Centre for Economic Performance

Discussion Paper

ISSN 2042-2695

No. 2212
August 2026

Explaining the predictive performance of police in cases of domestic abuse

Jeffrey Grogger
Andrew Jordan
Tom Kirchmaier

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

UKRI Economic and Social Research Council

Abstract
Police in England and Wales are asked to predict the likelihood of serious recidivism in domestic abuse cases, with little support beyond a flawed questionnaire. 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. Even after adjusting for censoring, 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 we find evidence consistent with representativeness bias, overreaction, and categorization and selective attention. We find that higher-skill officers use information not captured by the questionnaire to improve their predictions, whereas lower-skill officers use such information in ways that reduce accuracy.

Keywords: domestic abuse; risk assessment; screening problems; censoring; cognitive biases

This paper was produced as part of the Crime Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

We thank Pedro Bordalo, Alex Imas, and Andrei Shleifer for helpful discussions, participants at a number of seminars and conferences for comments, 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.

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, Centre for Economic Performance at the London School of Economics.

Published by
Centre for Economic Performance
London School of Economic and Political Science
Houghton Street
London WC2A 2AE

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 risk of serious harm 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, intended to reduce their risk 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 undertaken 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 a small share of 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 assigning them based on mistaken predictions.

Finally, we ask why officers make mistakes. One problem is the potentially misleading nature of some of the questionnaire items. Although a face-value reading of the ques-

tionnaire suggests that all the questions should positively predict risk, in fact nearly half of them are negatively predictive. Beyond that, 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 on outcomes, they can hardly be acting as Bayesian decision makers.

We consider alternative cognitive mechanisms through straightforward regressions that exploit several unique features of our data. First, we observe the officers' beliefs and the features from the assessment questionnaire that 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 that are observable to them. Second, under conditions which we discuss below, our imputation procedure allows us to consistently estimate an analogous regression of latent recidivism on the same variables. These regressions point to representativeness bias, overreaction, and categorization and selective attention as potential sources of prediction error. Our analysis also reveals heterogeneous processing of information. Higher-skill officers use information not captured by the questionnaire to improve predictions, whereas the lowest-skill officers use it in a way that reduces accuracy.

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.

The methodological issue stems from the censoring problem that may arise 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.

The second literature we contribute to considers similar censoring issues 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's symptoms truly stemmed from 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. The first difference is that, in the cases cited above, the censoring problem has been solved by virtue of random assignment of decision makers to cases. With random

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

assignment, cases assigned a maximally lenient decision maker may be used to estimate the latent outcome for all cases. Alternatively, random assignment can be used to narrow the worst-case bounds on average latent outcomes (Jordan, 2025).

This approach is not available to us, since our cases are not randomly assigned to police officers.[2] 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 decide to charge the perpetrator with a crime.

The second difference between our setting and others is that 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, any treatment effects are 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 from those based on imputed outcomes.

The third literature we contribute to is the vast array of empirical research on cognitive biases in decision making, especially regarding belief formation. Tversky and Kahneman (1974) posited that decision makers form beliefs not from the true distribution of features, but from a distribution distorted toward features representative of the outcome they are predicting. 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

[2]The officers in our data who classify the fewest cases as high-risk, and hence are the least censored, are clearly affected by this non-random assignment. When we calculate bounds on uncensored recidivism rates for individual lenient officers, the bounds are frequently disjoint and lie above the observed force-wide recidivism rate.

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 underreaction in some settings and overreaction in others. 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 theoretical models (Augenblick et al., 2025; Ba et al., 2024; Fan et al., 2022).

Beyond overreaction, we find that the DASH features interact strongly in influencing officers' beliefs. To rationalize these interactions, we turn to the model of categorization and selective attention from Bordalo et al. (2025). They 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. An ancillary feature triggers a mental category, which causes instability in beliefs and turns attention toward the subset of features that compose the category.

Several potential trigger features are suggested by a recent survey, which asked police officers which predictive features they found essential for assessing risk (Robinson et al., 2018). Several analyses point to a similar structure: the six essential features appear to reflect three partially overlapping binary categories into which officers classify incidents. Each of these features is associated with instability of beliefs and changes in attention.

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 a question asking whether the officer used additional information in classifying risk. 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.

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 draw on 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, from 2014 to mid-2019, all calls for service were recorded as an incident in GMP's command-and-control database.[3] 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 hetero-

[3]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.

sexual intimate partners, we restrict attention to those calls.[4]

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.[5] 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 (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, 2023).

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. Focusing exclusively on such a rare event was likely to reduce power, 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.

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

[5]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.

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. This shows that officers' predictions fall short of the serious recidivism rate. Even if officers were perfect decision makers and all the cases they designated as high-risk resulted in serious recidivism, nearly 30% of serious recidivism cases would get no protective services.

The observed false positive rate (FPR) is defined as the number of cases that 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.[6] 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 to 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. Because each officer-level point averages roughly 141 incidents, we use empirical Bayes methods to shrink the points toward the full-sample means, reducing noise (Gu and 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

[6]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).

appear to predict poorly. About 10 percent lie below the 45-degree line. These officers would produce better predictions by making random guesses.[7]

To benchmark officer performance, we estimated a simple linear regression predicting serious recidivism from the DASH features. The receiver operating characteristic (ROC) curve from this regression is depicted by the curve labeled "TPR_reg" It shows the best tradeoffs between FPR and TPR that can be obtained by predictions from a simple statistical technology using the same data as that available to the officers. Each point on the curve corresponds to a different threshold for predicted values from a linear regression of serious recidivism on the DASH items. Cases that exceed this threshold are classified as high risk. Predictions from the regression dominate the predictions of all officers lying below the ROC curve. For each such officer, the regression achieves a higher TPR, holding FPR fixed; a lower FPR, holding TPR fixed; or both. Put differently, only the 58 officers above the ROC curve exhibit predictive skill superior to that of the linear regression.

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).[8]

We use the CATEs to impute latent outcomes for cases designated as high-risk, cases where charges were filed, and for those both classified as high risk and charged. However,

[7]Similarly, Chan et al. (2022) report that roughly 2.5 percent of the decision makers in their sample lie below the 45-degree line.

[8]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 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.

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 3-way interactions between the high-risk classification and the decision to refer charges, for the 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 the case 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

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

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

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

We use imputed outcomes to estimate true and false positive rates as well as the regressions discussed below.

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 0.135 (= 20817/153934), 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 0.190 (=2629/13821), it still indicates that over four out of five high-risk designations are incorrect. Similarly, the AUC based on the imputed data, at 0.521, 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.[9]

The fact that our imputed outcomes are so similar to the observed outcomes reiterates a key finding from Black et al. (2023): protective measures triggered by a high-risk designation have little impact on serious recidivism. This similarity shows that the primary bottleneck in victim safety is prediction quality, not a censoring artifact from effective interventions. In our setting, correcting for censoring is logically necessary, but it is practically inconsequential. In settings with more potent interventions, such as bail decisions, the causal forest approach to imputation might have greater consequences for inference.

4.3. Conditional independence

We now discuss the conditional independence assumption the causal forest approach needs to consistently estimate latent recidivism rates and related 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. Because it involves unobserved counterfactuals, this is not a condition that can be proven. However, we believe several factors make it plausible in our setting.

First, the institutional setting implies that our observables should be highly predictive of officers' risk classifications: officers are instructed to administer the DASH and use the responses to form their judgments. Black et al. (2023) show that this is indeed true, and it can also be inferred from our results below. Beyond the DASH, many of

[9]This adjustment assumes that high-risk designations affect only whether recidivism actually occurs and not whether it is reported. When we look separately at estimated treatment effects for recidivism incidents reported by the victim and those reported by a third party, we find that the former tend to be negative and the latter positive, though the magnitudes of both effects are small. This suggests that the high-risk designation, and resulting protective services, may increase reporting via the service providers themselves. Insofar as this is true, it would imply that some recorded true negatives were actually false negatives. Recidivism occurred, but it was not detected for want of reporting by service providers. In this case, our estimates that ignore possible changes in reporting take a charitable view of officers' actual decision making performance.

our additional controls, constructed from criminal and domestic abuse histories of the individuals involved, are also highly predictive. Thus conditional independence implies that officer decisions must be effectively random after controlling for a highly predictive set of covariates. That is, at least some officer decisions must be influenced by information that does not actually predict recidivism.

The key question is thus, what else might predict officers' judgments? In other criminal justice settings, decision makers have been shown to act on information irrelevant to the case at hand. Dutch prosecutors were more likely to prosecute Muslim defendants in unrelated cases after Theo van Gogh was murdered by a Moroccan perpetrator (Bielen and Grajzl, 2021). Judges' decisions have been shown to be affected by the weather, upset losses in football games, and features of other cases on the judge's docket (Heyes and Saberian, 2019; Eren and Mocan, 2018; Chen et al., 2016). Jury decisions in Old Bailey exhibited positive autocorrelation across unrelated cases (Bindler and Hjalmarsson, 2019). Ludwig and Mullainathan (2024) show that features of the defendant's face are more important for predicting bail judges' detention decisions than whether the defendant was arrested for a violent vs. non-violent offense. It seems reasonable to think that the unpredictable part of police decisions could similarly be influenced by factors unrelated to a DA perpetrator's recidivism risk.

Finally, we estimate real and placebo treatment effects, in order to provide a specification test of the type advocated by Imbens (2015). We first use the estimated CATEs to estimate the treatment effect of the 3-way interacted interventions on recidivism. We then estimate placebo treatment effects by estimating the effects of those interventions on serious domestic abuse during the period 12 months prior to the interventions. Since an intervention logically cannot affect prior behavior, non-zero placebo treatment effects would indicate that the model was misspecified, casting doubt on the validity of the conditional independence assumption. Conversely, zero treatment effects would provide another piece of evidence suggesting its plausibility.

Estimated treatment effects are reported in Appendix Table A1. The first column presents estimates of the effects of the three interacted interventions, formed by averaging the CATEs over the treated observations. The estimates in the first column are very similar to those reported by Black et al. (2023). The estimates in the second column are also estimated on the real serious recidivism outcome, but on a sample that excludes each dyad's first incident. This is the sample from which we can estimate the placebo treatment effect, since it is based on a lagged dependent variable. Moreover, the criminal and domestic abuse histories used as covariates to compute the estimates in the second column were truncated to end at least 12 months before the intervention. The estimates in the second column are slightly larger but generally similar to those in the first. Finally,

the placebo treatment effects appear in column (3). They are based on the same sample and same specification as those used for column (2), but they involve the lagged dependent variable. All three of the estimates are small, and none is significant. This does not prove that CIA holds, but at the same time gives no reason to think that it is violated.

5. Mistakes or preferences?

Above we have framed officers' poor predictive performance as stemming from mistakes in risk estimation. An alternative explanation is that officers are working from accurate risk estimates but not assigning protective services in the optimal number of cases. In this section, we briefly discuss the conceptual differences between these two explanations before providing three tests that suggest that officers are indeed making mistakes that take them away from the rational baseline.

5.1. Conceptual Framework

Define an officer's estimate of the risk of violent recidivism in the absence of protective services as $\tilde{Y}_i(X_i)$. While forming this estimate, officers have access to the information in X_i, but they may not make full or correct use of it. Given this estimate of risk, the officer assigns protective services if the anticipated risk exceeds some threshold $\tilde{Y}_i(X_i) > t$.

In this framework, "mistakes" are deviations between the officer's prediction of risk and the true risk: $\tilde{Y}_i(X_i) \neq E[Y_i(0)|X_i]$. For example, if officers consistently believe that a pattern of drug and alcohol abuse does not signal a higher risk of violent recidivism, despite the fact that it does, they have made a mistake in risk estimation.

In contrast to mistakes, "preferences" are factors that shift the officer's threshold for assigning protective services, t. An officer who perceives a high cost of false positives will have a high t, whereas an officer who perceives a high cost of false negatives will have a low t. Given a full accounting of the costs of serious recidivism and of protective services, it would be possible to calculate a socially optimal threshold. However, officers may not adhere to it because risk assessments present them with uneven career risks. A false negative followed by a particularly bad outcome, such as murder of the victim, could permanently harm the officer's career. In contrast, excessive false positives only impose costs on the institutions providing protective services, not the officer personally. Alternatively, officers who perceive a binding capacity constraint on protective services even when one does not exist may set an inefficiently high threshold.

The key point of this section is that while the selection of the threshold for protective services requires navigating a tradeoff between the cost of those services and the potential costs of serious recidivism, estimation mistakes are purely inefficient. When officers make estimation mistakes, victims who are at an objectively lower risk of serious recidivism

receive protective services over those at a higher risk. Situations where officers are making mistakes therefore present ready opportunities for improvements in the efficiency of protective services.

5.2. 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 assign protective services in 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, though we speculated above that officers' personal preferences would lean the other way. Low rates of protective services could also be driven by a simple resource constraint on how many cases can be assigned those services.

However, we find that 9.1 percent of officers pair a high rate of false negatives with a high rate of false positives.[10] These officers fall below the 45-degree line in Figure 1 even after shrinkage. Recall that points on the 45-degree line can be achieved by officers who randomly assign protective services at the corresponding rate. The only necessary information is $\tilde{Y}(0)$ with no X_i, and this estimate need not even equal $E[Y(0)]$. That is, 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 in X_i to produce $\tilde{Y}_i(X_i) \neq E[Y_i(0)|X_i]$.

5.3. 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 random case assigned protective services in the original allocation, an officer below the 45-degree line 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 generalized 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 predictions by shifting protective services from the low-risk neighborhood to the high-risk neighborhood, even if the shifted cases are selected at random.

[10]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.

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 X_i, which in this case is the set of DASH features.[11] This procedure yields sets of covariate values $\mathbf{X^1} \ldots \mathbf{X^5}$ such that $E[Y_i(0)|X_i \in \mathbf{X^j}] < E[Y_i(0)|X_i \in \mathbf{X^{j'}}]$ for $j < j'$. Officers who knew which quintile each case fell into and nothing else would assign protective services to cases in the top quintile until those cases were exhausted, then move on to the 4th quintile, and so on.

Instead, the histogram in Figure 2 shows that officers assign protective services to cases in all 5 quintiles. Only 5,287 cases, or about 38 percent of those classified high risk (= 5,287/13,821) are among the 30,786 cases in highest quintile of the objective risk distribution. At the same time, each of the lower quintiles is given at least 12 percent of the cases designated as high risk. The red line in Figure 2 shows that the cases classified as high-risk in lower quintiles have lower recidivism rates than that of a randomly selected case from the fifth quintile, depicted by the horizontal line. Even in the fourth quintile, the recidivism rate of cases designated as high risk is 15.4 percentage points lower than that of the average case in the fifth quintile. Officers could improve average outcomes by taking protective services from cases in the bottom 4 quintiles and assigning them to unprotected cases in highest quintile. There are 30,786 cases in this highest quintile, so each of the 13,821 times officers assigned protective services, they could have assigned them to a case in this quintile. 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, leading to mistakes.

5.4. Simultaneous over- and under-prediction

Our third test for mistakes comes from Mullainathan and Obermeyer (2022), who show that irrational officers will under-predict recidivism in some situations but over-predict it in others. They, like us, model decision makers as classifying cases according to a threshold rule in risk. If this threshold is set too low, officers may generally over-predict recidivism, and they will likewise under-predict if the threshold is set too high. Seeing both errors in different circumstances implies different thresholds in different settings, which is not consistent with a rational evaluation of the costs and benefits involved.

To test this, we plot both recidivism and officer's beliefs about recidivism by the objective risk quintiles described above. See Figure 3. Officers over-predict by one percentage point at the lowest quintile and under-predict by increasing amounts at higher quintiles.

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

At the fifth quintile, they underpredict by 13.4 percentage points. Again, this indicates that officers are making mistakes.

An alternative to the "mistake" hypothesis is that officers act rationally but operate under misaligned loss functions. Consider the principal-agent issue discussed in Section 5.1, whereby career concerns may cause the officer to perceive a cost of false negatives that is greater than is socially optimal. If such preferences were the primary driver of behavior, we would expect to see systematic over-prediction across all risk levels. However, as shown in Figure 3, officers exhibit significant under-prediction in the highest risk groups. This contrast, namely over-prediction at the low end and under-prediction at the high end, provides strong empirical evidence for cognitive error rather than misaligned preferences.

5.5. Alternative Objectives and Preference Heterogeneity

The evidence above shows that officers are making prediction errors so long as they are trying to predict serious recidivism as we have defined it and so long as they have uniform preferences across each case. We feel that this is a reasonable standard to hold officers to, but officers may in practice disregard certain lesser forms of serious recidivism or prioritize cases involving certain victims/perpetrators. We now consider whether officers can be said to be making prediction mistakes even given this more flexible specification of their goal.

First, we consider an alternative recidivism outcome. As discussed in Section 3, our primary recidivism outcome includes all sex offenses and violence with injury. We will now consider a recidivism outcome that focuses only on Grievous Bodily Harm (GBH) to capture the possibility that officers may discount all but the most serious injuries when predicting risk to the victim. Second, we restrict our attention to cases involving particular perpetrator demographics (male vs. female and white vs. non-white) or initial offense codes assigned by emergency dispatchers (violence, domestic violence, and non-violence). This makes the set of cases within any given analysis more homogeneous and reduces the chances that apparent prediction mistakes are actually driven by preference heterogeneity.

Each row of Table 3 presents one of the scenarios considered in this section, aside from the first, which uses our baseline specification. The columns present summaries of the prediction mistake tests developed above. The first column presents the fraction of officers with points below the 45-degree line when using our shrunken, CATE-adjusted latent outcome.[12] Regardless of the outcome or subsample considered, we consistently find that at least 1.2% and as many as 9.2% of officers make objectively mistaken predictions.

[12]Because some of the case types considered are less common, our restriction that officers must be seen responding to at least 100 cases binds for different numbers of officers in each. Furthermore, we relax this restriction for some subsets: For cases initially marked violence or DV, we require 50 observations;

The second column of Table 3 presents the potential gains from moving protective services from a case in the fourth quintile of objective risk to a random un-protected case in the fifth quintile of objective risk. This is essentially the difference between the blue and red lines in Figure 2, evaluated at the fourth quintile. These gains are large and positive for each alternative considered.[13] This indicates that across the alternatives, officers are still making prediction mistakes in the sense of Rambachan (2024).

The third and fourth columns of Table 3 present evidence on over- and under-prediction. The third column shows beliefs minus actual recidivism in the first quintile of objective recidivism risk. Positive values indicate higher beliefs about recidivism than what actually occurs, i.e. over-prediction. The fourth column shows beliefs minus actual recidivism in the fifth quintile of objective recidivism risk. Negative values indicate lower beliefs about recidivism than what actually occurs, i.e. under-prediction. With two exceptions, the first-quintile difference is positive and the fifth-quintile difference is negative. One exception is the GBH outcome, in which officers appear to consistently over-estimate the risk of serious recidivism. This is driven by the fact that GBH arises in only about one percent of cases, so it is rare relative to the share of cases classified as high-risk. The other exception is cases with female perpetrators, in which officers consistently under-predict the risk of serious recidivism.

The patterns that established officer prediction mistakes above are therefore robust to considering alternative outcomes and subsets of cases. This evidence does not show once and for all that these patterns *must* be driven by prediction mistakes, but it provides us with additional confidence that this is the case.

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

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

for non-violence and non-white cases, we require 30 observations; for female perpetrators, we require 10 observations.

[13]The absolute size of the gains when using GBH as the recidivism outcome is smaller, but the GBH recidivism rate is only 1.1 percent.

where Y_i^* is the latent outcome as defined in the Appendix, Q_i is a vector of dummy variables capturing responses to the DASH questions, Z_i denotes control variables constructed from criminal and domestic abuse histories, and ε_i and u_i represent unobservable disturbance terms. The terms 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 databases. 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.[14] All of these regressions also include missing value dummies for each DASH feature. These equal one if the question was omitted and equal zero otherwise.

6.1. Questionnaire design

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 spikes centered on the end of each bar.[15]

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 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.

Several of these negative coefficients may stem from a faulty reference category. For example, Question 26 asks whether the perpetrator has breached a protective order. However, no question asks whether such an order was ever in place, meaning the omitted category consists of both perpetrators who never had a protective order against them and those who have had an order but never breached it. Suppose an order reduces risk, but breaching it raises risk relative to not breaching it. Then, absent a category for simply having an order in place, the dummy for breaching it can carry a negative coefficient whose sign depends on the two underlying coefficients and the shares of observations in each category. A similar logic could help explain the negative coefficients on Questions

[14]For comparison purposes, we also regress the observed outcome on the same set of variables. Results are reported in Appendix Table A2.

[15]Columns (1) and (3) of Appendix Table A2 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.

7 (conflict over children), 10 (children present who are not the perpetrator's), 11 (harm to children), and 12 (threats against children), if the presence of children per se, which is not recorded, reduces recidivism risk.

Another reason for the unexpected negative coefficients could stem from the manner in which the questionnaire was designed. Myhill et al. (2023) report that the questions were drawn in part from the study of domestic homicide reports, a rather extreme form of selection on the dependent variable. The result was questions capturing features that were prevalent among such cases. However, by Bayes rule, whether such features are actually predictive of recidivism depends not only on their prevalence in the recidivist sample, but also on that in the non-recidivist sample. Thus if most of the abusers who threaten to kill their partners do not go on to recidivate, death threats may be counter-predictive of recidivism, even if death threats are common among recidivist cases.

All this points to problems with questionnaire design. A prima facie reading of the questions in Table 1 suggests that each of them 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.[16]

Questionnaire design notwithstanding, the DASH features strongly influence officers' beliefs. Indeed, the coefficients from the beliefs regression tend to be much 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) and whether the perpetrator had threatened to kill the victim (Question 17). Also noteworthy are indicators for whether the victim felt isolated (Question 4), whether the perpetrator had threatened the children (Question 12), 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).

[16]This is not an issue of collinearity. In Appendix Table A3, 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 A2, we see that most of the variables with negative coefficients in the multivariate regressions also have negative coefficients in the bivariate regressions.

6.2. Representativeness bias

Here we begin to answer the question 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 representativeness should interact with prevalence. 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 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 means between 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 representative features from Figure 4 (and columns (1) and (3) of Appendix Table A2). 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 common features.

All of representative and rare features move officers' beliefs. Their coefficients are positive and significant and several are quite large. 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 negative, as discussed above.

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 that are representative and rare, which is broadly consistent with the prediction of the Bordalo et al. (2016) model.

Bordalo et al. (2016) note that representativeness bias can distort beliefs, but still stems from a "kernel of truth." Features that predict recidivism should still drive beliefs, but this relationship may be exaggerated. Thinking about this aspect of representativeness bias in our setting is complicated by the fact that some features that the DASH frames as predicting recidivism, which the officers seem to trust, do not actually predict recidivism. We therefore might expect that the objective representativeness of a feature as measured using actual recidivism outcomes may deviate significantly from perceived representativeness.

In this event, it is difficult to separate the effect on beliefs due to representativeness bias from the effect due to this more fundamental distortion. However, we think that even in this case the prevalence of a feature can be informative. When a feature is significantly more prevalent than recidivism itself, officers will intuitively understand that representativeness cannot be too high. Only 13 percent of perpetrators who officers encounter will become recidivists but, for example, 38 percent of them will be marked as having drug or alcohol problems. Even if all recidivists have drug or alcohol problems, that leaves nearly two thirds of perpetrators with alcohol or drug issues who must not be recidivists. So while perceived representativeness is unbounded for rare features, it is

naturally capped for prevalent features.[17]

When we expand our view of the data even to features that are not objectively representative - but are likely perceived as representative by officers - we see a strong negative correlation between beliefs and the prevalence of the features. Figure 7 plots DASH questions by their prevalence on the x-axis and their coefficient in Equation (5) on the y-axis. Officers' responses to the features fall with their prevalence, and above a rate of occurrence of 28%, common DASH features receive essentially no response from officers making risk assessment decisions.

Returning to Figure 6, the finding regarding alcohol and drug use is particularly important. Alcohol and drugs present problems for nearly two-fifths of the perpetrators. They also predict serious recidivism, increasing it by 0.014 (0.002), which 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 the largest coefficient, namely that for death threats to the victim (Question 17; $\hat{\beta} = 0.103$ (0.004)). 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

Several recent studies seek to explain overreaction to signals in belief formation problems. In Augenblick et al. (2025), cognitive noise makes it difficult for the decision maker (DM) to assess the strength of a signal. As a result, she adopts an intermediate signal strength as a default, which leads her to underreact to strong signals and overreact to weak ones. Ba et al. (2024) combine cognitive noise with representativeness bias. They show that in problems with binary outcomes and asymmetric priors, DM's may overreact, particularly in response to disconfirming signals, that is, signals that should move beliefs away from the more likely outcome. Fan et al. (2022) note that a Bayesian can predict an

[17]To see this formally, consider the representativeness of a feature k: $R(k) = \frac{Pr\{k|\text{Recidivist}\}}{Pr\{k|\text{Not Recidivist}\}}$. $R(k)$ is maxmized when $Pr\{k|\text{Recidivist}\} = 1$ and $Pr\{k|\text{Not Recidivist}\}$ is as low as possible. By the law of total probability, these terms must fulfill $pPr\{k|\text{Recidivist}\} + (1-p)Pr\{k|\text{Not Recidivist}\} = q$ where p is the probability of recidivism and q is the prevalence of k. The minimum $Pr\{k|\text{Not Recidivist}\}$ that makes this true is $\frac{q-p}{1-p}$. Thus, when $q > p$ (i.e. k occurs more frequently than recidivism), $R(k)$ can be no higher than $\frac{1-p}{q-p}$. When $q \leq p$, then $R(k)$ approaches infinity when $Pr\{k|\text{Not Recidivist}\} = 0$ and $Pr\{k|\text{Recidivist}\} = \frac{q}{p}$.

outcome in two steps. First, she uses the signal to infer the underlying state, then uses the Law of Iterated Expectations to predict the outcome. They propose that the DM may instead use a two-step heuristic they term "exact representativeness." First, the DM infers the state, but then instead of using the LIE, she predicts the outcome that is most representative of the state. Their experimental evidence shows that exact representativeness arises more often in settings where the signal is similar to the outcome.[18]

All of these models provide plausible explanations for officers' overreaction to the DASH features. Augenblick et al. (2025) potentially explains overreaction to weak signals such as Question 17 (death threats), with a representativeness less than one (see Figure 5). Ba et al. (2024) seems particularly germane to our setting, for two 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, exact representativeness strikes us as a plausible decision heuristic in our setting. Imagine that officers perceive two possible states for the perpetrator: "bad," which entails a low recidivism risk, or "worse," which entails high risk. Signals that are similar to serious recidivism, such as victim injury (Question 1), lead the officer to infer "worse," whose representative outcome is recidivism. The importance of similarity between signal and outcome may help explain why we observe the greatest overreacton to signals such as victim injuries and death threats (Question 17), and less overreaction to other signals that are less similar to serious recidivism.

6.4. Categorization and selective attention

The discussion above provides a number of potential explanations for officers' overreaction to the DASH features. Here we consider the role of categorization and selective attention. This model may explain not only overreaction, but also an important set of interactions among the 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 signals. Ancillary signals, which may or may not be predictive of the outcome, trigger a mental category from the DM's past experience. Categorization then leads the DM to attend to a subset of features that are relevant to the category. The predictions from the model are that: (i) the realization of a trigger feature should lead to instability in the DM's belief, and (ii) it should affect the distribution of features to which she attends.

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

Analyzing these predictions in practice requires two key pieces of information. First is the set of trigger features. Second is the number, and nature, of the categories. In an experimental setting, the category (e.g., inference v. prediction from a sequence of coin tosses) is part of the design, and the trigger features (e.g., the length of a run of heads) are manipulated by the experimenter (Bordalo et al., 2023, 2025). In an observational setting, we have no control over such matters. Instead, we use findings from prior research to specify the trigger features. We then ask whether these features generate instability of beliefs. The next step is to learn the number of categories that underlie the trigger features. For this we apply natural language processing techniques to the trigger questions themselves. We then estimate selective attention by trigger feature, and find evidence consistent with the same set of categories. We close with an admittedly speculative interpretation of the categories and two alternative explanations for our findings. Since we do not control key features of the environment, but rather learn them from the data, we do not view this exercise as a conventional test of the predictions. Rather, we view it as asking whether the patterns we observe are plausibly consistent with a model of categorization and selective attention.

For the trigger features, we turn to a survey that asked which features officers find to be essential for evaluating risk (Robinson et al., 2018). The five factors receiving the most support, the share of officers regarding them as essential, and the DASH features to which they most closely correspond, are: (i) using or threatening to use a weapon (67 percent, Question 16), strangulation (64 percent, Question 18), escalation of abuse (53 percent, Questions 13 and 14), physical assault resulting in injury (51 percent, Question 1), and making threats to kill (43 percent, Question 17). We first consider how these features affect the stability of officers' beliefs.

Since we only observe binary predictions from the officers, we cannot make statements about how the full distribution of posterior beliefs shifts in the presence of triggers. However, we can show shifts in binary predictions that far exceed a rational response to just the information from the trigger feature. Figure 8 shows the share of officers predicting high risk by each of the trigger variables. The smallest change in beliefs involves abuse happening more often (Question 13), where officers predict high risk roughly 22 percent of the time in its presence, but only 5 percent of the time in its absence. At the other end of the spectrum are death threats (Question 17). In the presence of past death threats, officers classify 36.5 percent of cases as high risk, compared to only 6.5 percent in its absence. All these features are associated with instability in officers' beliefs.

Our next question is: how many mental categories underlie these six features? To answer it, we analyze lexical and semantic similarities among the corresponding DASH questions. The idea is that questions that sound similar, either in terms of wording or

meaning, likely reflect the same category.

To begin, we construct the term frequency-inverse document frequency (TF-IDF) matrix of the six DASH questions. We then compute the cosine similarities among them, which measure the similarity of their wording. The similarities displayed in Table 4 suggest three categories. One pertains to the victim injury question (Question 1), which is almost completely dissimilar from the other questions. Another underlies Questions 13 and 14, which ask whether the abuse is happening more often or getting worse. These two questions have higher similarity than any other pair of questions, and at the same time, are largely unrelated to the others. Finally, the questions on past weapons use, death threats, and strangulation are considerably more similar to each other than to the other questions.

A similar pattern appears when we construct semantic, rather than lexical, similarities. Table 5 reports cosine similarities from embeddings produced by the all-MiniLM-L6-v2 model, a language model trained on a large sample of English text (Reimers and Gurevych, 2021). Similarities of such embeddings generally do a better job in capturing correlated meaning than do similarities based on the TF-IDF matrix. Overall, the similarities in Table 5 are higher than those in 4, reflecting that all these questions share background meaning about aspects of domestic abuse. Otherwise, the pattern in Table 5 is similar to that in Table 4. The relatively low similarities in the first column show that the victim injury feature is largely distinct from the others, the high similarity between Questions 13 and 14 indicates that their meaning, as well as their wording, is largely the same, and Questions 16, 17, and 18 are closely related semantically. Moreover, questions within these three groups are substantially more similar to each other than to those across groups. Both the semantic and lexical analyses point to the same three categories.

If these categories are meaningful, then trigger variables within the same category should raise attention to roughly the same set of features. To measure selective attention, we estimate a pair of beliefs regressions (as in equation 5) for each trigger feature, one from the subsample in which the feature is present, and the other from the subsample in which it is absent.[19] We then compute the difference, or contrast, between coefficients for each pair of regressions, and represent the contrasts in the form of a heatmap. The contrasts show how officers' attention changes as we move from a trigger feature being present to its being absent.

In Figure 9, the row variables are the DASH features and the column variables are the trigger features. Maroon cells represent positive contrasts and gray cells represent negative ones. Darker shading represents larger (absolute) contrasts. Solid-colored cells

[19]Appendix Figure A1 shows the coefficients on the DASH features for these two regressions, splitting the sample by victim injury (Question 1).

reflect contrasts that are significant at the 10 percent level; cross-hatched cells represent insignificant contrasts.

If we focus on larger contrasts, meaning those of magnitude 0.04 or greater, the patterns of selective attention in Figure 9 are roughly consistent with the three categories from the NLP analyses above. Victim injury (Question 1) raises attention to heightened victim fear (Question 2), pregnancy or recent childbirth (Question 9), prior use of weapons (Question 16), death threats (Question 17), strangulation attempts (Question 18), sexual humiliation (Question 19), and two other features. Questions 13 and 14 (more frequent and worsening abuse) raise attention to five features in common: victim injury (Question 1), pregnancy (Question 9), harm to children (Question 11), prior use of weapons (Question 16), and prior threats to kill (Question 17). Three of these features are also associated with the category underlying victim injury, but four features stimulated by a victim injury are not involved with this second category. Finally, questions 16, 17, and 18 all raise attention to six features: victim injury (Question 1), victim fear (Question 2), victim isolation (Question 4), perpetrator harassment (Question 8), worsening abuse (Question 14), and prior trouble with police (Question 27). Three of these features are completely distinct from the features whose attention is raised by the other two categories. The pattern in Figure 9 seems largely consistent with a model of three cognitive categories that raise attention to partially overlapping sets of features.

We now offer an admittedly speculative interpretation of the three categories. We first note that the sets of ancillary features that trigger the three different categories capture different temporal aspects of relationship dynamics. Victim injury (Question 1) captures a specific moment in the immediate past. The questions about the abuse becoming more frequent or getting worse (Questions 13 and 14) capture a trend in the status of the relationship. The last three questions, pertaining to past weapons use, death threats, and strangulation (Questions 16, 17, and 18) refer to events that occurred one or more times in the unspecified past. We refer to these categories as "immediate threat," "escalation," and "lethal threat," respectively.

Each category raises attention to features that trigger some of the other categories. Immediate threat raises attention to past weapons use, death threats, and strangulation, which may be suggestive of the potentially dire consequences to follow if the incident in the immediate past were to repeat. Escalation raises attention to an immediate injury and a history of past weapons use, death threats, and strangulation, all potentially outcomes of further escalation. Lethal threats raise attention to immediate injury and worsening abuse, potentially because they make those threats appear more credible. One interpretation of this pattern is that officers' attention is particularly piqued when they observe two or more of the temporal aspects of relationship dynamics pointing toward

greater risk at the same time.

Each category also raises attention to a distinct set of features. Immediate risk raises attention to victim fear (Question 2) and pregnancy (Question 9), both of which may indicate greater susceptibility to abuse. Escalation raises attention to potential spillover harm involving children (Questions 11 and 12), who might be the targets of further escalation. Lethal threat raises attention to the perpetrator harrassing the victim (Question 8) and trouble with police (Question 27), both of which may capture volatility, and higher risk, on the part of the perpetrator.

Of course, there are other potential explanations for these patterns as well. We consider two. The first is that the interactions we observe reflect officers' accurate beliefs about the data generating process of true risk. The second is that officers simply cluster similar-sounding questions and respond to them as a group, producing spurious correlations that mimic selective attention.

Figure A3 presents a heatmap analogous to Figure 9, except where the outcome variable is latent recidivism. None of the trigger variables interact much with the other DASH features. Most of the cells indicate insignificant contrasts, and the significant ones tend to be much smaller than their counterparts in Figure 9. The interactions between features in forming officers' predictions do not appear to reflect accurate beliefs about the data generating process.

To evaluate the similar-sounding-features hypothesis, we note that there is both supporting and contrary evidence. By necessity, all of the features involved in the interactions sound similar insofar as they all deal with aspects of domestic abuse. Beyond that, we see greater interactions between the three categories than within them. Since Question 1 is a single-question category, it doesn't contribute to this comparison. However, in Figure 9 we see all three questions in the lethal threat category (questions 16, 17, and 18) raising attention to less similar-sounding features outside the category, but they change attention to other features within the category little if at all. Finally, the questions in the escalating abuse category (Question 13 and 14), which sound more similar to each other than to others in the questionnaire, affect attention to each other very little while substantially raising attention to several other features, as discussed above.

Another factor that supports the categorization interpretation of our results are the findings of instability. Figure 8 shows that all of the essential features trigger sharp changes in officers' beliefs. This is a prediction from the model of Bordalo et al. (2025), but it does not necessarily follow from a general notion of responses to similar-sounding questions.

Although the evidence for these two alternative hypotheses is weak, a remaining cause for concern about the model of categorization and selective attention is the fact that

Bordalo et al. (2025) expect that trigger features will cause both an increase in attention to some features and a decrease in attention to others. Figure 9 shows both, but the decreases tend to be fairly small, and many are insignificant. One possibility, since we only observe attention to features recorded in our data, is that DASH-based triggers may nevertheless reduce attention to features that are observable only to officers on the scene.

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 A2. 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 that vary over incidents. 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: drop observations with Question 28 = 1. The remaining Question 28 = 0 subsample would yield consistent estimates, since officers used no other information to classify those cases.

Although a literal interpretation of Question 28 seems overly optimistic, it might be reasonable to conjecture 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.

These estimates are presented in columns (6) and (7) of Appendix Table A2. For the most part, they 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 based on them remain unchanged.

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 A4, 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 it was weakly less than or greater than the sample median. Appendix Figure A5 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 numerous counter-predictive questions. In Appendix Figure A6 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 high-skill group as officers above the ROC curve, and the medium-skill group as those in between. With these definitions, we restrict attention to officers with at least 100 calls for service involving DA.

Figure 10 displays coefficients from the skill-group-specific beliefs regressions. Coefficients for low-skill officers are gray, those for medium-skill officers are beige, and those for high-skill officers are maroon. All officers tend to overreact to the DASH features, but in many cases, the extent of overreaction falls with the level of skill. Prominent examples include victim injury (Question 1), whether the victim is frightened (Question 2), harm to children (Question 11), abuse getting worse (Question 14) , death threats (Question 17), choking (Question 18), sexual humiliation (Question 19), and harm to animals (Question 22). Despite the small sample sizes in the low- and high-skill groups (n = 5,952 and 8,225, respectively), many of the between-group differences are significant.

Comparing the coefficients of these beliefs regressions with those from the corresponding recidivism regressions reported in Appendix Figure A7 draws another distinction between officers of different skill levels. The last set of coefficients in the beliefs regressions (Figure 10) shows that all sets of officers revise upward in response to information not on the DASH form, as collected by Question 28. Low-skill officers respond the most, whereas

high-skill officers respond the least. However, that information predicts recidivism differently according to officer skill. In the last set of coefficients in the recidivism regressions (Appendix Figure A7), we see that that information positively predicts recidivism for the medium- and high-skill officers, but it negatively predicts it for their low-skill counterparts.[20] 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.

The finding that low-skill officers respond differently to private information may imply a violation of CIA. The important question becomes whether any violation is strong enough to be driving our results. To check, we re-ran all of our analyses dropping cases where low-skill officers checked the Question 28 box. Summary results regarding our tests for mistakes vs. preferences are reported in the last row of Table 3.

We see that the share of officers below the 45-degree line falls from 9.1 percent in the full sample to 2.5 percent when we drop low-skill officers who used private information. Thus misuse of private information helps explain worse-than-random officer performance, but does not eliminate it entirely. The other results from this sample are very similar to those from the full sample. Officers classify many objectively low-risk cases as high risk, as shown in column (2). They also overpredict low-risk cases and underpredict high-risk cases. The conclusion that poor officer predictions represent mistakes does not hinge on low-skill officers making poor use of private information.

Coefficients from beliefs and recidivism regressions for this sample are shown in Appendix Figure A8. They are quite similar to their full-sample counterparts in Figure 4. The subset pertaining to representative features, shown in Appendix Figure A9, is likewise similar to the full-sample results in Figure 6. Finally, Appendix Figures A10 and A11 reproduce Figures 8 and 9. Instability of beliefs is similar in the smaller sample. Tests for differences between the two heatmaps resulted in 9 rejections at the 10 percent level, out of 156 cells. The joint test for the significance of the difference between the two heatmaps yielded a p-value of 0.68. None of our main results is substantially changed by the presence of low-skill officers who misuse private information.

[20]The vast majority of the cross-group contrasts between coefficients in Appendix Figure A7 are insignificant. For each question, we tested the equality of all three pairs of coefficients. Of the resulting 84 tests, only four had p-values below 0.05. Two of these involved Question 28. One was for the difference in the coefficients between the low- and medium-skill groups, and the other involved the coefficients for the low- and high-skill groups. The p-value for the former contrast was 0.005, for the latter it was 0.026.

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 resembles that of workers in other settings asked to solve important screening problems with little 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 intervention that follows from the decision-maker's prediction. We propose using causal forests to make that adjustment. The adjustment was of little practical consequence in our setting, since the effectiveness of the intervention was so limited. We offer the approach as a proof of concept for settings in which interventions have more bite.

Our results 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 that are representative but rare and underreact to features that are representative and prevalent. One implication of this is that officers largely ignore prior alcohol and drug problems among perpetrators, although they 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), to the presence of weak signals and an asymmetric prior, as in Ba et al. (2024), or to the use of the "exact representativeness" heuristic (Fan et al., 2022).

Furthermore, we find evidence of selective attention in officer decision making. Survey research has shown that officers find a small set of features to be essential for predicting risk. Several sources of evidence suggest that these features constitute three mental categories into which officers classify incidents. These categories trigger instability in the officers' predictions and shifts of attention toward related features, consistent with the Bordalo et al. (2025) model of categorization and selective attention. 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 stems directly from selective attention. Questionnaire designers routinely test questionnaire items to make sure they function as planned (Sudman

and Bradburn, 1982). However, such testing is usually carried out on each question in isolation. In the presence of selective attention, a questionnaire item may perform well on its own, but at the same time, change the decision-maker's attention to other items, potentially causing over- or under-reaction to them. Put differently, in the presence of selective attention, the performance of the whole may be less than the sum of its parts. Effective questionnaire design requires questions to be tested as an ensemble, with an explicit eye toward changes in attention.

Our analysis of Question 28, where the officer indicates whether she used additional information beyond the DASH questionnaire, reveals important heterogeneity. All officers revise recidivism risk upward in response to such information, but higher-skill officers find information that positively predicts recidivism, whereas the information found by lower-skill officers negatively predicts it. Higher-skill officers seem to find predictive contextual information, whereas lower-skill officers latch onto misleading cues. This points toward training interventions focused on improving information processing, if one were intent on continuing with a risk assessment protocol along the lines of DASH.

Our findings raise a more fundamental question about the risk assessment process. Currently, officers combine predictors of risk in some unknown manner and then make a decision about the level of risk faced by the victim. Alternatively, the combining of information could be left to a statistical algorithm that would make a recommendation to the officer. The officer could override that recommendation; if she did so on the basis of information not available to the algorithm, her override might improve the prediction.

Our analysis revealed that a simple statistical algorithm, a linear regression based on the DASH features, outperforms most of the officers in our sample. Even a simple machine-learning algorithm would presumably do better still, preventing many of the failures documented here. Such systems are currently being used for decision support in areas such as pre-trial release and child protective custody decisions. Such technology could probably be fruitfully employed in predicting the risk faced by domestic abuse victims as well.

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.

Bielen, S., Grajzl, P., 2021. Prosecution or Persecution? Extraneous Events and Prosecutorial Decisions. Journal of Empirical Legal Studies , 765–800.

Bindler, A., Hjalmarsson, R., 2019. Path Dependency in Jury Decision Making . Journal of the European Economic Association , 1971–2017.

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., Conlon, J.J., Gennaioli, N., Kwon, S.Y., Shleifer, A., 2023. How people use statistics. Technical Report. URL: https://www.nber.org/papers/w31631.

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.

Chen, D.L., Moskowitz, T.J., Shue, K., 2016. Decision Making under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires. Quarterly Journal of Economics 131, 1181–1242.

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.

Eren, O., Mocan, N., 2018. Emotional Judges and Unlucky Juveniles. American Economic Journal: Applied Economics , 171–205.

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.

Fan, T.Q., Liang, Y., Peng, C., 2022. The inference-forecast gap in belief updating URL: https://web.stanford.edu/~tonyqfan/IF_gap_paper.pdf. working Paper.

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.

Gu, J., Walters, C., 2022. Emprical Bayes methods: theory and application. Technical Report.

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.

Heyes, A., Saberian, S., 2019. Temperature and Decisions: Evidence from 207,000 Court Cases. American Economic Journal: Applied Economics , 238–265.

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.

Imbens, G., 2015. Matching methods in practice: three examples. Journal of human resources 50, 333–419.

Jordan, A., 2025. Racial Patterns in Approval of Felony Charges. Working Paper. Washington University in St. Louis.

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.

Ludwig, J., Mullainathan, S., 2024. Machine learning as a tool for hypothesis generation. The Quarterly Journal of Economics , 1–77doi:10.1093/qje/qjad055. advance Access publication on January 10, 2024.

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.

Murphy, K.M., Topel, R.H., 1985. Estimation and inference in two-step econometric models. Journal of Business and Economic Statistics 3, 370–379.

Myhill, A., Hohl, K., Johnson, K., 2023. The 'officer effect' in risk assessment for domestic abuse: Findings from a mixed methods study in england and wales. European Journal of Criminology 20, 856–877. doi:10.1177/14773708231156331.

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.

Reimers, N., Gurevych, I., 2021. all-MiniLM-L6-v2: Sentence transformers model. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Fine-tuned on 1B sentence pairs using a contrastive learning objective; maps sentences to a 384-dimensional dense vector space.

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., Pinchevsky, G.M., Guthrie, J.A., 2018. A small constellation: risk factors informing police perceptions of domestic abuse. Policing and Society 28, 189–204. doi:10.1080/10439463.2016.1151881.

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., 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.

Appendix – for online publication

1. Causal forest imputation with two treatments

Here we generalize the discussion above to account for the fact that our setting involves two treatments. The first is the high-risk designation, the outcome of the risk assessment discussed above. The second is a criminal charge, the outcome of the criminal investigation carried out at the same time.

Define the first binary treatment variable as $D_1 = 1$ if the case is designated as high-risk, and $D_1 = 0$ otherwise, as above. Define the second binary treatment variable as $D_2 = 1$ if criminal charges are filed, and $D_2 = 0$ otherwise. Define the potential outcome as $Y(D_1, D_2) \in \{0, 1\}$, so $Y(0, 0)$ denotes the untreated potential outcome, $Y(1, 0)$ denotes the potential outcome if the case is classified as high-risk but not charged, $Y(0, 1)$ denotes the potential outcome if the case is charged but not designated as high-risk, and $Y(1, 1)$ designates the potential outcome if the case is both designated as high-risk and charged. Denote the treatment effects as

$$\begin{aligned}\tau_{10} &= Y(1,0) - Y(0,0)\\ \tau_{01} &= Y(0,1) - Y(0,0)\\ \tau_{11} &= Y(1,1) - Y(0,0).\end{aligned}$$

Then the observed outcome can be written in terms of the treatment variables, the untreated potential outcome, and the treatment effects as

$$\begin{aligned}Y &= D_1D_2Y(1,1) + D_1(1-D_2)Y(1,0) + (1-D_1)D_2Y(0,1) + (1-D_1)(1-D_2)Y(0,0)\\ &= Y(0,0) + D_1\tau_{10} + D_2\tau_{01} + D_1D_2[Y(1,1) - Y(1,0) - Y(0,1) + Y(0,0)]\\ &= Y(0,0) + D_1(1-D_2)\tau_{10} + (1-D_1)D_2\tau_{01} + D_1D_2\tau_{11}.\end{aligned}$$

Under conditional independence and a number of technical regularity assumptions, Athey et al. (2019) propose a consistent estimator for the conditional average treatment effects, or CATEs, which are given by

$$\tau_{jk}(x) \equiv E(\tau_{jk}|X = x) \qquad j, k = (1,0), (0,1)$$

Denoting the estimated CATEs by $\hat{\tau}_{jk}$, we impute the untreated potential outcome $Y(0, 0)$ by Y^*, where:

$$Y^* = Y - D_1(1-D_2)\hat{\tau}_{10} - (1-D_1)D_2\hat{\tau}_{01} - D_1D_2\hat{\tau}_{11}$$

We estimate the CATEs from three samples, each of which pairs one of the treatment groups with the comparison group.

2. Consistency of the latent outcome regression

Here we analyze the use of the estimated CATEs in the latent outcome regression. We return to the case of a single intervention for simplicity, where $Y_i = D_{i1}Y(1) + (1 - D_{i1})Y_i(0)$. Define the idiosyncratic treatment effect as $\tau_i \equiv Y_i(1) - Y_i(0)$. To begin, we assume the CATEs, given by $\tau(x_i) \equiv E(\tau_i|X_i = x_i)$, are known. Thus

$$Y_i(0) = Y_i - D_{i1}\tau_i.$$

Under conditional independence, $Y_i(0), Y_i(1) \perp\!\!\!\perp D_{i1}|X_i$, which is also a necessary condition to consistently estimate the CATEs, OLS would yield consistent estimates of the latent outcome regression

$$Y_i - D_{i1}\tau_i = X_i\beta + e_i$$

if τ_i were known. Replacing the unknown τ_i with the known $\tau(x_i)$ yields

$$Y_i - D_{i1}\tau(x_i) = X_i\beta + e_i + D_{i1}v_i, \tag{7}$$

where $v_i \equiv \tau_i - \tau(x_i)$. If $E(D_{i1}v_i|X_i) = 0$, then OLS applied to (7) should yield consistent estimates. Use the Law of Iterated Expectations to write

$$E(D_{i1}v_i|X_i) = P(D_{i1} = 1|X_i)E(v_i|X_i, D_{i1} = 1)$$

By conditional independence,

$$\begin{aligned} & \tau_i \perp\!\!\!\perp D_{i1}|X_i \\ \implies & \tau(x_i) + v_i \perp\!\!\!\perp D_{i1}|X_i \\ \implies & v_i \perp\!\!\!\perp D_{i1}|X_i, \end{aligned}$$

where the last expression holds since $\tau(x_i)$ is fixed given $X_i = x_i$. This implies that $E(D_{i1}v_i|X_i) = E(v_i|X_i) = 0$, where the last equality holds since v_i is defined as the difference between a random variable and its conditional expectation.

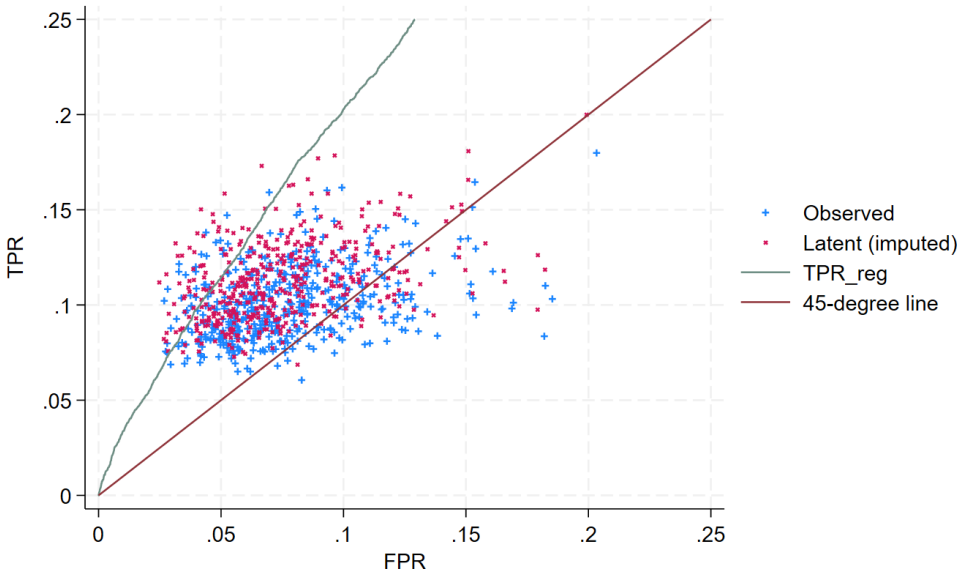
Now assume that $\tau(x_i)$ must be replaced by a consistent estimate $\hat{\tau}(x_i)$. Equation (7) becomes

$$Y_i - D_{i1}\hat{\tau}_i = X_i\beta + e_i + D_{i1}(v_i - \hat{\eta}_i), \tag{8}$$

where $\hat{\eta}_i = \hat{\tau}(x_i) - \tau(x_i)$. Although $\hat{\eta}_i$ is a function of X_i, by an argument along the lines of Murphy and Topel (1985), consistency of $\hat{\tau}(x_i)$ implies that $\hat{\eta}$ converges to zero and thus is uncorrelated with X_i asymptotically.

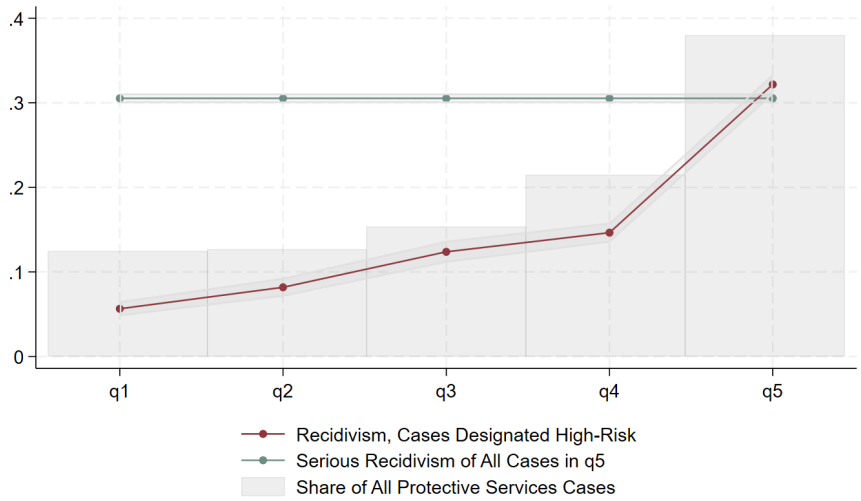
3. Figures and tables

Figure 1: Officer-level false vs. true positive rates, based on observed and latent outcomes



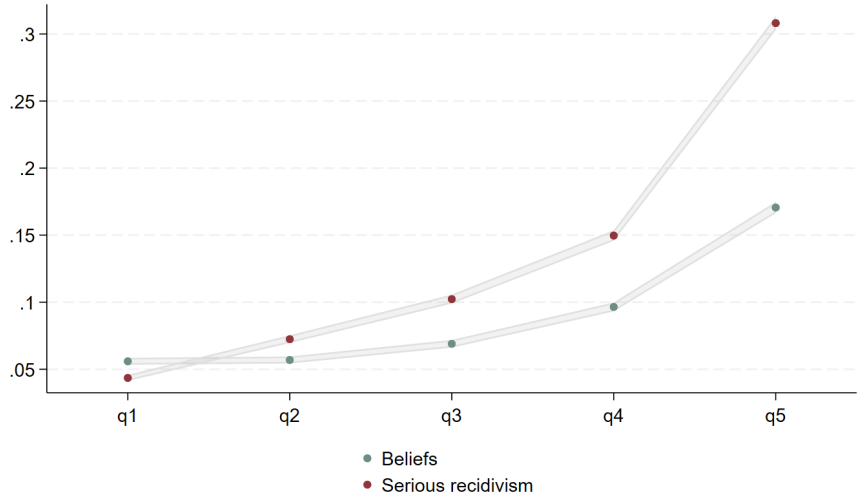
Note: Points plot (FPR, TPR) coordinates by officer for officers with at least 100 DA cases. Points have been shrunk towards the mean by empirical Bayes methods. The curve labeled Reg. ROC is the receiver operating characteristic curve from a linear regression of recidivism on the DASH features. It shows the best feasible tradeoff available between FPR and TPR for a given prediction technology at varying thresholds. It extends to the point (1,1), but is truncated here for the sake of clarity.

Figure 2: Share of cases classified as high risk, by objective risk quintile, and predicted high-risk share by objective risk quintile



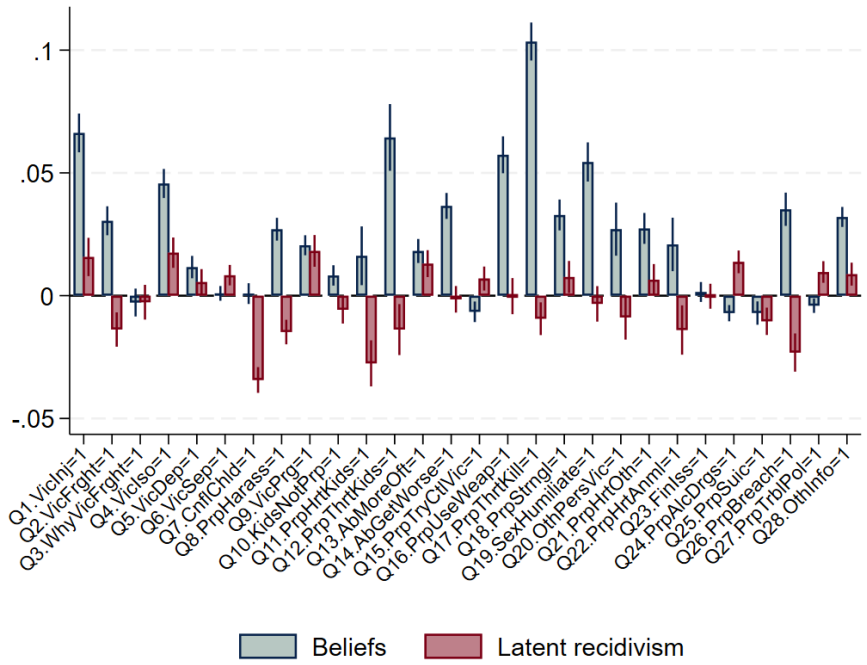
Note: This figure plots average latent recidivism by quintile of predicted risk from a regression of recidivism on case characteristics. The red line shows the serious recidivism rate of cases given protective services in each quintile. The horizontal line shows the serious recidivism rate of cases in the fifth quintile, including those not given protective services. The histogram shows the distribution of high-risk designations across quintiles.

Figure 3: Beliefs and mean recidivism, by risk quintile



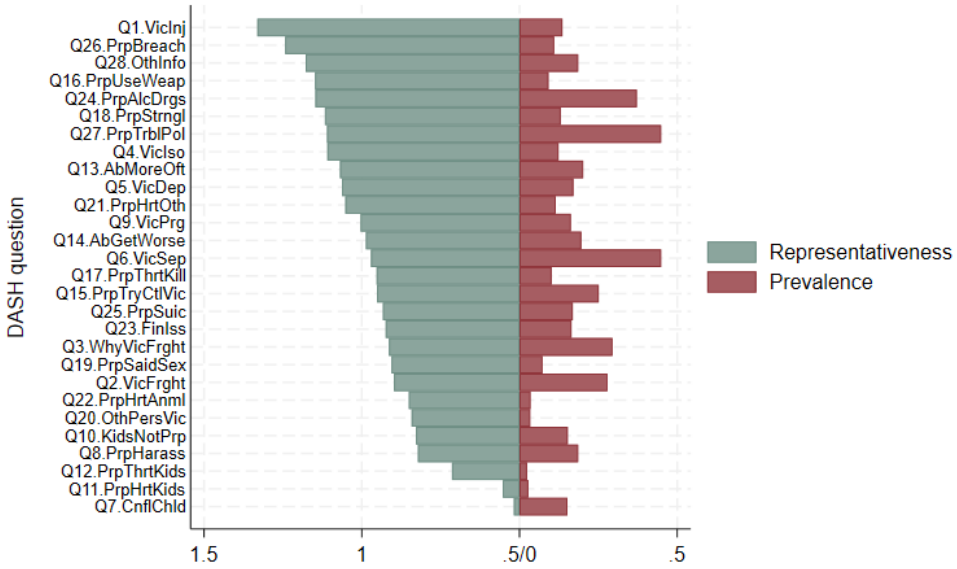
Note: This figure plots average latent recidivism and average rates of predicted risk by quintile of predicted risk from a regression of recidivism on case characteristics.

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



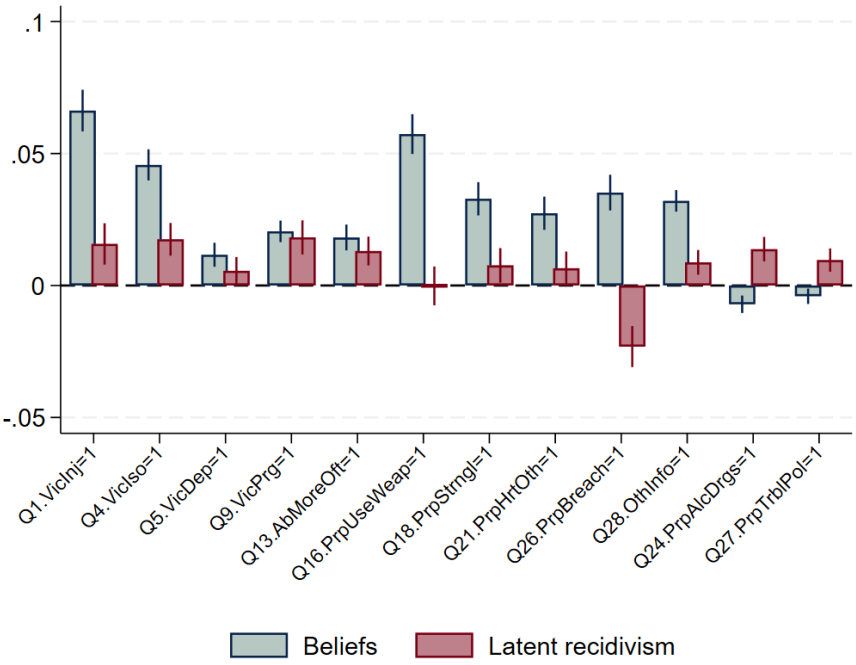
Notes: Coefficients from regressions of beliefs and latent recidivism on DASH features and other controls, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure 5: Representativeness and prevalence of DASH features



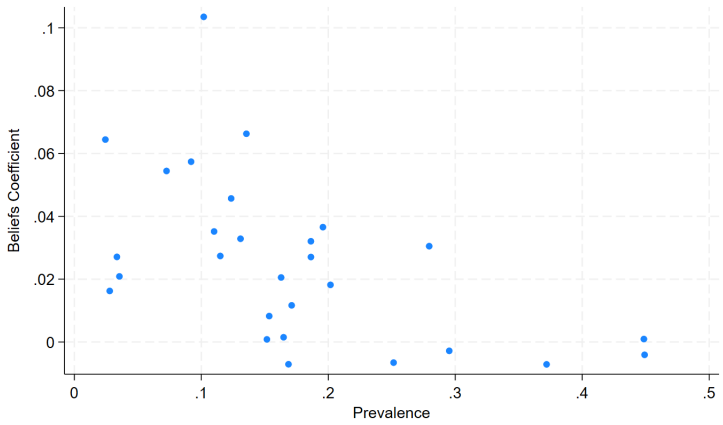
Note: Representativeness increases to the left, starting at .5. Prevalence increases to the right, starting at 0.

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: Scatterplot of responses of beliefs to prevalence of DASH features



Notes: The figure provides a scatterplot of the coefficients of the DASH features from the beliefs regression from Figure 4 against feature prevalence from Figure 5.

Figure 8: Share of officers predicting high risk, by trigger features

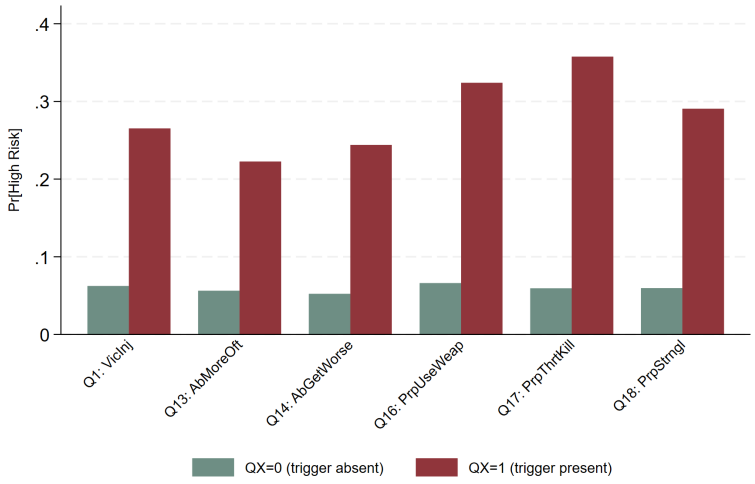
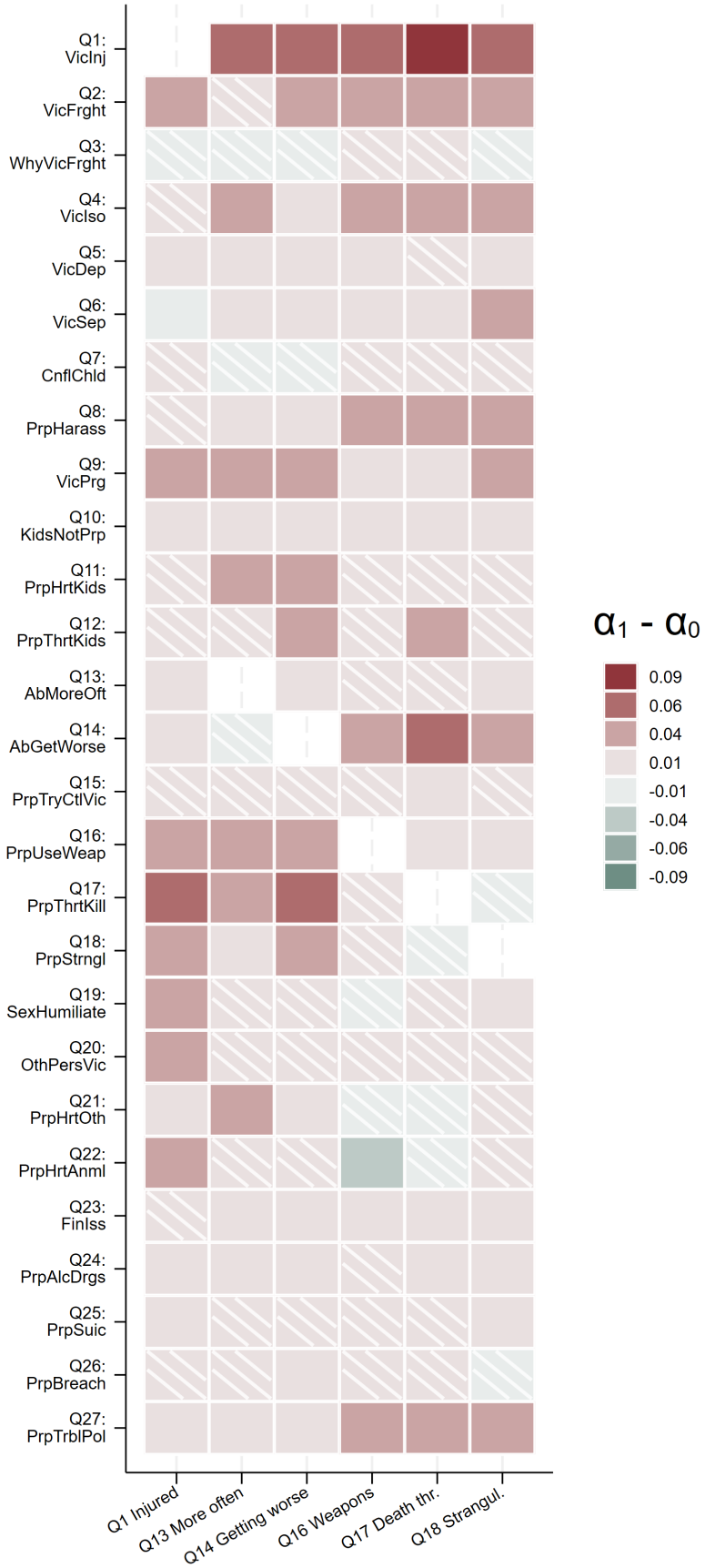
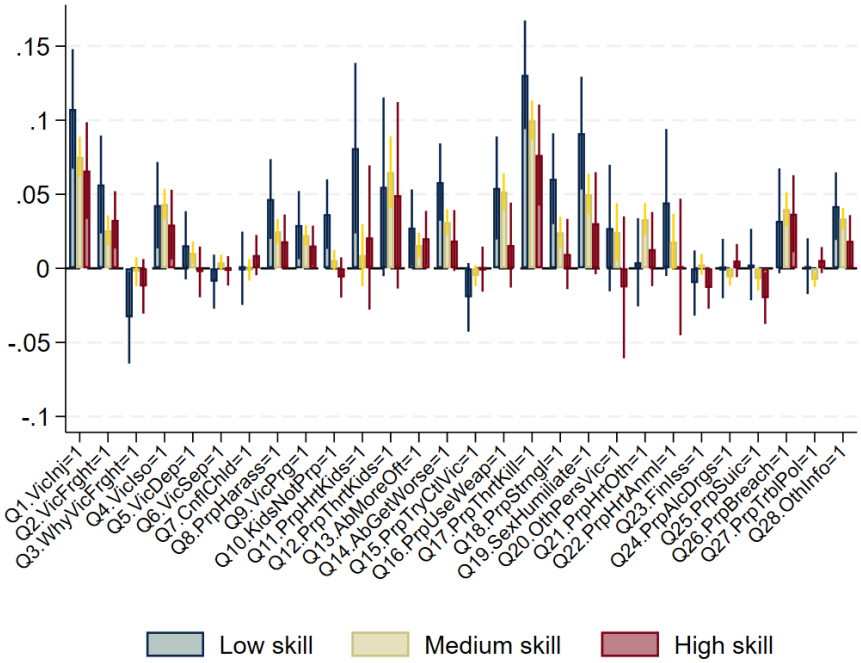


Figure 9: Heatmap showing contrast in beliefs coefficients, by trigger variable



Note: Each cell reflects the magnitude of the contrast between the coefficient of the row variable from two beliefs regressions: one in which the trigger (column) variable equals one and one (α_1) in which it equals zero (α_0). Solid cells indicate that the contrast was significant at the 10 percent level, cross-hatched cells indicate that it was insignificant.

Figure 10: 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. Officer skill is determined in terms of Figure 1: high-skill are points above the ROC curve; low-skill are points below the 45-degree line, and medium-skill are in between. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

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	121,925	11,192	133,117
	0.916	0.084	1
Yes	18,188	2,629	20,817
	0.874	0.126	1
Total	140,113	13,821	153,934
AUC	0.521		

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: Tests for Prediction Mistakes in Alternative Specifications

Specification	Pct Below 45-Degree Line	Q5 Recid - Selected Q4 Recid	Q1 Beliefs - Q1 Recid	Q5 Beliefs - Q5 Recid
Baseline	0.089	0.154***	0.010***	–0.133***
GBH	0.051	0.021***	0.063***	0.126***
Male Perp	0.092	0.148***	0.016***	–0.093***
Female Perp	0.031	0.203***	–0.031***	–0.318***
White Perp	0.081	0.165***	0.002	–0.162***
Non-white Perp	0.037	0.105***	0.022***	–0.062***
Violence	0.012	0.157***	0.026***	–0.099***
Domestic Violence	0.081	0.167***	0.000	–0.152***
Non-Violence	0.039	0.142***	0.028***	–0.154***
No Low-Skill Q28=Yes	0.025	0.152***	0.009***	–0.134***

This table presents tests for prediction mistakes across several alternative specifications. The specifications are specified by row. Baseline uses our full sample and standard recidivism outcome. GBH uses our full sample and a recidivism outcome that restricts only to Grievous Bodily Harm. Male Perp, Female Perp, White Perp, and Non-white Perp look at subsamples defined by the demographics of the perpetrator. Violence, Domestic Violence, and Non-Violence condition on the initial offense code issued when officers were dispatched to the scene. The final row removes cases where low-skill officers (those below the 45-degree line in ROC space) marked Q28, which indicates use of information outside the DASH questionnaire. The Pct Below 45-Degree Line column presents the fraction of officers with shrunken prediction outcomes below the 45-degree line in ROC space using observational data. The Q5 Recid - Selected Q4 Recid column presents the difference between the recidivism rate of a random case in the fifth quintile of objective recidivism risk and the recidivism rate of cases from the fourth quintile selected for protective services. The Q1 Beliefs - Q1 Recid column presents the difference between officer beliefs about recidivism in the first quintile of objective recidivism risk and actual recidivism. The Q5 Beliefs - Q5 Recidivism column presents the difference between officer beliefs about recidivism in the fifth quintile of objective recidivism risk and actual recidivism.

Table 4: TF-IDF cosine similarity of selected DASH questions

	Q1	Q13	Q14	Q16	Q17	Q18
Q1.CurIncInj	1.000					
Q13.AbMoreOft	0.011	1.000				
Q14.AbGetWorse	0.013	0.375	1.000			
Q16.AbUseWeap	0.055	0.010	0.012	1.000		
Q17.AbThrtKill	0.045	0.008	0.009	0.158	1.000	
Q18.AbStrngl	0.050	0.009	0.011	0.164	0.143	1.000

Note: The entries are cosine similarities between the term frequency-inverse document frequency vectors of the six trigger questions. They measure overlap in wording across questions, with IDF automatically down-weighting words that appear in many of the questions.

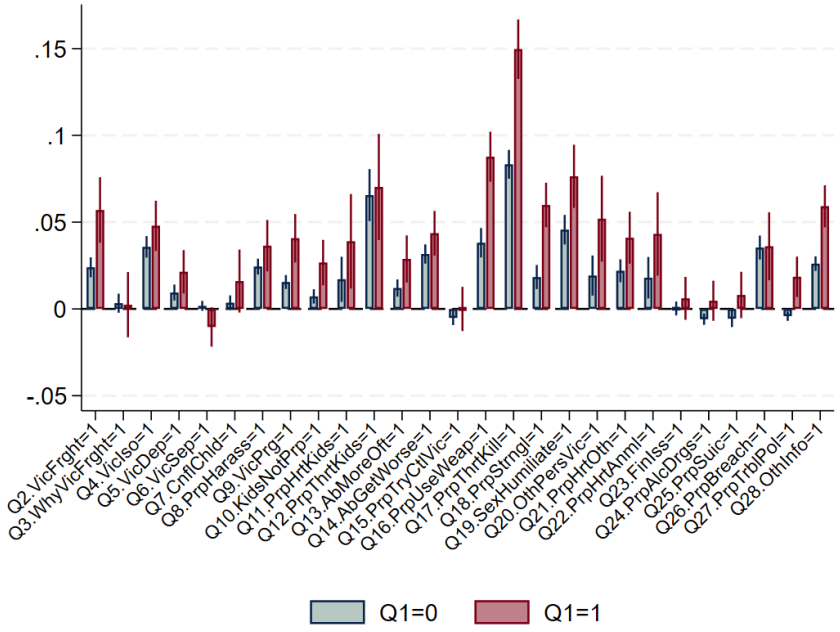
Table 5: Sentence-embedding cosine similarity of DASH questions

	Q1	Q13	Q14	Q16	Q17	Q18
Q1.InjFrmInc	1.000					
Q13.AbMoreOft	0.300	1.000				
Q14.AbEscalate	0.301	0.882	1.000			
Q16.Weapons	0.372	0.556	0.537	1.000		
Q17.ThrtKill	0.283	0.584	0.581	0.730	1.000	
Q18.Strangl	0.306	0.496	0.503	0.677	0.698	1.000

Note: The entries are cosine similarities between the embedding vectors of the six trigger questions. Each question is represented by a dense vector obtained from the all-MiniLM-L6-v2 model sentence-embedding model (Reimers and Gurevych, 2021), and the cosine measures how close two questions are in that vector space. This reflects semantic similarity rather than overlap in wording.

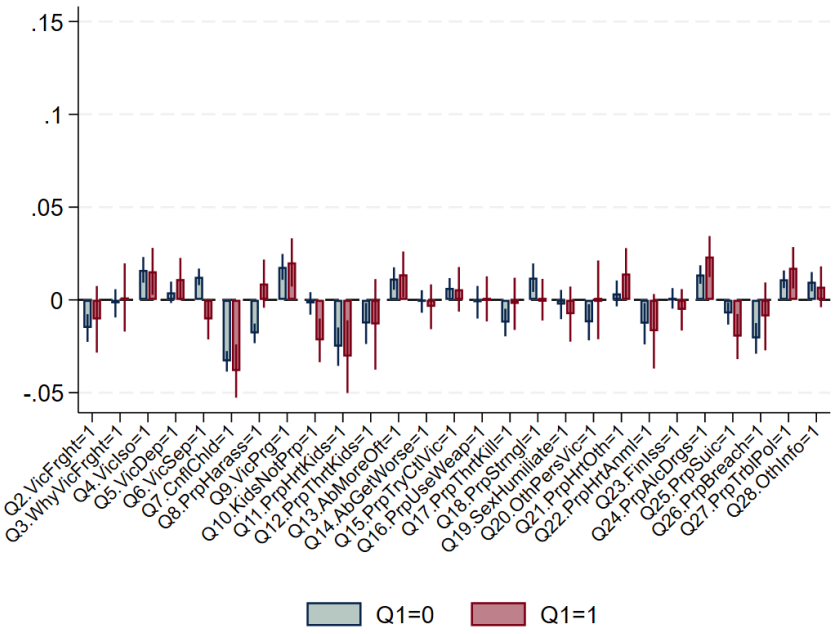
4. Online Appendix

Figure A1: DASH coefficients from beliefs regressions, by victim injury at current incident (Question 1)



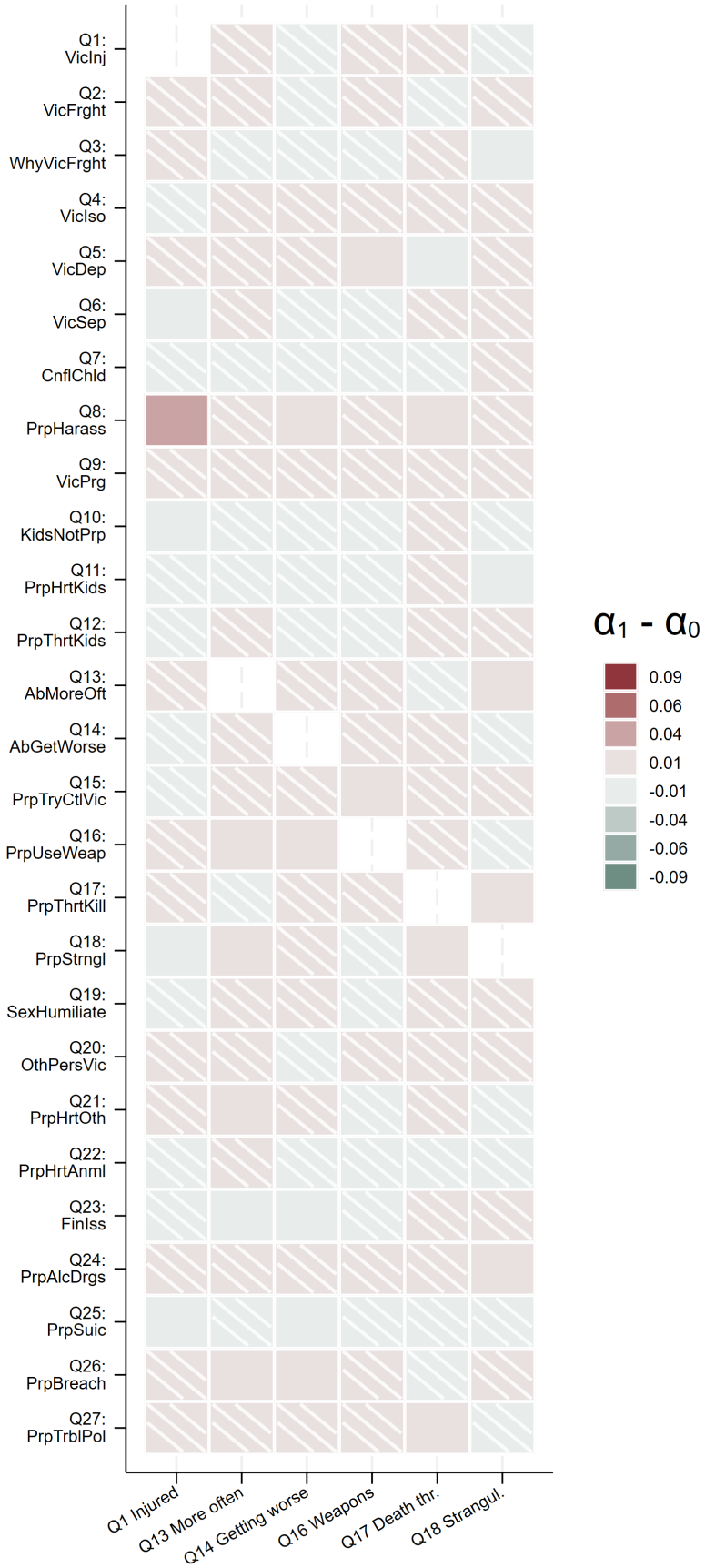
Notes: Coefficients from beliefs regressions on DASH features and other controls, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure A2: DASH coefficients from latent outcome regressions, by victim injury at current incident (Question 1)



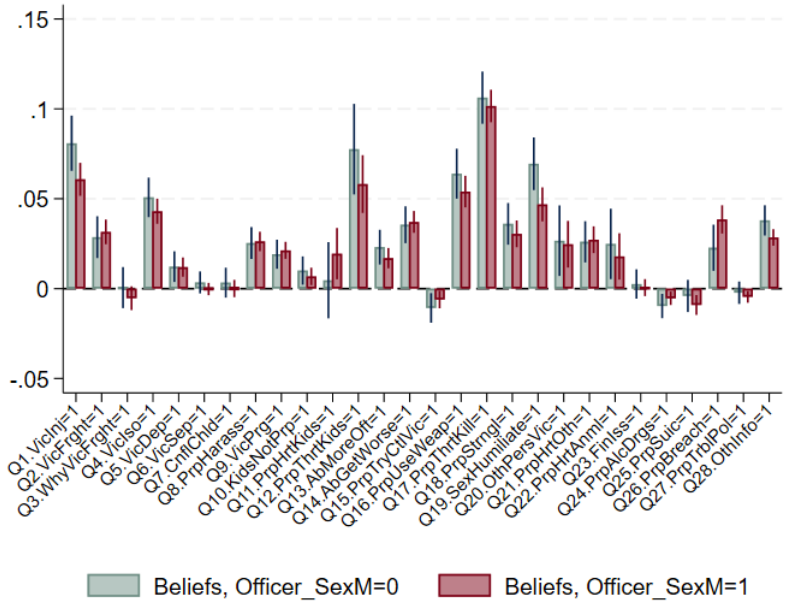
Notes: Coefficients from recidivism regressions on DASH features and other controls, by Question 1, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure A3: Heatmap showing contrast in latent outcome coefficients, by trigger variable



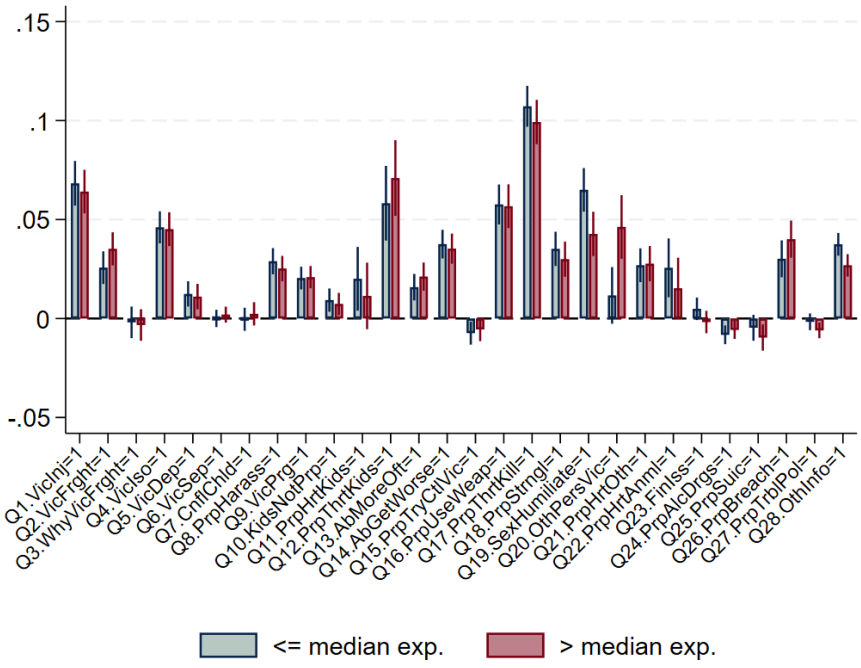
Note: Each cell reflects the magnitude of the contrast between the coefficient of the row variable from two recidivism regressions: one in which the trigger (column) variable equals one and one in which it equals zero. Solid cells indicate that the contrast was significant at the 10 percent level, cross-hatched cells indicate that it was insignificant.

Figure A4: DASH coefficients from beliefs regressions, by officer sex



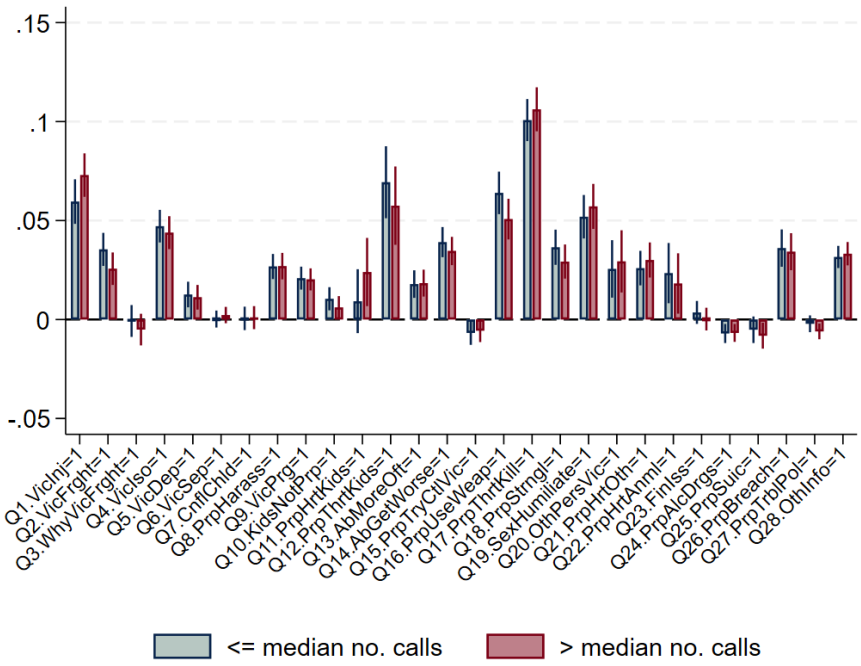
Notes: Coefficients from regressions of beliefs on DASH features and other controls, by officer sex, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure A5: DASH coefficients from beliefs regressions, by officer experience



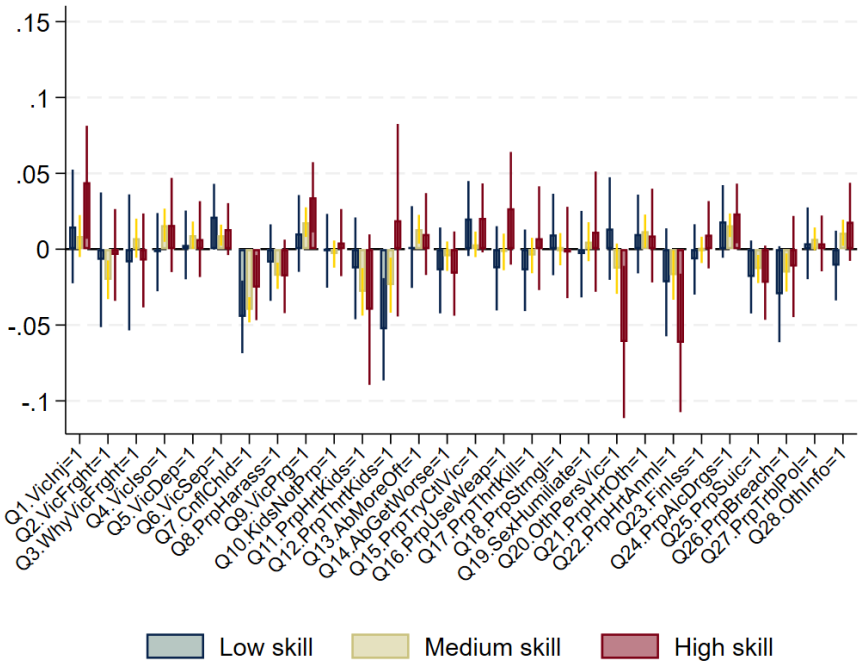
Notes: Coefficients from regressions of beliefs on DASH features and other controls, by officer experience, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure A6: DASH coefficients from beliefs regressions, by number of officer DA cases



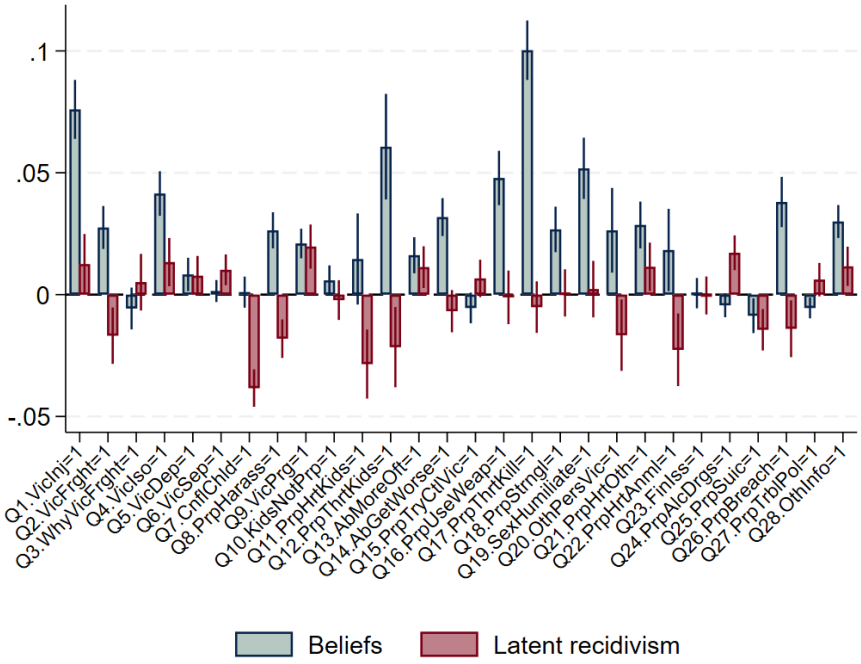
Notes: Coefficients from regressions of beliefs on DASH features and other controls, by officer experience, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure A7: DASH coefficients from latent outcome regressions, by officer skill, officers with at least 100 incidents



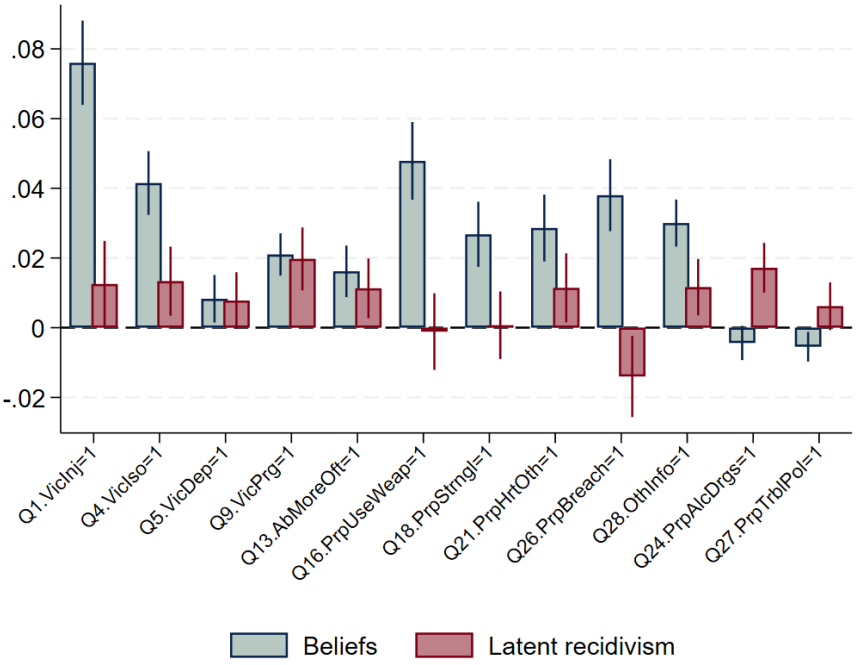
Notes: Coefficients from regressions of latent recidivism on DASH features and other controls, by officer skill, showing only coefficients from DASH features. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure A8: DASH coefficients from beliefs and latent outcome regressions, excluding incidents where low-skill officers checked Question 28



Notes: Coefficients from regressions of beliefs and recidivism on DASH features and other controls, showing only coefficients from DASH features. Incidents from low-skill officers who checked the Question 28 box are excluded from the sample. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure A9: DASH coefficients for representative features from beliefs and latent outcome regressions, excluding incidents where low-skill officers checked Question 28



Notes: Coefficients from regressions of beliefs and recidivism on DASH features and other controls, showing only coefficients from representative DASH features. Sample excludes incidents where low-skill officers checked Question 28. Vertical lines through ends of bars show 95 percent confidence intervals, based on standard errors clustered by dyad.

Figure A10: Share of officers predicting high risk, by trigger features, excluding incidents where low-skill officers checked Question 28

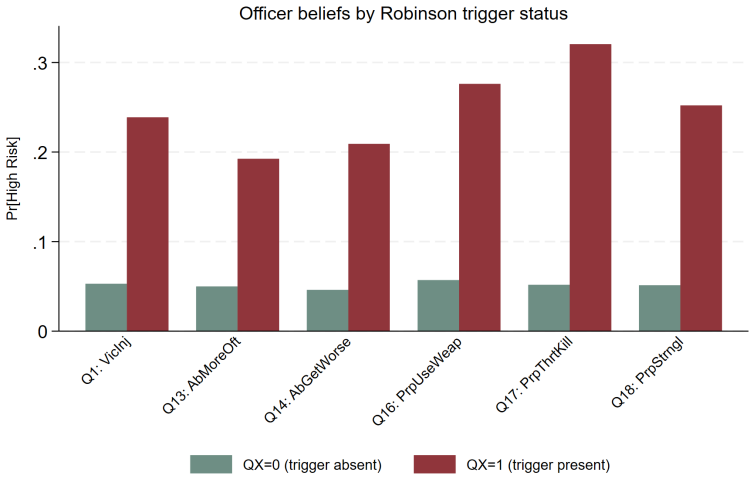
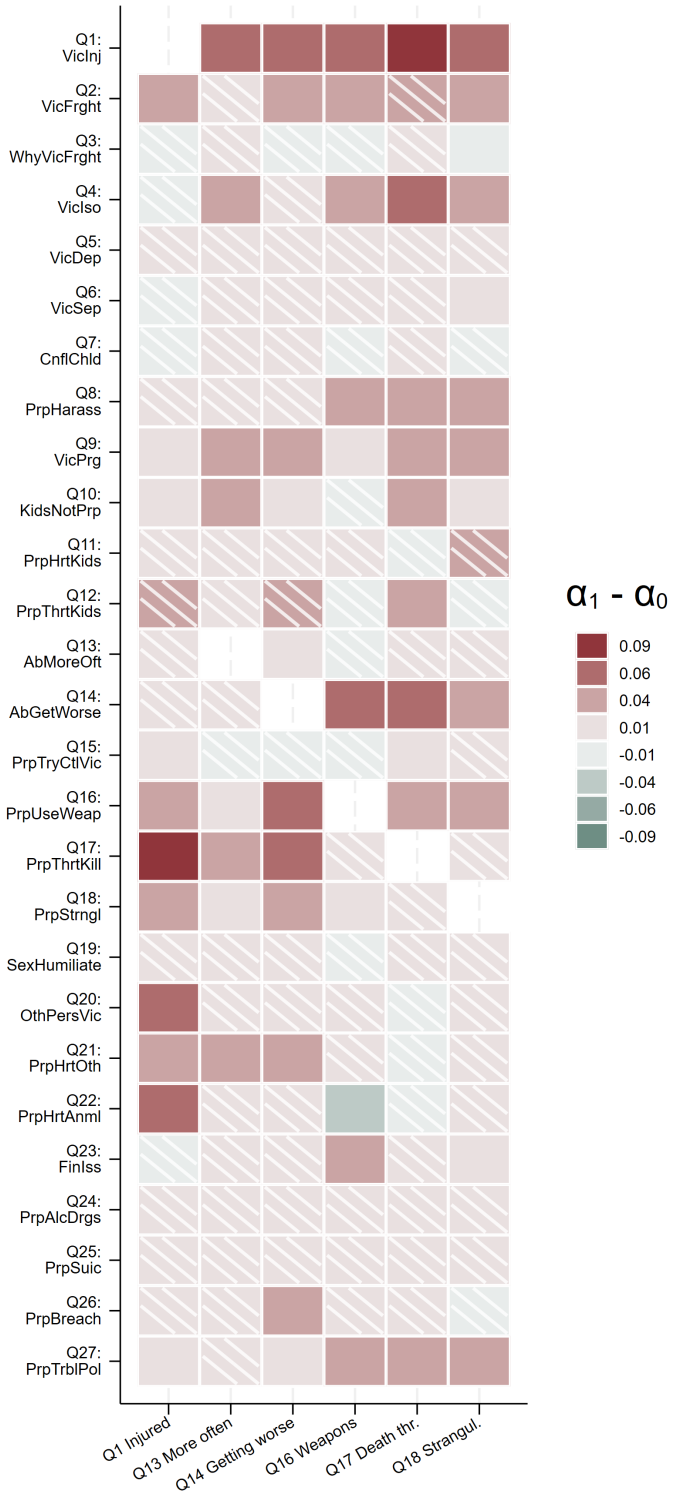


Figure A11: Heatmap showing contrast in beliefs coefficients, by trigger variable, excluding incidents where low-skill officers checked Question 28



Note: Each cell reflects the magnitude of the contrast between the coefficient of the row variable from two beliefs regressions: one in which the trigger (column) variable equals one and one in which it equals zero. Solid cells indicate that the contrast was significant at the 10 percent level, cross-hatched cells indicate that it was insignificant.

Table A1: ATT Estimates: Serious Recidivism and Placebo Tests

	Serious recidivism (full sample)	Serious recidivism (order≥2, placebo PS)	Lagged serious recidivism (order≥2, placebo outcome)
HR only	0.001	0.000	0.000
	(0.005)	(0.007)	(0.001)
	[N=136,137]	[N=73,604]	[N=73,604]
Charge only	-0.044	-0.058	-0.000
	(0.003)	(0.005)	(0.001)
	[N=140,273]	[N=74,867]	[N=74,867]
HR + Charge	-0.080	-0.098	0.002
	(0.006)	(0.008)	(0.002)
	[N=133,840]	[N=71,877]	[N=71,877]

Standard errors (in parentheses) clustered by id_dyad. Propensity scores estimated by grf.

Col. (2): same sample as Col. (3), DFViolence1yr outcome, propensity scores from Col. (3) model.

Col. (3): placebo outcome = any serious DV in prior 12 months; restricted to repeat calls (order≥2).

Table A2: 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	0.012	0.016	0.057	0.014	0.065	0.021
	(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	0.029	-0.009
	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)
Q3.WhyVicFrght=1	-0.003	-0.003	-0.003	0.000	-0.002	-0.003	-0.003
	(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	0.006	0.006	0.013	0.006	0.007	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	-0.034	-0.034	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	0.027	-0.016	-0.015	0.023	-0.016	0.022	-0.014
	(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	0.019	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	-0.004	0.009	-0.003
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)
Q11.PrpHrtKids=1	0.016	-0.028	-0.028	0.018	-0.028	0.021	-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	-0.015	0.058	-0.022
	(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
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Q14.AbGetWorse=1	0.037	-0.004	-0.001	0.034	-0.002	0.031	-0.003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Q15.PrpTryCtlVic=1	-0.007	0.007	0.007	-0.001	0.008	-0.004	0.008
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)
Q16.PrpUseWeap=1	0.057	-0.002	-0.000	0.050	-0.000	0.051	0.000
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Q17.PrpThrtKill=1	0.103	-0.012	-0.009	0.092	-0.009	0.090	-0.007
	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.005)	(0.004)
Q18.PrpStrngl=1	0.033	0.006	0.008	0.032	0.007	0.031	0.009
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Q19.SexHumiliate=1	0.054	-0.003	-0.003	0.050	-0.004	0.055	-0.005
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)
Q20.OthPersVic=1	0.027	-0.008	-0.009	0.026	-0.007	0.016	-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	0.003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Q22.PrpHrtAnml=1	0.021	-0.015	-0.014	0.024	-0.015	0.021	-0.013
	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)
Q23.FinIss=1	0.002	-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	-0.005	0.015
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
Q25.PrpSuic=1	-0.007	-0.011	-0.010	-0.002	-0.009	-0.006	-0.008
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Q26.PrpBreach=1	0.035	-0.028	-0.023	0.034	-0.023	0.034	-0.021
	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
Q27.PrpTrblPol=1	-0.004	0.009	0.010	-0.004	0.011	-0.004	0.011
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
Q28.OthInfo=1	0.032	0.009	0.009	0.034	0.011		
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)		
Observations	153934	153934	153934	153934	153934	103502	103502

Standard errors in parentheses
Notes: Standard errors, clustered by dyad, in parentheses.

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 A3: 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	-0.042
	(0.008)	(0.005)	(0.005)
Q12.PrpThrtKids=1	0.262	-0.031	-0.020
	(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	0.016
	(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	0.036
	(0.005)	(0.004)	(0.004)
Q17.PrpThrtKill=1	0.298	-0.002	0.012
	(0.004)	(0.003)	(0.003)
Q18.PrpStrngl=1	0.230	0.023	0.034
	(0.004)	(0.003)	(0.003)
Q19.SexHumiliate=1	0.224	-0.003	0.005
	(0.005)	(0.004)	(0.004)
Q20.OthPersVic=1	0.200	-0.009	-0.003
	(0.007)	(0.005)	(0.005)
Q21.PrpHrtOth=1	0.196	0.015	0.025
	(0.004)	(0.003)	(0.003)
Q22.PrpHrtAnml=1	0.205	-0.011	-0.001
	(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)
Observations			

Standard errors in parentheses
Notes: Standard errors, clustered by dyad, in parentheses.

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.