define and measure mass shootings, we replicate findings across three separate outcomes. This allows us to assess whether model performance varies across outcomes with different distributional characteristics (e.g., fatalities versus incidence) or levels of sparseness (e.g., mass public shootings vs. any mass shooting).

The set of quantitative methods evaluated in this study were selected to represent those that have been used in the research identified in a recent systematic review (Smart et al. 2026). See the supplemental materials for a detailed discussion of those studies. For our primary

analyses, we estimate policy effects using Poisson regression models that include a population offset so that effects are expressed on per capita rates. There is no evidence of overdispersion in these data and so we do not present negative binomial or quasi-Poisson methods. We present results from analogous linear models of mass shooting counts or population-weighted rates, findings of which conceptually replicate the main findings, in the Supplementary Materials. We fit all models using the fixest package (Berge 2023) in R 4.4.2 (R Core Team 2025). Model-based standard errors are computed using the "iid", "cluster", and "hetero" specifications in fixest (Berge 2025).

Data

Our three primary outcome measures are annual state-level counts of the number of mass shootings using four different definitions: (1) *mass pubic shooting incidents* where four or more people were killed with a firearm in a public location and the incident was unconnected to domestic violence or other criminal activities, (2) *mass shooting incidents* where four or more people were shot and killed regardless of location or other circumstances, and (3) *mass shooting fatalities* resulting from mass shooting incidents with four or more people killed with a firearm. These outcome measures, which were derived from data provided by Grant Duwe as an update to data used in Siegel et al. (2020), are measured annually from 1976 to 2020. In addition to these outcome measures, analyses include up to 19 state-level time-varying covariates that are commonly used in this literature (see Appendix S1).

Table 1 describes the distributional characteristics of our outcomes. Mass public shooting incidents are an extremely sparse outcome, with more than 90% of state-years experiencing zero incidents, and eight states experiencing zero incidents over the entire 45-year period. While all states have at least one mass shooting over the 45-year period, nearly three-quarters of state-year observations have zero mass shooting incidents or fatalities.

In supplementary analysis (Appendix S2), we present simulation results for an alternative mass shooting outcome measure. This measure, Gun Violence Archive (GVA) mass shooting incidents, is based on less stringent criteria—counting incidents where four or more people were killed *or injured* with a firearm—and hence less sparse. However, the GVA dataset only covers incidents starting in 2014 and thus has a substantially shorter time period over which to observe identifying policy variation.

Simulation design

In each simulation sample, a fictitious policy is generated, and a random sample of states is assigned to that policy, each one implementing the policy at a random year in the period and continuing the policy until the end of the period. We know that there is no causal effect of these policies on the real, historical mass shooting data because we assign entirely fictitious policies. A good causal effect estimation method should return effect estimates that are tightly clustered around zero across a set of these simulated policies. We can also modify the historical mass shooting data slightly to add a causal effect of a known size for each simulated policy. In this case, a good effect estimation method should return estimates tightly clustered around the true causal effect that was used to modify the data.

This simulation approach preserves the complex structure of the real mass shooting data (the univariate and multivariate distributions of the outcomes and the covariates) and allows us to assess the statistical properties of various causal effect estimators applied to those data. Specifically, we assess the quality of the estimated causal effects for any simulation condition and method by evaluating the root mean squared error of the estimate (RMSE). This metric represents the "standard error" of these estimates relative to a known true effect for any set of simulation conditions and method. We also assess the quality of the statistical inference using a range of common methods to obtain standard errors by assessing the Type I error rates when the null hypothesis is true. Finally, we assess correct rejection rates (i.e., the proportion of times the estimated effect is significant and in the correct direction) and Type S error rates (i.e., the proportion of times a nonzero effect is estimated to be statistically significant but has the wrong sign) for tests with a true $p = 0.05$ of a Type I error.

There are 21 conditions of the simulation, each with 10,000 samples of simulated policies. These conditions are the factorial combination of the number of treated states (5, 15, 35) and seven sizes of true effect. This range for the number of treated states represents a plausible range found in the literature evaluating the effects of state firearm policies on mass shootings.(Smart et al. 2024) Policy implementation dates were randomly selected with the constraint that there be at least one year pre- and post-implementation.

The range of simulated effect sizes represents symmetrical increases and decreases on a log rate ratio (RR) scale, corresponding to RRs = 0.80, 0.90, 0.95, 1.00, 1/0.95, 1/0.90, and 1/0.80, where an RR=1 represents the null effect in which the outcome data are unmodified. These sizes were selected as possible effects for state policies on mass shootings based on a previous survey of experts.(Smart et al. 2021) Specifically, the range of effect sizes covered in the simulation (±20%) extends to sizes that are larger than those that are typically expected even by the experts who support these types of policies (see Appendix S1). When modifying the real outcome data to incorporate a causal effect at a target RR, we add or remove counts from the treated state-years so that the total counts across all treated state-years after modification are as close as possible to the target IRR times total counts prior to modification. The modifications are made in a manner that is as homogeneous as possible across all treated state-years, within the constraints of discrete data (see Appendix S1).

Causal effect estimation methods evaluated

The existing literature estimating effects of state policies on mass shootings has mostly used model-based approaches (Smart et al. 2026), most commonly two-way fixed effects models or some simplification of that model. The range of models we explore roughly corresponds to the range of statistical methods that have been used in the current policy evaluation literature (Smart et al. 2026). We evaluate twelve model types that vary in complexity with two levels of geographic controls (state fixed effects, region fixed effects), three levels of time effects (linear effect of time, cubic B-spline of time with three degrees of freedom, year fixed effects), and two levels of state-level time-varying covariate inclusion (two, nineteen). When using only two covariates, those are: the fraction of suicides that use a firearm, commonly used as a proxy for gun ownership rates, and population density. The longer list of nineteen covariates represents a

range of demographic, geographic, and economic state characteristics that are commonly included in firearm policy research (see Appendix S1).

Additionally, for each of these twelve different methods to get causal effect estimates, we explore three different methods to conduct population inference: model-based standard errors assuming IID error, sandwich standard errors adjusted for clustering at the state level, and sandwich standard errors adjusted for heteroscedasticity and nonnormality of the residuals. These different types of variance estimates are used when evaluating the Type I error rate within the subset of simulations where the null hypothesis is true (i.e., RR=1). When evaluating correct rejection and Type-S errors rates it is critical that the inferential method yield the nominal Type I error rate; we would not want to claim a model yields a high correct rejection rate that is entirely attributable to effectively using a $p<0.5$ criteria for significance rather than the intended $p<0.05$. To ensure the correct Type I error rates in both the positive and negative direction when assessing correct rejection rates and Type-S error, we conduct significance testing using critical values derived from the 0.025 and 0.975 quantiles of the simulated distribution of effects found under the null for the same set of fictitious policies.

We summarize results across simulation settings and modelling choices in a few ways. To get an overview of sources of variability across the 252 simulation and modelling conditions, we regress the RMSE estimates onto those conditions separately for each outcome. More detailed summaries show barplots of a given quantity (e.g., correct rejection rate) across all considered models within a given simulation setting, where the height of the bar denotes average performance across models and a line range indicates minimum and maximum performance.

To avoid any single simulation condition having undue influence on our results, we filter out simulation runs that result in estimated effects with an $RR>1000$ or $RR<0.001$. Table S1 describes the models and percentage of simulation runs that required this filtering for each condition. These results occur only in the conditions with five treated states and state fixed-effect controls; these are likely degenerate cases in which there are zero counts in the treated state-years, meaning the estimate is lower-bounded only by the settings for the maximum likelihood estimation algorithm. We assume such results would often be accompanied by a warning message in software packages or produce effect sizes implausible enough that they alert the analyst to a potential problem. Thus, our RMSE measure excludes the simulations with the most error and represents a best-case scenario in which we assume researchers know not to interpret such results.

Results

Estimation Accuracy

Table 2 shows that RMSE is relatively poor for all simulation conditions, models, and outcomes assessed. Even the best-performing, lowest-RMSE models in the best simulation conditions have average error on the order of 0.10, expressed on a log RR scale. This implies that, under the best data and modelling conditions, even when the null hypothesis is true, we expect about 5% of effect estimates to be outside an RR range of approximately 0.80 to 1.20. Under typical conditions, the magnitude of the errors is approximately twice that large. RMSE is

relatively similar for *mass shooting incidents* and *mass shooting fatalities*, although worse on average for the rarer event of *mass public shooting incidents*. Across simulation conditions and models, RMSEs are largely driven by variance rather than bias. The absence of bias is perhaps unsurprising given our simulations used a best-case scenario where policies were randomly assigned to states.

For each outcome, there is considerable variation between the best and the worst RMSE across the various simulation and model conditions we explored. This variation is largest for *mass public shootings*, with the worst condition having almost six times the error as the best; for other outcomes, the worst is close to four times more error than the best. Regressing RMSE estimates onto the 252 simulation and modeling conditions with only main effects (Table 3) explains more than 90% of the observed variation in estimation error for all outcomes. The most important factor leading to the differing levels of error is the number of treated states that were available for estimating the effect. Going from 35 implementing states to 15 increases average error by around 0.08 for *mass public shootings*, and around 0.04 for the other outcomes. However, going from 15 down to 5 implementing states leads to a larger increase in error, 0.34 for *mass public shootings* and between 0.17 and 0.19 for the other outcomes.

The second most important factor contributing to variation in estimation accuracy is how geography is included in the model, although the magnitude of these effects varies across outcomes. For all our primary outcomes, more complex controls for geography (i.e., state fixed-effects versus region fixed-effects) increase RMSE; this is most notable for mass public shooting incidents where error increased by 0.27 on average.

Other simulation or modelling conditions matter much less for estimation accuracy. RMSE does not vary much across effect sizes, different ways of including time in the model, or the number of time-varying covariates. However, there is a pattern where more complex models (more detailed time trends and more covariates) are somewhat harmful with the sparsest outcome, *mass public shootings* (also see Figures S1-S4). Conclusions are broadly similar to mass shooting incidents when estimating effects on our GVA measure that uses a less stringent definition of mass shootings and thus has fewer sparsity issues (Tables S2-S3). One key distinction is that the RMSE for the GVA effect estimates has a lesser penalty for the inclusion of state fixed-effects, likely due to the shorter study timeframe available in the GVA data (2014-2022 vs. 1976-2020 for the other outcomes).

Replication results using linear models of both population rates (weighted) and counts (unweighted) are presented in Tables S4-S6. Conclusions are generally similar for those alternative specifications, although RMSE is expressed on a different scale. Across the 21 simulation conditions and 12 model conditions, the correlations between the RMSEs of the Poisson estimates and those from a population-weighted linear model of rates are over 0.88 for all outcomes. In general, these linear models do not appear superior to the Poisson model in estimation accuracy.

Type I error rates

Accurate standard error estimation is key for achieving desired Type I error rates. This accuracy varies substantially across modeling and simulation conditions, particularly as a function of the number of treated states and the specific outcome analyzed. As has been documented previously (Bertrand et al. 2004; Schell et al. 2018), clustered standard errors with few treated units are biased. They are consistently underestimated and substantially over-reject the null across all conditions and outcomes, with average type I error rates in the five treated state conditions ranging between 0.19 and 0.30 (see Table 4), far above the intended alpha of 0.05. However, even with 35 treated units, state-clustered standard errors tend to over-reject, with type I error rates more than double the intended rate.

Heteroskedasticity-robust standard errors tend to have the most accurate type I error rates across all outcomes, whereas the accuracy of conventional model-based standard errors assuming IID error varies substantively across outcomes. Standard errors assuming IID error severely over-reject the null for effects on mass shooting fatalities and moderately under-reject the null for estimated effects on mass public shooting rates. Notably, no method of standard error estimation that we assessed performs adequately for the GVA mass shooting outcome: for estimating effects on GVA mass shooting incidents, average type I error rates exceed the nominal rate more than threefold (see Table S2).

Patterns in the type I error rates for the linear models are broadly similar to the Poisson models except that IID standard errors are less inaccurate for mass shooting fatality and mass public shooting outcomes (Table S7).

Correct rejection and Type S Error

Figure 1 shows how correct rejection rates vary across different effect sizes, outcomes, and numbers of treated states. Correct rejection rates are highest for mass shooting incidents (average correct rejection rate of 16%), slightly lower for mass shooting fatalities (average of 13%), and substantially lower for mass public shootings (average of 8%). As expected, correct rejection rates increase with effect size magnitude and the number of treated states. With only five treated states, correct rejection rates are unacceptably low and never exceed 26%. Even with 15 or 35 treated states, unless effect sizes are very large, the best-performing models correctly reject the null less than 20% of the time. In the best-case scenario with 35 treated states and policies that generate effects on the RR scale of 0.80 or 1/0.80, the best-performing models for the least sparse outcome (mass shootings) correctly reject the null 68.5% of the time. Linear models show similar patterns (see Figure S5 and Figure S6).

Issues of low power are particularly severe for evaluating mass public shootings. Even with very large effect sizes (RRs of 0.80 or 1.25) and a large number (35) of treated states, the best-performing model correctly identifies a significant effect on mass public shootings less than 30% of the time. With five treated states, correct rejection rates are all below 12%.

The consequences of trying to estimate policy effects under conditions with such low power are highlighted by Figure 2, which shows Type S error rates (i.e., the probability that, conditional on rejecting the null hypothesis, the effect in the sample is the *opposite* direction of the true effects). Type S error averaged across models and simulation conditions is about 10% for

GVA mass shootings, mass shootings, and mass shooting fatalities; and over 20% for mass public shootings. With only five treated states, nearly one in five significant effects is in the wrong direction for mass shooting incidents, mass shooting fatalities, and GVA mass shootings; one in three significant effects is in the wrong direction for mass public shootings. With only five treated states, Type S error rates for some models evaluating mass public shooting incidents are nearly 50%, meaning it is essentially a coin flip whether a significant effect is in the right direction.

Discussion

It is important to identify and implement policies that can reduce rare, harmful events like mass shootings, terrorist attacks, fatal child maltreatment, hate crimes, and losses from natural disasters. Mass shootings account for less than 3% of all firearm deaths, but the loss of life and broader effects on affected families and communities carry enormous social and economic costs (Cook 2020). Terrorist attacks and fatal child maltreatment are similarly infrequent and similarly carry outsized psychological and societal effects. Despite the importance of preventing these harmful events, our study suggests that standard empirical methods may not be able to provide evidence for state law effects on this class of rare, costly outcomes. In our application, even using data that represents a plausible census of mass shooting incidents across the entire US over a 45-year period, estimates of policy effects using standard statistical methods are too inaccurate to provide useful guidance about which policies can effectively reduce mass shooting incidents or fatalities.

These conclusions apply only to the range of effect sizes we evaluated, which were informed by results from an expert survey about likely effects of state laws on mass shootings but that likely represent a reasonable range of effect sizes for the effects of state laws on other types of rare outcomes. We can only evaluate policies that are popular enough to pass legislatures and can actually be implemented under U.S. law. This substantially constrains both the range of policies that can be empirically investigated and the data available for investigation. This challenge is compounded by the fact that no single existing or proposed policy is likely to substantially affect most rare outcomes. Policies with widespread implementation, such as the federal assault weapons ban, may still be designed in ways that limit their potential effectiveness (Cook and Donohue 2022). Furthermore, because mass shootings (Smart and Schell 2021; Silva and Greene-Colozzi 2021) and terrorist attacks (Gill 2015) are diverse with respect to perpetrator characteristics, motivations, weapons, and physical environments, even a policy that is highly effective in preventing incidents by one type of perpetrator or in one type of location will only prevent a small proportion of all incidents.

In light of these inherent limits on likely effect sizes, our results show that commonly used methods for evaluating how state policies affect rare outcomes are subject to substantial error, both in the estimates themselves and in the standard errors commonly used for inference. The existing literature using standard quasi-experimental tools for evaluating state policy effects on mass shootings should thus be interpreted cautiously. Meaningfully important policies for preventing a negative rare event, even policies that would prevent 20% of incidents, are very likely to result in nonsignificant findings. Effects this large are widely believed to be implausible

for firearm policies (Smart et al., 2021) and may very well be implausible for policies affecting other high-stakes rare events as well. When studies do find statistically significant effects of a policy, they are likely to overstate the size of the effects; in a situation with such little power, many significant effects may even be in the wrong direction. In particular, the field should be suspicious of any study where a plausible effect size of the policy being studied is smaller than the RMSE from the analogous simulation condition or the estimated standard error of the causal effect estimate. Such a level of accuracy generally corresponds to having a correct rejection rate near 17% and a Type S error near 1% and suggests that the study has minimal scientific utility. Indeed, several studies of mass shootings (Smart et al. 2026) have standard errors of the estimates that are more than twice the size of the effect sizes experts expect for those policies (Smart et al. 2021), a level of statistical precision that corresponds to only 7% power and 9% of statistically significant results being in the opposite direction of the true effect.

These results are discouraging particularly because the simulations were designed to represent the best possible scenario for causal inference with these mass shooting data. The models' assumption about the functional form of the causal effects are exactly true, there are no actual confounds that would necessitate more complex models, effects are homogeneous across states and do not vary systematically over time, the causal effect has an instant effect when implemented, and there are always at least five treated states with both pre- and post-implementation data. We also drop the most extreme estimates from the simulation on the assumption that the analyst would know that an estimated RR greater than 1000 or less than 0.001 could not be interpreted. If we had included those effects, the RMSE and Type S error rates would be higher, and the correct rejection rate would be lower.

Our findings suggest two key ways to improve statistical power for accurately identifying the effects of state policies on rare outcomes: increase the prevalence of the outcome or increase the prevalence of states with the policy of interest. For these rare, negative outcomes, it is worth trying the latter to try to avoid the former. Even then, we may never have a conclusive body of empirical evidence for how state policies affect any given rare outcome. However, this is no excuse for policy inaction. Policy makers should not wait for definitive scientific guidance that may never come before making attempts to save lives. There are other forms of descriptive or theoretical evidence that provide less rigorous causal insights but can be useful for policy decisions, and there are Bayesian alternatives to null hypothesis testing that can avoid some of the pitfalls of frequentist approaches in contexts with such low statistical power (McElreath 2015). In many cases, looking at policy effects on analogous common outcomes may be preferred to attempting to study rare outcomes directly. For example, a stronger evidence base supports which policies can reduce firearm homicides, and research has shown that long-term trends in the mass murder rate generally mirror those for homicide in general (Duwe 2007). Basing policy decisions on well measured estimates of policy effects on broader, more common forms of firearm violence—rather than the rarest, most extreme forms—may be better calibrated to reduce mass shootings and will almost certainly save more lives.

References

Berge, Laurent. 2023. "Efficient Maximum Likelihood Estimation with Multiple Fixed-Effects." Accessed November 5. https://cran.r-project.org/web/packages/FENmlm/vignettes/FENmlm.html.

Berge, Laurent. 2025. "On Standard-Errors." Accessed November 5. https://lrberge.github.io/fixest/articles/standard_errors.html.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (1):249–75.

Black, Bernard, Alex Hollingsworth, Leticia Nunes, and Kosali SImon. 2022. "Simulated Power Analyses for Observational Studies: An Application to the Affordable Care Act Medicaid Expansion." *Journal of Public Economics* 213

Clauset, Aaron, and Ryan Woodard. 2013. "Estimating the Historical and Future Probabilities of Large Terrorist Events." *The Annals of Applied Statistics* 7 (4):1838–65.

Cook, Philip J. 2020. "Thinking About Gun Violence." *Criminology & Public Policy* 19 (4):1371–93.

Cook, Philip J, and John J Donohue. 2022. "Regulating Assault Weapons and Large-Capacity Magazines for Ammunition." *JAMA* 328 (12):1191–92.

Deaton, A., and N Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science and Medicine* 210:2–21.

Duwe, Grant. 2007. *Mass Murder in the United States: A History*. McFarland & Co., Inc.

Fang, Xiangming, Derek S. Brown, and Curtis Florence. 2012. "The Economic Burden of Child Maltreatment in the United States and Implications for Prevention." *Child Abuse and Neglect* 36 (2):156–65.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9 (6):641–51. https://doi.org/10.1177/1745691614551642.

Gill, P. 2015. *Lone Actor Terrorists: A Behavioural Analysis*. Routledge.

Griffin, Beth Ann, Megan S Schuler, Elizabeth A Stuart *et al.* 2021. "Moving Beyond the Classic Difference-in-Differences Model: A Simulation Study Comparing Statistical Methods for Estimating Effectiveness of State-Level Policies." *BMC Research Methods* 21 (1)

Griswold, Max, Beth Ann Griffin, Max Rubinstein *et al.* 2026. "Assessing Bias and Precision in State Policy Evaluations: A Comparative Analysis of Time-Varying Estimators Using Policy Simulations." *American Journal of Epidemiology* 195 (7):2010–19.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal*

McElreath, Richard. 2015. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*.

Olsen, R. B., L. L. Orr, S. H. Bell, and E. A Stuart. 2013. "External Validity in Policy Evaluations That Choose Sites Purposively." *Journal of Policy Analysis and Management* 32 (1):107–21.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for STatistical Computing.

Schell, Terry L, Beth Ann Griffin, and Andrew R Morral. 2018. *Evaluating Methods to Estimate the Effect of State Laws on Firearm Deaths*. RAND Corporation.

Siegel, Michael, Max Goder-Reiser, Grant Duwe, Michael Rocque, James Alan Fox, and Emma E Fridel. 2020. "The Relation between State Gun Laws and the Incidence and Severity of Mass Public Shootings in the United States, 1976–2018." *Law and human behavior* 44 (5):347.

Silva, Jason R, and Emily Ann Greene-Colozzi. 2021. "Mass Shootings and Routine Activities Theory: The Impact of Motivation, Target Suitability, and Capable Guardianship on Fatalities and Injuries." *Victims & Offenders* 16 (4):565–86.

Smart, Rosanna, Andrew R Morral, James P Murphy, Rupa Jose, Amanda Charbonneau, and Terry L Schell. 2026. *The Science of Gun Policy: A Critical Synthesis of Research Evidence on the Effects of Gun Policies in the United States, Fifth Edition*. RAND Corporation.

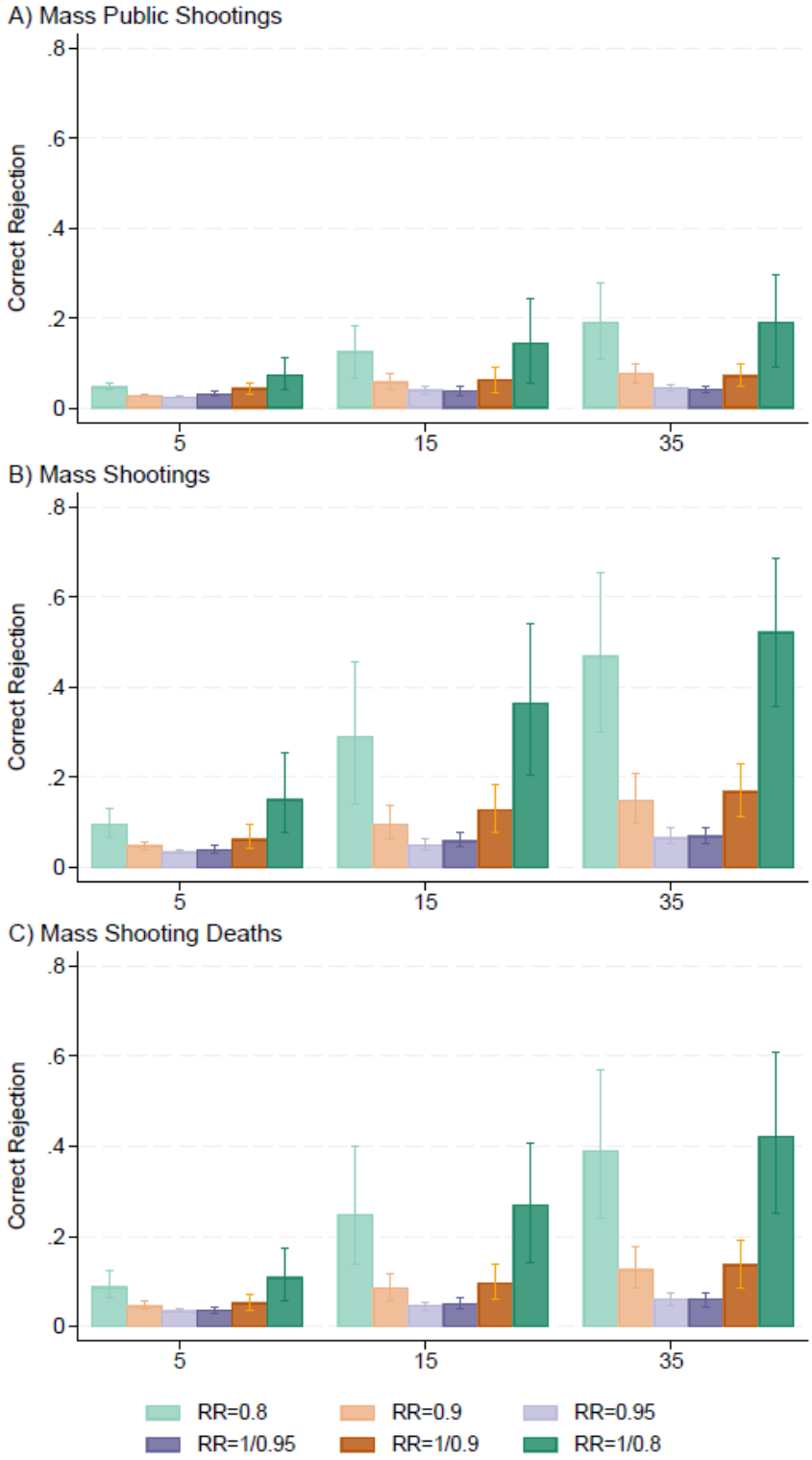
Smart, Rosanna, Andrew R Morral, James P Murphy, Rupa Jose, Amanda Charbonneau, and Sierra Smucker. 2024. *The Science of Gun Policy: A Critical Synthesis of Research Evidence on the Effects of Gun Policies in the United States, Fourth Edition*. RAND Corporation.

Smart, Rosanna, Andrew R Morral, and Terry L Schell. 2021. *The Magnitude and Sources of Disagreement among Gun Policy Experts*. RAND Corporation.

Smart, Rosanna, and Terry L Schell. 2021. "Mass Shootings in the United States." In *Contemporary Issues in Gun Policy: Essays from the Rand Gun Policy in America Project*. RAND.

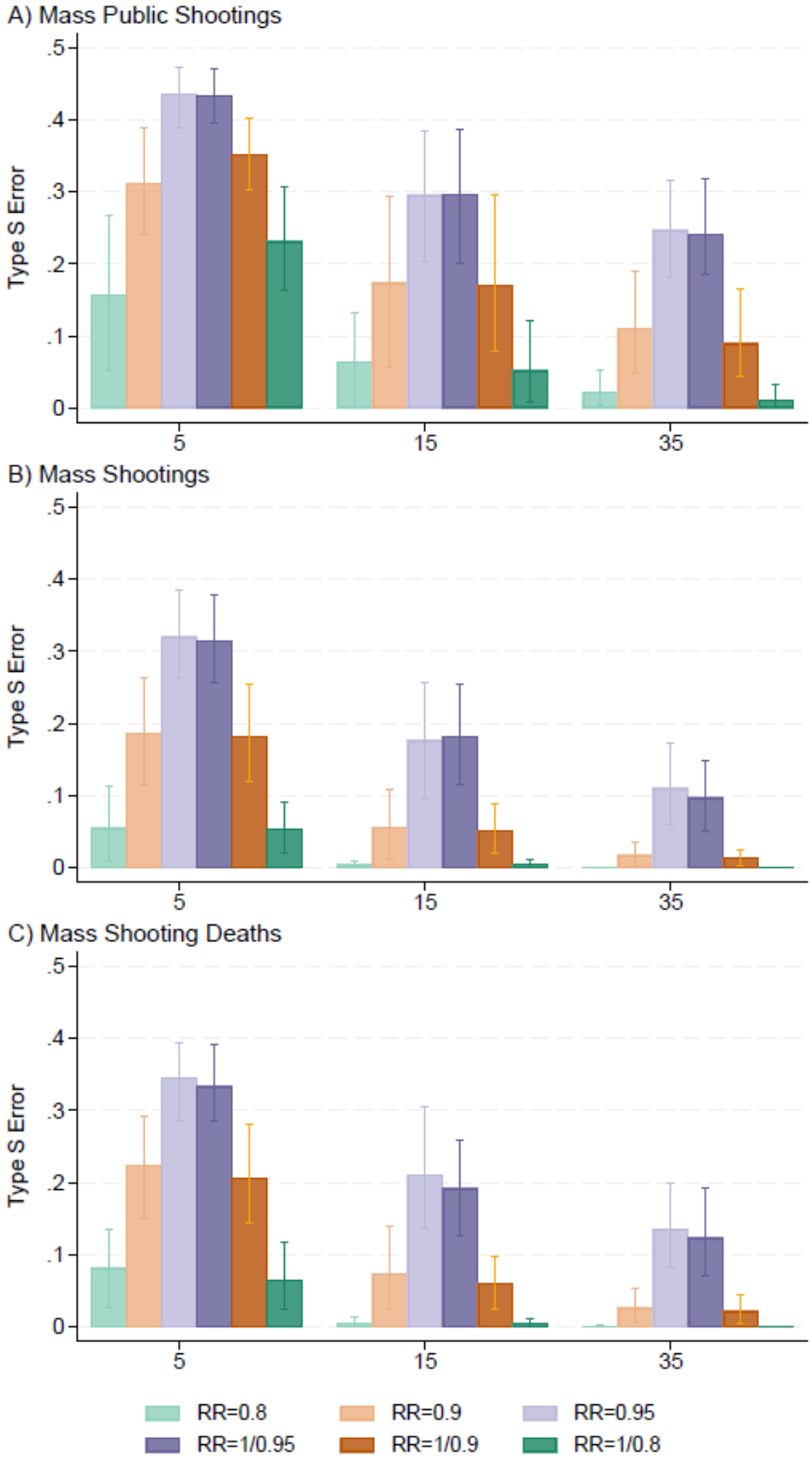
Figures

Figure 1. Correct Rejection Rates by Outcome and Simulation Condition



NOTES: Bar charts present the average correct rejection rate across all model specifications for a given simulation condition, and the error bars show the range from lowest to highest correct rejection rate on that margin. See Table S8 for numerical results.

Figure 2. Type S Error Rates, by Outcome and Simulation Condition



NOTES: Bar charts present average Type S error rates across all model specification variants for a given simulation condition, and the error bars show the range from lowest to highest Type S error rate on that margin. See Table S9 for numerical results.

Tables

Table 1. Distributional Characteristics of Outcomes Used in the Simulations

	Mass public shooting incidents	Mass shooting incidents	Mass shooting fatalities
Time period covered	1976–2020	1976–2020	1976–2020
Number of incidents or fatalities	166	897	4,456
Average count per state-year	0.07	0.40	1.98
Average count per state (full period)	3.32	17.94	89.12
Number (%) of states with 0 incidents	8 (16%)	0 (0%)	0 (0%)
Number (%) of state-years with 0 incidents	2,096 (93%)	1,634 (73%)	1,634 (73%)

NOTES: Excludes the District of Columbia.

Table 2. Summary of Root Mean Squared Error (RMSE), by Outcome

	Mass public shooting incidents	Mass shooting incidents	Mass shooting fatalities
All Conditions			
Best RMSE	0.166	0.098	0.109
Average RMSE	0.394	0.193	0.221
Worst RMSE	0.921	0.374	0.428
5 Treated States			
Best RMSE	0.375	0.213	0.242
Average RMSE	0.596	0.289	0.331
Worst RMSE	0.921	0.374	0.428
15 Treated States			
Best RMSE	0.201	0.121	0.137
Average RMSE	0.335	0.166	0.189
Worst RMSE	0.542	0.215	0.242
35 Treated States			
Best RMSE	0.166	0.098	0.109
Average RMSE	0.252	0.125	0.142
Worst RMSE	0.389	0.156	0.177

NOTES: RMSE = Root mean squared error of the estimates, expressed on a log IRR scale. The best, average, and worst RMSE refer to the distribution across the 252 RMSE estimates (21 simulation conditions x 12 effect estimation methods) from all conditions (top panel), or from the 84 RMSE estimates within each number of treated states condition.

Table 3. Regressions Predicting Root Mean Squared Error from Simulation Conditions by Outcome

	Mass Public Shooting Incidents	Mass Shooting Incidents	Mass Shooting Fatalities
Treated states			
35 Treated	Ref.	Ref.	Ref.
15 Treated	0.083	0.041	0.047
	(0.009)	(0.003)	(0.003)
5 Treated	0.344	0.165	0.189
	(0.009)	(0.003)	(0.003)
Effect Size (RR terms)			
0.8	0.008	0.002	0.000
	(0.014)	(0.005)	(0.005)
0.9	0.004	0.001	-0.000
	(0.014)	(0.005)	(0.005)
0.95	0.002	0.000	-0.000
	(0.014)	(0.005)	(0.005)
1.00	Ref.	Ref.	Ref.
1.05	-0.002	-0.001	-0.002
	(0.014)	(0.005)	(0.005)
1.11	-0.004	-0.002	-0.003
	(0.014)	(0.005)	(0.005)
1.25	-0.007	-0.003	-0.005
	(0.014)	(0.005)	(0.005)
Time-varying covariates			
2 Covariates	Ref.	Ref.	Ref.
19 Covariates	0.055	-0.001	-0.001
	(0.008)	(0.003)	(0.003)
Geography			
Region fixed effects	Ref.	Ref.	Ref.
State fixed effects	0.269	0.095	0.103
	(0.008)	(0.003)	(0.003)
Time			
Linear	Ref.	Ref.	Ref.
Spline	-0.000	-0.001	-0.006
	(0.009)	(0.003)	(0.003)
Year fixed effects	0.011	0.002	-0.002
	(0.009)	(0.003)	(0.003)
Constant	0.086	0.078	0.095
	(0.014)	(0.005)	(0.005)
Observations	252	252	252
Adjusted-R2	0.92	0.94	0.95

NOTES: Higher numbers indicate that predictor value is associated with greater error in the causal effect estimate relative to the reference value. Standard errors for individual coefficients in parentheses. RR=Rate ratio. Results for primary analysis, i.e., Poisson model of mass shootings with population offset.

Table 4. Average Rate of Type I Error by Type of Standard Error Estimate

	Number of Treated States		
	5	15	35
Mass Public Shootings			
State-clustered	0.192	0.132	0.116
Heteroskedasticity-robust	0.042	0.045	0.048
IID	0.026	0.034	0.038
Mass Shooting Events			
State-clustered	0.211	0.140	0.120
Heteroskedasticity-robust	0.079	0.079	0.076
IID	0.075	0.082	0.081
Mass Shooting Deaths			
State-clustered	0.234	0.147	0.128
Heteroskedasticity-robust	0.084	0.081	0.076
IID	0.497	0.503	0.497

NOTES: Standard errors were computed using the “cluster”, “hetero,” and “iid” specifications in fixest.

SUPPLEMENTARY MATERIAL

Appendix S1. Detailed Methods

Methods used in the literature: The simulation was designed to assess the range of statistical methods that have been employed in the literature assessing the causal effect of state-level policy on state-level mass shooting. We selected the range of methods we investigated by looking at ten studies identified within a recent systematic review (Smart et al., 2026) as estimating policy effects on some type of mass shooting outcome (any incidents, number of incidents, number of fatalities). All of these studies used model-based effect estimation of causal effects. Of these ten studies (Blau et al., 2026; Duwe et al., 2002; Fridel, 2021a, 2021b; Gius, 2015, 2024; Klarevas et al., 2019; Luca et al., 2020; Siegel et al., 2020; Webster et al., 2020):

- Five used a log link model (Poisson or Negative Binomial), three used a linear model, and two used a logistic link.
- The ten studies used 29 years of outcome data on average, ranging between 22 and 43 years.
- Five of the models were estimated using population weights (or population offsets in log link models), and five weighted all states-years equally.
- The number of states with policy transitions being evaluated was not always presented in the articles, but appears to range between five and 32 states, depending on the policy
- Seven included state fixed effects, two include region fixed effects, and one used no controls for geographic variation.
- Six used year fixed effects to control for time trends. Two used linear time trends and two used quadratic time trends.
- The studies used 13 covariates on average (in addition to geographic or temporal controls) ranging between 2 and 25 covariates.

Simulation details: Not all effect sizes are exactly realizable in discrete data. For example, if there are five events observed in treated state-years, we may realize an IRR of 1.2 by increasing the number of events in treated state-years to 6, but we cannot realize an IRR between 1 and 1.2. Therefore, for each simulated set of treated state-years, we first identified the potentially realizable IRRs and selected the target number of simulation events to correspond to the realizable IRR closest to the nominal IRR. Next, we added (or subtracted) counts to each state-year so as to preserve the major features of the observed data. We would prefer to add events, for example, to a state with an observed higher frequency of mass shootings than to a state that has never experienced one. Specifically, we added/subtracted events to/from state-years in proportion to a number which is the average of the actual count in that state-year and an estimated count in that state-year from a machine learning model fit to the actual data. This avoided difficulties with allocating to state-years with 0 observed events. This process was repeated until the target number of events was achieved.

In some cases, this process of incorporating causal effects into the data failed (e.g., the algorithm cannot create a negative effect when there were no actual incidents in any treated state-year). This occurred for two non-null policy randomizations out of 10,000. When this occurred, we dropped the same policy randomizations for all models, effect sizes, number of treated states, and outcomes. For this reason, our results are based on slightly fewer than 10,000 simulations.

Additionally, we conducted post-processing to filter out simulation runs that resulted in effect estimates indicative of degenerate solutions (i.e., rate ratios greater than 1000 or less than

1/1000). We anticipated such results would be accompanied by warning messages in many standard software packages or would alert the analyst to a potential problem with their models, and we did not want our assessment of a given model's performance to be highly influenced by such extreme outliers. Thus, we excluded these extreme estimates from our analyses. Table S1 details how many runs were subject to this exclusion criteria across all model specifications.

Model types and covariates: We evaluate twelve model types that vary in complexity with two levels of geographic controls (state fixed effects, region fixed effects), three levels of time effects (linear effect of time, cubic B-spline of time with three degrees of freedom, year fixed effects), and two levels of state-level time-varying covariate inclusion (two, nineteen). When using only two covariates, those are: firearm suicides as a proportion of all suicides commonly used as a proxy for gun ownership rates, and population density. When using 19 covariates, those are: firearm suicides as a proportion of all suicides, population density, incarceration rate, proportion of the population that identified as Black, proportion of the population ages 15–29, unemployment rate, poverty rate, education, per-capita alcohol consumption, the violent crime rate, Gini index, proportion of the population male, per capita income; and indicators for whether the state had a firearm trafficking prohibition law, firearm prohibition for individuals subject to domestic violence restraining orders, handgun permit requirement, age-21 limit for handgun purchase, and assault weapons ban. Covariate data came from Siegel et al. (2020).

Choice of effect sizes: The conclusions we draw are contingent on the range of effect sizes we chose to investigate. These sizes were selected as plausible effects of state policies on mass shootings based on a previous survey of experts in gun policy, including academic experts, legal experts, and individuals engaged in advocacy (Smart et al., 2021). Specifically, that study surveyed 173 firearm policy experts on the effects they expected on mass shooting outcomes for 19 common state firearm policies. The median absolute expected effect size across all policies was a 4% change in mass shootings among experts who favor restrictive firearm policies, and 0% change for those who favor permissive firearm policies. The policy with the largest expected effect on mass shootings for restrictive experts was a ban on the sale of new assault weapons and high-capacity magazines, for which the median restrictive expert expected a 15% reduction and the median permissive expert expected no effect. The policy with the largest expected effects for permissive experts was for arming school personnel in K–12 educational settings, for which permissive experts expected a 15% reduction and restrictive experts expected no effect.

Dominated model variants not shown: Our initial simulation design included several additional model variants that were found to perform worse or equivalently than similar models and thus were dropped from our primary results. These include Poisson models with population as a covariate rather than an offset; unweighted linear models of per capita rate outcomes; and population-weighted linear models of count outcomes. We found that Poisson models with a count outcome and population as covariate performed similarly as Poisson models with a count outcome and population as an offset. Our primary analysis thus focuses on the count model with population offset as this is more commonly seen in the literature. We found that the linear population-weighted rate model almost always outperformed the unweighted rate model in terms of RMSE, thus we dropped the unweighted linear rate model from consideration. In contrast, we found that the unweighted linear count models always outperformed the population-weighted linear count models, thus we dropped the population-weighted linear count model from

consideration. Finally, we also originally evaluated specification variants that controlled for state-specific intercepts and state-specific slopes. Regardless of outcome or model condition, null effect simulations showed that models with state-specific slopes were always dominated by models with only state fixed-effects or only region fixed-effects. The amount of error added with the introduction of state-specific slopes was substantial, about 1.2 in IRR terms for mass shooting events. Because of this, we do not discuss performance of these models further.

Table S1. Models and Simulation Conditions Subject to Filtering Due to Extreme Effect Estimates

Treated states	Outcome	Geography	Effect size	Percentage filtered
Five	Mass public shooting incidents	State FE	0.80	2.96%
Five	Mass public shooting incidents	State FE	0.90	2.96%
Five	Mass public shooting incidents	State FE	0.95	2.96%
Five	Mass public shooting incidents	State FE	1.00	2.96%
Five	Mass public shooting incidents	State FE	1.05	2.96%
Five	Mass public shooting incidents	State FE	1.11	2.96%
Five	Mass public shooting incidents	State FE	1.25	2.96%
Five	Mass shooting incidents	State FE	0.80	0.02%
Five	Mass shooting incidents	State FE	0.90	0.02%
Five	Mass shooting incidents	State FE	0.95	0.02%
Five	Mass shooting incidents	State FE	1.00	0.02%
Five	Mass shooting incidents	State FE	1.05	0.02%
Five	Mass shooting incidents	State FE	1.11	0.02%
Five	Mass shooting incidents	State FE	1.25	0.02%
Five	Mass shooting fatalities	State FE	0.80	0.02%
Five	Mass shooting fatalities	State FE	0.90	0.02%
Five	Mass shooting fatalities	State FE	0.95	0.02%
Five	Mass shooting fatalities	State FE	1.00	0.02%
Five	Mass shooting fatalities	State FE	1.05	0.02%
Five	Mass shooting fatalities	State FE	1.11	0.02%
Five	Mass shooting fatalities	State FE	1.25	0.02%
Five	GVA mass shooting incidents	State FE	0.80	0.30%
Five	GVA mass shooting incidents	State FE	0.90	0.27%
Five	GVA mass shooting incidents	State FE	0.95	0.29%
Five	GVA mass shooting incidents	State FE	1.00	0.26%
Five	GVA mass shooting incidents	State FE	1.05	0.24%
Five	GVA mass shooting incidents	State FE	1.11	0.24%
Five	GVA mass shooting incidents	State FE	1.25	0.22%

NOTE: Simulation runs for the Poisson models were filtered out if the estimated IRR was greater than 1000 or less than 0.001. The extent of filtering was only a function of the number of treated states, outcome, geographic controls, and effect size; filtering did not vary depending on how time trends or time-varying controls were included in the models.

Appendix S2. Primary results for alternative definition of mass shootings based on Gun Violence Archive (GVA) data)

There is no universally agreed upon definition of "mass shooting," although historically the fatality threshold of four or more has been the most common definition employed in the literature. The historical reliance on fatality thresholds results in part from data limitations: fatal shootings are measured far more accurately than nonfatal shootings. However, in the past decade, an alternative source of data has become more commonly used to measure multiple-victim shooting incidents without an imposed fatality threshold. In particular, the Gun Violence Archive (GVA) relies on a combination of administrative data and media reporting to attempt to inventory all injurious shootings nationwide, and it produces a data series on "mass shooting incidents" based on an incident definition that requires four or more people killed or injured with a firearm. GVA thus captures more incidents in each year because it includes both injuries and fatalities in its threshold criteria for what constitutes a mass shooting, but it is only available starting in 2014.

GVA incidents are far less sparse than our primary mass shooting outcomes, but there is an important tradeoff between more incidents per year and a shorter period covered. The shorter period is challenging to compare with the longer time series because the likelihood of a high number of states adopting a given policy within a narrow timeframe may be quite low, and the timing of policies will, by our simulation design, necessarily be more temporally clustered when analyzing the GVA data versus the other time series. Specifically, for the *GVA mass shooting* outcome, all policy implementations must occur in the years 2015-2021 while they may occur over a longer period (1977-2019) for the other outcomes. For this reason, the number of policy transitions available for effect estimation might be seen as unrealistically high in the GVA data, relative to the types of policies that have typically been studied, making the GVA results less comparable to the other outcomes.

That said, given the increasing use of GVA for understanding mass shootings, we below present results on model performance specific to the GVA mass shooting incident data. These results draw on data from GVA for 2014 through 2022, which covers 3,962 total mass shooting incidents (defined as incidents with four or more people killed or injured by a firearm). The average number of incidents per state-year is 8.8, 22 times higher than the average for mass shooting incidents with four or more killed (see Table 1). The GVA mass shooting outcome measure also has a far lower percentage of state-years with zero incidents, 22% (n=98) versus the 73% (n=1,634) of state-years with zero incidents in the mass shooting outcome measure evaluated in our primary analysis.

Table S2. Model Performance for Gun Violence Archive Mass Shooting Incident Data, Poisson Models

	Number of Treated States		
	5	15	35
Average rate of Type I error			
State-clustered	0.294	0.182	0.155
Heteroskedasticity-robust	0.223	0.183	0.157
IID	0.240	0.249	0.231
Root-mean squared error (rmse)			
Best RMSE	0.258	0.127	0.087
Average RMSE	0.299	0.167	0.122
Worst RMSE	0.383	0.123	0.182
Correct rejection rates			
Lowest correct rejection rate	0.030	0.044	0.049
Average correct rejection rate	0.065	0.154	0.246
Highest correct rejection rate	0.156	0.412	0.685
Type S error rates			
Lowest Type S error rate	0.016	0.000	0.000
Average Type S error rate	0.187	0.064	0.033
Highest Type S error rate	0.372	0.227	0.166

NOTES: IID=Independent and identically distributed. Standard errors were computed using the "cluster", "hetero," and "iid" specifications in fixest. The Best, Average, and Worst RMSE refer to the distribution across the 84 RMSE estimates within each number of treated states condition. The lowest, average, and highest correct rejection rates refer to the distribution across the 72 correct rejection rate estimates within each number of treated states condition (excluding the null effect simulations). The lowest, average, and highest Type S error rates refer to the distribution across the 72 estimates within each number of treated states condition (excluding the null effect simulations).

Table S3. Regressions Predicting Root Mean Squared Error from Simulation Conditions for Outcome of Gun Violence Archive Mass Shooting Incidents

Treated states (Ref: 35 treated)	
15 Treated	0.044
	(0.004)
5 Treated	0.177
	(0.004)
Effect size in risk ratio terms (Ref: 1.00)	
0.8	-0.000
	(0.006)
0.9	-0.001
	(0.006)
0.95	-0.001
	(0.006)
1.05	-0.002
	(0.006)
1.11	-0.004
	(0.006)
1.25	-0.006
	(0.006)
Time-varying covariates (Ref: 2 covariates)	
19 Covariates	-0.038
	(0.003)
Geography (Ref: Region fixed effects)	
State fixed effects	-0.053
	(0.003)
Time (Ref: Linear)	
Spline	-0.002
	(0.004)
Year fixed effects	-0.003
	(0.004)
Constant	0.171
	(0.006)
Observations	252
Adjusted-R2	0.90

NOTES: Higher numbers indicate that predictor value is associated with greater error in the causal effect estimate relative to the reference value. Standard errors for individual coefficients in parentheses. Results for Poisson model of mass shootings with population offset.

Appendix S3. Supplementary Tables and Figures

Table S4. Summary of Root Mean Squared Error (RMSE) for Linear Models with Outcomes Specified as Rates per 1,000 population

	Mass public shooting incidents	Mass shooting incidents	Mass shooting fatalities
All Conditions			
Best RMSE	0.235	0.704	3.893
Average RMSE	0.525	1.376	8.028
Worst RMSE	1.183	2.592	15.217
5 Treated States			
Best RMSE	0.528	1.490	9.168
Average RMSE	0.806	2.027	12.054
Worst RMSE	1.183	2.592	15.217
15 Treated States			
Best RMSE	0.294	0.879	5.053
Average RMSE	0.438	1.193	6.870
Worst RMSE	0.654	1.574	8.700
35 Treated States			
Best RMSE	0.235	0.704	3.893
Average RMSE	0.332	0.908	5.158
Worst RMSE	0.468	1.163	6.504

NOTES: RMSE = Root mean squared error of the estimates, expressed on a log IRR scale. The Best, Average, and Worst RMSE refer to the distribution across the 252 RMSE estimates (21 simulation conditions x 12 effect estimation methods) from all conditions (top panel), or from the 84 RMSE estimates within each number of treated states condition. RMSE cannot be compared across outcomes.

Table S5. Summary of Root Mean Squared Error (RMSE) for Linear Models with Outcomes Specified as Counts

	Mass public shooting incidents	Mass shooting incidents	Mass shooting fatalities
All Conditions			
Best RMSE	0.014	0.042	0.252
Average RMSE	0.026	0.076	0.458
Worst RMSE	0.051	0.135	0.818
5 Treated States			
Best RMSE	0.028	0.086	0.560
Average RMSE	0.038	0.110	0.674
Worst RMSE	0.051	0.135	0.818
15 Treated States			
Best RMSE	0.017	0.053	0.333
Average RMSE	0.023	0.067	0.401
Worst RMSE	0.031	0.082	0.485
35 Treated States			
Best RMSE	0.014	0.042	0.252
Average RMSE	0.018	0.051	0.300
Worst RMSE	0.024	0.060	0.366

NOTES: RMSE = Root mean squared error of the estimates, expressed on a log IRR scale. The Best, Average, and Worst RMSE refer to the distribution across the 252 RMSE estimates (21 simulation conditions x 12 effect estimation methods) from all conditions (top panel), or from the 84 RMSE estimates within each number of treated states condition. RMSE cannot be compared across outcomes.

Table S6. Regressions Predicting Root Mean Squared Error (RMSE) from Simulation Condition and Model Characteristics, by Outcome

	Linear Rate, Weighted (scaled)[a]				Linear Count, Unweighted			
	Mass public shootings	Mass shootings	Mass shooting deaths	GVA mass shootings	Mass public shootings	Mass shootings	Mass shooting deaths	GVA mass shootings
Treated States (Ref: 35)								
15 Treated	0.105	0.029	0.171	0.533	0.005	0.016	0.100	0.328
	(0.013)	(0.002)	(0.011)	(0.033)	(0.000)	(0.001)	(0.004)	(0.013)
5 Treated	0.474	0.112	0.690	1.818	0.020	0.059	0.374	1.122
	(0.013)	(0.002)	(0.011)	(0.033)	(0.000)	(0.001)	(0.004)	(0.013)
Effect Size (Ref: RR=1)								
0.80	-0.005	-0.002	-0.024	-0.073	-0.000	0.000	-0.009	-0.015
	(0.019)	(0.003)	(0.017)	(0.050)	(0.001)	(0.001)	(0.007)	(0.020)
0.90	-0.002	-0.001	-0.013	-0.054	-0.000	-0.000	-0.006	-0.028
	(0.019)	(0.003)	(0.017)	(0.050)	(0.001)	(0.001)	(0.007)	(0.020)
0.95	-0.001	-0.001	-0.007	-0.033	-0.000	-0.000	-0.003	-0.019
	(0.019)	(0.003)	(0.017)	(0.050)	(0.001)	(0.001)	(0.007)	(0.020)
1.05	0.001	0.000	0.001	-0.003	0.000	0.000	0.001	-0.011
	(0.019)	(0.003)	(0.017)	(0.050)	(0.001)	(0.001)	(0.007)	(0.020)
1.11	0.003	0.001	0.002	0.009	0.000	0.000	0.002	-0.000
	(0.019)	(0.003)	(0.017)	(0.050)	(0.001)	(0.001)	(0.007)	(0.020)
1.25	0.009	0.002	0.008	0.067	0.001	0.001	0.011	0.088
	(0.019)	(0.003)	(0.017)	(0.050)	(0.001)	(0.001)	(0.007)	(0.020)
Number of Covariates (Ref: 2)								
19 Covariates	0.066	-0.002	-0.016	-0.323	0.002	-0.000	0.001	0.029
	(0.010)	(0.002)	(0.009)	(0.027)	(0.000)	(0.001)	(0.004)	(0.011)
Geography (Ref: Region FE)								
State FE	0.314	0.065	0.337	-0.809	0.011	0.027	0.144	-0.591
	(0.010)	(0.002)	(0.009)	(0.027)	(0.000)	(0.001)	(0.004)	(0.011)
Time (Ref: Linear)								
Spline	0.001	0.000	-0.012	-0.029	0.000	0.000	-0.002	0.000
	(0.013)	(0.002)	(0.011)	(0.033)	(0.000)	(0.001)	(0.004)	(0.013)
Year FE	0.010	0.002	-0.001	-0.058	0.000	0.000	0.001	-0.002
	(0.013)	(0.002)	(0.011)	(0.033)	(0.000)	(0.001)	(0.004)	(0.013)
Constant	0.138	0.059	0.365	2.005	0.011	0.037	0.228	1.186
	(0.019)	(0.003)	(0.016)	(0.048)	(0.001)	(0.001)	(0.006)	(0.019)
Observations	252	252	252	252	252	252	252	252
Adjusted-R2	0.91	0.94	0.96	0.94	0.96	0.96	0.97	0.98

NOTES: RR=rate ratio. RMSE is on the scale of the outcome variable, and thus should not be compared across the count and rate model, nor across the outcomes. [a]For presentation purposes, in the rate model, coefficients and standard errors are multiplied by 100 (mass shootings, mass shooting deaths, GVA mass shootings) or by 1,000 (mass public shootings).

Table S7. Summary of Type I Error, Averages

	Linear Rate, Weighted			Linear Count, Unweighted		
Treated states	5	15	35	5	15	35
Mass Public Shootings						
Cluster	0.188	0.115	0.101	0.175	0.121	0.109
Hetero	0.047	0.041	0.044	0.051	0.048	0.050
IID	0.026	0.020	0.020	0.054	0.052	0.052
Mass Shooting Events						
Cluster	0.214	0.141	0.121	0.169	0.118	0.109
Hetero	0.080	0.078	0.074	0.096	0.098	0.090
IID	0.067	0.072	0.071	0.091	0.095	0.088
Mass Shooting Deaths						
Cluster	0.222	0.138	0.124	0.165	0.118	0.113
Hetero	0.078	0.071	0.070	0.098	0.097	0.095
IID	0.075	0.077	0.072	0.114	0.111	0.098
GVA Mass Shootings						
Cluster	0.288	0.185	0.157	0.205	0.139	0.125
Hetero	0.241	0.200	0.170	0.204	0.180	0.160
IID	0.219	0.239	0.210	0.153	0.179	0.163

NOTES: Standard errors were computed using the “cluster”, “hetero,” and “iid” specifications in fixest.

Table S8. Summary of Correct Rejection Rates, by Outcome

	Mass public shooting incidents	Mass shooting incidents	Mass shooting fatalities	GVA mass shooting incidents
All Conditions				
Lowest correct rejection rate	0.025	0.030	0.030	0.030
Average correct rejection rate	0.075	0.159	0.131	0.155
Highest correct rejection rate	0.297	0.685	0.609	0.685
5 Treated States				
Lowest correct rejection rate	0.025	0.030	0.030	0.030
Average correct rejection rate	0.042	0.072	0.061	0.065
Highest correct rejection rate	0.113	0.254	0.174	0.156
15 Treated States				
Lowest correct rejection rate	0.029	0.040	0.037	0.044
Average correct rejection rate	0.078	0.165	0.133	0.154
Highest correct rejection rate	0.243	0.540	0.407	0.412
35 Treated States				
Lowest correct rejection rate	0.034	0.052	0.044	0.049
Average correct rejection rate	0.104	0.241	0.199	0.246
Highest correct rejection rate	0.297	0.685	0.609	0.685

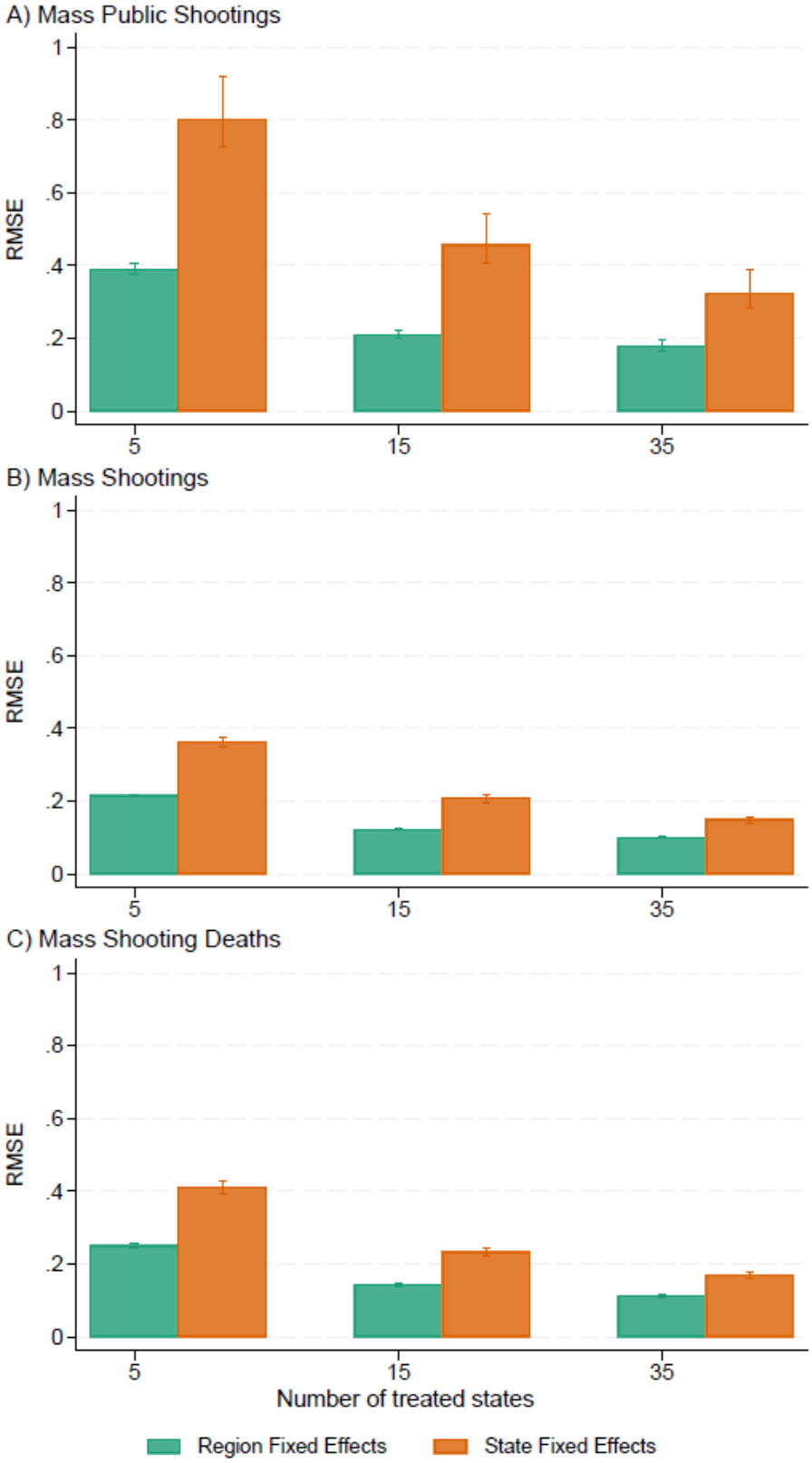
NOTES: The lowest, average, and highest correct rejection rates refer to the distribution across the 216 correct rejection rate estimates (18 simulation conditions x 12 effect estimation methods) from all conditions (top panel), or from the 72 correct rejection rate estimates within each number of treated states condition.

Table S9. Summary of Type S Error Rates, by Outcome

	Mass public shooting incidents	Mass shooting incidents	Mass shooting fatalities	GVA mass shooting incidents
All Conditions				
Lowest Type S error rate	0.002	0.000	0.000	0.000
Average Type S error rate	0.205	0.101	0.117	0.094
Highest Type S error rate	0.473	0.385	0.394	0.372
5 Treated States				
Lowest Type S error rate	0.053	0.010	0.024	0.016
Average Type S error rate	0.320	0.185	0.209	0.187
Highest Type S error rate	0.473	0.385	0.394	0.372
15 Treated States				
Lowest Type S error rate	0.002	0.000	0.000	0.000
Average Type S error rate	0.175	0.079	0.091	0.064
Highest Type S error rate	0.387	0.256	0.304	0.227
35 Treated States				
Lowest Type S error rate	0.002	0.000	0.000	0.000
Average Type S error rate	0.120	0.040	0.051	0.033
Highest Type S error rate	0.318	0.171	0.200	0.166

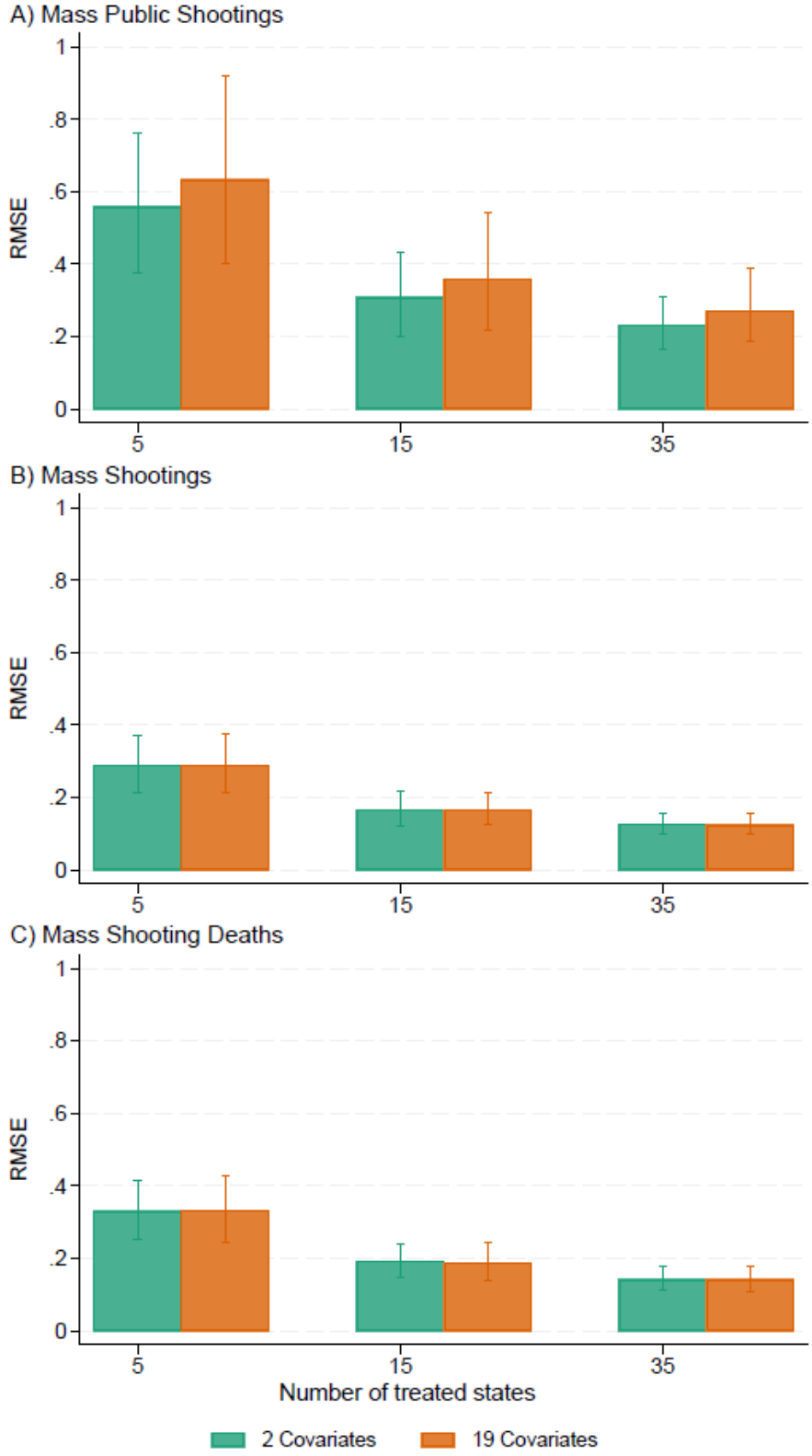
NOTES: The lowest, average, and highest Type S error rates refer to the distribution across the 216 Type S error rate estimates (18 simulation conditions x 12 effect estimation methods) from all conditions (top panel), or from the 72 Type S error rate estimates within each number of treated states condition.

Figure S1. Average and Range of Root Mean Squared Error, by Geography



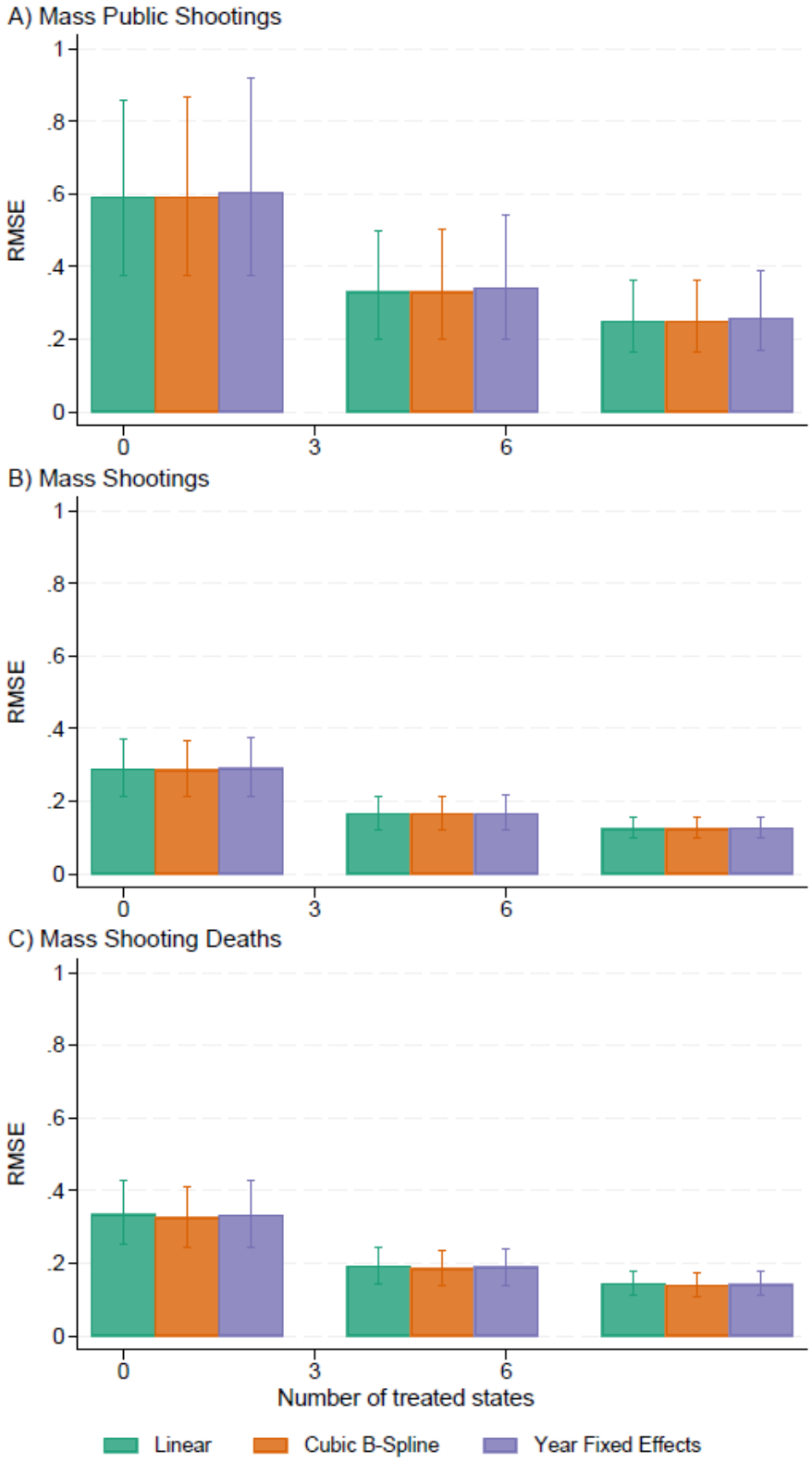
NOTES: Bar chart presents the average RMSE across all model specifications for a given simulation condition, and the error bar shows the range from lowest to highest RMSE on that margin.

Figure S2. Average and Range of Root Mean Squared Error, by Number of Time-Varying Covariates



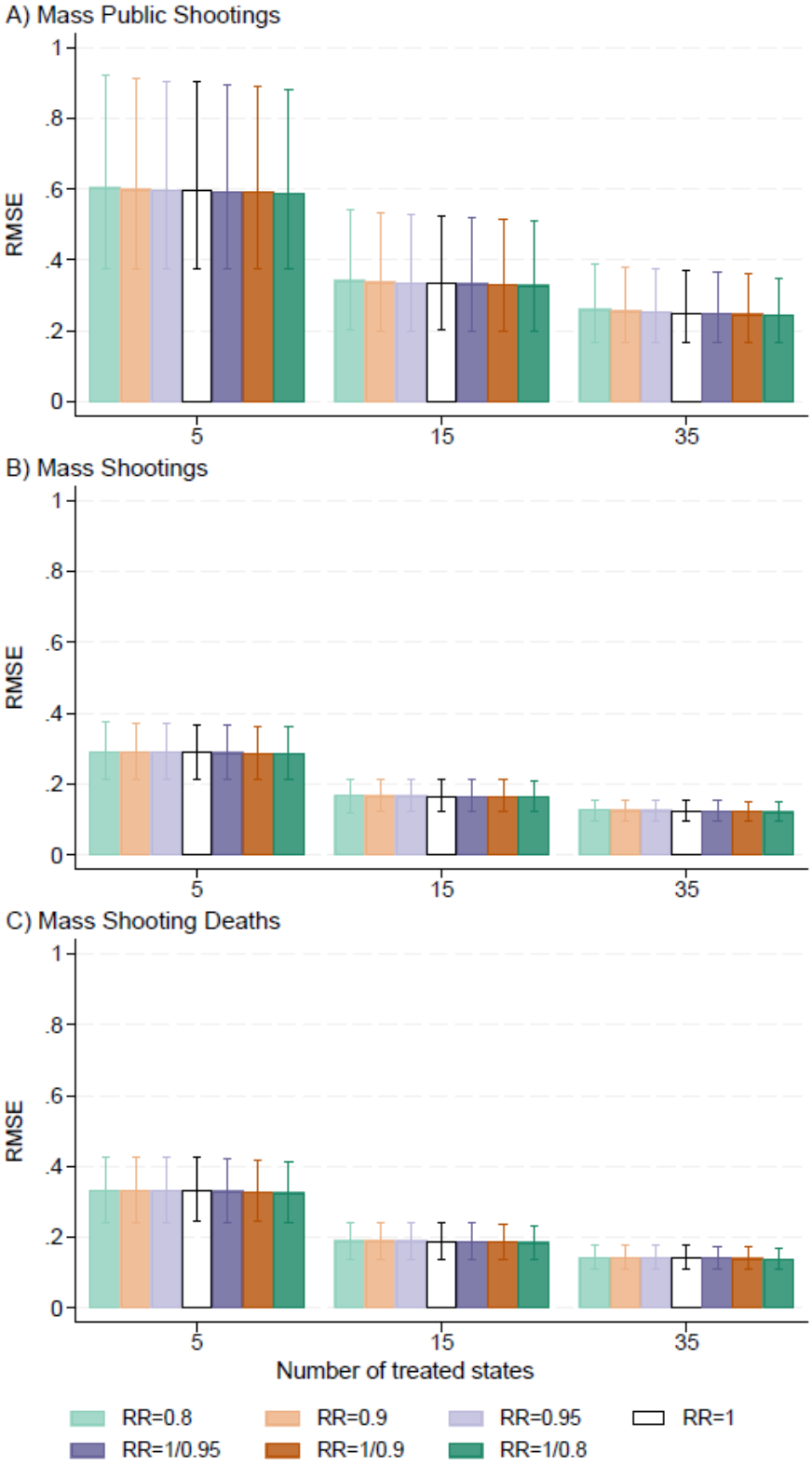
NOTES: Bar chart presents the average RMSE across all model specifications for a given simulation condition, and the error bar shows the range from lowest to highest RMSE on that margin.

Figure S3. Average and Range of Root Mean Squared Error, by Controls for National Trends



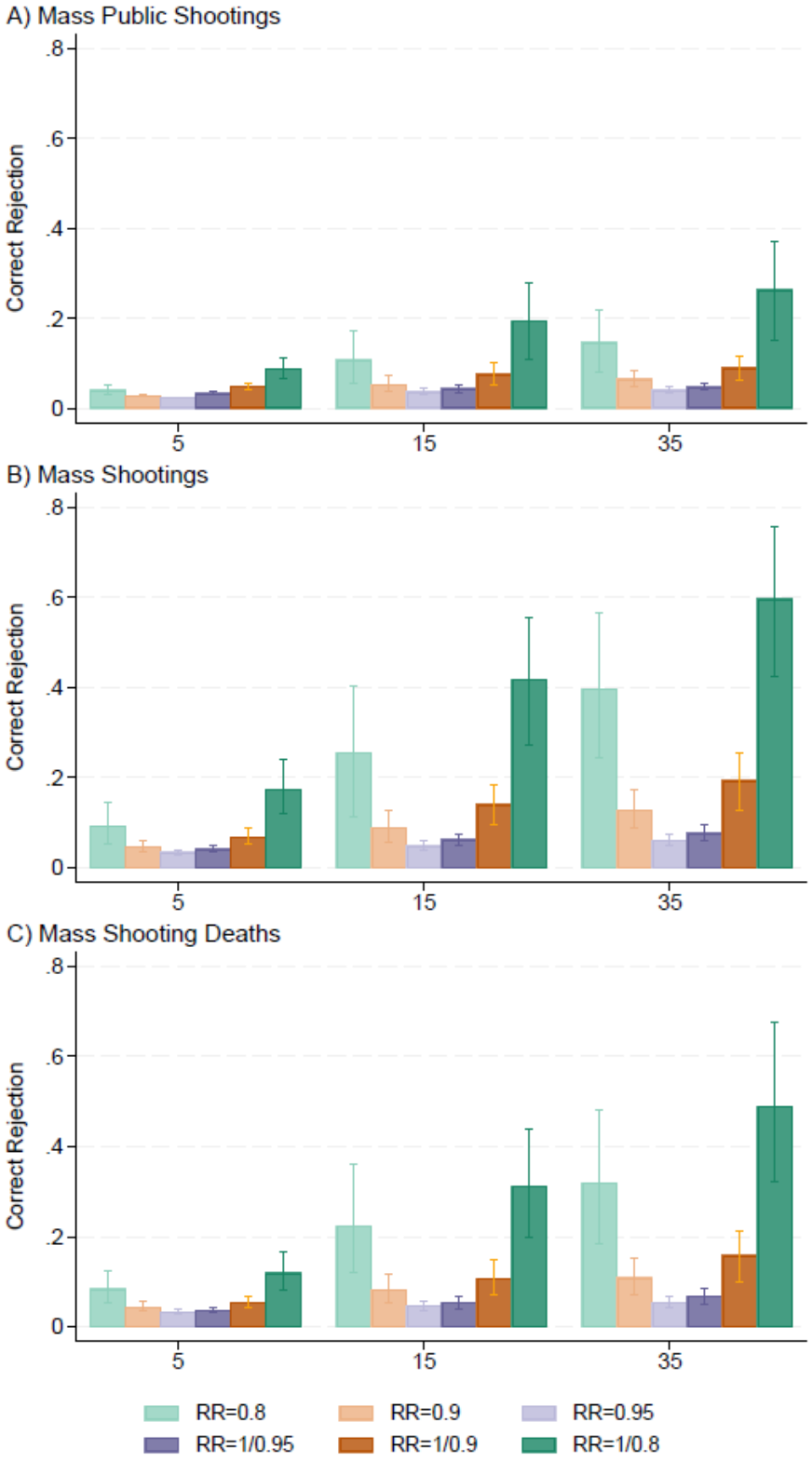
NOTES: Bar chart presents the average RMSE across all model specifications for a given simulation condition, and the error bar shows the range from lowest to highest RMSE on that margin.

Figure S4. Average and Range of Root Mean Squared Error, by Effect Size



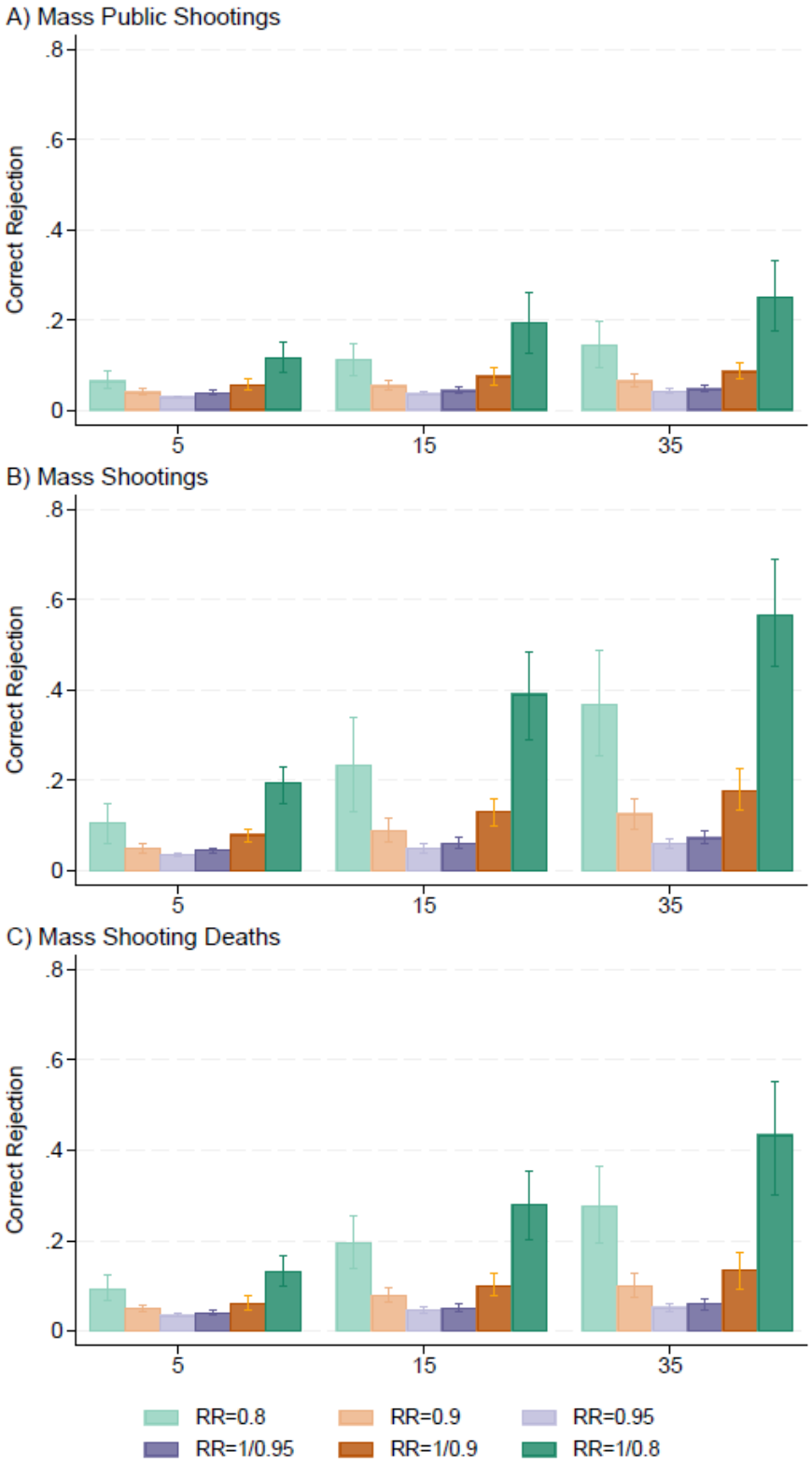
NOTES: Bar chart presents the average RMSE across all model specifications for a given simulation condition, and the error bar shows the range from lowest to highest RMSE on that margin.

Figure S5. Correct Rejection Rate for Linear Rate Models



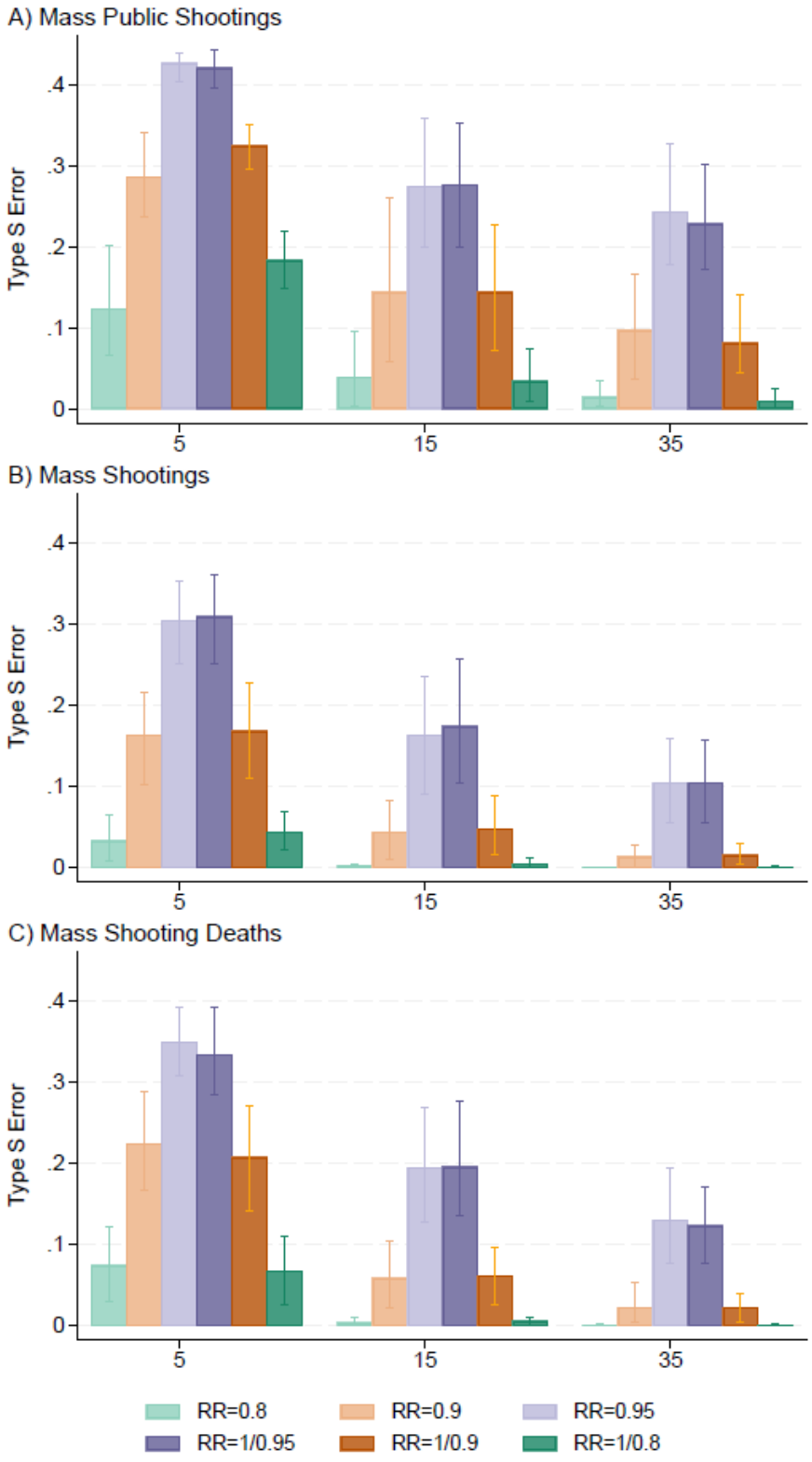
NOTES: Bar chart presents the average correct rejection rate across all model specifications for a given simulation condition, and the error bar shows the range from lowest to highest correct rejection rate on that margin.

Figure S6. Correct Rejection Rate for Linear Count Models



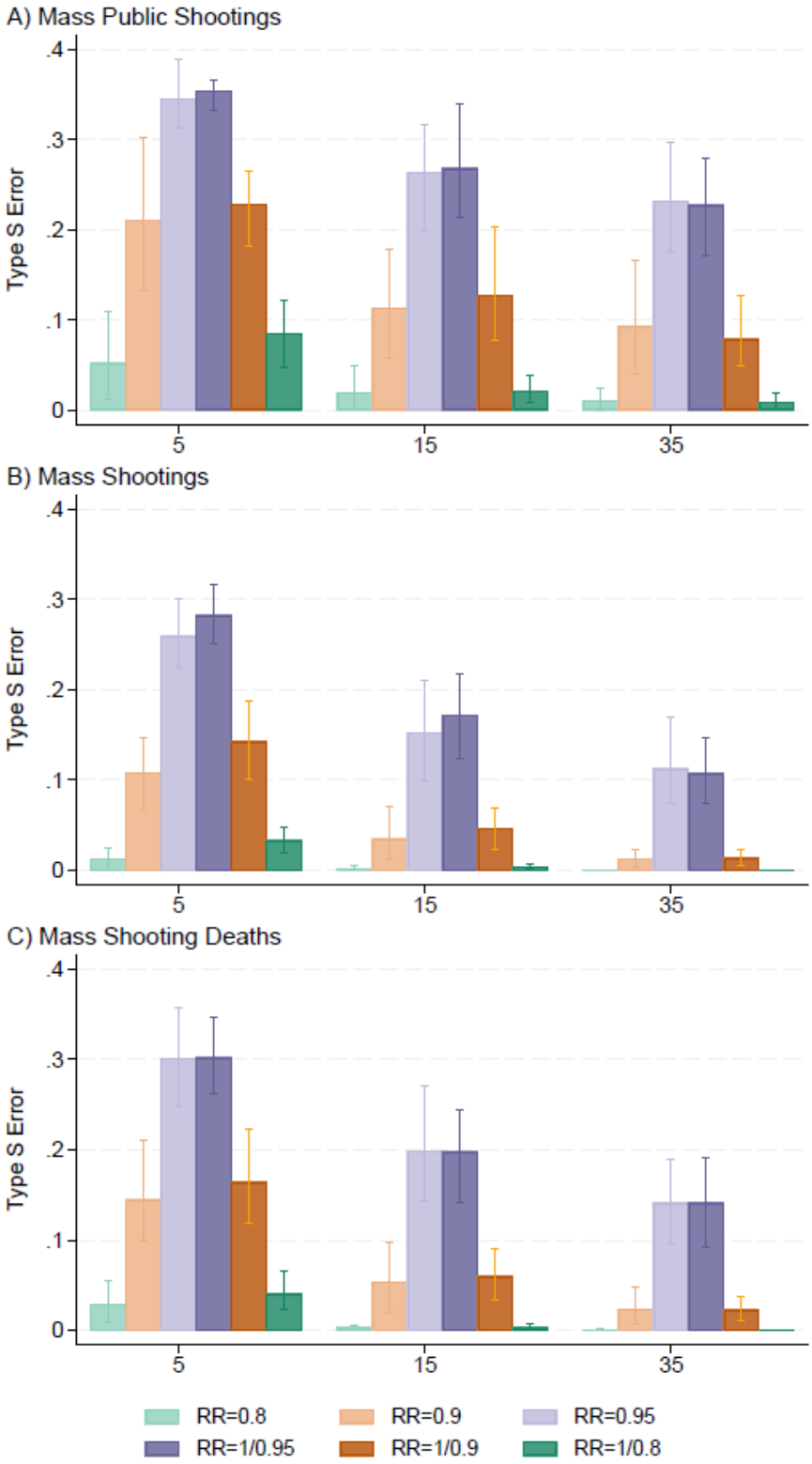
NOTES: Bar chart presents the average correct rejection rate across all model specifications for a given simulation condition, and the error bar shows the range from lowest to highest correct rejection rate on that margin.

Figure S7. Type S Error Rates for Linear Rate Models



NOTES: Bar chart presents average Type S error rate across all model specification variants for a given simulation condition, and the error bar shows the range from lowest to highest Type S error rate on that margin.

Figure S8. Type S Error Rates for Linear Count Models



NOTES: Bar chart presents average Type S error rate across all model specification variants for a given simulation condition, and the error bar shows the range from lowest to highest Type S error rate on that margin.

Supplemental Materials References

Blau, Benjamin M., Devon H. Gorry, and Chip Wade. "Guns, Laws, and Public Shootings in the United States." *Applied Economics* 48, no. 49 (2016).

Duwe, Grant, Tomislav Kovandzic, and Carlisle E. Moody. "The Impact of Right-to-Carry Concealed Firearm Laws on Mass Public Shootings." *Homicide Studies* 6, no. 4 (2002).

Fridel, Emma E. "Comparing the Impact of Household Gun Ownership and Concealed Carry Legislation on the Frequency of Mass Shootings and Firearms Homicide." *Justice Quarterly* 38, no. 5 (2021a).

Fridel, Emma E. "The Futility of Shooting Down Strawmen: A Response to Kleck (2020)." *Justice Quarterly* 38, no. 5 (2021b).

Gius, Mark. "The Impact of State and Federal Assault Weapons Bans on Public Mass Shootings." *Applied Economics Letters* 22, no. 4 (2015).

Gius, Mark. "Relationship Between Red Flag Laws and Mass Shootings." *Atlantic Economic Journal* 52, no. 1 (2024).

Klarevas, Louis, Andrew Conner, and David Hemenway. "The Effect of Large-Capacity Magazine Bans on High-Fatality Mass Shootings, 1990–2017." *American Journal of Public Health* 109, no. 12 (2019).

Luca, Michael, Deepak Malhotra, and Christopher Poliquin. "The Impact of Mass Shootings on Gun Policy." *Journal of Public Economics* 181 (2020).

Siegel, Michael, Max Goder-Reiser, Grant Duwe, Michael Rocque, James Alan Fox, and Emma E Fridel. 2020. "The Relation between State Gun Laws and the Incidence and Severity of Mass Public Shootings in the United States, 1976–2018." *Law and human behavior* 44 (5):347.

Smart, Rosanna, Andrew R Morral, James P Murphy, Rupa Jose, Amanda Charbonneau, and Terry L Schell. 2026. *The Science of Gun Policy: A Critical Synthesis of Research Evidence on the Effects of Gun Policies in the United States, Fifth Edition*. RAND Corporation.

Smart, Rosanna, Andrew R Morral, and Terry L Schell. 2021. *The Magnitude and Sources of Disagreement among Gun Policy Experts*. RAND Corporation.

Webster, Daniel W., Alexander D. McCourt, Cassandra K. Crifasi, Marisa D. Booty, and Elizabeth A. Stuart. "Evidence Concerning the Regulation of Firearms Design, Sale, and Carrying on Fatal Mass Shootings in the United States." *Criminology and Public Policy* 19, no. 1 (2020).