Appendix 2

Sampling for Alternative Study to the Combined Pilot

Institute for State and Local Governance[1]
Stanford Team[2]
May 18, 2020

To address the research questions for the alternative to the Combined Pilot, ISLG and the Stanford team have developed a new framework for sampling encounters to be reviewed by legal experts.[3] Specifically, in this alternative study, the two teams will need a sample of at least 2,500 police-citizen encounters that are *De Bour* Level 2 or above. The sample will be drawn in two waves with the first including approximately 1,500 encounters that will be reviewed by legal experts and the second approximately 1,000 encounters. The initial sample will capture a random sample of encounters that both ISLG and the Stanford team will utilize to answer primary research questions. In contrast, the second sample will be targeted in order to facilitate the use of machine learning techniques to identify Level 3 encounters. The following memo outlines core elements of the sampling plan. The sample will be drawn after the NYPD implements the two proposed policies that expand body worn camera (BWC) recording and categorization of encounters.

The sampling framework in the revised study differs substantially from that of the Combined Pilot because of fundamental changes to the research design. The Combined Pilot was a cluster randomized experiment testing the efficacy of two policy changes. As such, sampling was built around the selection of 16 commands for treatment and control and selecting 168 officer tours for observation across the selected commands. In contrast, the revised study is an observational study, which eliminates the need to focus on a limited number of comparable commands for treatment and control. Selection of encounters (rather than commands or tours) permits the direct selection of a much larger number of observations, which reduces the necessity of employing certain tactics such as stratification.

Sample 1

Both the ISLG and Stanford teams' studies require a random sample of police-citizen encounters that are *De Bour* Level 2 and above.[4] Observations will be randomly selected from the universe of encounters recorded by officers on their BWCs. The sample will be representative of police encounters with citizens in New York City that rise above Level 1.[5] In a

[1] Kathleen Doherty, Reagan Daly and Li Sian Goh.
[2] Nick Camp, Jennifer Eberhardt, and Rob Voigt.
[3] This memo specifically concerns the procedure to sample interactions to be reviewed by expert judges. Additional encounters or alternative sampling regimes (e.g., conditioned on citizen complaints or consent searches) may be necessary by ISLG or Stanford for questions that do not require legal expert judgments.
[4] Within an encounter, there may be interaction with multiple citizens. Legal experts will consider the legality of the officer's actions with regard to each citizen who interacts with the officer.
[5] There is one exception. Officers who consistently turn on their camera less or fail to turn on their BWC in a timely manner will be underrepresented in the study. This is only problematic if failing to turn on the BWC is associated with legally non-compliant behavior.

sample drawn from encounters, active officers will have more presence, because they engage in more encounters. This contrasts with a selection process based on officers, which is generalizable to the average officer. The proposed sample, however, will be more representative of the encounters, which will permit us to speak directly to how citizens experience policing.

The sample will include only encounters that reach Level 2 and above. Such a sample will generate more information on encounters that raise Fourth Amendment concerns relative to a simple random sample. The sample will also be more squarely focused on the Monitorship's mission of reducing unconstitutional Level 3 encounters. Without limiting the sample in this way, legal experts will review many encounters that do not raise constitutional concerns. As a result, we will have less information on more invasive encounters, which could leave us in a position where we cannot speak meaningfully to important questions about escalation or racial disparities.

Timing of Sampling

As mentioned above, the revised study will be executed after the NYPD has implemented its proposed policies of categorizing encounters and recording almost all encounters on BWCs. There will likely be a period of time where officers become accustomed to implementing the new policies. In consequence, we plan to commence data collection 4 weeks after the new policies take effect. This provides officers an adjustment period before data collection begins.

Selection will occur over an 8-week period. It is important to collect data over a period of time that is sufficiently long in order to minimize the influence of idiosyncratic events on the results. While improving the quality of the data collection, a longer time frame will increase the amount of time necessary to complete the study. To minimize delay, we propose to randomly select encounters at four points throughout the 8-week study period. One advantage to this approach is that ISLG and the Stanford team can begin coding, matching relevant reports, and transcribing encounters without having to wait until the conclusion of the study period.

Stratification

In the alternative to the Combined Pilot, it will be no longer necessary to stratify based on officer or precinct characteristics. The sampling regime will use stratification to address two distinct issues. First, we will stratify by officer categorization of *De Bour* level as a mechanism for including review of encounters labeled Level 1. Second, the study will stratify by citizen race to support our ability to detect disparities in officer behavior.

Random selection of encounters will produce a sample that reflects the population as the number of observations selected increases. Approximately 1,500 encounters Level 2 and above will ultimately be selected. As a result, encounters may be randomly selected without stratification on officer unit or command characteristics while yielding a sample where officer

units and commands will be approximately representative of their presence in the pool of encounters.[6]

Both the Stanford and ISLG studies will analyze encounters that reach Level 2 or above. Identifying a sample that excludes Level 1 is challenging. Researchers will have access to officer categorization of the *De Bour* level for each encounter, but officer categorization will be an imperfect indicator of the actual *De Bour* level if officers err in their identification of the *De Bour* level. It is not feasible to take a simple random sample of all *De Bour* levels because of the frequency with which we expect Level 1 encounters to occur. Level 1 encounters would comprise the vast majority of the sample and review of those encounters would be costly and an inefficient use of legal experts' expertise.

To address concerns related to officer miscategorization more efficiently, the study will stratify the sample selection by the officer's labeling of the *De Bour* level. This strategy permits the selection of a random sample of encounters labeled Level 2, 3 or 4 *and* a smaller random sample of encounters labeled Level 1.[7] The selected Level 1 encounters will be screened to determine whether each could plausibly be a higher-level encounter. ISLG in consultation with the Stanford team will develop a coding protocol for identifying encounters that are clearly Level 1. Any encounter that *potentially* rises above Level 1 will be included in the sample sent to legal experts. If legal experts determine that the *De Bour* level is 2 or above, the encounter will be included in the analyses. This approach ensures that a sample of encounters labeled Level 1 by officers are reviewed to determine whether they are in fact a higher level while limiting the costs associated with reviewing every Level 1 encounter captured in a random sample. This approach will also help the team ascertain, on a department-wide basis, the frequency with which higher-level encounters are miscategorized as Level 1 encounters.

There is one key issue associated with this approach. If encounters that were miscategorized as Level 1 differ significantly in compliance, escalation, and documentation from encounters that were accurately categorized, the results will not be representative of the city. Stratifying sample selection by the officer's labeling of the *De Bour* level may provide a skewed sample because Level 1 will be sampled below their prevalence in the actual universe of encounters. The extent to which stratifying sample selection by officer categorization skews the sample depends on the frequency with which higher-level encounters are miscategorized as Level 1. To address this

[6] Stratification by officer unit and characteristics of the command were particularly important in the Combined Pilot, because only 16 commands and 168 officers were selected for observation. As a result, a random sample was very likely going to be unrepresentative of the population.

[7] The number of encounters selected in each group has not yet been determined. There are two issues that affect the number of encounters selected: 1) the percent of BWC recordings that are be usable, 2) the rate of miscategorization of Level 1 encounters. For example, incomplete recordings will not be sent to judges for review. If recordings are often incomplete, then we may need to increase the number of encounters selected to ensure that we reach 1,500 encounters for legal expert review. We also do not know the rate of encounters that are miscategorized as Level 1 but are, in fact, Level 2 or above. If encounters are frequently miscategorized then we may need to select fewer encounters, because more of the encounters selected will be sent to legal experts. After the NYPD has implemented the new policy, there will necessarily be a pilot period where we examine BWC recordings and assess the appropriate number of encounters for selection in each group.

skew, the sample of higher-level encounters miscategorized as Level 1 can be weighted based on estimates of their frequency.

Stratification by citizen race will increase the ability of both studies to discern racial disparities in analyses on compliance, escalation and procedural justice. In a random sample of encounters of approximately 1,500 police-encounters, there may be too few encounters with white or Asian citizens captured in the sample to discern differences by citizen race. In 2019, black and Hispanic citizens constituted 89 percent of Level 3 encounters and 79 percent of arrests.[8] If we assume that citizens in Level 2 encounters exhibit a similar racial identity, there may be too few encounters with white or Asian citizens to meaningfully detect differences in legal compliance or officer language. This poses the danger that the study may find no evidence of racial disparities even if disparities are present, because the analysis is underpowered.

The primary means of addressing this issue is to stratify by citizen race, as labeled by officer, where such data are available.[9] We will select white and Asian citizens at a higher rate than their expected presence in a random sample to enhance our ability to detect differences across racial groups. While it may be counterintuitive to oversample white and Asian citizens when investigating Fourteenth Amendment violations, there must be sufficient encounters in each racial group to be able to discern meaningful differences. A very large increase in sample size of the study would also increase statistical power, but stratification is a more efficient means of increasing our ability to detect disparities without dramatically increasing the size and cost of the study. We will conduct power analysis to determine the appropriate numbers of encounters by race to include in the sample, taking care to oversample only to the minimum extent necessary.[10] We note that this will be difficult to implement if the NYPD's categorization of Level 2 encounters does not include information on citizen race.

Stratification by citizen race may have implications for the interpretation of results. For example, if whites are oversampled and encounters with white citizens are more likely to be compliant, the compliance rate of the sample will not be an accurate representation of compliance in the city. Fortunately, it is possible to address this issue by weighting encounters of different racial groups to reflect their presence in the population of encounters when producing estimates for compliance city-wide. Stratification can facilitate disparities analysis without detracting from the generalizability of the study to the city.

Selection Process

The NYPD has a record of Level 3 and 4 encounters as well as some Level 2 encounters (those with a consent search) through reports submitted by officers. Under the NYPD's proposal for documentation, there will be no database of information on all police-citizen encounters Level

[8] The data on Level 3 encounters were provided by the Monitor team and the data on arrests were found on NYC Open Data.

[9] Citizens' racial identification may differ from officers' perception of their race; however, officers' perception of race is more germane to questions of racial disparities in legal compliance and disrespect.

[10] We note that parity in the presence of racial groups is not necessary to be able to detect differences.

2 and above. Most Level 2 encounters will be documented only in the BWC system, unless the NYPD adopts a combined Level 2/Level 3 form. An investigative encounter form that documents both Level 2 and Level 3 encounters would be advisable, unless it would create insurmountable logistical or operational concerns. Without a combined Level2/Level 3 form, the primary data source with information on the universe of encounters is the BWC metadata. By BWC metadata, we refer to the record of each recording made by an officer, including the officer's name, date, time, length of video and sometimes information on the encounter. The NYPD proposed to include the *De Bour* level into the BWC system so that officers can categorize and report on the frequency of Level 1 and 2 encounters. (The NYPD has indicated that these data will be able to be accessed through the Axon system website evidence.com.) As a result, we will sample encounters from the BWC metadata for the study.

There are several unique challenges associated with selecting encounters through the BWC metadata in contrast to using a database of reported encounters. First, the system will often contain multiple records of the same encounter because each officer present will activate their camera. Before selection, recordings of the same encounter will have to be linked. If not, the probability of selecting an encounter for the study will increase as the number of officers present during the encounter increases. This would bias the study toward selecting encounters with more officers, which are more likely to be anti-crime officers and more serious encounters. The NYPD will need to support our efforts to link recordings of the same encounters prior to selection.[11]

Sample 2

An open question for the Stanford team is whether computational techniques can distinguish between correctly and incorrectly labelled Level 2 encounters through features in metadata, language, or video recordings. If so, this approach could scale legal experts' judgments to the interactions outside of our sampling regime. However, it is impossible to judge the feasibility of such a model until Sample 1 is collected. Supervised learning methods require judges to provide labels for a set of training data (i.e., whether Level 2-labelled encounters are actually Level 2 or Level 3). The model is trained against these labels and is then tested against a separate sample to determine its accuracy.

In addition to an initial sample of 1,500 encounters, we will select a second sample of 1,000 encounters for this purpose. If the Stanford team is able to build an adequately predictive model from the initial sample of encounters, then this second $_{\text{sample}}$ would consist of the 1,000 encounters most useful for refining the model (i.e., those closest to the boundary between correctly and incorrectly labelled Level 2 encounters).

[11] BWCs will recognize when another camera is activated nearby. However, it is not apparent from metadata which records may be of the same encounter.

However, if improperly categorized Level 2 encounters are rare, or if the differences between properly and improperly labelled encounters are exceedingly subtle, then a computational model may not be feasible for detecting compliance. If this is the case, we propose allocating these 1,000 additional encounters using the same sampling regime as Sample 1. This would allow for greater statistical power to measure the influence of covariates of interest (e.g., whether an encounter occurred during a housing patrol, the location in which an encounter occurred, etc.).

Proposed Study of Police-Citizen Encounters

Institute of State and Local Governance

May 28, 2020

Introduction

ISLG proposes to carry out a study that will address fundamental questions about the NYPD's compliance with the Fourth and Fourteenth Amendments in police-citizen encounters as an alternative to the Combined Pilot. Crucially, the study will examine legal compliance in police-citizen encounters, racial disparities in compliance and escalation, and appropriate documentation of encounters. All of which can be examined on a broader scale following the introduction of a new policy proposed by NYPD to record nearly all police-citizen encounters at every *De Bour* level with body-worn cameras (BWC). With a more complete record of police-citizen encounters available on BWC, the study will supply novel insights into officer compliance with the Fourth and Fourteenth Amendments and the degree of NYPD documentation of police-citizen encounters without direct observation of officers.

The study will explore the following research questions:

1) *Compliance*: How often does officer behavior in police-citizen encounters violate the Fourth Amendment or Fourteenth Amendment? What are the key reasons for officers' failure to comply with legal requirements? Are there racial disparities in the legality of police-citizen encounters?
2) *Escalation*: How often do police-citizen encounters escalate from lower to higher levels? Are there racial disparities in the legality of escalated encounters?
3) *Documentation*: Does the expansion of mandatory BWC recording to most Level 1 encounters increase the number of Level 3 encounters that are reported? Post-expansion of BWC recording, what is the documentation rate of Level 3 encounters? Are undocumented Level 3 encounters associated with race and ethnicity or legal non-compliance?

In addition to addressing these substantive research questions, the study will generate important descriptive information about the nature of police-citizen encounters. After the expansion of BWC recording, analysis of BWC footage by ISLG and legal experts will provide a more nuanced portrait of the police-citizen encounters than is currently available.

Approach for answering key research questions

1) Compliance with the Fourth and Fourteenth Amendments

One key area of responsibility for the Monitorship is legal compliance in police-citizen encounters. The study will investigate officer compliance with the Fourth and Fourteenth Amendments in encounters with citizens. To do so, ISLG will conduct three primary analyses. To address compliance with the Fourth Amendment, we will employ legal experts to assess encounters to determine whether there are any Fourth Amendment violations at each *De Bour* level, leading to an ultimate assessment of legal compliance for the encounters. To assess Fourteenth Amendment violations, we will examine whether

there are racial disparities in compliance across encounters and provide the legal experts the ability to indicate whether they found a Fourteenth Amendment violation in an individual encounter.

To establish the legality of officer compliance with the Fourth Amendment, a panel of retired judges (legal experts) will utilize BWC footage plus relevant documentation to determine and code the *De Bour* levels present in an encounter, whether officer actions were legally compliant and the reasons for non-compliance by *De Bour* level. Each encounter will be randomly assigned to at least two legal experts to establish consensus on legality.[1] With these data, ISLG will be able to analyze and report compliance of encounters by *De Bour* level.[2] In addition to documenting the rates of legal compliance, the assessments completed by the legal experts will also provide insight into the reasons that encounters were deemed non-compliant. Using the coded data, ISLG will report the most prevalent reasons cited by the legal experts for non-compliant encounters thus providing the NYPD with important information on the specific nature of Fourth Amendment violations.

The legal experts, as former judges, will be able to provide valuable analysis on Fourth Amendment concerns in police-citizen encounters. ISLG will utilize their expertise to supplement closed-ended coding. Each legal expert will complete an open-ended narrative that summarizes the most frequent causes of Fourth Amendment violations in the encounters that they analyzed and explain the nature of these violations. They can do so in greater depth and with more nuance than is possible in closed ended coding. These analyses will enable the NYPD to refine its training and oversight in ways that will improve the legality and quality of officer-citizen encounters.

Importantly, the study will assess encounters for Fourteenth Amendment violations in addition to Fourth Amendment violations, and this will be done in two ways. First, legal experts will be asked to identify any Fourteenth Amendment violations in each encounter they review. Second, ISLG will examine whether there are racial disparities in legal compliance across police-citizen encounters that are reviewed. We will assess the effect of citizen race on legal compliance while controlling for other aspects of the encounter that may affect compliance, thus determining whether black and Hispanic citizens are more likely to experience Fourth Amendment violations in their interactions with officers. It is important to note that this analysis will examine Fourteenth Amendment violations in encounters that occur, providing valuable information regarding disparities, or lack of disparities, in NYPD treatment of citizens in encounters.

2) Escalation

The expansion of BWC recording to most Level 1 encounters provides an opportunity to examine and better understand encounters that escalate from a Level 1 or 2 to Level 3 or 4. In this second analysis, ISLG will both explore the prevalence and legality of escalations in police-citizen encounters and examine any racial disparities in these encounters. Specifically, we will explore how often escalation occurs in encounters, whether escalated encounters are more likely to be non-compliant with Fourth

[1] A third legal expert will review any legal questions where there is disagreement in the initial review.
[2] We will also report descriptive information about encounters where there is disagreement among legal experts. This will provide nuance into those cases where a determination of compliance was complex and led to split decisions among the legal experts.

Amendment requirements, and whether there are racial disparities in the frequency and nature of escalations.

As described above, the first analysis will generate novel data on the *De Bour* levels contained in police-citizen encounters and legal compliance at each level. With these data, we will have the ability to describe *De Bour* escalation patterns in different ways. First, we will report the proportion of all encounters that escalated. This provides information on how common escalation through *De Bour* levels is in the universe of all encounters in the sample. Another analysis of interest is the proportion of Level 3 encounters that escalated from at least a Level 2. This will provide insight into the share of Level 3 encounters that resulted from escalation versus encounters where officers immediately detained an individual. After we have identified Level 3 encounters that escalated, we can compare compliance in escalated encounters relative to those encounters that did not escalate. Lastly, we will examine the proportion of Level 2 encounters that escalate to Level 3 or above. Using the definition of escalation above, encounters only escalate if they begin at Level 2 or below. This will establish the rate of escalation among those encounters where escalation is possible.

We will also examine racial disparities in *De Bour* escalation through two strategies. First, we will examine whether Level 1 or 2 encounters escalate more often if the citizen is black or Hispanic and whether there are racial disparities in the compliance rate of escalated encounters.[3] A second analysis will focus on describing racial differences in Level 3 encounters. Specifically, we will examine whether Level 3 encounters with black and Hispanic citizens are more likely to have escalated from a Level 1 or 2 as compared to white citizens. This will provide insight into whether there are differences by race in how Level 3 encounters are initiated. We will then examine whether Level 3 encounters that escalated from a Level 1 or 2 are less likely to be legally compliant and whether there are any differences by citizen race.

It is important to note that this analysis may be descriptive, because the sample may be small if escalation across *De Bour* levels is rare. While we will explore the possibility of looking at other factors and circumstances that may increase the likelihood of escalation, it may not be possible if there is not sufficient data.

A second analysis will consider escalation in a more fine-grained manner than through *De Bour* levels, which will permit a more nuanced assessment of escalation. The Fourteenth Amendment requires that similarly situated individuals be treated equally (see Liability Opinion, pg. 27). Examining escalation through *De Bour* levels may mask important differences in the intensiveness of encounters within *De Bour* level. Officers may take a range of actions within a *De Bour* level that affect the experience of the encounter for the citizen. For example, a Level 3 encounter may involve a frisk, but may not. Existing research has sequenced citizen and officer actions. We will use BWC footage to sequence the actions of an encounter to identify actions that might be associated with escalation across *De Bour* levels, especially inappropriate actions. By breaking down encounters into a sequence of actions, we can examine how officers interact with citizens, factors that trigger escalation in the actions that officers take, and inequalities in how interactions progress. This will also examine disparities in how

[3] The analysis will not be able to assess racial disparities in the probability of a citizen of having a Level 1 or 2 encounter.

encounters escalate. This work will complement analysis of language executed by the Stanford team. We will explore existing measures of escalation and consider how they might be adapted.

3) Documentation

The presence of undocumented stops limits the ability of the Monitor to assess the legality of Level 3 encounters. This is especially important if there is a relationship between documentation of an encounter and its legality. ISLG will examine the issue of officer documentation in two ways. First, we will assess the effect of expanding BWC recording to almost all encounters on the number of reported Level 3 encounters. Second, *after* BWC recording expands, we will determine the current rate of documented Level 3 encounters.

Changing the NYPD BWC recording policy to include most Level 1 encounters will increase the transparency of police-citizen encounters, which may affect an officer's decision to report a Level 3 encounter.[4] We will assess whether this policy change affects the number of reported Level 3 encounters. Because assessing reported Level 3 encounters relies on administrative data rather than BWC footage, we will be able to examine all officers on patrol rather than a sample of officers. The NYPD could provide a count of tours and reported Level 3 encounters by officer for approximately 12 weeks before and after the introduction of the policy. With these data, we will assess whether policy change is associated with an increase in the number of reported Level 3 encounters while controlling for other factors that affect the prevalence of Level 3 encounters. Ultimately, the strategy for analysis will depend on whether the City rolls out the new policy gradually by commands or citywide. We can better assess a causal effect between the policy and documented Level 3 encounters if the City rolls the policy out gradually. It permits differentiation between time trends and the effect of the policy on reported Level 3 encounters.

If the City rolls out the policy gradually by commands, ISLG can adopt an analytic strategy that better accounts for trends that change over time in addition to the policy change.[5] Using an officer-by-week unit of analysis would allow ISLG to control for individual officers' propensity to stop, as well as variations in time or seasonality—i.e., officer and week-fixed effects. In addition, ISLG would control for differences across precincts. NYPD can provide data on the number of stops each officer reported, as well as the number of tours each officer worked every week. From this, the number of stops per tour each officer reported every week can be calculated. Level 3 encounters will be calculated per tour as we must account for the opportunity to stop a citizen (or the number of tours worked).

[4] One outstanding issue is the form of documentation for Level 1 and 2 encounters. Categorization by *De Bour* level of these encounters would also affect transparency and may further affect the incentive to report. As yet, it is unknown what that categorization will look like or even whether it will be implemented with changes in BWC policy.

[5] One possibility is to conduct a difference-in-differences analysis in which the number of documented Level 3 encounters in commands which have implemented the policy, both prior to and after the policy change, are compared with commands which have not yet implemented the policy.

If the City chooses to implement the policy citywide, 'control' precincts in which the policy has not yet been implemented will be unavailable for comparison.[6] As discussed above, this approach is less desirable than a roll-out of the policy. Implementing the policy citywide provides only one pre-post point of comparison – the time period prior to roll-out and the time period after it. In comparison, rolling out the policy gradually would provide multiple points of comparison, in addition to allowing ISLG to control for variance in officer propensity to stop, seasonality, and crime trends.

A challenge common to both approaches is that it will be difficult to differentiate between an increase in actual Level 3 encounters and an increase in the reporting of those encounters. Moreover, we cannot differentiate between reporting and incidence with regard to the distribution of levels in stops, because Level 1 and 2 stops are not currently documented. A possible solution would be to conduct audits of potentially undocumented stops using radio transmissions to identify instances in which stops appear to have been made, but a stop report was not recorded, in each precinct during the pre- and post-intervention periods, as was done throughout the quarters between 2016–2018 (RAND audits).[7]

The second analysis on documentation will focus on the accurate documentation of Level 3 encounters *after* the BWC policy change is implemented. Legal expert coding will identify the *De Bour* level of all encounters for a sample of encounters. With that information, we will be able to identify Level 3 encounters that were not reported by officers and calculate the rate of undocumented Level 3 encounters. Once undocumented Level 3 encounters have been identified, it is possible to explore whether there is any relationship between documentation and legal compliance or citizen race. This analysis may be descriptive, because the sample size will likely be small.

* * *

Jennifer Eberhardt, Rob Voigt and Nick Camp (Stanford team) can build on this analysis by utilizing machine learning tactics to examine officer categorization of *De Bour* levels in a much wider set of encounters. We will provide legal expert coding and documentation information to them so that they may use legal expert identification of *De Bour* level to further their efforts.

Data Collection and Analysis:

The proposed analyses will rely on three sources of data: BWC footage, police reports, and administrative data. ISLG will descriptively code encounters, organize reports and administrative data by encounter, and then assign encounters to legal experts for review. This will generate a dataset of encounters that can be analyzed to generate new information on several distinct questions about compliance with the Fourth and Fourteenth Amendments, accurate documentation and the dynamics of escalation in encounters. BWC footage will serve as the primary data source for sampling and coding

[6] One possibility is to use an interrupted time-series design, in which trends in reported stop outcomes observed post-implementation will be compared to the trends observed prior to implementation.

[7] If the RAND audits suggest increased compliance in reporting after the policy change is implemented, then any increase in the number of reported stops is likely a result of increased reporting. However, if the reporting rate from the RAND audits remains constant, an increase in the number of reported stops will be the result of officers conducting more stops and not increasing the rate at which they document stops.

encounters. Therefore, the study will primarily be confined to documenting and analyzing encounters *after* the NYPD expands BWC recording to include Level 1 encounters. Importantly, ISLG will coordinate with the Stanford team on sampling, the coding of BWC videos, and the legal expert instrument to ensure the data collected serves the needs of ISLG's analyses as well as the Stanford team's analysis of language.

ISLG and the Stanford team will need a sample of at least 2,500 police-citizen encounters that are *De Bour* Level 2 or above. The sample will be drawn in two waves with the first including approximately 1,500 encounters and the second approximately 1,000 encounters. The initial sample will capture a random sample of encounters that both ISLG and the Stanford team will utilize to answer primary research questions. In contrast, the second sample will be narrower in order to facilitate the use of machine learning techniques to identify Level 3 encounters. The samples will be drawn after the NYPD implements the two proposed policies that expand body worn camera recording and categorization of encounters. A separate memo outlines the in detail the joint sampling plan for ISLG and the Stanford team.

Once a sample of encounters has been selected, ISLG will employ graduate student research assistants to code each encounter with a revised survey instrument. First, graduate students will assess whether the BWC footage for a selected encounter is sufficiently clear and complete for inclusion in the study. We will track the number of events that have to be excluded for failure to turn on BWC for the entirety of the encounters as well as any other reasons that an event needed to be excluded. After the footage has be deemed sufficient, graduate students will code key elements of each encounter. Coding will include information about citizens in the encounter as well as actions officers took. These data will provide descriptive information on encounters. The data will also aggregate features of encounters that will be important in the analyses conducted by both ISLG and the Stanford team. We will coordinate on the development of a new instrument to code data.

ISLG will also obtain several types of administrative data. First, we require the BWC metadata. These data are necessary for sample selection and in order to obtain important information about encounters, including their length. ISLG also requires other types of data from the NYPD on sampled encounters, including officer reports, ICAD data, and personnel information. For each encounter, ISLG will need to obtain any relevant reports submitted by officers (such as arrest reports, use of force reports). We will also extract the relevant portion of the ICAD data for the encounter. Importantly, ISLG must also acquire data on officer documentation for all Level 1 and 2 encounters. (The form of that documentation is not yet certain, but we note that it will be needed.)

The next step in the data collection will be legal expert review of sampled encounters. ISLG will bundle BWC footage with reports and ICAD data in preparation for legal expert review of the encounter. ISLG will then assign events to legal experts for evaluation through a web application. Assessment of legality will include identification of the *De Bour* level(s), a judgment of legality at each *De Bour* level in an encounter, and identification of Fourteenth Amendment violations. At least two legal experts will review each encounter. If there is a conflict between two experts on any legal judgments, a third legal expert will also review the encounter, but the third expert's evaluation will be confined to the questions that generated conflict between the first two legal experts. This process establishes a consensus opinion for each legal question. If officer actions are determined to be non-compliant, legal experts will provide the legal reason for non-compliance.

At the conclusion of their coding of individual encounters, the legal experts will have the opportunity to write a narrative analysis of the trends that they observed. The aim of this analysis is to provide the NYPD with more nuanced information on areas where officer behavior in encounters might be improved through training and oversight. The legal experts will be provided with a set of summary statistics on the encounters that they reviewed, including the number of encounters by *De Bour* level, the rate of non-compliance, and the reasons provided for non-compliance.[8] The legal experts will have the ability to review their coding for any specific encounter and to view relevant BWC footage. This may be useful, for example, if an expert wants to discuss a set of encounters which exhibited similar legal issues. They will complete an open-ended narrative setting forth their analysis on the nature of the Fourth Amendment violations that they observed. ISLG will supplement the quantitative analysis in its report with observations drawn from the legal experts' analyses.

The analysis on reported Level 3 encounters before and after the expansion of BWC recording will rely only on NYPD administrative data, which will permit analysis of the universe of officers on patrol. ISLG requires incident-level information on Level 3 stops, specifically the precinct number, the time and date, and an identification number for the officer who made the stop. All except the last is published in the publicly available SQF data. In addition, information on each tour that officers worked over the time period including precinct, date, shift duration (accounting for whether the shift went into overtime), officers' identification numbers, and the identification number of officers they were on tour with, is required. (The data can be de-identified as long as the same random ID number is provided for an officer across all the data.) If the policy is rolled out over time, NYPD would have to report when commands changed policy and the length of the data collection period may need to increase depending on the timing of the roll-out.

ISLG will draw on these different sources of data to conduct analysis for each research question listed above. Analysis for some research questions will utilize only a subset of the encounters that occur in the sample. For example, the analysis on documentation of Level 3 encounters will focus only on encounters that the legal experts identify as Level 3.

[8] The aim of providing legal experts with descriptive statistics is to minimize the salience of recent or unusual events in their consideration of trends.

Using Machine Learning Techniques to Analyze Body-Worn Camera Footage

Stanford Team
June 2020

Overview and Aims

Support for officer body-worn cameras (BWCs) is broad and growing. By some estimates, 95% of police departments across the country have invested in (or plan to invest in) them. Yet, despite the growing support for these cameras, and the large number of encounters they capture each day, BWC recordings are more often used as evidence to evaluate particular interactions than as data to inform practice and policy.

The Stanford team fills this gap by developing scalable computational tools for quickly and accurately analyzing police-community interactions captured by BWCs, and for examining broad patterns in those interactions. The NYPD, for example, could use such footage to examine the extent to which officers are complying with departmental practices and policies during encounters with the public. The footage could also be used to understand how such encounters might escalate unnecessarily. Because the first phase of Stanford's proposal relies on BWC videos and data already recorded, this part of the proposal can begin before the NYPD makes the changes outlined in the City's letter dated February 21, 2020 regarding an alternative to the combined pilot ordered by the Court.

The language officers use during investigative encounters is central to the above questions. Indeed, an officer's words are of great legal import: they can signal to a citizen whether they are detained or free to leave, request their consent, or provoke their complaint. More broadly, the manner in which officers relate to the public is consequential for building or eroding citizens' trust in the law, support for law enforcement, and even whether they personally cooperate with the police[1].

Police officers' treatment of the public during police-citizen encounters is also central to the monitorship and consent decrees in the *Floyd*, *Ligon* and *Davis* cases. The Court found the City liable for violating plaintiff's Fourth and Fourteenth Amendment rights. Not only were racially defined groups targeted for stops, but "[b]oth statistical and anecdotal evidence showed that minorities are indeed treated differently than whites. For example, once a stop is made, Blacks and Hispanics are more likely to be subjected to the use of force than whites." (p.13, Liability Opinion). The Court, in its finding regarding discrimination, highlighted the fact that Blacks who were subject to law enforcement action following their stop were about 30% more likely than Whites to be arrested (as opposed to receiving a summons) after a stop for the same suspected crime, even after controlling for relevant variables, and Blacks who were stopped were about 14% more likely than Whites to be subjected to the use of force (p.59-60, Liability Opinion). In the words of the Court, the Fourteenth Amendment's Equal Protection Clause "is essentially a direction that all persons similarly situated should be treated alike" (p.27,Liability Opinion). Thus, the question of whether officers treat Black and White citizens alike in encounters recorded on BWCs is paramount.

[1] Tyler, T. R., & Huo, Y. (2002). Trust in the law: Encouraging public cooperation with the police and courts. Russell Sage Foundation.

Here, we outline possible areas where a machine-learning approach to body-worn camera footage can help us understand and evaluate police-initiated pedestrian encounters. Specifically, we highlight the benefits of such an approach for evaluating four target areas:

- *complaints* - the actions taken on the part of an officer that lead encounters to result in a complaint
- *consent searches* - the language used by officers to request consent searches from citizens, and the impact of different language choices on community responses to that request
- *compliance* - the extent to which officers correctly classifying their encounters with public
- *escalation* - the manner in which an encounter becomes more fraught over time even when the De Bour level remains steady, and/or moves across De Bour levels, such as from Level 1 to 2 or Level 2 to 3.

We seek to answer these questions in two phases. The first phase aims to explore questions related to complaints and consent searches, which we can evaluate using existing data and footage from 2019. The second phase examines compliance and escalation using a sample of footage to be collected in coordination with ISLG and coded by legal experts.

Background

Our team has already used machine learning to develop a novel algorithm to extract and analyze officer language from a large corpus of body camera footage from the Oakland Police Department. Across nearly 1,000 routine traffic stops, we found that officers spoke more respectfully to White vs. Black community members, even after controlling for the location and outcome of the stop, the severity of the infraction, and the officer's race. This respect deficit, which is present from the beginning of the stop to the end, is significant because it can contribute to racial differences in reported experiences with law enforcement as well as community members' trust in the law.

In another series of studies, we extended our analysis from *what* words officers use during routine traffic stop interactions to *when* they use them during those interactions. Applying computational dialog methods to the same corpus of footage, we identified 11 discrete conversational actions that police officers take during stops (e.g., greeting the driver, asking for documents), and developed another novel algorithm to automatically tag footage for these acts. Disaggregating officer language and respect at the act (vs. stop) level can further help law enforcement agencies identify when community members are most likely to perceive disrespect or distrust, and to target social tactics training accordingly.

We have also begun to analyze officers' tone of voice during traffic stops. Once again, we find that officers speak to White drivers more respectfully than Black drivers. In addition, we find

that those community members who listen to short clips of officers' tone as they speak to White drivers (as opposed to Black drivers) state that they have greater trust in law enforcement more generally and expect more positive interactions with the police. These findings directly demonstrate how racial disparities in treatment during interactions could bolster or erode trust.

Proposed Directions

This approach can shed light on officer-initiated investigative encounters in New York City, and how they might differ by citizen race. Below, we discuss possible approaches we would take to explore particular questions related to complaints, consent searches, compliance, and escalation, using machine learning techniques to better understand and evaluate police-initiated pedestrian encounters. The scale of our analysis further lets us test the extent to which there are racial disparities in these aspects of police-citizen interactions. We propose two phases as follows.

Phase 1. In the first phase of this research, we constrain our sample to Level 3 stops from 2019. Since these Level 3 stops are already recorded and tagged by officers, we are able to sample and analyze interactions resulting in **complaints**, the linguistic characteristics of **consent searches**, and the conditions surrounding **escalation**. This work can begin immediately given existing department procedures for data collection and storage.

Phase 2. In the second phase of this research, in collaboration with the ISLG team, we propose to obtain a sample of Level 1, 2 and 3 encounters. As outlined in the ISLG sampling memo, this sampling requires NYPD officers to assign a De Bour level to each of their recordings. By obtaining judgments from legal expert coders evaluating the footage, it will be possible to study the linguistic dimensions of officer **compliance** and **De Bour escalation**.

Phase 1 - Analysis of Existing Data

In the first phase of research, we address a set of questions that can be examined under current procedures and policies for data collection, using existing data (encounters recorded in 2019). This will allow our team to begin the research process immediately, and thereby establish our cooperative relationship with the NYPD for exchanging data, as well as our research pipeline for data processing. This will facilitate sampling and matching recordings in collaboration with ISLG during Phase 2.

Prerequisites

The studies described in this section rely on the following:

- Computer-readable metadata, such as spreadsheets of stop reports, for all Level 3 encounters taking place during 2019
- Computer-readable records of citizen complaints taking place during 2019
- Computer-readable records of consent searches taking place during 2019
- Download access to footage from the sample for transcription and computational analysis

Sampling Procedure

For Study 1.a. (Retrospective Analysis of Complaints), we will sample the full population of BWC-recorded Level 3 encounters resulting in a citizen complaint. For a quasi-control group for comparison, we will use propensity score matching using automatically-extracted metadata and linguistic features of interactions to establish a comparable sample of equal size composed of interactions similar to the above but which did not result in a complaint.

For Study 1.b. (Linguistic Behavior in Consent Searches), we will sample the full population of BWC-recorded Level 3 encounters documented as including a consent search.

For Study 1.c. (Escalation in Level 3 Encounters), we will leverage the population of samples collected for Studies 1.a. and 1.b. to examine those in which escalation occurred.

We will begin this work by having these sampled encounters professionally transcribed, including per-utterance timestamps and speaker diarization (record of who is talking to whom for each utterance).

Study 1.a. Retrospective Analysis of Complaints

Research Questions: What elements of officer language and behavior in Level 3 encounters are most likely to result in citizen complaints?

A large number of citizen complaints against the police concern whether citizens feel that officers communicate appropriately and professionally with them. This study would seek to identify elements of police stops that are most predictive of citizen complaints, compared to otherwise similar stops that did not result in a citizen complaint. Specifically, we would focus our analysis on the subset of citizen complaints associated with Level 3 stops, where the encounter in question was captured on officer body-worn camera.

To address this question, we will train a machine learning model to learn how to distinguish between Level 3 encounters that led to complaints and those that did not. As in our previous work on officer respect, we can do this both with general distributional information about the words officers use, as well as with reference to a targeted set of linguistic features or strategies that are likely to be predictive. Our model will learn weights which quantitatively measure the influence of any given feature on a prediction of complaint/no-complaint, allowing us to determine which sorts of linguistic behaviors most strongly distinguish these kinds of encounters.

Such a model would shed light on the dynamics in police-citizen interactions that underlie complaints of unprofessional conduct or officer disrespect. An interesting possibility that such a model opens up is the ability to look for racial differences among complaints - perhaps a certain set of officer linguistic features are more likely to lead to complaints by White citizens, while another set is more predictive for Black citizens.

Study 1.b. Linguistic Behavior in Consent Searches

Research Questions: Do officers seek overt consent to search in Level 3 encounters? If so, when and how is consent sought? What language used by officers is associated with citizen consent versus refusal to search? When consent is granted, when and how is it done linguistically?

Upwards of 90% of warrantless police searches are conducted by means of the consent exception to the Fourth Amendment, but the conditions under which such searches are truly "voluntary" is a complex subject of legal inquiry that is ultimately tied to officers' communication in interactions.[2] Officers may voice a request to search in any number of ways, and citizens stopped and questioned by the police may not understand the full weight of the

[2] See, for instance: Simmons, Ric. "Not Voluntary but Still Reasonable: A New Paradigm for Understanding the Consent Searches Doctrine." Indiana Law Journal, vol. 80, no. 3, Summer 2005, p. 773-824

request or their option to refuse it. These challenges may be exacerbated by disparities in treatment and general miscommunication between police and members of the public.

Therefore we propose to examine the linguistic considerations surrounding consent searches by examining a transcribed sample of footage from encounters in which officers request a citizen's consent to search (both accepted and rejected consent searches). Note that in this study we aim to descriptively characterize the procedural facts surrounding searches so as to quantify potential disparity. To do so, we will develop algorithms that can automatically identify the occasion(s) of a request to search in the transcript of an interaction.

This will allow us to first ask basic distributional questions about these requests, such as: When in an interaction do they tend to occur? Considering the linguistic form such requests may take, are requests overt ("Can I search you?"), covert ("You mind if I check that out for a second?"), or even framed as commands or declarative statements ("Let me look in here.")? To the extent that the linguistic framings of these requests differ, are officers more likely to use more covert requests with Black and Hispanic citizens? And are some framings more likely to elicit a citizen's consent than others? When consent from the citizen is granted, is it clear and overt ("Yes, that's fine.") or indirect, hesitant, or ambiguous ("I guess." or "I'm good.")?

We can further examine the trajectories of the stops before and after consent searches take place, whether with reference to escalation as will be discussed below in Study 1.c., or in terms of conversational events in the encounter, like whether the officer provides the reason for the stop or an explicit reason for the search.

Study 1.c. Escalation in Level 3 Encounters

Research Questions: How can an officer's communication lead to the escalation of an encounter? When does citizen language lead officers to act more or less respectfully?

In order to determine which encounters escalate (e.g. low-level enforcement stops), and when they might escalate (e.g. in response to an insult) or de-escalate (e.g. once an officer gives the reason for a stop), we will map the trajectories of stops sampled for questions 1a and 1b. Specifically, we will measure the frequency of linguistic correlates of officer respect and citizen agitation over the time-course of a stop. A stop can be said to escalate when the officer becomes less respectful and a citizen more agitated over time. We will validate these linguistic ratings against human judgments of whether the encounter is becoming more or less fraught over time.

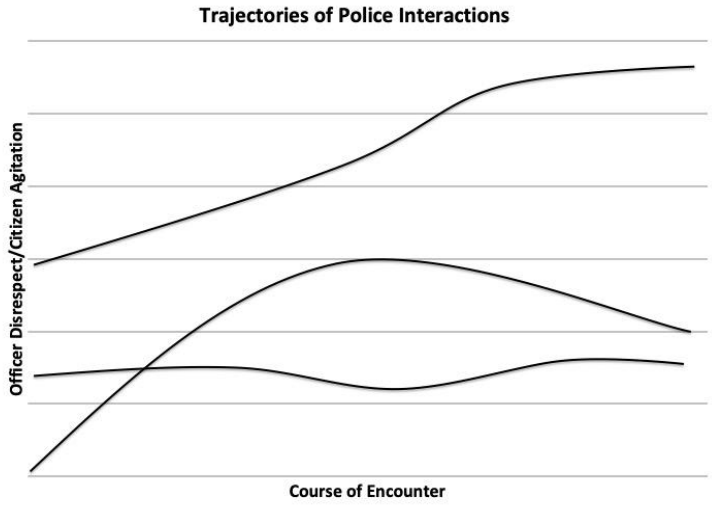

Figure 1. Schematic trajectories of police-citizen encounters. Each line represents a stop; note that interactions can vary in their starting position, ending position, and general trajectories

By charting the trajectory of each encounter, we can then ask what variables influence a) the starting state of Level 3 encounters (i.e. the intercept of this trajectory), b) the rate of escalation (i.e. the slope), and c) inflections in trajectories (i.e. going from a positive to a negative slope within an interaction). For example, does whether, when, and how officers state the reason for the encounter influence the likelihood that the encounter will escalate? Do citizens' procedural questions change the trajectory of police encounters?

Phase 2 - Analysis of Future Data

In the second phase of this research, in collaboration with ISLG, we aim to address questions surrounding the relationship between officer communication and legal standards of conduct: compliance with NYPD documentation and categorization of stops and De Bour escalation in officer encounters.

The questions we address in this phase require officers to categorize BWC recordings of Level 1 and Level 2 encounters as outlined in the ISLG sampling procedures. Further, since these questions concern the assignment and change in De Bour levels, they require legal experts' judgments.

Prerequisites

The studies described in this section rely on the following:

- Computer-readable metadata, such as spreadsheets of stop reports, for all encounters taking place during the period of the sample
- Interactions sampled by ISLG as per the attached sampling memo
- Download access to footage from the sample for transcription and computational analysis
- Race of citizen information, obtained either from officer-generated metadata (e.g. combined Level 2/Level 3 stop reports), ISLG coding, or both.

Sampling Procedure

The sampling procedure for Phase 2 will be coordinated with ISLG in the interest of obtaining generalizable population-level estimates of officer behavior with regards to compliance and escalation. This procedure is discussed in detail in the attached sampling memo.

We will begin this work by having the sample of encounters - those coded by ISLG and sent to legal experts for judgment - professionally transcribed, including per-utterance timestamps and speaker diarization (record of who is talking to whom for each utterance).

Study 2.a. Analysis of Officer Compliance

Research Question: How do officers, through their words, create a situation where a reasonable person would not feel free to leave?

NYPD patrol guidelines indicate that an officer "may not create a situation (either by words or actions) when a reasonable person would not feel free to walk away" during a Level 2

encounter; in contrast, in a Level 3 encounter, officers may detain the subject (i.e. "a reasonable person would not feel free to disregard the officer and walk away"). The goal of this study is to identify linguistic features that correspond to this distinction, confirm that they communicate appropriate levels of freedom or constraint on the part of the subject, and identify cases where legal experts judge that officers' language might create a situation where a reasonable person would not feel free to leave.

In coordination with ISLG, we will obtain legal expert judgments which either validate an officers' categorization of a stop or contradict it. We expect encounters labeled by officers as Level 3 but which legal experts judge to be Level 2 to be relatively rare. Therefore we distinguish three core types of encounters: basic Level 2 and Level 3 encounters, which are so categorized by officers and then validated by legal experts, and non-compliant Level 2 encounters (which we will refer to as NC-L2 encounters), which were categorized as Level 2 by officers but contradicted by legal experts who tagged them as Level 3 stops.

We can examine officer language and behavior across these different kinds of encounters by analyzing transcripts. Specifically, we will use machine learning techniques to build a model which aims to predict whether an encounter was Level 2 or Level 3 on the basis of the officers' words alone. This model will be trained on legal expert judgments.

Such a system inherently allows us to extract similarity between encounters to quantify the "severity" of an NC-L2: whether it more closely resembles a Level 2 or a Level 3 encounter. NC-L2 encounters that resemble Level 2 encounters linguistically (i.e., they are difficult to distinguish using officer words alone) may result from officer misinterpretation of NYPD policies which could be remedied with clarification. On the other hand, NC-L2 encounters that look very similar to Level 3 encounters, may indicate that officers are more aware of the non-compliance and are intentionally underreporting. Such a system would allow us to determine which of these scenarios is more prevalent, and to what degree. Moreover, we could identify whether a racial disparity exists within NC-L2 encounters: whether NC-L2 stops of White pedestrians more closely resemble Level 2 encounters than stops of Black pedestrians.

In our previous work on officer respect we identified linguistic strategies officers used which conveyed respect , developed algorithms to measure those linguistic strategies in transcripts, and used them as features for a computational model of respect. In this case we could follow that methodological paradigm, both by identifying which words were most predictive as described above, and by consulting with legal and police experts as well as the ISLG team as to key linguistic strategies for compliance. We would distill these findings down to a set of key features and train a model to quantify their usage in the three types of encounters, which could be evaluated in terms of racial disparity as well. We could find, for instance, that a higher rate of NC-L2 encounters among Black pedestrians could be explained by the use of certain phrases which officers are independently more likely to use when interacting with Black people. Additionally, this approach would provide concrete strategies for policy and training. For instance, if a certain phrase or set of phrases are highly indicative of NC-L2 encounters but

rarely occur in Level 2 and Level 3 encounters, a policy intervention that these be avoided could be implemented.

Study 2.b. De Bour Escalation

Research Questions: To what extent are officers' words predictive of improper escalation? To what extent does escalation, on the part of either the citizen or the officer, lead to an escalation in De Bour level?

A second area of interest concerns the escalation of Level 3 encounters (i.e. whether these encounters began at a lower level), the propriety of such escalation, and racial disparities in these trajectories. Our goal in these studies would be to identify key phrases or speech events that indicate escalation from legal expert judgments, build a predictive model to identify those features in a large amount of unseen data, and then examine the immediate context of these features (e.g. the officer observes contraband, a subject makes a complaint). Our power to answer these questions, and the granularity of our analyses depends on the prevalence of Level 3 encounters that began as Level 1 or Level 2 encounters.

To test this question, in coordination with ISLG we would ensure that for each encounter in our sample, legal experts viewing the footage determine whether the encounter starts at Level 3, and, if not, at what timepoint in the recording the interaction crosses that threshold.

As in our other studies, our goal would be to scale expert judgments to unseen data with computational modelling. Just as we have applied a computational dialog approach to categorize elements of traffic stops (greetings, document requests, sanctions, e.g.), we would identify phases of Level 3 encounters, including the point at which they cross the boundary from lower level encounters to Level 3 ones. Such a model serves two proximate goals. First, it would allow us to compare the trajectory of different encounters across citizen race (how quickly encounters move from introductory phases to the Level 3 boundary); second, it could potentially let us flag other Level 3 encounters with similar trajectories in unseen data.

If we are able to obtain a large enough sample of Level 3 encounters escalated from Levels 1 or 2 (either through expert judgment or model-assisted identification), we can examine what conversational events precede escalation. For example, an officer may develop reasonable suspicion in response to a citizen's answer to a question, or an officer may improperly detain a citizen who refuses to answer their questions in a common law right of inquiry. To address this question, we will create a coding scheme to categorize citizen language immediately preceding escalation and implement it algorithmically for analysis. While we acknowledge that some precursors to escalation are nonverbal (e.g. subject fleeing or an officer spotting contraband), this analysis could provide insight on officer-citizen dynamics in escalated encounters.

Concluding Remarks

The question of how officers communicate with the public is of critical legal, social, and policy importance: they literally give voice to the law. BWC recordings provide an unobtrusive means of observing how officers interact with citizens: how effectively law enforcement communicates citizens' rights, when encounters escalate, and when they might lead to citizen complaints against the police.

The analyses we propose here provide a means to address these questions at scale and with minimal cost by harnessing machine learning and expert judgment. There are huge advantages to taking a machine-learning approach to examine issues of compliance and escalation in investigative encounters. For example, once the transcripts are produced and the algorithms are set, human raters are no longer needed to judge each and every encounter. Moreover, since BWC recordings are constantly being generated by officers, our models can assess the efficacy and impact of policy change, various trainings and other interventions. Perhaps most importantly, the approach involves simply analyzing the data that is already being collected on a routine basis. That is, data collection and analysis will not require officers to do much beyond what they normally do.