



AIR OPPORTUNITY FUND

Systematic Review and Meta-Analysis Findings

Effectiveness of Reform Efforts
Focused on Police Use of Force
and Complaints

Abigail Jefferys, Bryant Crubaug, Candace Hester,
Jenna Collins, John Kochanek, Megan Iacocca,
Naomi Rockefeller, Reg Ledesma, Samuel DeWitt

MAY 2025

Contents

Executive Summary.....	1
Key Takeaways.....	3
Recommendations.....	4
Introduction.....	5
Background on Use-of-Force Interventions, Training, and Programs.....	7
Aim	11
Methods.....	12
Results.....	17
Summary of Findings and Recommendations for Implementation.....	33
References	38

Tables

Table 1. Study Characteristics, by Inclusion Status	17
Table 2. Summary of Main Meta-Analytic Results	24
Table 3. Summary of Moderator Analyses	27
Table 4. Publication Bias: Egger’s Regression Using CHE Model, Accounting for Dependency (CHE Sandwich).....	32

Figures

Figure 1. Chronological Overview of the Systematic Review and Meta- Analysis Process	12
Figure 2. PRISMA Diagram, Included Studies (k = 18)	19
Figure 3. Location and Name of Specified Law Enforcement Agencies	20
Figure 4. Effect Size Distribution by General and Specific Outcome Domains (k = 18, m = 106)	23
Figure 5. Orchard Plot of Overall Effects, by Outcome Domain	25
Figure 6. Orchard Plot of Effects, Moderated by Study Region	28
Figure 7. Orchard Plot of Effects, Moderated by Study Design	29
Figure 8. Orchard Plot of Effects, Moderated by Choice to Participate in the Study	30
Figure 10. Funnel Plot Assessing Publication Bias	32

Executive Summary

Police violence in the United States remains a concern, despite recent calls for reform. In 2024, more than 1,200 people were killed by police. Nearly all of these deaths (97%) were caused by firearms, with most other deaths caused by conducted energy devices (CEDs), physical force, and police vehicles (Sinyangewe et al., 2025).

While policies and training programs are in place at local and national levels to mitigate the prevalence of police violence and impose restrictions on the use of force, there is scant data supporting the adoption of specific policies or programs or offering insight into effective methods for reducing use of force. To date, researchers conducting systematic reviews in this domain have reported mixed results or have focused narrowly on a single type of reform (e.g., body-worn cameras [BWCs], de-escalation training).

We conducted a rigorous systematic review and meta-analysis to assess the impact of *any* type of policy, reform, program, or training on the prevalence of police use of force and citizen complaints in the United States (Pigott & Polanin, 2020). We used a correlated and hierarchical effects (CHE) model with robust variance estimation (RVE) to include *all* reported uses of force and complaint outcomes in a study (Pustejovsky & Tipton, 2022).

Our systematic review efforts yielded a wide range of eligible studies. We screened 4,510 records at the abstract level and 445 at the full text level and included 18 studies (reported across 22 publications) in the meta-analyses. Included studies evaluated policies, training, and legal rulings related to the prevalence of officer use of force and citizen complaints. We conducted quality appraisal using the Maryland Scientific Methods Scale, only retaining studies with a score of three or higher (Farrington et al., 2003). Half of the 18 included studies were randomized controlled trials (RCTs; $k = 9$, 50%) and half used a quasi-experimental design (QED; $k = 9$, 50%). Nearly all studies were peer reviewed ($k = 17$, 94%), and all studies were conducted (or secondary data were collected) between 2000 and 2021. Most studies were published within the last five years ($k = 11$, 61%).

We identified 27 unique intervention arms reported across the 18 studies (four studies evaluated more than one reform program). Over half of the studies evaluated the effects of BWCs ($k = 10$, 55%). Others tested various training programs ($k = 5$, 27%), the introduction or prohibition of intermediate force options (e.g., CEDs; $k = 1$, 6%), or minimum educational attainment hiring requirements ($k = 1$, 6%); or evaluated the effects of Department of Justice oversight or circuit-specific rulings against business-as-usual protocol, policies, or procedures ($k = 1$, 6%). Units of assignment to reform or control

conditions varied, but most studies ($k = 8$, 44%) assigned condition at the officer level or by officer–shift combination ($k = 3$, 17%). Sample sizes ranged from 54 to 7,785. Participation in reform efforts was mandatory in 10 studies (56%), unclear or not reported in six studies ($k = 6$, 33%), and voluntary in two studies (11%). Demographic characteristics of participating officers and/or law enforcement agencies were infrequently reported; only three studies (17%) reported officer sex, race, or tenure.

Our meta-analytic findings revealed important treatment effects. Following reform efforts, the relative rate of use-of-force incidents was estimated to be 11% lower for treatment groups than comparison groups, though this effect was not statistically significant (95% CI: -25%, + 5%). Regarding the effect on citizen complaints, the pooled estimate indicated that complaints were reduced by 18% for treatment groups versus comparison groups, though again this was not statistically significant (95% CI: -39%, +12%). The within- and between-study heterogeneity of the model indicates that one standard deviation is roughly +/- 53%. The 95% prediction interval (PI) for each outcome—i.e., the range of possible effects likely to be observed in future similar studies, or in existing studies that were not included in this synthesis—for use-of-force outcomes ranged from a mean decrease in incidents of up to 35% to an increase of up to 22%. For complaint outcomes, the 95% PI ranged from a reduction of 43% to an increase of 22%.

We conducted several moderator analyses to explore whether effects varied by specific intervention components (e.g., type of reform, location) or study-level components (e.g., design). Only participation (i.e., the choice to engage in a study or reform effort) was statistically significant ($F = 2.13$, $p = 0.02$). For studies in which officers or departments *voluntarily participated* in reform efforts, observed effects indicated a statistically significant reduction in use of force and complaints ($k = 2$, mean change -31%, 95% CI: -47%, -0.1%). Observed effects were not statistically significant in studies in which officers or departments were *mandated to participate* in reform efforts ($k = 10$; mean change -0.1%, 95% CI: -27%, +16%) or where *participation was unclear or not reported* ($k = 6$; mean change -14%, 95% CI: -30%, +4%). Notably, these three estimates do not differ statistically from one another, though this could potentially be attributed to a lack of power, coupled with low precision in point estimates. Both studies involving voluntary participation were RCTs. This suggests that bias arising from the study design is unlikely, and that observed effects may be partially driven by officers' motivation to engage in reform efforts or to more consciously consider their choice to use force. These are critical considerations for future reform efforts. It is impractical to administer reform efforts only among motivated officers. Therefore, program implementors should account for baseline officer motivation and engagement, measure or encourage compliance, and provide supports to increase motivation and compliance to yield more favorable outcomes.

Key Takeaways

1. Implementation of some type of reform or intervention was associated with a non-statistically significant trend of lower rates of use of force (11%) and citizen complaints (18%) among exposed officers.

- a. Prediction intervals—which can tell us the range of plausible effects for new reform efforts—ranged from a 43% reduction to a 22% increase in use-of-force or complaint outcomes. This finding indicates that, while the programs will more often reduce force, we cannot rule out the possibility of negative effects.
- b. Results from our moderation analyses indicated that only the voluntary nature of reform was associated with significantly lower use-of-force outcomes. Both studies driving this effect were RCTs, suggesting that officer motivation is an important component to address in use-of-force reform.

2. Fidelity of implementation is rarely assessed, providing scant insight into why some programs are effective and others are not.

- a. Few studies assessed fidelity of policy implementation (for BWCs, see Huff et al., 2020) and only one study used citizen surveys to assess intermediate training outcomes, such as how often officers used techniques from their trainings (White et al., 2023).
- b. The lack of implementation evaluations prevented us from fully contextualizing why the same policies adopted in different places appear to have substantively different effects.

3. With the exception of the evidence base on BWCs, RCTs and rigorous QEDs are rare.

- a. Within certain reform effort categories—such as crisis intervention training (CIT) and implicit bias awareness training (IBAT)—attitudinal outcomes were more commonly reported than behavioral outcomes.
- b. Rapid diffusion and implementation of certain reform efforts—such as de-escalation (Engel et al., 2020) and CIT (Seo et al., 2021)—make RCTs and high-quality QEDs impossible or impractical.

Recommendations

- **Use Rapid-Response Funding Streams for Police–Researcher Partnerships**
 - Multiple federal, state, and local agencies used rapid-response funding during the Covid-19 crisis to swiftly understand effective responses to the pandemic.
 - Using a similar approach to address pressing issues like officer use of force would enable researchers to generate critical knowledge about the impacts of training and interventions—including potential unintended consequences—before broad adoption occurs.
- **Pair Impact Evaluations With Implementation Evaluations More Frequently**
 - Implementation evaluations can better contextualize why the same (or very similar) reform efforts work in some places and not others. While implementation evaluations are not simple extensions to RCTs or high-quality QEDs, understanding **how** reform works is sometimes as important as understanding **why** it works.
 - With adequate funding, researchers could conduct implementation evaluations more frequently and gain a better understanding of which components of reform implementation account for differences in efficacy across different contexts, including compliance with treatment protocols.
- **Use Emerging Technology to Fill Gaps in Prior Work and Pursue Novel Outcome Measures**
 - Understanding attitudinal or perceptual changes resulting from training intended to transform officer beliefs (e.g., CIT, IBAT) is paramount to verifying a theory of change but represents only the first step toward establishing program effectiveness.
 - Emerging technologies, such as automated processing of BWC data, allow for more accessible and complete measurement of behavioral outcomes. New technologies may also allow researchers to explore additional outcomes we have not previously explored or thought to consider due to impracticalities in obtaining the necessary data (e.g., officer/citizen demeanor during interactions).

Introduction

Over the last 15 years, scholars have dedicated significant attention to police use of force, examining its varied antecedents or correlates (Gaston et al., 2021; Klinger et al., 2016; Paoline & Terrill, 2007) and exploring potential interventions (Worden et al., 2013) and training programs (Goh, 2021) to reduce the severity and prevalence of such events. Among the more prominent and widely adopted interventions are body-worn cameras (BWCs; Lum et al., 2020) and early intervention systems (EIS; Gullion & King, 2020). Among training programs, both de-escalation training (Engel et al., 2022) and crisis intervention training (CIT; Seo et al., 2021) have been broadly adopted across U.S. police departments. Other programs—such as implicit bias awareness training (IBAT)—are also prevalent, but most research concerning their effect is limited to officer attitudes (e.g., Kochel & Nouri, 2024) rather than actual behavior during interactions with citizens.¹

Despite increased scholarly attention, high-quality experimental and quasi-experimental research is lacking in all but one of these areas (BWCs), as recent systematic reviews and meta-analyses of EIS, de-escalation training, and CIT lament (Engel et al., 2020; Gullion & King, 2020; Seo et al., 2021). A small sample of nascent studies (e.g., Engel et al., 2022; Goh, 2021; White et al., 2023a) have begun to fill this critical gap and expand on the knowledge synthesized by Engel et al. (2020) and Lum et al. (2020). However, recent evidence has yet to be distilled into a more interpretable form for policymakers, police departments, and the general public.

Despite increased scholarly attention, high-quality experimental and quasi-experimental research remains is lacking in all but one of these areas (—BWCs), —as recent systematic reviews and meta-analyses of EIS, de-escalation training, and CIT lament (Engel et al., 2020; Gullion & King, 2020; Seo et al., 2021).

To that end, the American Institutes for Research (AIR) Opportunity Fund provided funding for a partnership between AIR and its Methods of Synthesis and Integration Center (MOSAIC) to develop an ongoing and publicly accessible evidence hub on officer use of force, as well as other prominent issues within the criminal legal system. The current project is intended to be the first step toward a *living meta-analysis* of the interventions, training programs, and individual, agency, and community characteristics that affect the magnitude and prevalence of police officer use-of-force behaviors. We are initially

¹ A very recent exception to this is an evaluation of an IBAT program with the New York Police Department conducted by Worden et al. (2024), which evaluated both perceptual and behavioral outcomes. We will include this paper in the next update of this meta-analysis.

prioritizing a focused update of recent meta-analyses (Engel et al., 2020; Lum et al., 2020), which will begin with experimental and high-quality quasi-experimental studies evaluating the impact of interventions and trainings on officer use-of-force behavior in the field, as opposed to changes in attitudes or behaviors in simulations.

This report is organized as follows. In the next section, we provide background information on common use-of-force interventions and training programs. We follow this with an explanation of the specific aims of this study, the methods we used for our meta-analysis, a review of our results, and a discussion of the limitations of our approach. Our report concludes with a discussion of the implications of our results and recommendations for future research.

Background on Use-of-Force Interventions, Training, and Programs

As the breadth of interventions and training programs we considered is quite wide, a comprehensive overview of each specific intervention or program is beyond the scope of this report. Instead, we focus on two primary categories that the interventions and training programs within our current review fall into: BWCs and de-escalation training. After providing an overview of these primary intervention categories, we briefly discuss the variety of interventions examined in the remaining studies included in the meta-analysis.

Body-Worn Cameras. The ubiquity of BWCs across modern U.S. police departments is difficult to overstate. According to the most recent Bureau of Justice Statistics report, using estimates from 2016 on the prevalence of BWCs (Hyland, 2018), almost half (47%) of the over 15,000 general purpose police agencies in the United States have adopted BWCs in their department at some level. The largest police agencies—those with 500 or more full-time sworn officers—had an 80% BWC adoption rate (Hyland, 2018). It is reasonable to assume that these numbers have increased substantially over the near decade since Hyland’s (2018) report, though we lack more recent national estimates to confirm this assumption.

The ubiquity of BWCs across modern U.S. police departments is difficult to overstate. According to the most recent Bureau of Justice Statistics report, using estimates from 2016 on the prevalence of BWCs (Hyland, 2018), almost half (47%) of the over 15,000 general purpose police agencies in the United States have adopted BWCs in their department at some level.

The widespread adoption of BWCs has prompted a corresponding increase in scholarly attention, with research on their effects—and particularly on police use of force—expanding substantially between 2015 and 2018. This included several efforts to synthesize the emerging body of evidence (Maskaly et al., 2017). While early meta-analyses and reviews indicated that BWCs were associated with reductions in officer use of force (Maskaly et al., 2017), a more comprehensive review later concluded that such findings were premature (Lum et al., 2020). While initial studies suggested promising outcomes, subsequent research produced mixed results. Possible explanations for this include pre-existing differences between early and later adopting agencies, which may moderate the potential effects of BWC adoption (Gaub & White, 2020); and officers’ discretion about when to activate their BWC (Huff et al., 2025).

De-escalation Training. Demand for police de-escalation policies and training was ignited by multiple high-profile killings of unarmed Black men at the hands of police officers in the

last 15 years, which sparked nationwide protests (Engel et al., 2020). The general purpose of de-escalation training is to introduce and reinforce officers' use of techniques that can stabilize tense encounters with citizens, reducing the immediacy of threats to bodily harm such that "more time, options, and/or resources become available" to the officer (Dunham et al., 2021, p. 401). Typical elements of de-escalation include maintaining physical distance from citizens during a tense encounter and using verbal techniques—active listening, asking questions, and building rapport—to defuse what might otherwise turn into a situation requiring the officer to use force (Todak & James, 2018).

However, the proliferation of de-escalation training programs in the field of policing, complete with increasingly realistic simulator programs (James et al., 2018), has substantially outpaced rigorous evidence of its practical impact on use-of-force behavior (Engel et al., 2020). Further, despite many large police departments requiring officers to de-escalate situations, how departments define and apply de-escalation principles is far from uniform (McKesson et al., 2016). Arguments against de-escalation training—suggesting that such policies may *increase* injury risk to officers and citizens alike—have also lacked the necessary empirical evidence to properly evaluate these claims (Engel et al., 2020).

The proliferation of de-escalation training programs in the field of policing, complete with increasingly realistic simulator programs (James et al., 2018), has substantially outpaced rigorous evidence of its practical impact on use-of-force behavior (Engel et al., 2020). Further, despite many large police departments requiring officers to de-escalate situations, how departments define and apply de-escalation principles is far from uniform (McKesson et al., 2016).

As Engel et al. (2020) observed at the time of their study, the empirical evidence on de-escalation was scarce. Indeed, in their meta-analysis of de-escalation program evaluations, featuring studies from over four decades, not a single study evaluated a police de-escalation program. Since making these observations, Engel et al. (2022) and others (Goh, 2021; White et al., 2023) have conducted rigorous experimental or quasi-experimental evaluations of de-escalation programs. Despite finding no effect at the officer level, Goh (2021) identified a significant 40% decrease in use-of-force events at the department level in her synthetic control analysis of the Camden, New Jersey police department. While such results may be discounted because the police department in Camden, New Jersey stands out as a unique context—the city recently dissolved and replaced the police department—subsequent randomized controlled trials (RCTs) by Engel et al. (2022) and White et al. (2023a) reported similar results. In two separate analyses, both sets of authors reported that de-escalation training resulted in a roughly 30% to 50%

reduction in officer use of force, without any negative consequences for officer safety. Indeed, Engel et al. (2022) reported that officer injuries *decreased* by over 30%.

Additional Interventions and Training Programs. The evidence base on officer behavior is limited for several of the interventions and trainings we describe below, at least with respect to more rigorous RCT and quasi-experimental design (QED) evaluations. For certain interventions—i.e., Department of Justice (DOJ) consent decrees and officer education—there are practical limitations preventing treatment randomization. For other interventions and trainings—such as CIT or procedural justice training (PJT)—the challenge is that adoption outpaces researchers’ ability to conduct timely evaluations to verify that these reforms achieve their intended results.

- **Crisis Intervention Training**

- CIT aims to (a) improve officer perceptions of individuals with mental illnesses, (b) teach de-escalation techniques to reduce use of force, and (c) promote diversion to mental health services (Watson et al., 2011).
- Very limited QED or experimental research exists, though a recent nationwide study by Fagan et al. (2020) suggests that CIT can reduce use of force, consistent with prior nonexperimental evidence (Morabito et al., 2017).

- **Department of Justice Consent Decrees, Investigations, and Monitoring**

- Consent decrees are a selectively employed tool of the federal government to reform police departments with a “pattern and practice” of violating citizens’ civil rights (Jiao, 2021).
- Due to the highly selective application of a DOJ consent decree, RCTs are not possible. A recent QED using the Fatal Encounters database reported that more intense federal oversight (DOJ investigations or oversight) was associated with a 30% reduction in officer-involved fatal shootings (Goh, 2020).

- **Intermediate Force Options (TASERS)**

- The introduction or removal of intermediate use-of-force options—such as conducted energy devices (CEDs)²—has received some empirical attention (Alpert & Dunham, 2010; Terrill & Paoline, 2011).
- Evidence from a recent QED analysis of a federal court decision substantially limiting the use of CEDs as an intermediate use-of-force option suggests that

² CEDs are often referred to as “Tasers,” though TASER is technically a brand of CED and not the name of the device. CEDs are typically conceptualized as less lethal force options, relative to service weapons, because they can still result in serious injury or death.

reducing officers' intermediate use-of-force options increased the number of officer-involved shootings by 2.90 per month in states affected by the court decision (Boehme et al., 2022).

- **Officer Education Requirements**

- Most of the evidence base concerning the relationship between officer education and use of force is correlational (Paoline & Terrill, 2007; Sherman & Blumberg, 1981). It remains unclear if more capable officers are choosing to pursue higher education to *signal* their pre-existing higher capability, or if pursuing higher education *makes* an officer more capable.
- A recent study (Johnson et al., 2022) leveraged a QED using data from 235 cities' police departments with different educational requirements for patrol officers. It demonstrated that departments with higher requirements reported 110% lower rates of police-related fatalities among Black citizens, compared with departments with lower educational requirements.

- **Procedural Justice Training**

- PJT has been around for some time (Tyler, 1988) but high-quality evidence on its impact is rare (Nagin & Telep, 2017). The outcome of most PJT studies is citizen perceptions of how fairly an officer treated them during an encounter. Use of force is a natural extension of the expectation that PJT should increase the civility of police–citizen interactions and therefore result in fewer circumstances where use of force is necessary.
- A recent RCT of PJT within the Chicago Police Department supports this expectation. Compared with officers who had yet to undergo training, treated officers exhibited use-of-force rates that were between 21% and 43% lower (Wood et al., 2020).

Aim

In funding this systematic review and meta-analysis of extant, high-quality RCTs and QEDs pertaining to officer use of force, the AIR Opportunity Fund aims to leverage AIR's expertise in evidence synthesis through MOSAIC, coupled with its investments in evidence-based and practice-informed solutions to address root causes of harm and promote safer communities. In pursuit of that goal, we systematically searched for recent, high-quality research evidence on BWCs, de-escalation training, and multiple other types of interventions and training programs to facilitate its integration into an ongoing, public-facing evidence synthesis.

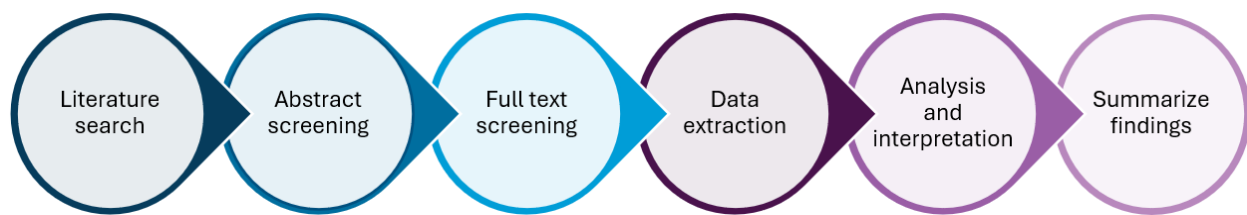
This meta-analysis, and its corresponding web tools and documentation, provide an initial foundation for an ongoing evidence hub on evidence-based practices to reduce inappropriate officer use of force. The current work will serve as the cornerstone for future updates and expansions to its original scope. The Opportunity Fund, in partnership with AIR and MOSAIC, aims to provide a publicly available, continually updated, and increasingly expansive review of the research evidence on inappropriate use of force and what can be done to curtail it for the betterment of communities and their police officers. Too often, interventions and training programs are adopted well before rigorous evaluations are accessible to practitioners, due to delays in the publication process or evidence being placed behind expensive paywalls. Reflecting the broader shift toward rapid-response and open science practices in peer-reviewed publishing,³ our ongoing meta-analysis and dynamic web tools will provide a routinely updated and publicly available evidence hub. This approach will minimize dissemination delays and promote stronger collaboration among academics, policymakers, and practitioners.

³ Examples of rapid-response journal policies include the availability of rapid communications submissions at the Police Practice and Research journal, which is meant to maintain rigorous peer-review standards while shortening the time between submission and online publication of a manuscript to as little as one week.
<https://www.tandfonline.com/action/authorSubmission?show=instructions&journalCode=gppr20#preparing-your-paper>

Methods

We followed best practices in systematic review and evidence synthesis search, screening, coding, and analysis processes (Pigott & Polanin, 2020). Figure 1 provides a high-level summary of the methodological steps involved in conducting a systematic review. The individual sections below provide overviews of the criteria used and decisions made during each step of the systematic review.

Figure 1. Chronological Overview of the Systematic Review and Meta-Analysis Process



Study Inclusion/Exclusion Criteria

We included studies evaluating any type of policy, procedure, training program, or legal oversight, provided the aim was to affect police use-of-force behavior. We further limited the scope so that our conceptualizations of the interventions and conclusions drawn from analyses were most generalizable to current procedures and potential solutions in American policing. The criteria for study inclusion were as follows:

- Published in 2000 or later (if the paper was a secondary analysis, the data must have been collected in 2000 or later)
- Available in the English language
- Of a scholarly nature (e.g., academic writing, dissertation, white or grey paper)
 - Did not need to be peer reviewed or published
 - Could not be from a news article, magazine article, op-ed, or other non-scholarly source
- Conducted in a U.S. setting
 - Studies conducted internationally were allowable only if U.S. data were disaggregated
- Used one of the following study designs:
 - Randomized controlled trial

- Quasi-experimental design with comparison group
- Aimed to evaluate an intervention, policy, reform, or other intervention type to affect use-of-force behavior
- Population included or was limited to law enforcement officers
 - Studies reporting citizen outcomes were eligible only if law enforcement data were disaggregated
- Reported outcomes reflected actual officer behaviors
 - Responses to simulation or vignette scenarios were excluded
 - Attitudinal or perceptual outcomes were excluded
- Eligible outcomes provided sufficient statistical information to calculate an effect size (specifically, a relative incident risk ratio or incident risk ratio)

Further details on the full inclusion and exclusion criteria and examples can be found in Supplemental File 1, Table 1.

Literature Search

We conducted a comprehensive and systematic search of the literature. A professional librarian at AIR assisted in the development of the search strategy. We used Boolean operators in the title, abstract, and main text fields to maximize results and account for variations in reporting and language among different disciplines and journals. We conducted our search in EBSCO's Academic Search Premier, Education Source, and ERIC databases; APA's PsycINFO; and eight databases within the Web of Science Core Collection. The full search strategy is detailed in Supplemental File 1.

Deduplication. Records were deduplicated in EndNote (The EndNote Team, 2013) and .csv files based on combinations of the fields “author,” “title,” and “DOI,” as available. Results where all three fields were flagged as duplicative with another record were assumed to be “true” duplicates and additional record(s) were deleted. Results where only one or two fields were classified as identical to another record were manually inspected for similarity.

Reference Harvesting. In addition to identifying potentially eligible records via the systematic database search, additional records were identified from studies screened during full text review. Given the time limitations, the team focused on conducting backward harvesting only (i.e., looking for all citations referenced in the focal paper). Forward harvesting was not conducted (i.e., looking for all other papers that referenced the focal paper). Backward harvesting was conducted based on inclusion status—if the study

was determined to be eligible after full text screening, its citations were matched against our search results to see if they were already captured; if not, they were added to our reference library and screened for inclusion—and for studies identified as meta-analyses or reviews during abstract screening.

Study Screening

We conducted screening in two phases. First, we reviewed titles and abstracts of all deduplicated search results. We then screened the full texts of publications deemed potentially eligible at the abstract screening stage. We provide both screening forms in Supplemental File 1.

Abstract Screening. We used the Systematic Review Data Repository (SRDR+; Saldanha et al., 2019), a free, web-based program, to conduct abstract screening. Six screeners contributed to abstract screening efforts. We piloted abstract screening on a random subset of 100 studies, with all six screeners participating to test the mechanics of the SRDR+ screening setup and ensure that all screeners had convergence on inclusion/exclusion status and could successfully identify and discuss disagreements. After the pilot round, two screeners independently evaluated the remaining records' titles and abstracts. We resolved disagreements via consensus; if consensus could not be reached, a third coder was consulted. We flagged any records that were missing title or abstract information or contained insufficient information to determine inclusion or exclusion status in SRDR+ and later evaluated these for relevance after locating full text citation information.

Full Text Screening. We conducted full text screening using MetaReviewer (Polanin et al., 2025), another free, web-based program. We procured the full article or report text for all articles identified as potentially relevant at the abstract screening stage and uploaded PDFs into a shared file drive for the team to access.

In some cases, we were unable to locate an English-language version of the full text, or purchasing the full text or using the Interlibrary Loan service was cost prohibitive or unavailable, respectively. In these cases, we excluded these studies ($n = 37$) as we were unable to access the full text. Like abstract screening, during the full text screening phase, two coders independently evaluated each record for eligibility. Disagreements were discussed and resolved via consensus, with a third coder weighing in if necessary.

Study Quality. All studies that were deemed eligible after full text coding underwent quality appraisal prior to data coding. We assessed study quality using the Maryland Scientific Methods Scale (MSMS; Farrington et al., 2003), which ranges from one to five (with higher scores reflecting stronger internal validity). One coder reviewed the study against MSMS

criteria and a second validated. Disagreements were resolved via conversation. Only studies with a rating of three or higher were retained for inclusion in the meta-analyses. Refer to Supplemental File 1, Table 2, for more information on the MSMS.

Data Extraction

We also conducted data extraction in MetaReviewer (Polanin et al., 2025). A coding form using a hierarchical relational structure allowed one form to be used for a single study with multiple groups, outcomes, and effect sizes, and allowed for linkage of related publications. We randomly selected three studies to be coded by all team members as a pilot exercise to ensure consistency in comprehending inclusion/exclusion criteria and navigating the MetaReviewer coding form. After the pilot phase, we equally divided assignments so that one coder extracted data and another validated the coding form responses. We resolved disagreements in weekly team meetings or ad hoc conversations, involving a third researcher if consensus could not be reached between the coder and validator. The full data extraction form can be found in Supplemental File 1.

Analysis Plan

Data Handling and Effect Size Information. The focal outcomes of use-of-force incidents and citizen complaints are not appropriately measured as continuous outcomes using a standardized mean difference (e.g., Cohen's d or Hedges' g). Instead, they are often reported as counts or rates, so we used the relative incident rate ratio (RIRR) as the effect size in all analyses. An extensive rationale for this approach is detailed in the Campbell systematic review and meta-analysis conducted by Lum and colleagues (2020). To convert our outcome data to RIRR and calculate the associated variance, we heavily relied on much of the R code and formulas developed by Lum et al. (2020), in addition to using the *metafor* package in R (Viechtbauer, 2010). We scaled all effect sizes so that a higher value corresponded to higher rates of the outcome. In cases where the authors reported outcomes in such a way that a higher value indicated an effect favoring the treatment group, we flipped the estimate by calculating its inverse (i.e., $1/\log \text{IRR}$). When available, we prioritized coding adjusted effects in favor of non-adjusted effects. We additionally captured if effects were cluster-adjusted and coded cluster size in addition to sample size, as well as the intraclass correlation coefficient (ICC) if provided.

Missing Data. We relied on information presented in each study to code reform, population, study, outcome characteristics, and effect size data. We used the first two steps of Pigott and Polanin's (2020) "infer, initiate, and impute" process in coding each study. Given the time and resource constraints, imputation of missing data was not a feasible option, but future updates to this review could implement that step. In the event a study was eligible for inclusion but not enough information was provided to calculate the

RIRR and associated variance, we conducted author queries to obtain the missing data so we could include the study in the analyses.

Meta-analytic Models. It is common for studies to report multiple effect sizes for the same construct or within the same domain in one paper and, until recently, there was not a sufficiently robust method of including dependent effect sizes within a single meta-analysis. However, modern meta-analytical techniques, such as correlated and hierarchical effects (CHE), allow for the inclusion of all eligible effect sizes within the same analysis, unlike older methods that require effect sizes to be grouped by outcome domain, timepoint, or other characteristics (Pustejovsky & Tipton, 2022). We adopted this analytic approach to maximize our analyses and include all eligible outcomes within a domain reported in a study. Using a CHE meta-analytic model is appropriate given the clustered nature of the data in many of the included studies (e.g., officers clustered by unit or geographic area). We also used robust variance estimation (RVE) to account for effect size dependencies. We fitted all analyses using the restricted maximum likelihood (REML) approach.

Exploring Sources of Heterogeneity. We explored variation in study effects using four pre-specified moderators. We selected these variables based on prior research, the likelihood that studies in our pool reported them, and their potential utility for policy considerations and guiding future research. Specifically, we assessed whether effects varied based on (a) whether participation in the reform effort was mandated or voluntary; (b) whether the study was an RCT or QED; (c) the type of primary reform evaluated—that is, a policy (e.g., must wear BWCs), completion of a training, mandated tactics, and so on; and (d) the region in which the study occurred. For moderator analyses, we used a CHE modeling approach with RVE, evaluating each of the four variables in separate models.

Addressing Additional Sources of Bias or Missing Information. We tested whether publication bias (i.e., the publication or non-publication of research findings, depending on the nature and direction of results; Higgins et al., 2024) was present, and to what extent, using methods consistent with our analytic models (i.e., accounting for correlated and hierarchical dependencies), namely Egger’s regression test with RVE (Rodgers & Pustejovsky, 2021) and visual inspection of the funnel plot for asymmetry.

Results

Included Studies

Our search yielded 4,510 unique articles after de-duplication. We screened all articles at the abstract level, during which we identified 482 as relevant and potentially eligible. We were unable to locate the full text for 37 records (8%) and therefore assessed 445 studies at the full text level. After full text screening, we deemed 24 studies (reported across 30 publications) eligible for inclusion. However, we only quantitatively synthesized 18 studies (reported in 22 publications) in our meta-analyses due to a lack of sufficient information to code an eligible effect size and/or non-response from author queries to obtain needed statistical information. Table 1 displays characteristics for both the 24 eligible studies and the 18 studies we quantitatively synthesized. The PRISMA diagram (Page et al., 2021) is provided in Figure 2. Additional characteristics for all studies—including the initially eligible 24 studies—are provided in Tables 2–4 in the Supplemental File. The list of studies included in syntheses and their related publications are also available in the Supplemental File.

Table 1. Study Characteristics, by Inclusion Status

Characteristic Level	All Initially Eligible Studies (k=24)	Studies in Meta-Analyses (k=18)
	k (%)	k (%)
Design		
<i>RCT</i>	11 (45.8)	9 (50)
<i>QED</i>	12 (50)	8(44.4)
<i>Mixed methods</i>	1 (4.2)	1 (5.6)
U.S. Region		
<i>Midwest</i>	4 (16.7)	3 (16.7)
<i>National or across regions</i>	4 (16.7)	3 (16.7)
<i>Northeast</i>	3 (12.5)	3 (16.7)
<i>South</i>	6 (25)	4 (22.2)
<i>West</i>	7 (29.2)	5 (27.8)
Years Conducted (range)	2000–2021	2000–2021
Peer Reviewed		
<i>Yes</i>	22 (91.7)	17 (94.4)
<i>No</i>	2 (8.3)	1 (5.6)

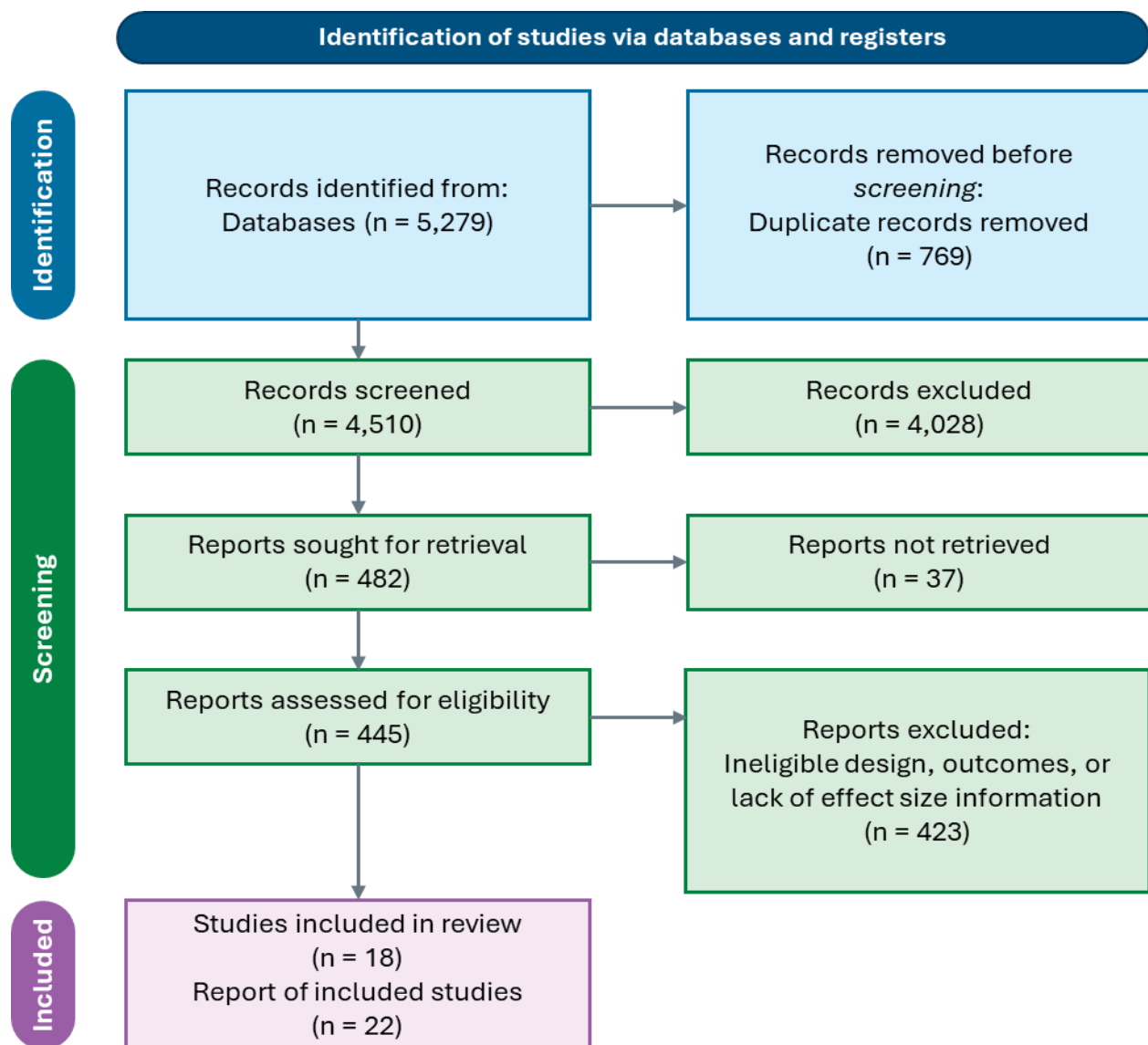
Characteristic Level	All Initially Eligible Studies (k=24)	Studies in Meta-Analyses (k=18)
Agency Type		
<i>Municipal only</i>	20 (83.3)	16 (88.9)
<i>Multiple</i>	4 (16.7)	2 (11.1)
Unit of Assignment		
<i>Agency</i>	1 (4.2)	1 (5.6)
<i>Department</i>	1 (4.2)	1 (5.6)
<i>Enforcement groups</i>	1 (4.2)	1 (5.6)
<i>Unclear</i>	1 (4.2)	1 (5.6)
<i>Officer</i>	11 (45.8)	8 (44.4)
<i>Officer–shift combination</i>	3 (12.5)	3 (16.7)
<i>Geographic area</i>	3 (12.5)	1 (5.6)
<i>Division</i>	1 (4.2)	1 (5.6)
<i>Shift</i>	1 (4.2)	1 (5.6)
<i>Other</i>	1 (4.2)	0 (0)
Sample Size (range)	54–7,785	54–7,785

Notes: k = number of studies; QED = quasi-experimental design; RCT = randomized controlled trial

General Characteristics. Of the 18 studies included in our analyses, half were RCTs (k = 9, 50%), eight were QEDs (44%), and one (6%) used a mixed-methods design that included both randomized and non-randomized components. Nearly all studies were peer reviewed (k = 17, 94%), and all studies were conducted (or secondary data were collected) between 2000 and 2021. Most studies were published within the last five years (k = 11, 61%). A total of 27 unique reform efforts were reported across the 18 studies (four studies included more than one intervention condition). We elaborate on the 27 specific reform arms in a later section of this report.

Of the 18 studies included in our analyses, half were RCTs (k = 9, 50%), eight were QEDs (44%), and one (6%) used a mixed-methods design that included both randomized and non-randomized components.

Figure 2. PRISMA Diagram, Included Studies (k = 18)



All studies were conducted in the United States, though there was regional variation in study location. The plurality of studies were conducted in law enforcement agencies in western states⁴ (k = 5, 28%), followed by southern states⁵ (k = 4, 22%). There were three studies each in the midwestern states⁶ (17%), northeastern states⁷ (17%), and nationally and/or across regions (17%). Only two studies (11%) were conducted in more than one type of agency; all other studies were conducted in a single municipal law enforcement agency (k = 16, 89%). Figure 3 shows the geographic spread of all named law enforcement

⁴ Western region states: AK, AZ, CA, CO, HI, ID, MT, NV, NM, OR, UT, WA, WY

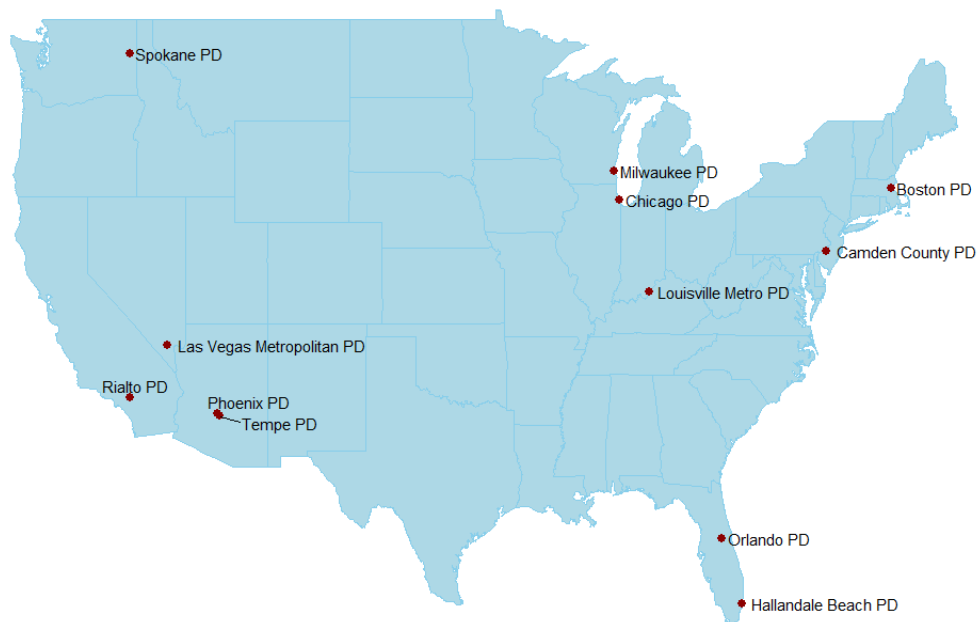
⁵ Southern region states: AL, AR, DE, FL, GA, KY, LA, MD, MS, NC, OK, SC, TN, TX, VA, WV, Washington, D.C.

⁶ Midwestern region states: IL, IN, IA, KS, MI, MN, MO, NE, ND, OH, SD, WI

⁷ Northeastern region states: CT, ME, MA, NH, NJ, NY, PA, RI, VT

agencies across the 18 studies. Some studies did not report a specific agency name due to confidentiality agreements, so not all 18 studies are represented in this figure.

Figure 3. Location and Name of Specified Law Enforcement Agencies



Across the 18 studies, units of assignment to the reform or control condition varied. The plurality of studies ($k = 8$, 44%) assigned reform efforts at the officer level, followed by officer–shift combinations ($k = 3$, 17%). One study each assigned reform efforts at the agency (6%), department (6%), enforcement group (6%), division (6%), or shift level (6%). One study grouped conditions using geographic area (6%), and it was unclear how one study (6%) assigned reform efforts. Sample sizes (in the respective unit of assignment) ranged from 54 to 7,785. Table 3 in the Supplemental File lists study characteristics in further detail, including design, location, agency type, and study aim.

Population Characteristics. Participants’ demographic information was seldom reported. Only three studies (17%) reported officer or department demographic characteristics regarding sex, gender, and race (Worden et al., 2013; Jennings et al., 2015; White et al., 2017). Across these three studies, the distribution of male officers ranged from 87% to 91.5%, and officer race reported as White ranged from 69% to 95.5%. Only two studies (Jennings et al., 2015; White et al., 2017) reported any metric of officer tenure. In White et al. (2017), the mean years of police service among intervention and control participants were 12.8 years and 14.3 years, respectively. In Jennings et al. (2015), the mean years of police service were 6.9 years and 5.9 years among intervention and control officers, respectively.

Intervention/Reform Characteristics. Twenty-seven distinct reform efforts were represented across the 18 studies included in the meta-analyses (four studies included at least two unique reform arms). Reform efforts were categorized by primary component (policy, training, counseling, or other), as well as whether the reform effort was single or multicomponent. No included studies evaluated counseling programs or “other” reform efforts that could not be classified as either policy or training alone or a combination of both. Only one study’s reform program explicitly focused on a specific use-of-force type: Boehme et al. (2022) evaluated whether the “Armstrong Rule” (i.e., restricting [TASER] use to only encounters that pose an “immediate safety risk” or “danger”) was associated with an increase in officer-involved shootings. The 17 other studies’ reform programs did not specify use-of-force types or focused on assessing broad uses of force in any policing context. See Tables 4 and 5 in the Supplemental File for brief and in-depth descriptions of interventions, respectively.

Policy. Twenty reform arms across 13 studies evaluated the effectiveness of a policy. Sixteen of these arms were single component efforts; the four multicomponent reform arms with policy as the main component also included training. Across these 13 studies, policies included BWCs (11 arms, $k = 10$), DOJ consent decree (one arm, $k = 1$), DOJ technical assistance (one arm, $k = 1$), DOJ monitoring (one arm, $k = 1$), DOJ investigation (one arm, $k = 1$), educational hiring requirements (four arms, $k = 1$), and TASER use (one arm, $k = 1$). Five reform arms in two studies evaluated policies that were developed by the department, agency, or police unit; all other policies resulted from changes in legal statutes or department-level policy on a broader scale. In 14 reform arms across seven studies, participation in the reform effort was mandated; it was unclear or not reported whether participation was mandated in four studies, and participation in policy reform efforts was voluntary in two studies. The intervention duration (i.e., the policy implementation period prior to evaluation) was infrequently reported, but across the 10 reform arms that reported it ($k = 9$), the policy delivery period ranged from 26 weeks to 78 weeks (mean = 46 weeks).

Twenty reform arms across 13 studies evaluated the effectiveness of a policy. Sixteen of these arms were single component efforts; the four multicomponent reform arms with policy as the main component also included training. Across these 13 studies, policies included BWCs (11 arms, $k = 10$), DOJ consent decree (one arm, $k = 1$), DOJ technical assistance (one arm, $k = 1$), DOJ monitoring (one arm, $k = 1$), DOJ investigation (one arm, $k = 1$), educational hiring requirements (four arms, $k = 1$), and TASER use (one arm, $k = 1$).

Training. Seven reform arms in five studies evaluated training programs (all were single component). Training programs included the Tempe De-escalation Program (one arm; White et al., 2023a); the Officer-Civilian Interaction School (one arm; Worden et al., 2013);

Integrating Communications, Assessment, and Tactics (one arm; Engel et al., 2022); de-escalation training (three arms, $k = 1$; Goh, 2021); and procedural justice training (one arm; Wood et al., 2020). Three training programs were developed by the department, agency, or unit ($k = 3$), and four training arms used existing training programs and/or curriculum for law enforcement ($k = 2$). Participation in the training program was mandatory in three studies, and unclear or not reported in the remaining two studies. The duration of the training program was reported in all five studies, ranging from two days to 24 weeks (mean = 3.6 weeks).

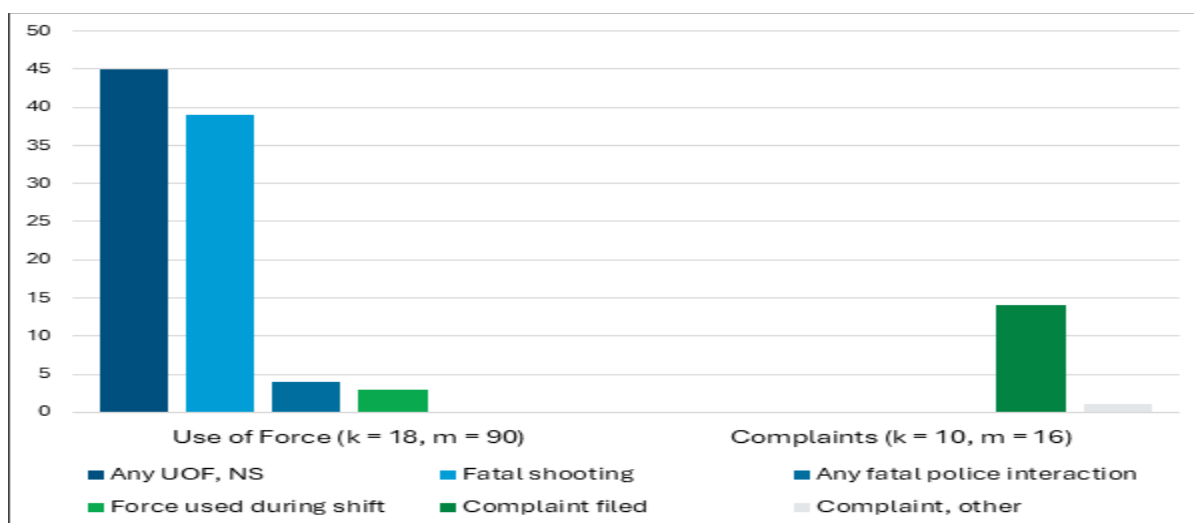
Outcome Characteristics. We categorized outcomes as belonging to one of two general domains: use of force or complaints. A total of 106 effect sizes and their variances were eligible for inclusion in analyses. The majority of effects ($m = 90$, 85%) represented use-of-force outcomes, with each of the 18 studies reporting at least one use-of-force outcome. The remaining effects represented citizen complaints ($m = 16$, 15%), with 10 studies reporting at least one complaint outcome. The number of effect sizes coded per study ranged from one to 30 (median = four). As evident in the coding form (see Supplementary File 1), we planned to code 16 distinct specific constructs within the use-of-force and complaint domains. However, very few studies reported outcome measurement characteristics in sufficient detail to allow us to granularly code specific constructs and use them in analyses.

Figure 4 depicts the distribution of the 106 effect sizes within and across general and specific domain categories. When available, we coded adjusted effects over unadjusted (i.e., statistical adjustment for baseline values). Sixty effect sizes (57%) in 15 studies were *not* adjusted, four effect sizes (4%) in one study did not have sufficient information to determine whether they were adjusted, and 42 effect sizes (40%) in six studies were author-reported as adjusted for baseline values of the outcome measure. Both across and within studies, effects were reported at various timepoints. Most effect sizes ($m = 66$, 62%, $k = 11$) reflected outcomes measured during the active reform period, 31 effects (29%) in 10 studies reflected outcomes measured six months or later after the reform period ended, eight effect sizes (8%, $k = 2$) reflected outcomes measured from less than a month to up to six months after the reform period ended, and only one effect size (1%) in a single study reflected an outcome measured immediately (<one month) after the reform period.

We also coded the method by which outcome data were collected: 42 effect sizes (40%) were drawn from administrative data, 52 (49%) were coded as an “other” collection method (in most cases, a study-specific method of assessing various sources of officer data that did not neatly fit into one of our categories), one effect size (1%) represented self-report survey or questionnaire data, two effect sizes (2%) had an unclear or not reported

method of data collection, and nine effect sizes (9%) represented observational data, all of which were sourced from the Fatal Encounters database (Burghart, 2025).

Figure 4. Effect Size Distribution by General and Specific Outcome Domains (k = 18, m = 106)



Missing Data

During data coding, we discovered that six studies did not contain sufficient statistical information to calculate a RIRR and variance for an eligible outcome. It was also not possible to make appropriate inferences with the information provided. For this reason, we conducted author queries. We sent five emails to five authors of these six studies seeking information on 10 effect sizes and sent a reminder email after two weeks of non-response. We ultimately received a single response from one author of two of the studies (total response rate of 20%). The lack of responses from the other author queries led us to drop two studies from the analysis. The other studies for which we conducted author queries had at least one usable effect size in one of the domains, so we were able to retain them.

During data cleaning, we discovered that eight studies reporting effects as Ordinary Least Squares or interrupted time series regression coefficients did not report group means (a requirement for using RIRR and variance calculation method 7c; see Supplemental File 1 and Appendix E in Lum et al. [2020]). Four of these studies had alternative measures of the same effect that we could use instead (e.g., counts, rates) or other eligible effects. Four studies did not have either of these, which meant that we had to drop four additional studies from the analyses. In all, six studies that were initially deemed eligible for inclusion were unable to quantitatively contribute to the overall model and were dropped from the analyses, as we did not impute missing data. These six studies are summarized in Tables 2–4 in the Supplemental File. In future updates of this review, we plan to conduct a second

round of author queries and explore additional methods of effect size estimation to ensure that all eligible studies and effects are included in syntheses.

Meta-Analytic Results

We performed all analyses with R Statistical Software (R Core Team, 2024) using various packages, specifically *metafor* (Viechtbauer, 2010) for multilevel meta-analyses and *clubSandwich* (Pustejovsky, 2016) to conduct robust variance estimation. We downloaded data from MetaReviewer and cleaned and re-coded, as necessary, in R using the *tidyverse* package (Wickham et al., 2019). We also created tables and plots using these resources, as well as other R packages, including *orchaRd 2.0* (Nakagawa et al., 2023), *svglite* (Wickham et al., 2015), *DiagrammeR* (Iannone & Roy, 2015), and *writexl* (Ooms, 2017). The full R script with cleaning and analysis syntax is available in Supplemental File 1.

Main Analyses

A total of 106 effect sizes from 18 studies were included in the meta-analyses. We included all effects in a single model, with the outcome domain (i.e., use of force or complaints) as a moderator to obtain a pooled effect for each outcome. Due to analyzing effects using a CHE model, we accounted for the potential correlation between and within domains, effects, and studies by specifying a variance-covariance matrix and using robust variance estimation techniques to create cluster-robust standard errors (Pustejovsky & Tipton, 2022). We did not find statistically significant effects of reform programs on reducing use-of-force behaviors or complaints, though estimates suggest effects trending toward favoring reform efforts in reducing use-of-force events and complaints. Full model details are available in Table 2. Figure 5 presents an orchard plot (Nakagawa et al., 2023), providing a visual summary of individual effect estimates and the pooled effect by outcome categories.

Table 2. Summary of Main Meta-Analytic Results

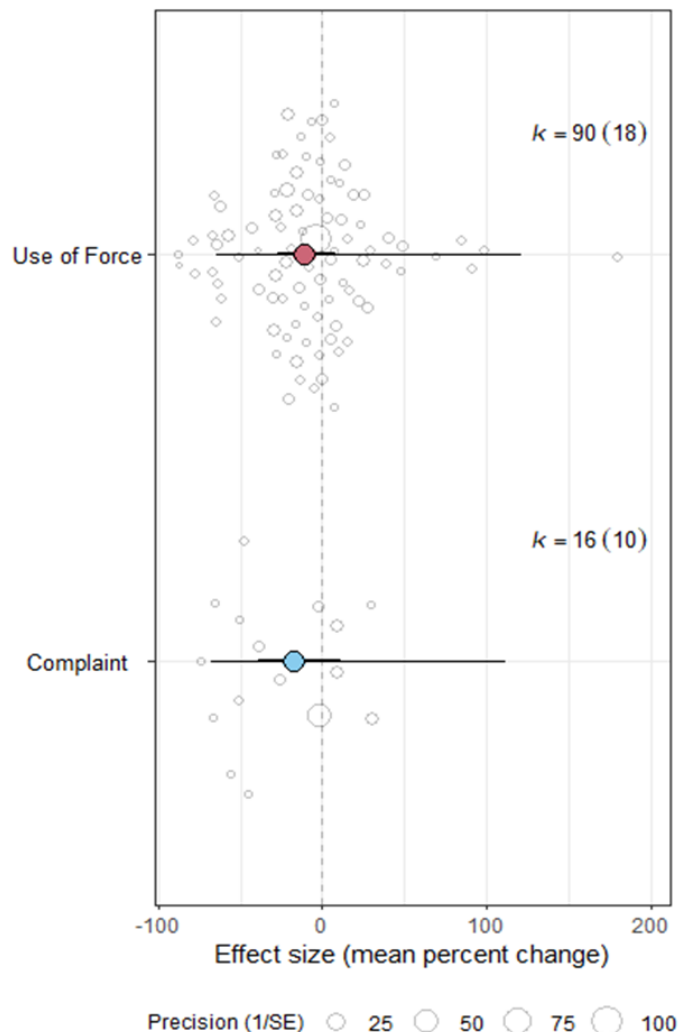
Outcome Domain	<i>k</i> (<i>m</i>)	Log RIRR, 95% CI	SE	Effect as Mean % Change (95% CI)	95% Prediction Interval	<i>t</i> -statistic <i>p</i> (<i>t</i>)	<i>T</i> (between)	<i>ω</i> (within)
Use of Force	18 (90)	-0.12 (-0.29, 0.05)	0.08	-11% (-25%, +5%)	-35%, +22%	-1.54 0.15	16%	37%
Complaints	10 (16)	-0.20 (-0.50, 0.11)	0.12	-18% (-39%, +12%)	-43%, +22%	-1.64 0.16		

Notes: CI = confidence interval; *k* = number of studies; *m* = number of effects; *p* = *p*-value; RIRR = relative incident rate ratio; SE = standard error; *t*-statistic = Student’s *t*-test result; *T* = tau; *ω* = omega

Use of Force. Though not statistically significant, the pooled effect of reform efforts on use-of-force outcomes indicated an effect favoring treatment officers (log RIRR = -0.12 [95% CI: -0.29, 0.05], $SE = 0.08$, $t = -1.54$, $p = 0.15$). That is, following reform efforts, treatment group officers had, on average, an 11% decrease in use-of-force events compared with control officers. The confidence interval indicates effects ranging from a 25% decrease in use-of-force events to a 5% increase following reform efforts. The 95% prediction interval for use-of-force outcomes—which indicates the range of values in which effects from future observations (i.e., studies) would likely fall—suggests a possible range from a 35% decrease to a 22% increase in use-of-force events following reform efforts in future studies.

Complaints. Like use-of-force outcomes, the pooled effect of reform efforts on reducing complaint outcomes was not statistically significant but trended towards an effect favoring fewer complaints among officers, squads, or divisions who received reform efforts, compared with control groups (log RIRR = -0.20 [95% CI: -0.55, 0.11], $SE = 0.12$, $t = -1.64$, $p = 0.16$). Among treatment officers, this estimate translated to an average reduction in complaints after reform efforts of 18%, with possible effects on complaints ranging from a 39% decrease to a 23% increase. The

Figure 5. Orchard Plot of Overall Effects, by Outcome Domain



Note on interpreting orchard plots. The pooled effect for each outcome is represented by the colored circle (in this figure, red and blue for use of force and complaint outcomes, respectively). Pooled effects are calculated from the individual effect sizes within each study, which are represented by the grey circles surrounding the pooled estimate. Grey circles (individual effect sizes) stacked along the y-axis are from the same study. The larger the grey circle, the more precise that individual effect. The confidence interval of the pooled effect is represented by the bolded (shorter) line, and the prediction interval by the narrower (and longer) line.
 k = number of effect sizes (number of studies)
 SE = standard error

prediction interval suggests that future studies may observe a 43% decrease to a 22% increase in complaint outcomes following reform efforts.

Evaluating Variation in Effects. The heterogeneity (or underlying variation) in effect sizes was estimated between studies (T , tau) and within studies (ω , omega; see Table 2). For the model discussed above—in which all effects were entered simultaneously, and where the outcome category (i.e., use of force or complaints) was considered a moderator—between-study variance equaled 0.15 (exponentiated from log RIRR to 16% change) and within-study variance equaled 0.32 (37% change). To translate to a simpler metric, one standard deviation in the unit of interpretation was $\pm 53\%$ (i.e., total heterogeneity), indicating a wide range of effects observed across the pool of studies. These heterogeneity results are not what we expected to find, as we anticipated that between-study variation would be larger than within-study variation due to differences in study design and participant characteristics. These statistics suggest that the underlying effect across studies is relatively consistent, but that effects within studies contain more variation. This is visually evident in Figure 5, where the spread of individual effects within a study can be observed on the same x-axis line. Another use of measuring heterogeneity is to calculate the prediction intervals—or rather, the range of effects we anticipate a new study might have. For use-of-force outcomes, the 95% PI indicates a maximum reduction of 35% to an increase of 22% following reform efforts. For complaints, the 95% PI ranges from -43% to +22% in complaints received following reform efforts.

The observed heterogeneity is potentially attributable to underlying differences in study characteristics, such as reform focus, intensity, program duration, and officer characteristics, as well as characteristics that were not measured or reported in the current pool of studies or are difficult to quantitatively capture. We explore variation in effects via moderator analyses, described in the following section.

Moderator Analyses

We explored whether effects on use of force and complaints varied by study location (U.S. region), study design (RCT, QED, mixed methods), primary reform component (policy or training), or whether participation in the reform effort was required (voluntary, mandatory, unclear or not reported). Since our main model indicated no statistically significant effects for use-of-force or complaint outcomes, and because the two estimates were not statistically different from each other, all outcomes for moderator models were entered into a single model, with the focal variable of interest replacing the outcome domain as the moderator. Summaries of all moderator analyses are presented in Table 3. Figures 6–9 visually depict effects grouped by moderator levels.

Table 3. Summary of Moderator Analyses

Moderator Level	k (m)	Average Effect Size (SE) ⁸	Average Effect Size (SE) ⁹	95% CI ⁹	F-statistic (p-value)	df
U.S. Region					0.61 (0.71)	2.6
National or multiple regions	3 (43)	-0.09 (0.08)	-8.61 (8.33)	-37, +32		
Northeast	3 (11)	-0.18 (0.24)	-16.47 (27.12)	-72, +153		
Midwest	3 (6)	0.02 (0.05)	2.00 (5.13)	-19, +28		
West	5 (35)	-0.26 (0.25)	-22.89 (28.40)	-65, +70		
South	4 (11)	-0.21 (0.10)	-18.94 (10.52)	-42, +13		
Design					1.03 (0.51)	2.39
RCT	9 (30)	-0.24 (0.21)	-21.34 (23.37)	-43, +8		
QED	8 (73)	-0.02 (0.06)	-1.98 (6.18)	-17, +15		
Mixed methods	1 (3)	-0.32 (0.14)	-27.39 (15.03)	-95, +938		
Primary Reform Component					1.68 (0.26)	6.61
Policy	13 (89)	-0.18 (0.09)	-16.47 (9.42)	-32, +4		
Training	5 (17)	-0.05 (0.09)	-4.88 (9.42)	-27, +23		
Participation					46.89 (0.02)	2.13
Voluntary	2 (6)	-0.37 (0.03)	-30.93 (2.53)	-48, -8*		
Mandatory	10 (62)	-0.08 (0.10)	-7.69 (10.52)	-27, +16		
Unclear or NR	6 (38)	-0.16 (0.07)	-14.79 (7.25)	-30, +4		

Notes: CI = confidence interval; df = degrees of freedom; k = number of studies; m = number of effects; NR = not reported; QED = quasi-experimental design; RCT = randomized controlled trial; SE = standard error; U.S. = United States.

⁸ Reported in log RIRR

⁹ Reported in mean % change

Figure 6. Orchard Plot of Effects, Moderated by Study Region

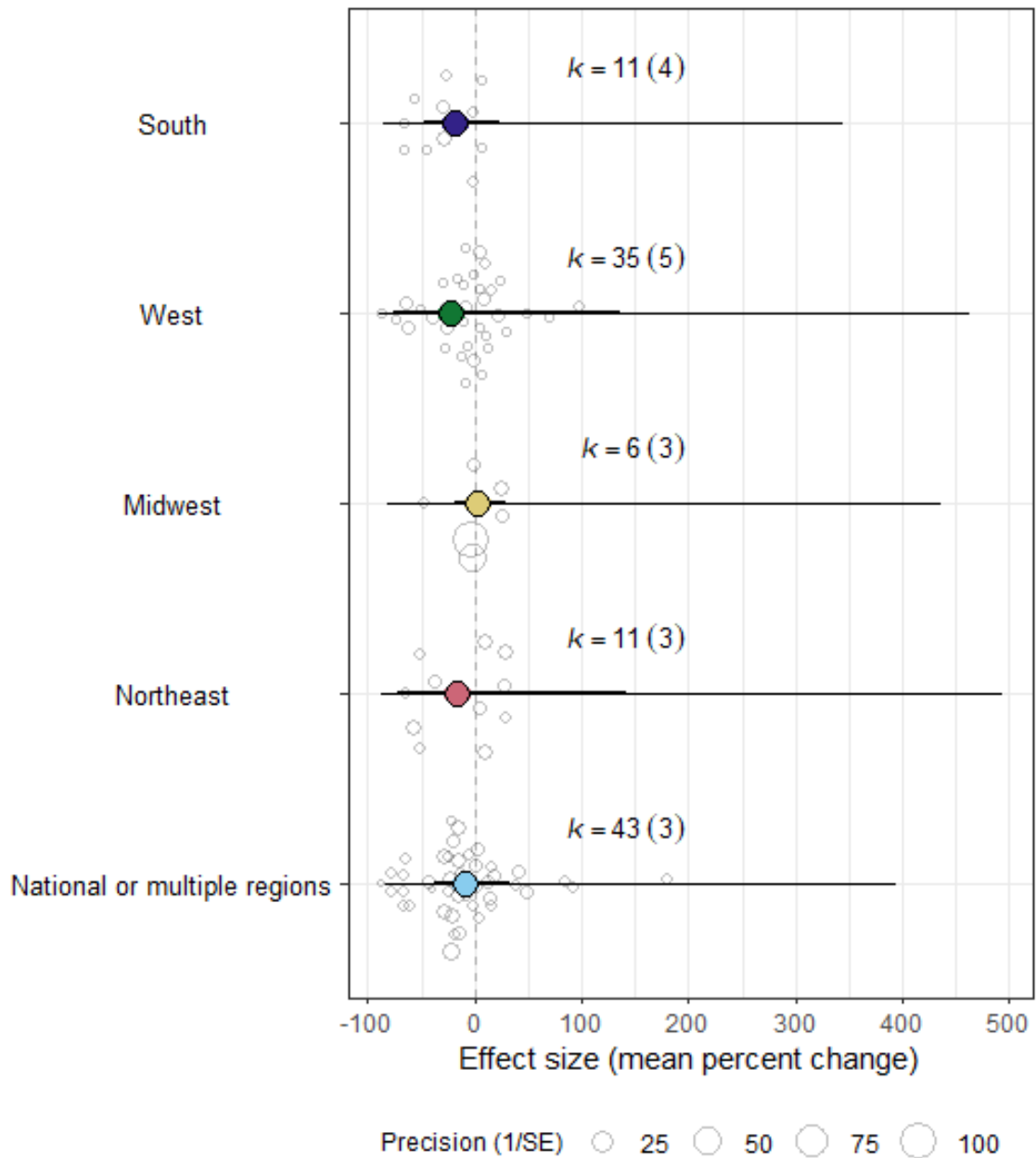
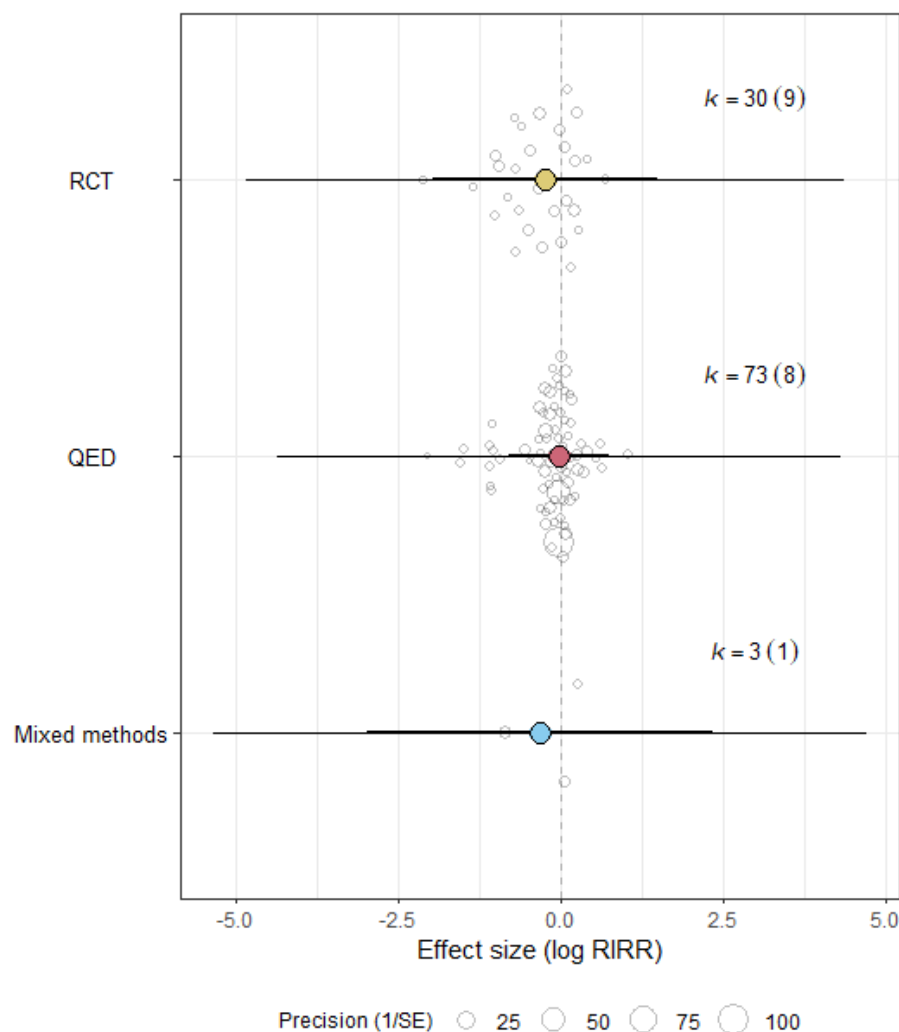


Figure 7. Orchard Plot of Effects, Moderated by Study Design



Choosing to participate in the study or reform effort was the only statistically significant moderator ($F = 46.89, p = 0.02$). Studies where officers volunteered to participate in an evaluation of a reform program had statistically significant reductions in all outcomes (mean decrease of 31% [95% CI -48%, -8*], $SE = 2.5\%$; $k = 2, m = 6$). Studies in which officer or unit participation was mandatory had a non-significant reduction in use of force and complaints (mean decrease of 8% [95% CI: -27%, +16%], $SE = 10.5\%$; $k = 10, m = 62$), as did studies in which the choice to participate was unclear or not reported (mean decrease of 15% [95% CI: -30%, +4%], $SE = 7.3\%$; $k = 6, m = 38$). However, it is important to

Choosing to participate in the study or reform effort was the only statistically significant moderator ($F = 46.89, p = 0.02$). Studies where officers volunteered to participate in an evaluation of a reform program had statistically significant reductions in all outcomes (mean decrease of 31% [95% CI -48%, -8*], $SE = 2.5\%$; $k = 2, m = 6$).

note that post-hoc comparisons revealed that these three estimates did not statistically differ from each other (all pairwise comparison $p > 0.10$). The two studies in which officer participation was voluntary (Braga et al., 2018; Jennings et al., 2015) were both RCTs that evaluated the effects of BWCs. The results of this moderator analysis should be carefully interpreted, as it is possible that volunteer officers, who might be more likely to comply with treatment protocols (e.g., activating their cameras), may benefit more from treatment than non-volunteer officers. We return to this issue in our discussion, where we elaborate further on how future research should address this possibility.

Figure 8. Orchard Plot of Effects, Moderated by Choice to Participate in the Study

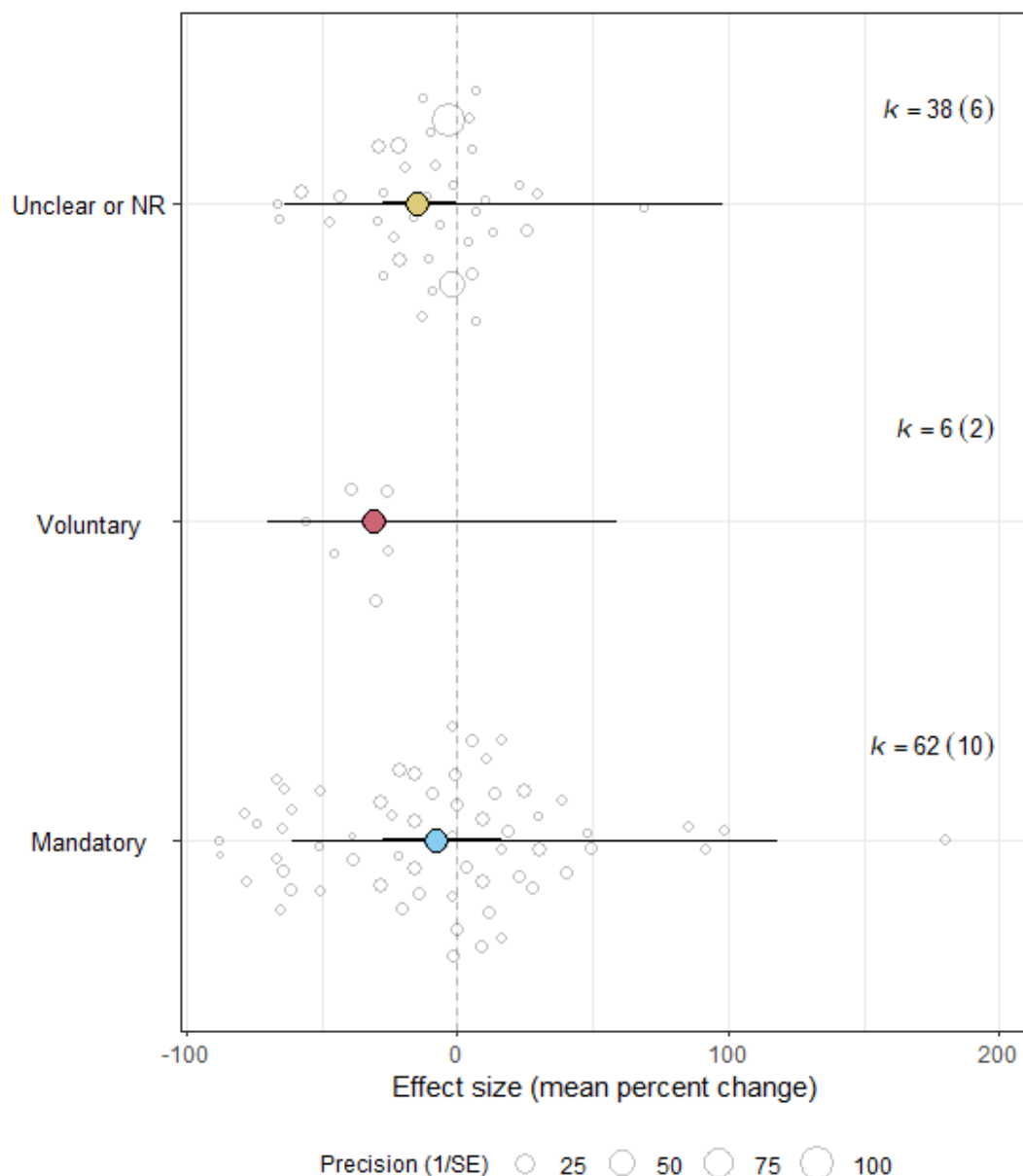
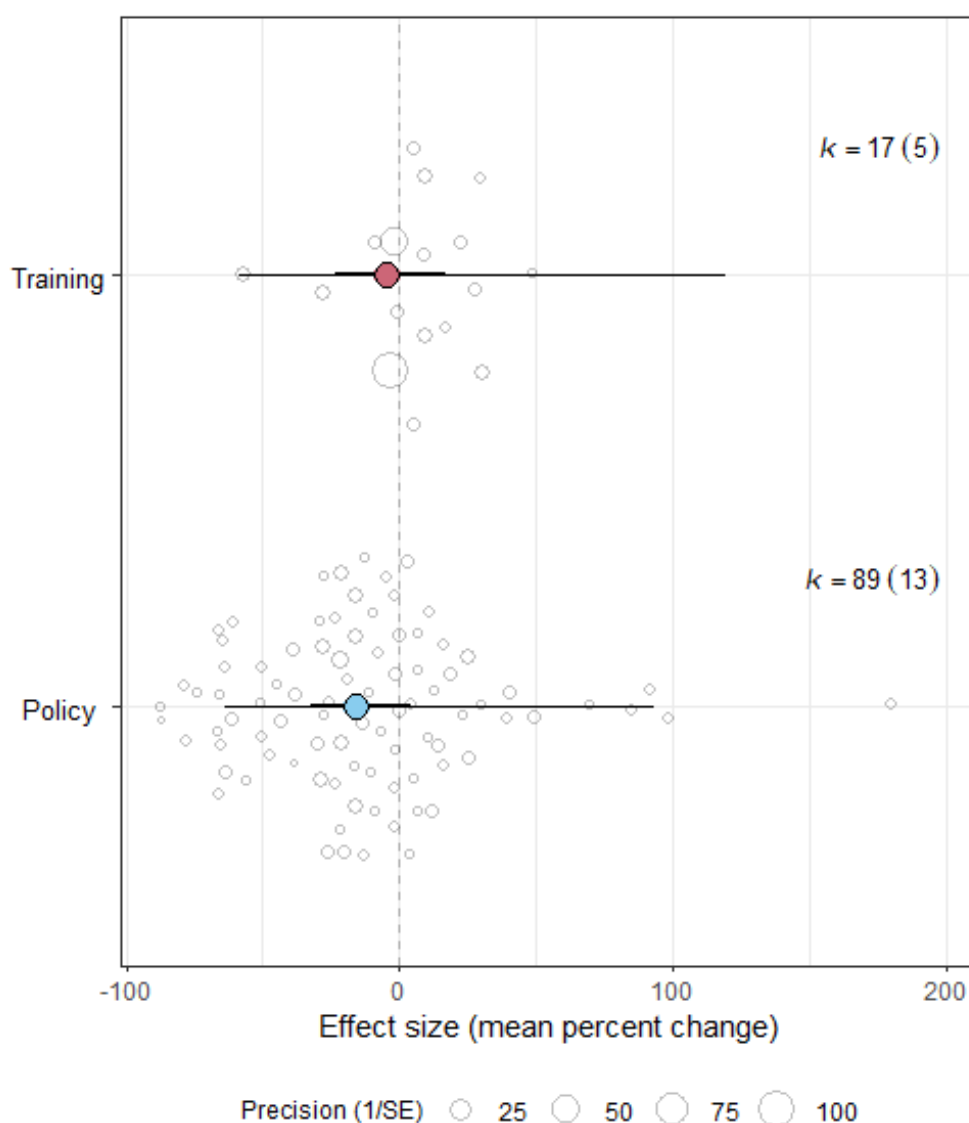


Figure 9. Orchard Plot of Effects, Moderated by Primary Reform Component



Assessing Publication Bias

As outlined earlier, our inclusion criteria allowed for both published and unpublished research to be included in this review. The only requirement was that the research was of a scholarly nature. Only one study (Peterson et al., 2018) of the 18 included in the analyses was not peer reviewed (a summary report published by the Urban Institute). However, even with the inclusion of unpublished or peer-reviewed literature, it is still best practice to assess publication bias (Polanin et al., 2016). Since we used a multilevel, non-independent model for meta-analyses, we used a CHE regression-based selective reporting detection method accounting for effect size dependency (Pustejovsky & Rodgers, 2019; Pustejovsky & Tipton, 2022). The model summary is available in Table 4 and the funnel plot is presented

in Figure 10. We did not find statistically significant evidence of selective reporting (intercept = -0.07 [95% CI: -0.25, 0.11], $SE = 0.07$, $t = -0.92$, $p = 0.39$). However, upon visual inspection of the funnel plot, it is likely that some publication bias is present due to the slight asymmetry.

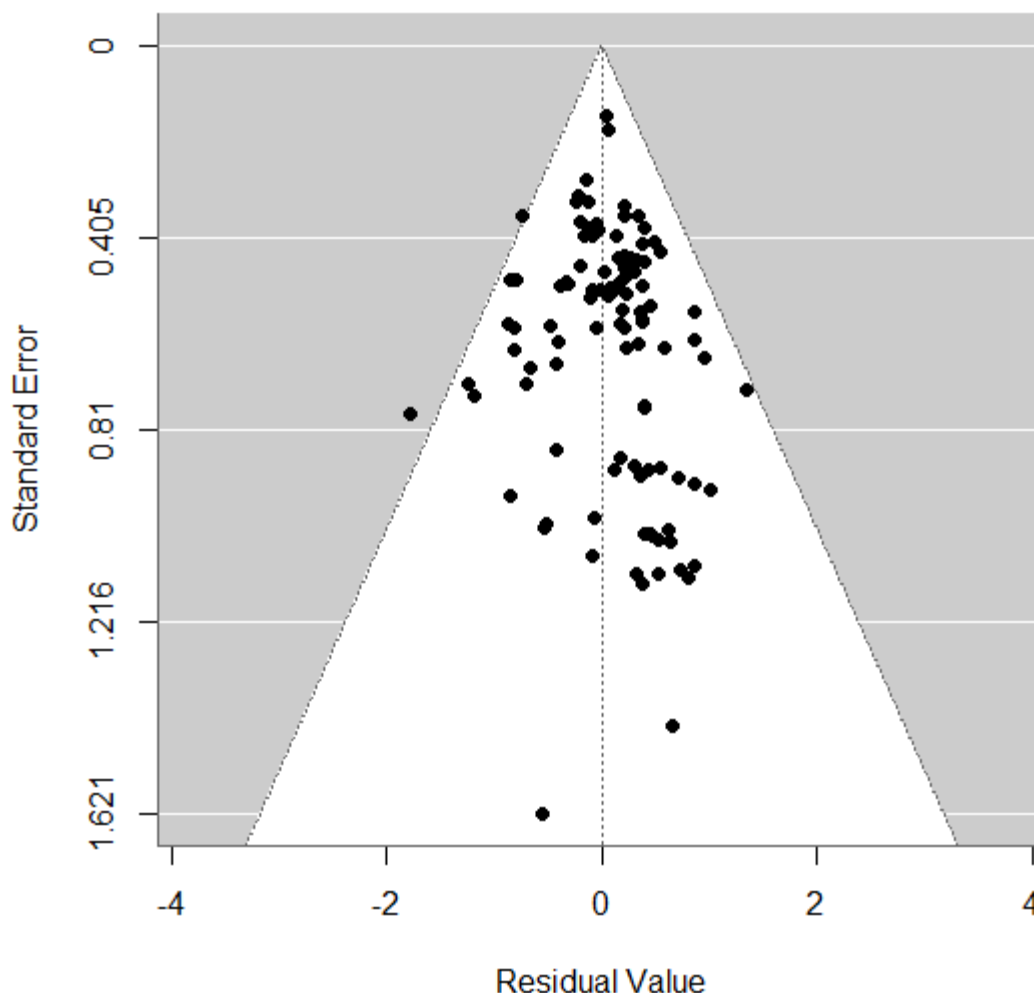
Table 4. Publication Bias: Egger's Regression Using CHE Model, Accounting for Dependency (CHE Sandwich)

Variable	Coefficient (SE)	t-statistic, <i>df</i>	p-value
Intercept	-0.07 (0.07)	-0.92, 6.07	0.39
sei	-0.46 (0.20)	-2.24, 3.85	0.09

Notes: CHE = correlated and hierarchical effects; *df* = degrees of freedom; SE = standard error; sei = modified standard error estimate in variance-covariance matrix

Non-significant coefficient indicates there is not enough evidence to support the presence of publication bias.

Figure 10. Funnel Plot Assessing Publication Bias



Summary of Findings and Recommendations for Implementation

Summary of Findings

Across 18 studies and 106 effects, we found that use-of-force reform efforts trended in a direction favoring the effectiveness of reform, despite failing to reach standard levels of statistical significance. Compared with officers and departments who were unaffected by reforms, treated officers and agencies reported lower rates of use of force (-11%, CI: -25%, + 5%) and citizen complaints (-18%, CI: -39%, +12%). However, confidence intervals around both estimates included zero, so we were unable to rule out null results.

Compared with officers and departments who were unaffected by reforms, treated officers and agencies reported lower rates of use of force (-11%, CI: -25%, + 5%) and citizen complaints (-18%, CI: -39%, +12%). However, confidence intervals around both estimates included zero, so we were unable to rule out null results.

Our moderation analyses revealed that only the voluntary nature of reform effort participation was significantly associated with variation across studies. Both studies driving this effect were RCTs that randomized treatment assignment within a sample of officers who volunteered to wear a BWC (Braga et al., 2018; Jennings et al., 2015). These studies were also among the first RCTs of BWCs to be conducted, during a time of rapid BWC policy adoption and revision. For these reasons, we are interpreting these findings with caution, as it is possible that volunteer officers simply benefit more from treatment than non-volunteer officers—meaning that these effects are indicative of treatment effect heterogeneity in studies with mandatory assignment—or that the novelty of BWC policy and experience at this time led officers to believe their actions would be scrutinized more strictly than supervisors were ultimately capable of in practice.

These findings largely align with those reported by recent meta-analyses (Lum et al., 2020) and individual studies of various reform efforts (for example, see Yokum et al., 2017). With regard to BWC research, despite promising early evidence of beneficial effects and systematic reviews supporting these findings, a more comprehensive review by Lum et al. (2020) found less consistent evidence in favor of BWCs. While scholars have taken this to indicate that there may be implementation issues (among others) affecting newer studies of BWCs (Gaub & White, 2020), we do not yet know enough about the differences between modern BWC policies and slippage in BWC activation requirements to understand how these issues may be affecting results from recent studies.

With respect to other types of reform efforts— de-escalation and procedural justice—we lack sufficient RCT and high-quality QED studies to estimate consistent effects, despite promising recent evidence supporting each of these reforms (Engel et al., 2022; Goh, 2021; White et al., 2023; Wood et al., 2020). Given the paucity of RCT and high-quality QED evidence evaluating programs such as these, our results should be interpreted with caution for de-escalation and procedural justice trainings (among other rarely evaluated programs and interventions), as many of the studies in the current meta-analysis strictly concern the impact of BWCs (n=10 of 18 total studies).

Limitations in the Existing Evidence Base

Despite longstanding scholarly attention, the evidence base on what works to reduce use of force remains limited, at least with respect to RCTs and high-quality QEDs. In part, this is due to reforms being adopted rapidly by police departments, often before proper empirical evaluations are possible. For example, although de-escalation training and CIT are used broadly across police agencies in the United States, much of the research concerning their impact is correlational (for example, see Rockwell et al., 2021), restricted to evaluating changes in officer attitudes and preferences (Compton et al., 2014), or, in the case of de-escalation training, more commonly evaluated in contexts other than policing (e.g., much of the de-escalation research occurs in the nursing context; see Engel et al., 2020). There are many reasonable explanations for these limitations, but the ultimate consequence is that the evidence base for use-of-force reform lacks a strong empirical footing, with many studies lacking both internal and external validity.

While conducting our meta-analysis, we also encountered difficulties obtaining the necessary statistical information to compute effect sizes. Of the 24 studies included in our initial sample of full texts, six studies required author queries for a total of 10 individual effect sizes. Although we did our best to include all eligible studies and effect sizes, our author queries were often unsuccessful. We understand that authors may be unwilling—or even unable—to revisit old analyses at the request of researchers conducting evidence syntheses, but we believe the lack of standard statistical information in some of these studies speaks to a general need for cross-disciplinary statistical reporting standards. Certain open science practices could also alleviate these issues, particularly the practice of making programming syntax and data publicly available on authors’ personal websites or on sites such as GitHub.

Limitations of Our Approach

It is important to note several limitations in the current meta-analysis that constrain our ability to speak more broadly about the use-of-force evidence base. First, due to time constraints, we were unable to assess additional outcomes beyond citizen complaints and

general use of force. Although researchers did not often report additional outcomes, several studies included breakdowns of the type of nonfatal force (minor, intermediate, major) and citizen complaints (filed, dismissed, sustained). Policymakers and practitioners alike would benefit from researchers recording these more nuanced outcomes and reporting the effects of reform efforts beyond simply fatal or nonfatal use of force or whether a citizen filed a complaint, regardless of its outcome. Second, we only included RCTs and certain qualifying QEDs in our analysis. Although the effects researchers report in less rigorous studies are likely affected by bias, they provide some information about the range of plausible effect sizes we might observe in future studies. Further, the ability of a department to conduct an RCT in partnership with researchers might be a form of selection bias in itself, though a recent study suggests such bias might be minimal (Gechter & Meager, 2021). Third, many of the studies in the current analysis focus on BWCs (10 of 18 studies). Although we used statistical models to account for different types of reform strategies, the limited number of studies on each other type of reform (e.g., de-escalation training, federal consent decree) means that our estimates should be interpreted with caution beyond RCT and high-quality QED evidence regarding BWCs.

Recommendations for Improving the Evidence Base for Use-of-Force Reform

Rapid-Response Funding Streams. As is often lamented by policing scholars (for example, see Engel et al., 2020), police reform efforts—including de-escalation and CIT—are broadly adopted without rigorous empirical evidence supporting their efficacy and confirming that they do not result in unintended consequences. Although we need more RCTs and high-quality QEDs on police reform efforts, we recognize that the realm of policing is not unique in this pattern of widespread adoption of reform outpacing empirical evidence. Scholars frequently voice the same criticism regarding other types of legal system reform, including novel approaches to community supervision violations (Cullen et al., 2018) and the use of actuarial tools and needs assessments in correctional populations (Klingele, 2016).

Yet another call for more timely and rigorous research evidence will not “move the needle” toward better evidence because it does not speak to one of the more prominent issues causing the lag between practitioners’ need to adopt novel reforms and the time required to conduct an RCT or high-quality QED. High-quality, timely research on promising reforms requires coordinating both *time* (having a partnership in place before a reform is implemented) and *resources* (having the funds required to conduct an evaluation), which often takes upwards of a year to arrange, if not longer. To properly respond to this issue, we need to work with policymakers at federal, state, and local levels to develop innovative

types of research funding streams that are more responsive to the time-sensitive needs of criminal legal system organizations, much like the opportunities offered at the outset of the Covid-19 pandemic to facilitate a timely understanding of its impact on public safety. Although this need is generally fulfilled by more agile philanthropic organizations, it is overdue for government organizations to enter this space with novel, rapid-response funding streams that simultaneously incentivize research partnerships with legal system practitioners to enable RCTs and rigorous QEDs and provide more expeditious responses to rapidly evolving reform efforts.

Pairing Impact Evaluations With Implementation Evaluations. Implementation evaluations can better contextualize why the same (or very similar) reform efforts work in some places and not in others. While we acknowledge that implementation evaluations are not simple additions to RCTs and high-quality QEDs, understanding **how** reform works is sometimes just as important as understanding **why** it works. For example, one of our key findings—that the voluntary nature of BWC participation is associated with significant reductions in use of force—might be attributable to differences in treatment protocol compliance across studies. Logically, officers who volunteer for treatment may comply with treatment protocols more consistently than officers who do not volunteer, potentially explaining why studies of mandatory interventions tend to exhibit wider ranges of effect sizes. Unfortunately, because most studies do not conduct implementation evaluations—including many of the studies in this meta-analysis—it is not possible for us to evaluate this hypothesis. With adequate funding, researchers could conduct implementation evaluations more frequently and gain a better understanding of which components of reform implementation account for differences in efficacy across different contexts, including compliance with treatment protocols.

Using Emerging Technology to Expand Prior Work and Pursue Novel Outcomes. Many studies that appeared eligible after abstract screening were ultimately excluded from the analyses because they only measured attitudinal outcomes following reform efforts. While understanding the theory of change is important, evaluations without behavioral outcomes only address the first step in that process. Moreover, it is unclear whether perceptual or attitudinal changes are necessary for all types of trainings to affect officer behavior. For example, both de-escalation and CIT curricula include practical components that teach officers new ways to respond in tense encounters with citizens—techniques that do not necessarily require perceptual or attitudinal change to achieve reductions in use of force.

The advent of new technologies to automatically process body-worn camera data—including Truststat and Truleo (Adams et al., 2024; Camp & Voight, 2025)—will place more complex and comprehensive measurements of behavioral outcomes within reach for

researchers evaluating use-of-force reform. For example, research will be able to rely on automated processing of BWC footage to collect incident-level data on how officers interact with citizens, without the need for prolonged manual footage review. Researchers are only beginning to understand the impact these tools may have on evaluating reform efforts (for example, see White et al., 2025) and, to our knowledge, are yet to explore how automated BWC footage processing could contribute to a better understanding of how officers successfully use (or do not use) de-escalation or CIT tactics in the field.

Conclusion

This meta-analysis represents a first step toward an ongoing and expanding review of extant research evidence on the effectiveness of use-of-force reform, in partnership with the AIR Opportunity Fund and MOSAIC. While we found that the overall effects of use-of-force reform appeared to trend in the direction favoring reform, the effects were insignificant. Moderation analysis also suggested that supportive effects were partially explained by officers volunteering to wear BWCs. Further, more than half of the studies included in this meta-analysis evaluated BWC policies, as there were very few RCTs or high-quality QEDs focused on other widely used reform efforts such as CIT and de-escalation training. Despite recent evidence supporting such programs, more research is needed to understand the impact of use-of-force reforms beyond BWC policies. More generally, nascent research on use-of-force reform needs to include implementation evaluation to properly contextualize why reform is or is not successful. Emerging technologies—such as automated processing of BWC footage—can be used to facilitate implementation evaluations and to explore additional reform outcomes previously thought impractical due to limitations in data collection.

References

* Indicates citation included in review

Adams, I. T., McLean, K., & Alpert, G. (2024). Improving police behavior through artificial intelligence: Pre-registered experimental results in two large US agencies. *CrimRxiv*. <https://www.crimrxiv.com/pub/q5fehcov>

Alpert, G. P., & Dunham, R. G. (2010). Policy and Training Recommendations Related to Police Use of CEDs: Overview of Findings From a Comprehensive National Study. *Police Quarterly*, 13(3), 235–259. <https://doi.org/10.1177/1098611110373993>

*Ariel, B., Farrar, W. A., & Sutherland, A. (2015). The Effect of Police Body-Worn Cameras on Use of Force and Citizens' Complaints Against the Police: A Randomized Controlled Trial. *Journal of Quantitative Criminology*, 31(3), 509–535. <https://doi.org/10.1007/s10940-014-9236-3>

*Ba, B. A. (2018). *A Dissertation Submitted to the Faculty of the Irving B. Harris Graduate School of Public Policy Studies in Candidacy for the Degree of Doctor of Philosophy*.

*Boehme, H. M., Kaminski, R. J., & Leasure, P. (2022). A comparative interrupted time-series assessing the impact of the Armstrong decision on officer-involved shootings. *Police Practice and Research*, 23(5), 614–622. <https://doi.org/10.1080/15614263.2022.2079508>

*Braga, A. A., Barao, L. M., Zimmerman, G. M., Douglas, S., & Sheppard, K. (2020). Measuring the Direct and Spillover Effects of Body Worn Cameras on the Civility of Police–Citizen Encounters and Police Work Activities. *Journal of Quantitative Criminology*, 36(4), 851–876. <https://doi.org/10.1007/s10940-019-09434-9>

*Braga, A. A., Sousa, W. H., Coldren, J. R., & Rodriguez, D. (2018). The Effects of Body-Worn Cameras on Police Activity and Police-Citizen Encounters: A Randomized Controlled Trial. *Journal of Criminal Law and Criminology*, 108(3). <https://scholarlycommons.law.northwestern.edu/jclc/vol108/iss3/3/>

Burghart, B. (2025). *Methodology – Fatal Encounters*. Fatal Encounters. <https://fatalencounters.org/methodology/>

- Camp, N. P., & Voigt, R. (2025). Body camera footage as data: Using natural language processing to monitor policing at scale & in depth. *Behavioral Science & Policy*, 23794607241308636. <https://doi.org/10.1177/23794607241308636>
- Compton, M. T., Demir Neubert, B. N., Broussard, B., McGriff, J. A., Morgan, R., & Oliva, J. R. (2011). Use of force preferences and perceived effectiveness of actions among Crisis Intervention Team (CIT) police officers and non-CIT officers in an escalating psychiatric crisis involving a subject with schizophrenia. *Schizophrenia Bulletin*, 37(4), 737–745.
- Cullen, F. T., Pratt, T. C., Turanovic, J. J., & Butler, L. (2018). When Bad News Arrives: Project HOPE in a Post-Factual World. *Journal of Contemporary Criminal Justice*, 34(1), 13–34. <https://doi.org/10.1177/1043986217750424>
- Dunham, R. G., Alpert, G. P., & McLean, K. (2021). *Critical issues in policing: Contemporary readings* (Eighth edition).
- *Engel, R. S., Corsaro, N., Isaza, G. T., & McManus, H. D. (2022). Assessing the impact of de-escalation training on police behavior: Reducing police use of force in the Louisville, KY Metro Police Department. *Criminology & Public Policy*, 21(2), 199–233. <https://doi.org/10.1111/1745-9133.12574>
- Engel, R. S., McManus, H. D., & Herold, T. D. (2020). Does de-escalation training work? A systematic review and call for evidence in police use-of-force reform. *Criminology & Public Policy*, 19(3), 721–759. <https://doi.org/10.1111/1745-9133.12467>
- *Fagan, J., & Campbell, A. D. (2020). RACE AND REASONABLENESS IN POLICE KILLINGS. *Boston University Law Review*, 100(3), 951–1016.
- Farrington, D. P., MacKenzie, D. L., Sherman, L. W., & Welsh, B. C. (2003). *Evidence-Based Crime Prevention*. Taylor and Francis.
- Gaston, S., Teti, M. J., & Sanchez, M. (2021). Does racial congruence between police agencies and communities reduce racialized police killings of civilians? *Criminology & Public Policy*, 20(4), 665–690. <https://doi.org/10.1111/1745-9133.12567>
- Gaub, J. E., & White, M. D. (2020). Open to Interpretation: Confronting the Challenges of Understanding the Current State of Body-Worn Camera Research. *American Journal of Criminal Justice*, 45(5), 899–913. <https://doi.org/10.1007/s12103-020-09518-4>

- *Gaub, J. E., Todak, N., & White, M. D. (2021). The distribution of police use of force across patrol and specialty units: A case study in BWC impact. *Journal of Experimental Criminology*, 17(4), 545–561. <https://doi.org/10.1007/s11292-020-09429-8>
- Gechter, M., & Meager, R. (2021). Combining experimental and observational studies in meta-analysis: A mutual debiasing approach. In *Technical Report*. Mimeo. https://www.academia.edu/download/83327733/MGRM_Combining_Experimental_and_Observational_Studies.pdf
- Goh, L. S. (2020b). *The Evaluation of Policy Measures to Reduce Police Use of Force*.
- *Goh, L. S. (2020a). Going Local: Do Consent Decrees and Other Forms of Federal Intervention in Municipal Police Departments Reduce Police Killings? *Justice Quarterly*, 37(5), 900–929. <https://doi.org/10.1080/07418825.2020.1733637>
- *Goh, L. S. (2021). Did de-escalation successfully reduce serious use of force in Camden County, New Jersey? A synthetic control analysis of force outcomes. *Criminology & Public Policy*, 20(2), 207–241. <https://doi.org/10.1111/1745-9133.12536>
- Gullion, C. L., & King, W. R. (2020). Early intervention systems for police: A state-of-the-art review. *Policing: An International Journal*, 43(4), 643–658. <https://doi.org/10.1108/PIJPSM-02-2020-0027>
- *Headley, A. M., Guerette, R. T., & Shariati, A. (2017). A field experiment of the impact of body-worn cameras (BWCs) on police officer behavior and perceptions. *Journal of Criminal Justice*, 53, 102–109. <https://doi.org/10.1016/j.jcrimjus.2017.10.003>
- Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Editors). (2024). *Cochrane Handbook for Systematic Reviews of Interventions*, Version 6.5. Cochrane. Available from training.cochrane.org/handbook.
- *Huff, J., Katz, C. M., & Hedberg, E. C. (2020). A randomized controlled trial of the impact of body-worn camera activation on the outcomes of individual incidents. *Journal of Experimental Criminology*, 18(2), 247–272. <https://doi.org/10.1007/s11292-020-09448-5>
- Huff, J., White, M. D., Malm, A. E., & Katz, C. M. (2025). An Officer-Level Examination of the Prevalence and Correlates of Police Body-Worn Camera Activation. *Police Quarterly*, 28(1), 45–77. <https://doi.org/10.1177/10986111241241751>

- Hyland, S. S. (2018). *Body-Worn Cameras in Law Enforcement Agencies, 2016* (No. NCJ 251775). US Department of Justice.
<https://bjs.ojp.gov/content/pub/pdf/bwclea16.pdf>
- Iannone, R., & Roy, O. (2015). *DiagrammeR: Graph/Network Visualization* (p. 1.0.11) [Dataset]. <https://doi.org/10.32614/CRAN.package.DiagrammeR>
- James, L., James, S., & Vila, B. (2018). Testing the impact of citizen characteristics and demeanor on police officer behavior in potentially violent encounters. *Policing: An International Journal*, 41(1), 24–40. <https://doi.org/10.1108/PIJPSM-11-2016-0159>
- *Jennings, W. G., Lynch, M. D., & Fridell, L. A. (2015). Evaluating the impact of police officer body-worn cameras (BWCs) on response-to-resistance and serious external complaints: Evidence from the Orlando police department (OPD) experience utilizing a randomized controlled experiment. *Journal of Criminal Justice*, 43(6), 480–486. <https://doi.org/10.1016/j.jcrimjus.2015.10.003>
- *Jennings, W. G., Fridell, L. A., Lynch, M., Jetelina, K. K., & Reingle Gonzalez, J. M. (2017). A Quasi-Experimental Evaluation of the Effects of Police Body-Worn Cameras (BWCs) on Response-to-Resistance in a Large Metropolitan Police Department. *Deviant Behavior*, 38(11), 1332–1339. <https://doi.org/10.1080/01639625.2016.1248711>
- Jiao, A. Y. (2021). Federal consent decrees: A review of policies, processes, and outcomes. *Police Practice and Research*, 22(1), 793–804.
<https://doi.org/10.1080/15614263.2020.1722664>
- *Johnson, T. L., Johnson, N. N., Sabol, W. J., & Snively, D. T. (2022). Law Enforcement Agencies' College Education Hiring Requirements and Racial Differences in Police-Related Fatalities. *Journal of Police and Criminal Psychology*, 37(3), 681–698.
<https://doi.org/10.1007/s11896-022-09534-6>
- Klinge, C. (2016). The promises and perils of evidence-based corrections. *Notre Dame Law Review*, 91(2), 2.
- Klinger, D., Rosenfeld, R., Isom, D., & Deckard, M. (2016). Race, Crime, and the Micro-Ecology of Deadly Force. *Criminology & Public Policy*, 15(1), 193–222.
<https://doi.org/10.1111/1745-9133.12174>
- Kochel, T. R., & Nouri, S. (2024). Impact of In-Service Implicit Bias Training: A Study of Attitudinal Changes and Intention to Apply Anti-Bias Techniques. *Police Quarterly*, 27(4), 561–581. <https://doi.org/10.1177/10986111241237512>

- Lum, C., Koper, C. S., Wilson, D. B., Stoltz, M., Goodier, M., Eggins, E., Higginson, A., & Mazerolle, L. (2020). Body-worn cameras' effects on police officers and citizen behavior: A systematic review. *Campbell Systematic Reviews*, 16(3), e1112. <https://doi.org/10.1002/cl2.1112>
- Maskaly, J., Donner, C., Jennings, W. G., Ariel, B., & Sutherland, A. (2017). The effects of body-worn cameras (BWCs) on police and citizen outcomes: A state-of-the-art review. *Policing: An International Journal of Police Strategies & Management*, 40(4), 672–688. <https://doi.org/10.1108/PIJPSM-03-2017-0032>
- McKesson, D., Sinyangwe, S., Elzie, J., & Packnett, B. (2016). *Police Use of Force Policy Analysis*. Campaign Zero. <https://oag.ca.gov/sites/all/files/agweb/pdfs/ripa/policy-analysis-sinyangwe.pdf>
- Methods of Synthesis and Integration Center (MOSAIC) at the American Institutes for Research. (2025). Evidence Gap Maps. https://airshinyapps.shinyapps.io/mosaic_egm/
- *Miller, J., & Chillar, V. F. (2022). Do Police Body-Worn Cameras Reduce Citizen Fatalities? Results of a Country-Wide Natural Experiment. *Journal of Quantitative Criminology*, 38(3), 723–754. <https://doi.org/10.1007/s10940-021-09513-w>
- Morabito, M. S., Socia, K., Wik, A., & Fisher, W. H. (2017). The nature and extent of police use of force in encounters with people with behavioral health disorders. *International Journal of Law and Psychiatry*, 50, 31–37. <https://doi.org/10.1016/j.ijlp.2016.10.001>
- Nagin, D. S., & Telep, C. W. (2020). Procedural justice and legal compliance: A revisionist perspective. *Criminology & Public Policy*, 19(3), 761–786. <https://doi.org/10.1111/1745-9133.12499>
- Nakagawa, S., Lagisz, M., O'Dea, R. E., Pottier, P., Rutkowska, J., Senior, A. M., Yang, Y., & Noble, D. W. A. (2023). orchaRd 2.0: An R package for visualising meta-analyses with orchard plots. *Methods in Ecology and Evolution*, 14(8), 2003–2010. <https://doi.org/10.1111/2041-210X.14152>
- Ooms, J. (2017). *writexl: Export Data Frames to Excel “xlsx” Format* (p. 1.5.3) [Dataset]. <https://doi.org/10.32614/CRAN.package.writexl>

- *Owens, E., Weisburd, D., Amendola, K. L., & Alpert, G. P. (2018). Can You Build a Better Cop? *Criminology & Public Policy*, 17(1), 1–259.
<https://doi.org/10.1111/1745-9133.12337>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, n71.
<https://doi.org/10.1136/bmj.n71>
- Paoline, E. A., & Terrill, W. (2007). Police Education, Experience, and the Use of Force. *Criminal Justice and Behavior*, 34(2), 179–196.
<https://doi.org/10.1177/0093854806290239>
- *Peterson, B. E., Yu, L., Vigne, N. L., & Lawrence, D. S. (2018). The Milwaukee Police Department’s Body-Worn Camera Program: Evaluation Findings and Key Takeaways. *Urban Institute*.
- Peterson, J., & Densley, J. (2018). Is Crisis Intervention Team (CIT) training evidence-based practice? A systematic review. *Journal of Crime and Justice*, 41(5), 521–534.
<https://doi.org/10.1080/0735648X.2018.1484303>
- Pigott, T. D., & Polanin, J. R. (2020). Methodological Guidance Paper: High-Quality Meta-Analysis in a Systematic Review. *Review of Educational Research*, 90(1), 24–46.
<https://doi.org/10.3102/0034654319877153>
- Polanin, J. R., Austin, M., Peko-Spicer, S., Ebersole, C., Michaelson, L., Iacocca, M. O., Clements, J., Soule, C., Lee, S., Ezzat, Y., Williams, S., Mitchell, S., & Williams, R. T. (2025). *MetaReviewer* (Version 1.3.0) [Computer software]. American Institutes for Research. <https://www.metareviewer.org/>
- Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the Difference Between Published and Unpublished Effect Sizes: A Meta-Review. *Review of Educational Research*, 86(1), 207–236. <https://doi.org/10.3102/0034654315582067>
- *Powell, Z. A. (2023). Did California Act to Save Lives? AB 392 and fatal officer-involved shootings. *Policing: A Journal of Policy and Practice*, 17, paad073.
<https://doi.org/10.1093/polic/paad073>

- Pustejovsky, J. E. (2016). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections* (p. 0.6.0) [Dataset].
<https://doi.org/10.32614/CRAN.package.clubSandwich>
- Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, 10(1), 57–71.
<https://doi.org/10.1002/jrsm.1332>
- Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Prevention Science*, 23(3), 425–438.
<https://doi.org/10.1007/s11121-021-01246-3>
- R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. [Computer software]. Foundation for Statistical Computing. <https://www.R-project.org>
- Rockwell, A. R., Bishopp, S. A., & Orrick, E. A. (2021). Do policy and training changes influence patterns of police use of force? An interrupted time-series analysis. *Policing: An International Journal*, 44(3), 469–482.
- Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, 26(2), 141.
- Saldanha, I. J., Smith, B. T., Ntzani, E., Jap, J., Balk, E. M., & Lau, J. (2019). The Systematic Review Data Repository (SRDR): Descriptive characteristics of publicly available data and opportunities for research. *Systematic Reviews*, 8(1), 334.
<https://doi.org/10.1186/s13643-019-1250-y>
- Seo, C., Kim, B., & Kruis, N. E. (2021). A Meta-Analysis of Police Response Models for Handling People With Mental Illnesses: Cross-Country Evidence on the Effectiveness. *International Criminal Justice Review*, 31(2), 182–202.
<https://doi.org/10.1177/1057567720979184>
- Sherman, L. W., & Blumberg, M. (1981). Higher education and police use of deadly force. *Journal of Criminal Justice*, 9(4), 317–331.
- Sinyangewe, S., Hammond, M., & Emerson, J. (2025). *2024 Police Violence Report*. Mapping Police Violence. <https://policeviolencereport.org/2024/>
- Taeho, K. (2021). *Essays in Personnel Economics of Policing*.
https://knowledge.uchicago.edu/record/2967/files/Kim_uchicago_0330D_15707.pdf

- Terrill, W., & Paoline Iii, E. A. (2012). Conducted Energy Devices (CEDs) and Citizen Injuries: The Shocking Empirical Reality. *Justice Quarterly*, 29(2), 153–182.
<https://doi.org/10.1080/07418825.2010.549834>
- The EndNote Team. (2013). EndNote. [Clarivate](#).
- Todak, N., & James, L. (2018). A Systematic Social Observation Study of Police De-Escalation Tactics. *Police Quarterly*, 21(4), 509–543.
<https://doi.org/10.1177/1098611118784007>
- Tyler, T. R. (1988). What is procedural justice? Criteria used by citizens to assess the fairness of legal procedures. *Law & Society Review*, 22(1), 103–135.
- Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the **metafor** Package. *Journal of Statistical Software*, 36(3). <https://doi.org/10.18637/jss.v036.i03>
- Watson, A. C., Ottati, V. C., Draine, J., & Morabito, M. (2011). CIT in context: The impact of mental health resource availability and district saturation on call dispositions. *International Journal of Law and Psychiatry*, 34(4), 287–294.
<https://doi.org/10.1016/j.ijlp.2011.07.008>
- White, M. D., Orosco, C., & Watts, S. (2023b). Beyond force and injuries: Examining alternative (and important) outcomes for police De-escalation training. *Journal of Criminal Justice*, 89, 102129. <https://doi.org/10.1016/j.jcrimjus.2023.102129>
- *White, M. D., Gaub, J. E., & Todak, N. (2017). Exploring the Potential for Body-Worn Cameras to Reduce Violence in Police–Citizen Encounters. *Policing*.
<https://doi.org/10.1093/police/paw057>
- *White, M. D., Orosco, C., & Watts, S. (2023a). Can police de-escalation training reduce use of force and citizen injury without compromising officer safety? *Journal of Experimental Criminology*. <https://doi.org/10.1007/s11292-023-09584-8>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686.
<https://doi.org/10.21105/joss.01686>

- Wickham, H., Henry, L., Pedersen, T. L., Luciani, T. J., Decorde, M., & Lise, V. (2015). *svglite: An “SVG” Graphics Device* (p. 2.1.3) [Dataset]. <https://doi.org/10.32614/CRAN.package.svglite>
- Wood, G., Tyler, T., Papachristos, A. V., Roth, J., & Sant’Anna, P. H. C. (2021). *Revised findings for “Procedural justice training reduces police use of force and complaints against officers.”* SocArXiv. <https://doi.org/10.31235/osf.io/xf32m>
- *Wood, G., Tyler, T. R., & Papachristos, A. V. (2020). Procedural justice training reduces police use of force and complaints against officers. *Proceedings of the National Academy of Sciences*, 117(18), 9815–9821. <https://doi.org/10.1073/pnas.1920671117>
- *Worden, R. E., Kim, M., Harris, C. J., Pratte, M. A., Dorn, S. E., & Hyland, S. S. (2013). Intervention With Problem Officers: An Outcome Evaluation of an EIS Intervention. *Criminal Justice and Behavior*, 40(4), 409–437. <https://doi.org/10.1177/0093854812458095>
- Worden, R. E., Najdowski, C. J., McLean, S. J., Worden, K. M., Corsaro, N., Cochran, H., & Engel, R. S. 2024. Implicit bias training for police: Evaluating impacts on enforcement disparities. *Law and Human Behavior*, 48(5-6): 338-355.
- *Yokum, D., Ravishankar, A., & Coppock, A. (2017). *Evaluating the Effects of Police Body-Worn Cameras: A randomized Controlled Trial*. The Lab at DC.



AIR OPPORTUNITY FUND

About the American Institutes for Research®

Established in 1946, the American Institutes for Research® (AIR®) is a nonpartisan, not-for-profit institution that conducts behavioral and social science research and delivers technical assistance both domestically and internationally in the areas of education, health, and the workforce. AIR's work is driven by its mission to generate and use rigorous evidence that contributes to a better, more equitable world. With headquarters in Arlington, Virginia, AIR has offices across the U.S. and abroad. For more information, visit [AIR.ORG](https://www.air.org).

AIR® Headquarters

1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
+1.202.403.5000 | [AIR.ORG](https://www.air.org)

Notice of Trademark: "American Institutes for Research" and "AIR" are registered trademarks. All other brand, product, or company names are trademarks or registered trademarks of their respective owners.

Copyright © 20xx American Institutes for Research®. All rights reserved. No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, website display, or other electronic or mechanical methods, without the prior written permission of the American Institutes for Research. For permission requests, please use the Contact Us form on [AIR.ORG](https://www.air.org).