A Randomized Controlled Trial of Social Interaction Police Training

Forthcoming in *Criminology & Public Policy*

Kyle McLean
Clemson University
Scott E. Wolfe
Jeff Rojek
Michigan State University
Geoffrey P. Alpert
University of South Carolina
Michael R. Smith
University of Texas at San Antonio

Keywords
police training, social interaction, de-escalation, procedural justice, evidence-based policing, randomized-controlled trials

Research Summary

We conducted a randomized-controlled trial (RCT) of a social interaction training program to determine its effectiveness in improving attitudes and behaviors among police officers. Survey data and a series of difference-in-difference tests found that participating in the training program improved attitudes with treatment group officers placing higher priorities on procedurally-fair communication during a hypothetical officer-citizen encounter. An interrupted-time series analysis of official use of force reports provided no evidence that the training program altered officer behavior.

Policy Implications

Policing scholars and reformers have increasingly called for improvements to police training that emphasize communication and de-escalation skills. While many programs addressing these issues exist, evidence of their effectiveness has been scarce. Our findings provide evidence that such training may improve police officer attitudes, but perhaps not behaviors.

This research was funded by the National Institute of Justice under Grant 2016-IJ-CX-0018. Points of view and opinions provided are those of the authors and do not represent the official position of the US Department of Justice. An earlier version of this manuscript was presented at the 2018 American Society of Criminology meeting in Atlanta, GA. The authors would like to thank their partners in the Fayetteville and Tucson police departments for making this evaluation possible. They also thank John MacDonald for his helpful comments on the manuscript. All errors belong to the authors. Please direct correspondence to Kyle McLean, Department of Sociology, Anthropology, and Criminal Justice, Clemson University, 132 Brackett Hall, Clemson, SC 29634 (email: kdmclea@clemson.edu).

In the wake of numerous controversial police use of force encounters with citizens in recent years, the need for improved social interaction and de-escalation skills has been featured as the centerpiece of police reform efforts (President's Task Force on 21st Century Policing, 2015). To improve these skills, many have suggested improvements in the quality and availability of police training on interactions with citizens. As such, academic researchers often argue that successful police training interventions should focus on improving officers' ability to apply procedural justice or effective communication principles consistently throughout their interactions with citizens (Mazerolle, Antrobus, Bennett, & Tyler, 2013; Sargeant, Antrobus, & Platz, 2017). We know, for example, that citizens are more likely to comply when they believe officers have treated them with respect, provided them a voice, and been unbiased in their decision-making—i.e., acted with procedural justice (Tyler, 2006; Walters & Bolger, 2018; Wolfe, Nix, Rojek, & Kaminski, 2016), so it is only logical that improving officers' behavior along these dimensions should improve police-citizen interactions. Yet, as Nagin and Telep (2017) recently noted, rarely is theory or research evidence integrated into police training, and even when it is, it is rarely subjected to rigorous scientific evaluation.

A critical development in the professionalization of policing over the past 50 years has been the implementation of data and science into the practices of police agencies (Sherman, 2013). The evidence-based policing movement has demanded a model of policing that "links external demands on police…to research evidence on how to meet those demands…" (Sherman, 2013:381) or as Alpert (1988:453) called it: "linking data to decisions." Yet with respect to police training, Skogan and colleagues (2015:320) accurately described the state of scientific knowledge by declaring, "We know virtually nothing about the short- or long-term effects associated with police training of any type." Similarly, in attempting a systematic review of studies of police training,

Huey (2018) concluded that there were too few studies on any single topic to conduct a review. Thus, modern policing is currently in a precarious situation, with external demands for police agencies to implement training programs for which little research evidence exists. Nowhere is the need for an evidence-base more acute than in the area of officer social interaction skills and de-escalation.

To this end, the present study used a randomized-controlled trial to evaluate a social interaction training program in two mid-size police departments in the United States. This program involved the repetitive practice of social interaction skills on a bi-weekly basis over several months. We collected data on both treatment and control officers' priorities in a hypothetical officer-citizen scenario before and after the training program, as well as official reports of use of force incidents before, during, and after the training program was implemented. The results indicate that the training program was effective at increasing the priority officers place on procedurally fair communication but was not effective at reducing the number of reported use of force incidents for officers undergoing the training program. Prior to discussing our experimental results, we first explore the literature on social interaction and related police training and detail the elements of the training program we evaluated.

Policing, Social Interaction, and the Use of Force

One of the fundamental challenges of police organizations in a democratic society is the exercise of the social control function in a way that reflects just action and restraint and is consistent with social and legal expectations. This challenge is most evident in police-citizen interactions that involve officers using some form of physical force. While extensive research has been conducted on police use of force (e.g., Alpert & Dunham, 2000; Eith & Durose, 2011; Garner & Maxwell, 2002; Kaminski et al., 2015; Paoline, Terrill, & Ingram, 2012; Terrill & Mastrofski,

2002), most of these efforts have treated these events as static. An interactionist approach to police use of force, however, conceptualizes officer-citizen interactions as dynamic exchanges that unfold with varying potential to escalate into a use of force event or de-escalate into a peaceful resolution (e.g., Alpert & Dunham, 2004; Sykes & Clark, 1975).

Drawing on Goffman's (1956, 1961) classic work on social interaction, Sykes and Clark (1975) asserted that encounters between individuals are governed by a set of exchange rituals that create order in interactions, including the mutual expectation that each individual will show respect and regard for the other. However, Sykes and Clark highlight that the rules for showing respect and regard, or deference, within police-citizen encounters is asymmetrical in nature given the officer's formal authority to enforce the law and maintain order. That is, officers expect a higher degree of deference shown to them that they do not feel they have to reciprocate. Officers seek to maintain the authority afforded through this asymmetry because it allows them to control a situation—including the ability to question, command, and even physically coerce if necessary—and arrive at outcomes that serve their interest (Bittner, 1967; Muir, 1977; Van Maanen, 1978). When citizens do not abide by this asymmetrical deference expectation, officers will attempt to re-assert their authority, which Sykes and Brent (1980) found was generally accomplished by taking verbal control of an interaction.

Alpert and Dunham (2004) subsequently refined the work of Sykes and Clark to explain more thoroughly how police-citizen interactions turn into use of force events. One of the notable distinctions of Alpert and Dunham's reconceptualization is their greater recognition of the citizen's role during interactions. Just as officers enter with the goals of maintaining their authority and controlling the interaction, citizens also have a set of expectations that can range from being treated with respect to actively avoiding an interaction to avoiding apprehension for an offense. Similarly,

where officers will become more coercive to overcome their goals being blocked, citizens will increasingly resist officers when their goals are not met. Alpert and Dunham assert that this blockage creates an action-reaction chain that can escalate into force and continue until one party acquiesces.

The benefit of Alpert and Dunham's (2004) model is the articulation of force events as an escalating or de-escalating exchange of coercion, resistance, or compliance. Thus, each interaction follows a trajectory, whereby the officer and citizen either escalate the interaction towards force or de-escalate it away from force. Using this conceptualization, police training should target skills and methods for turning the intensity or trajectory of the interaction away from force. However, basic police training typically focuses on physical tactics during citizen interactions. For example, a survey of state and local law enforcement academies found that academies spent an average of 71 hours training new recruits on firearms, 60 hours on self-defense, and only 21 hours on use of force policies, de-escalation tactics, and crisis intervention strategies combined (Reaves, 2016). Moreover, the asymmetric deference expectation is often engrained in officers through such training and socialization processes (e.g., "maintaining the edge," see Van Maanen, 1978). Traditional police training rarely spends time teaching officers that, when appropriate, listening to people's concerns, empathizing with their situation, maintaining respect, and the like, may be valuable social interaction skills that can help turn encounters away from the need to use force (thereby increasing officer safety). For example, Tyler's (1990) legitimacy theory suggests that when officers act in a procedurally-fair manner, individuals will have more favorable attitudes towards the police and be more likely to comply with – or show deference to –the police officer. Similarly, officers may employ de-escalation tactics, such as prioritizing communication and de-emphasizing physical control (Todak, 2017; Todak & James, 2018), to bend the trajectory of the

encounter away from the use of force. Notably, each of these suggestions relies on training officers to improve their social interaction skills.

Social Interaction Training

Research on training programs that focus on officers' social interaction skills has increased slightly in recent years (Hansson & Markström, 2014; Krameddine et al., 2013). Specifically, procedural justice theory has gained support among police researchers and some practitioners through promising findings based on training and subsequent evaluation. Nagin and Telep (2017) identified six such programs in a review of the application of procedural justice theory to policing. Most of the studies reviewed by Nagin and Telep showed promising outcomes, but the training programs varied on a number of dimensions.

First, the studies varied in the intensity of the training delivered. Specifically, four studies featured one-time delivery of communication skills training to in-service officers with no future re-training (Owens et al., 2016; Schaefer & Hughes, 2016; Skogan et al., 2015; Wheller et al., 2013), while the other two programs involved integrating communication skills training into the curriculum for new recruits (Robertson et al., 2014; Rosenbaum & Lawrence, 2013).[1] Second, studies varied in the outcomes measured. Several studies included attitudinal measures as outcomes (Rosenbaum & Lawrence, 2013; Schaefer & Hughes; Skogan et al., 2015), while others relied on behavioral measures of officers' social interaction skills (Owens et al., 2016; Wheller et al., 2013). Lonsway and colleagues (2001) noted that officers trained to improve interactions with victims of sexual assault did not report any changes in attitudes regarding sexual assault victims such as rape myth acceptance. However, officers involved in the training did see an improvement

[1] While the integration of training into the recruit program may have involved more repetitive training than the one-time delivery programs, they are still notably limited in that once recruit training is completed, the communication skills training is no longer repeated.

in ratings of their interview skills during a simulated interview with a sexual assault victim. Similarly, Rosenbaum and Lawrence (2013) evaluated a procedural justice training program and did not find significant differences in attitudes towards procedural justice or legitimacy. However, a review of videotapes of simulated officer-citizen interactions suggested a treatment effect on officer performance, though the sample of videotaped encounters was too small for strong conclusions. Thus, it is possible that officers may not report attitudinal changes but will change their behaviors (or vice versa) in officer-citizen interactions making the measurement of both attitudes and behaviors desirable in assessing training programs.

Finally, the training programs varied by their method of implementation. Some programs attempted to include scenario-based training (e.g., Wheller et al., 2013), but others still included lectures by university professors (see e.g., Rosenbaum & Lawrence, 2013) or were primarily discussion-based (e.g., Skogan et al., 2015). In sum, while we have done a better job in recent years attempting to translate empirically-supported theory into training practice, our knowledge of "what works" is largely limited to a single framework (i.e., procedural justice), and the limitations of programs and evaluations provide plenty of justification for evaluating larger and more sustained training programs. Still, the results of these studies are encouraging, as they provide preliminary evidence that officers can be trained on social interaction skills.

Tact, Tactics, and Trust (T3) Training Program

The current study focused on the Tact, Tactics, and Trust (T3) training program offered by Polis Solutions. The training is based on the "Good Stranger" program developed by the Defense Advanced Research Projects Agency (DARPA) to improve the social interaction skills of U.S. soldiers in Iraq and Afghanistan. The DARPA program emphasized the need for U.S. soldiers to be able to communicate effectively with individuals from varied backgrounds in a manner that was

safe and promoted the legitimacy of U.S. intervention in a foreign country. Polis Solutions recognized the similarities between this approach and the need for effective communication in a manner that is safe and promotes the legitimacy of the police. Thus, their training attempts to teach officers three core tenets of social interactions: Tact – procedural fairness, rapport building, self-control, and empathy; Tactics – delaying physical contact and limiting the reliance on physical force; and Trust – the need for creating a lasting positive impact on the citizens they contact.

Accordingly, T3 is a social-interaction-based training that focuses on developing officers' skills in decision-making, de-escalation, empathy, rapport building, and self-control. By building officers' skills in these areas, T3 aims to encourage officers to alter the trajectory of officer-citizen interactions away from the need to use force. The T3 program initially introduces officers to these concepts in a traditional classroom setting with the inclusion of examples and videos. However, the central focus of the program is to have officers move beyond this classroom instruction to more actively engage in decision exercises around these concepts.

To accomplish this active engagement, officers involved in T3 training are asked to observe videos of actual officer-citizen interactions (often derived from body camera footage) with set decision points built in. The videos show a portion of an interaction then automatically pauses at a pre-determined point. Officers are then asked to complete worksheets by writing what their priorities would be during the interaction at that moment. Importantly, officers are given a limited amount of time to write down their answers, simulating the need to make rapid decisions in the field. The videos are designed to be completed in a group setting—e.g., during roll call. After completing each decision point, officers are given about five minutes to discuss their views of the interaction with each other. The video then resumes until it reaches the next decision point (usually three per video). Each video exercise is designed to last about 45 minutes to allow for relatively

short training sessions during roll calls. This format limits the need for officers to be off the street for an extended period and allows for repeated training sessions over time. Delivering T3 during roll calls also eliminates the need to gather officers at a centralized training facility, which is a common obstacle that impedes agencies from pursuing many social interaction-based training programs, especially those that emphasize repeat, deliberate practice.

The video scenarios are facilitated by a department's own officers who have completed a train-the-trainer program led by Polis Solutions. Trainers are taught to concentrate group discussion during the stoppage points on the T3 principles which highlight procedurally-just communication skills, maintaining self-control during the encounter, and de-escalating the encounter by delaying physical contact with the subject until it is necessary to maintain either officer or citizen safety. The principles of T3 are reinforced to the officers undergoing training by departmental trainers throughout the program.

The T3 program differs from previous police social interaction training in several critical ways. First, the program was taught exclusively by law enforcement personnel with most sessions being taught by trainers from the police departments themselves.[2] Second, the program involved low intensity, high repetition training. While the training sessions took less than one hour to complete, officers participated in the training every other week for an extended time. These shorter sessions were also supplemented with half and full day training sessions that introduced additional concepts and scenario training opportunities (training length is discussed more below). Third, the program brought in an outside organization (Polis Solutions) to train selected officers on how to conduct T3 training in-house. Thus, the agencies participating in this program did not have to

[2] As described below, some sessions were taught by trainers from Polis Solutions. Polis Solutions trainers are current or former sworn law enforcement personnel, though they are from outside agencies rather than the specific training sites.

devote resources toward developing their own training programs. This approach is likely to be particularly appealing to smaller agencies that do not have the time or money to dedicate toward the development of their own innovative training programs. At the same time, it allows departments (regardless of size) to have greater ownership of the training because it is self-facilitated. Finally, Polis Solutions went to great lengths to base the T3 program and its principles on theory and empirical research from multiple disciplines (Wender, 2016; Wender & Lande, 2015) such as cognitive psychology (Gottman, 2011), linguistics (Damari et al., 2015; Logan-Terry & Damari, 2015), and social psychology (Cialdini, 1993).[3] As a result, the training is a good example of a practical application of Alpert and Dunham's (2004) asymmetric model. It focuses on teaching officers that their actions can impact the trajectory of citizen interactions and how they can leverage communication skills to alter interactions when needed.

Current Study

The present study evaluated Polis Solution's T3 program using a randomized controlled trial (RCT). Survey data were collected from experimental and control group officers at both research sites prior to and after the implementation of the training program. Additionally, official use of force reports from both agencies were collected from roughly one year prior to the implementation of the training program to one year after the completion of the training program. The survey data are used to examine changes in officers' attitudes and priorities during hypothetical officer-citizen encounters to determine if the training program impacted the importance officers placed on specific social interaction concerns. The use of force reports were then used to determine whether the training program successfully reduced the number of use of

[3] For an extended discussion of the components of the training program please see BLINDED FOR REVIEW.

force incidents among officers involved in the training program. The overarching goal of this study is to help build an evidence base of "what works" in police training.

Methods

Research Context

Data for this study come from a National Institute of Justice-funded evaluation of the T3 social interaction training program in two police departments in the United States, the Fayetteville (NC) Police Department (FPD) and the Tucson (AZ) Police Department (TPD). Both departments serve diverse populations, with FPD serving a racially diverse population that is primarily White (46%) and African American (42%) and TPD serving an ethnically diverse population that is primarily non-Hispanic White (47%) and Hispanic (42%). Additionally, both departments are sizeable but not excessively large; FPD had 164 patrol officers and TPD had 320 patrol officers at the time of the study. This is a critical point to make because the limited number of police training studies have occurred mostly in very large departments (e.g., Chicago PD). Studying a training program in FPD and TPD expands what is currently known about the impact of police training in an often-understudied population – medium-sized agencies. The generalizability of the results reported here is strengthened because FPD and TPD come from two very different regions of the United States with differently diverse populations, thereby providing distinct contexts under which the training program was evaluated.

Evaluating the T3 Program

T3 was developed by Polis Solutions for police departments to easily deliver social interaction training to officers during roll calls or other brief settings in whatever frequency fits with their operational tempo. The goal, of course, is to allow for repeated, low intensity trainings over long periods rather than one-time courses. We partnered with Polis Solutions in late 2016 to

evaluate their T3 program. Prior to doing so, Polis Solutions had provided the train-the-trainer program to several agencies across the United States, which included FPD and TPD, prior to the start of this study but had never subjected T3 to empirical scrutiny. With a cadre of officers trained to facilitate T3 in each agency, we developed an evaluation design. This resulted in a four component T3 course that we evaluated with an RCT. The first component was a one-hour introductory session where officers assigned to receive the T3 training were taught the core principles by departmental trainers. The second, and main, component involved repeated practice of these tenets through the video-based scenario sessions, referred to as "tactical-decision exercises (TDEs)" by Polis Solutions and throughout this paper. To avoid training fatigue but also to allow for repetition, the TDEs were delivered every other week at roll calls and involved officers viewing video footage of real police-citizen interactions. As discussed above, the videos automatically paused at pre-programmed points, and officers were instructed to answer a series of questions on a worksheet about what they would do in the situations based on what they knew at that point. After each stoppage point, the departmental trainers guided a discussion of the incident and focused on the core T3 principles. The third component was a 4-hour refresher course on the core tenets of T3 delivered by Polis Solutions personnel that occurred approximately three months into the training program. The final component was an 8-hour capstone session delivered by Polis Solutions personnel at the conclusion of the training program. Both the midpoint and conclusion sessions were designed to reinforce the key principles of T3 and ensure that the departmental trainers had conveyed the training in a manner consistent with Polis Solutions' intentions.

Design

Key to the evaluation of the T3 program was determining whether the training impacted officers' attitudes and behaviors regarding social interactions with citizens. The best method for

accomplishing this was to randomly assign patrol officers to either a treatment or control condition. Patrol officers were chosen because such officers attend roll call meetings at their district headquarters prior to going on (or after being on) patrol, providing an opportunity for training to be consistently delivered at the start or end of an officer's shift. The research team worked with departmental contacts to randomly assign officers to treatment and control conditions using each agency's patrol roster.[4] Across both agencies, a total of 224 officers were assigned to the treatment group and 227 were assigned to the control group. Officers were notified of their assignment to receive training approximately two months in advance of the first training to minimize absences from the training sessions.

The evaluation also attempted to assess whether training dosage impacted officer outcomes. To accomplish this, we worked with the departments to split the treatment sample into a high-dose group that would receive 6-months of T3 training (13 TDEs plus the refresher and capstone sessions) and a low-dose group that would receive 3-months of T3 training (7 TDEs plus the refresher session). Randomly assigning dosage at the officer level was determined to be impractical as it would involve too many different training sessions needing to be offered across officers. Instead, since training was to be delivered at roll calls within each patrol district, the best solution was to assign dosage at the district level. While there were too few districts to randomly assign these conditions, the authors worked with agency personnel to rank the districts within each agency by use of force rate. Researchers then alternated assigning high- and low-dosage to each

[4] Specifically, researchers went to both police departments and sat down with a member of the training department who was asked to bring a list of every patrol officer currently working in the agency. Researchers generated a list of the same length of random one's and zero's using statistical software and then merged the two lists. Officers with a one next to their name were allocated to the treatment group. Officers with a zero next to their name were allocated to the control group.

district. Thus, the dosage-level was not randomly assigned to districts, but the assignment of treatment or control was random.[5]

Researchers employed a pre-test post-test control group design in administering surveys to all patrol officers at FPD and TPD over the course of the study (Campbell & Stanley, 1963). All patrol officers were surveyed prior to the start of the program and after T3 training was completed in their jurisdictions. Accordingly, low-dose officers (and corresponding control group officers within the low dose districts) were surveyed at the beginning and end of the 3-month training assignment, and high-dose officers (and corresponding control group officers within the high dose districts) were surveyed at the beginning and end of the 6-month training assignment. In practice, this administration resulted in the experimental design depicted in Figure 1.

[Insert Figure 1 About Here]

Finally, official reports of uses of force were collected from both agencies. We asked both agencies to provide data for the year prior to the start of the training, the period the training occurred, and the year after training concluded. Accordingly, FPD provided 34 months of use of force data from March 2016 to December 2018. TPD provided 36 months of data from January 2016 to December 2018.[6] FPD and TPD generated unique identifiers so that officers involved in the use of force could be attributed to the treatment group, the control group, or as a non-patrol officer that was not affiliated with the study (e.g., a detective or supervisor), thus allowing for comparisons between groups and time periods in the number of use of force reports.

Procedure and Sample

[5] Officers in the control group were dismissed from the daily roll call prior to the delivery of T3 training to the treatment group officers, so there was no contamination in treatment delivery.

[6] TPD found it easier to query their system for three complete years rather than breaking it down to exactly one year prior to the start of the training. Since the inclusion of additional data was only beneficial, we included the two "extra" months in the analyses.

For the pre- and post-test survey data, members of the research team worked with agency personnel at both departments to attend roll calls for every shift in each patrol district over a two-week period to ensure that all patrol officers had an opportunity to be surveyed. Prior to survey administration, we provided a brief introduction to the project and the purpose of the questionnaire. We emphasized the anonymous nature of the survey and that honest feedback was necessary for a faithful evaluation of the training program. In other words, a finding that the training program did not work would not be a reflection on the officers or the agency, but on the training program itself. Officers were told that the survey was voluntary, all results would be reported in the aggregate, and that no one other than the researchers would have access to the raw data.

During the pre-test survey, 166 officers indicated that they had been randomly assigned to receive T3 training and 228 officers indicated that they were assigned to the control group. Accordingly, we were able to survey 74% of officers randomly assigned to the treatment condition. Some treatment group officers were missed due to days off, sickness, or other absences when the surveys were administered. Of those contacted, only two officers refused participation in the survey. Interestingly, more officers were in the control group than were originally assigned. This occurred due to the time gap between random assignment and the administration of surveys. Questionnaires were administered one week prior to the start of T3 training, but assignment occurred two months prior to the training program start date. Any officers who graduated from the academy or who were on long-term leave (e.g., military or maternity leave) during the assignment stage defaulted to the control group for the survey portion of the analysis. Still, a comparison of officers in the treatment and control groups at the pre-test revealed no significant differences in terms of gender, age, years of service, or race.

A meaningful amount of attrition was observed in the survey data throughout the study with 114 officers (representing 69% of the pre-test survey treatment group) indicating they were in the treatment group at the time of the post-test survey. Both departments experienced considerable turnover during the six-month study period, a problem that is common in police agencies. While it would have been preferable to collect identifiable officer data on the surveys so that their responses to the pre-test and post-test could have been matched together, the benefits of survey anonymity outweighed this option. On the questionnaires, officers were asked about their willingness to engage in a variety of behaviors and we felt anonymity was critical to receiving truthful responses rather than more socially desirable responses to these items.[7] Again, the treatment and control groups were compared, this time on the post-test (see Table 1). Respondents in each group were similar on gender, race/ethnicity, experience, rank, and military service. The only significant difference between the treatment and control group respondents at the post-test was in the age and education categories with the treatment group indicating that they were slightly older and more educated than the control group.[8]

In re-analyzing data from several criminal justice experiments, Berk and colleagues (2014) noted that the introduction of covariates did not substantively alter average treatment effect estimates when correlations between the covariate (here, age and education) and the outcome measures were greater than .4. For our study, these correlations were all non-significant with point estimates less than .1, suggesting that even though there were age and education differences due to attrition, it is unlikely these differences would impact the treatment effect estimates. In

[7] This is not to suggest that social desirability may not still be an issue, but rather that anonymity would decrease the amount of socially desirable answers that we received.

[8] As a robustness check, we estimated models including these demographic variables as controls in case they were related to the outcomes of interest. These models provided the same conclusions as the ones presented below and are available for review upon request to the lead author.

considering why these differences appear, they are likely attributable to the addition of new officers to the control group who graduated from the academy during the 6-month study period at both research sites.

[Insert Table 1 About Here]

Attrition likely affected the official use of force reports as well. However, unlike with the survey design where identifiers were not used, the official data did contain random unique identifiers tracking officers throughout the study period. While officers could still have transferred to other roles or left the department, this distribution should not have been correlated to the random assignment of treatment. Furthermore, officers graduating from the academy or returning from long-term leave would not have been assigned a unique identifier, and therefore, would not be included in the use of force analysis as a member of the treatment or control group. Thus, by employing an intent-to-treat design for the official use of force report analysis, we can be confident that the attrition does not threaten the validity of the findings.

Measures

Survey data.

Three survey measures were critical to the evaluation of the effectiveness of the T3 program. These measures were intended to tap into the priorities that an officer places on various aspects of a citizen encounter. To do so, our survey employed a vignette involving a hypothetical encounter with a citizen (see Appendix A). The officer responds to a vague suspicious person call in the vignette. The respondent was then asked how important several priorities would be in the ensuing interaction. Several survey questions were used to tap into the key tenets of the T3 program. *Procedural justice priorities* contained eight items that related to communicating and building rapport with the subject (α=0.88). *Maintaining self-control* contained seven items that

related to the officer remaining calm and thinking through his/her options (α=0.77). Finally, *physical control priorities* contained two items that related to the physical restraint of the suspect (α=0.60, r=0.42).[9] See Appendix B for a full list of the items contained in each scale.

This vignette approach to measuring the major outcomes in the survey data was chosen as a compromise between the use of attitudinal data and behavioral data. Uses of force are relatively rare among police officers (Adams, 1999). Furthermore, the decision to use force is influenced not only by the approaches highlighted in the T3 training program, but also by aleatory factors such as the context of the situation (Alpert & Dunham, 2004; Shjarback, 2018). As a result, the behavioral effects of the training are likely to be weaker than any attitudinal effects. Still, previous evaluations of training programs have found that officers may not report attitudinal changes while still demonstrating changes to their behavior (e.g., Lonsway et al., 2001). By measuring hypothetical behavior, as opposed to attitudes towards uses of force and de-escalation, we attempted to move past attitudinal changes by examining the psychological approach an officer would take to a hypothetical situation. Furthermore, the use of the hypothetical vignette allowed us to control the proposed situation and not allow outcomes to be influenced by differences in the circumstances of a situation (e.g., the suspect's behavior). In prior research, questions have been raised about the validity of attitudinal measures in accurately assessing behaviors (see e.g., Scott & Willits, 1994), however, studies have shown hypothetical choices in vignettes to have good predictive validity (see e.g., Alexander & Becker, 1978; Hainmueller et al., 2015). Vignettes also have been previously used in a variety of policing contexts to assess hypothetical officer behavior

[9] Cronbach's alpha is related to the number of items in a scale (Carmines & Zeller, 1979), so the low alpha value for physical control priorities is not unexpected. For this reason Pearson's r is also presented as a measure of the correlation between the two items.

(McLean, 2019; Nix, Picket, & Mitchell, 2019; Nix, Pickett, Wolfe, & Campbell, 2017; Phillips, 2009).

Official use of force data.

While the results from the survey-based vignette analyses will provide useful information, the data are still limited in that they do not tap into actual officer behavior. Ultimately, T3, like any other police training program, is intended to impact officer behavior. For the behavioral outcome analysis, databases for every use of force reported in both police departments from the time periods identified earlier were obtained.[10] Officer names were replaced with random, unique identifiers that could then be used to determine if an officer was in the treatment group, the control group, or not included in the study.[11] With these databases compiled, we constructed a new, time-series database that indicated the number of use of force incidents in a given month that treatment or control group officers were involved in. Accordingly, an incident involving two treatment officers would be scored as a 1 for the treatment group, since it is a single incident. Further, an incident involving a treatment officer and a control group officer would be scored as a 1 for both groups, since it was a single incident, but both a treatment and a control group officer was involved. This type of incident indicates a potential for contamination; however, in a study of a long-term training program, it is virtually impossible to avoid contamination in use of force incidents.[12]

[10] While a use of force incident does not, on its own, indicate a failure to de-escalate, our use of random assignment allows us to assume that over the three years of use of force data that we have access to, across the groups of treatment and control, officers will have experienced the same number of opportunities to de-escalate a use of force situation on average. Thus, an incident resulting in a use of force is not indicative of a failure in that incident or by that officer, but on average, we would expect to see fewer use of force incidents across the entire treatment group if they were more successful at de-escalation as a result of the training.

[11] Only patrol officers were included in the study, so officers assigned to other roles would not be included in the random assignment and are also not included in the analysis. This further ensures rigor in the analysis of official use of force reports because officers assigned to other roles (e.g., administrative duty or specialized field units) would have very different opportunities to be involved in use of force incidents.

[12] For comparison, a study currently being conducted by White et al. (2019) uses block randomization to randomize entire squads rather than individual officers to avoid contamination. This is possible in their study because training is being conducted at a single time-point, so a control group squad need only cover the squad's normal duties for the single training period. It would be considerably more difficult to randomize at this level in our study, as the squad's

Once the total number of incidents in a given month was determined, it was converted into a rate where the number of incidents was divided by the number of officers assigned to each group across the entire agency (low-dose treatment, high-dose treatment, and control) and then multiplied by 10. The values can be interpreted as the rate of use of force incidents per 10 officers in the group. We did not combine the agencies' data into a single database because the departments had substantial differences in their definitions of reportable uses of force. Most notable were differences in the need to report uses of force at lower levels. For example, in Tucson, any time an officer conducted a "takedown" on a citizen officers were required to report the incident as a use of force. However, in Fayetteville, the incident would only require a report if the citizen was injured. Given these differences, it was more appropriate to analyze the agencies separately.

Descriptively, this measurement strategy produces two databases – one for Fayetteville PD and one for Tucson PD. The Fayetteville PD database contains 34 months of data (March 2016 to December 2018). Officers involved in the study – that is, officers who were randomly assigned to either treatment or control in February 2016 – accounted for 119 reportable use of force incidents. Of these 119 incidents, 12 (10.1%) were "contaminated" in that officers from multiple groups were involved in the incident. The Tucson PD database contains 36 months of data (January 2016 to December 2018). Officers involved in the study accounted for 1,024 use of force incidents during this time-period with 190 (18.6%) "contaminated" incidents.

Analytic Strategy

Survey analysis.

normal duties would need to be covered by a control group squad every other week for six months. Thus, the White et al. (2019) study is less likely to be affected by contamination but does not contain the repetitive practice principle that is a key to the T3 program. Similarly other RCTs involving police departments, such as body-worn camera studies, have been able to minimize contamination by randomizing at the shift level. That is, an officer will wear a camera on his shift on Tuesday, but not during his shift on Wednesday, for example. This is obviously not possible in an RCT of police training as training concepts and practice are not something that can be removed once they have been conducted.

In RCTs, complex regression models with a series of covariates are not only overly complicated but may, in fact, inflate standard errors and bias estimates (Berk et al., 2014; Freedman, 2008). As a result, to evaluate the T3 program using the survey data, a series of difference-in-difference tests were conducted on the three scales identified above. Difference-in-difference scores examine whether individuals in the treatment group experienced the same changes from pre-test to post-test as individuals in the control group. In other words, it assumes that had the treatment group not been treated, they would have experienced the same changes from pre-test to post-test as the control group. If there is a difference in the changes between the treatment group and the control group, then a treatment effect is found (Lechner, 2011; Meyer, 1995). This is a common strategy used in economics that also has been successfully applied to criminological issues (see Branas et al., 2011; Smith & Petrocelli, 2018).

To generate the difference-in-difference estimates, a regression approach was utilized. Specifically, if we let Y_{igt} represent an officer i in group g at the time of survey t, where $T_{g(i)}$ is 1 if the officer is treated and $d_{t(i)}$ is 1 if the observation is in the post-period, then the difference-in-difference estimator can be found using a regression equation specified as:

$$Y_{igt} = \beta_0 + \beta_1 T_{g(i)} + \beta_2 d_{t(i)} + \beta_3 \left(T_{g(i)} \text{ X } d_{t(i)}\right) + \epsilon_{igt}$$

Since the difference-in-difference (DiD) value is conceptually,

$$DiD = \left(\overline{Treatment_{post}} - \overline{Treatment_{pre}}\right) - \left(\overline{Control_{post}} - \overline{Control_{pre}}\right)$$

the results of the regression equation can be used to estimate the difference-in-difference (DiD):

$$DiD = \left((\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_1)\right) - \left((\beta_0 + \beta_2) - \beta_0\right) = \beta_3$$

In other words, because the other terms cancel out, the difference-in-difference estimator can be found in the coefficient of the interaction term of treatment ($T_{g(i)}$) and time ($d_{t(i)}$). Further, the use of linear regression is advantageous here because it allows for the estimate of standardized coefficients which are potentially more meaningful in understanding the effect size, as well as estimates of significance that would not be available in the conceptual equation. In the first set of analyses, three models, one for the full sample and one for each of the two research sites, were generated using this method to test for a treatment effect. Since T3 encourages officers to communicate, develop rapport, think through their options, and delay physical contact until necessary, we would expect the difference-in-difference estimator would demonstrate that officers receiving training would:

1) Have higher prioritization of procedurally-just communication
2) Have higher prioritization of maintaining self-control, and
3) Have lower prioritization of physical control

The second set of survey analyses examined the differing effects of experiencing high dose and low dose treatment. To do so, the difference-in-difference regressions were altered so that there were two interaction terms for each level of dosage. Accordingly, if we use the same annotation as before and $L_{g(i)}$ is 1 if the officer is in the low dose treatment group (and 0 if the officer is in the high dose treatment group or the control group) and $H_{g(i)}$ is 1 if the officer is in the high dose treatment group (and 0 if the officer is in the low dose treatment group or the control group), then:

$$Y_{igt} = \beta_0 + \beta_1 L_{g(i)} + \beta_2 H_{g(i)} + \beta_3 d_{t(i)} + \beta_4 \left(L_{g(i)} \text{ X } d_{t(i)}\right) + \beta_5 \left(H_{g(i)} \text{ X } d_{t(i)}\right) + \epsilon_{igt}$$

Using the same proof as before, the difference-in-difference estimator for low-dose treatment is $\beta_{4,}$ and the difference-in-difference estimator for high-dose treatment is β_5. Importantly, these estimators will reveal significant differences between receiving low-dose or high-dose treatment and being in the control group—not significant differences between levels of treatment. This approach was chosen because the designation of high-dose or low-dose treatment was not randomly assigned. Any treatment effects found in these estimators can be compared to see if there is evidence that dosage matters, but definitive conclusions regarding differences in dosage will not be drawn from these analyses. Again, three models were estimated, one for the full sample and one for each of the two research sites.

Finally, a sensitivity analysis was conducted to determine whether treatment effects were limited to specific districts. To do this, the difference-in-difference models were re-estimated on samples limited to officers in a given patrol district. This approach was used because treatment was delivered during roll calls within the districts that officers patrol. Officers discussed the video scenarios with other officers in their district making variance in effects by district plausible.

Official use of force data analysis.

To analyze the official use of force reports, the time-series database outlined above was subjected to a series of interrupted time-series analyses using Stata's `itsa` command suite (see Linden, 2015). Newey-West standard errors were used to handle autocorrelation due to the time-series nature of the data, and the Cumby-Huizinga test was used to ensure that the correct lag was estimated for each model (Cumby & Huizinga, 1992; Linden, 2015). This approach is consistent with a difference-in-difference approach with the model examining multiple groups (low or high-dose treatment compared to control) and an intervention point for the month in which training began. Coefficients can then assess whether there were significant differences between the groups

at pre-test, significant differences at post-test, and a difference in the difference between pre- and post-tests.

Results

Survey Analysis

The results of the first set of difference-in-difference estimators can be found in Table 2. There is a significant treatment effect on procedural justice priorities in both the full sample ($\beta = .12$, $p < .05$) and in the Fayetteville sample ($\beta = .24$, $p < .01$). Despite these promising results, there were no significant results in the Tucson sample. Consider that the non-significant effects for maintaining self-control and physical control priorities in the Fayetteville sample is in the expected direction and may simply reflect a power issue given the smaller effect size compared to the other outcomes (there are only 108 officers in the entire Fayetteville post-test data).[13] In the Tucson sample, however, none of the effect sizes is greater than |0.04|, so a lack of power is unlikely to be the cause of the absence of effects among officers in that agency.

Figure 2 presents the full sample difference-in-difference changes graphically. The chart on the far left shows the significant effect for procedural justice priorities. Specifically, while procedural justice priorities decreased from the pre-test to the post-test for the control group, it increased for the treatment group. With respect to maintaining self-control, the control group again decreased, but the treatment group remained constant, rather than increasing. These changes were, however, not significant in the difference-in-difference analyses presented in Table 2. Finally, for physical control priorities, both the control group and the treatment group decreased from pre-test to post-test. The treatment group decreased at a slightly greater rate, but again, this change is not significant in the difference-in-difference analysis.

[13] A post-hoc power analysis for both estimates suggest that if the estimated effect size is accurate, we would only have 55% and 48% chance of detecting a significant effect in the reduced sample size.

[Insert Table 2 and Figure 2 About Here]

Next, Table 3 presents the difference-in-difference estimators with dosage considerations. In the full sample model there is a significant treatment effect for procedural justice priorities (β = .12, $p < .05$) and maintaining self-control (β = .11, $p < .05$) when receiving the lower dosage of treatment, and there is a significant treatment effect for physical control priorities (β = -.13, $p < .05$) when receiving the higher dosage of treatment. In the Fayetteville sample, there are significant treatment effects for procedural justice priorities (β = .21, $p < .05$) when receiving the lower dosage and physical control priorities (β = -.22, $p < .05$) when receiving the higher dosage of treatment. Additionally, note that the effect size for maintaining self-control in the Fayetteville sample is consistent with the effect size in the full model, but is not significant due to the reduced power of the smaller sample size. Finally, there are again no significant effects in the Tucson sample. However, unlike last time, the estimate for maintaining self-control when receiving the lower dosage of treatment does have an effect size that is not far removed from the full sample effect size (β = 0.09 compared to β = 0.11). The other effects are again small and non-significant.

Figure 3 presents these analyses graphically. The chart on the far left again shows the procedural justice priorities effects. There is a clear demonstration that the low-dose treatment group saw a substantial (and statistically significant) increase in their procedural justice priorities scores while the control and high-dose treatment groups remained relatively stable. A similar finding is seen in the middle chart for maintaining self-control. The high-dose treatment and control groups have very similar lines, while the low-dose treatment group experienced a substantial (and statistically significant) increase. Finally, the far-right chart shows the treatment effect on physical control priorities. While there is a difference in the overall levels between the low-dose treatment group and the control group, their trends are relatively stable. The high-dose

treatment group, on the other hand, experienced a large reduction in physical control priorities from the pre-test to the post-test that is statistically significant.

[Insert Table 3 and Figure 3 About Here]

Sensitivity analysis.

The last set of survey data analyses examined variations in the effect size across the seven different districts in Tucson and Fayetteville (see Table 4). Due to further reductions in sample size, it is not surprising that only two significant effects – the treatment effect on procedural justice priorities in Fayetteville's Central District ($\beta = .33$, $p < .05$) and the treatment effect on physical control priorities in Fayetteville's Campbellton District ($\beta = -.38$, $p < .05$) – were found. What is interesting, however, is the differences in effect sizes seen across the districts. The previous models suggested that Tucson did not experience any treatment effects. However, this analysis suggests that some of Tucson's patrol divisions experienced effects. Specifically, Tucson's East Division had point estimates similar to Fayetteville for procedural justice priorities and maintaining self-control though neither of these effects were statistically significant. Similarly, Tucson's South Division had relatively similar point estimates for procedural justice priorities and physical control priorities compared to Fayetteville's districts, though again neither were statistically significant. While it is difficult to draw conclusions from this analysis due to the reduced sample size, it seems apparent that the effects of training are likely dependent on the circumstances of the department and even the district where training is delivered.

These results are depicted graphically in Figure 4 as the standardized effect sizes for each outcome are plotted by district. Several conclusions are demonstrated in this figure. First, the effect of treatment on procedural justice is clearer than the other two effects as five of the seven points are on the hypothesized side of the line at 0.0, which represents no effect. Second, the points for

the Fayetteville districts tended to be further from 0.0 on the hypothesized side of the line (i.e., the effect sizes were larger in Fayetteville). Finally, Tucson's Midtown and West divisions clearly experienced no effect, while Tucson's South and East divisions showed similar effects to those observed in Fayetteville.

[Insert Table 4 and Figure 4 About Here]

Official Use of Force Data

The first set of interrupted time-series models examines the effect of treatment in the Fayetteville Police Department (see Figures 5 and 6). Reportable use of force incidents are rare in Fayetteville (n=119 for the entire study period) and neither model suggests that there are any differences between the treatment (either low-dose – Figure 5 or high-dose – Figure 6) and control groups or from before to after the training program began. Instead, use of force incidents appear to have remained relatively static across groups and time-periods (see Supplementary Materials for non-significant regression coefficients).

[Insert Figures 5 and 6 About Here]

The second set of interrupted time-series models examines the effect of treatment in the Tucson Police Department (see Figures 7 and 8). The results indicate that the control group experienced a statistically significant drop in the rate of reportable use of force incidents immediately following the introduction of T3 training ($b = -0.56$, $p < 0.05$). Visually, the treatment group for both high and low-dose training appears to have experienced a similar drop. While there is no specific coefficient for this drop in the multiple-group ITSA presented in the Supplementary Materials, a follow-up single-group ITSA was estimated to see if this drop was also significant. The results indicated that the decrease in use of force reports in the treatment groups was not statistically significant.

[Insert Figures 7 and 8 About Here]

Discussion and Policy Implications

Calls for improved police training always seem to follow any controversial police-citizen encounter. This is especially true when the incident involves the use of force. Our study was the first RCT of its kind designed to evaluate a long-term, repetitive police social interaction training program, and the survey results indicated the training showed promise in several ways. First, procedural justice priorities saw a significant treatment effect for officers receiving any treatment in the full sample. However, it is important to note that the magnitude of this effect size was rather modest. Second, when broken down by dosage, each outcome experienced a significant treatment effect: low-dose treatment resulted in improved procedural justice priorities and an increased emphasis on maintaining self-control, and high-dose treatment resulted in a de-prioritization of physical control. However, the behavioral analysis did not support the survey analysis with no significant differences detected in the number of reportable use of force incidents in Fayetteville or Tucson attributed to the T3 training program.

Despite these mixed findings, the results are encouraging as we now have further evidence that officers can be trained in social interaction skills with an eye toward using procedural justice and related concepts. We also found treatment effects at different levels of dosage for each of the three outcomes. This suggests that officers can be trained to see the potential value of establishing rapport, displaying empathy, and communicating with a subject before resorting to establishing physical control. As noted in previous literature, the term de-escalation has lacked a clear definition. However, prior work has suggested that it typically includes the prioritization of communication – especially procedurally-fair communication – and a de-emphasis on physical control (Todak, 2017; Todak & James, 2018). Taking such an approach may lead to less frequent

uses of force as officers start interactions on a trajectory that is likely to lead away from the use of force.

With that said, several sensitivity analyses in our study demonstrated that context may influence the degree of success obtained from T3. As discussed earlier, some patrol districts experienced better results from the training than others. This is not entirely surprising given the established literature on police subculture that suggests officer attitudes and behavior can vary by their immediate, small work group (e.g., shift or squad; see Klinger, 1999; Ingram, Paoline, & Terrill, 2018; Ingram, Terrill, & Paoline, 2018). This effect may have been further amplified in the present study where the training program included a significant discussion component where officers talked to other officers in their workgroup about how they would handle potentially fractious officer-citizen encounters. At the same time, it is also important to acknowledge that our data and analyses are limited in their ability to speak to what caused differences in training success by research site, dosage, and patrol district. These findings could have been the result of differences in training delivery, departmental culture, the workgroup environment, or some other unmeasured factor. Still, we would strongly recommend that departments have a solid understanding of how officers in certain divisions, shifts, squads, or other small groups may react to training to help departments tailor the programs (or set up strategies to "sell it") to improve officer motivation to train (see also, Wolfe et al., 2019). In short, police training is not a one size fits all solution, and the T3 program is no exception.

The findings from the comparisons of length of training provided additional context to the main findings. As noted previously, the high-dose and low-dose treatment groups experienced different attitudinal effects. Those officers that completed six months of T3 training—the high dose group—saw a treatment effect consistent with the de-prioritization of physical control during

the hypothetical scenario. However, we did not observe a treatment effect with respect to their prioritization of procedural justice communication or maintaining self-control in the scenario. In comparison, low dose officers experienced treatment effects for procedural justice communication and maintaining self-control. These divergent effects could reflect differences in the focus of the training for the second half of the program. Members of the research team observed training once a month for the duration of the program. In our observations, the focus of the training sessions seemed to shift from procedurally-just communication and self-control in the first three months to physical control in the final three months. Thus, the divergent results may reflect differences in the topics recently covered in training, though this in itself may be a troubling indicator that the effects of training erode over time as the content of the first three months was consistent across both groups. Still, we would be remiss to not reiterate that the assignment of high-dose and low-dose treatment was not random and we do not have a reliable method for estimating how much these differences may be caused by differences in training content, training delivery, or even workgroup culture. Thus, it is impossible to disentangle potential dosage effects from the effects of district context. Despite being limited in these conclusions we encourage future researchers to examine dosage in their officer training evaluations. Such information will prove valuable as we search for the optimal level of training for specific situations. This will help target officer outcomes in the most time- and cost-efficient manner.

With regards to the behavioral findings, Fayetteville PD did not see any reductions in reported use of force incidents. This may be due, in part, to the small number of reportable use of force incidents in Fayetteville prior to the start of the training program. That is, since there were fewer incidents reported, it is more difficult to detect reductions. Tucson PD saw reductions in the number of reported uses of force in both the treatment and control groups though the treatment

group's reduction was not statistically significant. This simultaneous reduction may be due to the effects of contamination or may be spurious due to changes that we were unaware of occurring in the agency at the time the training program started. As a result, we must conclude there is no clear evidence the T3 program reduced reportable use of force incidents at either police department involved in the study.

One of the primary reasons police training is rarely subjected to evaluation is that it is difficult to do so. While using an RCT was the most rigorous way to evaluate the T3 program, such a design came with unavoidable limitations. The main limitation we faced was attrition. Random assignment had to be done far enough in advance of the beginning of the training to allow adequate time for scheduling concerns to be addressed. Accordingly, during the period between random assignment and the start of training a meaningful number of treatment officers were reassigned to specialized units, promoted, or placed on leave (e.g., maternity or military leave). We also had a number of officers experience similar career changes during the observation period, which resulted in a few more cases of attrition. While this attrition should be taken into consideration when interpreting our results, we have no evidence that the officers were different from those that remained in the treatment and control group aside from the latter being slightly younger and more likely to have an Associate's degree rather than a Bachelor's degree. It would have been easier to conduct a full RCT without attrition in a laboratory setting or by surveying participants immediately before and after a single training session. However, this is likely not the best way to accomplish police training nor does it tell us if observed differences last more than a day. A project like ours will be subject to unavoidable attrition, but still provides us with evidence about the success of training programs.

Our second limitation is the effect of contamination on the results. Contamination of treatment and control group officers occurred in 10% of use of force incidents in Fayetteville PD and 18% of incidents in Tucson PD. However, while contamination presents a threat to the validity of the results, it again, needs to be reiterated that this evaluation is representative of how training is conducted in real-world conditions. There were limited numbers of shifts in the two police departments and the opportunity for use of force incidents necessarily vary according to these shifts (e.g., use of force incidents are less likely at 8 am on a Monday morning than at 9 pm on a Friday night), so randomization at the shift level was not realistic. Thus, contamination was unavoidable if treatment was to be randomly assigned.[14]

Moving forward we hope this study helps stimulate more evaluation research on police training. Forming a strong evidence base regarding what works in police training will help police departments make more informed decisions when pursuing training for their officers. Ideally, we would parse out exactly which components of the training were successful and which were not, but due to the research design (a program-level RCT), we know only the effects of the larger program but not the effects of its individual parts. Identifying successful elements of a police training program requires a large body of evidence on diverse training programs to see the common elements of successful (and unsuccessful) programs. By generating this body of evidence, the systematic reviews of police training that have either failed or relied on work outside of the field of policing (Engel, McManus, & Herold, 2020; Huey, 2018) can be revisited and provide stronger conclusions. These stronger conclusions will allow police departments to more responsibly

[14] We also considered removing contaminated incidents from the analysis to determine their effects on the results. However, T3 encourages officers to prioritize delaying physical contact. Thus, contamination may, in and of itself, be related to training, as officers going through T3 are more likely to delay an encounter so that more officers can arrive on scene. Unfortunately, the data we have do not indicate who was first on scene to see if contaminated incidents were the result of T3 officers on scene delaying encounters or control group officers delaying encounters.

allocate scarce police training resources by selecting training that has the greatest potential to improve officer and citizen safety.

Establishing a larger evidence base will require two things. First, researchers should seek to partner with police agencies to evaluate training they currently have or wish to pursue. Depending on the scale, this likely will require a stronger commitment from the federal government or other funding agencies to provide support for such practitioner-researcher partnerships (Hansen, Alpert, & Rojek, 2014; Rojek, Smith, & Alpert, 2012). Another option may be for agencies to evaluate their own training by leveraging "pracademics" (i.e., in-house officers with research expertise) in their own departments (see, e.g., NIJ's LEADS Scholar program). Second, and relatedly, we need police executives and officers to be open-minded about the virtues of evaluating police training. The Fayetteville and Tucson officers that helped us coordinate and participate in our evaluation clearly are at the forefront of policing innovation and evidence-based decision making. An evaluation of this scale would not have been possible without our partnership. At this point we know very little about what works in police training, but if chiefs and sheriffs are receptive to the idea of allowing training evaluations in their agencies, we can begin accumulating a solid base of evidence. We are confident that many police executives will be eager to pursue such opportunities.

In the end, we hope our study has provided researchers with two important conclusions: (1) social interaction training programs show promise, and (2) moving from the realm of theory to a realistic training program that can be implemented in agencies across the country is extremely difficult, but certainly achievable with the appropriate resources. In the case of T3, this study has found that officers are receptive of the training (BLINDED FOR REVIEW), that their attitudes could be improved, but that this did not lead to fewer reportable use of force incidents.

Theoretically-grounded training and evaluation are necessary to provide police executives with the best information possible when they are adjusting their own training curricula or pursing new options. Evidence-based police training will save money and, more importantly, improve officer and citizen safety.

Conflict of Interest Statement

The authors hold no conflicts of interest in evaluating the training program in this study. The authors acted as independent evaluators whose data collection efforts were separate from the program delivery by Polis Solutions.

References

Alexander, C.S. & Becker, H.J. (1978). The use of vignettes in survey research. *Public Opinion Quarterly, 42*, 93-104.

Alpert, G.P. (1988). Police pursuits – Linking data to decisions. *Criminal Law Bulletin, 24*, 453-468.

Alpert, G.P. & Dunham, R.G. (2004). *Understanding police use of force: Officers, suspects, and reciprocity*. New York, NY: Cambridge University Press.

Baddeley, A.D., & Longman, D.J.A. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics, 21*, 627-635.

Berk, R., Pitkin, E., Brown, L., Buja, A., George, E., & Zhao, L. (2013). Covariance adjustments for the analysis of randomized field experiments. *Evaluation Review, 37*, 170-196.

Branas, C.C., Cheney, R.A., MacDonald, J.M., Tam, V.W., Jackson, T.D., & Ten Have, T.R. (2011). A difference-in-differences analysis of health, safety, and greening vacant urban space. *American Journal of Epidemiology, 174*, 1296-1306.

Campbell, D.T. & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Belmont, CA: Wadsworth.

Carmines, E.G. & Zeller, R.A. (1979). *Reliability and validity assessment.* Thousand Oaks, CA: Sage Publications, Inc.

Cialdini, R.B. (1993). *Influence: Science and practice* (3rd ed.). New York, NY: HarperCollins.

Cumby, R.E. & Huizinga, J. (1992). Testing the autocorrelation structure of disturbances in ordinary least squares and instrumental variables regressions. *Econometrica, 60*, 185-195.

Damari, R.R., Rubin, G., & Logan-Terry, A. (2015). Navigating face-threatening terrain: Questioning strategies in cross-cultural military training scenarios. *Procedia Manufacturing, 3*, 4090-4097.

Engel, R.S., McManus, H.D., & Herold, T.D. (2020). Does de-escalation training work? A systematic review and call for evidence in police sue-of-force reform. *Criminology & Public Policy,* DOI: 10.1111/1745-9133.12467.

Freedman, D.A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics, 40*, 180-193.

Fridell, L. (2013). This is not your grandparents' prejudice: The implications of the modern science of bias for police training. *Translational Criminology, 5*, 10-11.

Fridell, L.A. (2016). Racial aspects of police shootings: Reducing both bias and counter bias. *Criminology and Public Policy, 15*, 481-489.

Gottman, J. M. (2011). *The science of trust: Emotional attunement for couples*. New York, NY: WW Norton & Company.

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences of the United States of America, 112*, 2395-2400.

Hansen, J. A., Alpert, G. P., & Rojek, J. J. (2014). The benefits of police practitioner–researcher partnerships to participating agencies. *Policing: A Journal of Policy and Practice, 8*, 307-320.

Hansson, L., & Markström, U. (2014). The effectiveness of an anti-stigma intervention in a basic police officer training programme: A controlled study. *BMC Psychiatry, 14(1)*, 55-63.

Huey, L. (2018). *What do we know about in-service police training? Results of a failed systematic review*. Western University: Sociology Publications. Retrieved from: https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1043&context=sociologypub.

Ingram, J. R., Paoline, E. A., & Terrill, W. (2013). A multilevel framework for understanding police culture: The role of the workgroup. *Criminology, 51*, 365-397.

Ingram, J. R., Terrill, W., & Paoline, E. A. (2018). Police culture and officer behavior: Application of a multilevel framework. *Criminology* DOI: 10.1111/1745-9125.12192.

Krameddine, Y., DeMarco, D., Hassel, R., & Silverstone, P. H. (2013). A novel training program for police officers that improves interactions with mentally ill individuals and is cost-effective. *Frontiers in Psychiatry*, *4*, 1-10.

Lechner, M. (2010). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics, 4*, 165-224.

Linden, A. (2015). Conducting interrupted time-series analysis for single- and multiple-group comparisons. *The Stata Journal*, *15*, 480-500.

Logan-Terry, A., & Damari, R. R. (2015). Key culture-general interactional skills for military personnel. *Procedia Manufacturing*, *3*, 3990-3997.

Lonsway, K.A., Welch, S., & Fitzgerald, L.F. (2001). Police training in sexual assault response: Process, outcomes, and elements of change. *Criminal Justice and Behavior, 28*, 695-730.

McLean, K. (2019). Revisiting the role of distributive justice in Tyler's legitimacy theory. *Journal of Experimental Criminology*, DOI: 10.1007/s11292-019-09370-5.

Meyer, B.D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics, 13*, 151-161.

Nagin, D.S., & Telep, C.W. (2017). Procedural justice and legal compliance. *Annual Review of Law and Social Science, 13*, 5-28.

Nix, J., Pickett, J.T., & Mitchell, R.J. (2019). Compliance, noncompliance, and the in-between: Causal effects of civilian demeanor on police officers' cognitions and emotions. *Journal of Experimental Criminology, 15*, 611-639.

Nix, J., Pickett, J.T., Wolfe, S.E., & Campbell, B.A. (2017). Demeanor, race, and police perceptions of procedural justice: Evidence from two randomized experiments. *Justice Quarterly, 34,* 1154-1183.

Owens, E.G., Weisburd, D., Alpert, G., & Amendola, K.A. (2016). *Promoting officer integrity through early engagements and procedural justice in the Seattle Police Department.* Washington, DC: Police Foundation.

Phillips, S.W. (2009). Using a vignette research design to examine traffic stop decision making of police officers: A research note. *Criminal Justice Policy Review, 20,* 495-506.

President's Task Force on 21st Century Policing. (2015). *Final report of the President's Task Force on 21st Century Policing.* Washington, DC: Office of Community Oriented Policing Services.

Reaves, B.A. (2016). *State and local law enforcement training academies, 2013.* Bureau of Justice Statistics. Retrieved from: https://www.bjs.gov/content/pub/pdf/slleta13.pdf.

Rojek, J., Smith, H. P., & Alpert, G. P. (2012). The prevalence and characteristics of police practitioner–researcher partnerships. *Police Quarterly, 15,* 241-261.

Robertson, A., McMillan, L., Godwin, J., & Deuchar, R. (2014) *The Scottish Police and Citizen Engagement (SPACE) trial: Final report.* Glasgow, UK: Glasgow Caledonian University.

Rosenbaum, D.P., & Lawrence, D.S. (2013). *Teaching respectful police-citizen encounters and good decision making: Results of a randomized control trial with police recruits.* Washington, DC: National Institute of Justice.

Schaefer, B., & Hughes, T. (2016). *Honing Interpersonal Necessary Tactics (H.I.N.T.): An evaluation of procedural justice training.* Louisville, KY: Southern Police Institute, University of Louisville.

Scott, D. & Willits, F.K. (1994). Environmental attitudes and behavior: A Pennsylvania survey. *Environment & Behavior, 26,* 239-260.

Sherman, L.W. (2013). The rise of evidence-based policing: Targeting, testing, and tracking. *Crime and Justice, 42,* 377-451.

Shjarback, J. (2018). "Neighborhood" influence on police use of force: state-of-the-art review. *Policing: An International Journal, 41*(6), 859-872.

Skogan, W.G., & Frydl, K. (2004). *Fairness and effectiveness in policing: The evidence*. Washington, DC: National Academies Press.

Skogan, W.G., Van Craen, M., & Hennessy, C. (2015). Training police for procedural justice. *Journal of Experimental Criminology, 11,* 319-334.

Smith, M.R., & Petrocelli, M. (2019). The effect of concealed handgun carry deregulation in Arizona on crime in Tucson. *Criminal Justice Policy Review, 30,* 1186-1203.

Sutton, R.M., Niles, D., Meaney, P.A., Aplenc, R., French, B., Abella, B.S., Lengetti, E.L., Berg, R.A., Helfaer, M.A., & Nadkarni, V. (2011). Low-dose, high-frequency CPR training improves skill retention of in-hospital pediatric providers. *Pediatrics, 128,* e145-e151.

Tyler, T.R. (2006). *Why people obey the law.* Princeton, NJ: Princeton University Press.

Walters, G.D. & Bolger, P.C. (2018). Procedural justice perceptions, legitimacy beliefs, and compliance with the law: A meta-analysis. *Journal of Experimental Criminology,* DOI: 10.1007/s11292-018-9338-2.

Wender, J. (2016). Enhancing officers' trust-building and tactical skills. *BJA NTTAC TTA Blog*. Accessed at https://www.bjatraining.org/media/blog/enhancingofficers%E2%80%99-trust-building-and-tactical-skills.

Wender, J. & Lande, B. (2015). Tact, Tactics, and Trust: Building the foundations for engagement-based policing. In *Engagement-Based Policing: the What, How, and Why of Community Engagement* (pp. 15-28). Major Cities Chiefs Association.

Wheller, L. & Morris, J. (2010). *Evidence reviews: What works in training, behaviour change and implementing guidance.* London, UK: National Policing Improvement Agency.

Wheller, L., Quinton, P., Fildes, A., & Mills, P.C.A. (2013). *The Greater Manchester Police procedural justice training experiment: The impact of communication skills training on officers and victims of crime.* Coventry, UK: College of Policing.

White, M.D., Engel, R.S., Alpert, G.P., Isaza, G., McManus, H.D., Herold, T.D…Orosco, C. (2019). *An overview of ongoing de-escalation training program evaluations.* Presented at the 2019 Annual Meeting of the American Society of Criminology. San Francisco, CA.

Wolfe, S.E., McLean, K., Rojek, J., Alpert, G.P., & Smith, M.R. (2019). Advancing a theory of police officer training motivation and receptivity. *Justice Quarterly*. DOI: 10.1080/07418825.2019.1703027.

Appendix A – Priority Vignette

While on patrol, you receive a call regarding a suspicious person in the parking lot of a busy strip mall. You have little information and do not know whether the subject has a weapon, but arrive at the scene and make contact with a male who fits the description you were given. He appears to be angry, is being loud, using profanity, and occasionally breaks eye contact and looks around the shopping area. The subject continues to slowly walk backwards away from you despite your order to stop.

Appendix B – Vignette Measures

How important is each of the following during the above interaction? (1=Not important to 5=Very Important; *Note: Items were mixed during survey administration and not clustered by measure.*)

Procedural justice priorities:

- Treating the subject respectfully
- Establishing rapport with the subject
- Explaining the reason you've made contact with the subject
- Treating the subject politely and with dignity
- Allowing the subject to explain his side of the story
- Considering the subject's side of the story
- Explaining to the subject the reasons for your decisions

- Earning the subject's trust

Maintaining self-control:

- Remaining calm
- Maintaining self-restraint
- Thinking about how my actions may impact people other than the subject
- Getting the subject to cooperate without using force
- Thinking through possible alternatives before I act
- Not making a decision about what to do until you've gathered all necessary information
- Trying to talk the subject into complying

Physical control priorities:

- Making the subject stop walking away
- Establishing physical control over the subject

Table 1. Post-Test Demographic Balance

	No Treatment	***Treatment***	***T-Test***
N	*195*	*114*	
Age			
21-24	12.3	5.3	2.04*
25-29	27.2	23.7	0.71
30-34	20.5	19.3	0.29
35-39	12.8	16.7	-0.91
40-44	8.2	11.4	-0.91
45-49	10.8	13.2	-0.61
50 and older	6.7	9.7	-0.93
Gender			
Male	83.6	85.1	0.14
Female	11.8	11.4	-0.14
Race/Ethnicity			
White	59.5	58.8	0.10
Hispanic	25.1	25.4	-0.08
Black	6.2	4.4	0.65
Native American	2.6	0.9	1.03
Asian	1.5	2.6	-0.67
Other	1.0	3.5	-1.53
Experience			
1-4 years	46.7	37.7	1.49
5-9 years	20.0	21.9	-0.44
10-14 years	14.9	17.5	-0.65
15-19 years	11.8	14.0	-0.60
20 years or more	5.6	7.0	-0.50
Rank			
Officer	85.6	90.4	-0.77
Specialist	5.1	6.1	-0.33
Other	7.2	3.6	1.37
Education			
High School	13.9	7.9	1.60
Less than 2 yrs	28.7	32.5	-0.65
Associate's	21.5	12.3	2.07*
Bachelor's	30.3	42.1	-2.08*
Graduate	4.1	4.4	-0.11
Military Service	30.3	30.7	0.10

Note: All numbers are percentages except for the T-value. $*p<0.05$

Table 2. Difference-in-Difference Estimates of the Impact of Social Interaction Training

	Full Sample		*Tucson Only*		*Fayetteville Only*	
Outcome	β	S.E.	β	S.E.	β	S.E.
Procedural Justice Priorities	0.12*	0.10	0.04	0.13	0.24**	0.16
Maintaining Self-Control	0.08	0.07	0.04	0.10	0.16^	0.12
Physical Control Priorities	-0.07	0.15	-0.03	0.18	-0.15	0.25

^p<0.10, *p<0.05, **p<0.01

Table 3. Difference-in-Difference Estimates with Dosage Consideration

	Full Sample		*Tucson Only*		*Fayetteville Only*	
Outcome	β	S.E.	β	S.E.	β	S.E.
Procedural Justice Priorities						
Low Dose	0.12*	0.13	0.06	0.16	0.21*	0.23
High Dose	0.04	0.14	-0.03	0.19	0.15^	0.19
Maintaining Self-Control						
Low Dose	0.11*	0.09	0.09	0.11	0.12	0.17
High Dose	0.01	0.10	-0.06	0.14	0.11	0.14
Physical Control Priorities						
Low Dose	0.02	0.19	0.02	0.22	0.04	0.35
High Dose	-0.13*	0.19	-0.06	0.26	-0.22*	0.29

^p<0.10, *p<0.05, **p<0.01

Table 4. Difference-in-Difference Estimates by District

	PJ Priorities		*Maintaining SC*		*Physical Control*	
District	β	S.E.	β	S.E.	β	S.E.
Tucson PD						
South (N=61)	**0.11**	**0.23**	**-0.05**	**0.17**	**-0.13**	**0.36**
East (N=64)	**0.11**	**0.26**	**0.17**	**0.19**	**0.11**	**0.36**
West (N=65)	-0.01	0.27	-0.02	0.21	-0.15	0.34
Midtown (N=53)	-0.02	0.30	0.03	0.20	0.11	0.41
Fayetteville PD						
Campbellton (N=57)	0.12	0.31	0.06	0.23	-0.38*	0.42
Central (N=56)	**0.33***	**0.27**	**0.22**	**0.20**	**0.10**	**0.37**
Cross Creek (N=43)	0.22	0.25	0.18	0.18	-0.17	0.52

Note: Low-dose districts in bold. N represents number of officers in each district at the pre-test survey as an indication of reduced sample size.
^p<0.10, *p<0.05, **p<0.01

Figure 1. Experimental Design

	Pre-Training (Early Spring 2017)		7 TDEs (Spring/Summer 2017)	Refresher (Summer 2017)		6 TDEs (Fall 2017)	Capstone (Winter 2017/8)	
Low-Dose Districts								
Treatment Group	R	O	X	X	O			
Control Group	R	O			O			
High-Dose Districts								
Treatment Group	R	O	X	X		X	X	O
Control Group	R	O						O

Note: Figure uses Campbell & Stanley's (1963) research design notation: R=random assignment, O=observation (surveys administered), X=treatment (training delivered).

Figure 2. Difference in Difference Plots

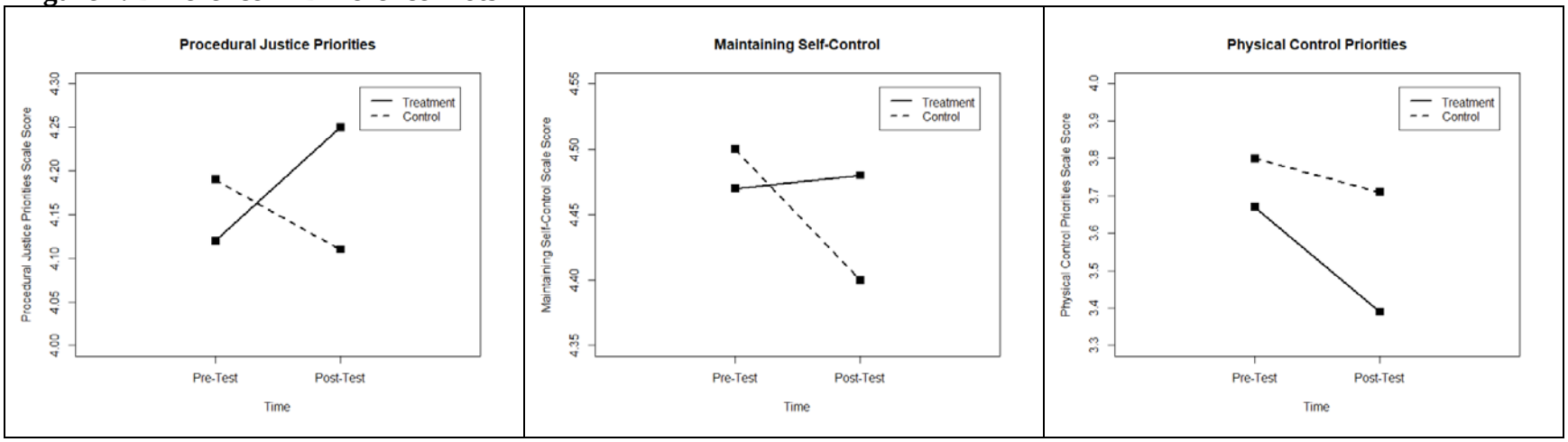

Figure 3. Difference in Difference Plots with Dosage

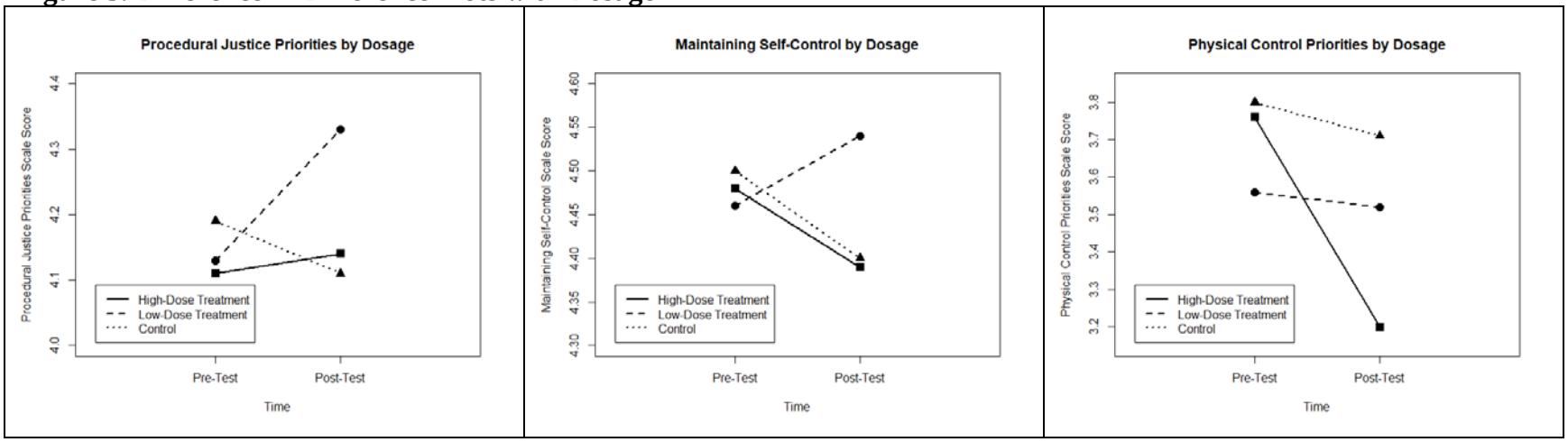

Figure 4. District Effect Sizes

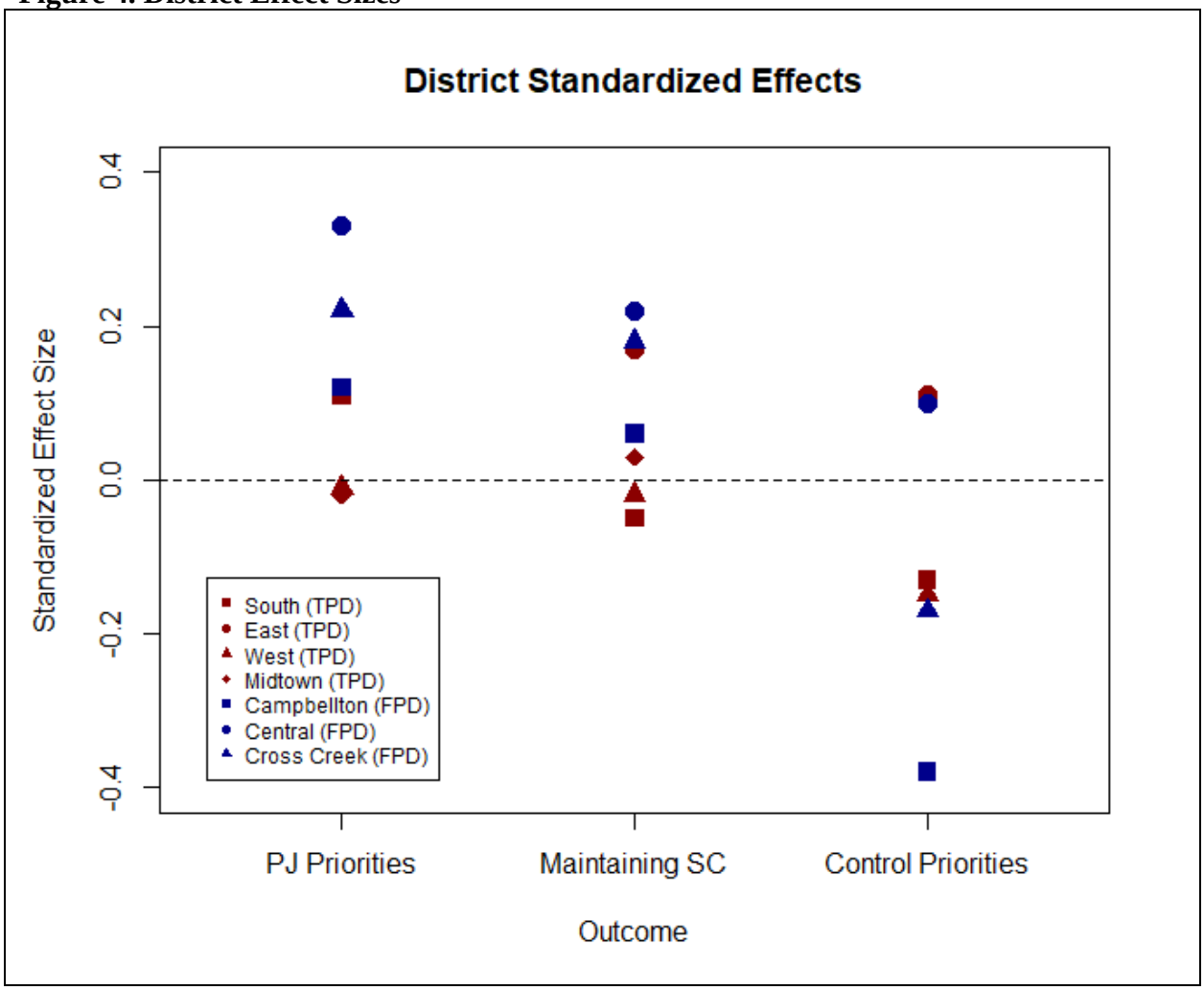

Figure 5. Low-dose Use of Force Reports (FPD)

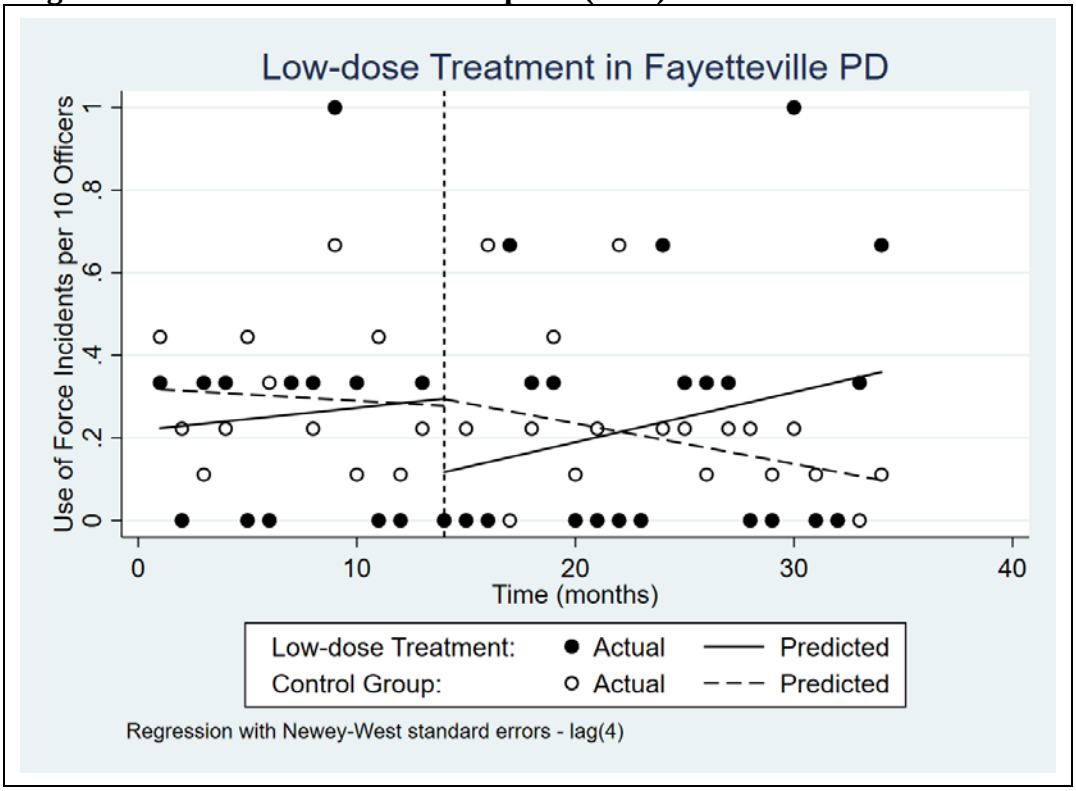

Figure 6. High-dose Use of Force Reports (FPD)

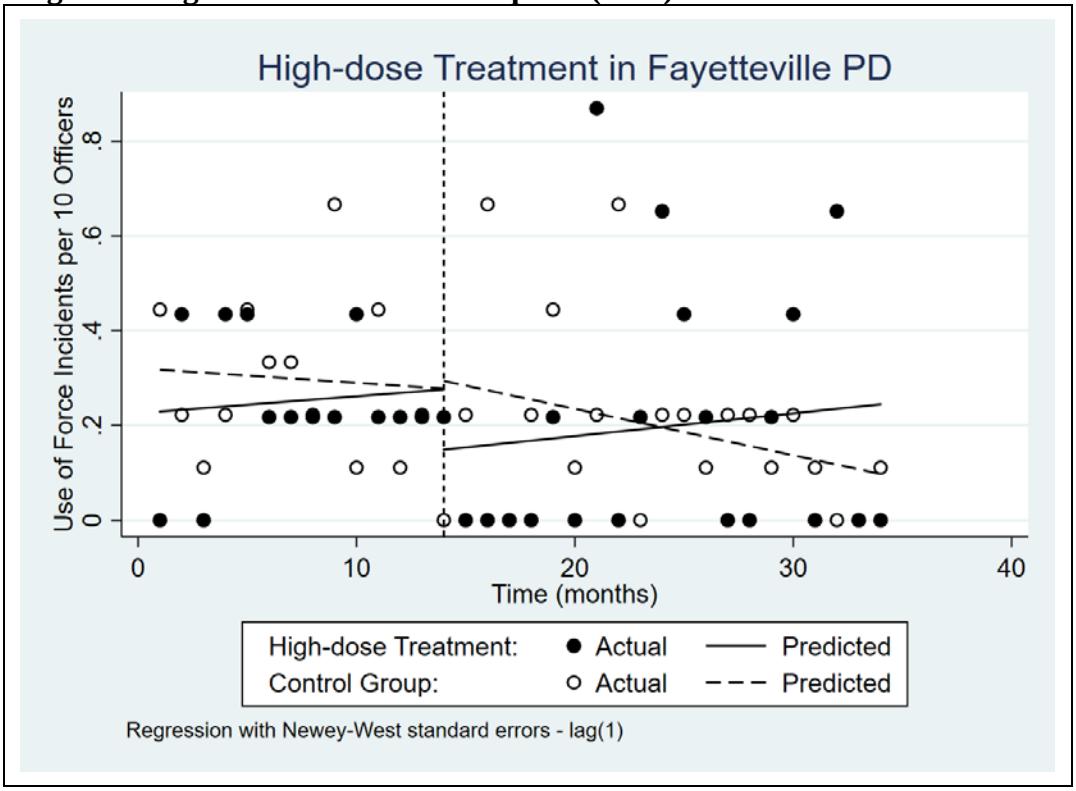

Figure 7. Low-dose Use of Force Reports (TPD)

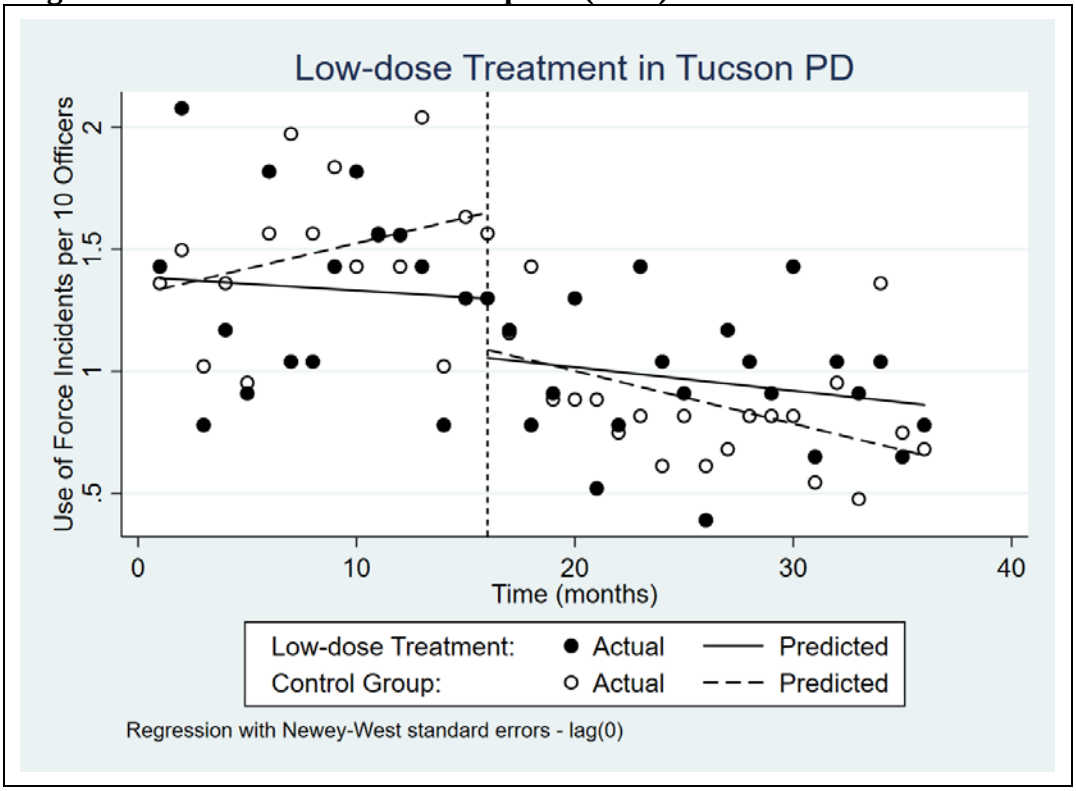

Figure 8. High-dose Use of Force Reports (TPD)