Explaining the Predictive Performance of Police in Cases of Domestic Abuse*

Jeffrey Grogger

University of Chicago, London School of Economics, National Bureau of Economic Research, and Institute for the Study of Labor

Andrew Jordan

Washington University in St. Louis

Tom Kirchmaier

Center for Economic Performance, London School of Economics

Abstract

Police in England and Wales are asked to predict the likelihood of serious recidivism in domestic abuse cases, without being given much support beyond a potentially misleading questionnaire on which to base their predictions. To analyze their decisions, we first develop methods to deal with a censoring problem that arises because the officer's prediction may change the outcome she was attempting to predict. We find predictive performance to be low, even lower in some cases than what one would expect by chance. We next ask whether their predictions represent mistakes, and provide several types of confirmatory evidence. We ask how officers formulate their predictions, and find evidence of representativeness bias, overreaction, and categorization.

Keywords: Domestic abuse, risk assessment, screening problems, censoring, cognitive biases

*We thank Alex Imas and Pedro Bordalo for helpful discussions and Smriti Ganapathi and Tobias Auer for outstanding research assistance. The views expressed here are our own, and do not necessarily reflect those of Greater Manchester Police.

December 15, 2025

1. Introduction

At every call for domestic abuse, English police carry out a difficult prediction task. They conduct a risk assessment by administering a lengthy questionnaire to the victim, then classify recidivism risk as low, medium, or high. Beyond the questionnaire itself, the officers are not given much guidance, other than to use the responses and their professional judgment in forming their belief (Her Majesty's Inspectorate of Constabulary, 2014). They are not provided with much feedback on the outcome of the case (Robinson et al., 2016), and moreover, as we show below, many of the questionnaire items are misleading.

At the same time, the officers' predictions are consequential. Victims in cases designated as high-risk are provided with protective services, reducing their risk and helping to keep them safe from further abuse. Thus poor predictions can endanger victims. They can also result in a substantial misallocation of resources, since domestic abuse accounts for one of every 12 calls for service (Her Majesty's Inspectorate of Constabulary, 2014).

In this paper we do three things. First, we measure the quality of the predictions made via the risk-assessment protocol, comparing the officer's belief to the outcome of the case. Beliefs are fully observed, but outcomes may be censored, since protective services and other interventions under taken by the police may reduce the likelihood of the recidivism that the protocol seeks to predict. We propose an imputation approach to censoring, which under certain conditions lets us predict the latent, or untreated, outcome. Other researchers may find this approach useful in settings like ours where cases are not randomly assigned to decision makers. Our analysis indicates that officers struggle to predict serious recidivism.

Second, we establish that officers' predictions fall short of the rational baseline. In many cases that will end in serious recidivism, officers fail to assign protective services, and in many cases without serious recidivism, officers do assign them. For certain officers, these prediction failures are so acute that they would have had better average outcomes from assigning protective services at random. We also demonstrate the presence of prediction errors in the sense of Rambachan (2024), where officers assign predictive services to objectively low-risk cases before all high-risk cases are covered. Following Mullainathan and Obermeyer (2022), we also show that officers over-predict recidivism risk when objective risk is low but under-predict it when risk is high. We conclude that high-risk cases are going without protective services not because those services are scarce or because officers are cautious about deploying them, but rather because officers are acting irrationally when assigning them.

Finally, we ask why they make mistakes. Specifically, we look for patterns in the data that suggest what types of cognitive processes police officers may be using as they form their judgments. Since they get little feedback about the outcomes, the notion that they

act as Bayesian decision makers is essentially ruled out. We consider alternative cognitive mechanisms by means of some straightforward regression analyses, for which we exploit several unique aspects of our data. First, we observe the officers' beliefs and the features from the assessment questionnaire which they are instructed to use in forming them. We regress beliefs on those features and numerous controls, which shows how officers react to key predictors of risk which are observable to them. Second, under conditions which we discuss below, our imputation procedure allows us to consistently estimate an analagous regression of latent recidivism on the same variables. These regressions point to representativeness bias, overreaction, and categorization as sources of prediction error.

Although the facts of our setting may seem unusual in some ways, we suspect that other real-world screening problems share important features with the one we study here. Hiring managers implicitly predict the likelihood of a job candidate's success when they make hiring decisions. In the process, they may use a protocol to guide job interviews, but may receive limited input on how to translate potentially lengthy responses into hiring decisions. It may also be difficult for human resource personnel to update their beliefs, since failures on the part of new hires may be due to actions taken by their supervisors, rather than mistakes in the hiring process. Moreover, protocol items may not be well suited to the task. Google was once famous for posing trick questions to its job candidates, only to drop them after determining that they failed to predict job performance (Ramba, 2013).

This paper contributes to three literatures. The first is the literature on police responses to domestic abuse. Domestic abuse affects roughly one-fourth to one-third of all women worldwide, with serious consequences for their well-being and that of their children (World Health Organization, 2021; Bhuller et al., 2021; Aizer, 2011; Gutierrez and Molina, 2020). In places like the US and the UK, roughly half of domestic abuse cases come to the attention of police, highlighting their potential to reduce risk and enhance the safety of the victims. A substantial literature has focused on the effect of arresting the perpetrator (Amaral et al., 2021; Black et al., 2023; Chin and Cunningham, 2019; Sherman and Berk, 1984). Evidence is somewhat mixed, but in the main suggests that arresting and charging the perpetrator can reduce recidivism.

Risk assessments of the type we study here are widely used by police in domestic abuse cases. They are conducted in hundreds of police jurisdictions in the U.S., nationwide in England and Spain, and extensively in other parts of Europe (European Institute for Gender Equality, 2019; Koppa, 2024). Considering their widespread use, the literature evaluating their effectiveness is surprisingly small. Messing et al. (2014) used a difference-in-differences approach to analyze the effect of protective resources allocated on the basis of a risk assessment. Unfortunately, that study suffered from such high levels

of attrition that it is difficult to draw conclusions from it. Robinson (2006) and Robinson and Tregidga (2007) each followed DA victims at high risk of re-victimization, although neither study included a comparison group. Whinney (2015) used an exact-matching approach to compare 539 high-risk DA victims to 539 lower-risk victims. He found that the protective resources allocated to high-risk victims had no effect on recidivism.

Based on much the same sample as we analyze here, Black et al. (2023) found that the English risk assessment protocol, called DASH, had essentially no effect on serious recidivism.[1] Grogger et al. (2021) and Turner et al. (2019) report that the protocol does a poor job predicting it. Those findings motivate our interest in the cognitive processes by which officers make their risk judgments. They also have methodological implications for our work.

Our key methodological contribution is to deal with the censoring problem that may stem from police involvement with the case. In assessing risk, the officer seeks to predict whether the perpetrator will recidivate in the absence of intervention. However, in the process of intervening, the officer may convert the case from a latent true positive into an observed false positive. This means that observed performance understates true performance. Censored outcomes may also distort regression analyses of recidivism.

A similar censoring issue arises in many other screening problems. Bail judges detain defendants whom they predict to be at high risk of flight or recidivism. In doing so, they censor the outcome, meaning that we never observe whether the defendant would have fled or recidivated had he been released (Angelova et al., 2023; Arnold et al., 2022; Kleinberg et al., 2018). If a radiologist predicts the presence of pneumonia, the attending physician treats it accordingly. As a result, it remains unknown whether the patient truly had the disease (Agarwal et al., 2023; Chan et al., 2022).

There are two main differences between the cases that have appeared in the literature and ours. First, in the cases cited above, the censoring problem has been solved by virtue of random assignment of decision makers to cases. With random assignment, cases assigned a maximally lenient decision maker may be used to estimate the latent outcome for all cases.

This approach is not available to us, since our cases are not randomly assigned to police officers. However, our data include many covariates that strongly predict officers' beliefs, including the features from the risk assessment. We exploit these predictors to impute latent outcomes to cases designated as high-risk by means of causal forests. Under a conditional independence assumption, causal forests provide consistent estimates of the expectation of the latent outcome (Athey et al., 2019; Wager and Athey, 2018). They also allow us to deal with the incidental censoring problem that may arise when police

[1] DASH stands for Domestic Abuse, Stalking, and Harrassment.

decide to charge the perpetrator with a crime.

Second, the degree of censoring in our setting is much less than that in the studies cited above. In the case of the bail judge, detaining the defendant sets his probability of flight or recidivism to zero. Put differently, the intervention is 100 percent effective. With modern medical treatment, the effect of a pneumonia diagnosis is similar.

In our case, there is reason to think that any treatment effect is much smaller. As noted above, an earlier study found that the DASH risk-assessment process had essentially no effect on serious recidivism (Black et al., 2023). In the absence of any effect, the latent and observed outcomes would be the same, and one could simply carry out the analysis based on the observed outcome. That same paper found that charging the perpetrator substantially reduced serious recidivism, raising the possibility of incidental censoring. However, filing charges was far less than 100 percent effective, and only about 12 percent of perpetrators were charged with a crime. As an empirical matter, we find that results obtained based on the observed outcome are not that different than those based on imputed outcomes.

Finally, we contribute to the vast empirical literature on cognitive biases in decision making, especially regarding belief updating. Tversky and Kahneman (1974) posited that decision makers may use not the true distribution of features to form beliefs, but rather a distribution distorted by features that are particularly representative of the outcome they are trying to predict. Bordalo et al. (2016) refine that prediction, pointing out that some representative features involve more distortion than others. Features that are not only representative, but also relatively rare, are those that are more likely to seriously distort beliefs. The reason is that they move the decision maker's perceived distribution away from the truth to a greater extent than do representative features that are prevalent. Our data provide examples of both prevalent and rare representative features, and yield evidence consistent with the hypothesized interaction. We believe this study is the first to do so.

Recent work on overreaction to signals has been motivated by Benjamin's (2019) extensive literature review, which documented a predominance of under-reaction. Since then, work has focused on explaining when agents under-react, which predominates in lab settings, and overreact, which is widely observed in asset markets. We find considerable overreaction in a setting distinct from both the lab and Wall Street, which we discuss in relation to recent models by Augenblick et al. (2025)and Ba et al. (2024).

Beyond overreaction, we find that the DASH features interact strongly in influencing officers' beliefs. To rationalize these interactions, we turn to Bordalo et al. (2025), who propose a model of bottom-up attention where a decision maker (DM) seeks to predict an outcome, again forming his belief on the basis of a number of features. In their model,

the DM may attend to only a subset of the features, where the subset he focuses on may be influenced an ancillary, or trigger, feature. The model has two key predictions. One is that the realization of the ancillary feature should lead to instability, that is, sharp changes in the DM's belief. The second is that it should trigger selective attention, that is, changes in the distribution of attention paid to the various features.

We find evidence that is broadly consistent with the prediction from Bordalo et al. (2025). Officers react sharply to a victim injury, a feature which is similar to the event which they are trying to predict. A victim injury greatly increases the likelihood that the officer will designate the case as high risk, and at the same time, raises attention to other features which indicate a willingness or ability to do harm to the victim.

One limitation to our approach is common to observational settings: we have no control over the decision-maker's environment. This makes identification fundamentally more challenging than in a laboratory setting. At the same time, there is no concern about whether our results reflect the artificial environment of a lab. The behavior we observe stems from consequential decisions made under challenging circumstances.

In the next section of the paper, we provide more background on how police respond to domestic abuse calls in England and Wales. We then discuss our data. In section 4, we spell out our approach to imputing latent outcomes. The fifth section presents evidence that officers' prediction errors are the results of mistakes. Section 6 considers the cognitive processes underlying officers' predictions, and the last section concludes.

2. Background on the Police Response to Domestic Abuse Calls in England and Wales

When an English police officer appears on the scene of a domestic abuse case, her first responsibility is to prevent any further harm. Beyond that, she carries out risk assessment by means of the Domestic Abuse, Stalking, and Harassment (DASH) protocol. DASH was first developed in 2009 by a committee consisting of representatives from the police, academics, and service providers. It was drawn from earlier assessments as well as reviews of intimate partner homicide cases (Robinson, 2011).

The DASH protocol consists of two steps. The first is to administer the questionnaire, which consists of 27 items, plus an additional question asking whether the officer has used any additional information in classifying the risk presented by the case. The questionnaire appears in Table 1, along with question numbers and abbreviations which we will make use of below. The first few questions assess the victim's current situation, such as whether she was injured at the current incident, whether she is frightened, frightened about future violence, is depressed, feels isolated, and the like. Another section of several questions establishes whether the perpetrator has harmed or threatened any children,

whether the abuse is becoming more frequent or more severe, whether the perpetrator attempts to control the victim's behavior, and whether the perpetrator has previously taken actions that suggest an intent or capacity to inflict physical harm. Other questions ask about alcohol or drug problems, financial problems, and whether the perpetrator has had previous trouble with the law. At face value, all of these questions would seem to positively predict serious recidivism. Finally, the officer is asked to indicate whether any additional information has influenced her views about the case.

In the second step of the protocol, the officer designates the case as standard-, medium-, or high-risk, where high risk implies that "[t]here are identifiable indicators of risk or serious harm. The potential event could happen at any time and the impact would be serious." Officers are instructed to use both the victim's responses to the questionnaire and their own professional judgment in order to classify the risk posed by the case (Robinson et al., 2016). Other than that, they are given little guidance on how to map the question responses into risk levels (Richards, 2009).

Victims graded as high-risk are assigned to a Multi-Agency Risk Assessment Conference (MARAC), which may involve service providers from several organizations (Coordinated Action Against Domestic Abuse, 2012). The MARAC assesses the victim's needs and provides a customized package of services designed to keep her safe and provide her with time to consider her options (Her Majesty's Inspectorate of Constabulary, 2014).

Beyond safekeeping and risk assessment, the officer initiates an investigation. One question for the investigation is whether a crime took place. The criteria for labeling an incident as a crime are that, "on the balance of probability: (a) the circumstances of the victims' report amount to a crime defined by law ...; and (b) there is no credible evidence to the contrary immediately available" (Home Office, 2023). If these conditions are met, and if there is "sufficient evidence to provide a realistic prospect of conviction," the police may refer charges to the prosecutor (Her Majesty's Inspectorate of Constabulary, 2014, p. 98). Roughly 40 percent of incidents in our sample were judged to be crimes, and about 30 percent of those, or 12 percent of the full sample, resulted in charges.

Although the risk assessment and the criminal investigation are carried out simultaneously, and may make use of some of the same evidence, they answer different questions. The criminal investigation is backward looking, seeking to establish whether a law was broken. The risk assessment is forward looking, seeking to predict whether another serious incident is likely to take place. Our focus here is on the forward-looking risk assessment.

3. Data

3.1. Data sources and key variable definitions

We analyze data provided by Greater Manchester Police (GMP), a police force serving a population roughly the same size as that of the city of Chicago. During the period we study, all calls for service were recorded as an incident in GMP's command-and-control database.[2] All incidents that were determined to be domestic abuse were recorded in a separate DA database, whether they were classified as crimes or not.

The DA database is a key source of data for our study. It includes information on the date, time, and location of the incident, other characteristics of the incident, whether it was classified as a crime, whether the police pursued charges against the perpetrator, and if so, the charge referred to the prosecutor.

Each incident in the DA database can be linked to victim and suspect databases, which provide basic information about the parties involved. Because incidents involving the same parties can be linked together, we can construct measures of recidivism as well as histories of past crimes and domestic abuse. The DA file can also be linked to information about the responding officer and the DASH reports. The DA, victim, suspect, and DASH databases are the sources of the data we analyze here.

In England and Wales, domestic abuse is broadly defined to include incidents between individuals age 16 years or older who are or have been intimate partners or family members (Brown, 2020). This can include incidents between siblings, incidents between adult children and parents, or intimate-partner incidents involving current or past spouses or romantic partners. Since the great majority of domestic abuse incidents involve heterosexual intimate partners, we restrict attention to those calls.[3]

The DASH protocol asks the responding officer to predict whether the incident will result in serious recidivism. Here we define serious recidivism as a future domestic abuse incident involving either violence with injury or a sex offense that is reported to police within 12 months.[4] Our goal in defining serious recidivism this way is to capture a reasonable notion of the serious harm which DASH is intended to predict, given the offense categories recorded in the GMP data.

Violence with injury consists primarily of two classes of offenses: Wounding with Intent to do Grievous Bodily Harm (GBH) and Assault Occasioning Actual Bodily Harm

[2] In mid 2019, GMP adopted a completely new recording system. It is not backward compatible with the system that was in place during our sample period.

[3] Intimate partners include ex-partners, partners, wives, girlfriends, ex-wives, husbands, boyfriends, ex-husbands, and civil partners.

[4] We ignore multiple calls on the same day. Of course, many DA incidents are not reported to police; we are unable to analyze those incidents. Since our data come from GMP's command and control database, we may also miss any incidents that take place in another police jurisdiction.

(ABH). GBH includes injuries resulting in permanent disability; permanent, visible disfigurement; compound fractures; substantial loss of blood; lengthy treatment, or psychiatric injury. GBH clearly accords with any notion of serious harm; the maximum sentence for GBH is life imprisonment. ABH also includes injuries involving considerable harm, such as broken noses, broken fingers, loss of teeth, and shock, as well as lesser injuries such as grazes, swelling, and black eyes. It carries a maximum sentence of five years. It is important to note that violence with injury does not include Common Assault, which may entail slaps, punches, or other attacks that leave no visible mark or injury, and which carries a maximum sentence of six months (Home Office 2020).

As a measure of serious harm, ABH may seem too inclusive, whereas GBH seems too restrictive. A final statistical consideration led us to focus on violence with injury: GBH accounts for only about 1 percent of the domestic abuse cases in our sample. Analyzing such a rare event was unlikely to generate useful findings, considering the size of our sample.

3.2. Other covariates

In addition to our serious recidivism measure, the officer's belief, and features from the DASH questionnaire, our data include many predictors of both the outcome and beliefs from GMP's administrative records. These include perpetrator and victim domestic abuse histories, criminal histories, age, the location of the incident, the priority score assigned to the call for service, and characteristics of the responding police officer. We make use of these covariates both in our imputation procedure and in the regressions we estimate to analyze cognitive biases.

3.3. True and false positive rates and predictive performance

Panel A of Table 2 cross-tabulates officers' predictions with the observed outcomes, presenting cell counts and row shares. The total in the second column shows that 13,821 cases, or nine percent of the total, were classified as high-risk. The total in the second row shows that 20,003 cases, or 13 percent of the total, resulted in serious recidivism.

The observed false positive rate (FPR) is defined as the number of cases which were predicted to recidivate but did not, divided by the number of cases that did not recidivate. It equals 8.6 percent. The observed true positive rate (TPR) is defined as the number of cases which were predicted to recidivate and did, divided by the number of cases that recidivated. At 11.4 percent, the TPR is not much higher than the FPR.

Also worth noting is the positive predictive value (PPV), which is the ratio of correct positive predictions to total positive predictions. The PPV is 0.165 (= 2276/13821). Put differently, based on the observed outcome, we would conclude that 83.5 percent of the officers' high-risk designations are incorrect.

Our measure of predictive performance is the area under the curve (AUC), which is a function of the FPR and TPR. It gives the probability that the officer would rank a randomly chosen recidivist case higher than a randomly chosen non-recidivist case. That is, the AUC shows whether the officer can tell the cases apart. For this (FPR, TPR) point, the AUC is 0.514.[5] This is not much higher than 0.5, which is the AUC associated with random guessing.

We can make similar calculations for each officer, which lets us examine heterogeneity in their predictive performance. Here, we restrict attention officers who have responded to at least 100 DA incidents and plot each of their predictions in so-called ROC space. Figure 1 depicts each officer's (FPR, TPR) point by a blue plus-sign. Since these officer-level points are based on an average of roughly 141 incidents, we have used empirical Bayes methods to reduce noise by shrinking the points toward the full-sample means (Walters 2022). Typically, one would plot such points from 0 to 1 along both the x- and y-axes. We have truncated the axes because of the restricted range of our data (the raw data were similarly restricted).

Some of the points are well above the 45-degree line. However, most are fairly close to it, along which TPR = FPR and AUC = 0.5. Based on the observed data, most officers appear to predict poorly. About 10 percent lie below the 45-degree line. These officers would produce better predictions by making random guesses.

4. Accounting for censoring

Of course, we would expect measures that are based on observed outcomes to understate true predictive performance, precisely because the observed outcomes reflect the effects of efforts to alter them. If the protective resources deployed after the high-risk designation were effective, they would convert latent true positives to observed false positives. We propose to solve this censoring problem by imputing the latent outcome for cases designated as high-risk.

4.1. Conditional average treatment effects and the latent outcome

Our approach to imputation makes use of causal forests, which are an adaptation of random forests (Athey et al., 2019; Athey and Wager, 2019). Athey et al. (2019) show that under conditional independence and a number of technical conditions, the estimates from causal forests provide consistent estimates of conditional average treatment effects (CATEs).[6]

[5] Plotting the TPR against the FPR, the AUC for the point (FPR', TPR') is given by 0.5 plus the area of the triangle formed by connecting the points (0, 0), (FPR', TPR'), and (1, 1).

[6] We also took a second approach to imputation, which is simple and intuitive. We restricted attention to cases which were not designated as high-risk and where the perpetrator was not charged with a crime.

We use the CATEs to impute latent outcomes both for cases designated as high-risk and for those where charges were filed. However, to most simply illustrate how causal forests can be used to estimate the latent outcome, we first ignore the prospect of criminal charges, assuming the officer decides only whether to classify the case as high-risk. We leave the actual case, involving both the high-risk classification and the decision to refer charges, for the (forthcoming) Appendix.

Let D_{i1} denote the officer's belief. Then $D_{i1} = 1$ if the officer predicts the ith incident is likely to result in serious recidivism, that is, if she classifies it as high-risk. Otherwise, $D_{i1} = 0$. Define potential outcomes $Y_i(D_{i1})$ as a function of the officer's prediction, so $Y_i(1)$ represents the outcome for incident i if it is classified as high-risk, and $Y_i(0)$ represents the latent outcome, that is, the outcome if it is not classified as high-risk. The observed serious recidivism outcome is given by $Y_i = D_{i1}Y_i(1) + (1 - D_{i1})Y_i(0)$, which shows that we only observe the potential outcome associated with the observed risk classification. That is, we only observe the latent outcome for cases not designated as high-risk. Let Q_i denote a vector of 28 dummy variables encoding the DASH questions, and let Z_i denote a vector of control variables derived from the criminal and DA history information in GMP's administrative records. Let $X_i = [Q_i, Z_i]$.

Causal forests estimate conditional average treatment effects (CATEs), defined as

$$\tau(X_i) = E[Y_i(1)|X_i] - E[Y_i(0)|X_i]. \tag{1}$$

Rewrite the observed outcome in terms of the potential outcomes as

$$\begin{aligned} Y_i &= Y_i(0) + D_{i1}[Y_i(1) - Y_i(0)] &(2)\\ &= Y_i(0) + D_{i1}\tau_i, &(3)\end{aligned}$$

defining the idiosyncratic treatment effect τ_i. Rewriting and replacing the unknown τ_i with an estimate $\hat{\tau}(X_i)$ of $\tau(X_i)$ from the causal forest, we base estimates of the latent outcome, denoted by Y_i^*, on

$$Y_i^* = Y_i - D_{i1}\hat{\tau}(X_i). \tag{4}$$

It is straightforward to show that, under conditions that ensure the consistency of the CATEs, the imputed outcomes can be used to consistently estimate mean latent recidivism rates, true positive rates, and false positive rates.

We regressed the serious recidivism outcome, which was unaffected by the interventions, on the DASH features and other control variables. We used predicted values from this regression to impute the outcome for cases which were classified as high-risk or where charges were filed. These predictions yielded results similar to those from the CATEs.

4.2. True and false positive rates and predictive performance based on imputed data

Panel B of Table 2 cross-tabulates imputed outcomes by the officer's prediction, where we have imputed latent outcomes for cases classified as high-risk and for those where charges were referred to the prosecutor. The recidivism rate is 0.136 (= 20864/153934), slightly higher than 0.130 estimated from the observed data. Considering that the interventions are intended to reduce recidivism, we would expect the imputed recidivism rate to exceed the observed rate. In the imputed data, the true positive rate is higher, and the false positive rate is lower, than in the observed data. This raises the PPV a bit, although at 0.188 (=2605/13821), it still indicates that over four out of five high-risk designations are incorrect. Similarly, the AUC based on the imputed data, at 0.520, is only a bit higher than that based on the observed data.

In Figure 1, each officer's (FPR, TPR) point, based on the imputed data, is depicted by a red asterisk. The cloud formed by these points lies above and to the left of that formed by the observed data. The latent FPR is lower, and the latent TPR higher, than their observed counterparts for almost every officer.

The small difference between observed and latent prediction rates raises the question of why imputation makes so little difference. The reason has to do with the effectiveness and prevalence of the interventions. As noted above, the high-risk designation has been found to have little effect on average. Pressing charges, which is correlated with the high-risk designation, has a more substantial effect, equal to roughly 5 percentage points, or about 40 percent of the mean recidivism rate. However, charges were levied in only about 12 percent of the cases. In applications where the intervention is more effective, imputation would have a greater quantitative effect. For example, when a bail judge detains a defendant before trial, he virtually eliminates the possibility that the defendant will fail to appear in court. When the interventions are less effective, the resulting censoring is less severe. As a result, our findings below are largely the same whether we analyze the observed or imputed data.

4.3. Evidence on conditional independence

Finally, we discuss evidence in support of the key conditional independence assumption needed for our the causal forest approach to consistently estimate latent recidivism rates and associated magnitudes. Conditional independence requires the potential outcomes to be independent of the high-risk classification, given the predictors. Put differently, it requires police officers' predictions to be effectively random, given the observed data. Black et al. (2023), who studied the effects of the risk assessment and criminal charges on serious recidivism, argue that these data may satisfy this condition on a number of grounds. The first is that officers are directed to use the responses to the DASH

questionnaire, plus their professional judgment, to predict the risk facing the victim. On the basis of this important institutional feature, they expect (and show) that the DASH items explain considerable variation in the high-risk designation. Likewise, for estimating the propensity of being charged, several of the DASH items are strong indicators that a crime has been committed, which is a pre-condition for charges to be filed.

Second, they provide several sets of specification tests in support of this condition. First, they estimate the propensity score for the high-risk designation using several different methods, and show that, regardless of the method used, the propensity score greatly reduces imbalance in X_i between the high-risk and lower-risk groups. For several specifications, it eliminates the imbalance. Moreover, the treatment effects they estimate are largely invariant to the choice of propensity score. They then carry out "out-of-model" balance tests using variables which are highly predictive of the outcome, but which were not used to estimate the propensity score. Again, the estimated propensity scores greatly reduce imbalance on these variables, in some cases eliminating it. Finally, they estimate placebo regressions, showing that the interventions do not affect lagged values of the outcome. We refer the reader to Black et al. (2023) for more details.

5. Mistakes or preferences?

Above we have framed officers' poor predictive performance as stemming from mistakes. An alternative explanation is that officers are acting rationally but have inadequate information or perceive a high cost of false positives relative to false negatives, perhaps because the supply of protective services is constrained. Here, however, we discuss three tests that suggest that officers are indeed making mistakes that take them away from the rational baseline.

5.1. Location in ROC space

Officers assign protective services in only nine percent of cases, which is lower than the rate of serious recidivism in the sample and leads to a high rate of false negatives. Failing to cover every case of serious recidivism is not necessarily a decision-making failure if it is driven by preferences that place a high relative weight on false positives or a simple constraint on how many cases can be given protective services.

However, we find that many officers pair a high rate of false negatives with a high rate of false positives.[7] These officers fall below the 45-degree line in Figure 1. Recall that points on the 45-degree line can be achieved by officers who randomly assign protective

[7]Figure 1 depicts false positive rates (FPR) and true positive rates (TPR). The false negative rate is given by FNR = 1 - TPR, so low TPR implies high FNR.

services at the corresponding rate. Thus, these points are accessible even to officers who have no useful information to aid their decision. For any officer who values true positives and devalues false positives, there are points on the 45-degree line that dominate any point below it. Thus, the officers below the 45-degree line must be actively misusing the information available to them.

5.2. Misallocation by Risk Group

Another way to think about the shift from a point below the 45-degree line in ROC space to a point on the line is as a series of exchanges. Given a case assigned protective services in the original allocation, the officer is better off removing protective services from that case and giving them to a randomly selected case that does not yet have them. Following an insight from Rambachan (2024), this idea of exchanges can be extended to multiple well-defined groups of cases. Suppose for example that an officer works in both a high-risk neighborhood and a low-risk neighborhood, and that in the high-risk neighborhood, even cases that are not assigned protective services have a higher recidivism rate than cases in the low-risk neighborhood that do get protective services. In this hypothetical, the officer could improve average outcomes by shifting protective services from the low-risk neighborhood to the high-risk neighborhood, even if the shifted cases are selected at random.

In Figure 2, we look for mistakes of this kind in our data. Here, "neighborhoods" are quintiles of predicted recidivism risk. We obtain these predictions by regressing recidivism on the DASH features.[8] We find that officers could be taking protective services from cases in the bottom 4 quintiles and assigning them to un-protected cases in higher quintiles. These exchanges are available to any officer who understands the relationship between the variables in our prediction regression, which are known to her, and recidivism. Our findings instead suggest that officers are misinterpreting that information.

5.3. Simultaneous over- and under-prediction

Our third test for mistakes comes from Mullainathan and Obermeyer (2022), who show that rational officers will not under-predict recidivism in some situations but over-predict it in others. This is because rational officers will classify cases as high-risk according to a threshold rule in recidivism risk. If this threshold is set too low, officers many generally over-predict recidivism, and likewise under-predict if the threshold is set too high. Seeing both errors in different circumstances implies different thresholds for each.

To test this, we plot both recidivism and officer's beliefs about recidivism by the objective risk quintiles described above. See Figure 3. We see a seemingly small amount

[8]From here on, we use the term "recidivism" to mean "latent serious recidivism."

of over-prediction at the lowest quintile and increasing levels of under-prediction at higher quintiles.

6. What explains officers' predictive performance?

All three of the analyses above indicate that officers are making mistakes. The key question is: why? Our main objective in this section is to understand what factors underlie officer's poor predictive performance.

To do so we start by estimating two regressions, which we write as

$$D_{i1} = Q_i\alpha + Z_i\phi + \varepsilon_i \tag{5}$$

and

$$Y_i^* = Q_i\beta + Z_i\psi + u_i \tag{6}$$

where Y_i^* is the latent outcome as defined in the (forthcoming) Appendix, ε_i and u_i represent unobservable disturbance terms, and α, β, ϕ, and ψ represent vectors of regression coefficients to be estimated. Equation (5) regresses the officer's belief, that is, her high-risk designation, on dummy variables for all of the DASH features which she is instructed to use in making her prediction. It also includes numerous control variables constructed from GMP's administrative data bases. These consist of indicators for the dyad's and perpetrator's criminal and domestic abuse histories, as well as age and other characteristics of the incident not reported on the DASH form. Equation (6) regresses latent recidivism on the same variables.[9] All of these regressions also include missing variable dummies for each DASH feature. These equal one if the question was omitted and equal zero otherwise.

6.1. Bad information

Figure 4 displays estimates of α and β from equations (5) and (6). Coefficients from the beliefs regression are depicted by gray bars; those from the latent recidivism regression are in maroon. Ninety-five percent confidence intervals are shown as vertical lines centered on the end of each bar.[10]

There is a substantial divergence between the predictors of beliefs and the predictors of the latent outcome. The vast majority of the coefficients in the beliefs regression

[9]For comparison purposes, we also regress the observed outcome on the same set of variables. Results are reported in Appendix Table A1.

[10]Columns (1) and (3) of Appendix Table A1 report these coefficients and their standard errors, as well as those from the regression based on the observed outcome in column (2). Throughout the paper, standard errors are clustered at the level of the dyad. Those clustered at the level of the police officer were similar.

are positive and most are significant. Those that are negative are generally small. In contrast, thirteen of the coefficients in the latent outcome regression are negative, and most of those are significant.

This points to a problem with questionnaire design. A prima facie reading of the questions in Table 1 suggests that each of the questions should positively predict serious recidivism. Yet half the questions are counter-predictive, negatively predicting serious recidivism risk.

Since the officers receive little feedback on their predictions, it is largely impossible for them to learn which items predict negatively. Even with regular feedback, it would be very difficult to distinguish which features were counter-predictive in a questionnaire with so many items. Poor questionnaire design is making an already difficult prediction problem that much harder.[11]

Questionnaire design notwithstanding, the DASH features strongly influence officers' beliefs. Indeed, the coefficients from the beliefs regression tend to be larger than those from the latent outcome regression, even when both coefficients are positive. The DASH features with particularly strong effects on beliefs include the indicators for whether the victim was injured (Question 1); whether the perpetrator has threatened the children (Question 12), and whether the perpetrator has threatened to kill the victim (Question 17). Also noteworthy are indicators for whether the victim felt isolated (Question 4), whether the perpetrator had ever used a weapon (not necessarily at the current incident; Question 16), and whether the perpetrator had ever humiliated the victim sexually (Question 19).

6.2. Representativeness bias

Here we ask whether cognitive biases may help explain officers' reactions to the DASH features. We first consider representativeness bias. Tversky and Kahneman (1974) propose that decision makers may form beliefs on the basis of features that most readily come to mind, such as features that are most representative of the event that the decision-maker is trying to predict. Bordalo et al. (2016) point out that not all representative features are equally problematic. Features that are representative but also prevalent in the population may involve little distortion. In contrast, features that are representative but rare may distort decisions considerably. The reason is that putting more weight on a prevalent feature may not move the decision-maker's perceived distribution much, relative to the

[11] This is not an issue of collinearity. In Appendix Table A2, we present coefficients from a sequence of bivariate regressions of the dependent variables on the DASH questions. Comparing those coefficients with their counterparts in Appendix Table A1, we see that most of the variables with negative coefficients in the multivariate regressions also have negative coefficients in the bivariate regressions.

true distribution of features. In contrast, placing more weight on a rare feature may lead to a large divergence between the perceived and actual distributions.

Following Tversky and Kahneman (1974) and Bordalo et al. (2016), we define the representativeness of a feature as the ratio of group-specific means in the two groups. That is, the representativeness of one of the DASH features is equal to the mean of the feature in the recidivist group divided by the mean in the non-recidivist group.

Figure 5 displays the representativeness and prevalence of each of the DASH features. The figure is sorted by the features' representativeness, which is depicted by gray bars that increase in magnitude as one moves to the left from the value 0.5. Prevalence is depicted in maroon bars, which increase as one moves to the right from 0.

Twelve of the features have representativeness greater than one. These include all features down to Victim Pregnant (Question 9), whose representativeness is 1.004. Fortunately for our strategy, these features vary in their prevalence. Ten of them are relatively rare, with prevalence rates below 0.17. Two of them, perpetrator involvement with alcohol/drugs (Question 24) and prior trouble with the police (Question 27), are much more common, with prevalence rates of 0.38 and 0.45. This gives us the chance to look for the key interaction predicted by Bordalo et al. (2016). According to that model, the ten features that are representative and rare should move officers' beliefs, but the two that are representative and prevalent should not.

Results are presented in Figure 6, which excerpts the subset of prevalent features from Figure 4 (and columns (1) and (3) of Appendix Table A1). As above, coefficients from the beliefs regression are depicted by gray bars and those from the latent recidivism regression are in maroon. Here we have reordered the features so that the estimates for the 10 rare features precede those for the two prevalent features.

All of the coefficients on the representative and rare features move officers' beliefs. They are all positive and significant and several are quite large in absolute terms as well. A victim injury (Question 1) raises the likelihood that the officer predicts serious recidivism by 0.066 (standard error = 0.004). If the victim reports feeling isolated (Question 4), the officer's belief increases by 0.046 (0.003). If the perpetrator has ever used a weapon (Question 16), it rises by 0.057 (0.004). A perpetrator having strangled the victim, hurt a third party, or breached a restraining order (Questions 18, 21, and 26) raises the likelihood of a high-risk designation by about three percentage points. These are all large magnitudes in relation to the nine percent share of cases that are designated as high-risk.

They are also large in relation to the coefficients of these variables in the latent outcome regression. A victim injury raises the likelihood of serious recidivism by 0.015 (0.004), whereas an isolated victim increases it by 0.018 (0.003). Prior weapon use by the perpetrator has essentially no effect on serious recidivism. The strangling and third-

party harm coefficients (Question 18 and 21) have small positive coefficients, whereas the coefficient on breaching a restraining order is actually negative.

In contrast to the sizable beliefs coefficients on the features that are representative and rare, those on the features that are representative and prevalent are negative and significant but quite small. The coefficient on perpetrator alcohol/drug problems (Question 24) is -0.007 (0.002), and the one on prior trouble with the police (Question 27) is -0.004 (0.001). These features move officers' beliefs much less than do the features which are representative and rare, which is broadly consistent with the prediction of the Bordalo et al. (2016) model.

The finding regarding alcohol and drug use is particularly important. Alcohol and drugs present problems for nearly 40 percent of perpetrators in the sample. They also predict serious recidivism, increasing it by 0.014 (0.002). This amounts to about 10 percent of the baseline recidivism rate. Nevertheless, police ignore this feature in forming their beliefs. The result is false negatives, that is, victims who are at risk but who are not provided with protective services.

Thus representativeness bias helps explain the magnitude of several of the coefficients that stood out in the beliefs regression in Figure 4. These include victim injury (Question 1), victim isolation (Question 4), weapons use (Question 16), and attempted strangulation (Question 18). However, representativeness bias cannot help explain the magnitude of one of the most sizable coefficients, namely that for death threats to the victim (Question 17). The reason is that those death threats are not representative of serious recidivism. As a result, we consider other models of over-reaction to signals.

6.3. Overreaction

In Augenblick et al. (2025), cognitive noise makes it difficult for the DM to assess the strength of a signal. Ba et al. (2024) combine cognitive noise with representativeness. Augenblick et al. (2025) find evidence of overreaction to weak signals, such as Question 17, because cognitive noise leads them to adjust according to the "typical" signal strength, which is greater than the actual strength. In the case of binary outcomes, such as ours, Ba et al. (2024) find overreaction in the presence of asymmetric priors, particularly in response to signals that should move the DMs belief away from the more likely outcome. This seems particularly germane to our setting, for three reasons. First, the latent recidivism rate is 13 percent, meaning that the likelihood of not recidivating is 87 percent. Even if these numbers are not known exactly to police officers, one would expect their priors to be highly asymmetric. Second, the officers clearly perceive that affirmative responses to the DASH questions raise the likelihood of recidivism, the less likely outcome. Finally, a death threat is a weak signal, with representativeness less than

one (see Figure 5).[12]

6.4. Categorization and selective attention

The discussion above indicates that representativeness bias, cognitive noise, and weak signals may help explain key patterns in the coefficients of the beliefs regression presented in Figure 4. Here we consider the role of categorization and selective attention. This model may explain not only the findings from Figure 4, but also an important set of interactions among the DASH features.

Bordalo et al. (2025) present a model where a DM seeks to predict an outcome, forming her belief on the basis of a number of features. However, the DM may pay attention to only a subset of the available features, where the subset she focuses on may be influenced an ancillary, or trigger, feature. The ancillary feature may or may not be directly predictive of the outcome the DM seeks to forecast. The key prediction from the model is that the realization of the trigger feature should both lead to instability in the DM's belief and affect the distribution of features to which she attends.

To carry out this analysis requires us to specify the ancillary feature. We choose the feature on the basis of its similarity to the outcome. Similarity provides a natural approach, since it plays a fundamental role in psychological theories of knowledge and behavior (Tversky, 1977). It has played a key part in general models of categorization, both conceptual and computational (Nosofsky, 1986; Kruschke, 1992). A recent example comes from Evers et al. (2022), who theorize and provide evidence showing that similarity underlies the formation of categories on which mental accounting is based.

Although measures of similarity entail an exercise of judgment, in our case we believe that a victim injury (Question 1) is the DASH feature most similar to the outcome the officer is attempting to predict. The goal of DASH is to predict "serious harm" to the victim. In cases where the victim has been injured, it is easy to imagine that the officer perceives her goal as to prevent the same thing from happening again. The salience of the question, as the first item on the questionnaire, likely helps to reinforce its similarity to serious recidivism.

Table 3 cross-tabulates recidivism and the officer's belief by whether the victim was injured at the current incident. The top panel shows that injury is predictive of recidivism, which rises from 12 percent among cases without injury to 17.1 percent among cases where the victim was injured. The bottom panel shows substantial instability in officers' beliefs. They classify only 6.5 percent of cases with no victim injury as high-risk, compared to 26.5 percent of cases where the victim has been injured.

[12]Over-reaction to disconfirming signals has also been reported by Kieren et al. (2024).

Figure 7 addresses the question of selective attention arising from victim injuries. It reports coefficients from two beliefs regressions. One regression, in maroon, is estimated from the subsample of cases in which the victim was injured. The other, in gray, is from cases in which she was not. The victim injury raises officers' attention to several of the other features, including whether the victim is frightened (Question 2), whether the perpetrator has used weapons (Question 16), and whether the perpetrator has ever strangled the victim (Question 18). Most notable is Question 17, indicating whether the perpetrator has ever threatened to kill the victim. In summary, a victim injury at the current incident raises officer attention to other features that indicate a willingness to harm the victim.

In Figure 7, a past death threat (Question 17) is the feature that most influences officers beliefs. It also bears similarity to the event the officer is trying to predict, suggesting a willingness to seriously harm the victim. In Table 4, we cross-tabulate recidivism and the officers' beliefs with death threats. Death threats are not very predictive of recidivism, but they are highly predictive of officers' beliefs. Officers classify 35.8 of cases as high-risk if the perpetrator has ever threatened to kill the victim, but so classify only 6 percent of the cases if he has not.

Figure 8 reports regression coefficients from samples split on Question 17. A prior death threat by the perpetrator raises officer attention to several features. The greatest increase involves a victim injury at the current incident (Question 1), mirroring how a victim injury most raised attention to death threats. It also raises attention to features such as whether the victim is frightened (Question 2), isolated (Question 4), or whether the abuse is getting worse (Question 14). All this suggests that officers categorize cases as either threatening or routine on the basis of a victim injury or death threat. Threatening cases in turn raise attention to other features that indicate a willingness to harm the victim or victim vulnerability.

Thus our data yield evidence that is consistent with categorization and selective attention along the lines of Bordalo et al. (2025). This helps explain the overreaction observed in Figure 4. In a regression without interactions, the constrained coefficient will be a rough weighted average of the underlying interacted coefficients. A large coefficient in the presence of the interaction will contribute to a large coefficient in the constrained regression with no interaction terms.

7. Robustness

The consistency of our regression estimates requires that the unobservables in equations (5) and (6) be uncorrelated with the regressors, including the DASH features. We consider two partial tests of this assumption: one based on police officer fixed effects

and another that makes use of Question 28, where the officer indicates whether she used additional information not captured by the DASH features to classify risk.

Since officers are not randomly assigned to incidents, one might be concerned about selective sorting of police officers to cases. In that case, there would be correlation between officer-specific unobservables and the DASH features. To check this possibility, we added officer fixed-effects to the regressions in equations (5) and (6). Results are reported in columns (4) and (5) of Appendix Table A1. The estimates of both equations are very similar to those from specifications that omit the officer fixed-effects, reported in columns (1) and (3). We conclude that selective sorting of officers to incidents is not biasing our estimates.

Even if officer-specific unobservables are not an issue, unobservables that vary across incidents remain a potential problem. We use Question 28 to construct a rough test for bias due to potentially correlated unobservables. In Question 28, the officer tells us whether she used information other than that on the DASH form to classify the risk level of the incident. If we could take the answer literally, the test would be simple: we would just drop the observations for which Question 28 = 1 from the sample. Estimates based on the subsample for which Question 28 = 0 would yield consistent estimates, since the officers would have used no other information in those cases to classify risk.

Unfortunately, the evidence suggests that we cannot interpret Question 28 literally. When we estimate our regressions from the subsample for which Question 28 = 0, many of the coefficients on the control variables remain significant. Nevertheless, it might be reasonable to conclude that officers make less use of unobservables when Question 28 = 0 than when Question 28 = 1. In this case, comparing estimates from the full sample to those from the subsample for which Question 28 = 0 may give us a rough sense of the degree to which correlated unobservables drive the findings reported above.

Estimates are presented in columns (6) and (7) of Appendix Table A1. For the most part, these are similar to the coefficients from the full sample in columns (1) and (3). Since these exercises do little to change the coefficients in the beliefs regressions, our conclusions regarding representativeness bias remain unchanged.

To check the robustness of our selective attention results, we estimate these two additional specifications by the presence and absence of the various trigger features. Estimates are shown in Appendix Table A4. For both specifications, the key patterns persist.

Finally, although the implications of representativeness bias and selective attention focus on the the beliefs regressions, one might wonder how the recidivism regressions are affected by our imputation strategy. Comparing columns (2) and (3) of Appendix Table A1 shows that most of the coefficients are similar whether the dependent variable is the

observed or latent (i.e., imputed) outcome measure. We attribute this to the fact that the risk assessment process had little effect on serious recidivism, and that the effect of criminal charges, although larger, affected only 12 percent of our sample.

8. Heterogeneity

Here we consider how various dimensions of officer heterogeneity affect how officers' beliefs respond to the DASH features. We start by estimating beliefs regressions by officer sex. Results are reported in Appendix Figure A1, where coefficients for female officers are in gray and those for males are in maroon. We see that female officers are slightly more attentive to victim injuries (Question 1), threats toward children (Question 12), perpetrator's use of weapons (Question 16), and sexual humiliation (Question 19), but the differences are not very large, and most of the differences are not significant.

We next consider officer experience, first dividing the sample according to whether the officer's experience was weakly less than or greater than the the sample median. Appendix Figure A2 displays coefficients from the corresponding beliefs regressions, with those for less experienced officers in gray and those for more experienced officers in maroon. Most differences in attention are quite small. This is consistent with the notion that officers do not get much feedback about their prediction, and would face a difficult learning environment even if they did, due to the presence of counter-predictive questions. In Appendix Figure A3 we divide the sample by a different measure of experience, namely whether the officer had more or less than the median number of DA calls during our sample period. Again, differences in attention are small.

Finally, we consider differences by skill group. We define the low-skill group as officers with (FPR, TPR) points below the 45-degree line in Figure 1. The higher-skill group consists of officers above the 45-degree line. With these definitions, we restrict attention to officers with at least 100 calls for service involving DA.

Appendix Figure A4 displays coefficients from the skill-level-specific beliefs regressions. Coefficients for higher-skill officers are in gray and those for lower-skill officers are in maroon. Although there are a few exceptions, we see that lower-skill officers generally overreact more to the DASH features than higher-skill officers to the DASH features. Prominent examples include victim injury (Question 1), whether the victim is frightened (Question 2), harm to children (Question 11), death threats (Question 17), choking (Question 18), and harm to animals (Question 22). Most of the differences in those coefficients between groups are significant.

Comparing the coefficients of these beliefs regressions with those from the corresponding recidivism regressions reported in Appendix Figure A5 draws another distinction between higher- and low-skill officers. The last pair of coefficients in the beliefs regressions

(Appendix Figure A4) show that both sets of officers respond similarly to information not on the DASH form, as collected by Question 28. However, that information predicts recidivism very differently according to officer skill. In the last pair of coefficients in the recidivism regressions (Appendix Figure A5), we see that that information positively predicts recidivism for the higher-skill officers, but negatively predicts it for their low-skill counterparts. This is analogous to a finding by Angelova et al. (2023). In their study of release decisions by bail judges, they found that when higher-skill judges overrode recommendations from a predictive algorithm, they did so on the basis of private information that positively predicted an adverse outcome. In contrast, lower-skill judges were more likely to base decisions on less predictive demographic characteristics.

Finally, we compare differences in selective attention between higher- and low-skill officers. To do so, we estimate separate beliefs regressions according to whether the victim was injured at the current incident for each of the groups. Results are reported in Figure 9 for higher-skill officers and Figure 10 for low-skill officers. In both cases, coefficients from the sample with no victim injury are in gray, whereas those from cases with a victim injury are in maroon.

First, the features involved in selective attention differ somewhat between the two groups. Among low-skill officers, a victim injury increases attention to whether there are children in the household who are not the perpetrator's (Question 10), and whether the perpetrator has sexually humiliated the victim (Question 19). Among higher-skill officers, victim injuries increase attention toward threats to children (Question 12), and whether the perpetrator has harmed a third party (Question 20) or an animal (Question 22).

To features such as whether the victim is frightened (Question 2) or the perpetrator has threatened the victim (Question 17), low-skill officers are more attentive than their higher-skill counterparts, even in the absence of a victim injury. At the same time, the increase in attention in response to a victim injury is greater for the low-skill than the higher-skill officers. In response to a victim injury, the coefficient on whether the victim is frightened rises from 0.054 (0.016) to 0.124 (0.052) for the low-skill officers, but from 0.017 (0.004) to 0.064 (0.016) for the higher-skill officers. The corresponding changes to the death threats coefficients are from 0.106 (0.020) to 0.195 (0.045) for the low-skill officers and from 0.073 (0.007) to 0.154 (0.015) for their higher-skill counterparts. Although these analyses are limited by the small number of low-skill officers, they suggest that differences in skill stem in part from differences in selective attention.

9. Conclusion

Police officers in England and Wales are charged with a difficult task: predicting serious recidivism in domestic abuse cases. Beyond a poorly designed questionnaire, they are given little guidance and little feedback. Their situation is similar to that of workers in other settings who are asked to solve important screening problems without receiving much support.

Analyzing how workers in such settings make their predictions requires not only data on beliefs and outcomes, but also a means for adjusting observed outcomes to account for the fact that they may be affected by the prediction the decision-maker is asked to make. We propose using causal forests to make that adjustment. We suspect this approach would be useful not only here, but in the presumably numerous settings where decision-makers are not randomly assigned to cases.

Results based on both adjusted and unadjusted outcomes suggest that the typical officer's prediction is just a bit better than random. We show that these poor predictions stem from mistakes rather than preferences. To understand why officers make mistakes, we consider various types of cognitive biases. We test a key implication of representativeness bias, which is that the representativeness of features should interact with their prevalence in influencing decision-makers' beliefs (Bordalo et al., 2016). We find strong evidence of this interaction. Officers both overreact to features which are representative but rare, and underreact to features which are representative and prevalent. One implication of this is that officers largely ignore prior alcohol and drug problems among perpetrators, which are prevalent in our sample and raise the objective risk of recidivism. This contributes to false negative predictions.

We also observe overreaction to features which are not representative of serious recidivism. Overreaction in these cases may arise due to cognitive noise, as in Augenblick et al. (2025), or due to the presence of weak signals and an asymmetric prior, as in Ba et al. (2024).

We also find evidence of selective attention in officer decision making. When the victim is injured, officers are much more likely to classify the case as high risk. They are also much more attentive to other features which which indicate a willingness to harm the victim. The strongest increase in attention involves death threats against the victim, which in turn raise attention to victim injuries. These findings are largely consistent with Bordalo et al. (2025) model of categorization. Such elevated attention greatly exceeds the effects that the features have on recidivism risk, leading to false positive predictions.

Our findings have important implications for how one designs predictive questionnaire instruments. The first is obvious, which is that counter-predictive questions should be avoided. The second draws from the notion of selective attention. Some questionnaire

items may change the decision maker's attention, causing her to focus on only a subset of the items, or alternatively to focus too much attention on them. If so, the question could lead to poor predictive performance, even if it performs well on its own. It has long been standard practice in the survey design literature to test the functioning of individual questionnaire items (Sudman and Bradburn, 1982). However, in the presence of selective attention, the performance of the whole may be less than the sum of its parts. An effective instrument requires good performance from the items as an ensemble, rather than just good performance of individual items.

References

Agarwal, M., Moehring, A., Rajpurkar, P., Salz, T., 2023. Combining human expertise with artificial intelligence: experimental evidence from radiology. Technical Report. URL: http://www.nber.org/papers/w31422.

Aizer, A., 2011. Poverty, violence, and health: The impact of domestic violence during pregnancy on newborn health. Journal of Human Resources 46, 518–538. doi:10.3368/jhr.46.3.518.

Amaral, S., Bhalotra, S., Prakash, N., 2021. Gender, Crime and Punishment: Evidence from Women Police Stations in India.

Angelova, V., Dobbie, W., Yang, C., 2023. Algorithmic recommendations and human discretion. Technical Report.

Arnold, D., Dobbie, W., Hull, P., 2022. Measuring racial discrimination in bail decisions. American Economic Review 112, 2992–3038.

Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. Annals of Statistics 47, 1148–1178.

Athey, S., Wager, S., 2019. Estimating Treatment Effects with Causal Forests: An Application. Observational studies 5, 1–10.

Augenblick, N., Lazarus, E., Thaler, M., 2025. Overinference from weak signals and underinference from strong signals. The Quarterly Journal of Economics 140, 335–401. doi:10.1093/qje/qjae032.

Ba, C., Bohren, J.A., Imas, A., 2024. Over- and underreaction to information Working paper.

Benjamin, D.J., 2019. Errors in probabilistic reasoning and judgment biases, in: Bernheim, B.D., DellaVigna, S., Laibson, D. (Eds.), Handbook of Behavioral Economics: Applications and Foundations 1. Elsevier. volume 2, pp. 69–186.

Bhuller, M., Dahl, G.B., Løken, K.V., Mogstad, M., 2021. Consequences of Domestic Violence for Victims and their Families.

Black, D., Grogger, J., Kirchmeier, T., Sanders, K., 2023. Criminal charges, risk assessment, and violent recidivism in cases of domestic abuse. Technical Report. URL: https://www.nber.org/papers/w30884.

Bordalo, P., Coffman, K., Gennaioli, N., Shleifer, A., 2016. Stereotypes. Quarterly Journal of Economics , 1753–1794.

Bordalo, P., Gennaioli, N., Lanzani, G., Shleifer, A., 2025. A Cognitive Theory of Reasoning and Choice. Working Paper 33466. National Bureau of Economic Research.

Brown, K., 2020. Domestic Abuse. URL: https://www.cps.gov.uk/crime-info/domestic-abuse.

Chan, D.C., Gentzlow, M., Yu, C., 2022. Selection with variation in diagnostic skill: evidence from radiologists. Quarterly Journal of Economics , 729–783.

Chin, Y.M., Cunningham, S., 2019. Revisiting the effect of warrantless domestic violence arrest laws on intimate partner homicides. Journal of Public Economics 1, 1–10.

Coordinated Action Against Domestic Abuse, 2012. A place of greater safety. Technical Report. Coordinated Action Against Domestic Abuse. URL: www.caada.org.uk/commissioning.

European Institute for Gender Equality, 2019. Risk assessment and management of intimate partner violence in the EU. Technical Report. European Institute for Gender Equality. Vilnius, Latvia.

Evers, E.R.K., Imas, A., Kang, C., 2022. On the role of similarity in mental accounting and hedonic editing. Psychological Review 129, 777–789. doi:10.1037/rev0000325.

Grogger, J., Gupta, S., Ivandic, R., Kirchmaier, T., 2021. Comparing Conventional and Machine-Learning Approaches to Risk Assessment in Domestic Abuse Cases. Journal of Empirical Legal Studies 18, 90–130.

Gutierrez, I.A., Molina, O., 2020. Does domestic violence jeopardize the learning environment of peers within the school? Peer effects of exposure to domestic violence in urban Peru. doi:10.1016/j.econedurev.2021.102147.

Her Majesty's Inspectorate of Constabulary, 2014. Everyone's business: Improving the police response to domestic abuse. Technical Report. Her Majesty's Inspectorate of Constabulary. London.

Home Office, 2023. Crime recording rules for front line officers and staff. Technical Report. URL: https://www.gov.uk/government/publications/counting-rules-for-recorded-crime.

Kieren, P., Müller-Dethard, J., Weber, M., 2024. Disconfirming information and overreaction in expectations. SSRN Electronic Journal URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3644226. sSRN Working Paper No. 3644226.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S., 2018. HUMAN DECISIONS AND MACHINE PREDICTIONS. The Quarterly Journal of Economics 133, 237–293. doi:10.1093/qje/qjx032.Advance.

Koppa, V., 2024. Can information save lives? effect of a victim-focused police intervention on intimate partner homicides. Journal of Economic Behavior and Organization 217, 756–782. doi:10.1016/j.jebo.2023.11.010.

Kruschke, J.K., 1992. ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review 99, 22–44. doi:10.1037/0033-295X.99.1.22.

Messing, J.T., Campbell, J., Wilson, J.S., Brown, S., Patchell, B., University, A.S., Work, S.o.S., 2014. Police Departments' Use of the Lethality Assessment Program: A Quasi-Experimental Evaluation. Technical Report. Report to the National Institute of Justice. URL: https://www.ncjrs.gov/pdffiles1/nij/grants/247456.pdf.

Mullainathan, S., Obermeyer, Z., 2022. Diagnosing physician error: A machine learning approach to low-value health care. The Quarterly Journal of Economics 137, 679–727. doi:10.1093/qje/qjab046.

Nosofsky, R.M., 1986. Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General 115, 39–57. doi:10.1037/0096-3445.115.1.39.

Ramba, S., 2013. Identifying Prediction Mistakes in Observational Data. abcnews.com URL: https://abcnews.go.com/blogs/business/2013/06/google-skips-waste-of-time-brainteaser-interview-questions.

Rambachan, A., 2024. Identifying prediction mistakes in observational data. The Quarterly Journal of Economics 139, 1665–1711. doi:10.1093/qje/qjae013.

Richards, L., 2009. Domestic Abuse, Stalking and Harassment and Honour Based Violence Risk Identification Checklist. URL: http://www.dashriskchecklist.co.uk/uploads/V-DASH2010-2015.pdf.

Robinson, A.L., 2006. Reducing Repeat Victimization Among High-Risk Victims of Domestic Violence. Violence Against Women 12, 761–788.

Robinson, A.L., 2011. Risk and intimate partner violence, in: Kemshall, H., Wilkinson, B. (Eds.), Good Practice in Assessing Risk: Current Knowledge, Issues and Approaches. Jessica Kingsley Publishers, London and Philadelphia. chapter 7, pp. 119–138.

Robinson, A.L., Myhill, A., Wire, J., Roberts, J., Tilley, N., 2016. Risk-led policing of domestic abuse and the DASH risk model. Technical Report. College of Policing. URL: http://www.college.police.uk/News/College-news/Documents/Risk-led{_}policing{_}of{_}domestic{_}abuse{_}and{_}the{_}DASH{_}risk{_}model.pdf.

Robinson, A.L., Tregidga, J., 2007. The perceptions of high-risk victims of domestic violence to a coordinated community response in Cardiff, Wales. Violence Against Women 13, 1130–1148. doi:10.1177/1077801207307797.

Sherman, L.W., Berk, R.A., 1984. The Specific Deterrent Effects of Arrest for Domestic Assault Author (s): Lawrence W . Sherman and Richard A . Berk Published by : American Sociological Association Stable URL : http://www.jstor.org/stable/2095575 ARREST FOR DOMESTIC ASSAULT *. American Sociological Review 49, 261–272. URL: https://www.jstor.org/stable/2095575.

Sudman, S., Bradburn, N., 1982. Asking questions: a practical guide to questionnaire design. Jossey-Bass, Hoboken, NJ.

Turner, E., Medina, J., Brown, G., 2019. Dashing Hopes? the Predictive Accuracy of Domestic Abuse Risk Assessment by Police. The British Journal of Criminology 59, 1013–1034. doi:10.1093/bjc/azy074.

Tversky, A., 1977. Features of similarity. Psychological Review 84, 327–352. doi:10.1037/0033-295X.84.4.327.

Tversky, A., Kahneman, D., 1974. Judgment under Uncertainty : Heuristics and Biases Linked references are available on JSTOR for this article : Judgment under Uncertainty : Heuristics and Biases. Science 185, 1124–1131. doi:10.1126/science.185.4157.1124.

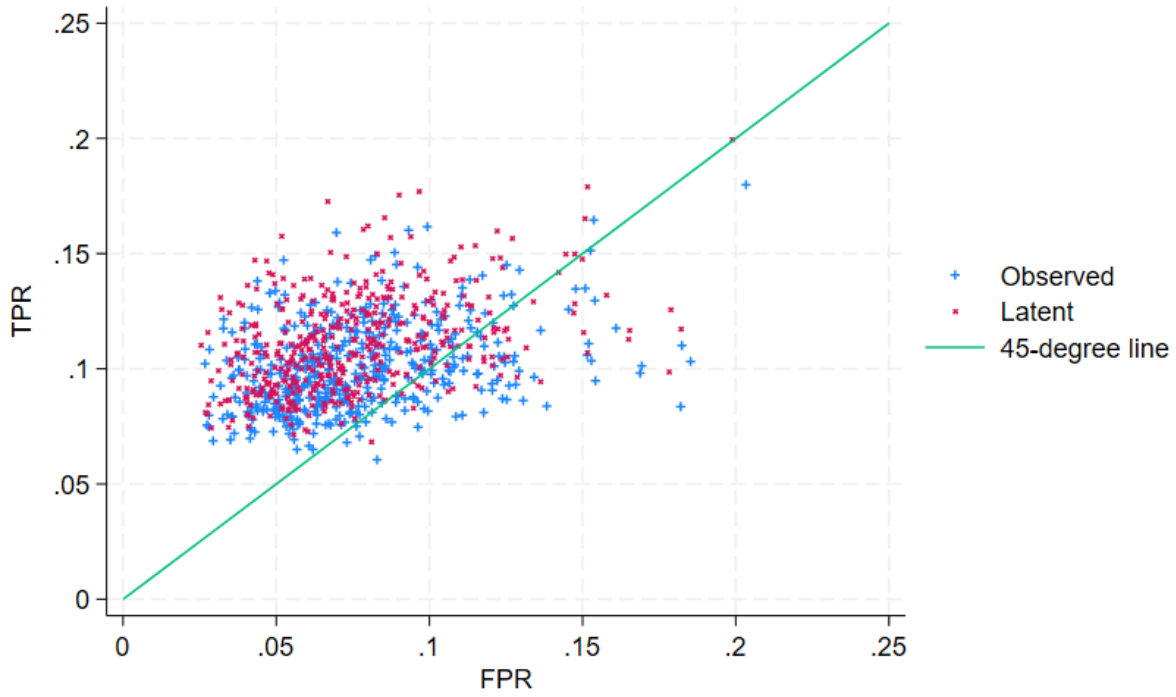
Wager, S., Athey, S., 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association 113, 1228–1242. doi:10.1080/01621459.2017.1319839, arXiv:1510.04342.

Whinney, A., 2015. A descriptive analysis of Multi-Agency Risk Assessment Conferences (MARACs) for reducing the future harm of domestic abuse in Suffolk. Ph.D. thesis.

World Health Organization, 2021. Violence against women Prevalence Estimates, 2018. Global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women. Technical Report. World Health Organization. Geneva.

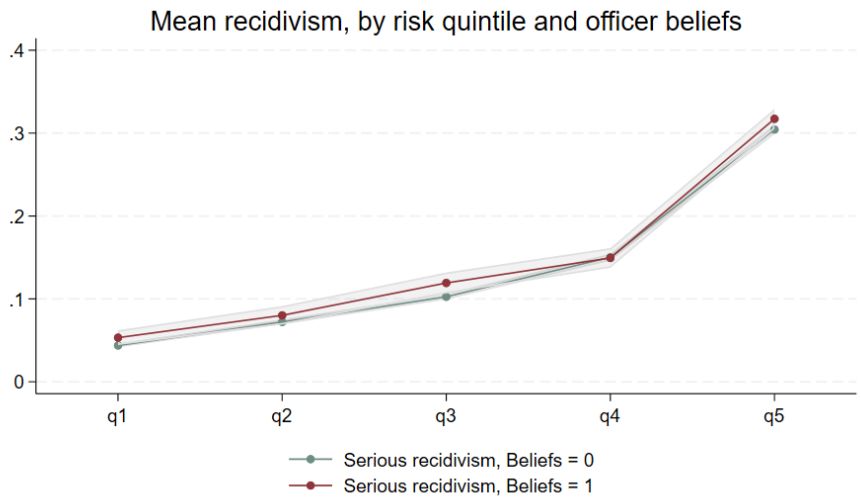
10. Figures and tables

Figure 1: Officer-level ROC plot, based on observed and latent outcomes



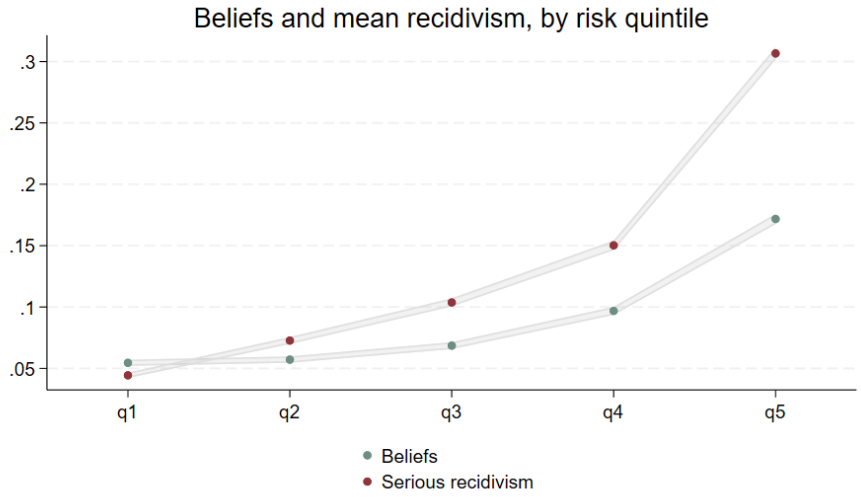
Note: Points plot (FPR, TPR) coordinates by officer for officers with at least 100 DA cases. Points have been shrunk towards the mean by empirical Bayes methods.

Figure 2: Mean recidivism, by risk quintile and officer beliefs



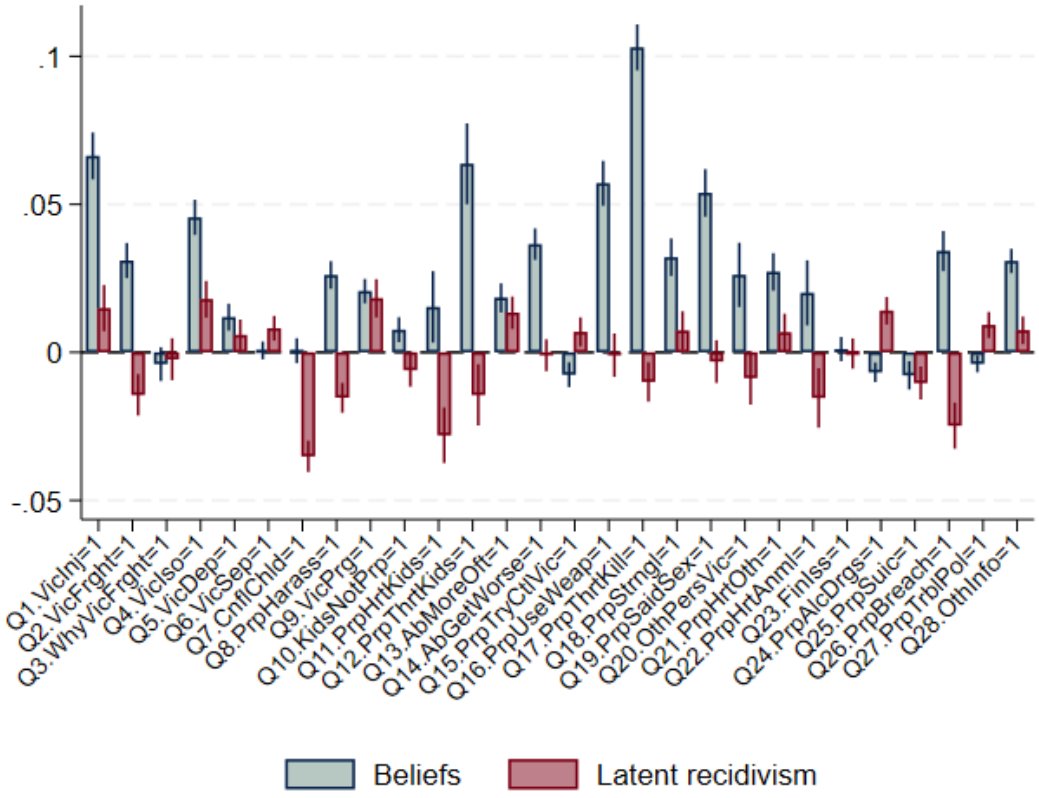
Note: This figure plots average latent recidivism by quintile of predicted risk from a regression of recidivism on case characteristics. The two lines separately plot cases where officers did and did not predict a high risk of recidivism.

Figure 3: Beliefs and mean recidivism, by risk quintile



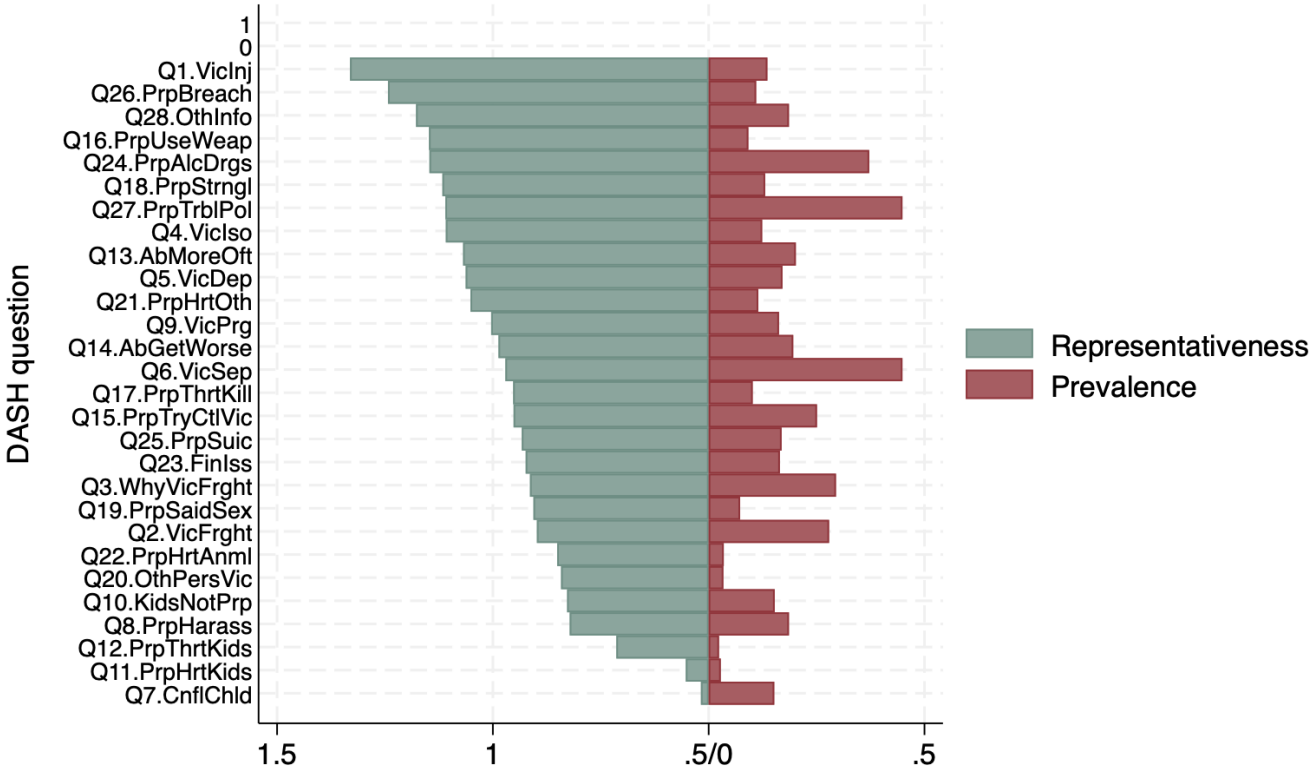
Note: This figure plots average latent recidivism and average rates of predicted risk by quintile of predicted risk from a regression of recidivism on case characteristics.

Figure 4: Coefficients from regressions of beliefs and latent recidivism on DASH questions



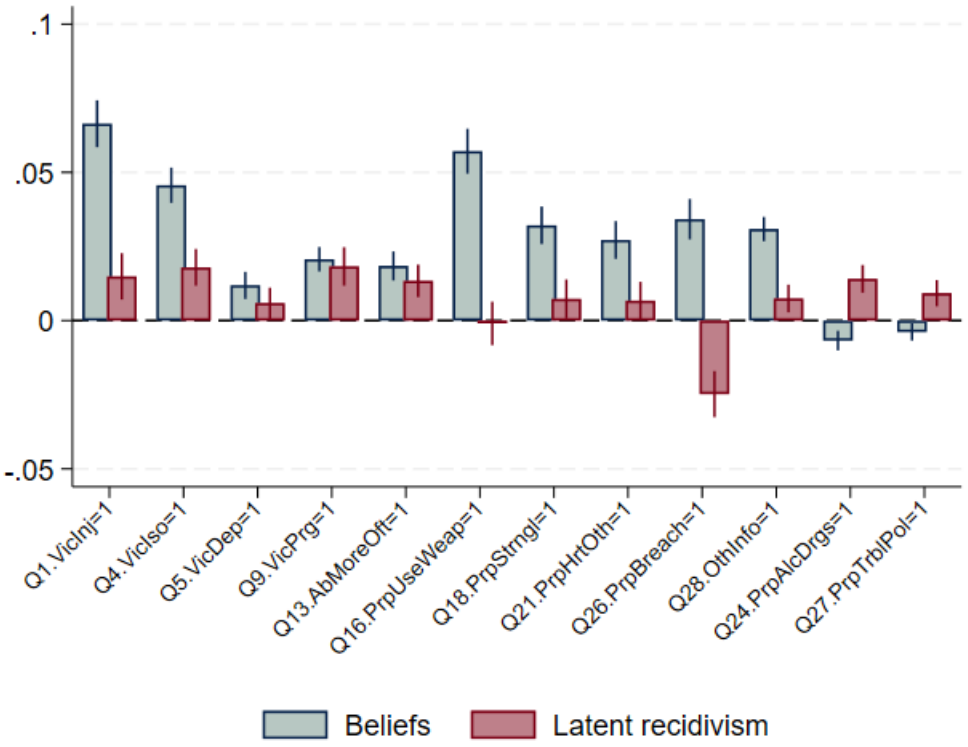
Notes: Coefficients from regressions of beliefs and latent recidivism on DASH features and other controls, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure 5: Representativeness and prevalence of DASH features



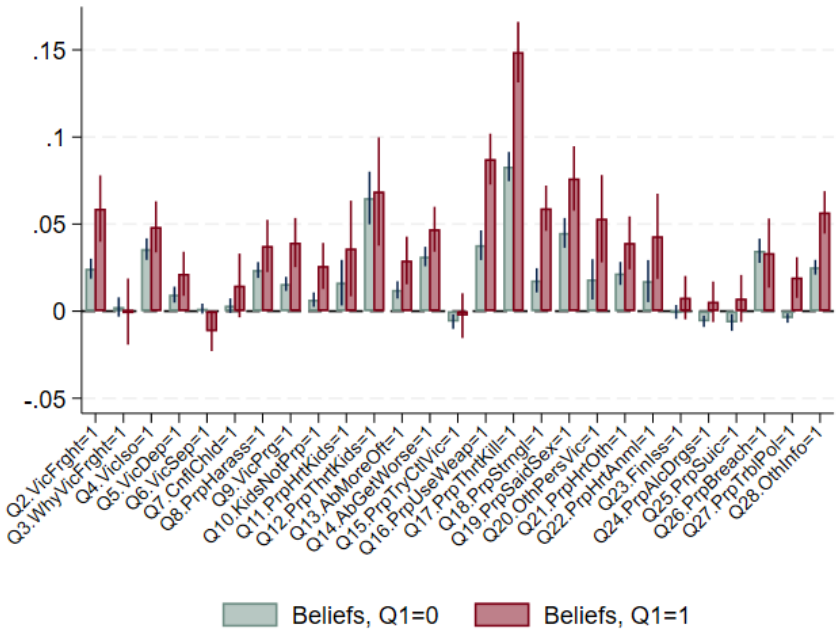
Note: Representativeness increases to the left, starting at .5. Prevalence increases to the right, starting at 0.

Figure 6: Coefficients from regressions of beliefs and latent recidivism on representative DASH features



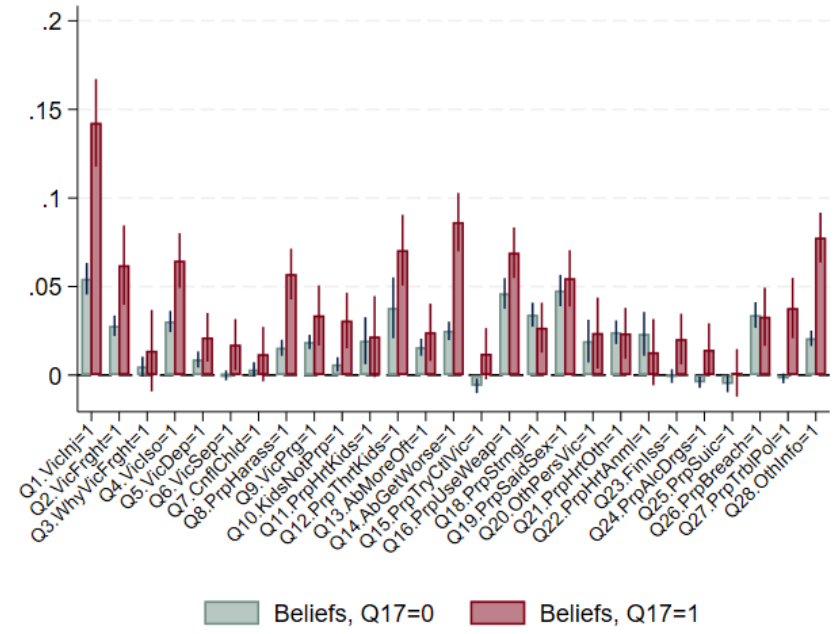
Notes: Coefficients from regressions of beliefs and latent recidivism on DASH features and other controls, showing only coefficients from DASH features that are representative of serious recidivism. The last two features, reflecting perpetrator alcohol and drug problems (Question 24) and prior trouble with the police (Question 27), are relatively common; the other features are all relatively rare. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure 7: DASH coefficients from beliefs regressions, by victim injury at current incident (Question 1)



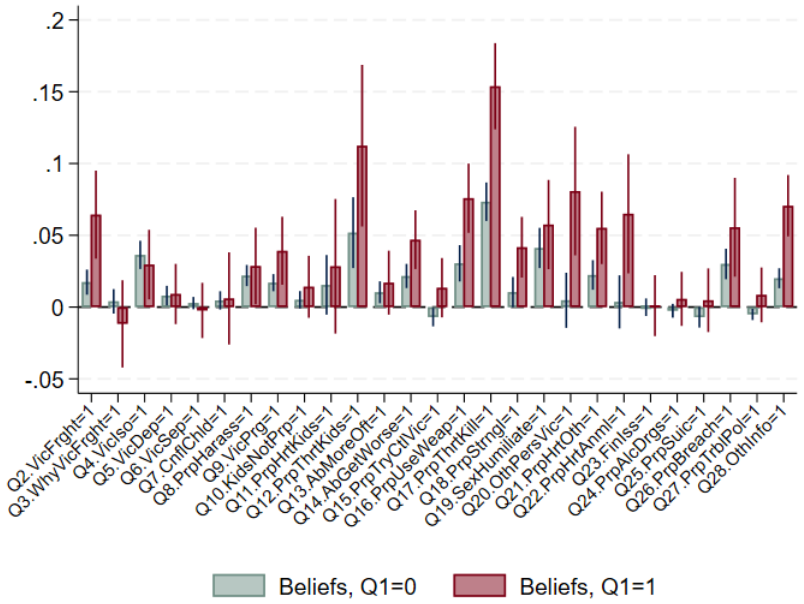
Notes: Coefficients from beliefs regressions on DASH features and other controls, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure 8: DASH coefficients from beliefs regressions, by death threats (Question 17)



Notes: Coefficients from beliefs regressions on DASH features and other controls, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure 9: Coefficients from regressions of beliefs on DASH features, by victim injury at current incident, higher-skill officers



Notes: Coefficients from beliefs regressions on DASH features and other controls, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure 10: Coefficients from regressions of beliefs on DASH features, by victim injury at current incident, low-skill officers

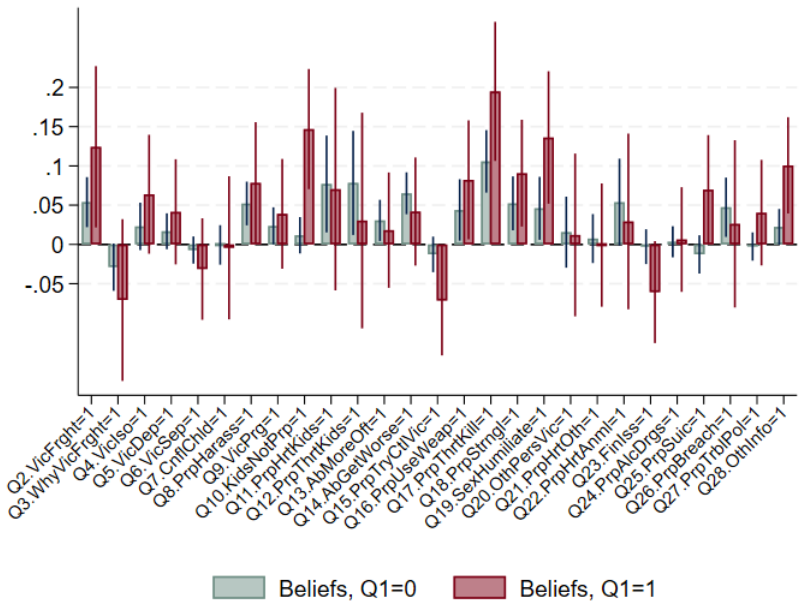


Table 1: Text of questions from DASH questionnaire

Question number	Abbreviation	Question text
1	VicInj	Has the current event resulted in injury?
2	VicFrght	Are you very frightened?
3	WhyVicFrght	What are you afraid of? Is it further injury or violence?
4	VicIso	Do you feel isolated from family/ friends i.e. does (name of abuser(s).....) try to stop you from seeing friends/family/Dr or others?
5	VicDep	Are you feeling depressed or having suicidal thoughts?
6	VicSep	Have you separated or tried to separate from (name of abuser(s)....) within the past year?
7	CnflChld	Is there conflict over child contact? (please state what)
8	PrpHarass	Does (.....) constantly text, call, contact, follow, stalk or harass you?
9	VicPrg	Are you currently pregnant or have you recently had a baby in the past 18 months?
10	KidsNotPrp	Are there any children, step-children that aren't (.....) in the household? Or are there other dependants in the household (i.e. older relative)?
11	PrpHrtKids	Has (.....) ever hurt the children/dependants?
12	PrpThrtKids	Has (.....) ever threatened to hurt or kill the children/dependants?
13	AbMoreOft	Is the abuse happening more often?
14	AbGetWorse	Is the abuse getting worse?
15	PrpTryCtlVic	Does (.....) try to control everything you do and/or are they excessively jealous? (In terms of relationships, who you see, being 'policed at home', telling you what to wear for example. Consider honour based violence and stalking and specify the behaviour)
16	PrpUseWeap	Has (.....) ever used weapons or objects to hurt you?
17	PrpThrtKill	Has (.....) ever threatened to kill you or someone else and you believed them?
18	PrpStrngl	Has (.....) ever attempted to strangle/choke/suffocate/drown you?
19	PrpSaidSex	Does (....) do or say things of a sexual nature that makes you feel bad or that physically hurt you or someone else?
20	OthPersVic	Is there any other person that has threatened you or that you are afraid of?
21	PrpHrtOth	Do you know if (.....) has hurt anyone else ?
22	PrpHrtAnml	Has (.....) ever mistreated an animal or the family pet?
23	FinIss	Are there any financial issues? For example, are you dependent on (.....) for money/have they recently lost their job/other financial issues?
24	PrpAlcDrgs	Has (.....) had problems in the past year with drugs (prescription or other), alcohol or mental health leading to problems in leading a normal life?
25	PrpSuic	Has (.....) ever threatened or attempted suicide?
26	PrpBreach	Has (.....) ever breached bail/an injunction and/or any agreement for when they can see you and/or the children?
27	PrpTrblPol	Do you know if (.....) has ever been in trouble with the police or has a criminal history?
28	OthInfo	Other relevant information (from victim or officer) which may alter risk levels

Table 2: Confusion matrix for observed and latent serious recidivism by high-risk status

	High risk		
A. Observed recidivism	No	Yes	Total
No	122,386	11,545	133,931
	0.914	0.086	1
Yes	17,727	2,276	20,003
	0.886	0.114	1
Total	140,113	13,821	153,934
AUC	0.514		

	High risk		
B. Latent recidivism	No	Yes	Total
No	121,854	11,216	133,070
	0.916	0.084	1
Yes	18,259	2,605	20,864
	0.875	0.125	1
Total	140,113	13,821	153,934
AUC	0.520		

Note: Within each cell, the first entry is the number of cases and the second is the row percentage. The row percentages in the yes column give the false positive (first row) and true positive (second row) rates.

Table 3: Cross-tabulation of latent recidivism and high-risk status by victim injury (Question 1)

	Victim injury		
A. Lat.recid.	No	Yes	Total
No	99,915	17,285	117,200
	0.880	0.829	.904
Yes	13,612	3,576	17,188
	0.120	0.171	.128
Total	113,527	20,861	134,388
	1	1	1

	Victim injury		
B. High risk	No	Yes	Total
No	106,159	15,328	121,487
	0.935	0.735	.904
Yes	7,368	5,533	17,188
	0.065	0.265	.096
Total	113,527	20,861	134,388
	1	1	1

Note: Within each cell, the first entry is the number of cases and the second is the column percentage. Lat. recid. stands for latent recidivism.

Table 4: Cross-tabulation of latent recidivism and high-risk status by death threats (Question 17)

	Death threats		
A. Lat.recid.	No	Yes	Total
No	94,244	13,645	107,889
	0.881	0.870	.879
Yes	12,753	2,042	14,795
	0.119	0.130	.121
Total	106,997	15,687	122,684
	1	1	1

	Death threats		
B. High risk	No	Yes	Total
No	100,581	10,077	110,658
	0.940	0.643	0.902
Yes	6,416	5,610	12,026
	0.060	0.358	0.098
Total	106,997	15,687	122,684
	1	1	1

Note: Within each cell, the first entry is the number of cases and the second is the column percentage. Lat. recid. stands for latent recidivism.

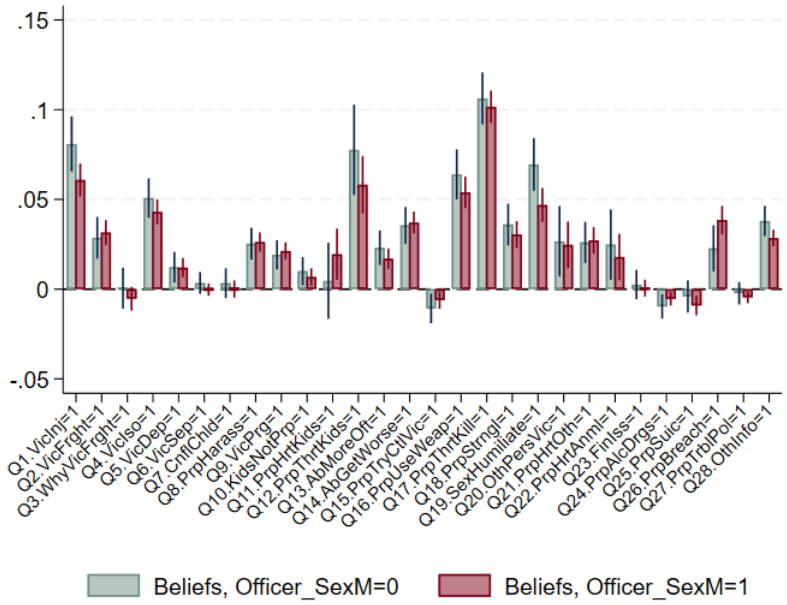
Table 5: Cross-tabulation of latent recidivism and high-risk status by recent high-risk classification

	Recent high-risk		
A. Lat.recid.	No	Yes	Total
No	122,626	10,444	133,070
	0.878	0.734	.864
Yes	17,081	3,782	20,864
	0.122	0.266	.136
Total	139,707	14,227	153,934
	1	1	1

	Recent high-risk		
B. High risk	No	Yes	Total
No	130,925	9,188	140,113
	0.937	0.646	0.910
Yes	8,782	5,039	13,821
	0.063	0.354	0.090
Total	139,707	14,227	153,934
	1	1	1

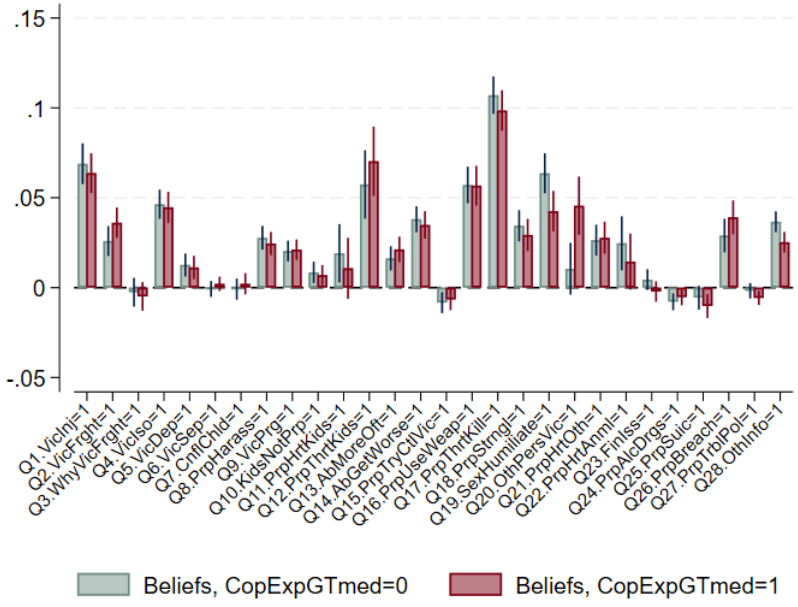
Note: Within each cell, the first entry is the number of cases and the second is the column percentage. Lat. recid. stands for latent recidivism.

Figure A1: DASH coefficients from beliefs regressions, by officer sex



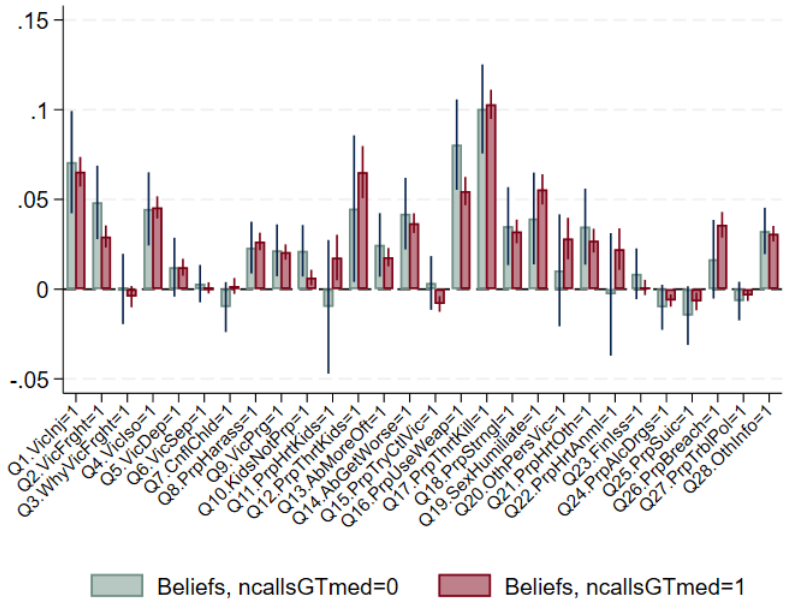
Notes: Coefficients from regressions of beliefs and latent recidivism on DASH features and other controls, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure A2: DASH coefficients from beliefs regressions, by officer experience



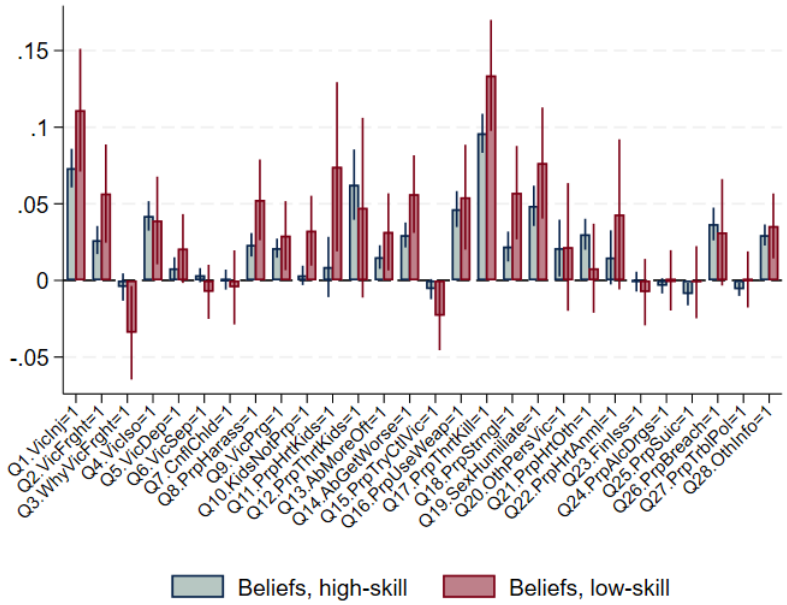
Notes: Coefficients from regressions of beliefs and latent recidivism on DASH features and other controls, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure A3: DASH coefficients from beliefs regressions, by number of officer DA cases



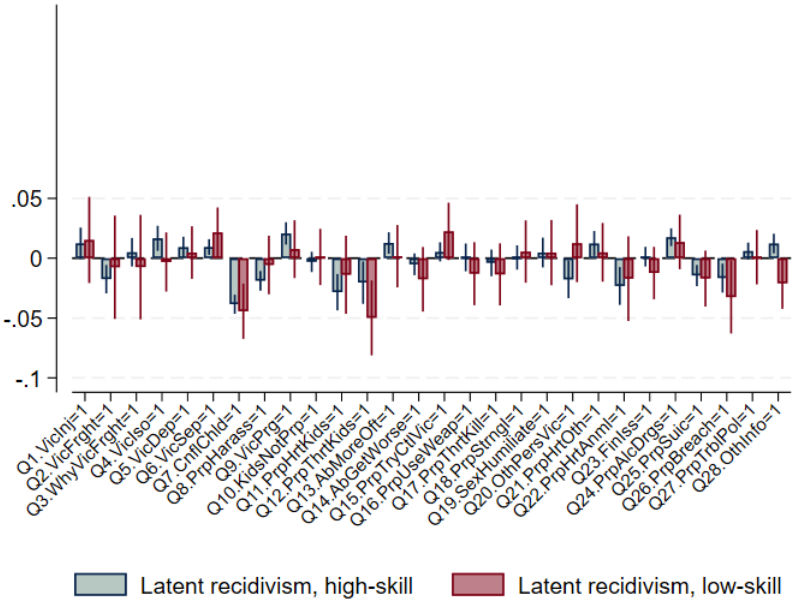
Notes: Coefficients from regressions of beliefs and latent recidivism on DASH features and other controls, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure A4: DASH coefficients from beliefs regressions, by officer skill, officers with at least 100 incidents



Notes: Coefficients from regressions of beliefs on DASH features and other controls, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure A5: DASH coefficients from latent outcome regressions, by officer skill, officers with at least 100 incidents



Notes: Coefficients from regressions of latent recidivism on DASH features and other controls, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Table A1: Coefficients on DASH features from beliefs and outcome regressions, various specs.

	(1) Beliefs	(2) Obs. recid.	(3) Lat. recid.	(4) Beliefs	(5) Lat. recid.	(6) Beliefs	(7) Lat. recid.
Q1.VicInj=1	0.066	0.011	0.015	0.057	0.013	0.065	0.020
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Q2.VicFrght=1	0.031	-0.015	-0.014	0.026	-0.015	0.030	-0.010
	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)
Q3.WhyVicFrght=1	-0.004	-0.003	-0.002	0.000	-0.002	-0.004	-0.002
	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)
Q4.VicIso=1	0.046	0.017	0.018	0.040	0.017	0.038	0.017
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
Q5.VicDep=1	0.012	0.007	0.006	0.013	0.005	0.008	0.005
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Q6.VicSep=1	0.001	0.008	0.008	0.001	0.009	0.000	0.009
	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
Q7.CnflChld=1	0.001	-0.035	-0.035	0.002	-0.034	0.000	-0.035
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)
Q8.PrpHarass=1	0.026	-0.016	-0.015	0.023	-0.016	0.021	-0.015
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Q9.VicPrg=1	0.021	0.018	0.018	0.019	0.018	0.020	0.018
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)
Q10.KidsNotPrp=1	0.008	-0.006	-0.006	0.009	-0.005	0.009	-0.003
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)
Q11.PrpHrtKids=1	0.015	-0.028	-0.028	0.018	-0.028	0.020	-0.026
	(0.006)	(0.005)	(0.005)	(0.006)	(0.005)	(0.007)	(0.005)
Q12.PrpThrtKids=1	0.064	-0.016	-0.014	0.055	-0.016	0.057	-0.023
	(0.007)	(0.005)	(0.005)	(0.007)	(0.005)	(0.008)	(0.006)
Q13.AbMoreOft=1	0.018	0.012	0.013	0.017	0.013	0.016	0.012
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Q14.AbGetWorse=1	0.037	-0.004	-0.001	0.035	-0.001	0.031	-0.003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Q15.PrpTryCtlVic=1	-0.008	0.007	0.007	-0.001	0.008	-0.006	0.008
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)
Q16.PrpUseWeap=1	0.057	-0.002	-0.001	0.050	-0.001	0.051	-0.000
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Q17.PrpThrtKill=1	0.103	-0.013	-0.010	0.092	-0.010	0.090	-0.008
	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.005)	(0.004)
Q18.PrpStrngl=1	0.032	0.007	0.007	0.032	0.007	0.031	0.009
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Q19.PrpSaidSex=1	0.054	-0.003	-0.003	0.050	-0.004	0.054	-0.005
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)
Q20.OthPersVic=1	0.026	-0.007	-0.009	0.026	-0.007	0.015	-0.010
	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)
Q21.PrpHrtOth=1	0.027	0.006	0.007	0.028	0.007	0.023	0.004
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Q22.PrpHrtAnml=1	0.020	-0.015	-0.015	0.024	-0.016	0.020	-0.014
	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)
Q23.FinIss=1	0.001	-0.000	-0.000	0.003	-0.000	0.002	-0.003
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)
Q24.PrpAlcDrgs=1	-0.007	0.014	0.014	-0.005	0.014	-0.004	0.015
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
Q25.PrpSuic=1	-0.008	-0.010	-0.010	-0.003	-0.010	-0.006	-0.008
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Q26.PrpBreach=1	0.034	-0.030	-0.025	0.035	-0.024	0.032	-0.022
	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
Q27.PrpTrblPol=1	-0.004	0.009	0.009	-0.004	0.011	-0.004	0.010
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
Q28.OthInfo=1	0.031	0.007	0.007	0.034	0.011		
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)		
Observations	153934	153934	153934	153934	153934	103502	103502

Notes: In addition to the variables shown, all specs. include controls from administrative records. Columns (4) and (5) also include officer fixed effects. Columns (6) and (7) are estimated from cases where Question 28 = 0. Standard errors, clustered by dyad, in parentheses. Obs. recid. = Observed recidivism; Lat. recid. = Latent recidivism.

Table A2: Coefficients on DASH questions from bivariate beliefs and outcome regressions

	(1) Beliefs	(2) Obs. recid.	(3) Lat. recid.
Q1.VicInj=1	0.200	0.037	0.052
	(0.003)	(0.003)	(0.003)
Q2.VicFrght=1	0.194	-0.010	0.000
	(0.002)	(0.002)	(0.002)
Q3.WhyVicFrght=1	0.192	-0.006	0.005
	(0.002)	(0.002)	(0.002)
Q4.VicIso=1	0.193	0.022	0.030
	(0.004)	(0.003)	(0.003)
Q5.VicDep=1	0.138	0.021	0.026
	(0.003)	(0.003)	(0.003)
Q6.VicSep=1	0.078	0.017	0.021
	(0.002)	(0.002)	(0.002)
Q7.CnflChld=1	0.029	-0.057	-0.057
	(0.003)	(0.003)	(0.003)
Q8.PrpHarass=1	0.147	-0.016	-0.009
	(0.003)	(0.003)	(0.003)
Q9.VicPrg=1	0.044	0.014	0.015
	(0.003)	(0.004)	(0.004)
Q10.KidsNotPrp=1	0.045	-0.013	-0.010
	(0.003)	(0.003)	(0.003)
Q11.PrpHrtKids=1	0.172	-0.049	-0.043
	(0.008)	(0.005)	(0.005)
Q12.PrpThrtKids=1	0.262	-0.031	-0.021
	(0.008)	(0.005)	(0.005)
Q13.AbMoreOft=1	0.166	0.020	0.029
	(0.003)	(0.002)	(0.002)
Q14.AbGetWorse=1	0.192	0.005	0.005
	(0.003)	(0.002)	(0.002)
Q15.PrpTryCtlVic=1	0.155	0.005	0.013
	(0.002)	(0.002)	(0.002)
Q16.PrpUseWeap=1	0.255	0.024	0.035
	(0.005)	(0.004)	(0.004)
Q17.PrpThrtKill=1	0.298	-0.002	0.011
	(0.004)	(0.003)	(0.003)
Q18.PrpStrngl=1	0.230	0.023	0.034
	(0.004)	(0.003)	(0.003)
Q19.PrpSaidSex=1	0.224	-0.003	0.005
	(0.005)	(0.004)	(0.004)
Q20.OthPersVic=1	0.200	-0.009	-0.004
	(0.007)	(0.005)	(0.005)
Q21.PrpHrtOth=1	0.196	0.015	0.024
	(0.004)	(0.003)	(0.003)
Q22.PrpHrtAnml=1	0.205	-0.011	-0.003
	(0.007)	(0.005)	(0.005)
Q23.FinIss=1	0.067	0.005	0.007
	(0.003)	(0.003)	(0.003)
Q24.PrpAlcDrgs=1	0.106	0.047	0.054
	(0.002)	(0.003)	(0.003)
Q25.PrpSuic=1	0.121	0.003	0.009
	(0.003)	(0.003)	(0.003)
Q26.PrpBreach=1	0.210	0.033	0.047
	(0.005)	(0.005)	(0.005)
Q27.PrpTrblPol=1	0.116	0.047	0.055
	(0.002)	(0.002)	(0.002)
Q28.OthInfo=1	0.107	0.032	0.036
	(0.003)	(0.003)	(0.003)

Notes: Each coefficient is from a separate bivariate regression of the dependent variable in indicated column on the DASH feature in the indicated row. Sample size is 153,934. Standard errors, clustered by dyad, in parentheses. Obs. recid. = Observed recidivism; Lat. recid. = Latent recidivism.

Table A3: Coefficients on DASH questions from beliefs regressions, by coercive control

	(1) CC=0	(2) CC=1
Q1.VicInj=1	0.032	0.096
	(0.005)	(0.006)
Q2.VicFrght=1	0.022	0.039
	(0.004)	(0.004)
Q3.WhyVicFrght=1	0.008	-0.006
	(0.004)	(0.004)
Q5.VicDep=1	0.000	0.018
	(0.003)	(0.003)
Q6.VicSep=1	-0.000	0.012
	(0.002)	(0.003)
Q7.CnflChld=1	0.002	0.007
	(0.002)	(0.003)
Q9.VicPrg=1	0.008	0.032
	(0.002)	(0.004)
Q10.KidsNotPrp=1	0.002	0.016
	(0.002)	(0.003)
Q11.PrpHrtKids=1	0.003	0.023
	(0.009)	(0.007)
Q12.PrpThrtKids=1	0.027	0.066
	(0.013)	(0.008)
Q13.AbMoreOft=1	0.006	0.024
	(0.004)	(0.003)
Q14.AbGetWorse=1	0.023	0.039
	(0.004)	(0.003)
Q16.PrpUseWeap=1	0.042	0.058
	(0.006)	(0.005)
Q17.PrpThrtKill=1	0.051	0.101
	(0.008)	(0.004)
Q18.PrpStrngl=1	0.013	0.032
	(0.005)	(0.004)
Q19.PrpSaidSex=1	0.045	0.048
	(0.008)	(0.005)
Q20.OthPersVic=1	0.016	0.023
	(0.010)	(0.006)
Q21.PrpHrtOth=1	0.012	0.026
	(0.005)	(0.004)
Q22.PrpHrtAnml=1	-0.009	0.020
	(0.010)	(0.006)
Q23.FinIss=1	-0.000	0.007
	(0.002)	(0.003)
Q24.PrpAlcDrgs=1	-0.005	0.002
	(0.002)	(0.003)
Q26.PrpBreach=1	0.024	0.028
	(0.005)	(0.004)
Q27.PrpTrblPol=1	-0.003	0.005
	(0.002)	(0.003)
Q28.OthInfo=1	0.014	0.048
	(0.002)	(0.003)
Observations	92670	61264

Notes: In addition to the variables shown, all specs. include controls from administrative records. Standard errors, clustered by dyad, in parentheses. CC stands for coercive control.

Table A4: Coefficients on DASH questions from beliefs regressions, various specs., by coercive control

	(1) Beliefs, CC=0	(2) Beliefs, CC=1	(3) Beliefs, CC=0	(4) Beliefs, CC=1
Q1.VicInj=1	0.029	0.079	0.044	0.085
	(0.005)	(0.006)	(0.006)	(0.007)
Q2.VicFrght=1	0.018	0.030	0.020	0.038
	(0.004)	(0.004)	(0.005)	(0.004)
Q3.WhyVicFrght=1	0.009	0.000	0.009	-0.009
	(0.004)	(0.004)	(0.004)	(0.004)
Q5.VicDep=1	0.004	0.016	-0.003	0.015
	(0.003)	(0.003)	(0.003)	(0.003)
Q6.VicSep=1	0.000	0.011	0.000	0.009
	(0.001)	(0.003)	(0.002)	(0.003)
Q7.CnflChld=1	0.001	0.009	0.000	0.005
	(0.002)	(0.003)	(0.002)	(0.004)
Q9.VicPrg=1	0.009	0.028	0.007	0.033
	(0.002)	(0.003)	(0.002)	(0.004)
Q10.KidsNotPrp=1	0.002	0.016	0.003	0.016
	(0.002)	(0.003)	(0.002)	(0.004)
Q11.PrpHrtKids=1	0.004	0.028	0.004	0.028
	(0.008)	(0.007)	(0.010)	(0.009)
Q12.PrpThrtKids=1	0.023	0.056	0.036	0.057
	(0.012)	(0.007)	(0.014)	(0.009)
Q13.AbMoreOft=1	0.007	0.023	0.006	0.021
	(0.003)	(0.003)	(0.004)	(0.004)
Q14.AbGetWorse=1	0.021	0.035	0.024	0.031
	(0.004)	(0.003)	(0.004)	(0.004)
Q16.PrpUseWeap=1	0.036	0.048	0.034	0.053
	(0.006)	(0.004)	(0.007)	(0.005)
Q17.PrpThrtKill=1	0.044	0.089	0.046	0.088
	(0.007)	(0.004)	(0.008)	(0.005)
Q18.PrpStrngl=1	0.015	0.031	0.016	0.030
	(0.005)	(0.004)	(0.005)	(0.004)
Q19.PrpSaidSex=1	0.038	0.047	0.043	0.049
	(0.008)	(0.004)	(0.009)	(0.005)
Q20.OthPersVic=1	0.014	0.025	0.013	0.012
	(0.009)	(0.006)	(0.011)	(0.007)
Q21.PrpHrtOth=1	0.014	0.028	0.010	0.022
	(0.005)	(0.004)	(0.005)	(0.004)
Q22.PrpHrtAnml=1	-0.003	0.024	0.000	0.019
	(0.010)	(0.006)	(0.011)	(0.007)
Q23.FinIss=1	0.001	0.009	0.000	0.008
	(0.002)	(0.003)	(0.002)	(0.003)
Q24.PrpAlcDrgs=1	-0.004	0.003	-0.002	0.003
	(0.002)	(0.003)	(0.002)	(0.003)
Q26.PrpBreach=1	0.025	0.032	0.025	0.028
	(0.005)	(0.004)	(0.005)	(0.005)
Q27.PrpTrblPol=1	-0.003	0.005	-0.001	0.001
	(0.001)	(0.003)	(0.002)	(0.003)
Q28.OthInfo=1	0.019	0.048		
	(0.002)	(0.003)		
Observations	92670	61264	58375	45127

Notes: In addition to the variables shown, all specs. include controls from administrative records. Columns (1) and (2) also include officer fixed effects. Columns (3) and (4) are estimated from cases where Question28 = 0. Standard errors, clustered by dyad, in parentheses. CC stands for coercive control.