Analysis Plan

Project Name: Measuring the impact of the Appropriate Response Initiative on arrest rates in Ramsey County, MN

Project Code: 2308

Date Finalized: 11/12/2024

Date Revised: 2/25/2025

Note: This is an updated version of a prior Analysis Plan which was posted on 11/12/24. The original Analysis Plan is available upon request by emailing oes@gsa.gov.

Project description

The U.S. Department of the Treasury ("Treasury") is administering the State and Local Fiscal Recovery Funds program (SLFRF). SLFRF funds provide substantial flexibility for each recipient to meet local needs within several eligible use categories described in the 2022 Final Rule and the 2023 Interim Final Rule, and many communities are using SLFRF funds to invest in approaches to violence prevention. The Office of Evaluation Sciences (OES) at the U.S. General Services Administration is partnering with Treasury and Ramsey County, Minnesota to understand the impact of Ramsey County's SLFRF-funded Appropriate Responses Initiative (ARI) on arrest rates. Under ARI, emergency dispatchers will be able to send, when appropriate, mental and public health services and community responders instead of, or in addition to, traditional responses when responding to 911 calls. Through ARI, Ramsey County hopes to decrease the number of overall arrests in the County.

OES will evaluate the impact of dispatching ARI responses on arrests using a quasi-experimental design. The evaluation design leverages call-taker variation in the likelihood of sending traditional versus ARI responses for observationally equivalent 911 calls. This primary study will use data through the end of 2024. As exploratory analyses, we list additional approaches either as robustness checks to the main analysis or to be used in case the main analysis is not likely to capture all possible outcomes.

Figure 1a. Traditional response model

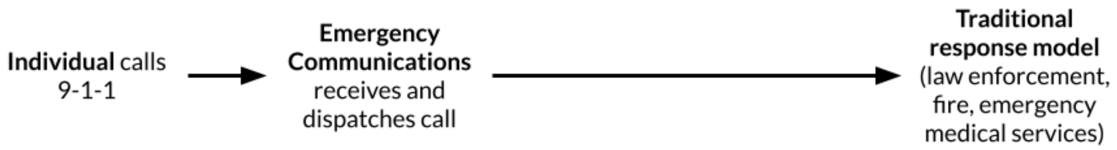
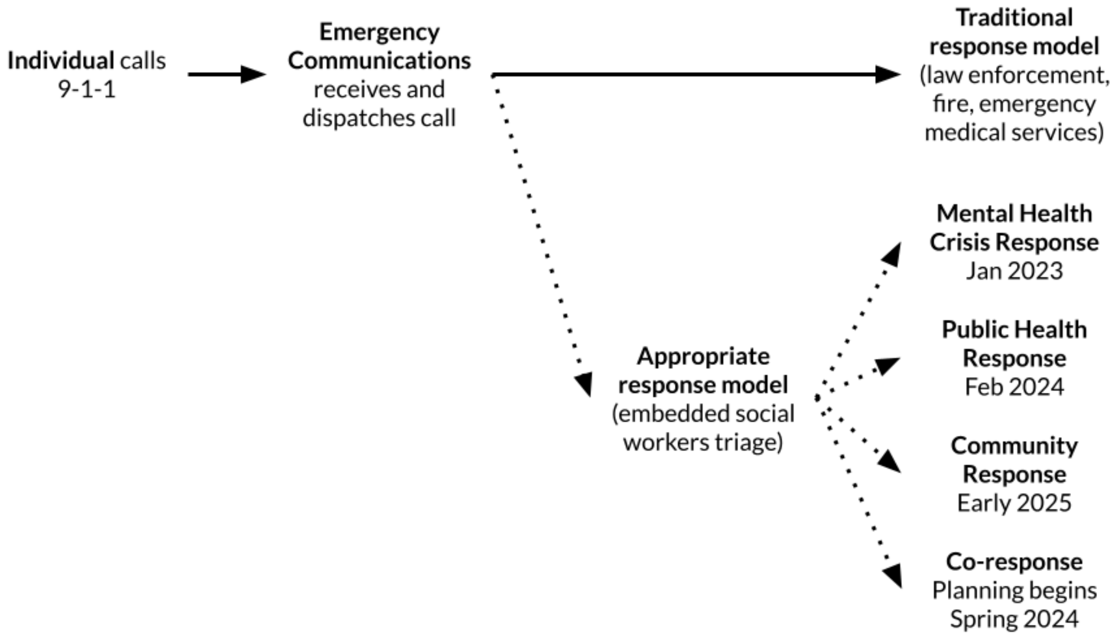

Individual calls 9-1-1
Emergency Communications receives and dispatches call
Traditional response model (law enforcement, fire, emergency medical services)

Figure 1b. Appropriate response model


Individual calls 9-1-1
Emergency Communications receives and dispatches call
Traditional response model (law enforcement, fire, emergency medical services)
Appropriate response model (embedded social workers triage)
Mental Health Crisis Response Jan 2023
Public Health Response Feb 2024
Community Response Early 2025
Co-response Planning begins Spring 2024

Preregistration details

This Analysis Plan will be posted on the OES website at oes.gsa.gov before outcome data are analyzed. In addition, this project will be preregistered in the Open Science Framework Registry at https://osf.io/registries/.

To increase credibility of our estimates, before running any analyses OES separated the project team into two sub-teams: one working on the outcomes data (from Bookings Records) and the other working on 911 call data, with no research team members able to access both through the OES shared drive. We will maintain this separation until this Analysis Plan is posted. This separation is not absolute: from January until April 2024, all team members had access to 911 Call records for 2023 as well as downloaded pdfs of bookings records. The bookings data is also publicly available in pdf form here. However, the bookings pdf data was not transformed into machine-readable data tables until after the folders and teams were split by an independent (within OES but outside of the research team for this project) “data steward” on March 21, 2024. Once the Analysis Plan is posted, we will 1) obtain National Incident-Based Recruiting System (NIBRS) data, and 2) make the bookings data and 911 call records accessible to all team members in order to link them and begin analysis.

Hypotheses

Our primary hypothesis is that 911 calls routed to ARI responses (including behavioral/mental health, public health, community, and co-responses) through the ARI result in fewer arrests than similar calls routed to traditional responses.

Data and data structure

This section describes variables that will be analyzed, as well as changes that will be made to the raw data with respect to data structure and variables.

Data source(s):

The data used for this study will come from three sources: 1) The Computer Aided Dispatch (CAD) data from the Ramsey County Emergency Communications Center (ECC), which covers details of 911 calls coming in and how they are routed, 2) Publicly available arrest records of people booked into the Ramsey County Adult Detention Center reported by the Ramsey County Sheriff's Office (Bookings), and 3) Minnesota Bureau of Criminal Apprehension's (BCA) reporting to the National Incident-Based Reporting System (NIBRS). We outline these three below:

1. CAD

This data covers details of incoming 911 calls, descriptions of the issue made by call-takers, how the calls were routed, and personnel involved at each step. From this data, we will use:

- Case Number (which is generated when a police dispatch unit is assigned to an incident)
- Call-taker ID (the individual who initially answers the call)
- Date and time of the call
- Location of the call: longitude/latitude, automatically generated through phone system
- Information about the reason for the call:
    - Incident type
    - Problem
    - Priority description
- Response type sent
    - Law enforcement
    - Fire/Emergency Medical
    - ARI responses

2. NIBRS

This data is reported to NIBRS by BCA in Minnesota. NIBRS includes information on incidents, offenses, and arrests. In NIBRS, an "incident" is defined as one or more offenses committed by the same offender, or group of offenders acting in concert, at the same time and place. NIBRS distinguishes between two types of offenses: Category A offenses (e.g., burglary, assault, sex offenses) and Category B offenses, which are lesser offenses (e.g., DUI, disorderly conduct, liquor law offenses, non-violent family conduct). Category B offenses do not always result in an arrest, and even when they do, we anticipate that more details of the arrest will be missing in the NIBRS data, including location. We do not yet know how often this information will be missing for either type.

We expect the data to include the following information for all age groups:

- For incidents
  - Name
  - Date of birth
  - Incident date
  - Incident time
  - Incident location: Address (which will be geocoded into latitude and longitude using the Google Maps, geocodio, and / or OpenStreetMap APIs in R)
  - Offense type
- For arrests
  - Date
  - Arrestee age
  - All of the above information on the incident
  - Case number/incident number (which can link to Bookings data)

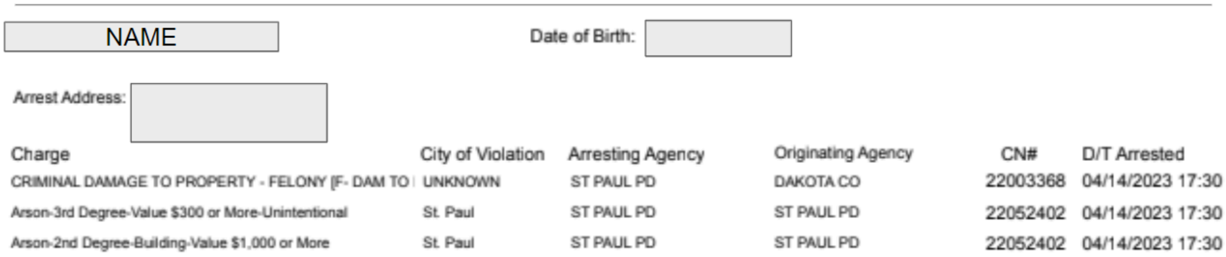
NIBRS has the benefit of including data on all age groups as well as the location of incidents.

3. Bookings

The arrest records are downloaded from the Ramsey County Sheriff's Office website and display people who have been booked into the Adult Detention Center within the timeframe indicated in the individual reports. These include:

- Name and age of arrestee
- Date/time of arrest
- Arrest location: Address (which will be geocoded into latitude and longitude using the Google Maps, geocodio, and / or OpenStreetMap APIs in R)
- The arresting agency and the originating agency
- The charge(s) (e.g., theft)
- Case number (which can link to NIBRS data and to CAD data). This number is generated in the ECC when a police dispatch unit is assigned.

This information does not include anybody under the age of 18. Below is an example of a single record:



NAME

Date of Birth:

Arrest Address:

Charge	City of Violation	Arresting Agency	Originating Agency	CN#	D/T Arrested
CRIMINAL DAMAGE TO PROPERTY - FELONY [F- DAM TO	UNKNOWN	ST PAUL PD	DAKOTA CO	22003368	04/14/2023 17:30
Arson-3rd Degree-Value $300 or More-Unintentional	St. Paul	ST PAUL PD	ST PAUL PD	22052402	04/14/2023 17:30
Arson-2nd Degree-Building-Value $1,000 or More	St. Paul	ST PAUL PD	ST PAUL PD	22052402	04/14/2023 17:30

We expect that neither the NIBRS nor the bookings data will contain a perfect and complete set of geolocalizable arrest records. Bookings data does not contain information on the different incidents that led to an arrest or any information for juveniles. Bookings data also only include arrest location and not incident location. NIBRS only includes incident location, and this field may be missing or incomplete for some fraction of the data. When merging with the CAD data, incident location is more likely to be geographically closer to the address where the CAD calls originated than the arrest location.

Given that neither the Bookings nor the NIBRS data set is complete, we plan to use both arrest data listed in NIBRS and Bookings data to identify arrests that occurred (and refer to the combination as “Arrests data”). However, we will de-duplicate the data by taking out arrests that occur in both Bookings and NIBRS. We can easily identify and remove these duplicate observations since both Bookings and NIBRS include a case number field.

Outcomes to be analyzed:

There will be two primary outcome measures, each reflecting potentially different effects of calls sent to ARI:

1. Whether there was an arrest that was linked - via Case Number - from an incoming call to an arrest. This will be based on the Arrests data (combining both the Bookings data and the NIBRS data).

Whether there was an arrest within 0.1 miles and in the 3 hours following an incoming call, given that call’s latitude, longitude, and timing.[1] This will be based on the Arrests data (combining both the Bookings data and the NIBRS data).

The first measure, based on *linked* arrests, is clean and will reflect any change in arrests from law enforcement dispatched for that call. However, it may miss some arrests that are not linked to the call and subsequent dispatching. This could happen if - following an ARI response - another call is placed about the same concern, and then law enforcement is dispatched, and the calls are not linked in the ECC.[2] This could also happen if there are spillover arrests due to an officer being dispatched. Thus, the linked arrests may over or understate the true effect of the policy.

[1] We have chosen this radius and timing through conversations with those familiar with law enforcement data and the time from dispatching to associated arrests. We believe that this approximates the distance in which an officer would observe behavior, and the upper end of the time period before an officer would be dispatched again. We will also present the same results for alternative distances and time lags (0.2 and 0.5 miles, 1.5 and 6 hours).

[2] The ECC attempts to link calls about the same incident.

The second measure, based on arrests within the vicinity of an incoming call will include these *linked* arrests (as long as they happen within the given radius and time lag) as well as:

- arrests for prior offenses made because an officer dispatched for a call finds and arrests someone with an outstanding arrest (known as warrant arrests)
- other spillover arrests in the vicinity because an officer has been dispatched to the area
- unrelated arrests (also known as false positives or type 1 errors).

Together, these two outcome measures will paint a more complete picture of the effects of ARI dispatching on arrests.

Transformations of variables:

Call data will be coded to include indicators, based on the arrests data, for 1) whether there was an arrest linked to the Case Number associated with a call, and 2) whether there was an arrest within the 0.1-mile circular radius of and 3 hours following a 911 call.

Call dispositions refer to the event that concluded a call. Many dispositions refer to the results of sending traditional responses (e.g., "X - Fire Incident Responded," "TP - Transported" — taken by ambulance—, and "G - Gone on arrival"). To comply with HIPAA regulations, details of mental and public health responses are not included in the CAD data that are shared with OES.

Calls dispatched to ARI responses will be identified by having, at any point during the call, a call disposition of:

1. "MH - Mental Health" (routed to Mental Health/Crisis),
2. "CR - XX" (routed to Public Health), or
3. Call Disposition including "ARI-XX" (routed through the ESW).

In addition, we will include other call dispositions or other call characteristics that are chosen by the ECC in the future to indicate an ARI response, including community responders and co-responses.

Data exclusion:

For our primary analysis, we will include only calls that could plausibly have been routed to ARI responses based on the characteristics of the call. We will therefore exclude the following call types:

- Calls which could not result in an ARI response (e.g., a fire or a car accident) and calls for which no response is needed or expected (e.g., pocket dials and test calls). We will include details of how this is done in a technical appendix.
- Calls routed to call-takers who take fewer than 100 calls throughout the study period. The identification strategy relies on estimates of call-taker propensities to route calls differently, which requires enough observations.
- Calls that were not answered at call-takers consoles. Calls that were not answered at call-taking consoles come directly from police to dispatchers, and not from individuals calling 911.

For arrests outcomes, we will include those for which we have information about the location from either the bookings or NIBRS data. We will exclude arrests made at jails with the expectation that these either are unrelated to 911 calls or do not reflect the location where the incident occurred.

Treatment of missing data:

Our sample is all 911 calls in the CAD system, subject to the exclusions listed above. Our outcome is an arrest recorded in the arrest data that is linked via Case Number or that occurred within 0.1 miles and 3 hours of each call. However, we note that both measures will be measured with error. The first, based on a case number link, will miss any arrests that are not linked to the call and subsequent dispatching. This could happen if - following an ARI response - another call is placed about the same concern, and then law enforcement is dispatched. The second outcome measure, based on proximity, will also be measured with error. There can be some arrests that happen within the vicinity of a call and have nothing to do with the call itself (false positives). For example, another officer within 0.1 miles can make an arrest for an unrelated incident. Arrests may also result from a 911 call but not in the time or geographic vicinity of the call (false negatives). For example, an arrest could be made more than 0.1 miles from a call that was made or more than 0.1 miles away from a reported incident. We expect that such false positives and false negatives will be unrelated to the variation we rely on for causal identification, and so we do not plan to address them explicitly as, under these conditions, they should not change the inferences we make based on the estimates. If these errors are common, which we do not expect, the resulting measurement error could bias our estimates toward zero.[3]

Descriptive statistics, tables, and graphs

We will present descriptive statistics of the 911 calls, including:

- Number of calls routed to fire or emergency medical services (FEMS), law, and ARI responses in the full sample
- Number of calls routed to ARI responses in each month
- Number of call-takers, calls per call-taker, distribution of propensities to dispatch to ARI responses

Statistical models and hypothesis tests

This section describes the statistical models and hypothesis tests that will make up the analysis — including any follow-ups on effects in the main statistical model and any exploratory analyses that can be anticipated prior to analysis.

Statistical models:

Our identification strategy uses variation in the propensity of 911 call-takers (telecommunicators) to dispatch calls to ARI responses in order to identify the causal effect of ARI responses (vs.

[3] If these errors appear to be more common, we may be able to estimate the bias using the method outlined in Meyer and Mittag (2017).

traditional responses) on arrests. Conditional on the time, date, and city of origin of the call, the assignment of calls to call-takers should be uncorrelated with any other observable or unobservable call characteristics. This plausible randomness comes from an automated call assignment system that routes each incoming call to the 911 call-taker who is available and has been off the phone for the longest. No information about the call is used in making this assignment.[4]

The question we want to answer is whether sending an ARI response, instead of a traditional response, changes the likelihood of an arrest in the area of and time following a 911 call.

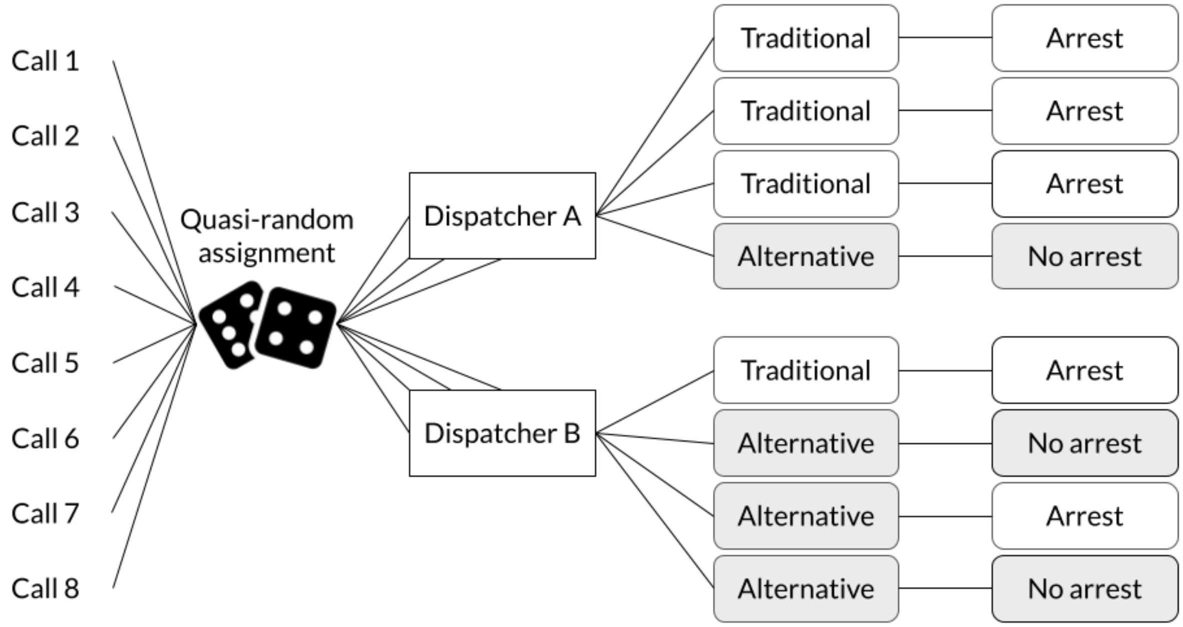
This is challenging to estimate, because 911 calls that lead to law enforcement being dispatched are likely very different from 911 calls that lead to ARI responses being dispatched. For example, law enforcement are explicitly more likely to be sent if a weapon is present. Therefore, outcomes of calls going to law enforcement and outcomes of calls going to ARI responses are likely to be very different from one another, regardless of who is sent. So comparing outcomes across different call types will lead to a mis-estimation of the effect of sending ARI responses, because the difference will include both this effect in which we are interested and the differences in outcomes due to differences in underlying characteristics of the calls. Some of these differences that determine how calls are routed can be observed and accounted for in the data (e.g., if the call-taker writes a note that a weapon is present), but many cannot.

Therefore, we will use only the variation in the likelihood of an ARI response being sent that derives from the variation in the propensity of different 911 call-takers to dispatch ARI responses. Following an instrumental variables methodology, we will estimate the relationship between the likelihood of an arrest and the dispatching of an ARI response, using only the dispatching variation due to the call-taker, and not to features of the call. We estimate "dispatching variation due to the call-taker" using the predictions from a regression that explains the likelihood of dispatching an ARI response as a function of the call-taker who happened to get the call, controlling for measurable features of the call.

The intuition for this strategy is demonstrated in Figure 2, in which two hypothetical dispatchers have a different propensity to send traditional and ARI responses, possibly resulting in differences in the likelihood of arrests for calls routed to each dispatcher.

[4] OES team members visited Ramsey County in December 2023 and toured the Emergency Communications Center to see how call assignment and routing worked and discuss background descriptions of the center and systems with call-takers and other personnel.

Figure 2. Stylized example of how incoming calls assigned to dispatchers influence outcomes



Following the methodology in Chyn, Frandsen, and Leslie (2024), we will conduct a residualized (or covariate-adjusted) Jackknife Instrumental Variables Estimation (JIVE), which is similar to what is referred to as a "leave-one-out mean." We use the residualized Unbiased JIVE (UJIVE) (Kolesár, 2013), because our random assignment assumption depends on conditioning on call characteristics (time of day, day of week, month, city of origin). We use JIVE instead of traditional two-stage least squares (2SLS) because of the many instruments problem (Angrist, et al., 1999; Kolesár, 2013) prevalent in leniency-designs such as ours, as JIVE has been shown to have better bias and coverage properties with multiple instruments even in the presence of heteroskedasticity (Poi, 2006) compared to 2SLS.[5] We do not implement clustered-JIVE because we do not have assignment of calls to call-takers in batches or clusters.

To construct the first-stage instrument, we proceed in two steps. First, we regress a binary indicator for whether an ARI response was sent on a set of call-taker fixed effects, controlling for a set of control variables (hour of the day, day of the week, month, and originating city), leaving out one call at a time. This gives us a predicted value for the omitted call from these regressions. Next, we partial out the same covariates (time of day, day of week, month, city of origin) from the resulting predicted values, again using a set of jack-knifed regressions (in which one call is left out at a time). This provides the instrument, representing the call-taker-specific difference in propensity to route to an ARI response.

This will also allow us to see the magnitude of the effect of a call taken by a call-taker with a higher propensity to route to ARI responses on the likelihood an individual call is routed to an ARI response.

[5] Note that in our context, the risk of bias from multiple instruments is probably low, as we have many calls per call-taker and so the influence of any one call on the call-taker propensity is likely small.

In the second stage, we regress the outcome – an indicator for a linked arrest, and an indicator for an arrest within 0.1 miles and 3 hours of a call – on the predicted propensity of an ARI response being sent, as constructed above.

Our coefficient of interest will be the coefficient on the predicted propensity to send an ARI response. This number reflects the percentage point difference in the likelihood of arrest between calls routed to ARI responses and the same type of calls (based on included covariates) routed to traditional responses. If it is positive and statistically significantly different from zero, we will conclude that dispatching ARI responses increases arrest rates relative to similar calls dispatched to traditional responses. If it is negative and significant, we will conclude that dispatching ARI responses reduces arrest rates for similar calls. If it is not statistically significantly different from zero, we will not be able to reject the null hypothesis of no relationship.

For this identification strategy to be credible, we rely on the following standard assumptions:

1. *Conditional random assignment*: we assume that, within a given hour of the day, day of the week, month, and city from which the call originated, which call-taker receives a call is uncorrelated with the potential arrest outcomes for that call. Calls are assigned to call-taker through an automated system that directs a call to the call-taker at a console that has been off the phone for the longest. Different types of calls are not routed to different call-takers. The set of control variables included in our instrumenting regression should take account of the correlations between who receives the call and the likelihood that the call results in arrest that we *do* think exist, primarily because of the correlation of these variables by *time*. For example, while there may be differences in who works more overnight shifts (e.g., if some call-takers take more night shifts) and the types of calls that come in during different times of the day, we condition all estimation on the time of day. Thus we believe this assumption is valid. Using call location characteristics (e.g., pre-ARI arrests counts) and descriptions of incoming calls that we believe will not be determined by the call-taker (e.g., whether a fire is reported), we can conduct covariate balance-type tests by regressing call location factors (factors excluded from the residualization procedure described above) against call-taker propensities. The absence of any relationship between call-taker propensities and call characteristics will lend credibility to this conditional ignorability assumption.

2. *Exclusion restriction*: we assume that the only way that a call-taker alters the outcomes of interest in this study is through how a call is routed. This would be violated if, for example, call-takers who were more likely to send ARI responses were also somehow able to alter the likelihood of subsequent arrests. We believe this is unlikely. In our visit to the ECC in Ramsey County, this was a primary question of ours asked of many call-takers and other ECC staff. We were repeatedly reassured to hear that there simply did not exist channels through which call-takers could influence outcomes after routing a call. In addition, call-takers are quite busy, so are encouraged to quickly move on to the next call after one has been routed.

3. *Relevance*: in order to estimate how sending traditional vs. ARI responses affects arrests based on the call-takers' differing propensities to send these different responses, it must

be the case that there exists variation across call-takers in their propensity to route calls to (non-)traditional responses. Based on conversations with the ECC, we believe this to be the case. For example, on our tour of the ECC, we heard that a call-taker sitting closer to the Embedded Social Worker may be more likely to route calls to that person, simply because their presence and that option is more salient. We have structured our data so that we can estimate call-taker propensities before linking to outcomes (see above), which will let us determine whether the first stage is strong enough before proceeding. To clarify, the first stage strength is based on whether the call-taker indicators explain a statistically significant degree of variation in the responses sent to calls, independently of the different control variables we include. In contrast to traditional tests of this assumption which use first stage tests based on F-statistics, Angrist and Kolesár (2024) recommend proceeding if the association between predicted propensities and observed treatment (i.e., sending an ARI response) is greater than zero.

4. *Monotonicity*: Our estimates could be biased if call-takers with a high propensity to route some types of calls to ARI responses are less likely than other call-takers to route other types of calls to ARI responses. We can use the Stata command, *testjfe* (Frandsen, Lefgren, and Leslie, 2023) to test this. *testife* tests for violations of pairwise monotonicity. Pairwise monotonicity (the assumption that any call-taker who is more likely to route to ARI responders would be more likely to do so in all calls compared with any call-taker who is less likely to do so) is a strong assumption and likely violated in practice. Average monotonicity requires only that call-taker propensity is positively correlated with call-taker routing.[6] Any violations (which, based on the previous literature, are likely to occur), will be addressed as recommended by Frandsen, et al. (2023) and Chyn, Frandsen, and Leslie (2024). A test of average monotonicity is fairly straightforward to implement; it is the same as the first-stage test of relevance conducted on pre-intervention subsamples. For example, using pre-ARI location-specific arrest rates – the same covariate we would use to test for conditional random assignment to call-taker propensities – we can test whether the positive association between predicted call-taker propensity and sending an ARI response is positive. If, on average, across subsamples the association remains positive, then the assumption of average monotonicity is more credible.

As mentioned earlier and emphasized by Chyn, Frandsen, and Leslie (2024), the IV estimator in the context of leniency designs needs to be especially considerate of the following:

1. Within-cluster correlations: Similar studies sometimes need to adjust estimated standard errors (via clustering) when cases (calls in our case) are assigned in batches to the same decision-maker (call in our case). In this study, each call is separately assigned to a call-taker. Therefore, this is not a threat to our estimates for which we will need to adjust. Using a randomized study as an analogy, the call-takers are blocks, not clusters.

2. Too many instruments: Many studies using a similar model face the issue of too many decision-makers (call-takers in our case) relative to the number of decisions (calls in our

[6] Average monotonicity limits interpretation of the IV estimand. It is no longer interpretable as a marginal treatment effect, but still estimates a weighted average effect across all call-takers.

case). With more than 1000 calls per day and just over 100 call-takers, this is much less of a concern in our case. Still, we follow the norms in this literature and use the Jackknife Instrumental Variables Estimation strategy, which addresses this concern.

3. An additional threat not mentioned in Chyn, Frandsen, and Leslie (2024) is the likelihood of heterogeneous effects across subgroups for which fixed effects are included. When this is the case, the weights in a standard regression can mean that the estimate of the overall effect is biased. To address this issue, we conducted a simulation study in R. The simulation was constructed so that the alternative response is much more effective at reducing arrest rates in the night than in the day, but also so that “high-propensity call-takers” (those most likely to send alternative responses) are much more likely to work the day shift. These conditions are highly likely to generate bias due to the regression weights introduced by the fixed effect (see, for example, Aronow & Samii, 2016). However, even under these relatively extreme conditions, the bias was negligible in comparison to even moderate local average treatment effect sizes, and none of the approaches set up to deal with this issue (such as inverse propensity weighting) produced substantial improvements in mean squared error. Therefore, we take no steps to address this threat directly beyond what is described above.

Robustness analysis:

To confirm that our estimates are not sensitive to specific choices of thresholds, we will repeat the analysis described above, but with the following variations:

1. Alter the radius around a call to 0.2 and 0.5 miles,
2. Alter the time lapsed from a call to 1.5 hours and to 6 hours,
3. Replace the outcome with a continuous measure of the number of arrests. To do so, we will remove duplicates between the NIBRS and bookings data using Case IDs.

Exploratory analysis:

Alternative identification strategies:

Below we describe three alternative identification strategies that we may use if, upon implementing the primary strategy described above, we have concerns about its validity due to unforeseen complications or challenges. We may also use these strategies to corroborate our main findings and/or to ask additional research questions, notably estimating effects of ARI on outcomes for different types of calls.

Each of these identification strategies requires a different set of assumptions and answers a slightly different question. The strategies also rely on variation in ARI responses being sent, and thus they cannot reflect – except through large assumptions – the effect of ARI on a larger set of call types.

Alternative strategy 1. Capacity constraints as an *instrumental variable*

The first alternative identification strategy relies on capacity constraints in staffing of the ARI responders. Controlling for a set of fixed effects, we will use idiosyncratic variation in ARI call volume that removes the 911 call-takers' option to quickly route a call to ARI responders, so that calls are defaulted to a traditional response (e.g., when the ESW is busy and cannot take the next call). This strategy will provide an estimate of the effects of ARI service availability on any type of call that could be routed to an ARI response but is routed to a traditional response if the ARI responders are unavailable. Among others, this strategy relies on the assumption that – conditional on observable characteristics about an incoming call (including timing, location, and problem type) – the ARI call volume at the time of the call is unrelated to how the call would be routed, except through the capacity constraints on ARI responses.

With limited staff for ARI, sometimes the ESW or other members of the ARI response team are busy or absent, meaning that call types that would have been sent ARI responses are instead sent to the traditional response team. We assume that some of this variation is pseudo-random, conditional on the day and time of calls. Then we can instrument for the use of ARIs using a measure of constrained capacity (e.g., the ESW still being on the last call) with the following:

$$Prob(ARIResponse\ =\ 1)_{ijt} = \boldsymbol{\beta}_1\{Capacity\ Constrained\}_{ijt} + X_{ijt} + \boldsymbol{\varepsilon}_{ijt}$$

Where *i* again indexes a 911 call for phone responder *j* in time *t* and $\boldsymbol{1\{Capacity\ Constrained\}}_{ijt}$ is an indicator for the ESW being still on a call or absent.

We use the predicted variation in police probabilities to estimate:

$$Prob(Arrest\ =\ 1)_{ijt} = \alpha_0 ARI\widehat{Response}_{ijt} + X_{ijt} + \xi_{ijt}$$

Alternative strategy 2. Public health Difference-In-Differences (DID)

The second alternative identification strategy estimates the effect of the availability of a public health response on public-health-relevant calls using a DID design. Ramsey County rolled out a public health response in February 2024. The calls which are eligible to be routed to a public health responder are clearly specified substance use, no threat of violence, no overdose ongoing. Thus, these call types will not be captured by the main estimation strategy.

The clear rules for routing to public health responders suggest that there is minimal variation in routing across call-takers, but a substantial change in which responses are sent following the introduction of the public health response in February 2024. Therefore, we will compare arrest rates for the types of calls that will go to public health response, before and after the rollout of public health response (February 2024). As a comparison group, we will use non-opioid calls, which should not change – in routing or arrests – with the rollout of the public health response.

We will estimate the increase in arrest rates following opioid-related calls, compared to the increase in arrest rates for other types of calls with similar arrest rates.

Alternative strategy 3. Embedded social worker timing Regression Discontinuity (RD)

The third alternative identification strategy will estimate the effect of the ESW using a regression discontinuity in time of day, as the ESW is only on duty from 8am-10pm. This strategy will compare calls that come in just before and just after 8am and 10pm, with the expectation that those that come in outside of the ESW hours are *discontinuously* less likely to receive an ARI response.

This approach relies on the assumption that the potential arrest outcomes of calls, alongside any related covariates we condition on, are continuous at the two temporal cutoffs. This assumption can be inspected using indirect tests, such as checking to see that other characteristics of calls are continuous at the cutoff(s). If these tests lend plausibility to the assumption of continuity, we will estimate the discontinuity in arrests based on ESW presence using a regression discontinuity estimator.

Inference criteria, including any adjustments for multiple comparisons:

We will use p-values from the IV specification of the main effect of ARI and the interaction of ARI with the arrestee being non-white. The null hypothesis is that the average effect of the treatment is zero and the test is two-tailed. We will use an alpha of 0.05 to determine statistical significance.

Limitations:

The primary limitation of this planned study is that the design will not allow us to estimate the effects on all potential results of ARI. In particular, we will not be able to measure escalation or de-escalation of violence, incidents that do not result in an arrest, tickets and fines, long-term costs or benefits from participation in the justice system or receipt of mental health services, or changes in public trust. Such outcomes are either not measured, cannot be linked to specific calls, or are sufficiently rare that statistical inference with any certainty will be impossible. That said, as information on these outcomes becomes available or feasible to connect with calls, we may incorporate additional exploratory analysis.

Citations:

Angrist, Joshua, Guido W. Imbens, and Alan B. Krueger. “Jackknife instrumental variables estimation.” *Journal of Applied Econometrics* 14, no. 1 (1999): 57-67.

Angrist, Joshua, and Michal Kolesár. "One instrument to rule them all: The bias and coverage of just-ID IV." *Journal of Econometrics* 240, no. 2 (2024): 105398.

Aronow, Peter M., and Cyrus Samii. "Does regression produce representative estimates of causal effects?" *American Journal of Political Science* 60.1 (2016): 250-267.

Chyn, Eric, Brigham Frandsen, and Emily C. Leslie. Examiner and Judge Designs in Economics: A Practitioner's Guide. No. w32348. *National Bureau of Economic Research*, 2024.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. "Judging judge fixed effects." *American Economic Review* 113, no. 1 (2023): 253-277.

Kolesár, Michal. Estimation in an instrumental variables model with treatment effect heterogeneity. No. 2013-2. 2013.

Meyer, Bruce D., and Nikolas Mittag. "Misclassification in binary choice models." *Journal of Econometrics* 200, no. 2 (2017): 295-311.

Poi, Brian P. “Jackknife Instrumental Variables Estimation in Stata.” *The Stata Journal* 6.3 (2006): 364-376.