THE UNIVERSITY OF CHICAGO

OPERATIONAL ISSUES IN LARGE JAIL AND JUDICIARY SYSTEMS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
RUSSELL HANNIGAN

CHICAGO, ILLINOIS
JUNE 2022

Copyright © 2022 by Russell Hannigan

All Rights Reserved

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acknowledgments

This Ph.D. was only possible due to the hard work of my incredible wife, Katie, over the last 6 years. She funded our life by climbing a rung-light corporate ladder while carrying me on her shoulders. Katie, you work so hard and care for the people around you so deeply. You are my role model. Thank you for everything, love.

My family has believed in me every step of the way. Mom, Dad, Lilly, Cora, Ben, Tom, Betsy, and Jake, thank you for the vote of confidence. I love you all.

I am fortunate to have an incredible network of friends that have kept me going through this. I can't mention everyone, but I appreciate the help regardless. Marcel Tuijn, Lisa Hillas, Monty Montgomery, Cat Buckley, James Traina, Uyen Tran, Gizem Yilmaz, Deniz Akturk, Nasser Barjesteh, Amir Alwan, Kelsi Morse, Chris Stewart, and more were all there throughout the degree.

It is some combination of Kim and Blair Flicker's fault that I got into this mess. Thank you both for your friendship and your amazing kids. Thank you, Blair, for the fun times researching.

Baris Ata is my advisor and coauthor for the first two chapters of this dissertation. Baris, thank you for the over-the-top training and support. I've never even *heard* of an advisor that devotes so much time and effort to their students. I think I needed that help more than most. The good parts of this dissertation are largely due to your guidance, and the growth I've had throughout this degree is largely thanks to you.

Like most academics, I was the beneficiary of fantastic teaching when I was younger. Thank you Todd Rorem, Chris Neuman, Jim Talley, Mieczyslaw Dabkowski, Victor Worsfold, and Tim Redman for influencing me to be curious men like you.

The Cook County Sheriff's Office has been incredibly generous with their time and resources, which were crucial for this research. Thank you to the many people on staff there who helped us transfer data and understand the criminal justice system. The Rustandy Center at Booth facilitated this relationship. In particular, Salma Nassar's professionalism

and poise allowed this work to happen.

Finally, I was blessed with a fairy godmother who gave me the opportunity to have a world-class education. Thank you to Mrs. McDermott and the Scholars Program for many of the blessings in my life.

ABSTRACT

This dissertation is primarily about improving the criminal justice system, focusing on pretrial detention. Chapter 1 attempts to leverage detainees' (and their attorneys') utility over different detention locations in conjunction with traditional operational improvements to save costs, reduce turnarounds (people whose pretrial detention is longer than their eventual sentence), and improve efficiency. Chapter 2 develops a model to identify turnarounds before they occur so their cases can be intervened upon. Chapter 3 is an experiment which studies people's utility over time, attempting to ascertain what people care about when evaluating potential waits, and how those waits' presentations impact people's utility.

In Chapter 1, we consider excessive pretrial detention, which is often caused by inefficient case processing. Pretrial detention is expensive for both the taxpayer in terms of housing costs and for detainees in terms of perceived costs. In the extreme, detainees can be incarcerated longer pretrial than their sentence requires. Using data from the Cook County Sheriff's Office, we explore the drivers of delays in case processing and policies which can reduce the consequences of excessive pretrial detention. We develop a model of detainee behavior that affects their case lengths, and hence, the duration of their pretrial duration. Taking it to the data obtained from the Cook County Sheriff's Office, we estimate detainees' perceived costs of being detained in jail, prison, and on EM. We find that prison is perceived as the most costly housing location, followed by jail, and then EM. Costly housing locations may induce unnecessary delays in case processing. We consider four counterfactual interventions and study their impact. First, we consider operational improvements to court processes that may lower the number of court visits by the detainees. Removing one court visit from detainees' cases can save the jail over $20 million annually and reduce turnarounds by 10.9%. Second, we consider paying the bonds of detainees with lower level charges. Simple fund-allocation policies can reduce the pretrial jail population by 2% and can save taxpayers four times what is paid toward bail. Third, we consider split sentencing: in which sentences are split between incarceration and supervision or probation. We estimate that the jail could save save $8.5 million and the courts 2,900 visits annually. Finally, we consider the impact of reducing the perceived costs of being detained in prison. We find that this can shorten case lengths by 193 years annually, remove 2,523 court visits each year, and cut turnarounds by over 40%.

In chapter 2, we consider targeted intervention to reduce the incidence of turnarounds. But because turnarounds account for less than 5% of the detainee population, detainees who receive this intervention would need to be selected carefully. This paper attempts to score detainees using data available to jails to predict turnarounds before they happen and prioritize intervention. We develop a scoring method that predicts turnarounds before they occur, using data about the detainee, their case, and their current case length. We also extend this scoring method to prioritize detainees whose cases are predicted to end after a lead time of up to 28 days. These scoring methods rely on two tools: First, a classification algorithm which determines detainees' probability of being a turnaround given their attributes and case length. Second, a proportional hazards model which predicts detainees' probability of their case ending at a certain case length given their current case length. Testing this scoring method with immediate intervention on 100 detainees each month for four months in 2016 results in 58 turnarounds identified per month, 10.1 years of dead days removed each month, and an associated excess housing cost of over $525,000 per month. Incorporating a 28 day lead time for the intervention to be effective results in 52 turnarounds identified per month, 8.2 years of dead days removed each month, and an associated excess housing cost of over $429,000 per month.

Finally, in Chapter 3, in a conjoint analysis study, we analyze the relative import of mean duration, variability, line length, and reward for "everyday" waits: those of moderate duration (less than twenty minutes) and modest reward (approximately five dollars). We find that mean duration and variability are the key drivers of people's disutility over waits. The latter suggests that customers are risk averse with their time, a phenomenon rarely included in queueing models. We also find that the information about a wait—how it is presented and customers' beliefs about it—strongly influences customer utility. We identify three primary information effects: (1) Mean duration appears twice as costly when wait times are presented in aggregate (like ride sharing apps) than when presented per-person (like grocery store lines). (2) People familiar with a wait defer to their prior beliefs in lieu of posted statistical information. And as a corollary to the previous item, (3) posting information about a wait's duration or variability does little to induce more sensitivity to that feature for customers familiar with the wait. The interaction of information and fundamentals can connect people's utility over waits to their behavior in queueing systems. We capture these interactions in a utility function for modelers desiring an empirically grounded specification of people's utility. Finally, we provide a series of managerial insights for practical use by managers and researchers alike.

CHAPTER 1
UNINTENDED CONSEQUENCES OF PRETRIAL DETENTION

1.1 Introduction

In his award-winning book *Courtroom 302*, Bogira (2005) documents the bleak, everyday events of a Chicago courthouse. Illustrating the overburdened court system, he tells the story of Amy Campanelli:

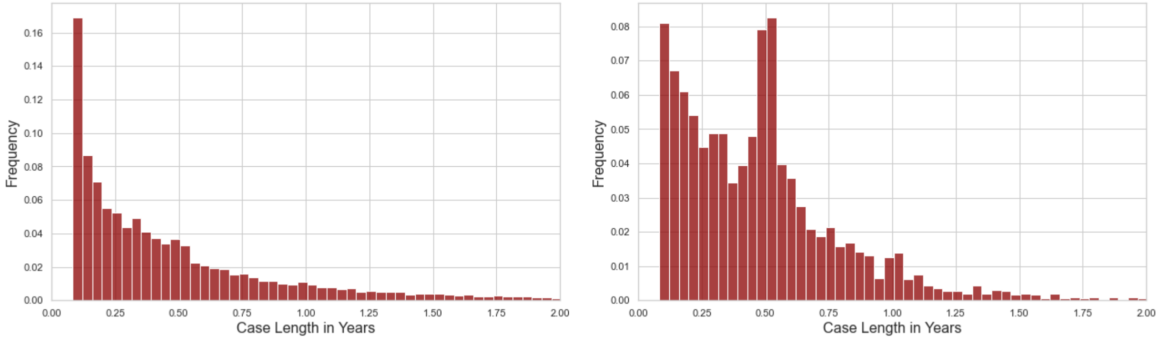
> Amy Campanelli loves criminal defense work, but she's burned out by the caseload of a courtroom public defender. So early in 1998 the ten-year veteran decides to quit.
>
> Fifty cases would be manageable, she says in her office on a February afternoon, as she packs boxes on her final day at work, but she had more than a hundred... So she had to repeatedly ask for continuances. (p. 124)

At any given time, there are about 7,000 people detained in the Cook County Jail and 2,500 people on Electronic Monitoring (EM). The vast majority are detained pretrial, presumed innocent of their charges. But as with Amy Campanelli's clients, these defendants' cases are repeatedly continued; tabled month-by-month while an overloaded court system slowly processes them.

Cases begin when a person is first detained and continue while the courts work to determine their guilt or innocence. This can be time consuming: evidence must be discovered and shared between the prosecution and defense, administrative motions must be filed and processed, and arguments must be developed by both sides regarding the disposition of the detainee. When a court visit is scheduled but adjudication isn't complete, the case is continued until a future date. Ostensibly, cases end with a trial when this fact-finding portion of the case is complete. But in practice, over 95% of cases end in plea bargains (Foxx, 2018).

By accepting a plea, detainees waive their right to trial in return for negotiated concessions from the prosecutor. If the defendant is found guilty, any time they spent detained in jail or on EM counts toward their sentence duration. If they are found not guilty, they are let free; any time incarcerated pretrial was wasted. Of course, regardless of the verdict, locking up defendants cost the tax payers. The Board of Commissioners of Cook County (2021) estimate that it costs $143 per day on average to hold a detainee in jail.

It isn't uncommon for cases to endure longer than a year (see Figure 1.1a). In the extreme, cases can take so long that the detainee is locked up in the jail longer pretrial than their sentence eventually requires. Deemed "turnarounds," these detainees are brought to prison, are photographed, fingerprinted, and booked, and are immediately released. Compensated with one-way bus fare, they "turn around" from the state prison in Joliet, Illinois back to the city.[1] Within our data, in 2016, there were approximately 2,000 turnarounds released from the Cook County Jail to the state prison (the Illinois Department of Corrections (IDOC)). The excess time they spent incarcerated—their "dead days"—amounted to over 675 years, and cost Cook County over $35 million in housing costs.



(a) All Detainees

(b) Class 4 Felons Sentenced to Prison

Figure 1.1: **Histograms of Case Lengths.** Case lengths are restricted to be longer than 30 days and less than 2 years for visual clarity.

In this paper, we consider policies designed to alleviate the delays that keep people detained pretrial. But to do so, we need to understand detainee behavior and the drivers

[1] `https://www.chicagotribune.com/investigations/ct-jail-prison-turnaround-met-20150412-story.html`. Accessed on 10/16/2021.

of case lengths. Generally, longer cases become less likely as time goes on. But in some cases, such as in Figure 1.1b, there are conspicuous “spikes” in the detainees’ case length distribution which occur at common sentence durations. Is it possible that there is more to it than a simple congestion story? Perhaps some detainees are intentionally delaying their cases to spend less time at undesirable sentencing locations. For example, consider a detainee awaiting trial on EM: at home, able to work, and getting credit against an eventual prison sentence. It may be preferable for their case to continue while they are on EM so they spend less time in prison. Detainees have loose control over the balance of their time they spend in their pretrial location vs. their post-sentencing location. Because most detainees plea, they have a good signal of the sentence they will receive, and can accept the plea to end their case when they desire. Roughly speaking, by continuing their case, they reduce the time they spend in their post-sentencing location in favor of their current pretrial housing location. For example, the highlighted spikes in the case-length histograms in the bottom-right panel of Figure A.3 in Appendix A.3.3 for detainees on EM, who have class 3 or 4 felony charges or mesdemeanor charges, are consistent with such behavior.

In Section 1.5, we develop a model of detainee behavior when detained pretrial, and use it to estimate the perceived costs of being detained on EM, in jail, and in prison. We then apply these cost estimates alongside data provided by the Cook County Sheriff’s Office (CCSO) to study the effects of four different sets of counterfactual changes: straightforward improvements to the case processing at the Cook County Courts, paying the bonds of detainees with low-level charges, split sentencing, and reducing the perceived cost of prison.

Straightforward improvements to the Cook County Court system could curtail unnecessary administrative court visits for detainees. We model these style of improvements as a reduction in the number of court visits required to resolve detainees’ cases. We show that reducing the required number of court visits by one can reduce annual Cook County Jail housing costs by $20.1 million, turnarounds by 10.9%, and total case lengths by over 515 years each year. The reduction in case length also helps ameliorate some of the load of the

Cook County courts system.

Detainees are held in jail pretrial if they can't meet their bond conditions. Many people's bonds are small—one or two hundred dollars—and yet they are detained pretrial for months. We show that for a wide range of bonds, it is less expensive to pay the detainees' bonds than house them during their trial. We suggest a prioritization method for detainees who don't pose a threat to society. This method is easy to implement and results in housing cost savings that are more than four times what it spends on paying for bonds. A yearly million-dollar investment, for example, could reduce the jail population by approximately 1.7%, and save Cook County roughly $4.5 million in annual housing costs.

Split sentencing divides detainees' sentenced time between incarceration and release. It reduces the jail population by directly reducing sentence durations and reducing detainees' incentives to delay. This results in significant reductions in EM, jail, and prison populations, saves money, and reduces the load on the courts. For example, splitting sentences by half would save 870 years of detainee sentence time each year, reduce annual housing costs from the jail by $8.5 million, and reduce number of court visits by 2903. This policy also impacts the turnaround population. Measured against their original sentences, splitting sentences by one half would reduce the turnarounds population by 12.6%. But precisely because detention time in detainees' sentences is reduced, there would be an increase in "new turnarounds": detainees whose case length is longer than their split sentence. We find that splitting sentences in half would nearly double new turnarounds, causing a new 946.7 years of dead days, when measured against the detainees new, shorter detention portion of their sentences. This suggests that while the policy would make both the jail and detainees better off, it should be paired with polices which reduce case lengths to avoid the creation of new turnarounds.

Finally, we study the impact of reducing detainees' perceived cost of prison. By reducing its perceived cost to be equal to that of jail, case lengths would be reduced by 193 years each year, there would be 2523 fewer court visits each year. Turnarounds would be reduced by 41.5%, and the detainees would have 63 fewer years worth of dead days each year. In addition to the reduced housing and court administration costs, the disutility borne by detainees in prison would be reduced. By observing payouts from the state for wrongful imprisonment, we associate a dollar cost with time detained in prison: $14,285 per year at the lowest, and $400,000 per year at the highest. The associated reduction in disutility each year for reducing the cost of prison to jail would range from nearly $12 million at the lowest and $335 million at the highest, just for detainees in prison which originated from the Cook County Sheriff's Office.

1.2 Literature Review

This paper is at the intersection of the criminology and operations management literatures; see Berger et al. (2005) for an introduction to criminology. The criminology literature on incarceration is vast as the history of incarceration goes as far back as that of human civilization; see for example, Morris and Rothman (1995) for a history of prisons. Elsner (2006) provides an account of the correctional system in the U.S. circa 2005. It also highlights various important challenges it faces. Clear (2009) documents the vicious generational cycle of imprisonment that affects the disadvantaged neighborhoods of large U.S. cities through ethnographic studies. We refer readers to BJS (2021) and Myers and Lough (2014) for further background.

The operations management literature that focuses on criminal justice is thin. Maltz (1994) and Maltz (1996) provide overviews circa 1990; also see Blumstein (2007) for an overview of his and his collaborators' contributions to this field. Early work in the field attempted to broadly model the criminal justice system, see Blumstein and Larson (1969), Reich (1973), Nagel and Neef (1976), Brantingham (1977), Harris and Moitra (1978), Cassidy (1985). More recently, Dabbaghian et al. (2014) model the criminal justice system of British Columbia at a high-level. Zooming in, some work has been devoted to police staffing, patrolling, and dispatch, see Freeman (1992), Swersey (1994), and Green and Kolesar (2004). Seepma (2020) and Hancock and Raeside (2010) analyze communication processes within the criminal justice space. Bray et al. (2016) study the optimal scheduling of court visits. Combining criminal justice, healthcare, and operations, Ayer et al. (2019) study hepatitis C treatment in U.S. prisons and propose effective policies.

Within the intersection operations and criminal justice, a few papers are concerned with detention, as with our paper. Usta and Wein (2015) is the most relevant. They study the effectiveness of the pretrial release and split sentencing policies by estimating the flows between various segments of the criminal justice system. They then evaluate how these policies would trade off reductions in the jail population with increased recidivism risk for the population. They show that split sentencing is more effective for Los Angeles' estimated process flow. Their work influenced our counterfactual study in Section 1.7.3. Master et al. (2018) extends this work, assuming that jails may not exceed their population cap by renting space from neighboring precincts, and characterize approximate performance measures for policies which offer pretrial release and split sentencing to detainees. Finally, Korporaal et al. (2000) analyze prison capacity in the Netherlands.

Mathematics is often used in criminology, and because the field of criminal justice can be so tethered to issues of operations, operations and non-operations problems in the field can be difficult to separate. We list some reviews as well as some individual papers which are focused on criminal justice, are quantitative, and are operations-esque for the interested reader. Avi-Itzhak and Shinnar (1973) reviews quantitative models in crime control circa 1970. More recently, Weisburd (2017) reviews quantitative methods in criminology. Pratt (2014) collects several papers which expemplify the contribution quantitative methods can have on criminal justice. Risk assessment is common throughout the criminal justice space, and has become more quantitative over time; see Yang et al. (2010) for a review of nine risk assessment tools. Wang and Wein (2018) and Wang et al. (2020) study and propose policies to reduce the backlog of untested sexual assault kits in the USA. Wang et al. (2017) and Wang et al. (2018) analyze ballistic imaging systems, and propose policies which pair firearms and cartridge cases from crime scenes and test fires more efficiently.

Methodologically, our model of detainee costs resembles those seen in the structural estimation literature. The seminal papers Rust (1987) and Berry et al. (1995) are among the first in this area, also see Nevo (2000). In particular, the market share constraints in our model are analogous to those of Berry et al. (1995) and Nevo (2000). More recently, structural estimation has been used for a wide range of applications in operations management. Olivares et al. (2008) study the structural estimation of a newsvendor model and apply it to operating room scheduling, also see Musalem et al. (2010) for structural estimation of stock-outs. Similarly, Akşin et al. (2013, 2017) and Ata et al. (2017) study structural estimation of the delay sensitivity of call center customers and surrounding theoretical questions. Li et al. (2014) study the behavior of customers in the air-travel industry, making use of a structural estimation model to impute the fraction of strategic customers in the population. Moon et al. (2018) empirically study markdown pricing using structural estimation; also see Bimpikis et al. (2020). Bray et al. (2019) explores consequences of the bullwhip effect using structural estimation. Buchholz (2018) and Ata et al. (2019) use structural estimation to study the behavior of taxi drivers in New York City using ride data. Dong et al. (2020) stududy mobile money markets. Shen et al. (2020) studies a healthcare application. The authors use a structural model to demonstrate differences in emergency departments' admission behavior during peak periods, and suggest policies to alleviate the inefficiencies caused by this behavior. Also in the healthcare domain, Agarwal et al. (2021) and Ata et al. (2020) use structural estimation to study the deceased-donor kidney allocation system in the U.S. We refer the reader to Musalem et al. (2017) for a recent, more detailed review of this stream of literature.

Finally, our structural model relies on estimating delaying detainees' case length distributions from a set of positive and unlabeled data. We refer to Bekker and Davis (2020) for a comprehensive review on this subject.

1.3 The Criminal Justice System Through an Operations Management Lens

In this section, we provide an operations-focused summary of how detainees move through the (often complex) criminal justice system. Following the convention of the Bureau of Justice Statistics (BJS) (BJS, 2021), a detainee's case can be thought of in three parts: prosecution and pretrial services, adjudication, and sentencing, which roughly translate to the beginning, middle, and end of their case, see Figure 1.2. We refer to "pretrial" as the entire duration before sentencing, and "post-sentencing" as everything afterward. We focus on the processes and outcomes which are most common and are relevant to the analysis in this paper. For more detailed descriptions, we refer readers to (BJS, 2021) and (Myers and Lough, 2014). We use the terms "person," "defendant," and "detainee" to refer to the accused individual as appropriate during their case.

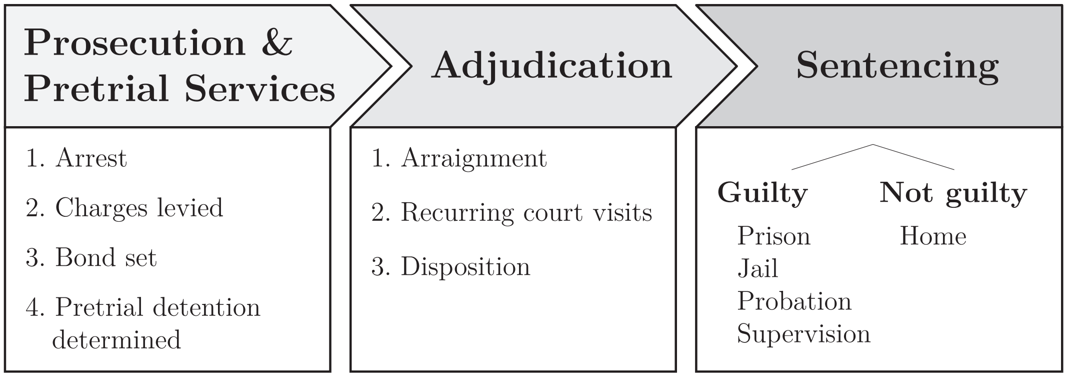


Figure 1.2: **A Three Step Process Through the Criminal Justice System.**

Prosecution and Pretrial Services. A person's case typically begins after an arrest or grand-jury indictment. Within a few days, the courts move through pretrial administrative procedures which allow the case to proceed. First, the prosecutor files charges against the defendant, enumerating the laws they are accused of breaking. The defendant is assigned a defense attorney, i.e. public defender, if needed. Then, the charges are reviewed in a preliminary hearing to ensure that there is probable cause for the case to continue—otherwise charges are dropped or dismissed. During this time, "pretrial services" collect relevant data about the defendant, such as their criminal history, residence, and drug use.

If the case continues, the defendant is quickly brought to bond court. This will determine the defendant's pretrial detention status. They can either be released, detained on Electronic Monitoring (EM), or detained in jail. The bond court judge evaluates the defendant and their case on two metrics: their likelihood to return for trial and their likelihood of being a danger to the community. The judge then assigns a bond, which stipulates the defendant's conditions for pretrial release, if any. These conditions are typically monetary, but can incorporate special qualifications such as EM, home visits from police, or surrendering of passports.

The most common bonds are monetary bail bonds. The three primary types are I, D, and cash. Each lists a dollar amount, such as $50,000-D, which indicates the penalty for not appearing for court. The three differ by what fraction of the listed amount the detainee must post (pay) up front for release. I-bonds (individual recognizance bonds) require no up front payment. D-bonds (deposit bonds) require 10% up front. Cash bonds require 100%. If the defendant is present for court, any up front payments are returned, although court fees are sometimes taken from the posted bond. If they cannot pay the required amount for a D or cash bond, defendants are detained in jail pretrial.

As mentioned above, another condition which can be imposed on detainees is EM. On EM, defendants may return home pretrial but are monitored by a GPS device. Depending on the circumstances of the case, defendants may be allowed to move between approved areas, such as home and work. Leaving the approved areas or removing the GPS device violates their bond conditions and can lead to more severe charges.

The judge may also decide to detain the defendant until their case is complete. This is referred to as "no bail." In this case, there are no conditions the defendant can meet to be released pretrial.[2]

[2] In accordance with the severity of this bond, when no bail is set, the judge reads a script which outlines the test that they use to determine that no bail is appropriate. It reads: "The proof is evident and the presumption great that the defendant is guilty of the alleged crimes. It is clear that the defendant

Adjudication. Adjudication is the portion of the case devoted to determining the guilt or innocence of the detainee. It takes place after the bond hearing, and begins with an arraignment hearing. It ends when the detainee receives a verdict of guilty or not guilty. This is generally the longest part of the case (see Section 1.4 for more detailed information about case lengths).

The defendant initially pleads guilty or not guilty to their alleged charges during the arraignment hearing. A plea of not guilty is most common during arraignment. However, defendants frequently switch their plea to guilty when accepting plea bargains later in their case.[3]

After the arraignment hearing, defendants visit court multiple times before their case ends. These visits span a wide range of purposes, such as administrative motions in the case, discovery of evidence, coordination of witnesses, and communication between attorneys. The defendant's case persists during this time because of "continuances"—motions by the attorneys that table the case to a future date to ensure it is properly adjudicated. The prosecution and judge are limited in this capacity because defendants have a right to a speedy trial, see Appendix A.2. Thus, most continuances come from the defense. The recurring court visits that arise due to these continuances are typically spaced 3-5 weeks apart. It is not uncommon for cases to extend longer than a year during this time.

Adjudication ends when the defendant is found guilty or not guilty. Defendants are guaranteed the right to a trial by jury. They may also forgo the jury and instead opt for a bench trial, where the judge serves the jury's role. However, the vast majority (over 95%) of cases conclude because of a plea bargain. In these cases, the defendant admits guilt for a set of charges in exchange for a known sentence negotiated with the prosecution. If the defendant is found not guilty, no sentence is imposed. If they are found guilty, the judge

represents a clear and present danger to the community, and there are no conditions which can reasonably ensure the defendant's return."

[3] At any stage, a guilty plea must be evaluated by the judge to ensure the defendant was not coerced and understands the implications of the plea. If accepted, the judge may then determine the defendant's sentence.

hands down a sentence during their sentencing hearing.

Sentencing. The judge administers a sentence for any detainee found guilty. They are composed of two parts: a location and a duration. Prison is the most common location, followed by jail. Jail sentences are typically reserved for misdemeanors or short felony sentences. In both cases, any time the detainee spent incarcerated in jail or on EM pretrial counts toward their sentence duration. These sentences are also subject to sentence credit, which allows detainees to serve as little as half of their sentence duration if found guilty of low-level crimes, see Appendix A.1 for further details.

Supervision and probation are less severe sentences. They do not require detention post-sentencing, but still restrict detainees following a guilty verdict. Supervision typically stipulates that the detainee not reoffend during a set period. If successful, their charges are often eligible for removal from their criminal record. Probation is more severe. In addition to not reoffending, detainees are typically required to meet with a probation officer and pass regular drug tests. Charges which result in probation are usually not eligible for removal from the detainee's criminal record.

1.4 Data

Our dataset primarily consists of data retrieved from the Cook County Office Offender Management System (CCOMS) via the Cook County Sheriff's Office (CCSO). We also make use of data from the Illinois Department of Corrections (IDOC), the state's prison system. We use five data files which collectively provide the data fields listed in Table 2.1. Each contributes the following information: the beds file lists the detainees' pretrial housing location, the bonds file lists the detainees' bond type and amount, the courts file lists the detainees' court dates, the IDOC file has data about people detained in prison, which we use to determine sentence durations for detainees sentenced to prison, and he main CCOMS file provides all remaining data fields in Table 2.1.

In summary, our dataset provides information about detainees' cases from booking through sentencing. We focus attention on detainees booked in 2015 and 2016 and who remained under the CCSO's purview either on EM or in jail. They correspond to 98,882 rows in our dataset, each row representing one detainee's booking. This allows us to follow them through the completion of their detention in jail on on EM under CCSO custody. We use different portions of this data for various elements of this paper. We describe each data field in detail in Appendix A.3.

Detainee	**Housing**	**Case**	**Sentencing**
Inmate ID	Booking ID	Docket number	Sentence location
Criminal history	Booking date	Crime class	Sentence duration
Prison history	Pretrial housing location	Case length	Turnaround status
	Security classification	Court dates	
	Bond type		
	Bond amount		

Table 1.1: **Data Fields.**

Our classification and estimation procedures in Appendix A.4, 1.5, and 1.6 focus on intentional delay behavior and the relative perceived cost of EM, jail, and prison. To estimate these relative costs, we concentrate on detainees who plead guilty, were detained in jail or on EM, whose case lengths were greater than 60 days, and were sentenced to prison. We use turnarounds as signals for delaying behavior for developing a classifier in Appendix A.4. We also restrict our focus to detainees whose primary charges are classes 1, 2, 3, 4, and A; removing classes X and M because they are too severe and classes B and C because they rarely result in pretrial detention. Finally, we remove detainees whose sentence was greater than 4 years pre-sentence credit because of the implied severity of the crime. In combination, these restrictions allow us to focus on detainees who potentially delayed.

Our counterfactual analysis in Section 1.7 focuses on detainees who entered CCSO custody during 2015-2016 and uses the whole dataset corresponding to those years to calculate the jail population. If delaying behavior is analyzed, as done in counterfactuals 1, 3, and 4, we simulate delaying behavior for the subset of detainees for whom we have the case length distributions and perceived housing costs. Namely, detainees who are housed in jail or on EM, were sentenced to jail or prison, whose crime classes are 1, 2, 3, 4, or A, whose sentence was less than 4 years pre-sentence credit, and whose case length was greater than 60 days. The rest of the detainees either remain in the CCSO's custody for a short duration if at all, e.g., crime classes B, C, or their charges are too severe, e.g. crime classes X, M, and hence the nature and the evolution of their cases is very different. The latter group of detainees constitute a negligible portion of the jail population.

1.5 Structural Estimation of Detainees' Location Costs

In this section, we develop a structural estimation model to estimate detainees' relative perceived costs of being detained in EM, jail, and prison. We assume a detainee's cost is linear in their length of stay, but his cost rate can differ across different locations. Because pre-trial detention time counts against sentence duration, detainees can loosely balance the amount of time they spend at different housing locations by either intentionally delaying their case or letting it follow its natural course. This choice results in different case length distributions for each detainee. Such variation in the data allows us to identify the cost parameters. In doing so, we allow (unobserved) heterogeneity among detainees' cost rates and estimate the distribution of the cost rate per time unit for each location. To be more specific, we seek the cost parameters that maximize the likelihood of detainees' case lengths observed in the data. In doing so, we restrict attention to detainees who are sentenced to prison and were either on EM or held in jail pretrial. The results of this Maximum Likelihood Estimation are presented in Section 1.6 and are used in our counterfactual analysis in Section 1.7.

Detainees have two phases in which they can be incarcerated and thus accrue housing costs: pre-sentencing (or phase 1) and post-sentencing (or phase 2). During phase 1, detainees can be held in jail (J) or electronic monitoring (EM). During phase 2, detainees are incarcerated in prison (P).

Let $l_i \in \{EM, J\}$ denote detainee i's phase 1 location and P denote his phase 2 location of prison. He incurs a linear cost of $c_i(l)$ in location l per unit of time he is incarcerated there. We denote his case length by W_i and his sentence duration by S_i.[4] Because the detainees we focus on plea guilty, they know their sentence duration in phase 1. As mentioned earlier, the time detainees spend incarcerated pre-trial counts toward their sentence duration. Thus, detainee i spends w_i in location $l_i \in \{EM, J\}$ and $(s_i - w_i)^+$ in prison. If the detainee's case length exceeds their sentence, he is immediately released upon sentencing. In particular, he is a turnaround.

Detainee i chooses an action $a_i \in \{D, N\}$, representing intentionally delaying or not, respectively. These actions result in different case length distributions, W_i, with cdfs (pdfs) $F_{a_i}(w_i)$ $(f_{a_i}(w_i))$ from which each detainee's case length is drawn.

Cost Structure and Probability of Delaying. Let $C_i(A)$ denote the expected cost of incarceration for detainee i, who chooses action $a_i = A$, with phase 1 housing location l_i and sentence duration s_i:

$$C_i(A) = c_i(l_i)\mathbb{E}_A[W_i] + c_i(P)\mathbb{E}_A[(s_i - W_i)^+], \tag{1.1}$$

where the expectation is taken over W_i under F_A for $A \in \{D, N\}$.

Detainees are heterogeneous in their perceived costs of detention. We assume that each detainee draws his cost parameters from a Gaussian distribution whose mean and standard deviation depend on the detention location. To be specific, we assume that

$$c_i(l) \sim \mathcal{N}(\mu_l, \sigma_l^2), \quad l \in \{EM, J, P\}. \tag{1.2}$$

The following proposition is immediate from Equations (1.1)-(1.2).

Proposition 1 *Detainee i's expected cost associated with action A has a Gaussian distribu-*

[4] We use lower case w_i and s_i to denote the realized case length and sentence duration.

tion with mean $\tilde{\mu}_A(i)$ and variance $\tilde{\sigma}_A^2(i)$ for $A \in \{D, N\}$, where

$$\tilde{\mu}_A(i) = \mu_{l_i}\mathbb{E}_A[W_i] + \mu_P\mathbb{E}_A[(s_i - W_i)^+],$$
$$\tilde{\sigma}_A^2(i) = (\sigma_{l_i}\mathbb{E}_A[W_i])^2 + (\sigma_P\mathbb{E}_A[(s_i - W_i)^+])^2,$$

and $\mathbb{E}_A$ is taken over case length W_i with respect to the cdf F_A.

Detainees seek to minimize their expected costs by choosing between intentionally delaying their case (D) or not (N). We let p_i denote the probability that detainee i (with pre-sentencing location l_i and sentence duration s_i) choose to intentionally delay his case. We have that

$$p_i = \mathbb{P}\left(C_i(D) \le C_i(N)\right).$$

The following proposition characterizes p_i.

Proposition 2 *Detainee i's probability of intentionally delaying is given by*

$$p_i = \Phi\left(\frac{\tilde{\mu}_N(i) - \tilde{\mu}_D(i)}{\sqrt{\tilde{\sigma}_N^2(i) + \tilde{\sigma}_D^2(i)}}\right),$$

or equivalently, in terms of the location cost parameters:

$$p_i = \Phi\left(\frac{\mu_{l_i}\left(\mathbb{E}_N[W_i] - \mathbb{E}_D[W_i]\right) + \mu_P\left(\mathbb{E}_N[(s_i - W_i)^+] - \mathbb{E}_D[(s_i - W_i)^+]\right)}{\sqrt{\sigma_{l_i}^2\left(\mathbb{E}_N[W_i]^2 + \mathbb{E}_D[W_i]^2\right) + \sigma_P^2\left(\mathbb{E}_N[(s_i - W_i)^+]^2 + \mathbb{E}_D[(s_i - W_i)^+]^2\right)}}\right). \tag{1.3}$$

Estimation Formulation. We maximize the likelihood of the observed detainee case lengths given the cost structure outlined above. Detainee i's case length w_i is drawn from cdf F_D if he intentionally delays his case, which occurs with probability p_i. Otherwise, w_i is drawn from cdf F_N. Letting $L_i(w_i)$ denote the likelihood of detainee i's case length w_i, we have that

$$L_i(w_i|\mu_{EM},\mu_J,\mu_P,\sigma_{EM},\sigma_J,\sigma_P) = p_i f_D(w_i) + (1-p_i) f_N(w_i), \tag{1.4}$$

where f_D and f_N are the pdfs associated with F_D and F_N, respectively.

The likelihood of observing case lengths $w_1,...,w_I$, denoted by $\mathscr{L}(\mu_{EM},\mu_J,\mu_P,,\sigma_{EM},\sigma_J,\sigma_P)$, is then given by

$$\mathscr{L}(\mu_{EM},\mu_J,\mu_P,,\sigma_{EM},\sigma_J,\sigma_P) = \prod_{i=1}^{I} L_i(w_i|\mu_{EM},\mu_J,\mu_P,,\sigma_{EM},\sigma_J,\sigma_P) \tag{1.5}$$

We also require that

$$\frac{1}{|N_l|}\sum_{i\in N_l}\Phi\left(\frac{\tilde{\mu}_N(i)-\tilde{\mu}_D(i)}{\sqrt{\tilde{\sigma}_N^2(i)+\tilde{\sigma}_D^2(i)}}\right) = \frac{1}{|N_l|}\sum_{i\in N_l}\hat{y}_i, \quad l\in\{EM,J\} \tag{1.6}$$

where the left-hand side is the average probability of intentional delay predicted by our model, whereas the right-hand side is the average predicted probability of delaying derived from the SAR-EM method, see Appendix A.4. We impose this for each pre-trial housing location; N_l denoting the set of detainees housed in those locations pre-trial. Conceptually, this ensures that the proportion of delaying detainees are consistent between our two estimation methods. This constraint can be thought of as similar to the market share constraints in the formulations of Berry et al. (1995) and Nevo (2000).

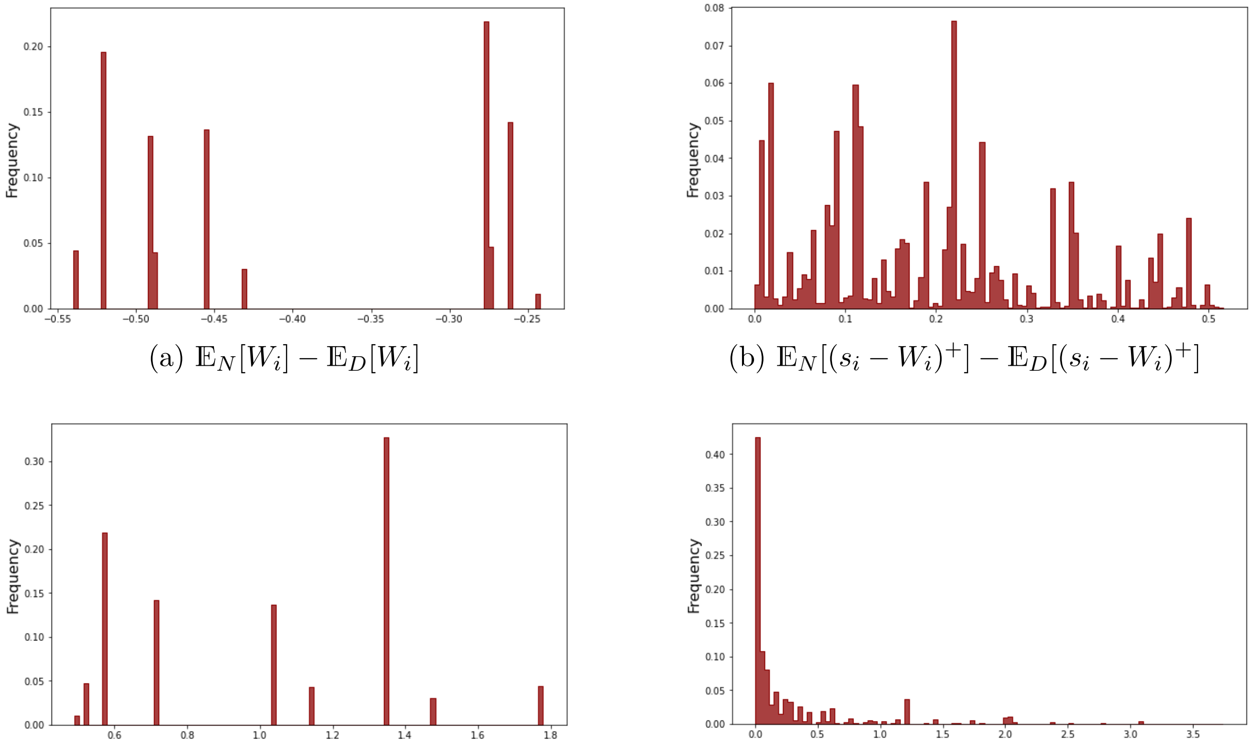
Moreover, as can be seen from Equation (1.3), the probability of intentionally delaying is left unchanged if we scale all cost parameters proportionally. Therefore, for identification purposes, we restrict the sum of σ's across each location to be one. That is,

$$\sigma_{EM}+\sigma_J+\sigma_P = 1. \tag{1.7}$$

Then the resulting MLE formulation is given as follows:

$$\begin{aligned} \max_{\mu_{EM},\mu_J,\mu_P,\sigma_{EM},\sigma_J,\sigma_P} \quad & \log\left(\mathscr{L}\left(\mu_{EM},\mu_J,\mu_P,\sigma_{EM},\sigma_J,\sigma_P\right)\right) \\ \text{subject to} \quad & (1.6)-(1.7). \end{aligned} \tag{1.8}$$

Identification. The structural parameters in our model drive changes in p_i, the detainee's probability of delaying. This, in turn, drives changes in the likelihood function, as long as $f_D \neq f_N$ for most case lengths observed in the data. Simple inspection of the histograms of estimated case lengths in Figure A.7 show that this condition holds. Thus, changes in the likelihood function are driven by p_i.



(a) $\mathbb{E}_N[W_i] - \mathbb{E}_D[W_i]$

(b) $\mathbb{E}_N[(s_i - W_i)^+] - \mathbb{E}_D[(s_i - W_i)^+]$

(c) $\mathbb{E}_N[W_i]^2 + \mathbb{E}_D[W_i]^2$

(d) $\mathbb{E}_N[(s_i - W_i)^+]^2 + \mathbb{E}_D[(s_i - W_i)^+]^2$

Figure 1.3: **Histograms of Coefficients of μ and σ.** Figures 1.3a and 1.3c plot the coefficients of μ_l and σ_l for jail and EM. Similarly, Figure 1.3b and Figure 1.3d plot the coefficients of μ_P and σ_P for prison.

As previously mentioned, p_i remains unchanged if all cost parameters are scaled proportionally. After restricting the sum of σ's across location to be one, variation in the coefficients of the cost parameters drives their identification, see Equation (1.3). Specifically, variation in the expectations in the numerator drive the identification of μ's, and variation in the expectations in the denominator drive the identification of σ's. We plot histograms of the calculated values of the four coefficients in Figure 1.3. They exhibit significant variation to identify the parameters. The coefficients of the pretrial housing locations have 10 possible realizations based on the five crime classes and two pretrial housing locations used to develop F_D and F_N. Recall that the case length distributions depend on both the crime class (1, 2, 3, 4, A) and the housing location (EM or jail). The coefficients of the prison cost parameters are even more varied, as they incorporate sentencing data in addition to crime class and housing location.

Intuitively, these coefficients represent the difference in time detainees are incarcerated at the three detention locations. For our model to identify detainee costs, there must be significant variation in this time depending on the detainee's choices. Because our estimated delaying and non-delaying distributions are quite different from each other, and differ between crime classes and pretrial housing locations, our model can identify the parameters of detainees' location costs.

1.6 Estimation Results

Location	Estimated μ	Estimated σ
EM	0.370	0.241
	(0.091)	(0.110)
Jail	1.378	0.446
	(0.274)	(0.091)
Prison	1.835	0.313
	(0.360)	(0.188)

Table 1.2: **Estimated Location Cost Parameters.** Estimates are truncated at three decimal places for readability. Standard errors are listed in parentheses below the estimates.

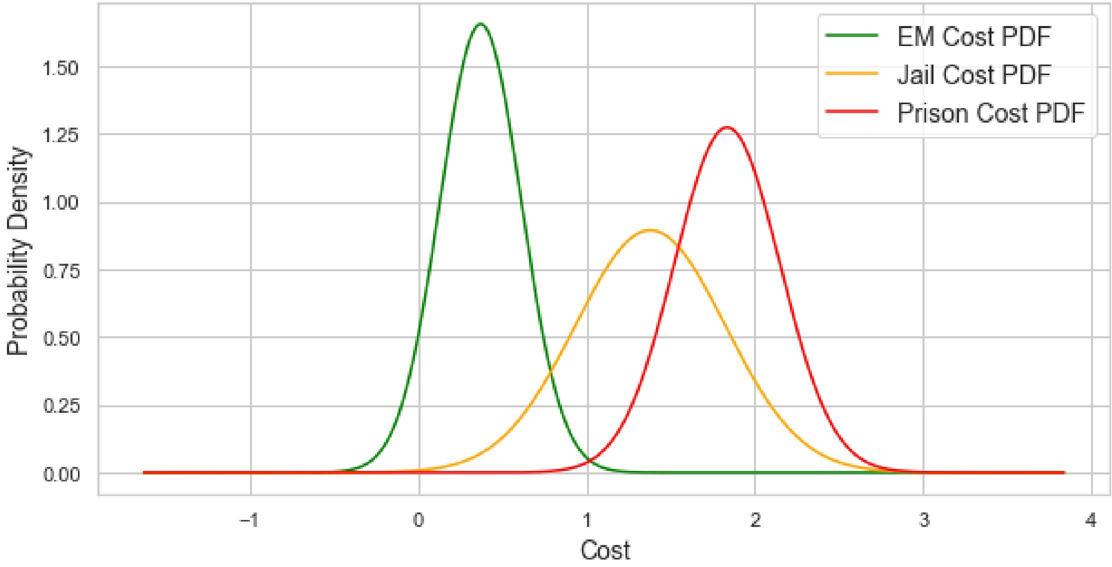


Figure 1.4: **Estimated Location Cost pdfs.** The plotted densities, from left to right, represent the estimated distribution of location costs for EM, jail, and prison.

We maximize the log-likelihood formulation of Equation 1.8 using the nonlinear optimization solver KNITRO (Byrd et al., 2006) via the AMPLpy interface (Brandão, 2021).[5] The resulting parameters of the location cost distributions are given in Table 1.2 and the resulting pdfs are plotted in Figure 1.4. To compute each parameter's standard error, we perform the non-parametric bootstrap method (Horowitz, 2001). We generate 500 simulated datasets with the same size as our dataset by drawing from ours with replacement. We then estimate parameters of the simulated datasets and compute their standard errors, which are listed in parentheses in Table 1.2. Appendix A.6 describes a Monte Carlo simulation study to illustrate the identification of our model; we are able to recover the true parameters when using our estimation procedure on simulated data resulting from those parameters.

1.7 Counterfactual Analysis

This section studies four counterfactuals designed to reduce pretrial detention and save costs. They each address this issue by tackling one of the following: case lengths, bonds, split sentencing, and improving prison conditions. When calculating housing costs for jail, we use $143 per inmate per day in jail (Board of Commissioners of Cook County, 2021).

The counterfactual studies in Sections 1.7.1, 1.7.3, and 1.7.4 involve simulating the detainees' intentional delaying behavior. Whenever this is needed, we simulate their behavior using 50 replications and report the average.

1.7.1 Straightforward Operational Improvements

Straightforward improvements to the Cook County court system could curtail unnecessary administrative court visits for detainees. Many continuances result from the court's outdated, non-digital adjudicatory practices. In their review of the Cook County courts' pre-trial detention, (Staudt, 2020) Chicago Appleseed noted:

[5] To assist in finding a locally feasible point, and to improve the robustness of our results, we consider 40 multistart points.

> “In the federal courts, and most other major state court systems, pieces of evidence are exchanged electronically as soon as they are received by the prosecutor. The Cook County Circuit Court continues to adhere to the outdated practice of physically exchanging paper copies of documents and CDs and DVDs of audio and video recordings only in court, not between court dates. As the amount of digital evidence in cases rapidly increases, this process is even more cumbersome than it was a decade ago.”

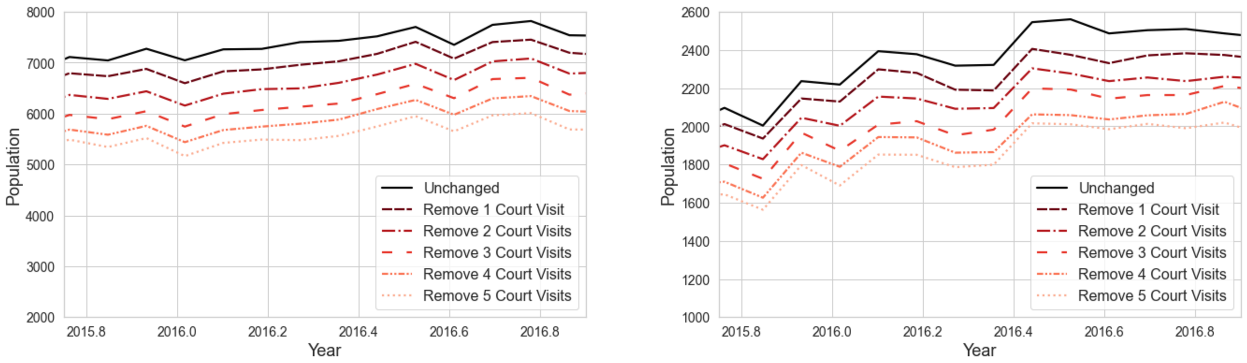
In addition to exchanging evidence electronically, automatically delivering common pieces of evidence, such as body camera footage, can speed up discovery by weeks. Centralizing where various types of court visits take place make scheduling more efficient.

In a 2019 audit of the courts, the National Center for State Courts identify the benefits Cook County could reap from these style of improvements (National Center for State Courts, 2019). These benefits include the following: a reduction in case continuances and postponements, quicker and more case resolutions prior to trial, reductions in needless delays in case processing, integrated online information sharing (e.g. e-discovery exchange), multi-party access to digital files at the same time.

We model these style of improvements as a reduction in the number of court visits required to resolve detainees’ cases. Specifically, we say that improvements reduce the number of court visits by n. Detainees’ court visits are reduced by the interarrival time of their removed court visits, ξ_j, $j = 1, ..., n$ down to a minimum of 30 days (we do not suppose that these improvements can shorten a case to less than a month). That is, we delete their first n court visits. If a case was already shorter than 30 days, their case length is unaffected.

These improvements to the court system will be effective if detainees choose to not delay, so we modify their non-delaying case length distribution in the following manner: $\tilde{w} = max(30, w - \sum_{j=1}^{n} \xi_j)$ for all case lengths w. We refer to this new distribution as $\tilde{F}_N$. Their delaying case length distribution, F_D remains unaffected. Note that this is a conservative analysis: improvements to the court system may have some impact on detainees’ ability to delay. As such our analysis provides a lower bound on the effects of reducing case lengths. We simulate detainee's choice to delay or not, drawing their location costs from our estimates in Section 1.6. Depending on their choice, we draw their case length from $\tilde{F}_N$ or F_D.

For the non-delaying detainees, we simply reduce their case length by $\sum_{j=1}^{n} \xi_j$, down to a minimum of 30 days. If their case length was less than 30 days, it remains unaffected.



(a) Jail Population Over Time

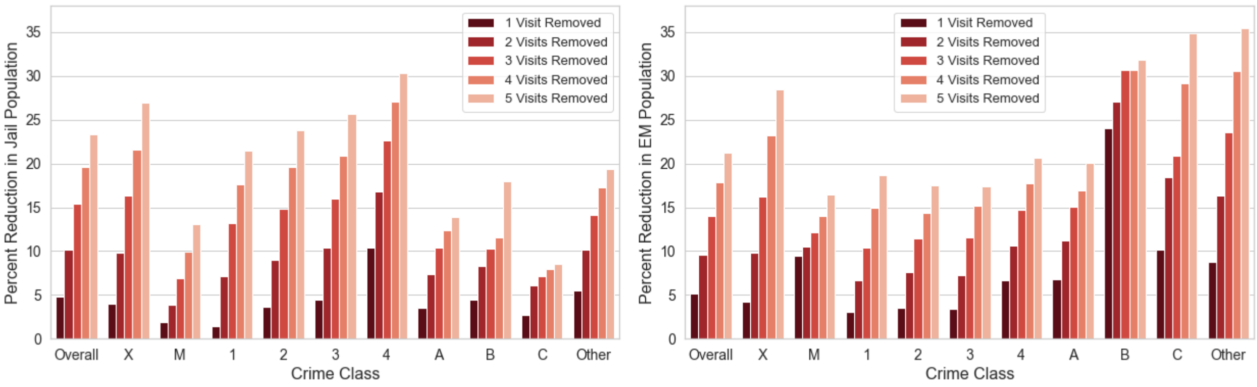
(b) EM Population Over Time

Figure 1.5: **Reductions in Jail and EM Populations Over Time via Reduced Court Visits.** Focusing attention on detainees who entered CCSO custody during 2015-2016, Figures 1.5a and 1.5b display the average predicted detainee populations over time. We restrict our attention between between September 2015 and December 2016 when the simulated population achieves steady state.

The plots in Figure 1.5 show the resultant drop in jail and EM populations from removing detainees' first n court visits. Figure 1.6 displays the overall reduction as well as that for each crime class both for jail and EM populations. It shows that even a moderate reduction in the number of court visits results in significant reductions in both the jail and EM populations. At only one court visit reduced, the jail population is reduced by 4.84% on average in steady state, and 5.22% for EM.[6]

These improvements are impactful for reducing the turnarounds population. Table 1.3 displays the benefits of reducing the number of court visits for the turnarounds population.

[6] Note that the magnitude of the change for EM is larger, despite people on EM having slightly longer cases if crime class is fixed. This is because the interarrival times of court visits can be slightly longer for detainees on EM, and that the proportion of less-severe cases is larger for detainees who are qualified to be released on EM. The more severe, longer cases tend to be for detainees who are detained in jail.



(a) Average Reductions in Jail Population

(b) Average Reductions in EM Population

Figure 1.6: **Average Reductions in Detainee Populations in Steady State via Reduced Court Visits.** These plot the average reduction in jail (Figure 1.6a) and EM (Figure 1.6b) populations due to reduced court visits during the steady state period between September 2015 and December 2016. "Overall" provides statistics for all data. The remaining provide statistics by crime class.

Removing one court visit per detainee would cut turnarounds by 10.9%. This benefit would be associated with 29.8 years worth of reduced dead days each year, saving the jail $1.6 million in excess housing costs per year.

	Yearly Reductions in:			
# Removed	Turnarounds (Net)	Turnarounds (Percent)	Excess Housing Time "Dead Days" (Years)	Excess Jail Housing Costs Due to Turnarounds (Millions of Dollars)
1	244.0	10.9%	29.8	$1.6
2	347.4	15.5%	62.9	$3.3
3	437.5	19.5%	92.9	$4.9
4	517.9	23.1%	125.5	$6.6
5	575.0	25.6%	145.8	$7.6

Table 1.3: **Reduction in Turnarounds due to Reduced Court Visits.**

These reductions come with significant cost savings. Housing detainees in jail pretrial is expensive: $143 per inmate per day (Board of Commissioners of Cook County, 2021). Thus, the aforementioned reductions in jail populations also substantially reduces costs. For

	Yearly Reductions in:			
# Removed	Case Length EM (Years)	Case Lengths Jail (Years)	Total Jail Time Served (Years)	Jail Housing Costs (Millions of Dollars)
1	92.9	413.5	385.6	\$20.1
2	198.7	826.7	785.8	\$41.0
3	311.0	1252.8	1202.1	\$62.7
4	409.6	1624.9	1566.9	\$81.7
5	500.7	1940.0	1876.0	\$97.9

Table 1.4: **Yearly Case Length, Incarceration Time, and Cost Reductions due to Reduced Court Visits.**

a single court visit removed from detainees' cases, Cook County would see a reduction of 413.5 years of total jail time served. This would achieve a cost savings of \$20.1 million from housing costs every year (see Table 1.4 for estimated housing cost savings for 1-5 court visits removed). And, because the courts are overloaded, even modest reductions in the detainee's number of court visits could have major operational improvements and cost savings there as well.

Our analysis is conservative in that we assume these improvements do not affect the detainees' ability to delay. That is, we draw their case lengths from the original F_D distribution estimated in Appendix A.4 if they choose to delay. This underestimates the reductions associated with this policy. And because turnarounds often arise from delaying behavior, this underestimation is particularly acute for that population. To see how changes in incentive structure affect both the turnarounds population and the detained population at-large, see the counterfactual analysis in Sections 1.7.3 and 1.7.4.

1.7.2 Paying Bonds of Lower Level Detainees

The cost of pretrial detention in jail far exceeds the cost of paying for detainee's bonds. Figure 1.7 displays CDFs of detainee detention costs grouped by the twelve most common "effective bonds," the amount necessary for the detainee to be released. Nearly all detainees

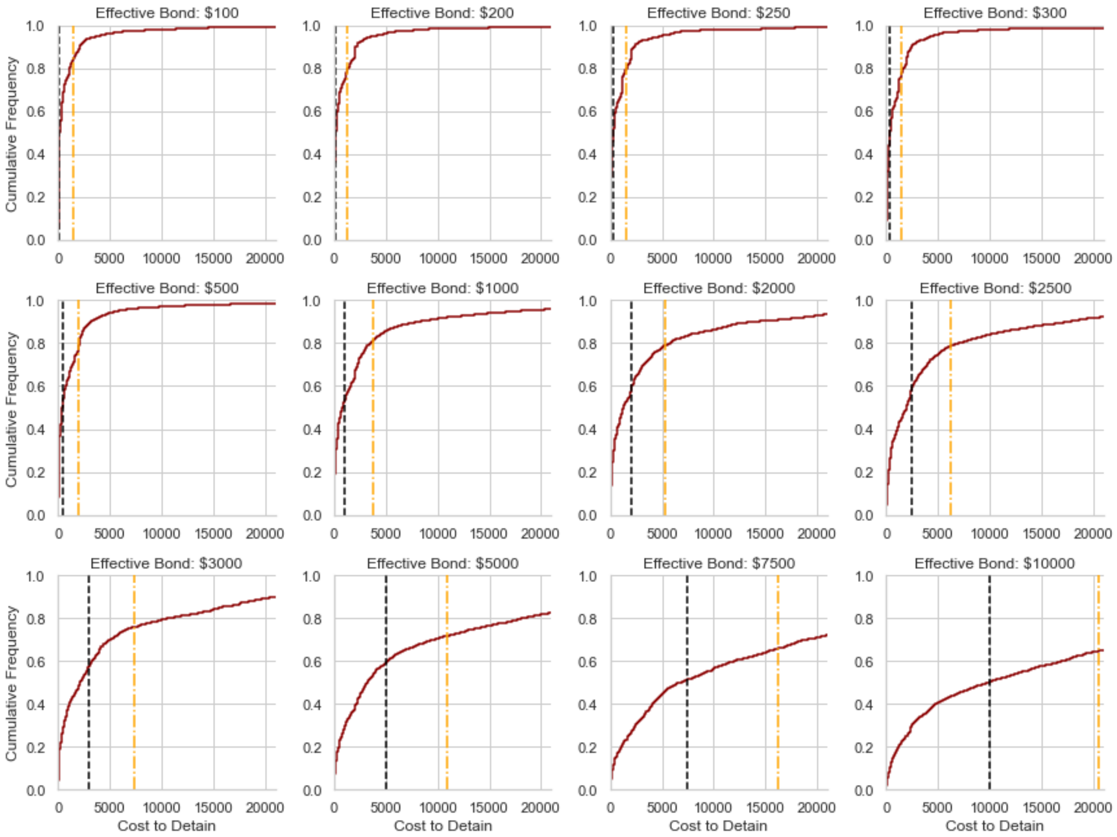


Figure 1.7: **Empirical CDFs of Pretrial Detention Costs by the Twelve Most Common Effective Bond Amounts.** The black dashed line on each CDF represents the effective bond—the amount needed for detainee's of that group to be released. The orange dash-dotted line on each represents the groups' mean costs to detain pretrial. The x-axis is truncated at $21,000 for readability.

with small (yet common) bonds, such as $100 or $200, are more expensive to house than their bond. But even for groups of detainees with large effective bonds of $5,000 or $7,000, 40%-50% are more expensive to detain than their effective bonds. In every case, the mean cost to detain pretrial for the group is greater than the detainees' effective bonds.

In this section, we consider how to pay bonds to reduce the number of detainees held in jail pretrial and analyze the associated cost savings.This policy considers all detainees held in jail, and is applied to subsets of those detainees if they meet the conditions of the policies described below. Suppose a third party (an NGO, Cook County, etc.) had a yearly budget of d dollars with which to pay detainee's bonds. This yearly budget can be thought of as renewing, a-la a county's budget, or revolving due to people returning bails, a-la a revolving bail fund. In a more seasonal vein, charitable parties often pay some detainees' bails near Christmas.[7] If the party knew detainee's case lengths a-priori, they could maximize case length mitigated per dollar by prioritizing detainees by their "efficiency": case length divided by effective bond. Unfortunately case lengths are not known a-priori. A suitable proxy for efficiency is to divide a detainee's expected case length based on crime class (given in Table 1.5) by their effective bond. We deem this metric "approximate efficiency," and make use of it in the policies below.

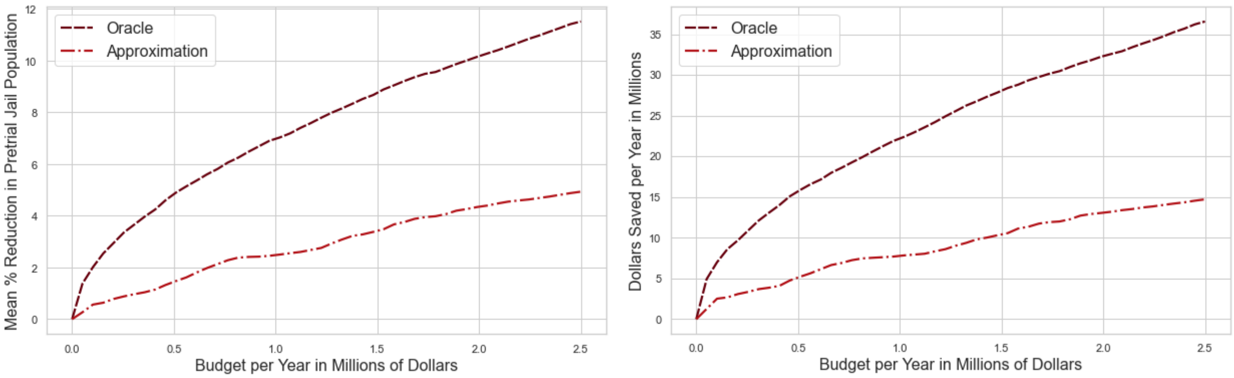
Crime Class	Mean Case Length in Days
M	512
X	201
1	117
2	126
3	88
4	47
A	19
B	13
C	12

Table 1.5: **Mean Case Length in Days by Crime Class.**

We display the results of two policies using these metrics in Figure 1.8. In both, we rank all detainees which enter the jail in a year by the associated metric: the "Oracle" policy ranks by efficiency, and the "Approximation" policy ranks by approximate efficiency. Then, detainees' bonds are paid up to a budget d according to this ranking, the most efficient being paid first. In Figure 1.8a, we show that for a one million dollar yearly budget, the pretrial jail population would be reduced by 6.8% (Oracle) and 2.5% (Approximation). Figure 1.8b displays the associated cost savings.[8]. A one million dollar per year budget would save Cook

[7] `https://chicago.suntimes.com/news/2019/12/25/21037383/messages-of-hope-resonate-in-cook-county-jai-on-christmas`. Accessed on 12/1/2021.

[8] Note that $143 per day per inmate is a lower bound on the cost of detaining people in the Cook County Jail. Special accommodations, mental disability, and more can increase the cost of detention in the jail (Board of Commissioners of Cook County, 2021).



(a) Reduction in Pretrial Jail Population

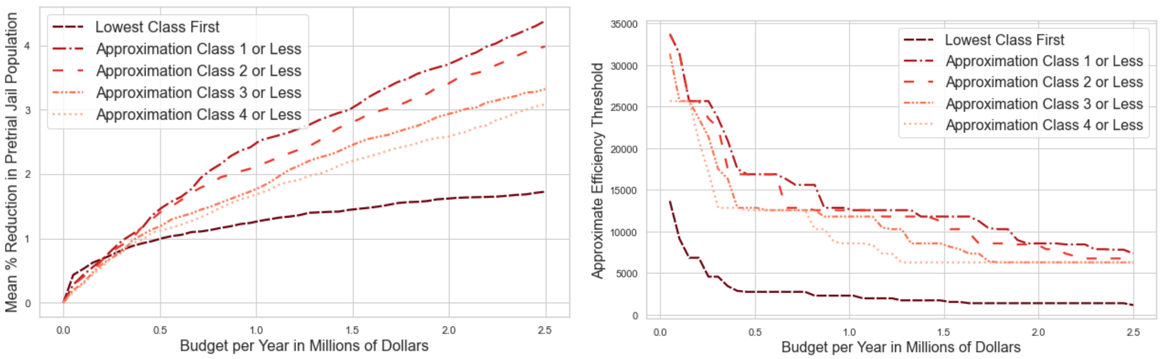
(b) Mean Yearly Housing Cost Savings

Figure 1.8: **Jail Reductions with $$d$ Yearly Budget.** In each Figure, "Oracle" represents the policy priortizing efficiency, while "Approximation" represents the policy prioritizing approximate efficiency. Figure 1.8a displays the average reduction in jail population in steady state. Figure 1.8b shows the mean yearly housing cost savings in steady state. For housing cost calculations, etainees who are released pretrial but are sentenced to jail are presumed to begin their sentence when their case ended before this counterfactual analysis.

County an estimated $26 million under the Oracle policy and $10 million in the Approximate policy. Regardless of budget, reductions in the pretrial jail population and the associated savings are significant.

However, solely focusing on efficiency may release detainees deemed too severe. Detainees accused of class M felonies—first degree murders—rank highest in efficiency by nearly six times the other classes! It is necessary to consider policies which are implementable given information known at the outset of a case, and carefully weigh which detainees are eligible for release.

To that end, we consider five bond payment policies. For each, we consider all detainees who are detained in the jail in a year, but rank them differently. Their rankings are as follows: "Lowest Class First": Pay bonds prioritizing lowest crime class first, then approximate efficiency. And max-severity prioritizations, such as "Approximation Class 1 or Less": Prioritize approximate efficiency, focus only on class 1 or less. We implement the latter for classes 1, 2, 3, and 4. Once ranked, detainees' bonds are paid up to the budget d. This style of policy establishes a threshold of approximate efficiency above which people's bonds are paid if they meet the restrictions of the policy. We plot the threshold in Figure 1.9b. Making use of this threshold allows for a straightforward heuristic to implement these policies in practice.



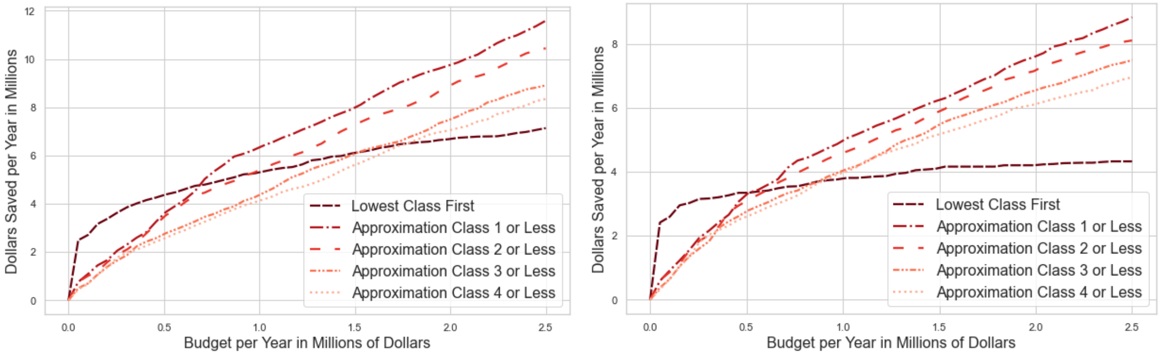
(a) Reduction in Pretrial Jail Population (b) Approximate Efficiency Threshold over d.

Figure 1.9: **Results of Bond Payment Policies with $$d$ Yearly Budget.** Figure 1.9a displays the average reduction in jail population in steady state. Figure 1.9b shows the associated minimum approximate efficiency threshold used by each policy for each value of d.

As seen in Figure 1.9a all three policies are effective at reducing the pretrial detention population. Using the "Approximation Class 3 or Less" policy with a budget of one million dollars results in a near 2% reduction in the pretrial jail population in steady state. At 2.5 million dollars, this reduction exceeds 3%. That is equivalent to a total of 320 years of time detainees would have been detained pretrial per year.

The savings in housing costs resulting from these policies is significant. Figure 1.10a displays these savings for the jail. For all policies and all budgets the savings in housing costs exceed the cost of paying bail. For all policies, a yearly budget of 1 million dollars would result in over 4 million dollars in savings for the Cook County jail. For the the policies which restrict based on maximum crime class, the cost savings are nearly linear in the range we study. Thus, a dollar invested in paying bails under this policy results in a four- to six-fold savings in terms of housing costs. And this is a conservative estimate, as these policies focus on lower-level detainees.

Some detainees would have been sentenced to prison or jail following their trial, incurring housing costs during those periods. However, for detainees sentenced to supervision, or sentenced to probation, their post-sentencing cost to society is much lower. For detainees found not guilty, their costs are nonexistent. In Figure 1.10b, we show housing cost savings, presuming that detention costs in prison are equal to that of jail, and that supervision, probation, and “Charge Dropped or Finding of Not Guilty” have a cost of zero. Bond payments policies are highly effective on returns to society, returning nearly $4 million in housing costs for $1 million in paid bonds.



(a) Yearly Housing Cost Savings for Jail (b) Yearly Housing Cost Savings for Society

Figure 1.10: **Cost Savings with $$d$ Yearly Budget.** Figure 1.10a shows the mean yearly housing cost savings in steady state for the jail. Figure 1.10b displays the reduction in housing costs to the taxpayer, presuming prison housing costs are equal to jail, and supervision and probation cost zero.

Pretrial release is studied by Usta and Wein (2015) within the Los Angeles criminal justice system. They create a model of the jail and courts system, and estimate the flows into, between, and from each element of that system. They find that pretrial release is not as efficient as split sentencing policy when evaluated on a metric of mitigated time served pretrial vs. risk of recidivism when calibrated to the data of Los Angeles. We do not have the necessary data available, namely an equivalent to the California Static Risk Assessment tool, to repeat their analysis and judge its appropriateness for the Chicago setting. However, we find that pretrial release can be an effective cost saving measure. And, the policies we suggest prioritize small bonds and low-level crimes first, which reduces recidivism risk. Our analysis of split sentencing, the policy Usta and Wein (2015) deem more effective for Los

Angeles, is given in Section 1.7.3. We find it to be a powerful cost-saving tool, in line with Usta and Wein (2015).

1.7.3 Split Sentencing

An "x-split sentence" divides a detainee's total sentence duration s between incarceration (in jail or prison) and release (on supervision and/or probation). The parameter x represents the fraction of the detainees' sentence which is dismissed. Thus, a detainee with a split sentence first spends their case length w in their pretrial housing location. Then, they spend any remaining jail or prison time, $((1-x)s-w)^+$, in that location.[9] Finally, they are released to supervision or probation.

As with our model of detainee costs in Equation (1.1), with split sentences, detainees may choose to delay or not. We assume detainees don't incur any costs after they are released. That is, the cost associated with supervision and probation are zero. Thus, for an x-split sentence, detainee i's costs are given by:

$$C_i(A) = c_i(l_i^1)\mathbb{E}_A[W_i] + c_i(l_i^2)\mathbb{E}_A[((1-x)s_i - W_i)^+],$$

where expectations are taken under F_N and F_D as before and l_i^1 and l_i^2 are their phase 1 and 2 housing locations.

Because detainee's sentences are released for part of their sentence, federal sentencing guidelines restrict eligibility for split sentencing to detainee's whose charge and criminal history are not too severe. Specifically, split sentencing is only available to detainees who fall into Zone C or below of the Federal Sentencing Table (U.S. Sentencing Comm'n, 2018), which is reproduced in Appendix A.7, Figure A.8. Detainees in this zone have a maximum sentence of 18 months (prior to the application of any credit time). We use this threshold

[9] The detainee's credit time for pretrial detention counts toward their sentence duration, first reducing any time they must spend incarcerated, then reducing any remaining time they would be monitored on release.

to determine whether detainees are eligible for split sentencing when evaluating its effects.

We evaluate how offering various levels of split sentences to the eligible detainee population affects the EM and jail population.[10] The detainees eligible for this sentencing policy are detainees whose sentence includes incarceration (jail and prison) and whose original sentence duration is 18 months or less. We draw detainees' location costs using the structural parameters estimated in Section 1.6, and calculate their costs for delaying and not delaying. They choose the option which has the lowest cost. If they choose to delay, their case length is drawn from F_D. Otherwise, it is drawn from F_N.[11]

	Total Yearly Reductions in:					
x	Case Length EM (Years)	Case Length Jail (Years)	Court Visits	Jail Sentences (Years)	Prison Sentences (Years)	Detention Costs in Jail (Millions of Dollars)
0.1	34.7	23.9	764	25.2	144.3	\$2.6
0.2	72.0	42.3	1490	46.2	285.1	\$4.6
0.3	106.9	52.9	2084	65.1	421.6	\$6.2
0.4	137.3	62.1	2600	82.4	547.5	\$7.5
0.5	158.0	64.7	2903	97.5	661.7	\$8.5
0.6	170.4	64.4	3061	110.5	760.1	\$9.1
0.7	174.3	67.0	3146	120.8	838.4	\$9.8
0.8	173.8	65.8	3124	128.4	895.6	\$10.1
0.9	173.4	66.0	3121	133.0	932.9	\$10.4

Table 1.6: **Yearly Benefits from Split Sentencing.** Case length is the yearly total of time people are detained in jail or on EM pretrial. Court visits are assumed to be spaced 28 days apart during the detainees' pretrail detention. Jail and prison sentence time represent the yearly reduced total time people would be incarcerated in those locations due to their sentence being split. Finally, housing costs in jail incorporate reductions in both pretrial case length and reduced jail time.

Post-Sentencing Effects. Split sentencing reduces detainees' sentence durations by a

[10] The Sentencing Guidelines suggest that no more than half of the sentence be split to release (U.S. Sentencing Comm'n, 2018) (i.e. x should be less than 0.5). We will vary x more broadly, across the full range of (0,1), to more completely evaluate split sentencing's impact.

[11] We assume that cases which resolved in 60 days or less are not likely to have delayed, and consequently their case lengths do not change. However, their sentence is still split according to x.

factor of x. These post-sentencing reductions, parameterized by x, are given in the columns "Jail Sentences" and "Prison Sentences" of Table 1.6. The reductions are presented yearly, so for $x = 0.5$, 97.5 detainee post-sentencing detention years would be removed from the jail each year. For the same x, 661.7 detainee years would be removed from the prison each year.[12] The associated housing cost savings for the prison, if similar to that of jail, would be approximately $34.5 million annually. The entirety of the reduction for prison is due to this post-sentencing effect. For jail, post-sentencing effects account for about 60% of the reduction in jail detention time.

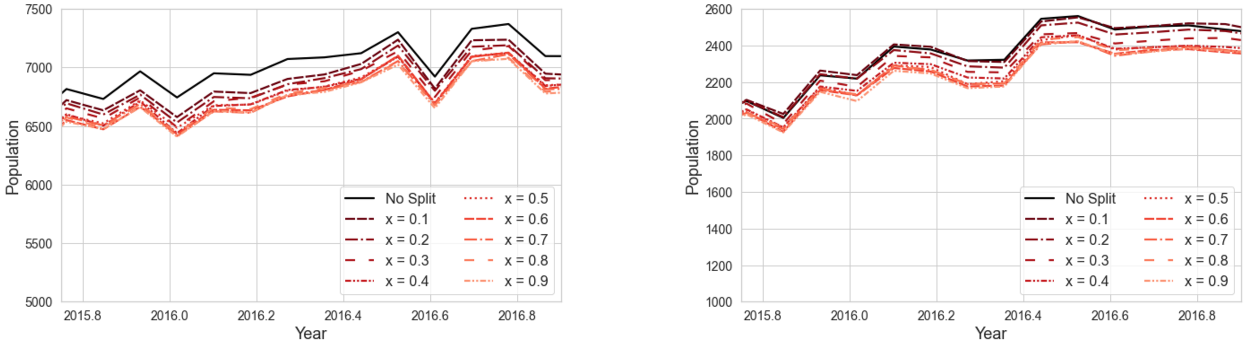
Pretrial Effects. Detainees delay less often when offered split sentences, which reduces aggregate pretrial detention time. Yearly pretrial case length reductions for EM and jail are given in the columns "Case Length EM" and "Case Length Jail" in Table 1.6. For $x = 0.5$, EM and Jail detention time would be reduced by 158.0 and 64.7 detainee years each year. Pretrial effects account for all of the reduction in EM incarceration time and approximately 40% of the reduction in jail time. In total, these pretrial reductions can reduce the load on the courts system significantly. For $x = 0.5$, court visits would be reduced by 2903 each year.

Outcomes on the Detained Populations. The plots in Figure 1.11 show the resultant drop in jail and EM populations from offering split sentencing to eligible detainees for various values of x[13]. The average reduction in jail and EM populations during the steady state period are plotted in Figures 1.12a and 1.12b.

Offering split sentencing significantly reduces the jail and EM populations. At $x = 0.5$, the jail population is reduced by 2.73% and EM is reduced by 5.68%, see Figures 1.12a and 1.12b. The reductions in the jail and EM populations are most significant for low level crimes—classes 4, 3, and A. This targeted effect is due to the restrictions on split sentencing,

[12] While we do not study the prison population, its population would also be reduced due to this decrease in total post-sentencing incarceration time.

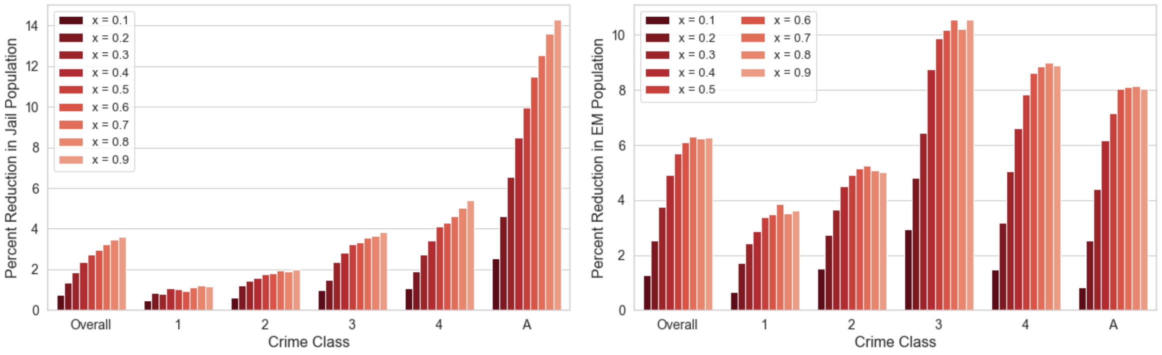
[13] As before, because we consider detainees who enter CCSO custody in 2015-2016, these appear non-stationary, but the detained population achieves a steady state in the period shown in Figures 1.11a and 1.11b. Implementation of this policy would resemble this steady state period.



(a) Jail Population Over Time

(b) EM Population Over Time

Figure 1.11: **Reductions in Jail and EM Populations Over Time via Split Sentencing.** Focusing attention on detainees who entered CCSO custody during 2015-2016, Figures 1.11a and 1.11b display the average predicted detainee populations over time. In these figures, we restrict our attention between September 2015 and December 2016, where steady state is achieved.



(a) Average Reductions in Jail Population

(b) Average Reductions in EM Population

Figure 1.12: **Average Reductions in Detainee Populations in Steady State via Split Sentencing.** These plot the average reduction in jail (Figure 1.12a) and EM (Figure 1.12b) populations due to reduced court visits during the steady state period between September 2015 and December 2016. "Overall" provides statistics for all data. The remaining provide statistics by crime class. Classes X and M are excluded because they are unaffected by the policy. Classes B and C are excluded from the figure as they represent very few detainees in the data.

not affecting detainees whose sentence represents a crime which is too severe. Reductions in the jail population directly reduce housing costs. At $x = 0.5$, the Cook County Jail could save \$8.5 million per year.

These benefits are substantial across a wide range of x values. However, most of the effects are seen once the detainees' sentences are mitigated by half (x = 0.5), and generally have diminishing returns after this point. This suggests that feasible applications of split sentencing which are well within the U.S. Sentencing Guidelines can realize most of its benefits.

	Yearly Reductions in:			
x	Turnarounds (Net)	Turnarounds (Percent)	Excess Housing Time "Dead Days" (Years)	Excess Jail Housing Costs Due to Turnarounds (Millions of Dollars)
0.1	72.0	3.5%	33.8	$1.8
0.2	136.3	6.7%	67.1	$3.5
0.3	189.8	9.3%	92.4	$4.8
0.4	233.7	11.5%	112.8	$5.9
0.5	257.9	12.6%	124.9	$6.5
0.6	272.9	13.4%	129.9	$6.8
0.7	275.7	13.5%	133.7	$7.0
0.8	277.1	13.6%	133.1	$6.9
0.9	276.8	13.6%	132.7	$6.9

Table 1.7: **Reduction in Turnarounds due to Split Sentencing Using Old Sentences as Metric.** Here, to determine whether or not someone is a turnaround, we use their original sentence duration, pre split-sentence.

The Impact of Split Sentencing on Turnarounds. Split sentencing reduces delaying behavior and thus detainees' case lengths. When measured against their original sentences, this reduces turnarounds; see Table 1.7. For example, for $x = 0.5$, turnarounds would be reduced by 12.6%, with an annual reduction of 124.9 years worth of dead days and $6.5 million of housing costs during that time. But split sentencing reduces the incarceration portion of detainees' sentences. Measured against this new sentence duration, $(1 - x)s$, this policy creates "new turnarounds," those who are incarcerated in jail longer than their split sentence would require. Table 1.8 reports these increases. Splitting sentences by half, $x = 0.5$, would create 2006.3 additional new turnarounds each year, a 98.4% increase. Measured against the detention portion of their sentences, this would make 946.7 years worth of new dead days per year. The cost associated with that time would be $49.4 million.

Detainees, in general, would be better off with this policy because of shorter case lengths

	Yearly Increases in:			
x	New Turnarounds (Net)	New Turnarounds (Percent)	Excess Housing Time "Dead Days" (Years)	Excess Jail Housing Costs Due to New Turnarounds (Millions of Dollars)
0.1	263.0	12.9%	114.2	$6.0
0.2	576.7	28.3%	251.8	$13.1
0.3	961.0	47.1%	429.2	$22.4
0.4	1451.7	71.2%	656.3	$34.3
0.5	2006.3	98.4%	946.7	$49.4
0.6	2691.0	132.0%	1317.6	$68.8
0.7	3457.2	169.6%	1787.2	$93.3
0.8	4269.3	209.4%	2390.6	$124.8
0.9	5034.8	247.0%	3156.9	$164.8

Table 1.8: **Increases in New Turnarounds due to Split Sentencing Using New Split Sentences as Metric.** Here, to determine whether or not someone is a turnaround, we use the detainees' new sentence durations post-split-sentence.

and less costly sentences. The jail would also save money. But it would increase instances that people are detained longer than their required detention time, precisely because their required detention time is reduced. To avoid this, split sentencing should be applied in tandem with operational improvements which reduce case lengths.

1.7.4 Reducing the Relative Cost of Imprisonment

As discussed above, delaying behavior is largely driven by the high cost of prison relative to jail. To reduce it and increase social welfare, we consider policies which may reduce the relative cost of prison to jail, and explore how that may effect pretrial detention. To compare dollars-to-dollars, we estimate society's cost for being housed in prison using payouts for wrongful imprisonment from the state.

Prison's perceived cost per unit time could be lowered relative to jail in quite a few ways. For example:

1. **Subsidize family visits to prison.** Some defendants may prefer jail to prison because

the Cook County Jail is located within the city, while prison is located far outside of the city. Consequently, family visits are possible in jail but not as easily in prison. If this changed, people may be more willing to serve their time in prison.

2. **Ensure quality living conditions in prison.** Living conditions in prison may need to be improved. As a result, fewer people would avoid it.
3. **Promote prison-oriented educational/remedial opportunities which aren't available in jail.** If there are educational opportunities in prison rather than jail (or more in prison than in jail), some people who suspect that they will get a prison sentence will opt to serve it there so they can get the benefit of the opportunities there. This has the additional benefit that prison sentences have definite beginning and end dates, allowing for more structured educational programs.[14]

To assign a dollar cost to prison, we consider payments from the state for wrongful imprisonment. As a conservative estimate, Illinois law stipulates that people wrongfully detained in prison for 5 years or less can be awarded up to $85,350, fourteen years or less can be awarded $170,000, and more than fourteen years can be awarded $199,150.[15] These adjust with cost of living increases, and include some additional funds for attorney fees, education, and job placement assistance. Federal guidelines suggest a higher rate of pay of $63,000 per year imprisoned, with an additional $63,000 for each year on death row.[16] Awarded amounts can be higher. For example, in 2020, three men were awarded approximately $280,000 per year of wrongful imprisonment.[17] In 2011, DNA tests exonerated five men who were awarded approximately $400,000 per year for nearly 20 years of wrongful imprisonment, totaling $40

[14] Educational programs exist within prisons at the moment and detainees can earn credit against their sentence for participating in them.

[15] https://www.ilga.gov/legislation/ilcs/fulltext.asp?DocName=070505050K8 Accessed on 11/5/2021.

[16] https://www.law.umich.edu/special/exoneration/Documents/CompensationByState_InnocenceProject.pdf Accessed on 11/15/2021.

[17] https://www.chicagotribune.com/news/breaking/chi-40m-wrongful-conviction-settlement-the-money-is-almost-beside-the-point-20140625-story.html Accessed on 11/5/2021.

million in their joint-wrongful conviction settlement.[18]

To model the social welfare gained from reducing the relative cost of prison, we make use of the costs listed above. We consider three different dollar-equivalent costs for time in prison: \$14,285 per year ("Low"—Illinois' payout rate), \$63,000 per year ("Medium"—the Federal payout rate), and \$400,000 per year ("High"—a representative large payout awarded to wrongfully imprisoned detainees). Costs for other housing location are scaled at the same proportion as the mean estimated costs in Section 1.6. That is, EM is 0.370/1.835 times less costly than prison, and jail is 1.378/1.835 times less costly than prison. We model the effect of the policies as reducing the mean cost of prison either halfway to that of jail, or entirely to that of jail. Then, in light of these reduced costs, we simulate detainee behavior to evaluate the effect on the criminal justice system.

	Total Yearly Reductions in:			
Cost Reduced to	Case Lengths (Years)	Court Visits	Jail Case Length (Years)	EM Case Length (Years)
Jail	193.5	2523	155.4	38.1
Halfway to Jail	96.6	1260	81.0	15.6

Table 1.9: **Yearly Detention Time Reductions due to Prison Cost Reduction.** Case length is the yearly total of time people are detained in jail or on EM pretrial. Court visits are assumed to be spaced 28 days apart during the detainees' pretrail detention. Jail and prison sentence time represent the yearly reduced total time people would be incarcerated in those locations. Note that a negative value indicates an increase in that statistic.

Reducing the perceived cost of prison reduces delaying behavior. This shortens case lengths, as seen in Table 1.9. By reducing the cost of prison to be (on average) equal to the cost of jail, case lengths would be reduced by 193.5 years annually, and the total number of annual court visits by approximately 2523. The majority of this contribution comes from people detained in jail pretrial whose incentive to delay was reduced. These case

[18] https://www.chicagotribune.com/news/breaking/chi-40m-wrongful-conviction-settlement-the-money-is-almost-beside-the-point-20140625-story.html Accessed on 11/5/2021.

length reductions pair with significant reductions in the overall jail and EM populations. By reducing the cost of prison to jail (halfway to jail), in steady state the jail population would be reduced by 2.00% (1.05%) and the EM population by 1.51% (0.65%). Note that for people sentenced to prison, a reduction in case length results in a commensurate increase in prison time. We find that this increase in time would be 129 years and 64 years annually for a reduction in prison cost all the way to jail and halfway to jail respectively.

	Total Yearly Reductions in:			
Cost Reduced to	Turnarounds (# People)	Percent of Turnarounds	Dead Years	Dead Years Housing Costs (Millions of Dollars)
Jail	932	41.5%	63.6	$3.3
Halfway to Jail	859	38.2%	32.5	$1.7

Table 1.10: **Yearly Turnaround Reductions due to Prison Cost Reduction.**

The reduced delaying behavior also significantly reduces turnarounds. Turnarounds would be reduced by 38.2% by reducing prison costs halfway to that of jail, and 41.5% by reducing them to be equal to that of jail. Their associated dead days would be also reduced. By reducing the cost of prison to that of jail, there would be an associated annual savings of $3.3 million in housing costs from dead days, and 63.6 years of dead days annually.

These reductions in (perceived) housing costs also increase social welfare. We display the change in perceived costs for people detained by the Cook County Sheriff's Office Table 1.11. As shown in the table, reducing the cost of prison to be equal to that of jail could, with a conservative estimate, reduce annual perceived detention costs by $11.9 million across detainees. At the high end, this could amount to over $335.5 million. Note that while we do not have data on people in prison who were not detained by the CCSO, these changes would effect them as well, and so these benefits are a conservative estimate.

Our results suggest that changes to prisons which lower their perceived detention cost could have significant positive effects for society. In the practical sense, delaying behavior

	Total Yearly Reductions in:		
	Disutility	Disutility	Disutility
Cost Reduced to	\$14,285/year	\$63,000/year	\$400,000/year
	(Millions of Dollars)	(Millions of Dollars)	(Millions of Dollars)
Jail	\$11.9	\$52.4	\$335.5
Halfway to Jail	\$6.0	\$26.6	\$169.0

Table 1.11: **Yearly Dollar-Equivalent Disutility Reductions due to Prison Cost Reduction.** Note that these estimates are only for people detained by the CCSO.

and pretrial case lengths could be decreased, and as a result, housing costs and court cost would be decreased. Additionally, by incorporating detainee preferences, social welfare could be increased.

1.8 Concluding Remarks

This paper is concerned with delays that cause excessively long pretrial detention, using data from the Cook County Jail to study this issue. This detention wastes detainees' time, costs taxpayers money, and needlessly burdens the courts system. In Section 1.5 we introduce a model of detainee behavior in which they can delay their cases to loosely control the balance of time they spend in their pretrial housing location and post sentencing location. Using this model, we estimate the structural parameters of detainees' location costs. We find that prison is most costly, followed by jail, then EM. In a series of four counterfactual analysis, we find that improving the operations of the courts and changing incentives to reduce delaying behavior, in tandem, can be effective tools to improve pretrial operations, save costs, and curb turnarounds.

CHAPTER 2

IDENTIFYING TURNAROUNDS BEFORE THEY OCCUR

2.1 Introduction

Turnarounds are detainees whose pretrial detention is longer than their sentence duration at prison. That is, while detained pretrial, their case lasts longer than the sentence they receive when it concludes. Time served pretrial counts toward time post-sentencing, so turnarounds have no remaining time to serve in prison. Nevertheless, these detainees are brought to prison, are booked, and are then immediately released. They "turn around" from the prison in Joliet, Illinois back to Chicago. These detainees are not compensated for the excess time they spent incarcerated.

People are held in jail pretrial if they can't meet their bond conditions. At the conclusion of their case they receive a disposition: either guilty or not-guilty. If guilty, they receive a sentence in the form of a location they must spend time, and the duration of time they must spend there. For more details, an operational view of the criminal justice system is outlined in Appendix 1.3.

Individualized intervention may be an effective tool for curbing turnarounds, resolving cases earlier than they might resolve naturally. In Cook County, the Sheriff's office has established the Justice Institute, a team "charged with enhancing the delivery of justice across every aspect of the Sheriff's office (jail, police, courts)."[1] This institute, or other parties like it, are able to address individual cases within the criminal justice system on a limited basis due to manpower constraints. Supplying this group with a list of detainees who have a high likelihood of being valuable to intervene allows them to bring stakeholders together, such as prosecutors, defenders, and judges, to help cases resolve before the detainee becomes a turnaround.

This paper is concerned with providing the Cook County Sheriff's Office (CCSO) a tool

[1] `https://www.cookcountysheriff.org/hardship-project/` Accessed on 3/30/2022.

to score detainees under their purview on the likelihood that they will eventually become a turnaround. We develop a model which is composed of two parts: a classification algorithm which takes detainees' attributes and case lengths into account to determine their probability of being a turnaround, and a Cox regression to determine the likelihood that a detainee's case will endure to a certain case length. In tandem, these two are used to identify detainees whose cases would be valuable to intervene upon.

We consider two intervention and scoring approaches. Both evaluate all detainees under the CCSO's purview on the first of each month. The first method considers the case that intervention happens immediately upon evaluating detainees, thus terminating their cases. This shows an ideal situation applying this tool. The second method adds a lead time to the intervention, to account for the time necessary to intervene on these cases. Implementing this model on sample data shows that targeted intervention using these models may successfully reduce turnarounds, their associated dead days, and the housing costs arising from that excess detention time. Without any lead time (and thus immediate intervention) we estimate this would reduce dead days by 10.1 years, saving $525,000 in housing costs for detainees identified each month. With a lead time of 2 weeks, these figures would be 6.8 years, saving $353,000 of housing costs each month. And with a lead time of 4 weeks, these figures would be 8.2 years, saving $429,000 of housing costs each month.

2.2 Data

Our dataset primarily consists of data retrieved from the Cook County Office Offender Management System (CCOMS) via the Cook County Sheriff's Office (CCSO). We also make use of data from the Illinois Department of Corrections (IDOC), the state's prison system. We use five data files which collectively provide the data fields listed in Table 2.1. Each contributes the following information: the beds file lists the detainees' pretrial housing location, the bonds file lists the detainees' bond type and amount, the courts file lists the detainees' court dates, the IDOC file has data about people detained in prison, which we use to determine sentence durations for detainees sentenced to prison, and the main CCOMS file provides all remaining data fields in Table 2.1.

In summary, our dataset provides information about detainees' cases from booking through sentencing. We focus attention on detainees booked in 2015 and 2016 and who remained under the CCSO's purview either on EM or in jail. They correspond to 98,882 rows in our dataset, each row representing one detainee's booking. This allows us to follow them through the completion of their detention in jail or on EM under CCSO custody. We use different portions of this data for various elements of this paper. We describe each data field in detail in Appendix A.3.

Detainee	**Housing**	**Case**	**Sentencing**
Inmate ID	Booking ID	Docket number	Sentence location
Criminal history	Booking date	Crime class	Sentence duration
Prison history	Pretrial housing location	Case length	Turnaround status
	Security classification	Court dates	
	Bond type		
	Bond amount		

Table 2.1: **Data Fields.**

In this application, we use k-fold cross validation to ensure our results are robust. Because our analysis is emulating the effect of the CCSO observing its current jail population, we slightly modify the traditional cross validation methodology to test our method on snapshots of the jail population. Within our dataset, the jail population reaches steady state after 5 months (May 2015). After it reaches steady state, we consider the first day of each month for the remainder of the dataset (i.e. June 2015 - December 2016). The jail population on that day constitutes the testing data for that day's cross-validation fold. The remaining detainees in the dataset constitute the training data. In each fold, we train the classification algorithm (see Section 2.3.1) and Cox regression (see Section 2.3.2) on the training data of that fold. We then test the models on the testing data. The results are recorded, the models are discarded, and the process is repeated on the next fold. The results reported in

Section 2.4 are the mean (and associated 95% confidence intervals[2]) of these 19 folds.

The classification algorithm, referred to as f in Section 2.3.1, makes use of the average interarrival times of detainees first five court visits, sentence duration, case length, pretrial housing location, crime class, security classification, criminal history, sentence duration, and sentence location. This resembles a case where available plea bargains are known by the intervention team. The Cox regression outlined in Section 2.3.2 uses the same data fields, excluding case length.

2.3 Model

In this section, we develop a model that identifies turnarounds before their case ends. If a detainee becomes a turnaround, we say $\tau = 1$, otherwise $\tau = 0$. This model makes use of the detainees' attributes, x, which are described in detail in Section 2.2, and include information like crime class, security classification, and EM status. It also makes use of the detainees' *current* time incarcerated, t, so that agents who are looking at the current detainee population (and thus cannot know the detainees' case lengths) can still identify detainees who are likely to be turnarounds.

Using these data, we model the detainee's probability of being a turnaround given their current case length and attributes, $P(\tau = 1 \,|\, x, t)$. We begin by decomposing this probability into two parts:

$$P(\tau = 1 \,|\, x, t) = \sum_{l=t}^{\infty} pr(\tau = 1 \,|\, x, l) pr(l \,|\, x, t).$$

That is, given the detainee's current case length, we look forward to each subsequent time period, determine the probability that their case terminates in that period, and if it did, the likelihood of being a turnaround given that case length.

We model these component probabilities in the following way. First, for any terminal case length, l, we use a classification procedure f to model the probability a detainee will

[2] These 95% confidence intervals represent the mean plus and minus 1.96 standard deviations of the 19 folds.

be a turnaround given their attributes. That is $f(\tau = 1 \,|\, x, l) = pr(\tau = 1 \,|\, x, l)$. Because this uses terminal case lengths, this classifier can be trained on historical data. Second, we will use a Cox proportional hazards model, g, to model the probability the detainee's case will terminate at time l given their attributes x and current time in jail t. That is, $g(l \,|\, x, t) = pr(l \,|\, x, t)$. Given f and g, $P(\tau = 1 \,|\, x, t)$ is modeled by:

$$P(\tau = 1 \,|\, x, t) = \sum_{l=t}^{\infty} f(\tau = 1 \,|\, x, l) g(l \,|\, x, t). \tag{2.1}$$

We describe the estimation of f and g in the subsequent Sections 2.3.1 and 2.3.2. We outline an extension of this model which incorporates a lead time to detection in Section 2.3.3.

2.3.1 Estimation of f

To estimate f, we train a classification algorithm on the data as described in Section 2.2, with the goal of classifying turnarounds. This classification algorithm makes use of x, and l as data, with turnaround status, τ as the target variable. This data is described in Section 2.2. The resulting model outputs a score for each detainee as a function of x and l which we will use to model their probability of being a detainee given those characteristics, $f(\tau = 1 \,|\, x, l)$.

To determine suitable hyperparameters, we partitioned all detainees into a train and test set, which represent 80% and 20% of the data, respectively, as in (Lee and Liu, 2003). Within this split, we ensure that the same proportion of detainees were labeled positively in the train and test sets.[3] We begin by training the models on the train set. Then, to evaluate the models, we use the trained models to classify the test set. The models are ranked by their AUC score. The list of candidate models and hyperparameters are given in Table 2.2. The best performing model and hyperparameter combination is used to model f. Logistic regression with a regularization parameter of 1 performed best with an AUC of .91, and is "known to predict well-calibrated probabilities" (Niculescu-Mizil and Caruana, 2005). The

[3] That is, we seperate positive and negative samples, then select 20% from each group to create the test set.

ROC curve associated with this classifier is given in Figure 2.1. What's especially important about this ROC curve is that the true positive rate is quite high for the most highly-scored detainees; this accuracy will contribute to good performance in Section 2.4. We fix the choice of classifier and hyperparameters, and use those when training our models in each fold of our k-fold cross validation (see Section 2.2 for a description of our cross validation methodology).

Model	**Hyperparameter**	**Range**
Logistic Regression	Regularization	[0.001, 1000]
Random Forest	# Estimators	[50, 200]
	Max Features	sqrt(# features), log_2(# features)
	Max Depth	[50, 110]
	Min Samples per Split	{2, 5}
	Min Samples per Leaf	{1, 2, 4}
	Bootstrap	{True, False}
Deep Neural Net	Hidden Layers	{5, 10, 50}
	Nodes per Layer	{10, 50, 100}
	Dropout Percentage	{.1, .3, .5}
	Loss Function	Binary Crossentropy
	Activation Function	relu
	Epochs	1, 5, 10
	Batch Size	{32, 64, 500, 1000}

Table 2.2: **Tested Models and Hyperparameters for f.** Logistic Regression and Random Forest classifiers are implemented in Scikit Learn. The Deep Neural Net classifier was implemented in TensorFlow's Keras. 100 combinations of the above hyperparameters were tested for each model.

2.3.2 Estimation of g

In each fold, we use the training detainees to model $g(l|x,t)$ (see Section 2.2 for a description of our cross validation methodology). To do so, we will estimate a survival function, $S(l|x) = 1-F(l|x)$ where F is the CDF of the detainee's case length given x. Discretizing the survival function, $g(l \mid x,t)$ is given by

$$g(l|x,t) = \frac{S(l|x) - S(l+1|x)}{S(t|x)}.$$

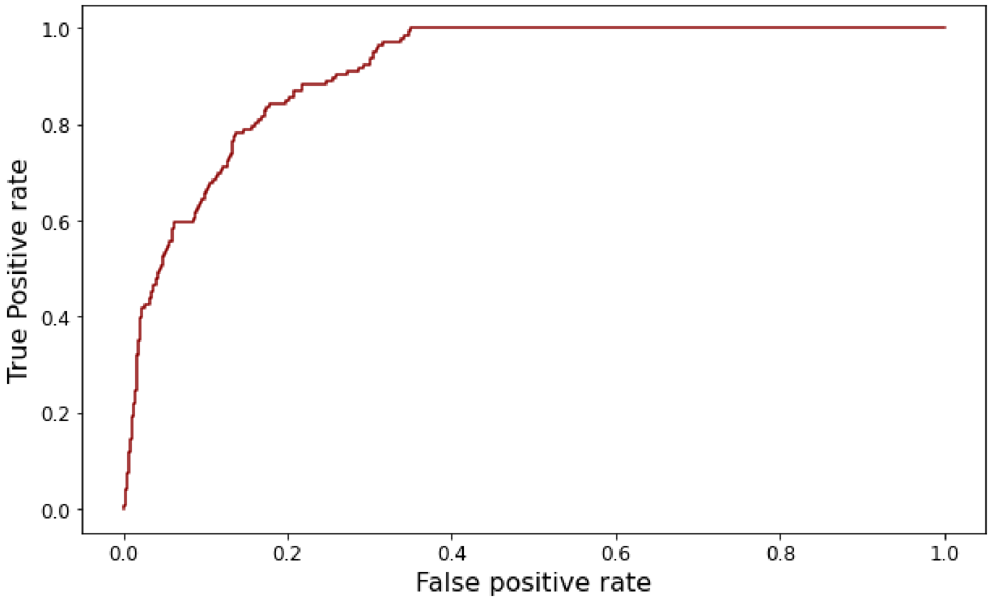


Figure 2.1: **ROC Curve of Fit Classification Function.** The AUC associated with this ROC curve is .911.

In this application, each period will represent one day.

The survival function can be constructed from the outputs of a Cox regression, which are

$$h(l|x) = h_0(l)e^{\sum_{j=1}^{p} \beta_j x_j}.$$

Namely, we will use the estimated parameters β and the cumulative baseline hazard function, $H_0(t) = \int_0^t h_0(u)du$. Given these, and noting that $S(t|x) = e^{-H(t|x)}$ (Rodríguez, 2007), we can decompose g into two interpretable parts: the baseline cumulative hazard function, which gives an understanding of any case's likelihood of ending at time t, and β's, which give an understanding of how the detainees' characteristics effect this baseline hazard. This decomposition is given in Proposition 3.

Proposition 3 *The function $g(l|x,t)$ can be expressed as a function of the baseline cumulative hazard function and estimated β's from the Cox regression in the following manner:*

$$g(l|x,t) = e^{-e^{\beta x}[H_0(l)-H_0(t)]} - e^{-e^{\beta x}[H_0(l+1)-H_0(t)]}.$$

Where the baseline hazard function is given by $H_0(t) = \int_0^t h_0(u)du$.

Proof. Begin by observing that the β's can be pulled out of the integral in the cumulative hazard rate function in the following manner:

$$
\begin{aligned}
H(t|x) &= \int_0^t h(u|x)du \\
H(t|x) &= \int_0^t h_0(u)e^{\beta x}du \\
H(t|x) &= e^{\beta x}\int_0^t h_0(u)du \\
H(t|x) &= e^{\beta x}H_0(t)
\end{aligned}
$$

Then, also noting that $S(t|x) = e^{-H(t|x_i)}$, we can substitute this into our definition of $g(l|x,t)$. Rearranging like terms gives the result.

$$
\begin{aligned}
g(l|x,t) &= \frac{S(l|x) - S(l+1|x)}{S(t|x)} \\
g(l|x,t) &= \frac{e^{-e^{\beta x}H_0(l)} - e^{-e^{\beta x}H_0(l+1)}}{e^{-e^{\beta x}H_0(t)}} \\
g(l|x,t) &= e^{-e^{\beta x}H_0(l)+e^{\beta x}H_0(t)} - e^{-e^{\beta x}H_0(l+1)+e^{\beta x}H_0(t)} \\
g(l|x,t) &= e^{-e^{\beta x}H_0(l)+e^{\beta x}H_0(t)} - e^{-e^{\beta x}H_0(l+1)+e^{\beta x}H_0(t)} \\
g(l|x,t) &= e^{-e^{\beta x}[H_0(l)-H_0(t)]} - e^{-e^{\beta x}[H_0(l+1)-H_0(t)]}
\end{aligned}
$$

■

We implement the Cox proportional hazards regression model using the statsmodels package in Python (Seabold and Perktold, 2010). In each fold, it is fit on the training detainees' data, as described in Section 2.2. For ease of interpretability, a sample baseline cumulative hazard for time in days t is given in Figure 2.2 (these functions differ little across folds). Detainees' case lengths are approximately exponentially distributed, and thus have an approximately constant hazard rate. As a result, the cumulative hazard function is approximately linear.

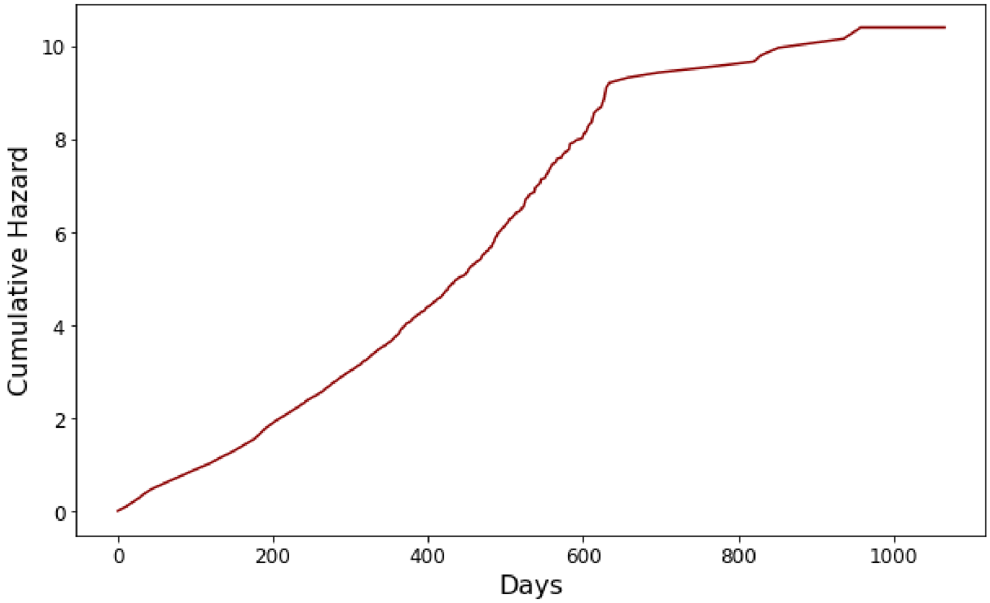


Figure 2.2: **Fit Baseline Cumulative Hazard Function.** Gaps in the domain of the estimated baseline cumulative hazard function, which are more common for long case lengths, are interpolated linearly. This represents one fit baseline cumulative hazard function from one fold.

The baseline cumulative hazard function is modified as a function of the detainees' attributes, x, and the fit parameters, β, from the Cox regression. A set of sample βs from one of the folds are given in the column "HR" in Table 2.3. They can be more easily interpreted via the first numerical column, "log HR", which gives the log hazard ratio. A positive number indicates that the increasing the associated variable, or the incidence of the categorical variable increases the hazard. A negative number indicates that the hazard decreases. As observed in the data, we note that increased sentence duration extends cases, as does being on EM, having a more severe crime class, and more severe security classification. Criminal history does not have a large effect on hazard. Larger average interarrival time of the detainees' first five court visits also increases hazard.

	log HR	log HR SE	HR	t	P> $\|t\|$	[0.025	0.975]
Average Interarrival Times	0.31	0.05	1.37	6.12	0.00	1.24	1.52
Sentence Duration	-12.75	1.05	0.00	-12.15	0.00	0.00	0.00
On EM	-0.57	0.02	0.56	-32.94	0.00	0.54	0.58
Class 2	-0.03	0.03	0.97	-1.01	0.31	0.92	1.03
Class 3	0.07	0.03	1.07	2.51	0.01	1.02	1.13
Class 4	0.38	0.02	1.46	15.52	0.00	1.39	1.53
Class A	0.73	0.03	2.07	27.65	0.00	1.97	2.18
Medium Security	0.47	0.04	1.60	10.69	0.00	1.47	1.75
Minimum Security	0.67	0.04	1.96	14.98	0.00	1.79	2.14
Low Criminal History	0.08	0.02	1.09	5.09	0.00	1.05	1.12
Medium Criminal History	-0.04	0.02	0.96	-2.06	0.04	0.93	1.00
High Criminal History	-0.04	0.03	0.96	-1.44	0.15	0.91	1.01
Supervision	0.88	0.04	2.41	19.91	0.00	2.21	2.63
Probation	0.30	0.03	1.35	8.93	0.00	1.26	1.44
Jail	0.05	0.02	1.05	2.17	0.03	1.01	1.10
Prison	-0.39	0.02	0.68	-16.90	0.00	0.65	0.71

Table 2.3: **Results of Cox Regression.** Categorical variables are one-hot-encoded, and the first category is dropped. For visual clarity, this is a sample set of fit parameters from one fold of the crossvalidation.

2.3.3 Incorporating Lead Time

In practice, intervening in cases may take some time. We extend our model to identify turnarounds before they occur, while focusing on detainees whose cases will end at least N days away. That way, when implemented, practitioners may be more confident that their efforts will have maximum effect.

We change our scoring metric to reduce the weight given to detainees who are predicted to have their cases end within N days. Denoted $\hat{P}(x, t, N)$, it is given by:

$$\hat{P}(x,t,N) = \sum_{l=t}^{t+N} f(\tau = 1 \,|\, x, l)(1 - g(l \,|\, x, t)) + \sum_{l=t+N+1}^{\infty} f(\tau = 1 \,|\, x, l) g(l \,|\, x, t). \tag{2.2}$$

As a result, this will highlight detainees who are predicted to be turnarounds, but whose cases are more likely to end after N days. We will use this formulation for scoring detainees for intervention with a lead time in Section 2.4.2.

2.4 Results.

The model outlined in Section 2.3 and extended in Section 2.3.3 allows us to determine detainees' probability of being a turnaround (or being a turnaround after some lead time N). We test our method using k-fold cross validation, as described in Section 2.2. The reported results are the means (and associated 95% confidence intervals) of each statistic across the 19 folds. Within each fold, we score each detainee in the test set. For Section 2.4.1, we use their probability of being a turnaround. In Section 2.4.2 we use their probability of being a turnaround with a lead of 7, 14, 21, and 28 days. Then, we evaluate the impact of intervening and resolving the top 10, 20,..., 100 highest scoring cases.[4]

2.4.1 Scoring Probability of being a Turnaround Without Lead Time

We score detainees by their probability of being a turnaround, which is given in Equation 2.1. In this case, we assume that intervention occurs immediately, resolving the cases when the intervention occurs.

The results of this scoring method given this intervention are given in Table 2.4. With 100 intervened cases, 58.3% of the intervened cases would be turnarounds, and intervention on a monthly basis would save 10.1 years of dead days per month, and save over $525,000 due to excess housing costs per month. The success rate of identifying turnarounds is relatively consistent across different values of intervened cases, so if resources were limited to only intervening on a fraction of 100 cases, the resulting benefits are approximately proportional to that fraction.

[4] Notice that as the number of intervened cases increases, the number of those cases which would have been turnarounds must also increase monotonically. However, the the average % of detainees who would eventually become turnarounds does not necessarily increase monotonically, as each group of 10 marginal detainees maybe identified with more or less accuracy than their proceeding groups. For example, consider a group of 10 detainees in which only one would eventually become a turnaround; this would increase the number of turnarounds intervened, but would likely decrease the average % of turnarounds, as 9 of the group would have "missed".

Intervened Cases	Would-be Turnarounds Intervened Upon		% of Interventions Which Would-be Turnarounds		Dead Days Reduced (In Years)		Reduced Dead Days Housing Costs ($1,000s of Dollars)	
10	5.3	[4.3, 6.2]	52.6	[43.4, 61.9]	1.0	[0.8, 1.3]	54.4	[40.5, 68.4]
20	10.7	[9.4, 12.1]	53.7	[47.0, 60.3]	2.2	[2.0, 2.4]	113.7	[103.3, 124.2]
30	17.6	[15.8, 19.5]	58.8	[52.6, 64.9]	3.3	[3.1, 3.6]	174.7	[160.9, 188.5]
40	23.7	[21.4, 25.9]	59.2	[53.6, 64.8]	4.4	[4.0, 4.9]	232.1	[209.7, 254.5]
50	29.4	[27.0, 31.8]	58.7	[54.0, 63.5]	5.5	[4.9, 6.2]	288.9	[256.6, 321.3]
60	34.9	[32.1, 37.8]	58.2	[53.4, 63.1]	6.7	[5.8, 7.6]	351.0	[304.4, 397.6]
70	41.0	[37.6, 44.4]	58.6	[53.7, 63.4]	7.8	[6.7, 8.8]	405.0	[350.8, 459.2]
80	46.5	[42.8, 50.1]	58.1	[53.5, 62.7]	8.3	[7.2, 9.4]	433.0	[374.2, 491.8]
90	52.1	[47.9, 56.2]	57.8	[53.2, 62.4]	9.2	[8.0, 10.4]	479.1	[415.2, 543.1]
100	58.3	[53.5, 63.0]	58.3	[53.5, 63.0]	10.1	[8.8, 11.4]	525.7	[457.3, 594.0]

Table 2.4: **Monthly Reductions from Intervention Without Lead Time.** Averages are taken across 19 folds. Figures in brackets represent 95% confidence intervals. Housing costs during dead days are assumed to be $143 per day. (Board of Commissioners of Cook County, 2021)

2.4.2 Scoring by Probability of being a Turnaround With Lead Time

Intervened Cases	Would-be Turnarounds Intervened Upon		% of Interventions Which Would-be Turnarounds		Dead Days Reduced (In Years)		Reduced Dead Days Housing Costs ($1,000s of Dollars)	
10	4.6	[3.8, 5.3]	45.8	[38.1, 53.5]	0.5	[0.3, 0.7]	26.4	[16.4, 36.4]
20	9.1	[7.7, 10.5]	45.5	[38.7, 52.3]	1.1	[0.8, 1.3]	56.9	[44.2, 69.7]
30	13.2	[11.3, 15.2]	44.0	[37.5, 50.6]	1.7	[1.3, 2.0]	87.1	[69.1, 105.1]
40	18.7	[16.4, 21.1]	46.8	[40.9, 52.8]	2.3	[2.0, 2.6]	120.3	[104.6, 136.1]
50	24.1	[21.3, 26.9]	48.2	[42.6, 53.8]	3.0	[2.6, 3.4]	157.0	[135.7, 178.4]
60	29.2	[25.7, 32.6]	48.6	[42.8, 54.4]	3.5	[3.0, 4.0]	182.8	[155.6, 210.0]
70	33.9	[30.2, 37.7]	48.5	[43.1, 53.9]	4.1	[3.6, 4.7]	215.9	[188.6, 243.3]
80	38.6	[34.3, 43.0]	48.3	[42.9, 53.7]	4.6	[4.1, 5.2]	242.6	[213.5, 271.8]
90	42.9	[38.4, 47.4]	47.7	[42.7, 52.7]	5.2	[4.6, 5.9]	273.9	[241.9, 305.8]
100	47.4	[42.7, 52.1]	47.4	[42.7, 52.1]	5.8	[5.1, 6.5]	300.5	[264.2, 336.8]

Table 2.5: **Results from Scoring with 7 Day Lead.** AAverages are taken across 19 folds. Figures in brackets represent 95% confidence intervals. Housing costs during dead days are assumed to be $143 per day. (Board of Commissioners of Cook County, 2021)

We score detainees by their probability of being a turnaround after a lead of N days, as given in Equation 2.3.3. We also assume that intervention occurs after N days, resolving the

Intervened Cases	Would-be Turnarounds Intervened Upon		% of Interventions Which Would-be Turnarounds		Dead Days Reduced (In Years)		Reduced Dead Days Housing Costs ($1,000s of Dollars)	
10	4.6	[3.7, 5.4]	45.8	[37.4, 54.2]	0.6	[0.4, 0.8]	29.0	[18.8, 39.3]
20	9.3	[8.1, 10.5]	46.6	[40.6, 52.5]	1.2	[1.0, 1.4]	62.5	[51.0, 74.0]
30	14.1	[12.1, 16.0]	46.8	[40.4, 53.2]	1.8	[1.5, 2.0]	92.3	[77.7, 106.9]
40	19.4	[16.8, 22.0]	48.6	[42.1, 55.0]	2.5	[2.1, 2.8]	128.5	[110.4, 146.5]
50	24.6	[21.7, 27.5]	49.2	[43.4, 54.9]	3.2	[2.8, 3.7]	169.0	[145.2, 192.9]
60	29.8	[26.4, 33.1]	49.6	[44.1, 55.2]	3.8	[3.3, 4.3]	199.6	[174.4, 224.8]
70	35.0	[31.1, 38.9]	50.0	[44.4, 55.6]	4.6	[3.9, 5.2]	237.5	[205.4, 269.5]
80	39.7	[35.1, 44.3]	49.7	[43.9, 55.4]	5.4	[4.5, 6.2]	280.7	[236.5, 324.9]
90	45.2	[40.1, 50.3]	50.2	[44.6, 55.9]	6.2	[5.3, 7.1]	324.4	[278.1, 370.7]
100	50.2	[44.7, 55.7]	50.2	[44.7, 55.7]	6.8	[5.8, 7.7]	353.6	[305.0, 402.2]

Table 2.6: **Results from Scoring with 14 Day Lead.** Averages are taken across 19 folds. Figures in brackets represent 95% confidence intervals. Housing costs during dead days are assumed to be $143 per day. (Board of Commissioners of Cook County, 2021)

Intervened Cases	Would-be Turnarounds Intervened Upon		% of Interventions Which Would-be Turnarounds		Dead Days Reduced (In Years)		Reduced Dead Days Housing Costs ($1,000s of Dollars)	
10	4.7	[3.9, 5.6]	47.4	[39.2, 55.6]	0.5	[0.3, 0.7]	27.4	[16.4, 38.4]
20	9.6	[8.3, 10.9]	47.9	[41.4, 54.4]	1.2	[1.0, 1.5]	63.5	[50.3, 76.6]
30	14.4	[12.5, 16.2]	47.9	[41.6, 54.2]	2.0	[1.6, 2.3]	103.1	[85.9, 120.2]
40	19.6	[17.1, 22.1]	49.1	[42.8, 55.4]	2.8	[2.4, 3.2]	147.2	[125.5, 169.0]
50	24.9	[21.9, 28.0]	49.9	[43.8, 55.9]	3.6	[3.1, 4.1]	189.2	[162.4, 215.9]
60	30.6	[27.2, 34.0]	51.1	[45.4, 56.7]	4.6	[4.0, 5.3]	242.3	[206.9, 277.6]
70	36.2	[32.1, 40.3]	51.7	[45.9, 57.6]	5.4	[4.6, 6.1]	281.4	[242.3, 320.5]
80	41.2	[36.5, 45.9]	51.5	[45.6, 57.4]	6.0	[5.2, 6.9]	315.8	[271.2, 360.3]
90	46.3	[41.2, 51.3]	51.4	[45.8, 57.0]	6.9	[5.9, 7.9]	360.0	[310.2, 409.8]
100	51.7	[46.4, 57.1]	51.7	[46.4, 57.1]	7.6	[6.7, 8.6]	397.6	[348.3, 447.0]

Table 2.7: **Results from Scoring with 21 Day Lead.** Averages are taken across 19 folds. Figures in brackets represent 95% confidence intervals. Housing costs during dead days are assumed to be $143 per day. (Board of Commissioners of Cook County, 2021)

cases when the intervention occurs.

The results of this scoring method given this intervention are given in Table 2.5-Classification Results 4. In general, classifying turnarounds with lead time does not perform as well as classifying without lead time. However, the performance remains strong, allowing for sig-

Intervened Cases	Would-be Turnarounds Intervened Upon		% of Interventions Which Would-be Turnarounds		Dead Days Reduced (In Years)		Reduced Dead Days Housing Costs ($1,000s of Dollars)	
10	4.7	[3.8, 5.5]	46.8	[38.2, 55.5]	0.5	[0.3, 0.7]	26.6	[16.8, 36.4]
20	9.8	[8.4, 11.2]	48.9	[42.0, 55.9]	1.3	[1.0, 1.6]	69.5	[54.6, 84.3]
30	14.9	[13.0, 16.8]	49.6	[43.4, 55.9]	2.3	[1.9, 2.7]	120.5	[98.9, 142.1]
40	20.3	[17.7, 22.8]	50.7	[44.3, 57.0]	3.2	[2.7, 3.6]	165.2	[141.0, 189.3]
50	25.6	[22.6, 28.7]	51.3	[45.1, 57.4]	4.3	[3.6, 4.9]	222.2	[188.4, 256.1]
60	31.4	[28.0, 34.9]	52.4	[46.6, 58.1]	5.2	[4.5, 5.9]	270.1	[234.2, 306.0]
70	36.4	[32.2, 40.6]	52.0	[46.0, 58.0]	5.8	[5.0, 6.7]	304.2	[261.1, 347.2]
80	41.9	[37.2, 46.6]	52.4	[46.4, 58.3]	6.6	[5.8, 7.5]	346.4	[302.0, 390.9]
90	47.4	[42.3, 52.5]	52.7	[47.0, 58.4]	7.5	[6.6, 8.4]	390.3	[342.9, 437.7]
100	52.5	[47.0, 57.9]	52.5	[47.0, 57.9]	8.2	[7.2, 9.2]	429.7	[377.1, 482.4]

Table 2.8: **Results from Scoring with 28 Day Lead.** Averages are taken across 19 folds. Figures in brackets represent 95% confidence intervals. Housing costs during dead days are assumed to be $143 per day. (Board of Commissioners of Cook County, 2021)

nificant amounts of turnarounds to be identified in advance, saving unnecessary dead days and housing costs. And, the magnitude of the reductions increases as lead time is increased, which suggests that implementing this policy with the delays associated with intervention is feasible. To give a sense of results, we summarize the consequence of intervening on 100 cases each month after 7 days (28 days). Of these, 47.4% (52.5%) would be turnarounds, resulting in 5.8 (8.2) years of reduced dead days. As a result, the taxpayer would save $300,000 ($429,000) in excess housing costs during those dead days each month.

2.5 Concluding Remarks

Turnarounds are costly for the taxpayer: they are detained pretrial longer than the total sum of time their sentence would require them to be detained at all. At $143 per day, these excess housing costs can become large. Targeted intervention to help resolve the cases of detainees likely to become turnarounds can help curb these costs, reducing unnecessary dead days for those defendants in the process.

In this paper, we develop a model which can identify turnarounds before they occur, leveraging both their attributes and their current case length. We extend this model to incorporate a lead time, to account for the time it would take to intervene on the cases and help them get resolved. Implementing this model on sample data shows that targeted intervention using these models may successfully reduce turnarounds, their associated dead days, and the housing costs arising from that excess detention time. Without any lead time (and thus immediate intervention) we estimate this would reduce dead days by 10.1 years, saving $525,000 in housing costs for detainees identified each month. With a lead time of 2 weeks, these figures would be 6.8 years, saving $353,000 of housing costs each month.

Individualized intervention on cases for applications beyond turnarounds is an interesting avenue for future research. For this application, alternative scoring metrics could be useful, including some which attempt to predict the magnitude of predicted turnarounds' dead days ahead of time, to prioritize those cases and save on housing costs. Additionally, this could be approached from a utility standpoint, attempting to reduce the disutility of excess time incurred by the detainees due to dead days.

CHAPTER 3

ON PEOPLE'S UTILITY OVER WAIT FUNDAMENTALS AND INFORMATION

3.1 Introduction

"Waiting in line sucks," reviewed one participant after completing our experiment. We *all* know the feeling. But despite decades of queueing research, we don't have a good explanation of *why*. Modeled in the traditional ways, our participant's response is a bit puzzling. They reported a compensation of \$20/hour, more than triple their average wage on MTurk (Hara et al., 2018). Shouldn't this wait have been an unmitigated success? What aspects of people's disutility are we missing? This paper explores the connection between a prospective wait's fundamentals and people's perceived disutility. Further, it demonstrates that the way a wait is presented also significantly affects utility.

The literature contains little empirical research on how people assess prospective waits. Life experience provides some intuition, of course—people dislike long waits, variable waits, and long lines. To invoke such preferences, scholars often reference David Maister's *The Psychology of Waiting Lines* (1985). While the article contains many keen observations, it provides no formal evidence: no data or statistical tests; no analytical model; no attempt to measure the relative importance of the phenomena described. And the subsequent decades have brought almost no progress on this front. "Evidently, it exceeds the scope of most individual research papers to identify individual-level behavioral (ir)regularities, explain their underlying drivers, and establish their implications for system-level behavior," concede Allon and Kremer (2018, p. 325) in a recent chapter on queueing.

In this paper, we estimate a series of utility functions from individual-level preferences over candidate waits. We explore the space of wait fundamentals, simultaneously manipulating reward amount, mean wait time, wait variance, and line length. To do so, we adopt conjoint analysis, a straightforward yet powerful technique from the marketing literature that is rarely used in behavioral operations. We also manipulate the way in which we *describe* wait fundamentals (e.g., as a queue with 6 people vs. an aggregate wait of about 10 minutes) and find that this "information layer" moderates the relationship between fundamentals and utility in consistent and meaningful ways. Our resulting utility functions can be incorporated directly into more complex models of service systems. Doing so could add nuance and credibility to analytical queueing models.

Our results stem from a simple experiment: Participants see a panel of waiting scenarios from which they choose their favorite. The scenarios are meant to represent "everyday" waits—those with moderate duration (less than 20 minutes) for a modest reward (worth about five dollars). We achieve incentive-compatibility—an elusive gold standard for conjoint studies (Ding, 2007)—by conditioning participants' rewards on enduring one of their chosen waits in real time.

The fitted utility functions confirm the intuitive—people like larger rewards, and shorter, less-variable waits. But they go further by revealing how people trade off these competing interests. In most queues, wait time variance is just as consequential as mean duration. This implies that people are risk averse with their time. And while risk aversion is frequently incorporated into models throughout operations (Davis, 2018), economics (Pratt, 1964), and finance (Markowitz, 1959), it "has been ignored in most of the past studies on queueing models" (Wang and Zhang, 2018, p. 1198). Some studies have provided support for the existence of wait-induced risk aversion (Leclerc et al., 1995; Pazgal and Radas, 2008). Our utility estimates measure its magnitude relative to mean wait time and reward—a critical step that makes our results accessible to modelers. That said, we understand that incorporating risk aversion will be infeasible in many models due to mathematical intractability. But this is no reason to eschew study of risk aversion over waits altogether.

We find that different presentation schemes influence people's utility, independent of wait fundamentals. This is important because waits have no "plain" or "default" presentation. A busy restaurateur could tell arriving guests their estimated wait time or, alternatively, their position in the queue. Both schemes are reasonable, yet customers may evaluate the two prospects (or rather, the two presentations of the same prospect) quite differently. We study several such schemes in our eight experimental treatments. They have identical fundamental characteristics but differ in how they are presented. We find substantial differences between two common regimes: per-person lines (as in grocery stores and banks) and aggregate countdown timers (as in ridesharing apps and pizza delivery). People are about twice as sensitive to a wait's duration in the latter. To mimic the fact that real world waits often have no posted numerical information about wait duration, we induce non-quantitative beliefs about wait time in some treatments by posting sample draws from the wait's distribution. People defer to this non-quantitative "sense" of a wait—to their detriment—even when they have precise numerical information available. And we find that when precise numerical information is available, it does little to prime subjects to be more sensitive to the posted characteristics.

Finally, we distill our results into a set of managerial insights in the spirit of Maister (1985). We hope that our paper will be the first of many data-driven refinements to his influential work. And for researchers and practitioners who wish to adopt a utility model which incorporates these insights, we provide a simple, parsimonious version that organizes all of our results.

The remainder of this paper is organized as follows: Section 3.2 introduces our framework for understanding wait fundamentals, wait information, and people's utility over the two. Section 3.3 reviews relevant literature. Section 3.4 details our experiment. Section 3.5 covers the implementation of our experiment and the benefits of hosting it on Amazon's MTurk. Section 3.6 briefly outlines our methods for maximum likelihood estimation. Section 3.7 reviews our estimation results and describes the impact of our eight informational treatments. Section 3.8 illustrates the managerial insights to be gained from our work. Finally, we offer concluding remarks in Section 3.9.

3.2 A Waiting Framework

A wait has three fundamental elements: a duration, a reward, and a context. Each is an established area of research in its own right. Queueing theorists study how systems dictate a wait's duration. Marketers and economists study how rewards like products and services influence a customers' utility. And psychologists study how various waiting contexts (e.g., priority boarding at the airport) affect people's emotions and behavior. Given the attention they've received individually, these fundamentals—together—may seem sufficient to describe people's utility over waits. But they aren't. Our experiments demonstrate that information about a wait—its presentation and people's prior beliefs about it—influences people's utility, independent of wait fundamentals.

Wait fundamentals are just incomprehensible for most people. Consider the seemingly simple example of a coffee shop. Would you wait in a five-person queue with i.i.d. inverted beta service times and parameters (10.7, 3.2)? If you only have 10 minutes to spare, will you get your drink in time? Does the menu item reading *K. Yirgacheffe (2,150m) AA; V60, 20s bloom* provide any objective information about its value as a reward? No, you wait because the shop has a nice atmosphere.

People depend on intermediary information to make sense of these complex wait fundamentals. We are well aware of this signaling game over reward fundamentals: we call it *marketing* or *advertising*. A process of similar complexity sits atop the communication of waits. We dub this the *information* layer (shown in Figure 3.1). Firms choose a (necessarily incomplete) presentation scheme, and customers fill in the blanks with their prior beliefs. People make informed decisions about whether they will join a wait or balk by estimating—non-quantitatively—the wait duration, the reward value, and the impact of the context. Customers do not compute the convolution of five i.i.d. inverted beta distributions; they just consider the five-person line and conclude that a fancy coffee at an inviting shop is worth a few minutes. We aim to study this decision.

Our results provide evidence that information about a wait influences people's utility,

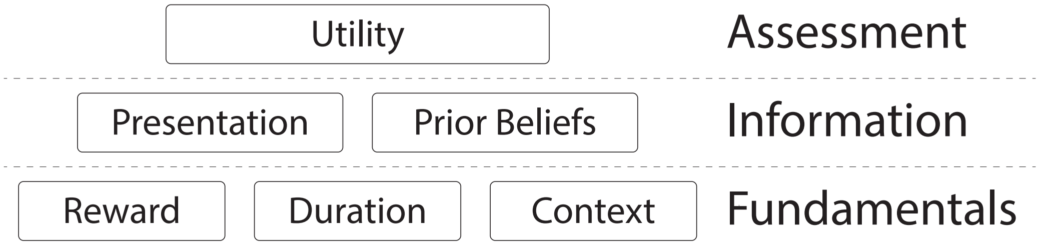


Figure 3.1: **A Framework for Organizing People's Utility Over Waits.**

independent of the wait's fundamentals. Holding fundamentals constant, we conduct a panel of experiments that differ only in presentation scheme and induced prior beliefs. Each treatment specifies wait fundamentals completely and accurately. Yet participants behave differently when the *same wait* is presented in different ways. Practitioners can use these patterns to influence behavior and system dynamics. For theorists, our results provide an empirically tuned utility specification. And finally, behavioral researchers can use the framework presented in Figure 3.1 to organize future work.

To reinforce the importance of wait information, we close this section with two short examples that demonstrate its impact intuitively.

A presentation scheme must be chosen, and it matters.

When customers call an Uber, they receive a single estimate of total trip duration. Uber chooses this simple presentation because it's useful for their customers. The underlying wait fundamentals are too complex: They arise from a constantly updating array of customer and driver locations underneath a complex matching and pricing algorithm. Even *if* Uber could communicate all this information to customers, it would be useless. People aren't characters in *The Matrix*—they need clear presentation schemes, not waterfalls of data.

Prior beliefs matter because posted information is rare.

Imagine finding a kid's lemonade stand on a hot day. Their sign reads: "Fresh Lemunade ♡ Squezed Wile You Wait." Neither you nor the kid knows anything about your service time distribution. Regardless, you choose to join the queue—it is a hot day, after all. Prior beliefs provide enough distributional knowledge of quality, price, and production time to let you assess the wait. To illustrate the power of these beliefs, imagine coming upon another, nearly identical child-run food stand advertising "Raw Sooshi ♡ Rolled Wile You Wait." You would be wise to balk at this offer. Once again, it's a hot day.

3.3 Related Work

The matter of how people assess waits is important for practitioners and scholars, alike. After all, "in many service systems customers behave strategically" (Ata and Peng, 2018, p. 163). *Strategic* customers make decisions in pursuit of their own objectives. But what do people "want" when it comes to waiting? Our study attempts to systematically explore this question.

Theorists' efforts to incorporate strategic customers into queuing models began a half-century ago (e.g., Naor 1969) and continue to this day—see Ward (2012) for a survey. Early work connected balking behavior to features that "characterize the queue, such as queue length and waiting time"(Haight, 1957, p. 360). Decades on, these two features are still the most common to appear in behavioral descriptions of waiting (Allon and Kremer, 2018). We build on this literature by incorporating people's preferences over the space of wait fundamentals and information structures.

Our paper contributes to a sparse literature on people's evaluation of prospective waits. "This is an area that needs to be explored further via different experimental designs" (Akşin et al., 2019, p. 25). More is known about people's behavior *during* waits. Experimental studies have linked waiting experience to reneging (Akşin et al., 2019; Ülkü et al., 2020); field studies, in turn, have linked reneging to long-run firm profitability (De Vries et al., 2018). The most established context for empirical queueing work—telephone call centers (Gans et al., 2003; Jouini et al., 2011; Akşin et al., 2017; Yu et al., 2017)—tends to focus more on aggregate customer behavior than on individual microeconomic decisions. Associated research considers the equilibrium behavior of customers who strategically time their arrivals to a congested system (Rapoport et al., 2004; Seale et al., 2005). And the behavior of customers with private information about wait fundamentals can serve as a signal to others (Veeraraghavan and Debo, 2009; Kremer and Debo, 2016).

3.3.1 Risk Aversion

Though risk sensitivity is a foundational principle of financial (Markowitz, 1959) and economic (Pratt, 1964) theory, it is often left out of queueing models, typically as a matter of mathematical tractability. There are few papers devoted to modeling risk-sensitive customers in queueing systems. In one, Wang and Zhang (2018) argue that "customer risk attitude can be one of the most critical factors affecting the performance of a stochastic service system" (p. 1198). However, most "existing models do not consider how customers include uncertainty... to decide which [queues] to join" (Delgado et al., 2011, p. 1720). Our results suggest that people are risk averse with their time and that this risk aversion has a substantial influence on customer behavior.

Kumar et al. (1997) conduct an experimental study of wait time guarantees using a risk-sensitive theoretical framework much like our own. They assume that when a "customer arrives at a service facility, he or she has some prior beliefs about the distribution of service times..." (p. 298). We induce prior beliefs explicitly in several of our treatments. Their model, like ours, involves normally distributed service times. Under negative exponential utility, they find that the "expected utility from waiting time reduces to a modified mean–variance model" (p. 300). We also adopt a mean–variance specification when modeling utility but focus on a different stage of the waiting process: They investigate the impact of wait time guarantees on customer satisfaction during the wait. We study the impact of fundamentals and information on customers' utility before the wait.

Current evidence of people's risk preferences over wait duration points toward risk aversion. Leclerc et al. (1995) and Weber and Milliman (1997) find that survey respondents avoid risky options when asked to choose among hypothetical travel delays. Furthermore, de Palma and Picard (2005) find that about sixty percent of travelers are risk averse while a third are "risk lovers." Leclerc et al. (1995) go on to offer a plausible explanation for risk aversion in waiting (i.e., lost time) contrary to the phenomenon of risk-*seeking* behavior regarding lost *money* (Kahneman and Tversky, 1979): The inability to store or transfer time induces risk aversion when considering prospective waits.

3.3.2 Wait Presentation and Bounded Rationality

Queues are going virtual. Google now provides real-time expected wait durations in its search results. Yelp lets customers "join a restaurant's waitlist from the comfort of [their] own home" (Yang, 2019).

Customers today often encounter online *descriptions* of queues rather than the queues themselves. But there is little research linking a wait's presentation to its appeal. Hui and Tse (1996) insert a 5–15 minute lag in software that participants are "evaluating" and present it as either a delay of about X minutes or as a queue with Y customers. They find that "duration information may not be the most effective tool to minimize consumer dissatisfaction" (p. 89). While their results are based on post-wait satisfaction measures, our estimates of prospective wait preferences would agree: We find that the per-minute wait penalty is roughly twice as high when people have duration (as opposed to queue) information. Similarly, Batt and Terwiesch (2015) argue that customers facing a queue-framed wait "might naively estimate the waiting time to be short and thus join a queue that they would not otherwise join if they were informed about the expected waiting time" (p. 39).

The literature on bounded rationality offers some explanation for why people might evaluate a queue-framed wait differently than its duration-framed analog. Huang et al. (2013) interpret "bounded rationality as the incapability of estimating expected waiting time in a service setting" (p. 264). Customers may form inaccurate beliefs about a wait's expected duration from a small sample of past experiences (Tong and Feiler, 2017) or based on anecdotes from others (Huang and Chen, 2015). Of course, bounded rationality primarily affects queue-framed waits; *no* computation is required when the expected duration is explicitly posted. This explains our finding that the preferences we measure in duration-framed waits are more consistent with the typical behavior of MTurk workers (as reported in other studies) while preferences in queue-framed waits appear to be distorted.

3.3.3 Implementing Waits in Controlled Experiments

To study waiting in the lab, researchers might want to subject participants to a large number of waits. But this approach is too slow. One solution is to have participants respond to "realistic hypothetical scenarios" (Leclerc et al., 1995, p. 111). Allon and Kremer (2018) warn, however, that "hypothetical snapshots ... may not capture well the dynamics of real wait time experiences" (p. 330). A second solution is to simulate waits (i.e., to resolve them instantly with a waiting cost assessed per unit of "time"). This transforms waiting problems into optimization problems: "One alternative is often better than the other. However, it requires cognitive abilities to correctly evaluate" them (Conte et al., 2016, p. 2). Some studies use a "hybrid" approach where participants actually wait but are also assessed a monetary waiting "fee" (Pazgal and Radas, 2008; Akşin et al., 2019, Studies 1 and 2). But by conflating time and money, hybrid waits make it difficult to identify preferences between the two.

"Real waits" allow for the most direct study of people's preferences by avoiding induced waiting costs. Instead, they make people wait in real time imposing no costs other than their natural distaste for waiting. Ülkü et al. (2020) achieve this standard across many contexts, using the real wait as a moderator to understand the relationship of waiting time and purchase intentions. In Study 3, Akşin et al. (2019) eliminate induced waiting costs so that participants can "make their own tradeoff between the monetary cash reward...and their own waiting cost" (p. 14). Our study uses a real wait to achieve incentive compatibility.

3.3.4 Conjoint Analysis

Conjoint analysis may be the "most significant development in marketing research over the last 30 years" (Rao, 2014, p. 1). The choice-based version works by describing many potential variants of a product (called *profiles*), arranging these profiles into choice sets, and then asking participants to select their favorite profile from each set. Maximum likelihood estimation allows researchers to impute the participant's utility as a function of product features (called *attributes*). With this imputed utility function, researchers can model the participant's response to *any* product variant.

The strength of conjoint analysis is its ability to generate a utility model over fundamental product attributes without having to ask contrived questions about the attributes directly. A researcher might be interested in the product features that drive apple sales. But asking about these drivers explicitly—"How important is the height of an apple, independent of its circumference, roundness, and color?"—is unlikely to prove illuminating. Conjoint analysis provides an elegant solution: the researcher can present participants with a set of apples, let them choose their favorite, and then mathematically *impute* the importance of an apple's stature.

The technique, however, has a substantial limitation: establishing incentive compatibility is difficult because the number of potential profiles is typically large. It is unrealistic for researchers to have an apple of every shape, size, and color available to incentivize choices (and even more so when each *Apple* is a $1,000 phone). "Almost without exception," Ding (2007) writes, "conjoint data have been collected in hypothetical settings that offer no consequences for participants' decisions" (p. 214). Proper incentivization, however, improves the reliability of results (Katok, 2018). We achieve incentive-alignment by making participants experience one of their selected profiles as a real wait.

Many questions of interest to the operations community are amenable to straightforward conjoint incentive-alignment using a technique like ours. Whereas marketing research explores preferences over *products* and *services* (which are hard to realistically emulate or bring into the lab), experiments in operations focus on *decision making* and *trade-offs* (which can more easily be linked to meaningful consequences). Furthermore, the optimal design (Sauré and Vielma, 2019) and analysis (Chen and Hausman, 2000) of conjoint studies provide problems of theoretical interest. As such, we see broader adoption of conjoint methods as a fruitful direction for operations researchers.

3.4 Experiment

Our experiment is a simple conjoint analysis study designed to measure how people evaluate prospective waits. Participants see a set of three potential waiting scenarios and choose their most preferred. This process repeats twenty times, and we call this elicitation our "questionnaire." We use these choices to estimate a utility model as a function of the wait fundamentals (see Section 3.6). Our scenarios represent "everyday" waits: participants exchange a moderate amount of time for a modest amount of money. To incentivize truth-telling, participants must endure one of their chosen waits in real time to earn its associated reward. Our experiment is comprised of eight treatments—all using the *same underlying waits*—that only differ in how we inform participants about the waits. This design allows us to understand people's utility over fundamental wait characteristics and the impact of information on that utility.

The waits are elemental. They specify a wait time distribution and a fixed monetary reward for enduring it. To isolate people's utility over reward and duration, there is no cover story (i.e. context) outside of the experiment itself. Rewards vary from $4 to $5. The wait durations are normally distributed with means varying between 4m 36s and 13m 38s and variances of 0, .25, or 1m^2. Some treatments present waits as queues with associated line lengths of 2, 5, or 8 people. Other treatments present waits as clocks (i.e., countdown timers) with participants receiving updates on their remaining wait time 2, 5, or 8 times (for symmetry). These attributes are fully listed in Figure 3.2.

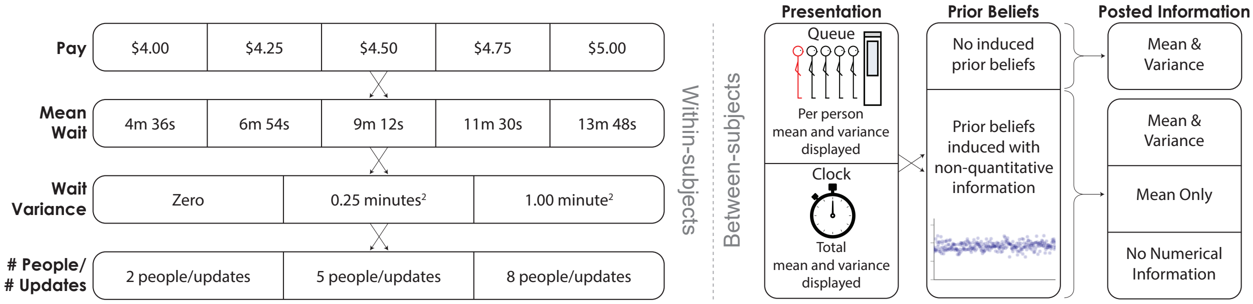


Figure 3.2: **Overview of Experimental Design.**

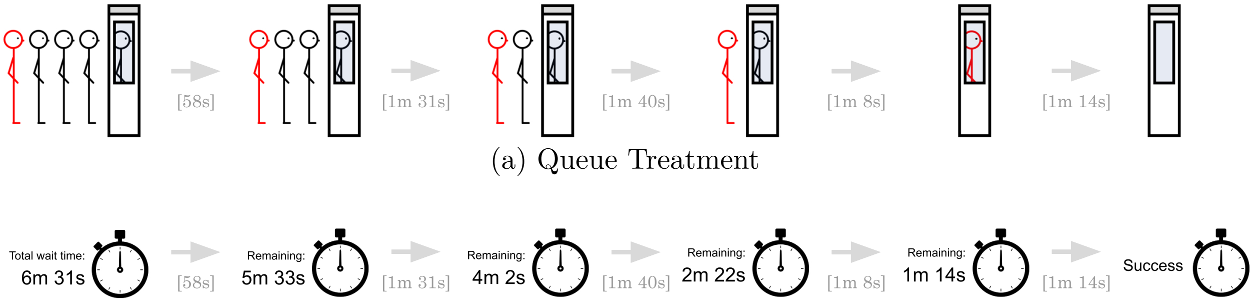
In total, there are $5 \times 5 \times 3 \times 3 = 225$ distinct waits ("profiles") created from these attributes that could be arranged into roughly 225^3 (over 11 million) choice sets. The careful selection of attributes, profiles, and choice sets—called the study's "design"—is a well-studied problem in the marketing literature (Rao, 2014). The goal is to achieve high statistical power with few questions to avoid participant fatigue. Each question should compare waits that are meaningfully different. The waits must also span the attribute space in a balanced manner. Our design (detailed in Appendix B.1) has a relative D-efficiency of 86%. We randomize the question sequence and the profile A/B/C labels to control for order and presentation effects.

Participants begin our study by reading instructions that describe how to interpret each wait profile, how their choices influence the selection of their real wait, and how their payment will be determined. We make clear that the majority of their potential earnings depends on successful completion of the real wait. We check understanding of these mechanisms with a four-question quiz. Participants must correctly answer at least three of these questions to continue with the experiment (see Table 3.1). For the interested reader, a sample set of instructions and quizzes for both the Queue and Clock treatments is available in the

	Queue				Clock			
Information	Attempted Quiz	Passed Quiz	Completed Real Wait		Attempted Quiz	Passed Quiz	Completed Real Wait	
Full-NoPrior	153	127	111	87%	131	104	97	93%
Full+Prior	111	75	63	84%	111	82	65	79%
Mean+Prior	109	75	66	88%	103	73	59	81%
None+Prior	126	85	70	82%	108	77	61	79%

Table 3.1: **Participants by Treatment.** Passing the comprehension quiz was an inclusion criteria. Participants who failed the quiz were not included in any analysis.

Appendix.



(a) Queue Treatment

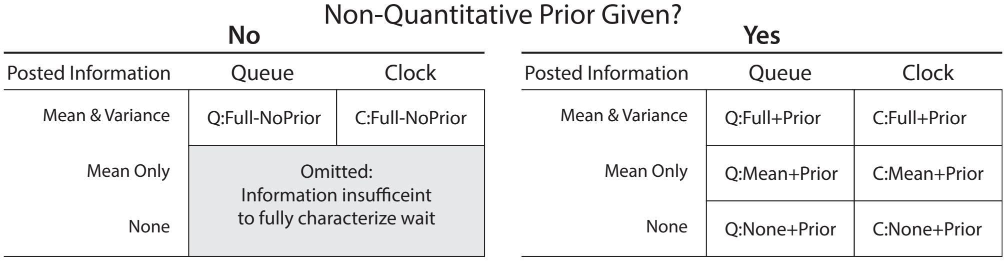
(b) Clock Treatment

Figure 3.3: **Example Incentivized Real Waits (Representing the Same Underlying Wait).**

After the questionnaire, we randomly select one of the participant's preferred waits for them to experience in real time (illustrated in Figure 3.3). While waiting, participants must complete regular attention checks (a button click every 15–30 seconds). If they complete the wait without missing an attention check, they earn the wait's associated reward. 80–90% of participants who begin a real wait complete it successfully (see Table 3.1). Participants earn a $.50 fixed wage for participation, $.25 for each correct answer on the comprehension quiz, and $4–5 for successfully completing the real wait.

3.4.1 Treatments

Our study design operates within the space of everyday wait fundamentals. Our treatments operate in the information space. They arise from three manipulations which represent common information structures. They are (1) Queues vs. Clocks, (2) Prior vs. No Prior, and (3) the posted information about wait time. We discuss these manipulations in detail below. Within this framework, we impose the restriction that we must fully inform participants about the true wait time distribution, either quantitatively or non-quantitatively. This results in the eight treatments listed in Figure 3.4.



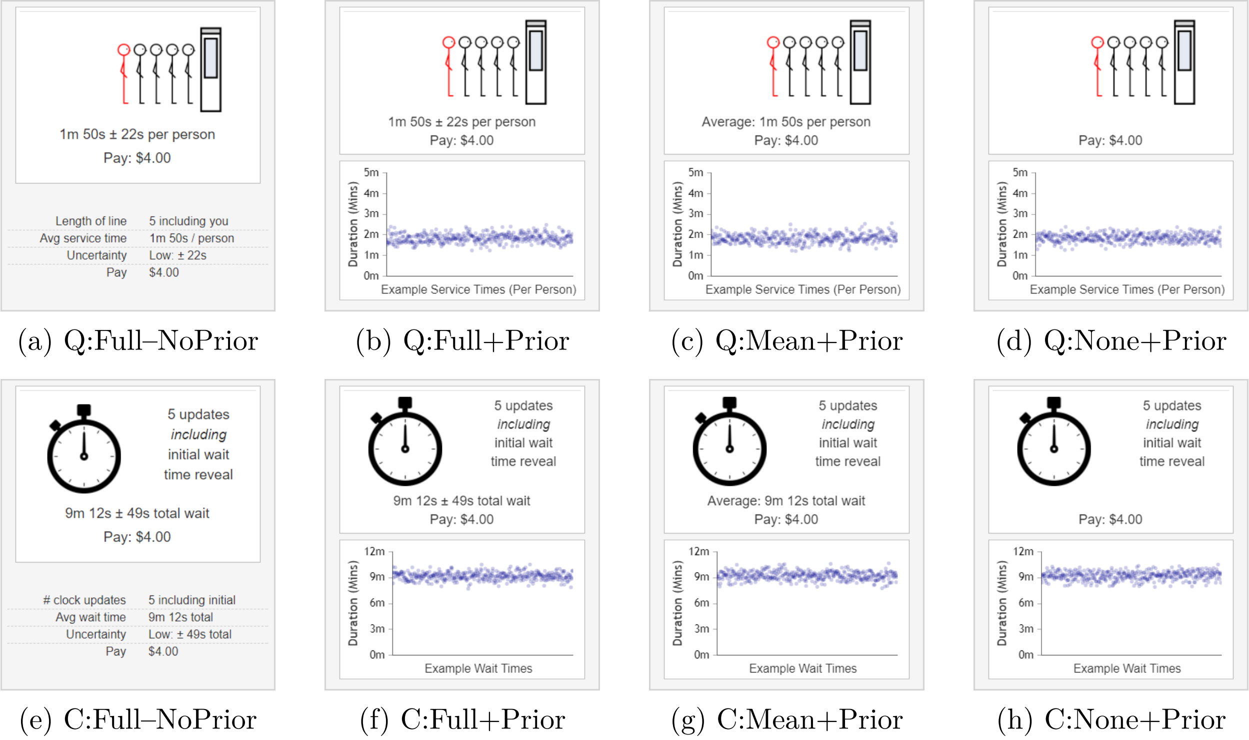
Non-Quantitative Prior Given?

No			Yes		
Posted Information	Queue	Clock	Posted Information	Queue	Clock
Mean & Variance	Q:Full-NoPrior	C:Full-NoPrior	Mean & Variance	Q:Full+Prior	C:Full+Prior
Mean Only	Omitted: Information insufficeint to fully characterize wait		Mean Only	Q:Mean+Prior	C:Mean+Prior
None			None	Q:None+Prior	C:None+Prior

Figure 3.4: **An Illustration of all Experimental Manipulations and Resulting Treatments.**

Queues vs. Clocks emulates two common presentation schemes for everyday waits: those presented as lines and those presented as aggregate timers. *Queues* are like grocery store and bank teller lines where the line length is visible, and customers typically join at the end. *Clocks* are like ridesharing apps and pizza deliveries where wait times are presented in aggregate, and firms often report an estimated duration or end time.

In our experiment, the way we inform participants about the fundamentals of each wait differs between Queues and Clocks. Queues have the feature “line length” (i.e., the number of people in line). All information posted about the wait time refers to individual service times and thus is presented on a per-person basis (see Figures 3.5a–3.5d). Clocks have the feature “number of updates,” and we present total wait time information in aggregate (see Figures 3.5e–3.5h). “Number of updates” tells participants how many times they will be informed about their remaining wait time during the real wait.


1m 50s ± 22s per person
Pay: $4.00

Length of line 5 including you
Avg service time 1m 50s / person
Uncertainty Low: ± 22s
Pay $4.00

(a) Q:Full–NoPrior

1m 50s ± 22s per person
Pay: $4.00

5m
4m
3m
2m
1m
0m
Duration (Mins)
Example Service Times (Per Person)

(b) Q:Full+Prior

Average: 1m 50s per person
Pay: $4.00

5m
4m
3m
2m
1m
0m
Duration (Mins)
Example Service Times (Per Person)

(c) Q:Mean+Prior

Pay: $4.00

5m
4m
3m
2m
1m
0m
Duration (Mins)
Example Service Times (Per Person)

(d) Q:None+Prior

5 updates *including* initial wait time reveal
9m 12s ± 49s total wait
Pay: $4.00

clock updates 5 including initial
Avg wait time 9m 12s total
Uncertainty Low: ± 49s total
Pay $4.00

(e) C:Full–NoPrior

5 updates *including* initial wait time reveal
9m 12s ± 49s total wait
Pay: $4.00

(f) C:Full+Prior

5 updates *including* initial wait time reveal
Average: 9m 12s total wait
Pay: $4.00

(g) C:Mean+Prior

5 updates *including* initial wait time reveal
Pay: $4.00

(h) C:None+Prior

Figure 3.5: **A Sample Wait as Presented in Every Treatment.** Ranges for variance are a 90% confidence interval around the mean. To convert this range from aggregate (as in the Clock treatments) to per-person (as in the Queue treatments), divide the radius of the aggregate range by the square root of the line length.

The Queues vs. Clocks manipulation also dictates how we present the real wait. In Queues, we place participants in the back of a simulated first-come, first-served line with line length serving as the only indication of progress (see Figure 3.3a). We don't tell participants how much time they have remaining. Interservice times are drawn according to the distribution represented by the profile selected for their real wait. In Clocks, a timer displays the total amount of time that the participant must wait. It refreshes periodically according to the number of updates in the participant's selected wait profile (see Figure 3.3b). We draw the interarrival times of these updates from the same distribution as the interservice times in Queues. In other words, each Clock wait is goverend by an unobservable queue.

Prior vs. No Prior Given manipulates whether we provide a non-quantitative (yet accurate and informative) signal about the wait's duration. To do so, we plot 400 sample draws from the wait time distribution (see Figures 3.5b–3.5d and 3.5f–3.5h).

Of course, this does not perfectly match how people's priors operate in practice. Real priors are much more complex. The two are similar, however, in that they contain wait time information but are not based on explicit statistics. In our study, providing sample draws from the wait time distribution allows us to quickly induce an accurate yet non-quantitative "sense" about wait time just as priors can do for waits in everyday settings.

Posted Information About Wait Time varies the statistics we use to describe each wait. We provide either: (1) no information, (2) mean wait time only, or (3) mean and a 90% confidence interval (which we refer to as "full" information). This emulates information provided about everyday waits. Some, like grocery store checkout lines, provide customers with no information. Others, like ridesharing apps, provide an estimate of mean wait time only. And uncertainty, if acknowledged at all, is presented as a range (e.g., 10 to 15 minutes). We mimic this presentation scheme when providing variance information. People have little intuitive understanding of variance; a range is a more interpretable way to communicate variability.

In Clocks, this posted information speaks directly to the aggregate wait time. But to have the same in Queues requires difficult arithmetic. Participants must multiply the per-person wait time statistics (given in minutes and seconds) by the line length, then somehow transmogrify the result back into something comprehensible using modulo arithmetic. This isn't trivial. And we don't adopt the minute/second notation as an obstruction—it's simply the way people describe time at this scale. Most people could not parse the easier-to-multiply range "0.833–1.167 minutes." It is much more natural to describe this range as "50s to 1m 10s." Further, understanding aggregate uncertainty requires additional difficult calculation. The per-person ranges we give for uncertainty are proportional to the standard deviation of wait time. To convert these to aggregate ranges, participants must multiply them by the *square root* of line length. To see these conversions of wait time statistics concretely, refer to Figure 3.5 which presents a single underlying wait presented as both Queue and Clock.

3.5 Experimental Setting

We ran our experiment on Amazon's Mechanical Turk (MTurk) using SoPHIE software (Hendriks, 2012). 952 participants entered into our study. 698 (73%) passed the comprehension quiz and are included in all of our analyses. 592 (85%) of these participants completed their real wait. Table 3.1 gives these statistics by treatment.

Running our study on MTurk provides the validity of a field experiment while avoiding potential issues from observational or laboratory studies in this setting:

Most observational data is too censored to study people's utility over wait fundamentals. First, people's priors are almost completely inaccessible—it wouldn't even help to ask about them. To understand people's preferences *given* their beliefs about a wait, these beliefs must be precisely induced. Second, in the field, people's decisions are obscured from the researcher. Consider balking. Many customers balk before they even approach a wait because they *believe* it will be too long. It would be difficult for a researcher to observe this decision. A controlled setting like ours allows us to completely observe people's actions and control their choice sets, giving us a clearer picture of their utility than would be available otherwise.

The traditional laboratory setting allows for more control but introduces a large fixed *time* cost to participants. Days in advance of their session, students must commit to a defined block of time and then travel to and from the lab. This dilutes the effect of the relatively short waits feasible within the experiment. It would be difficult to differentiate the subjects' preferences over the experiment's wait time and the time involved in the experiment overall. Because MTurk workers are already on the platform, implementing our experiment only requires some short instructions. The majority of time spent is tied to the real wait.

Our setting provides the precision of a laboratory while being minimally invasive to subjects' behavior. MTurk workers are constantly choosing among various options (HITs) to trade a little of their time for some money. Our experiment just asks them to do this same thing twenty times. Because our questionnaire is a scaled-down version of the MTurk platform itself and workers make their choices based only on their natural preferences and distaste for waiting, the study can be seen as a framed field experiment (Harrison and List, 2004) of how MTurk workers evaluate prospective waiting tasks. The average MTurk worker completes hundreds of HITs per month (Hara et al., 2018)—with such a diversified portfolio, workers would be wise to pursue a risk-neutral, wage-maximizing strategy. Thus, any behavioral deviations from this approach (e.g., risk aversion) that we observe among MTurk workers are likely to be even stronger in the general population.

3.6 Maximum Likelihood Estimation

We use maximum likelihood estimation (MLE) to tune the parameters of a model of people's utility over prospective waits. We will briefly outline this procedure—see Rao (2014) or Ben-Akiva and Lerman (1985) for a deeper discussion. Our proof sketch draws heavily from the former. For visual clarity, we refer to the symmetric attributes *line length* and *number of updates* as "#" in this section.

We begin by assuming that each participant, i, has a deterministic utility function, v_{ik}, over the pay, wait characteristics, and line length or number of updates of each profile, k. Here, *pay* refers to a profile's reward, *wait* refers to the mean aggregate wait time, and *var* refers to its aggregate variance. The deterministic utility functions we use for our estimation have the following form:

$$v_{ik} = \beta_{pay}\,(pay)_{ik} + \beta_{wait}\,(wait)_{ik} + \beta_{var}\,(var)_{ik} + \beta_{\#}\,(\#)_{ik} \tag{3.1}$$

Using the form $\beta_{wait}\,(wait) + \beta_{var}\,(var)$ to model people's utility over wait time is a "mean–variance" utility specification, which dates back to Markowitz (1952, 1959). It is useful for its interpretability for both scholars and practitioners. Titans of economics and finance have argued about the appropriateness of mean–variance utility because it requires a strong assumption: normally distributed returns (Borch, 1969; Liu, 2004; Johnstone et al., 2013). If this assumption is satisfied, mean–variance utility arises from the common exponential utility specification (Kumar et al., 1997; Johnstone et al., 2013). Critically, our waits *are* normally distributed—this is key to the symmetry of our Queue and Clock treatments. Conjoint analysis requires the ability to independently vary attributes (e.g., mean and variance) when generating profiles. This imposes a slight distributional constraint in Clocks. But in Queues, the aggregate wait time must be presented as the sum of i.i.d. service times. Maintaining the same underlying wait distributions across all treatments requires normality. Given any aggregate mean and variance, normally distributed aggregate waits can be deconvolved into an arbitrary number of i.i.d. normal random variables (i.e., service times). And, as an additional benefit, this allows the use of a mean–variance specification for our utility function.

We add a random shock, ϵ_{ik}, to the deterministic utility function v_{ik} to account for unobserved factors such as idiosyncratic preferences. We assume these shocks follow i.i.d. Type-1 Extreme Value distributions. Participant i's total (random) utility is then given by $u_{ik} \equiv v_{ik} + \epsilon_{ik}$.

We show participants choice sets, $s \in \{1, 2, \ldots, 20\}$, composed of three waits each. Participants choose the wait which has the greatest utility, $u_{ik},\ k \in \{1, 2, 3\}$. We denote participant i's choice in set s as $y_{is} \in \{1, 2, 3\}$. The probability that the participant will choose option 1 (the same arguments hold for options 2 and 3) is the probability that option 1 has the maximum utility in the choice set. That is, the sum of option 1's deterministic utility, v_{i1}, and its random shock, ϵ_{i1}, is greater than the total utility of both of the other options. Formally,

$$P(y_{is} = 1) = P\Big((u_{i1} \geq u_{i2}) \wedge (u_{i1} \geq u_{i3})\Big). \tag{3.2}$$

Ben-Akiva and Lerman (1985) show that this choice probability is a function of only the deterministic part of the participant's utility and has the following form:

$$P(y_{is} = 1) = \frac{e^{v_{i1}}}{\sum_{k=1}^{3} e^{v_{ik}}} \tag{3.3}$$

This is the well-known multinomial logit choice model. We use these choice probabilities to tune the β's in the participants' deterministic utility functions. The likelihood, l_i, of seeing participant i's history of chosen waits, $\{y_{i1}, y_{i2}, ..., y_{i20}\}$, is equal to the product of the probability of choosing those waits: $l_i = \prod_{s=1}^{20} P(y_{is})$. The likelihood of seeing the history of chosen waits for *all* participants is the product of their individual likelihoods: $L = \prod_{i=1}^{N} l_i$. These L's are functions of the attribute parameters (i.e., the corresponding β's). We maximize the likelihood of observing the participants' choices by maximizing L over the parameters β_{pay}, β_{wait}, β_{var}, and $\beta_{\#}$ for each treatment separately.

3.7 Experimental Results

We use the data from our experiment and the estimation methodology described in Section 3.6 to calibrate our utility functions (see Equation 3.1). The resulting coefficients appear in Appendix B.2. Dividing each parameter by its corresponding β_{pay} normalizes it to the dollar scale. Doing so makes treatments easier to compare—the value of a dollar is universal. The resulting pay elasticities, listed in in Table 3.2, represent the marginal payment required to offset a unit increase in each attribute. For ease of exposition, we will refer to these scaled coefficients for the remainder of this section.

	Queue				Clock			
Information	Pay	Mean	Variance	# People	Pay	Mean	Variance	# Updates
Full–NoPrior	1.00^{***}	$-.063^{***}$	$-.224^{***}$	$-.021^{***}$	1.00^{***}	$-.105^{***}$	$-.084^{***}$	$-.006$
Full+Prior	1.00^{***}	$-.051^{***}$	$-.104^{***}$	.001	1.00^{***}	$-.093^{***}$	$-.086^{**}$	$-.010$
Mean+Prior	1.00^{***}	$-.056^{***}$	$-.101^{***}$	$.013^{***}$	1.00^{***}	$-.121^{***}$	$-.091^{***}$	$-.011^{*}$
None+Prior	1.00^{***}	$-.053^{***}$	$-.111^{**}$	.000	1.00^{***}	$-.110^{***}$	$-.072^{***}$	$-.012^{***}$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.2: **Model Estimates Converted to Dollar Scale.**

We believe our estimated utility models provide the clearest evidence for our results, but we will corroborate our claims by referencing participants' choices. Table 3.3 lists the average value of participants' chosen profiles for each attribute as well as these average values for all 60 profiles in our design. Recall that each treatment presented waits with the same underlying fundamentals.

	Queue				Clock			
Information	Pay	Mean	Variance	# People	Pay	Mean	Variance	# Updates
All Profiles	4.50	9.23	.417	5.00	4.50	9.23	.417	5.00
Full–NoPrior	4.69	8.22	.305	4.66	4.67	7.40	.358	4.85
Full+Prior	4.68	8.47	.365	4.97	4.66	7.88	.370	4.83
Mean+Prior	4.67	8.36	.380	5.20	4.61	7.34	.379	4.81
None+Prior	4.68	8.52	.351	4.73	4.66	7.86	.380	4.80

Table 3.3: **Average Attribute Values of Chosen Profiles by Treatment.**

The remainder of this section is structured as follows. In Section 3.7.1 we discuss the commonalities among our treatments' different presentation schemes. These suggest some universal characteristics of people's utility over wait fundamentals. In Sections 3.7.2–3.7.4, we discuss the effects of our three manipulations: Queues vs. Clocks, posted wait time information, and Prior vs. No-Prior.

3.7.1 People's Utility Over Wait Fundamentals

Each of our eight treatments presents the same wait fundamentals under a different presentation scheme (see Figure 3.4). And across all eight we find commonalities: Participants choose waits with higher than average pay, shorter than average duration, and lower than average variance (Wilcoxon signed-rank test, $p < 0.0001$ for each attribute in each treatment; the participant is the unit of analysis for all direct comparisons of choices in this paper). These are not surprising. Of course people prefer larger monetary rewards. Of course they avoid long and uncertain waits. But these commonsense results lay a useful foundation: Preferences are qualitatively and directionally consistent across an array of information schemes. And people are risk averse with their time in all of them.

Our utility models confirm the consistent main effects described above. Pay has a positive coefficient in all treatments ($p < 0.001$, see Table 3.2). Mean wait time and wait variance both have negative coefficients in all treatments ($p < 0.01$). The penalty on mean wait ranges from \$0.05–0.12/minute while the penalty on variance ranges from \$0.07–0.22/minute^2.

We will close this section with discussion of our fourth (and heretofore ignored) attribute: line length (Queues)/number of updates (Clocks). Its impact on participants' choices is inconsistent (see Table 3.3): Participants choose profiles with a mean value significantly below 5 (the average of all profiles) in some treatments but indistinguishable from 5 in others. This is to be expected in our study because line length is, by construction, divorced from wait time. Accordingly, our estimated coefficients for this attribute are small and not statistically different from zero in most treatments, though two estimates are significantly negative and one is significantly positive ($p < 0.001$, see Table 3.2). Given this inconsistency, we refrain from drawing conclusions about this attribute.

3.7.2 Queues vs. Clocks

Our participants choose different waits depending on the presentation scheme (Queues vs. Clocks). Specifically, they behave as if their per-minute waiting cost is nearly twice as high when facing Clocks. And this difference is robust—the pattern holds across each of our four informational conditions.

The differential influence of Queues vs. Clocks has practical implications for service systems. These two formats are ubiquitous: Checkout lines are like Queues. Ridesharing apps are like Clocks. But often, system designers must *choose* a presentation scheme. A busy restaurateur could tell an arriving customer (1) to expect a forty-minute wait or (2) that they are the ninth party in line. Both approaches are perfectly reasonable. Yet our results suggest that this decision will have a dramatic impact on how customers evaluate the prospect of waiting.

Participants are more sensitive to mean wait time when evaluating Clocks and less sensitive when evaluating Queues, independent of fundamentals. Mean wait imparts a penalty of roughly \$0.10/minute in Clocks but only \$0.05/minute in Queues (see Table 3.2). Converting these elasticities from minutes to hours yields estimates of our participants' reservation wages: the compensation required to offset an hour-long wait. Among the four Clock treatments, we find reservations wages of about \$5.50–7.25/hour; for Queues, the range is \$3.00–3.75/hour. Hara et al. (2018) report an *actual* mean wage of \$6.19/hour for MTurk workers while on task. Again, participants' choices in Clocks are more consistent with behavior observed outside of our experiment.

Looking at choices directly provides additional detail. Participants choose mean wait times that are about 10% longer when facing Queues compared to Clocks (see Table 3.3). This is significant in each of the four information conditions (Wilcoxon rank-sum test, $p < 0.05$). These longer waits, however, are not offset by substantially higher pay. In only one condition (Mean+Prior) is chosen pay different at the 0.05 significance level: In this condition, pay is 1.3% higher for Queues. Pay trends higher for Queues in the other three conditions, but the relative differences is only about 0.5%. Together, these results indicate that participants earn lower hourly wages in Queues. This is true in each of our four information conditions: The average wage for those assigned to Queues is \$2.33–4.17/hour lower than for Clocks. And earning a high wage is the primary objective for MTurk workers (Kaufmann et al., 2011; Savage et al., 2020). Participants' choices in Clocks are more consistent with the preferences they display in other HITs.

A natural question is: What accounts for these differences between Queues and Clocks? Our experiment is designed to measure preferences, not explain them. We have no objective benchmark for "correctness" or "optimality." But the literature on bounded rationality offers a compelling explanation for the behavior we observe. Huang et al. (2013) interpret bounded rationality as customers' "incapability of estimating expected waiting time in a service setting" (p. 264). In Clocks, expected waiting time is easy to estimate—it is explicitly posted in three of four conditions and visually represented in the fourth. In Queues, participants must compute the product of mean service time and line length. And the way that humans demarcate time—demanding arithmetic in modulus 60 for seconds and minutes—makes this particularly difficult. It follows that our participants should appear more "rational" in Clocks and less so in Queues. And if we operationalize "rationality" as consistency with preferences and behavior exhibited in other settings, this is exactly what we find.

3.7.3 Posted Information About Wait Time

One might expect the choice of which statistics to report about a wait to have a substantial impact on how people evaluate it. People may appear averse to long wait durations when only the mean wait time is posted (making it more salient) and averse to uncertain waits when the variance is posted. We find some evidence of this, but the magnitude of this "priming" effect is small. Only a few of the observed differences are statistically significant.

3.7.4 Prior vs. No Prior Given

Prior beliefs are central to the way that humans assess waiting prospects. We find that people rely on their non-quantitative "sense" of a wait—even to their detriment—despite being provided accurate wait time statistics. This effect is small in Clocks: decision making is quite consistent across all information conditions. In contrast, the effect is substantial in Queues: behavior is consistent among the three conditions with non-quantitative information but notably different in the fourth. Participants choose profiles with slightly higher pay (n.s.) and shorter mean wait time (n.s.) but much lower variance (highly significant) when they lack induced prior beliefs. Computing the aggregate wait time from individual service time statistics (as required in Queues) is laborious—and avoided when possible—yet effective at identifying objectively good (lucrative, short, low-variance) waits.

A key feature of the non-quantitative prior is that it allows us to fully inform participants about expected wait times and uncertainty (concepts inherent to everyday waits) without having to explicitly post wait time statistics (rare in everyday waits, especially for uncertainty). The None+Prior condition provides participants with no numerical wait time information, yet the mean values of chosen profiles (see Table 3.3) are indistinguishable from those in Full+Prior (no significant difference for any attribute at the 0.05 level in either Queues or Clocks) despite the availability of a full numerical characterization of the wait time distribution in the latter. When participants have non-quantitative priors, there appears to be no marginal value of getting full posted information. When priors are taken away (Full–NoPriors), however, participants tend to choose objectively better waits. Relative to Q:Full+Prior, the average chosen profile in Q:Full–NoPrior pays \$0.01 more (n.s.), is about 15 seconds shorter (n.s.), and has nearly 20% less variance (Wilcoxon rank-sum test, $p < 0.0001$). And relative to C:Full+Prior, the average profile in C:Full–NoPrior pays \$0.01 more (n.s.), is about a half-minute shorter ($p < .05$), and has about 4% less variance (n.s.). Again, the posted information in these two treatments is identical; Full+Prior just illustrates the same information in pictorial format as well. But this provision of non-quantitative information causes participants to choose waits that tend to be lower-paying, longer, and more variable.

An exaggerated impact of non-quantitative priors on Queues (as opposed to Clocks) is clear from our scaled model estimates (see Table 3.2). In Clocks, the Full–NoPrior condition has elasticities that are similar to the other conditions: Its pay elasticity of mean wait time ranks second out of the four treatments. Its elasticity of wait variance ranks third. Within the Queue treatments, however, Full–NoPrior is an outlier: It has the largest elasticity of mean wait time (12.5% larger than the next-largest) and the largest elasticity of wait variance (more than twice that of the next-largest). These extreme elasticities have several ramifications: Participants in Q:Full–NoPrior are particularly discerning in their choices (hence the objectively good chosen waits). Further, for *any* Queue wait (irrespective of fundamentals), the imputed waiting cost (mean wait penalty plus variance penalty) will be largest in the Full–NoPrior condition. Finally, the extreme elasticity of variance (\$0.22/minute^2, the largest penalty we report) indicates particularly acute risk aversion.

This pattern of results suggests that people select waits based on their non-quantitative sense unless forced to do otherwise. In practice, customers may rely on their prior beliefs except in the most novel of waiting situations. Perhaps this is why posted information (especially information about uncertainty) is rare in everyday waits—most customers would just ignore it, anyway.

3.8 Managerial Insights

We will model this section on David Maister's essay, *The Psychology of Waiting Lines* (1985). Its high citation count (currently over 1,200 on Google Scholar) suggests that many find the style useful. Additionally, this section will demonstrate some simple applications of our utility functions.

3.8.1 Illustrative Examples

Consider a single-server system with $n = 5$ customers and an exponentially distributed service rate of $\mu = 1/2$. An arriving customer's wait time, T, follows an Erlang distribution with $\mathbb{E}[T] = n/\mu = 10$ minutes and $\text{Var}[T] = n/\mu^2 = 20$ minutes^2. The customer's deterministic utility, therefore, is

$$v = r + \mathbb{E}[T]\,\beta_{wait} + \text{Var}[T]\,\beta_{var} = r + n\left(\frac{1}{\mu}\,\beta_{wait} + \frac{1}{\mu^2}\,\beta_{var}\right) = r + 10\,\beta_{wait} + 20\,\beta_{var} \quad (3.4)$$

where r is the dollar-value of the service received. We omit the term corresponding to line length/number of updates because its effect is small and not significantly different than zero in many treatments (see discussion in Section 3.7.1). Here, the β coefficients refer to the dollar-scaled versions shown in Table 3.2, which change according to the setting.

We will now describe three brief vignettes based on the system described above. The fundamentals remain constant across all three, but the information available to the customer changes. We use the scaled coefficients from our most applicable experimental treatment to compute the dollar-value cost of the wait as appraised by the customer upon encountering it.

Call Center.

The system models a busy call center. A customer familiar with the system calls and immediately receives a message stating that the expected wait time is 10 minutes. From prior experience, the customer knows that (even conditional on this message), actual waiting times are highly variable. This is like our Clock:Mean+Prior treatment.

$$v_{\text{call center}} = r + 10 \times (-\$.12) + 20 \times (-\$.09) = R - \$3.00 \tag{3.5}$$

Coffee Shop.

The system models a local coffee shop. The shop does not report any wait time statistics. An arriving customer who visits the shop frequently observes the queue length and evaluates the waiting prospect based on prior beliefs. This is like our Queue:None+Prior treatment.

$$v_{\text{coffee shop}} = r + 10 \times (-\$.05) + 20 \times (-\$.11) = R - \$2.70 \tag{3.6}$$

Gondola Ride.

The system models a physical queue in an unfamiliar setting: the line to board a gondola in Venice. A newly arrived tourist encountering the line reads from a travel guide: “It may take up to six minutes for a Gondola to arrive. On average they come every two minutes.” This is like our Queue:Full–Prior treatment.

$$v_{\text{gondola ride}} = r + 10 \times (-\$.06) + 20 \times (-\$.22) = R - \$5.00 \tag{3.7}$$

3.8.2 Maister’s Original Observations

First, we will use our utility functions to add nuance to two of Maister’s original observations.

Uncertain waits are longer than known, finite waits,

Maister (1985) writes, and we agree. Wait variance imparts significant disutility in every treatment we tested. Our estimates expound upon Maister's sentiment by describing *how much* worse those uncertain waits are.

At the call center, the variance of wait time accounts for 60% of total disutility (see Equation 3.5). This figure is 80% at the coffee shop (see Equation 3.6). Not knowing when the wait will end is a greater source of disutility than the (mean) duration of the wait itself.

The more valuable the service, the longer the customer will wait,

states Maister (1985). Of course this is true, though a more precise observation might be: "Increasing the value of service makes customers willing to endure a more *costly* wait."

Will customers wait twice as long for twice the reward? Our utility functions show that this depends on the source of the delay. Doubling the length of the coffee shop line to $n = 10$ would double both its mean duration and disutility. In contrast, doubling the mean service time (i.e., cutting μ to $1/4$) would nearly *quadruple* disutility while still only doubling mean duration. Why? Utility is linear in n but quadratic in $1/\mu$ (see Equation 3.4).

Management should keep this in mind when tackling increased customer wait times. If longer waits are due to an influx of new customers creating longer lines, then a relatively small increase in the reward (e.g., a free tote bag or a modest discount) may offset the marginal disutility of waiting. But overcoming this disutility will be much more difficult in response to a drop in service rate. Managers should be cautious about adding time-consuming administrative tasks (e.g., having checkout clerks describe the benefits of the company's credit card to each customer).

3.8.3 Novel Observations

Clock waits are intimidating.

Clocks are are easier to parse than queues because they present the aggregate wait time explicitly, and this allows customers to make objectively better choices. For example, participants choose higher-wage waits in our Clock treatments relative to Queues (see Section 3.7.2 for full discussion). But the accessibility of clock waits also makes them intimidating: Customers can envision the entire (painful) waiting experience, right up front. The disutility per minute of mean wait time is twice as high for clock waits compared to queues. Total disutility in the clock-like call center is "only" 10% higher than in the coffee shop queue because the high variance of exponential service times dominates (see the first point in Section 3.8.2). In a lower-variance setting, a clock wait could induce up to twice the disutility of an analogous queue.

Everything is going digital these days, and that includes waiting "lines." Clock-like Uber waits are replacing physical taxi queues. Restaurant patrons receive predicted wait times with their Google and Yelp search results and can join wait lists virtually. Management must consider how this migration will change customer behavior. Intimidating clock waits may lead to higher balking rates. Reporting instead the length of a virtual queue may be better at getting customers "through the door."

Queues are more disappointing.

This is a corollary to the previous observation. Queues are relatively difficult to parse. Our results suggest that queue-like presentations cause people to underestimate the duration or pain of long waits (see Section 3.7.2). Customers are therefore more likely to join long waits when they are presented as queues. In this sense, queues may "trick" customers into joining a wait that would have made them balk had it been described more accessibly, like a clock. But this trick may leave customers disappointed. In the course of waiting, tricked customers will eventually realize the true nature of a painful wait, at which point they may renege or grow resentful. Perhaps this is why our participant complained that “waiting in line sucks.” But a firm that can turn a wait into a pleasant experience (e.g., a restaurant with live music that serves drinks to waiting customers) can make use of this trick: queues can get people in the door despite a long or variable wait.

It’s the variance that kills in service systems.

This stems from two facts: (1) people are quite sensitive to wait variance and (2) most queueing models involve high wait variance. As for (1), a minute^2 of variance imparts two to four times the disutility of a minute of mean wait time for queues. The import of the two is roughly equal in clocks. As for (2), exponential service times—probably the most common model used for service—result in variance which is the square of the mean. In tandem, it is straightforward to see why variance accounts for the majority of people’s disutility.

Managers should aim to reduce customers’ perceptions of wait time variance. This can be achieved via operational improvements to standardize service times, though it may require many interactions for customers to perceive any difference. Policies that are more visible to customers such as wait time guarantees (e.g., pizza delivered in 30 minutes or it’s free) may provide a more immediate approach to alleviate concerns over uncertainty. Note that these the two methods are complimentary. Guarantees can draw in wary customers who are risk averse with their time, and their experience with predictable wait times will keep them coming back.

Beliefs dictate customer behavior.

People rely heavily on prior beliefs when evaluating queues. They appear to defer to their non-quantitative “sense” of a queue—to their detriment—even when given precise service time statistics (see Section 3.7.4). Thus, improving queue fundamentals is only the first step in changing customer behavior.

Consider an institution famous for long and variable queues: state Departments of Motor Vehicles or DMVs. Imagine that a major system overhaul resulted in short and predictable wait times at all DMVs. How long would it take to update the DMV's *reputation* in the minds of its customers? Probably many years, during which time customers may make decisions based on outdated beliefs.

So how should a firm communicate improved queuing fundamentals to customers? Our results suggest that communicating wait time statistics (e.g., 90% of customers now wait less than 45 minutes) is unlikely to work. Instead, management should aim to give customers experiences that force them to update their priors. As discussed above, a wait time guarantee (e.g., if you have to wait more than 20 minutes, you will get a $20 voucher) might help draw in skeptical customers.

People are more risk averse in unfamiliar settings.

This final insight is closely related to the previous one—that people rely on prior beliefs to evaluate queues. When people encounter a queue for which they *don't* have relevant prior beliefs (i.e., when they are in an unfamiliar setting), we find that they become highly risk averse with their time. The disutility of the gondola queue is nearly twice that of the coffee queue despite the fact that they model nearly identical physical queues. This is because the disutility of a minute^2 of variance in our queue treatment with no priors is more than twice as high as that measured in any of the other seven treatments.

Firms bringing novel services to market should consider communicating their waits as clocks rather than queues to avoid inducing extreme risk aversion. And companies that do present their waits as queues should avoid measures that might "alienate" customers. For example, the new owners of a restaurant might be inclined to post a sign reading "Now under new management!" in the hopes of signaling better operations (e.g., better food, shorter wait times). This sign, however, may make customers feel that they can no longer rely on their prior beliefs. If so, it might induce severe risk aversion which could, in turn, lead to increased balking—just the opposite of its intended purpose.

3.8.4 A Model of Waiting Utility

We close this section by offering a single, parameterized utility specification for queueing theorists interested in incorporating the behaviors described above into their models. It is intended to capture the first-order features our eight fitted utility functions without relying on precisely tuned parameters. As far as we know, this is the only empirically grounded utility model of waiting available. We understand that it will not be appropriate in all settings. We hope that it will prove useful for some.

As with the examples in Section 3.8.1, our generalized model (see Equation 3.8) describes the utility of a customer encountering a wait with total duration distributed according to a random variable T. They receive a dollar-equivalent reward of r for completing the wait and have a prevailing wage rate of w (e.g., the rate of pay for some outside option). The two key features that contribute to the customer's appraised disutility of the wait are expected wait time, $\mathbb{E}[T]$, and variance of wait time, $\mathrm{Var}[T]$. Often, customers are modeled as only minimizing waiting costs in systems. We provide a reward-less cost function (Equation 3.9), which varies the relative import of $\mathbb{E}[T]$ and $\mathrm{Var}[T]$ according to the wait's information scheme.

We define two parameters which model the information given to customers about a wait (and thus change the relative import of its fundamentals). These parameters allow modelers to match their model to the nearest treatment in our study, thus capturing their setting's information effects on customer utility. First, we define α: a measure of the cognitive demand required to compute the wait time distribution T. If T is described explicitly (as in our Clock treatments), $\alpha = 0$. If T must be computed (as in our Queue treatments), $\alpha = 1$. Future research may explore other settings which vary this parameter more granularly. Second, we define γ: the gap between prior beliefs about T and T itself. If a customer has prior beliefs that perfectly characterize T (as in our treatments with induced non-quantitative priors),

$\gamma = 0$. This models a customer in a very familiar setting. If a customer has prior beliefs that differ from T or has no prior beliefs at all (as in our treatments with no induced priors), $\gamma > 0$. We assign our No Prior treatments $\gamma = 1$. This models a customer encountering a novel wait. As with α, this parameter may be explored in more detail in future research. We provide the following empirically grounded utility function which balances wait fundamentals (via a wait's mean and variance) and information (via α and γ).

$$v(T; r, w, \alpha, \gamma) = r - w\left(\frac{\mathbb{E}[T]}{1+\alpha} + \ (1+\alpha\gamma)\,\mathrm{Var}[T]\right) \equiv r - w\ c(T; \alpha, \gamma) \tag{3.8}$$

$$c(T; \alpha, \gamma) = \frac{\mathbb{E}[T]}{1+\alpha} + \ (1+\alpha\gamma)\,\mathrm{Var}[T] \tag{3.9}$$

Our general utility model in Equation 3.8 connects individual-level behavior to system-level operations. It is a parsimonious function of the first two moments of T, fundamentals that are well-defined in nearly all queueing systems. And the parameters governing information easily map onto many waits of practical importance: To model the call center (see Section 3.8.1), use $(\alpha = 0, \gamma = 0)$. To model the coffee shop, use $(\alpha = 1, \gamma = 0)$. And to model the gondola ride, use $(\alpha = 1, \gamma = 1)$.

When a utility model is needed, Equation 3.8 provides a flexible and descriptive candidate. It approximates the estimates from our eight experimental treatments without overly specific coefficients (see Figure 3.6). Instead, it models the relative import of mean duration and variability as a function of the modeled wait's information. In some settings, a function of mean wait time without a variance term may be required for mathematical tractability. We provide one in Equation 3.10. It gives an empirically grounded estimate of a customer's linear waiting cost as a function of their prevailing wage. For MTurk workers, we use $w \approx \$6.00$ (Hara et al., 2018). For retail workers, scholars could instead use use minimum wage (e.g., $w = \$15.00$ in New York City) or another measure of typical wages.

$$v(T; r, w, \alpha) = r - \frac{w}{1+\alpha}\mathbb{E}[T] \tag{3.10}$$

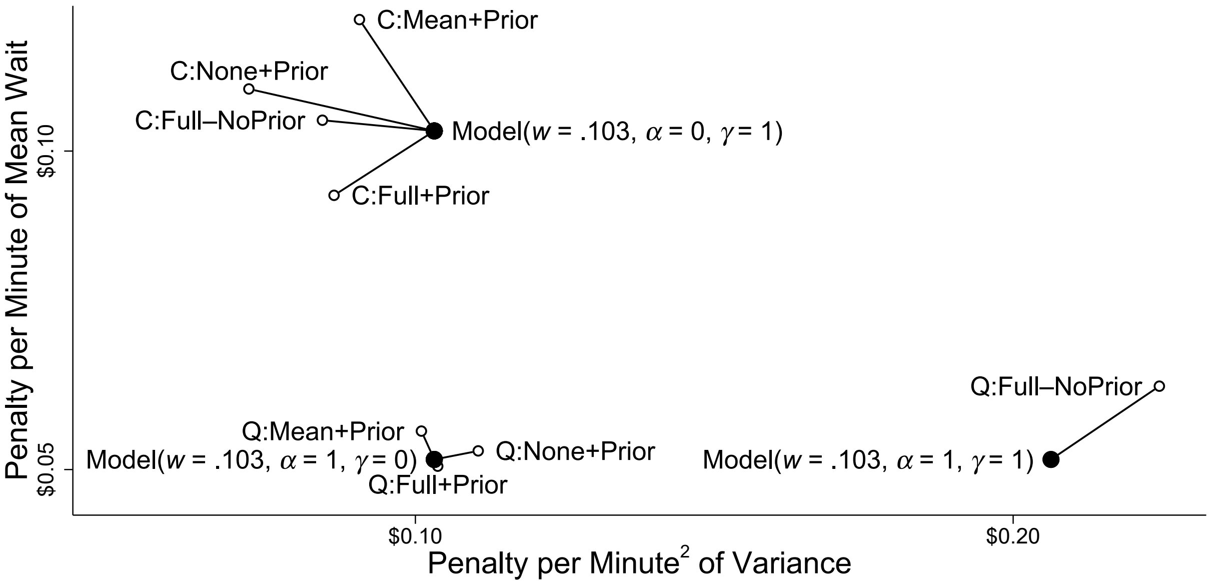


Figure 3.6: **Plot of the Penalty per Minute of Mean Wait over Penalty per Minute^2 of Variance.** A plot of the estimated pay elasticities of mean and variance for all eight experimental treatments and their relationship to the general utility model of Equation 3.8.

3.9 Conclusion

In this paper, we address the open research question of how wait fundamentals affect people's utility. Because there is no "regular" or "default" presentation scheme for waits, this pursuit required us to define and study the mediating effect of an information layer. We find some behavioral commonalities across all presentation schemes and information conditions: People dislike waiting and enjoy rewards. People are risk averse with their time—an effect that is intuitive yet rarely modeled. And, controlling for wait duration and uncertainty, people tend to be indifferent to line length.

The differences among our treatments show how information influences people's utility. There are two primary effects: (1) Mean wait time is twice as costly when times are presented in aggregate (like ride sharing apps; our Clocks) than when presented per-person (like grocery store lines; our Queues). Perhaps this is due to bounded rationality. Customers have difficulty translating per-person statistics into an aggregate wait time. Their tendency to underestimate the duration or pain of waiting makes them appear unusually patient. (2) People familiar with a wait rely on their prior beliefs rather than posted statistical information. And (3), posting information about a wait's duration or variability does not induce significantly more sensitivity to that feature for customers familiar with the wait. We summarize the effects of all of our treatments into a utility function which is ready to use for modelers who desire an empirically grounded specification for how customers evaluate waiting.

Perhaps the most cited work in this area is Maister (1985), whose pithy comments on people's psychology in waiting lines have helped researchers and managers alike. To that end, in Section 3.8 we add empirically grounded nuance to two of his observations: uncertain waits are longer than known, finite waits and the more valuable the service, the longer the customer will wait. And we supply five novel observations in Maister's style drawing from observed participant behavior: (1) clock waits are intimidating (2) queues are more disappointing, (3) it's the variance that kills in service systems, (4) beliefs dictate customer behavior, and (5) people are more risk averse in unfamiliar settings. We hope that these managerial insights can also be useful for both researchers and practitioners.

The goal of our study is to describe the links among customer utility, wait fundamentals, and information. Further work could give deeper insights into the cognitive and psychological drivers of customer behavior. We study people's utility for "everyday" waits under some common presentation schemes. There is room for much broader exploration. We test waits where variance fluctuates independently from mean wait time. Future studies may focus on alternative concepts of variability, such as the ratio of variance to mean wait time. More complex reward structures should be incorporated, and other common waiting regimes and contexts should be studied. Important waits occur on many timescales: a few minutes for a checkout line; several hours for a grocery delivery; and many days for a package delivery. All merit attention. Our study keeps the context as neutral as possible. We suspect that more colorful contexts and presentation schemes would influence behavior.

We see this paper as just a first step in understanding people's utility over waits. We look forward to more research unpacking the seemingly simple feeling that “waiting in line sucks.”

APPENDIX A

APPENDICES FOR CHAPTERS 1 AND 2

A.1 Credit Days

Detainees in jail or on EM accumulate credit time which counts against their sentence duration. There are three main ways of getting credit time, Statutory Sentence Credit, Program Sentence Credit, and Supplemental Sentence Credit.[1]

Statutory Sentence Credit reduces the percentage of their sentence detainees must serve based on the severity of the crime. Common percentages are 50%, 75%, 85%, or 100%. These percentages are determined by statute and by judge: different crimes have their minimum percentage of time served enumerated in law, and judges can increase the required percentage on a case-by-case basis. Detainees may lose this credit due to negative behavior while in custody.

Program Sentence Credit is credit for participating in certain programs while in jail or prison. Common programs are for education, drug rehab, life skills courses, behavioral modification, re-entry planning, and Illinois correctional industries. If an detainee participates, they earn a half of a day per day in program, conditional on completing the program. Not everyone is eligible for this program, and detainees can lose this credit due to bad behavior.

Finally, Supplemental Sentence Credit gives up to 180 days of credit for good behavior in prison, and is the sole discretion of the Director of Department of Corrections.

A.2 Appendix: Speedy Trial Act

Defendants' rights to a speedy trial are ensconced in the sixth amendment: "In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial." And in Illinois, the legislature has also guaranteed this right by statute with the "Speedy Trial Act" (725

[1] `https://www2.illinois.gov/idoc/aboutus/Pages/faq.aspx#qst1` Accessed on 10/18/2021.

ILCS 5/103-5). This act establishes a time limit for cases. After that time is expired, defendants may call for an immediate trial.

For defendants who are released on bail the limit is 160 days. For detainees on CCSO custody, the limit is 120 days. However, “Delays occasioned by the defendant” do not contribute to the 120- or 160-day speedy trial act clock. Any pretrial motion filed by the defendant (such as requests for information) is a delay occasioned by the defendant. The time the motion takes to resolve does not contribute to the speedy trial clock. “Agreement continuances” also count as delays occasioned by the defendant. Most often, these come from pretrial negotiations between the prosecution and defense. They may involve plea negotiations, witness availability, trial stipulations, and other things which take time. During these negotiations, the prosecution and defense meet with a judge to update them about the case and continue the case to complete these negotiations.

A.3 Detailed Data Description

A.3.1 Detainee Information

Inmate ID. Each detainee is assigned a unique identifier which does not change between cases or bookings.

Criminal History. We score defendants’ criminal history with a metric that approximates the United States Sentencing Commission’s (USSC) criminal history metric (USSC, 2019). Our metric is approximate because we only have access to criminal data from Illinois, while the USSC Criminal History metric incorporates national data.

In our metric, each detainee is given a criminal history score. They are assigned 3 points for each prior prison sentence, 2 points for each prior jail sentence, and 1 point for each prior probation or supervision sentence. We bucket these criminal history scores into four descriptive bins: “None”, “Low”, “Medium”, and “High”, each accounting for about 25% of the data.

Prison History. This data field indicates if the detainee has been sentenced to an Illinois prison in the past.

A.3.2 Housing Information

Booking ID. The CCSO assigns detainees a unique booking ID distinguishing each time they are booked in jail or on EM.

Booking Date. This is the date the detainee is placed under the responsibility of the CCSO for each booking. We use this date as the beginning of both the detainee's case and detention.

Pretrial Housing Location. There are two pretrial housing locations under the CCSO's purview: EM and jail. In our data, 13.8% of detainees are on EM pre-trial and 86.1% are in jail pre-trial.

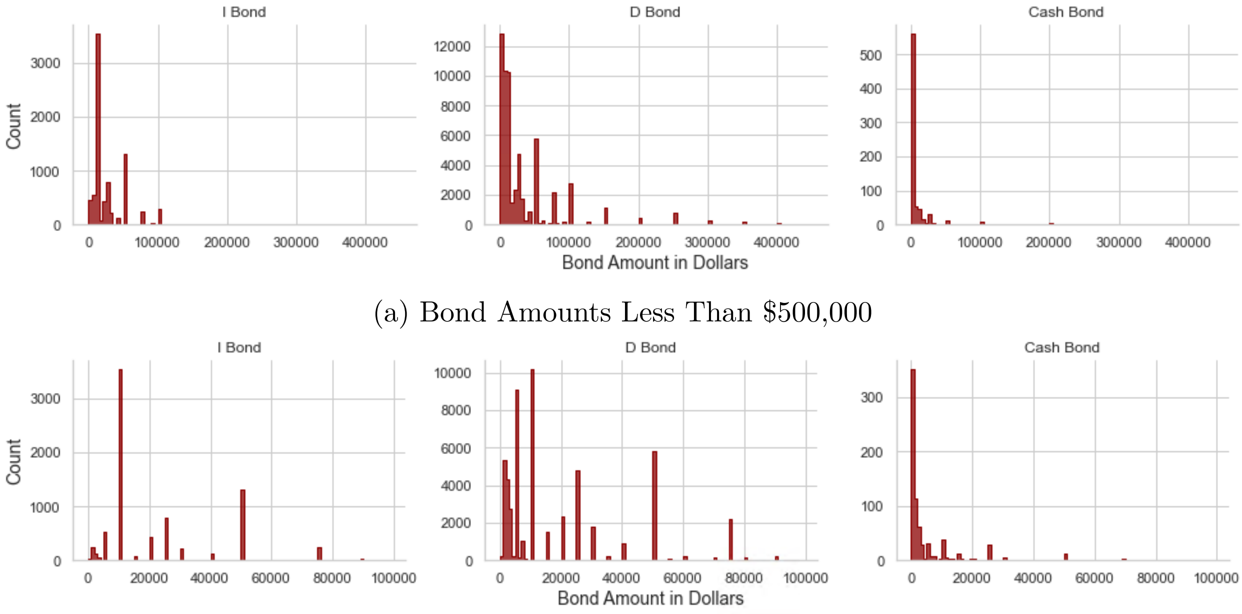
Security Classification. The CCSO classifies each detainee as either minimum, medium, or maximum security while under their purview.

Bond Type. As mentioned in Section 1.3, detainees are assigned a bond during their bond hearing, which outlines conditions they must meet for release from jail. The relative frequency of bond types is given in Table A.1.

Bond Type	Percent
D Bond	70.3%
No Bond	18.2%
I Bond	5.8%
I Bond with EM	3.7%
D Bond with EM	1.0%
Cash Bond	0.9%

Table A.1: **Bond Types.** Bond types listed in decreasing order of frequency for detainees housed in jail. "I Bond with EM" and "Deposit Bond with EM" indicate that the detainee will be released onto EM if bond is posted.

Bond Amount. As mentioned in Section 1.3, I, D, and cash bonds are associated with dollar amounts. Figure A.1 displays histograms of these amounts for the bonds in our dataset. Recall that for these bonds, detainees must pay 0%, 10%, and 100% for release, respectively, but are responsible for the full amount if they violate their bond by not arriving for court or re-offending during release. Histograms of bond amounts are given in Figure A.1. Note that the masses near zero are not exactly zero—bonds are often listed for values less than $1,000.



(a) Bond Amounts Less Than $500,000

(b) Bond Amounts Less Than $100,000

Figure A.1: **Histograms of Bond Amounts by Bond Type.** (A.1a) displays histograms of bond amounts less than $500,000 in the dataset. (A.1b) displays histograms of bond amounts less than $100,000 for easier readability of smaller bond amounts. Both are grouped by I, D, and Cash Bonds. I and D bonds include those with EM conditions. In all cases, these amounts are as listed on the bond; bond type determines the amount a detainee would need to play for release.

A.3.3 Case Information

Docket Number. The docket number is a unique ID number assigned by the courts for each case.

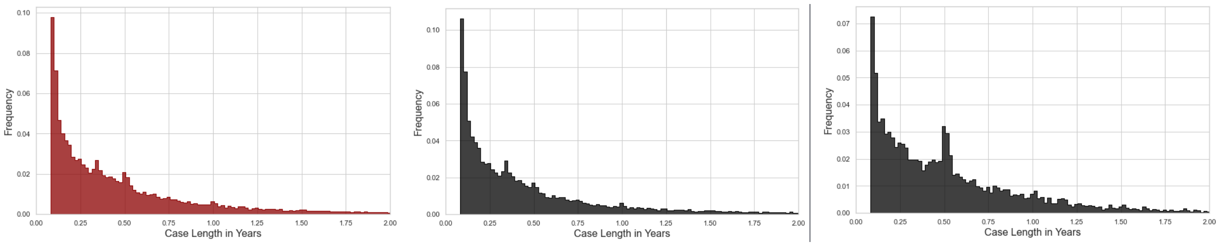
Crime Class. At the outset of a case, the prosecution levies charges which enumerate the laws the defendant is accused of breaking. Charges are bucketed into different "crime classes," which represent the severity of the crime. This class incorporates a combination of the alleged illegal act(s), details about the case, and the defendant's criminal history. During sentencing, crime class often determines the minimum and maximum sentence the judge can assign. There are nine primary crime classes in Illinois: six for felonies (M, X, 1, 2, 3, and 4) and three for misdemeanors (A, B, and C) (Divito, 2001). Table A.2 lists each with their frequency in the data and example charges. In the data, 12% of charges fall into other miscellaneous categories, such as petty crimes.

Class	Frequency	Classification	Example Charge	Sentence Duration
M	0.6 %	Felony	First-degree murder	20-60 Years
X	0.6 %	Felony	Aggravated criminal sexual assault	6-30 Years
1	5.5 %	Felony	Second-degree murder	4-15 Years
2	8.0 %	Felony	Kidnapping, arson	3-7 Years
3	7.7 %	Felony	Perjury	2-5 Years
4	25.4 %	Felony	Stalking	1-3 Years
A	30.1 %	Misdemeanor	Criminal trespass	<1 Year
B	2.5 %	Misdemeanor	Aggravated speeding	<.5 Years
C	1.6 %	Misdemeanor	Disorderly conduct	<30 Days

Table A.2: **Crime Classes in Illinois.** Crime classes in Illinois in descending order of severity. Example charges are listed for each crime class. Sentence duration is given as a range of the typical minimum and maximum possible sentence for that crime class (Divito, 2001).

Case Length. A detainee's case length represents their time between booking (at the outset of their case) and release from pretrial detention (at their time of disposition and sentencing, if applicable). Figure A.2a displays a histogram of all case lengths within the dataset. Figures A.2b and A.2c display case length grouped by pretrial housing location. Cases in which detainees are on EM tend to be longer, and also have a characteristically different shape.

Notice the "spike" in frequency of cases that ended at half a year for detainees on EM. These correspond to one of the most common prison sentences: half a year. This spiking



(a) **All Detainees** (b) **Detainees in Jail** (c) **Detainees on EM**

Figure A.2: **Histograms of Case Lengths.** Histograms are grouped in the following manner: on the left (in red) (A.2a) displays a histogram of all detainees in the data. The two histograms on the right (in black) partition the data into detainees in jail (A.2b) and detainees on EM (A.2c). Bins for these histograms are one week wide. Case lengths are truncated at two years for readability.

behavior is most stark on EM, but can also be seen for detainees in jail convicted of class 4 felonies and sentenced to prison. We highlight some of these spikes in Figure A.3.

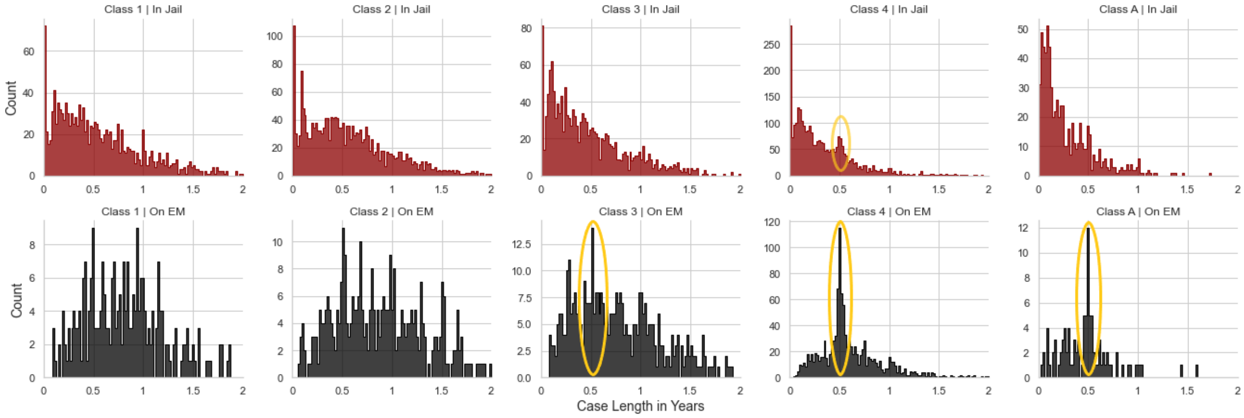


Figure A.3: **Case Lengths of Detainees Sentenced to Prison by Crime Class and Pretrial Housing Location.** Histograms are grouped by crime class (decreasing in severity from left to right, focusing on classes 1, 2, 3, 4, and A) and pretrial housing location (in jail on top in red, on em on bottom in black). We've added orange ovals to highlight prominent spikes at common sentence durations for prison. Case lengths are truncated at 2 for readability.

Detainees with different sentence locations, crime classes, and housing locations have characteristically different case length distributions. More severe crime classes tend to have longer cases. Prison sentences are associated with the longest cases. Cases where detainees are housed in jail tend to end earlier than cases where detainees are housed on EM. Also, detainees on EM have larger probability masses for case lengths near common sentence locations.

Court Dates. The dataset lists each detainee's court dates. From these, we compute the number of court visits each detainee had, and the court visits' interarrival times. We also use consecutive court visits at the end of a case to impute which detainees went to trial and which detainees accepted plea bargains.

Number of Court Visits. Defendants in the dataset visit the courts 4.9 times on average per case. As the severity of the crime increases, so do the average number of court visits. Table A.3 shows descriptive statistics regarding the number of court visits for each crime class.

Class	Mean	Median	Std. Dev.
M	25.6	7	13.3
X	11.4	19	26.5
1	8.6	5	10.1
2	8.5	5	11.2
3	6.8	4	7.7
4	4.1	2	5.0
A	2.2	1	4.1
B	1.8	1	2.9
C	1.6	1	2.0

Table A.3: **Descriptive Statistics of Court Visits by Crime Class.**

Court Visit Interarrival Times. The interarrival times of detainee's court visits are typically an integer multiple of a week. We plot histograms of the interarrival times of detainees' first five court visits in Figure A.4.

Plea Bargains. The majority of cases end in plea bargains, otherwise, they end in a trial. We do not observe this data directly, but can impute it from the detainees' court visits. Trials are often multi-day events, followed by a not guilty verdict or a sentencing hearing. We detect trials by looking at the interarrival times of the final court visits for detainees found guilty. We say that two or more court visits on consecutive weekdays within

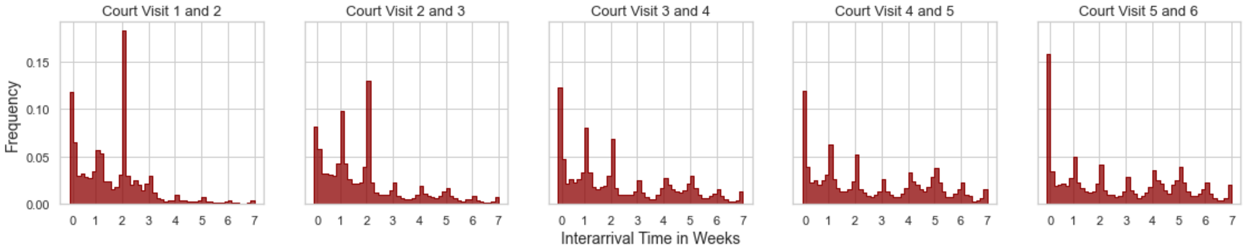


Figure A.4: **Interarrival Times of Court Visits.** Histograms truncated at 7 weeks for readability.

a detainee's final three court visits indicates a trial. Otherwise, we say the detainee plead guilty. 4% of cases in the data is detected to have gone to trial using this method, which is similar to reported statistics by the State's Attorney Office (Foxx, 2018).

A.3.4 Sentencing Information

Sentence Location. Sentence location refers to the location the detainees must spend the remainder of their sentence durations following the conclusion of their cases. Table A.4 displays the frequency of different sentence locations by crime class in our dataset. More severe crime classes are associated with more restrictive sentence locations.

Class	Charge Dropped or Finding of Not Guilty	Jail	Prison	Probation	Supervision
All	29%	17%	29%	18%	4%
M	25%	3%	67%	4%	0%
X	15%	2%	67%	17%	0%
1	13%	5%	54%	27%	1%
2	11%	6%	53%	29%	1%
3	17%	8%	46%	27%	2%
4	36%	8%	31%	23%	2%
A	39%	34%	7%	9%	11%
B	44%	36%	2%	4%	14%
C	52%	30%	4%	4%	10%

Table A.4: **Sentence Locations by Crime Class.** Percentages are based on crime class, i.e. rows sum to one.

Sentence Duration. Sentence duration is the amount of time the detainee is incarcerated as a result of a guilty sentence. As mentioned above, time incarcerated in jail or on EM counts toward sentence duration. Additionally, sentences to jail and prison are subject to sentence credit—various programs which reduce the portion of the sentence that the defendant must serve. The most significant is statutory sentence credit, which allows detainees to serve a fraction of their sentence duration, see Appendix A.1 for further details. We present sentences net of their sentence credits—that is, the time the detainee would actually serve.

Histograms of jail and prison sentence durations for the most common classes (1, 2, 3, 4, and A) are given in Figure A.5. Prison sentence durations are longer and have large probability masses on the year, half-year, and quarter-year marks. Jail sentence durations are shorter, but can take on many different values. We do not make use of sentence durations for probation and supervision in our analysis. A sentence location of "Charge Dropped or Finding of Not Guilty" indicates that the detainee has no sentence duration, i.e. they will no longer be detained.

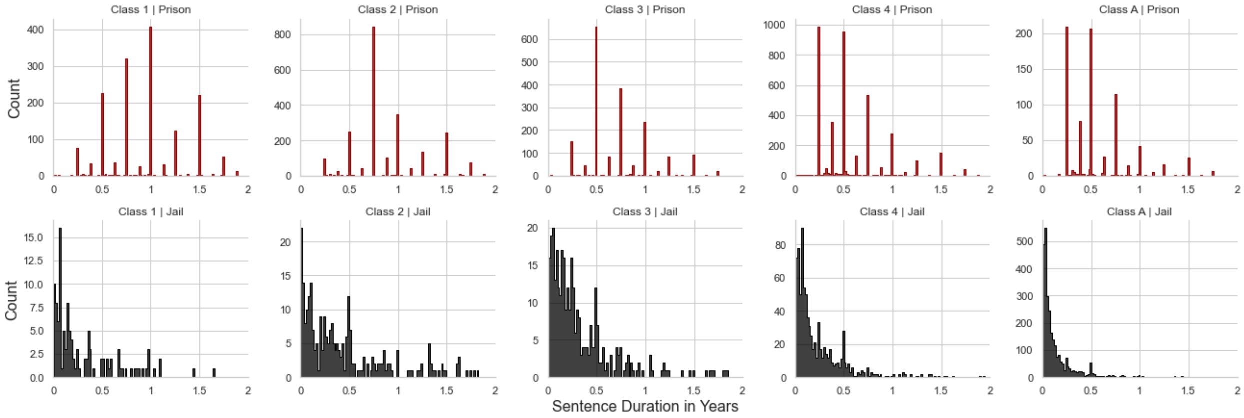


Figure A.5: **Sentence Durations in Years by Sentence Location and Crime Class.** Histrograms of prison sentence durations given on top in red, and jail sentences are given in bottom in black. Histogram bins are 1 week. Sentences are truncated at 2 years for readability.

Turnaround Status. A detainee is a "turnaround" if they are sentenced to prison and their sentence duration (including credit time) is less than their case length. That is, if they spent more time incarcerated pretrial than they were required to spend incarcerated due to their sentence.

A.4 Case Length Distributions

In our model, each detainee decides to either intentionally delay or not. Our analysis of intentional delay behavior requires an understanding of detainees' case length distributions when they intentionally delay (D) or not (N). In this section, we describe how we use the SAR-EM method to estimate those distributions.

The intentional delaying behavior is unobserved in the data. Instead, we receive a signal, τ, which partially labels the delaying detainees. That is, if the detainee delayed, $y = 1$, then there is a chance they are labeled as such: $\tau = 1$. But the remainder of the delaying detainees and all of the non-delaying detainees are unlabeled, $\tau = 0$. In other words, if a detainee is labeled, i.e. $\tau = 1$, then he delayed intentionally, i.e. $y = 1$. This exemplifies a Positive and Unlabeled (PU) dataset.

Recall that some detainees' case lengths may exceed their sentence duration, who are referred to as turnarounds. We adopt turnaround status as a signal of intentionally delaying within our dataset. That is, if the detainee is a turnaround, $\tau = 1$.

We calculate an expected probability of intentionally delaying for each detainee, $\hat{y}$, using an Expectation Maximization algorithm developed by Bekker and Davis (2018) called SAR-EM. As a preliminary to describing the algorithm, we first review the underlying probabilistic primitives. We consider our data described by a tuple (x, y, τ) whose distribution is governed by $Pr(x, y, \tau) = Pr(x)Pr(y|x)Pr(\tau|x, y)$.

Our implementation of the SAR-EM algorithm uses two machine learning models: the expected classification model $f(x|\theta)$ and the propensity score model $e(x|\phi)$. The expected classification model is used to approximate $Pr(y|x)$. The propensity score model is used to approximate $Pr(\tau|x, y)$. Both models f and e are selected from a list of classification models. They are parameterized by vectors θ and ϕ, respectively. Given the models f, e, the SAR-EM procedure starts with initial values for these parameters and updates them iteratively through the expectation and maximization steps. Each iteration starts with the current parameters, denoted by θ^{old} and ϕ^{old}. Then the expectation step calculates $\hat{y}$ for each detainee, i.e. their expected probability of intentionally delaying, using θ^{old}, ϕ^{old}. Next, given $\hat{y}$ for each detainee, the maximization step reoptimizes the parameters yielding θ^{new} and ϕ^{new}, which replace θ^{old} and ϕ^{old} as the current parameters. These two steps are performed iteratively until the algorithm converges.[2]

Next, we describe the expectation and maximization steps formally.

Expectation Step. In this step, we find $\hat{y}$ for each detainee given our current models f and e. These models are fit with the current parameters θ^{old} and ϕ^{old}. For a detainee i, we set

$$\hat{y}_i^{new} = Pr(y_i = 1|\tau_i, x_i, \theta^{old}, \phi^{old}) = \tau_i + (1-\tau_i)\frac{f(x_i|\theta^{old})(1-e(x_i|\phi^{old}))}{1-f(x_i|\theta^{old})e(x_i|\phi^{old})}.$$

In practice, the propensity score e is "decayed" by a parameter $d \in [0,1]$ to avoid local maxima where f returns 1 for any input. So, $\hat{y}$ is given by:

$$\hat{y}_i^{new} = Pr(y_i = 1|\tau_i, x_i, \theta^{old}, \phi^{old}, d) = \tau_i + (1-\tau_i)\frac{f(x_i|\theta^{old})(1-d\,e(x_i|\phi^{old}))}{1-f(x_i|\theta^{old})\,d\,e(x_i|\phi^{old})}.$$

Maximization Step. Given the updated $\hat{y}^{new}$, we find the model parameters which maximize the log-likelihood of observing x and τ. Bekker and Davis (2018) show that the models f and e which achieve this maximum satisfy the following two equations:

$$\theta^{new} = \arg\max_{\theta} \sum_{i=1}^{I} [\hat{y}_i^{new} ln\, f(x_i|\theta) + (1-\hat{y}_i^{new}) ln(1-f(x_i|\theta))]$$

$$\phi^{new} = \arg\max_{\phi} \sum_{i=1}^{I} \hat{y}_i^{new} [\tau_i ln\, e(x_i|\phi) + (1-\tau_i) ln(1-e(x_i|\phi))]$$

Then we update the parameters: $(\theta^{old}, \phi^{old}) \leftarrow (\theta^{new}, \phi^{new})$, and iterate until convergence. Convergence occurs when the change of outputs between iterations of e is smaller than some ϵ. Once the algorithm converges, letting θ^*, ϕ^* denote the final parameter values, we set

[2] At the outset of the algorithm, f and e are initialized with a short procedure training them directly on the labels τ.

the probability of delaying for detainee i as $\hat{y}_i = f(x_i|\theta^*)$. After doing this for all detainees, we use those probabilities to create binary predictions of delaying behavior by comparing them with a threshold α. Predicted probabilities greater than or equal to α are said to be delaying, $\tilde{y} = 1$, otherwise they are not, $\tilde{y} = 0$. Formally, we have

$$\tilde{y} = \begin{cases} 0 & \text{if } \hat{y} < \alpha, \\ 1 & \text{if } \hat{y} \geq \alpha. \end{cases}$$

For further details of the implementation, see Bekker and Davis (2020).

In order to choose the threshold α, we follow Lee and Liu (2003), who developed a metric for our setting (positive and unlabeled data) that approximates the traditional F_1 metric[3] and is often used in the literature. It relies on a modified recall $\hat{r} = Pr(\tilde{y} = 1|\tau = 1)$, and is given by:

$$F_1 = \frac{\hat{r}^2}{Pr(\tilde{y} = 1)}$$

The term $\hat{r}$ is the fraction of detainees labeled as intentionally delaying by the SAR-EM algorithm among those who were positively labeled in the data, i.e. $\tau = 1$. The denominator represents the fraction of detainees which SAR-EM labels positively. This metric has qualitative features that are similar to the traditional F_1 metric—for it to be high, precision and recall must be high. The higher this metric the better the classifier performs.

This modified F_1 score is a function of binary predictions, $\tilde{y}$, thus it is dependent on the threshold used to determine classification, α. The α^* which maximizes this metric is used to classify detainees for our estimation procedure.

Implementation and Resulting Case Length Distributions. We implemented the SAR-EM algorithm using Python, building on the code developed by Bekker and Davis

[3] The traditional F_1 metric, $F_1 = 2pr/(p + r)$, is the harmonic mean of precision ($p = Pr(y = 1|\tilde{y} = 1)$) and recall ($r = Pr(\tilde{y} = 1|y = 1)$). Where $\tilde{y}$ are the binary predicted classifications from f. Notice that for a high F_1 score, precision and recall must be high. However, in the PU setting that information on y is obscured, so this metric cannot be used. Thus a similar metric is necessary for PU settings.

(2020). The data used is briefly summarized in Table A.5 and is discussed in detail in Section 1.4. In particular, we restrict our attention to detainees with prison sentences, as turnarounds are used as a signal for classification purposes.

Features
Phase 1 Housing Location
Crime Class
Case Length
Sentence Duration
The Ratio of Case Length and Sentence Duration
The Average of the First Five Court Visit Interarrival Times
Number of Court Visits
Security Classification
Z-score of Case Length Grouped by Crime Class

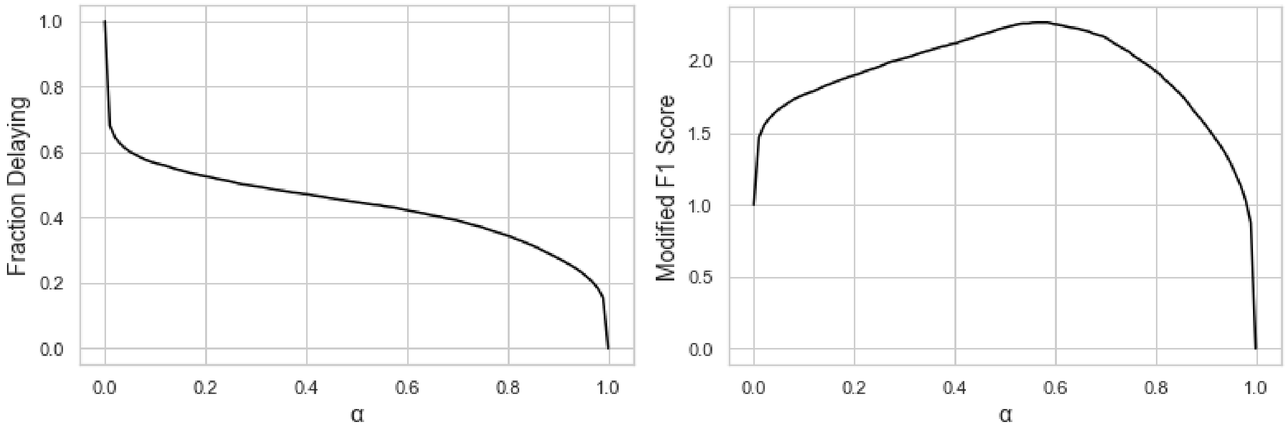
Table A.5: **Features Used in SAR-EM Classification.** The listed features are used as data in SAR-EM. We also include polynomial transformations of all features up to quadratic terms. That is, linear terms for each, interaction terms between each, and quadratic terms for each. The resulting dataset has 8,346 rows and 209 features.

We discuss the selection of models and hyperparameters in Appendix A.5. Logistic regression performed best for both models.[4] The threshold $\alpha^* = 0.556$ achieves the maximum modified F_1 score with these parameters. The fraction of detainees labeled delaying and the performance of the classification, both over α are given in Figures A.6a and A.6b.

We classify each detainee in the dataset using our threshold α^*. After classification, 43.4% of detainees are predicted to have delayed. We group their case length distributions by both crime class and pre-trial housing location so they depend on detainee characteristics.[5] Histograms of the resulting 20 case length distributions for delaying and non-delaying

[4] For f, the classification model, the regularization parameter is 2.15. For e, the propensity score model, the regularization parameter is 100. The propensity decay score is 0.8.

[5] These two covariates provide the most information about detainees' case length distributions while keeping the sample size of each group large enough for the empirical distribution to be reliable. We discuss their effects on the case length distributions in Section 1.4. To adapt the empirical distributions into F_D and F_N, we bucket case lengths into month-wide bins. This mimics the accuracy of detainees' ability to control their case lengths—average interarrival times of court dates are nearly a month—and helps ensure that the



(a) **Fraction Delaying over α.**

(b) **Modified F_1 Score Over α.**

Figure A.6: **Results of SAR-EM Classification.** SAR-EM produces a predicted probability of delaying for each detainee. As α changes, so does the performance of the classification and fraction of detainees labeled positively. The maximal modified F_1 score of 2.267 is achieved at $\alpha^* = 0.556$.

detainees are given in Figure A.7. These are used as F_D and F_N in our estimation described in Section 1.5. Notice that non-delaying distributions are smaller than delaying distributions, and tend to decrease monotonically as case length increases. Case lengths for delaying detainees tend to be longer and have increased probability densities near common sentence durations.

A.5 Hyperparameter Selection for SAR-EM Algorithm

We tested the panel of models and hyperparameters outlined in Table A.6. We also varied the propensity score decay parameter d in $\{.5, .6, .7, .9, 1\}$. To do so, we split our data into a training set and a test set, following (Lee and Liu, 2003), accounting for 80% and 20% of the overall data, respectively. Within this split, we ensure that the same proportion of detainees were labeled positively in the train and test sets.[6] We begin by training the

cdf of the distribution has robust support along its domain.

[6] That is, we seperate positive and negative samples, then select 20% from each group to create the test set.

models on the train set. Then, to evaluate the models, we use the trained models to classify the test set. The models are ranked by the maximum modified F_1 score they achieve on the test set, where a larger modified F_1 score indicates better performance. Once the best performing model and hyperparameter combination is selected, we train the best performing model on the entire dataset to classify the data for use in our model. Logistic regression performed best. Logistic regression is also used by Bekker and Davis (2018), as it is "known to predicted well-calibrated probabilities (Niculescu-Mizil and Caruana, 2005)".

Model	**Hyperparameter**	**Range**
Logistic Regression	Regularization	[0.001, 1000]
Random Forest	# Estimators	[50, 200]
	Max Features	sqrt(# features), log_2(# features)
	Max Depth	[50, 110]
	Min Samples per Split	{2, 5}
	Min Samples per Leaf	{1, 2, 4}
	Bootstrap	{True, False}
Deep Neural Net	Hidden Layers	{5, 10, 50}
	Nodes per Layer	{10, 50, 100}
	Dropout Percentage	{.1, .3, .5}
	Loss Function	Binary Crossentropy
	Activation Function	relu
	Epochs	1, 5, 10
	Batch Size	{32, 64, 500, 1000}

Table A.6: **Tested Models and Hyperparameters for f and e in SAR-EM.** Logistic Regression and Random Forest classifiers are implemented in Scikit Learn. The Deep Neural Net classifier was implemented in TensorFlow's Keras. 100 combinations of the above hyperparameters were tested for each model.

A.6 Monte Carlo Experiments

This section uses Monte Carlo experiments to evaluate the maximum likelihood estimation procedure described in Section 1.5 to identify true location cost parameters. We generate 100 datasets assuming the parameters listed in Table 1.2 are true, then estimate new parameters from the simulated data. We confirm that the true parameters fall within a 95% confidence interval of these estimates.

To simulate our data, we independently draw each detainee's location costs assuming the true μ and σ parameters for each location. Then, holding our estimates of the case length distributions F_D and F_N constant, we determine each detainee's cost for delaying and not delaying using their cost function (Equation (1.1)). They choose the action with the lower cost. Using their housing location, crime class, and action, we draw their case length w (with replacement) from the appropriate F. Finally, we re-estimate their location cost parameters, $\hat{\mu}$ and $\hat{\sigma}$, as before. We construct a 95% confidence interval by removing the upper and lower 2.5% of the re-estimated parameters. The true parameters and 95% confidence interval of the re-estimated parameters are listed in Table A.7.

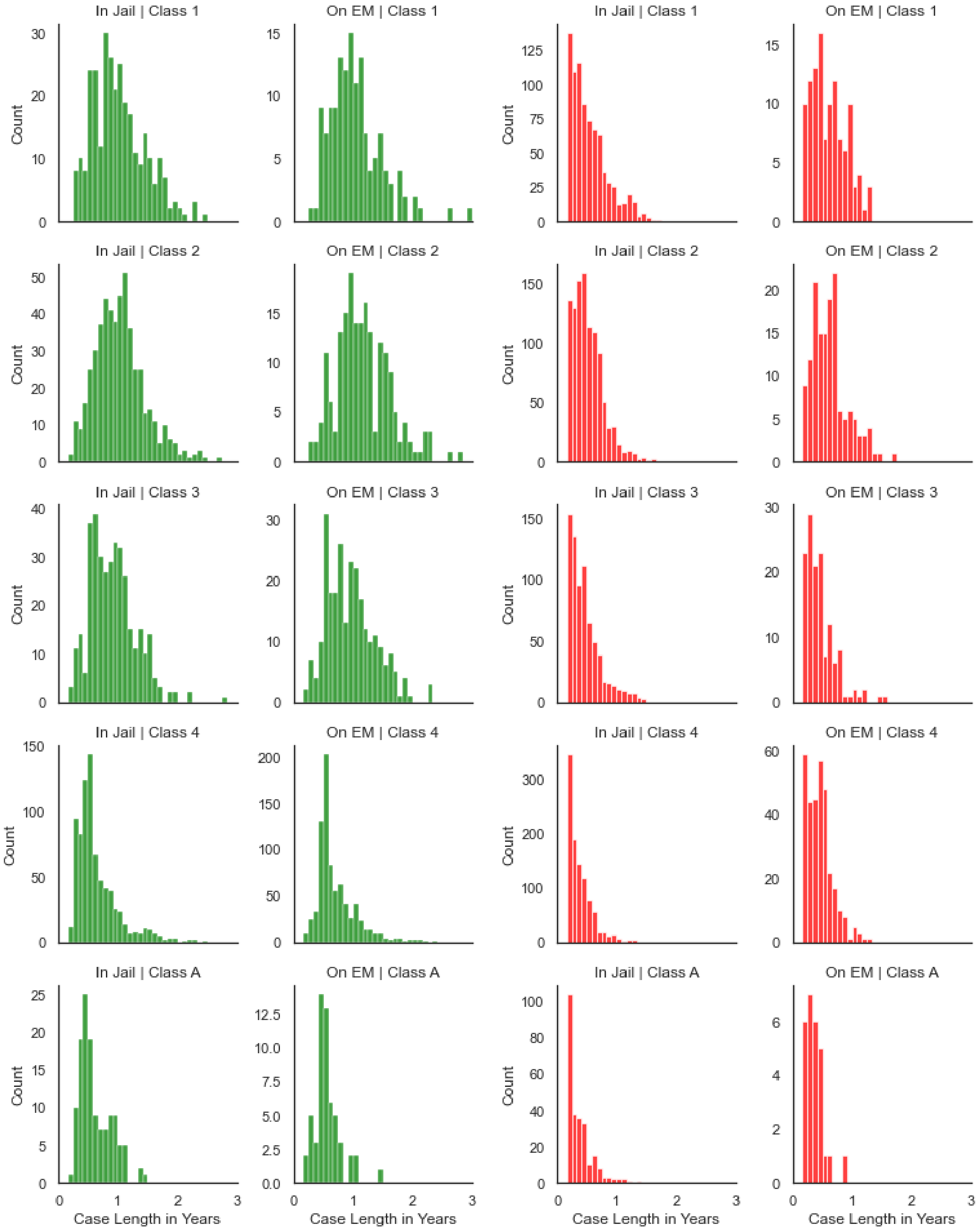
Location	True μ	95% CI	True σ	95% CI
EM	0.370	[0.211, 0.457]	0.241	[0.138, 0.333]
Jail	1.378	[1.034, 1.614]	0.446	[0.365, 0.631]
Prison	1.835	[1.325, 2.024]	0.313	[0.107, 0.470]

Table A.7: **Results of Monte Carlo Experiments.** The true parameters are the location cost parameters listed in Table 1.2. Each is presented alongside the associated 95% confidence interval of the re-estimated parameters from simulated data assuming these true parameters.

We observe that the true estimated parameters fall within the confidence interval for each location. This demonstrates that our estimation procedure can successfully recover the true structural parameters from the data.

A.7 Federal Sentencing Guidelines

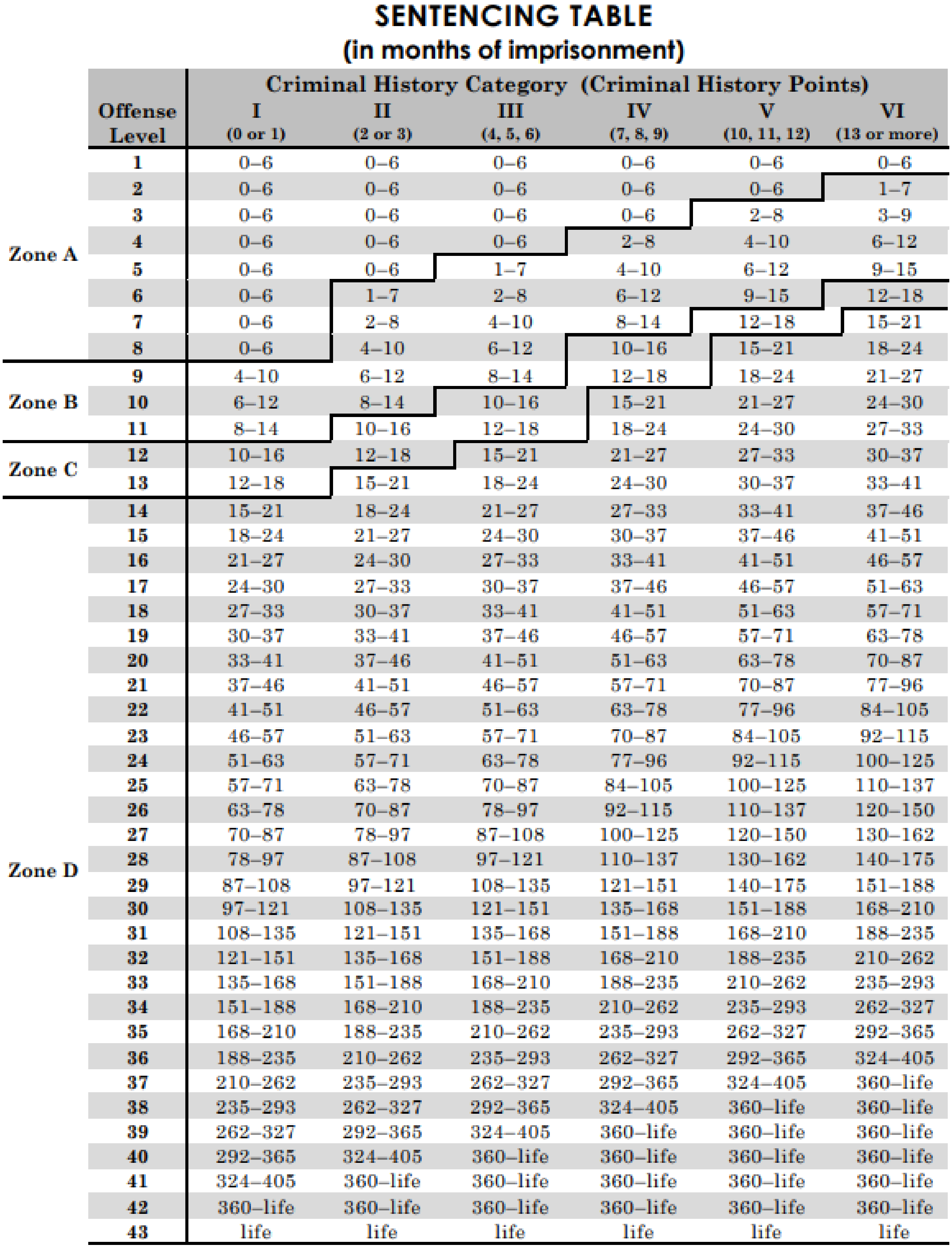
Figure A.8 is a reproduction of the Sentencing Table set forth in the U.S. Sentencing Guidelines §5A (U.S. Sentencing Comm'n, 2018). These guidelines use "offense level", which are a similar, but more granular classification of charges' severity to Illinois' crime classes. Higher offense levels indicate more serious charges.



(a) **Delaying Case Lengths:** F_D

(b) **Non-Delaying Case Lengths:** F_N

Figure A.7: **Histograms of Estimated Case Lengths.** Histograms of F_D and F_N following the SAR-EM classification. In this classification procedure, we restrict our attention to detainees with prison sentences.



SENTENCING TABLE
(in months of imprisonment)

	Offense Level	Criminal History Category (Criminal History Points) I (0 or 1)	 II (2 or 3)	 III (4, 5, 6)	 IV (7, 8, 9)	 V (10, 11, 12)	 VI (13 or more)
Zone A	1	0–6	0–6	0–6	0–6	0–6	0–6
	2	0–6	0–6	0–6	0–6	0–6	1–7
	3	0–6	0–6	0–6	0–6	2–8	3–9
	4	0–6	0–6	0–6	2–8	4–10	6–12
	5	0–6	0–6	1–7	4–10	6–12	9–15
	6	0–6	1–7	2–8	6–12	9–15	12–18
	7	0–6	2–8	4–10	8–14	12–18	15–21
	8	0–6	4–10	6–12	10–16	15–21	18–24
Zone B	9	4–10	6–12	8–14	12–18	18–24	21–27
	10	6–12	8–14	10–16	15–21	21–27	24–30
	11	8–14	10–16	12–18	18–24	24–30	27–33
Zone C	12	10–16	12–18	15–21	21–27	27–33	30–37
	13	12–18	15–21	18–24	24–30	30–37	33–41
Zone D	14	15–21	18–24	21–27	27–33	33–41	37–46
	15	18–24	21–27	24–30	30–37	37–46	41–51
	16	21–27	24–30	27–33	33–41	41–51	46–57
	17	24–30	27–33	30–37	37–46	46–57	51–63
	18	27–33	30–37	33–41	41–51	51–63	57–71
	19	30–37	33–41	37–46	46–57	57–71	63–78
	20	33–41	37–46	41–51	51–63	63–78	70–87
	21	37–46	41–51	46–57	57–71	70–87	77–96
	22	41–51	46–57	51–63	63–78	77–96	84–105
	23	46–57	51–63	57–71	70–87	84–105	92–115
	24	51–63	57–71	63–78	77–96	92–115	100–125
	25	57–71	63–78	70–87	84–105	100–125	110–137
	26	63–78	70–87	78–97	92–115	110–137	120–150
	27	70–87	78–97	87–108	100–125	120–150	130–162
	28	78–97	87–108	97–121	110–137	130–162	140–175
	29	87–108	97–121	108–135	121–151	140–175	151–188
	30	97–121	108–135	121–151	135–168	151–188	168–210
	31	108–135	121–151	135–168	151–188	168–210	188–235
	32	121–151	135–168	151–188	168–210	188–235	210–262
	33	135–168	151–188	168–210	188–235	210–262	235–293
	34	151–188	168–210	188–235	210–262	235–293	262–327
	35	168–210	188–235	210–262	235–293	262–327	292–365
	36	188–235	210–262	235–293	262–327	292–365	324–405
	37	210–262	235–293	262–327	292–365	324–405	360–life
	38	235–293	262–327	292–365	324–405	360–life	360–life
	39	262–327	292–365	324–405	360–life	360–life	360–life
	40	292–365	324–405	360–life	360–life	360–life	360–life
	41	324–405	360–life	360–life	360–life	360–life	360–life
	42	360–life	360–life	360–life	360–life	360–life	360–life
	43	life	life	life	life	life	life

Figure A.8: **United States Sentencing Table.**

APPENDIX B

APPENDICES FOR CHAPTER 3

B.1 Experimental Design for All Treatments

Set	Pay	Wait	Var	#	Set	Pay	Wait	Var	#
1	$4.75	4.6	0.25	5	11	$5.00	4.6	0.00	2
	5.00	6.9	1.00	8		4.75	11.5	0.25	8
	4.25	13.8	0.00	2		4.25	6.9	1.00	5
2	4.50	9.2	0.00	5	12	4.75	4.6	0.00	5
	4.25	11.5	1.00	2		5.00	9.2	1.00	2
	4.00	13.8	0.25	8		4.00	6.9	0.25	8
3	4.75	9.2	0.00	8	13	4.75	9.2	0.25	2
	5.00	13.8	1.00	5		4.50	4.6	1.00	8
	4.50	11.5	0.25	2		4.00	13.8	0.00	5
4	5.00	9.2	0.25	2	14	4.50	6.9	0.25	5
	4.75	4.6	1.00	5		4.75	11.5	1.00	8
	4.25	13.8	0.00	8		4.00	9.2	0.00	2
5	5.00	11.5	0.00	5	15	5.00	6.9	0.00	5
	4.25	6.9	0.25	8		4.50	4.6	0.25	8
	4.50	13.8	1.00	2		4.25	13.8	1.00	2
6	5.00	13.8	0.25	8	16	4.50	11.5	0.00	2
	4.25	6.9	0.00	2		4.25	9.2	0.25	8
	4.00	9.2	1.00	5		4.00	6.9	1.00	5
7	4.75	13.8	0.00	5	17	4.75	6.9	0.00	2
	4.25	4.6	1.00	8		4.50	9.2	1.00	5
	4.00	11.5	0.25	2		4.00	11.5	0.25	8
8	5.00	11.5	1.00	8	18	5.00	4.6	0.00	8
	4.50	9.2	0.25	5		4.75	13.8	0.25	2
	4.00	4.6	0.00	2		4.00	11.5	1.00	5
9	4.50	6.9	0.00	2	19	4.50	11.5	0.00	8
	4.25	4.6	0.25	5		4.25	9.2	0.25	5
	4.75	9.2	1.00	8		4.00	4.6	1.00	2
10	4.75	6.9	1.00	2	20	5.00	4.6	0.25	2
	5.00	13.8	0.25	5		4.25	11.5	0.00	5
	4.00	9.2	0.00	8		4.50	13.8	1.00	8

Table B.1: **Experimental Design for All Treatments.** Wait and Var refer to aggregate mean wait time and wait variance, respectively

B.2 Raw Model Estimates by Treatment

	Queue				Clock			
Information	Pay	Mean	Variance	# People	Pay	Mean	Variance	# Updates
Full–NoPrior	2.74^{***}	$-.173^{***}$	$-.615^{***}$	$-.056^{***}$	3.71^{***}	$-.389^{***}$	$-.313^{***}$	$-.021$
	(.052)	(.009)	(.058)	(.010)	(.178)	(.018)	(.064)	(.013)
Full+Prior	2.32^{***}	$-.117^{***}$	$-.241^{***}$	.003	2.24^{***}	$-.208^{***}$	$-.193^{**}$	$-.023$
	(.107)	(.010)	(.068)	(.012)	(.111)	(.011)	(.067)	(.012)
Mean+Prior	2.22^{***}	$-.125^{***}$	$-.224^{***}$	$.028^{***}$	2.00^{***}	$-.241^{***}$	$-.181^{***}$	$-.022^{*}$
	(.014)	(.007)	(.047)	(.009)	(.017)	(.008)	(.029)	(.009)
None+Prior	2.21^{***}	$-.117^{***}$	$-.244^{**}$	.001	2.12^{***}	$-.232^{***}$	$-.153^{***}$	$-.026^{***}$
	(.097)	(.007)	(.094)	(.009)	(.069)	(.007)	(.040)	(.007)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table B.2: **Raw Model Estimates by Treatment.**

B.3 Sample Choice Sets by Treatment

Table B.3 shows the attribute levels for three example waits of our experiment. The full details for all 20 choice sets appears in Table B.1 of the main paper. Table B.4 lists the same attribute levels, and, where appropriate, converts them to per-person levels, as is used in the Queue treatments in our experiment. Table B.5 lists the attribute levels and associated average wait time and uncertainty range as used in the Clock treatments in our experiment. Pictorial examples of how these three profiles would be presented in our experiment are shown in the following pages.

Set	Pay	Wait	Var	#
	$4.75	4.6	0.25	5
1	5.00	6.9	1.00	8
	4.25	13.8	0.00	2
	4.50	9.2	0.00	5
2	4.25	11.5	1.00	2
	4.00	13.8	0.25	8
	4.75	9.2	0.00	8
3	5.00	13.8	1.00	5
	4.50	11.5	0.25	2

Table B.3: **Attribute Levels for Three Example Choice Sets.**

Set	Pay	Wait/#	Var/#	#	Avg Service Time/#	Uncertainty Range/#
	$4.75	0.92	0.05	5	55s	±22s
1	5.00	0.86	0.13	8	51s	±34s
	4.25	6.90	0.00	2	6m 53s	n/a
	4.50	1.84	0.00	5	1m 50s	n/a
2	4.25	5.75	0.50	2	5m 45s	±1m 9s
	$4.00	1.73	0.03	8	1m 43s	±17s
	4.75	1.15	0.00	8	1m 9s	n/a
3	5.00	2.76	0.20	5	2m 45s	±44s
	$4.50	5.75	0.13	2	5m 45s	±34s

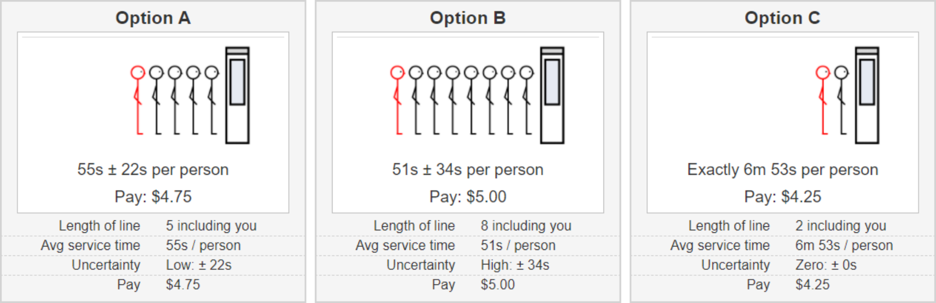
Table B.4: **Three Example Choice Sets with Per-Person Wait Statistics as Used in Queues.**

Set	Pay	Wait	Var	#	Avg Wait Time	Uncertainty Range
	$4.75	4.60	0.25	5	4m 36s	±49s
1	5.00	6.90	1.00	8	6m 53s	±1m 38s
	4.25	13.80	0.00	2	13m 47s	n/a
	4.50	9.20	0.00	5	9m 12s	n/a
2	4.25	11.50	1.00	2	11m 30s	±1m 38s
	4.00	13.80	0.25	8	13m 47s	±49s
	4.75	9.20	0.00	8	9m 12s	n/a
3	5.00	13.80	1.00	5	13m 47s	±1m 38s
	4.50	11.50	0.25	2	11m 30s	±49s

Table B.5: **Three Example Choice Sets with Aggregate Wait Statistics as Used in Clocks.**

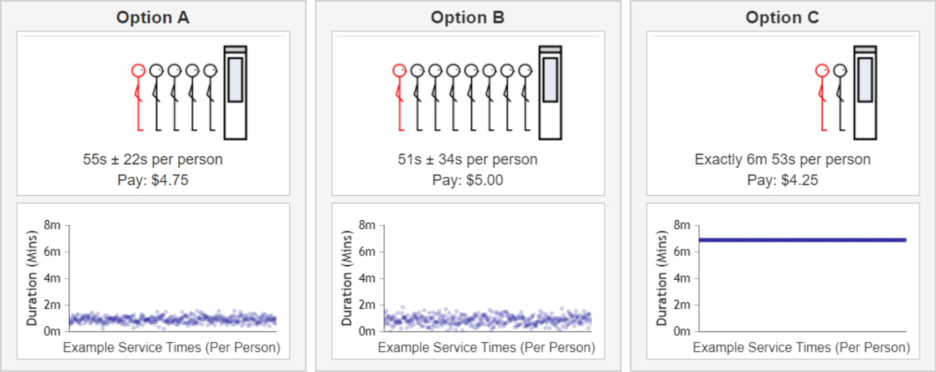
Choice Set 1, Queues

Figure B.1



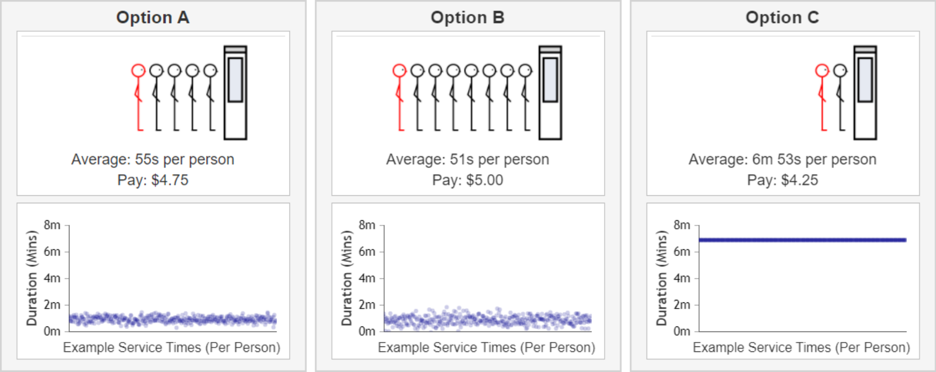
	Option A	Option B	Option C
	55s ± 22s per person	51s ± 34s per person	Exactly 6m 53s per person
	Pay: $4.75	Pay: $5.00	Pay: $4.25
Length of line	5 including you	8 including you	2 including you
Avg service time	55s / person	51s / person	6m 53s / person
Uncertainty	Low: ± 22s	High: ± 34s	Zero: ± 0s
Pay	$4.75	$5.00	$4.25

Q:Full–NoPrior



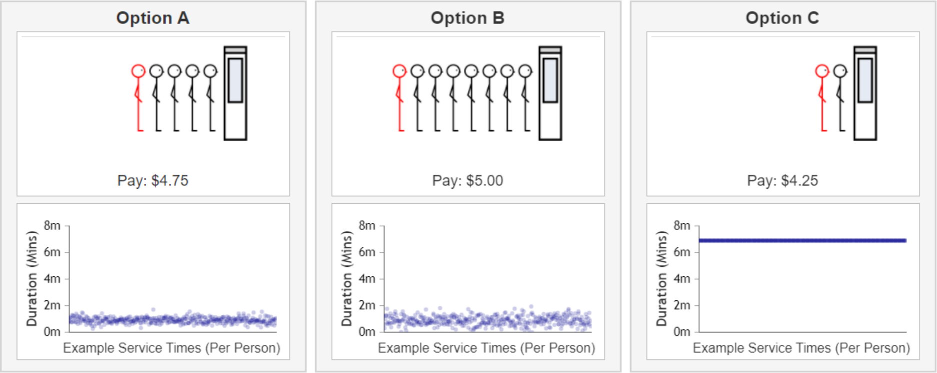
Option A	Option B	Option C
55s ± 22s per person	51s ± 34s per person	Exactly 6m 53s per person
Pay: $4.75	Pay: $5.00	Pay: $4.25

Q:Full+Prior



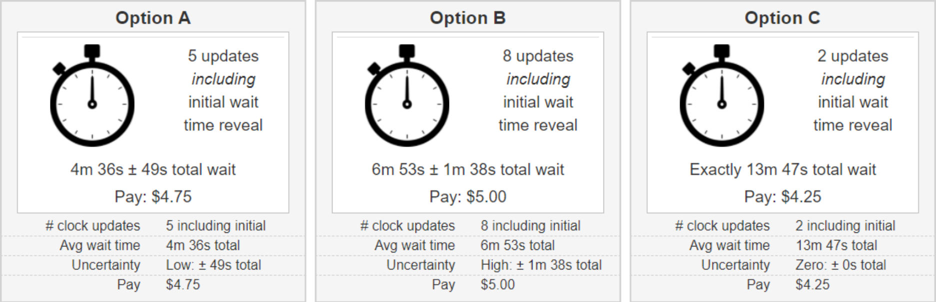
Option A	Option B	Option C
Average: 55s per person	Average: 51s per person	Average: 6m 53s per person
Pay: $4.75	Pay: $5.00	Pay: $4.25

Q:Mean+Prior


Option A
Pay: $4.75
Duration (Mins)
8m
6m
4m
2m
0m
Example Service Times (Per Person)
Option B
Pay: $5.00
Duration (Mins)
8m
6m
4m
2m
0m
Example Service Times (Per Person)
Option C
Pay: $4.25
Duration (Mins)
8m
6m
4m
2m
0m
Example Service Times (Per Person)

Q:None+Prior

Choice Set 1, Clocks



Option A

5 updates *including* initial wait time reveal

4m 36s ± 49s total wait

Pay: $4.75

# clock updates	5 including initial
Avg wait time	4m 36s total
Uncertainty	Low: ± 49s total
Pay	$4.75

Option B

8 updates *including* initial wait time reveal

6m 53s ± 1m 38s total wait

Pay: $5.00

# clock updates	8 including initial
Avg wait time	6m 53s total
Uncertainty	High: ± 1m 38s total
Pay	$5.00

Option C

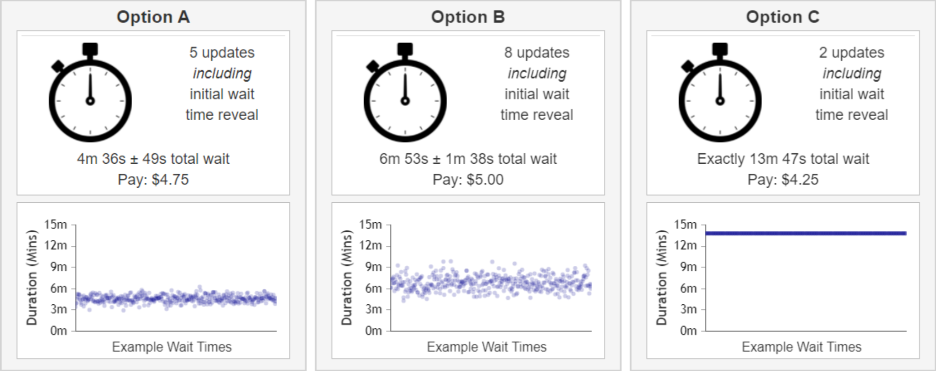
2 updates *including* initial wait time reveal

Exactly 13m 47s total wait

Pay: $4.25

# clock updates	2 including initial
Avg wait time	13m 47s total
Uncertainty	Zero: ± 0s total
Pay	$4.25

C:Full–NoPrior



Option A

5 updates *including* initial wait time reveal

4m 36s ± 49s total wait

Pay: $4.75

Option B

8 updates *including* initial wait time reveal

6m 53s ± 1m 38s total wait

Pay: $5.00

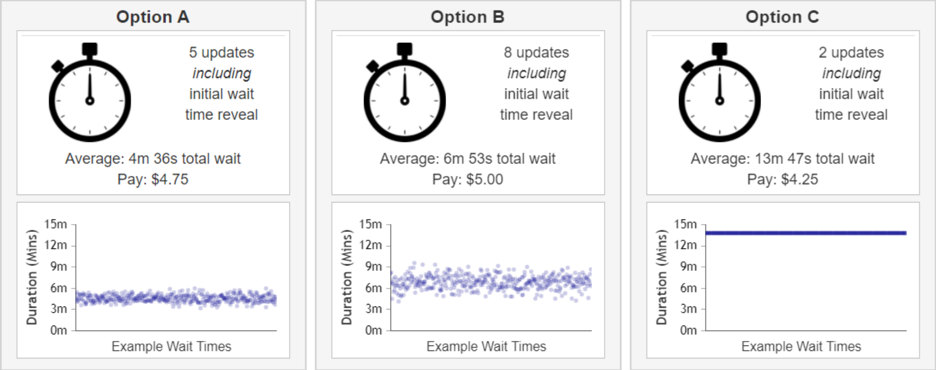
Option C

2 updates *including* initial wait time reveal

Exactly 13m 47s total wait

Pay: $4.25

C:Full+Prior



Option A

5 updates *including* initial wait time reveal

Average: 4m 36s total wait

Pay: $4.75

Option B

8 updates *including* initial wait time reveal

Average: 6m 53s total wait

Pay: $5.00

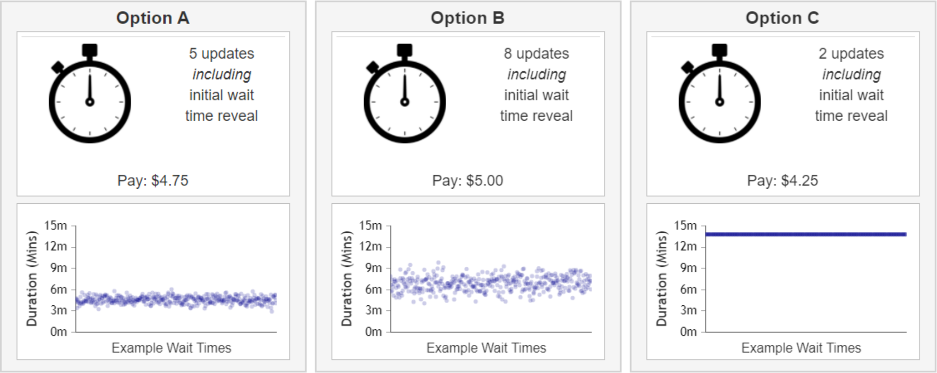
Option C

2 updates *including* initial wait time reveal

Average: 13m 47s total wait

Pay: $4.25

C:Mean+Prior



Option A

5 updates *including* initial wait time reveal

Pay: $4.75

Option B

8 updates *including* initial wait time reveal

Pay: $5.00

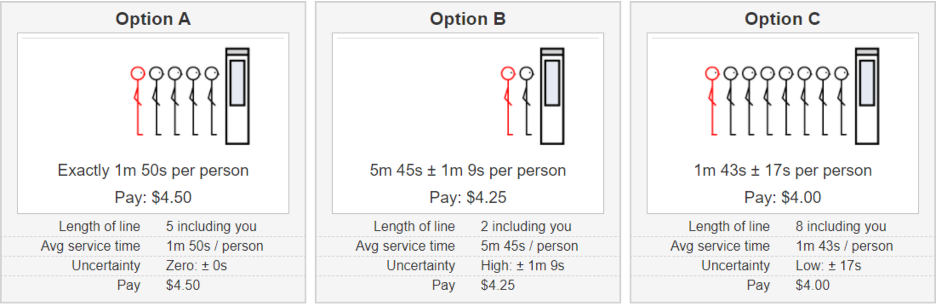
Option C

2 updates *including* initial wait time reveal

Pay: $4.25

C:None+Prior

Choice Set 2, Queues



Option A

Exactly 1m 50s per person

Pay: $4.50

Length of line	5 including you
Avg service time	1m 50s / person
Uncertainty	Zero: ± 0s
Pay	$4.50

Option B

5m 45s ± 1m 9s per person

Pay: $4.25

Length of line	2 including you
Avg service time	5m 45s / person
Uncertainty	High: ± 1m 9s
Pay	$4.25

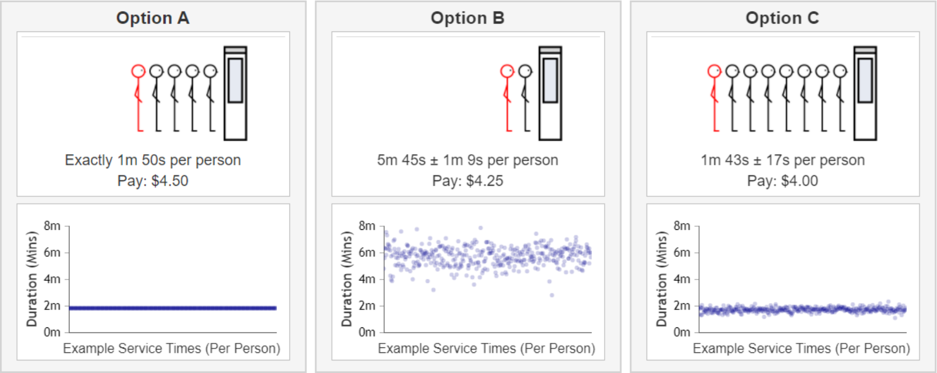
Option C

1m 43s ± 17s per person

Pay: $4.00

Length of line	8 including you
Avg service time	1m 43s / person
Uncertainty	Low: ± 17s
Pay	$4.00

Q:Full–NoPrior



Option A

Exactly 1m 50s per person

Pay: $4.50

Option B

5m 45s ± 1m 9s per person

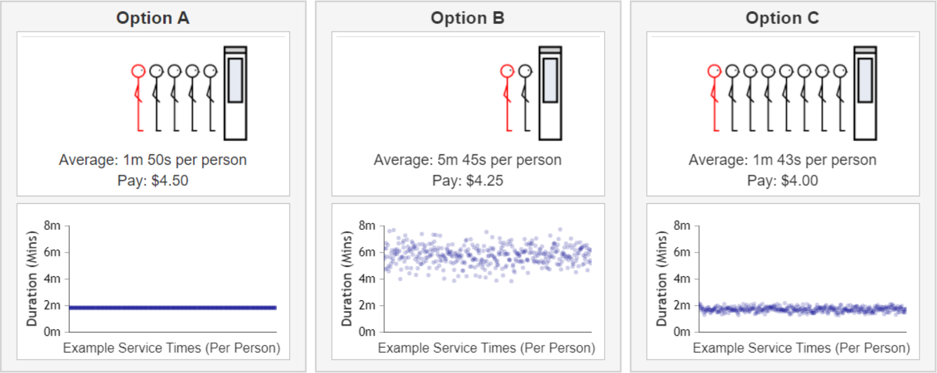
Pay: $4.25

Option C

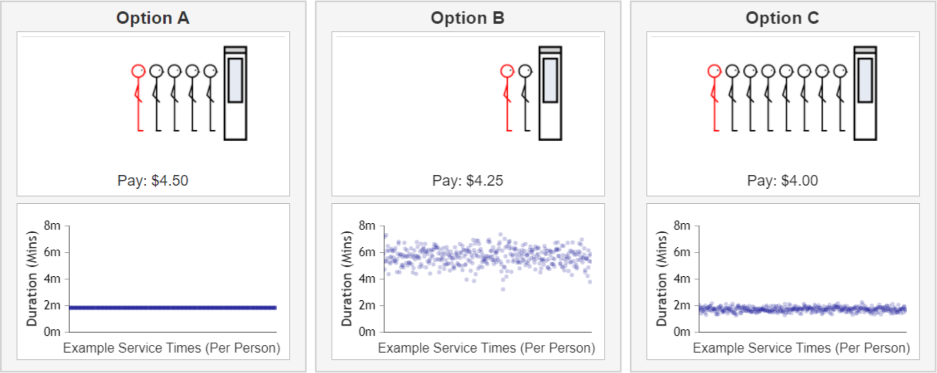
1m 43s ± 17s per person

Pay: $4.00

Q:Full+Prior

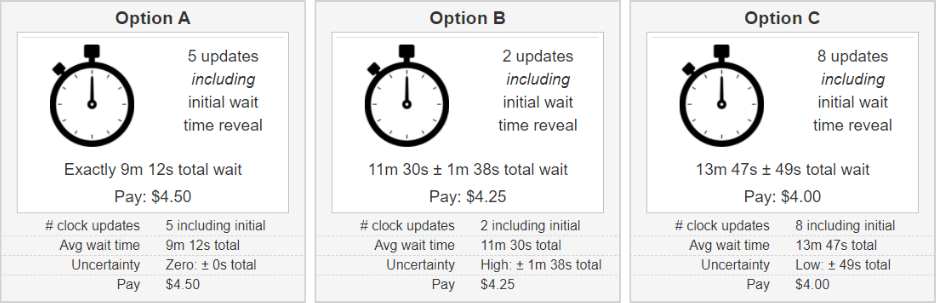

Option A
Average: 1m 50s per person
Pay: $4.50
8m
6m
4m
2m
0m
Duration (Mins)
Example Service Times (Per Person)
Option B
Average: 5m 45s per person
Pay: $4.25
8m
6m
4m
2m
0m
Duration (Mins)
Example Service Times (Per Person)
Option C
Average: 1m 43s per person
Pay: $4.00
8m
6m
4m
2m
0m
Duration (Mins)
Example Service Times (Per Person)

Q:Mean+Prior



Q:None+Prior

Choice Set 2, Clocks



Option A

5 updates *including* initial wait time reveal

Exactly 9m 12s total wait

Pay: $4.50

# clock updates	5 including initial
Avg wait time	9m 12s total
Uncertainty	Zero: ± 0s total
Pay	$4.50

Option B

2 updates *including* initial wait time reveal

11m 30s ± 1m 38s total wait

Pay: $4.25

# clock updates	2 including initial
Avg wait time	11m 30s total
Uncertainty	High: ± 1m 38s total
Pay	$4.25

Option C

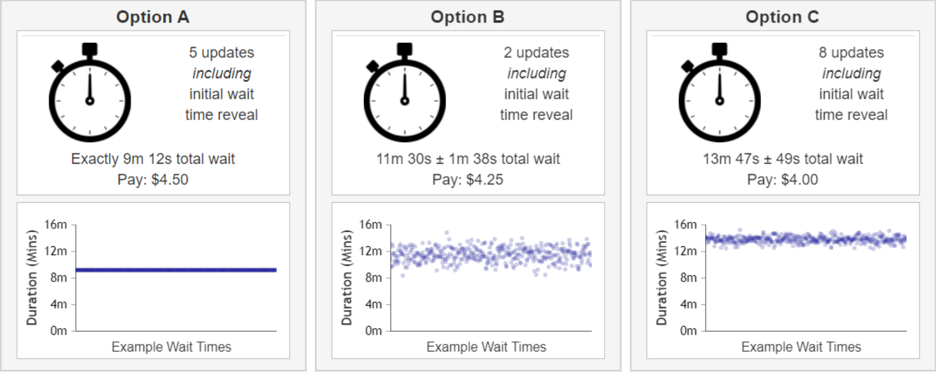
8 updates *including* initial wait time reveal

13m 47s ± 49s total wait

Pay: $4.00

# clock updates	8 including initial
Avg wait time	13m 47s total
Uncertainty	Low: ± 49s total
Pay	$4.00

C:Full–NoPrior



Option A

5 updates *including* initial wait time reveal

Exactly 9m 12s total wait

Pay: $4.50

Option B

2 updates *including* initial wait time reveal

11m 30s ± 1m 38s total wait

Pay: $4.25

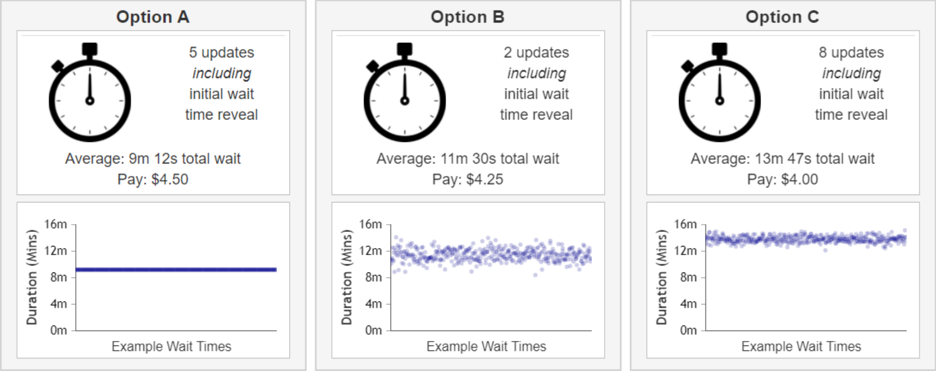
Option C

8 updates *including* initial wait time reveal

13m 47s ± 49s total wait

Pay: $4.00

C:Full+Prior



Option A

5 updates *including* initial wait time reveal

Average: 9m 12s total wait

Pay: $4.50

Option B

2 updates *including* initial wait time reveal

Average: 11m 30s total wait

Pay: $4.25

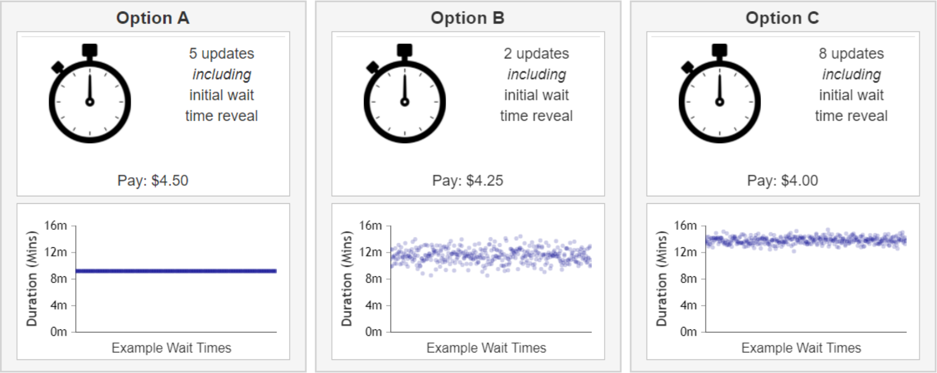
Option C

8 updates *including* initial wait time reveal

Average: 13m 47s total wait

Pay: $4.00

C:Mean+Prior



Option A

5 updates *including* initial wait time reveal

Pay: $4.50

Option B

2 updates *including* initial wait time reveal

Pay: $4.25

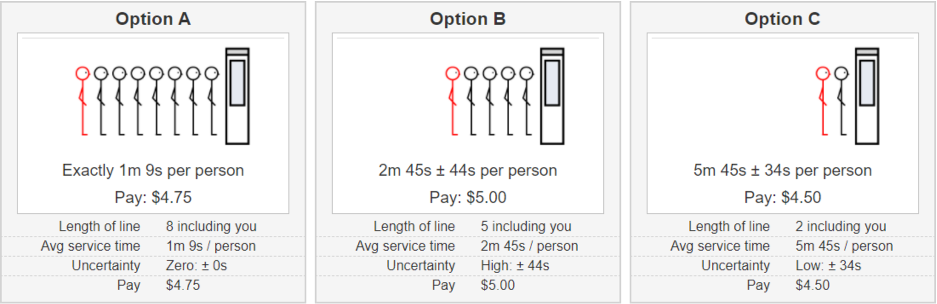
Option C

8 updates *including* initial wait time reveal

Pay: $4.00

C:None+Prior

Choice Set 3, Queues



Option A

Exactly 1m 9s per person

Pay: $4.75

Length of line	8 including you
Avg service time	1m 9s / person
Uncertainty	Zero: ± 0s
Pay	$4.75

Option B

2m 45s ± 44s per person

Pay: $5.00

Length of line	5 including you
Avg service time	2m 45s / person
Uncertainty	High: ± 44s
Pay	$5.00

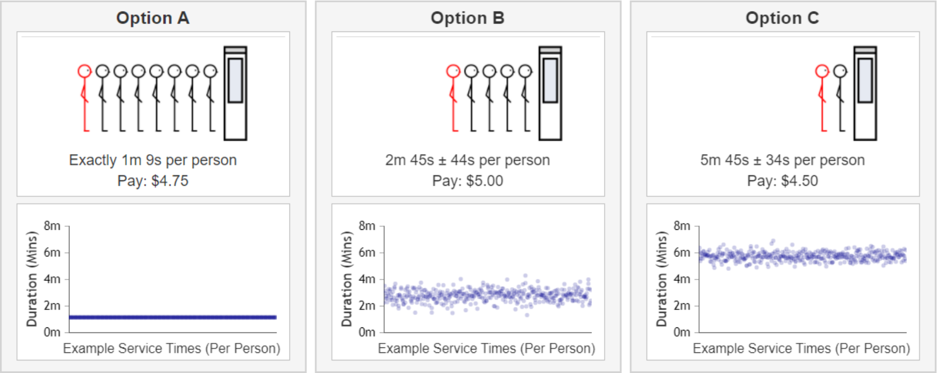
Option C

5m 45s ± 34s per person

Pay: $4.50

Length of line	2 including you
Avg service time	5m 45s / person
Uncertainty	Low: ± 34s
Pay	$4.50

Q:Full–NoPrior



Option A

Exactly 1m 9s per person

Pay: $4.75

Option B

2m 45s ± 44s per person

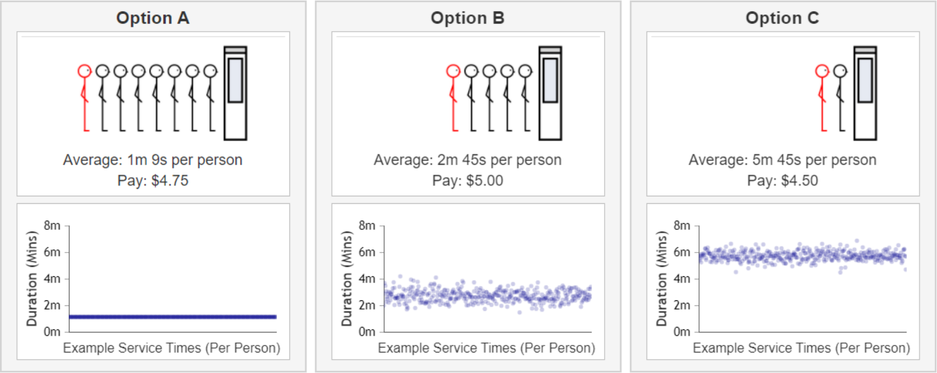
Pay: $5.00

Option C

5m 45s ± 34s per person

Pay: $4.50

Q:Full+Prior


Option A

Average: 1m 9s per person
Pay: $4.75

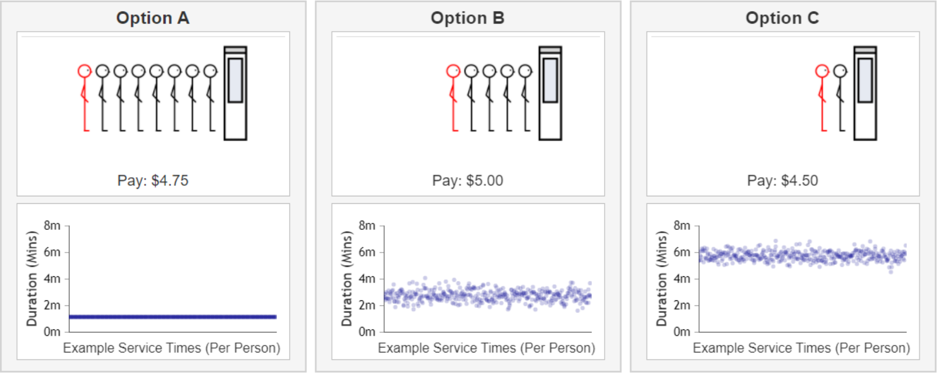
Option B

Average: 2m 45s per person
Pay: $5.00

Option C

Average: 5m 45s per person
Pay: $4.50

Q:Mean+Prior


Option A

Pay: $4.75

Option B

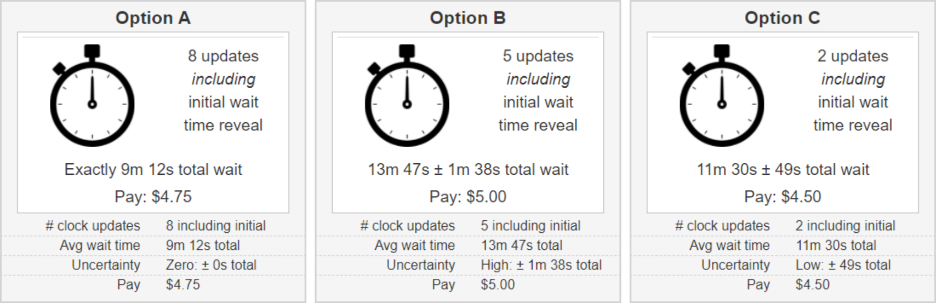
Pay: $5.00

Option C

Pay: $4.50

Q:None+Prior

Choice Set 3, Clocks



Option A

8 updates *including* initial wait time reveal

Exactly 9m 12s total wait

Pay: $4.75

# clock updates	8 including initial
Avg wait time	9m 12s total
Uncertainty	Zero: ± 0s total
Pay	$4.75

Option B

5 updates *including* initial wait time reveal

13m 47s ± 1m 38s total wait

Pay: $5.00

# clock updates	5 including initial
Avg wait time	13m 47s total
Uncertainty	High: ± 1m 38s total
Pay	$5.00

Option C

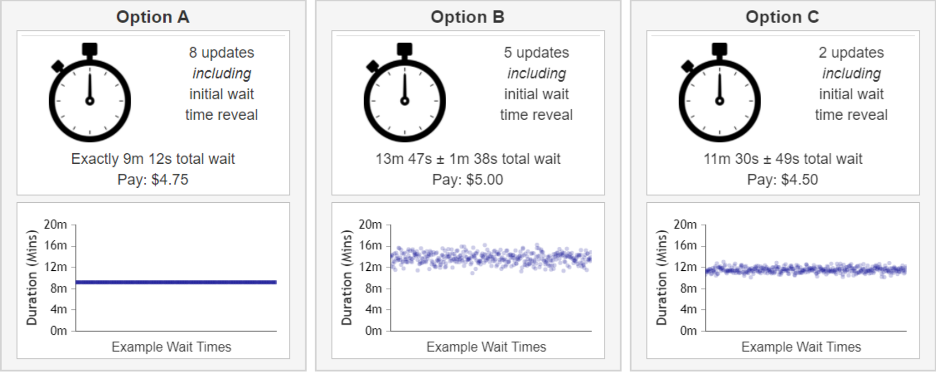
2 updates *including* initial wait time reveal

11m 30s ± 49s total wait

Pay: $4.50

# clock updates	2 including initial
Avg wait time	11m 30s total
Uncertainty	Low: ± 49s total
Pay	$4.50

C:Full–NoPrior



Option A

8 updates *including* initial wait time reveal

Exactly 9m 12s total wait

Pay: $4.75

Option B

5 updates *including* initial wait time reveal

13m 47s ± 1m 38s total wait

Pay: $5.00

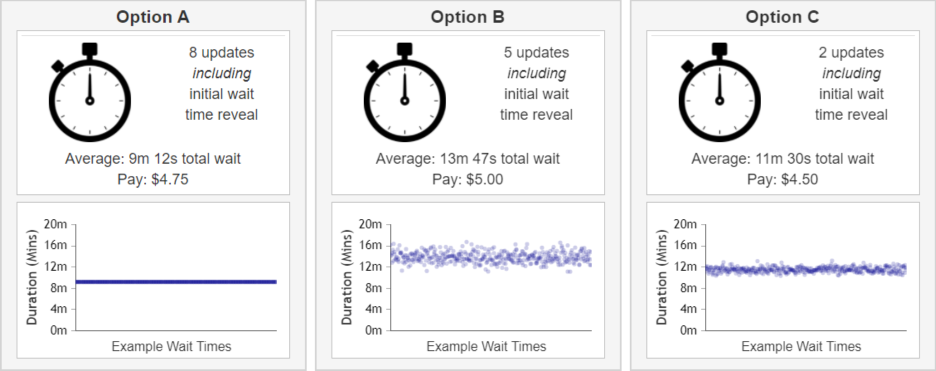
Option C

2 updates *including* initial wait time reveal

11m 30s ± 49s total wait

Pay: $4.50

C:Full+Prior



Option A

8 updates *including* initial wait time reveal

Average: 9m 12s total wait

Pay: $4.75

Option B

5 updates *including* initial wait time reveal

Average: 13m 47s total wait

Pay: $5.00

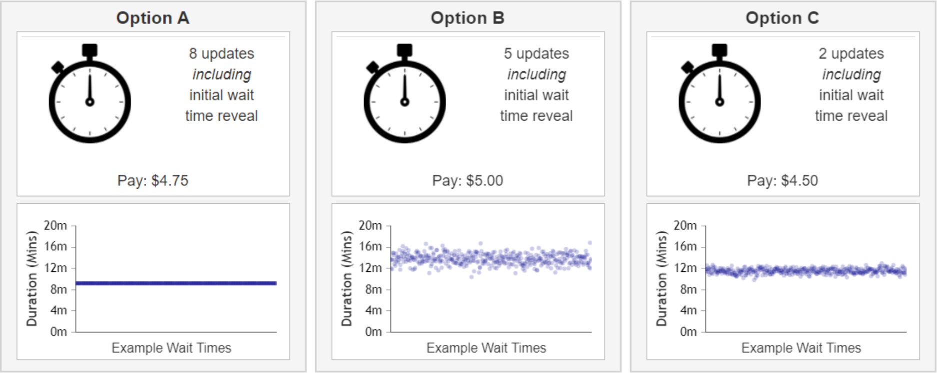
Option C

2 updates *including* initial wait time reveal

Average: 11m 30s total wait

Pay: $4.50

C:Mean+Prior



Option A

8 updates *including* initial wait time reveal

Pay: $4.75

Option B

5 updates *including* initial wait time reveal

Pay: $5.00

Option C

2 updates *including* initial wait time reveal

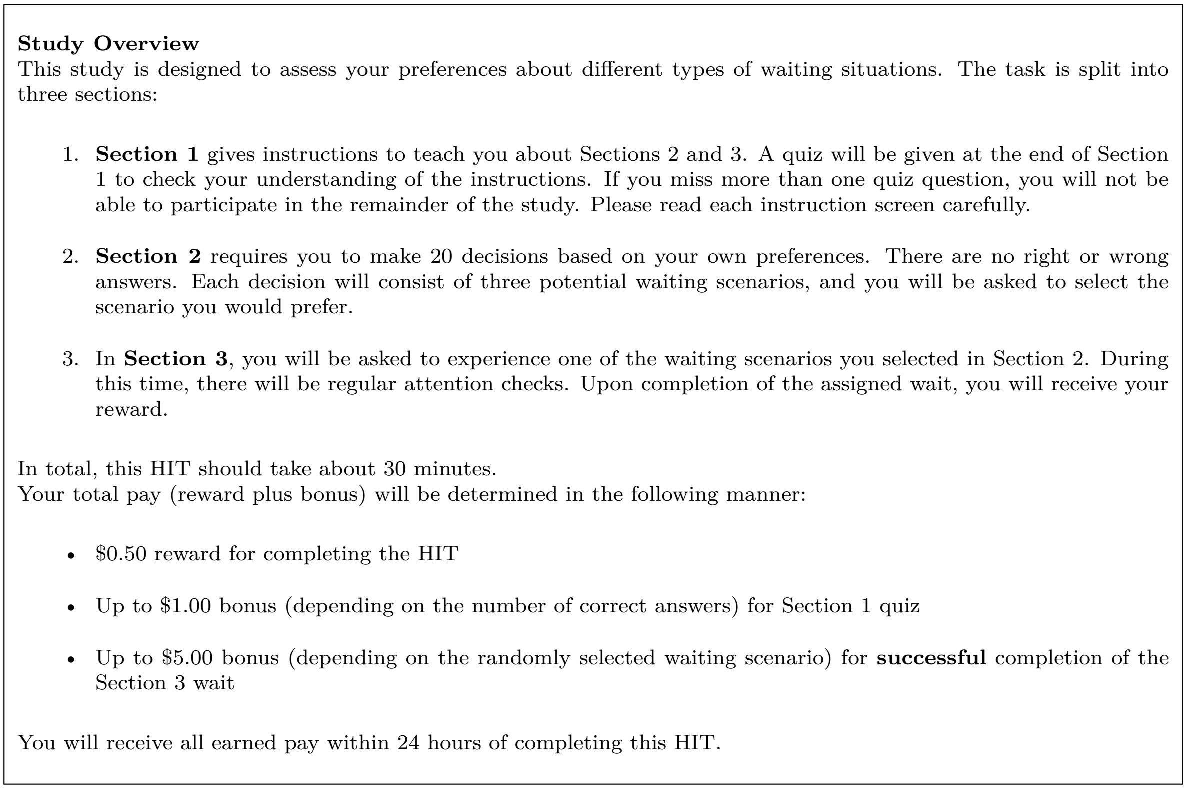
Pay: $4.50

C:None+Prior

B.4 Instructions

Here we provide comprehensive instructions for two treatments: Queue:Full+Prior and Clock:Full+Prior. The figures and descriptions were adjusted to accurately reflect the information provided in the other treatments.

Figure B.2



Study Overview
This study is designed to assess your preferences about different types of waiting situations. The task is split into three sections:

1. **Section 1** gives instructions to teach you about Sections 2 and 3. A quiz will be given at the end of Section 1 to check your understanding of the instructions. If you miss more than one quiz question, you will not be able to participate in the remainder of the study. Please read each instruction screen carefully.

2. **Section 2** requires you to make 20 decisions based on your own preferences. There are no right or wrong answers. Each decision will consist of three potential waiting scenarios, and you will be asked to select the scenario you would prefer.

3. In **Section 3**, you will be asked to experience one of the waiting scenarios you selected in Section 2. During this time, there will be regular attention checks. Upon completion of the assigned wait, you will receive your reward.

In total, this HIT should take about 30 minutes.
Your total pay (reward plus bonus) will be determined in the following manner:

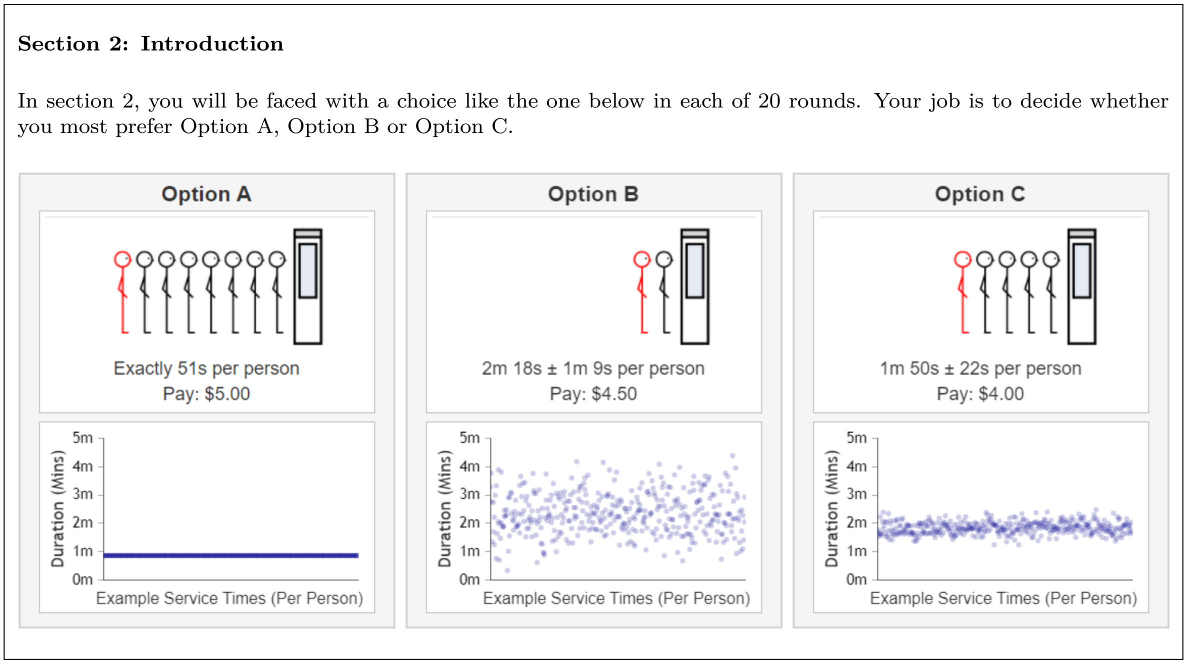
- $0.50 reward for completing the HIT

- Up to $1.00 bonus (depending on the number of correct answers) for Section 1 quiz

- Up to $5.00 bonus (depending on the randomly selected waiting scenario) for **successful** completion of the Section 3 wait

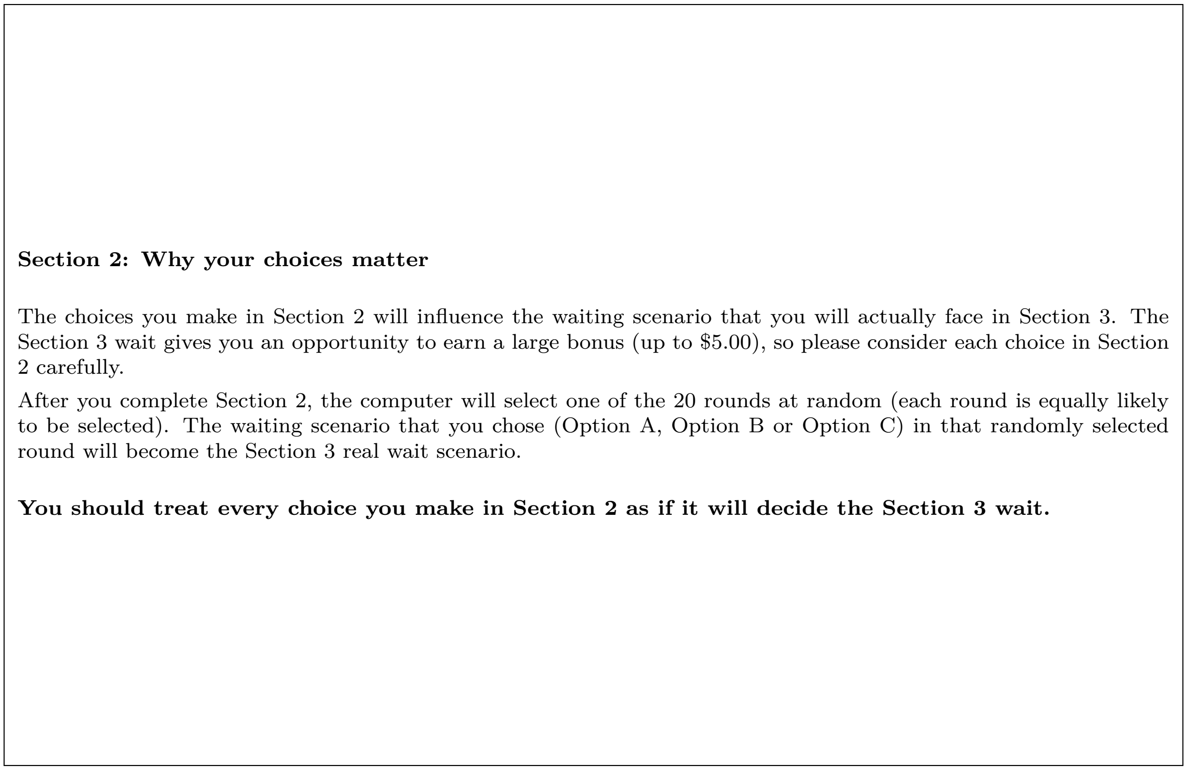
You will receive all earned pay within 24 hours of completing this HIT.

Queue:Full+Prior Instructions



Section 2: Introduction

In section 2, you will be faced with a choice like the one below in each of 20 rounds. Your job is to decide whether you most prefer Option A, Option B or Option C.



Section 2: Why your choices matter

The choices you make in Section 2 will influence the waiting scenario that you will actually face in Section 3. The Section 3 wait gives you an opportunity to earn a large bonus (up to $5.00), so please consider each choice in Section 2 carefully.

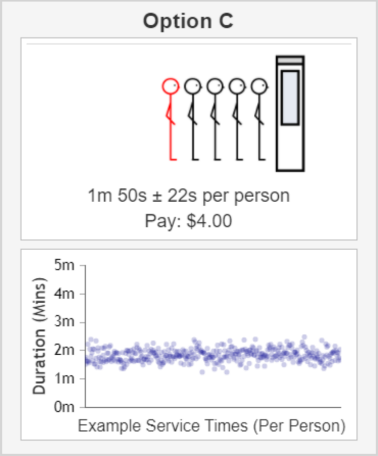
After you complete Section 2, the computer will select one of the 20 rounds at random (each round is equally likely to be selected). The waiting scenario that you chose (Option A, Option B or Option C) in that randomly selected round will become the Section 3 real wait scenario.

You should treat every choice you make in Section 2 as if it will decide the Section 3 wait.

Section 2: Detailed look waiting scenarios

Each of the waiting scenarios will require you (represented by the red figure) to join at the back of a line. Every person in line (including you) requires service (some amount of time in the booth). When the booth is empty, the person at the front of the line enters. During your wait, you will always know how many people are in front of you. However, you will never be told how much time remains in your wait.

How long will the wait take?
The total duration of a wait depends on the number of people (including you) in line and how long each person spends in the service booth. You will always know the number of people in line. In the scenario at right, there are 5 people in line (you plus 4 others).



Information about how long each person will take in the service booth is provided in two ways.

First, information about service durations appears beneath the image of the waiting line ("1m 50s ± 22s per person" in the example at right). This tells you that service takes 1m 50s on average, but that each person will take a little longer or a little shorter than this time. The "± 22s" range means that there is a 90% chance that each person's service time will be within 22 seconds of the average. In other words, 90% of the people in line will have service times between 1m 38s and 2m 12s.

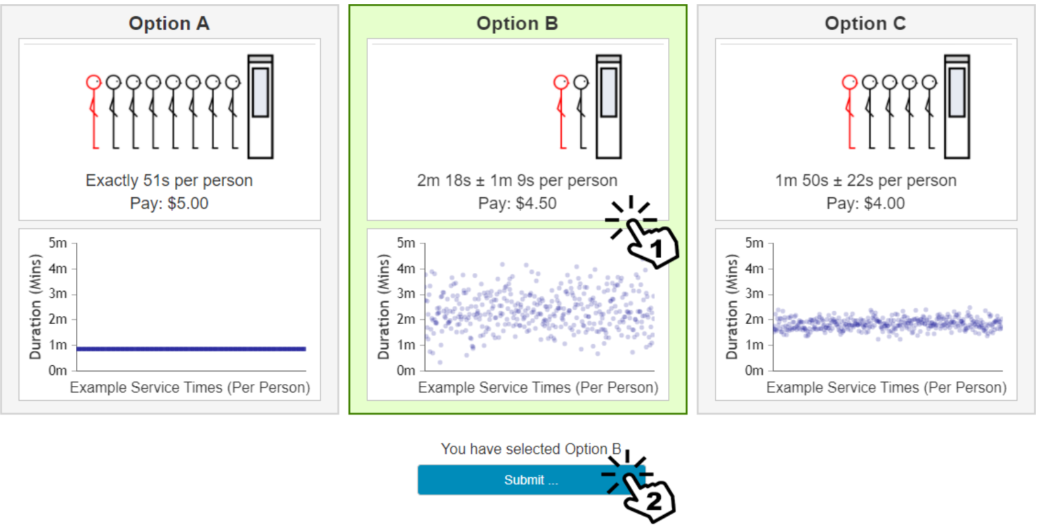
Second, information about service durations is given visually in the plot with the blue dots. Each dot represents a possible service time. You can see that most dots are slightly below 2m (because the average service time is 1m 50s), and that nearly all fall within the range of about 1.5m to just over 2m. Each person in line will have a service duration chosen at random from these dots.

How do I successfully complete a wait?
A wait is considered a success when you finish receiving service. When you successfully complete the wait, you will earn a bonus equal to the Pay listed ($4.00 in the scenario at right).

Section 2: Making your selection

In each of the 20 rounds in Section 2, you will be presented with three waiting scenarios. Your job is to carefully consider which wait you would most prefer to actually experience in Section 3.

To indicate the option (A, B or C) that you prefer, simply click anywhere inside of its box (see **1** below). Text will appear at the bottom of the screen indicating which option you have indicated. You may change your selection by clicking on another box. To finalize your choice, click the "Submit" button (see **2** below). You will then move on to the next round.



Section 3: Actual wait

After you complete Section 2, the computer will select one of the 20 rounds at random (each round is equally likely to be selected). The waiting scenario that you chose (Option A, B or C) in that randomly selected round will become the Section 3 real wait scenario.

You can think of this random selection process in another way. You can imagine that each time you choose a waiting scenario as your more preferred option in Section 2, it goes into a "bin" of preferred waits. The real wait in Section 3 will be chosen at random out of this "bin."

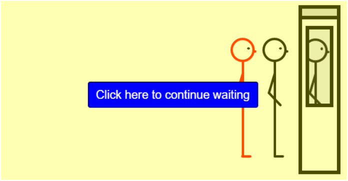
Since your choices in Section 2 directly affect the Section 3 real wait selection, you should carefully consider each decision you make.

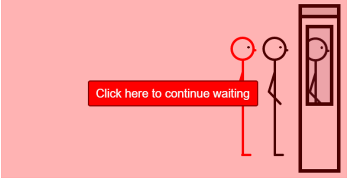
Section 3 attention checks

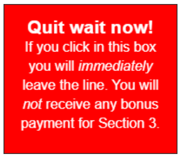
To successfully complete a wait, you must wait until you complete your service. During this time, there will be regular attention checks. Periodically, the screen will begin to flash yellow and a button will appear that says "Click here to continue waiting." You have 15 seconds to click the button. When 5 seconds remain, the screen will turn solid red.

To successfully complete the wait and earn the Section 3 pay, you must pass *all* attention checks.

You may exit the Section 3 wait at any time by clicking on the *Quit wait now!* button (see below). Clicking this button means that you will *not* earn any pay from Section 3. However, whether you successfully complete the Section 3 wait or not, you are still entitled to the HIT reward and your earnings from the quiz on Section 1.







Quit wait now!
If you click in this box you will *immediately* leave the line. You will *not* receive any bonus payment for Section 3.

Section 3: Examples of actual waits

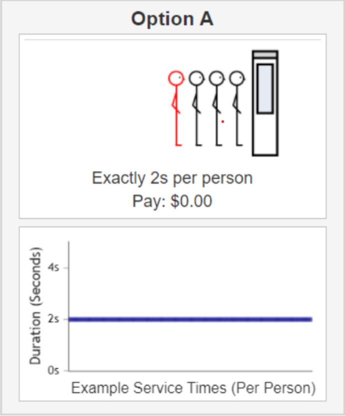
On the next screens you will see examples of two waiting scenarios. On each screen, you will see a description of a short waiting scenario in the way it would appear in Section 2. Additionally, a short animation (about 10 seconds) will illustrate what the wait would look like if it were chosen as the real wait in Section 3.

The duration of each of these example waits is *much* shorter than those you will really be considering in Section 2. However, these animations will give you a good feel for how real waits will proceed, both with and without uncertainty.

Please pay close attention to each animation. The animations will automatically repeat, so you may watch as many times as you like.

Example 1

If you saw this description in Section 2 ...

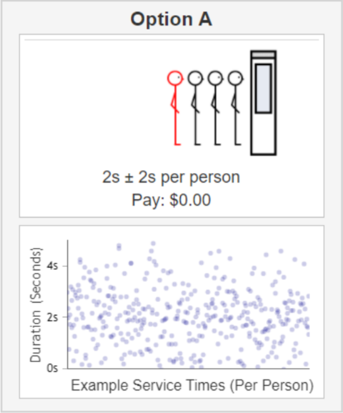


... the wait would look like this in Section 3 if it was picked.

[Animated GIF showing evolution of this wait]

Example 2

If you saw this description in Section 2 ...



... the wait would look like this in Section 3 if it was picked.

[Animated GIF showing evolution of this wait]

Comprehension quiz, Queue treatments

Quiz question 1

Which of the following is true about the real wait you will experience in Section 3?

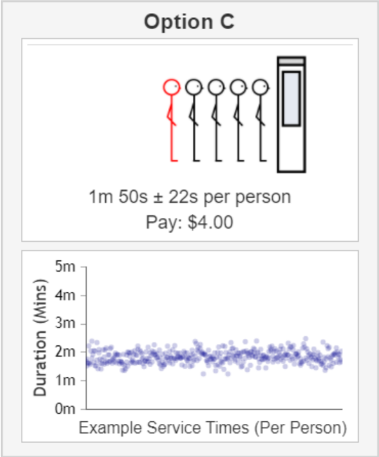
○ There will be frequent attention checks. If you miss one, you will earn no pay for Section 3.

○ There will be a *Quit wait now!* button. If you press it, you will earn no pay for Section 3.

○ You will will know exactly how many people are in front of you at all times.

● **All** of the above are **true**

Quiz question 2



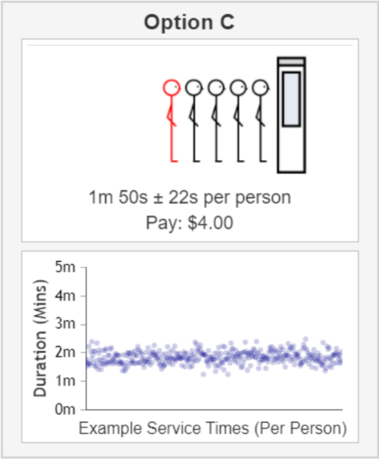
Consider the wait above. Imagine you just started this waiting scenario in Section 3. Which of the following is true about what you would know about the precise duration of your wait once you started experiencing it in Section 3?

○ You will be told the precise duration of the wait at the very beginning of Section 3.

● There is variability in the amount of time each person takes, so there is *no way* for you to calculate the precise wait duration.

○ Each person will take the *exact* same amount of time in service, so you can compute the precise wait duration.

Quiz question 3



Option C

1m 50s ± 22s per person
Pay: $4.00

Based on the waiting scenario described above, which of the following is **most likely** to be the time it will take for *one person* to complete service?

○ 1 minute
● 2 minutes
○ 3 minutes
○ 4 minutes

Quiz question 4

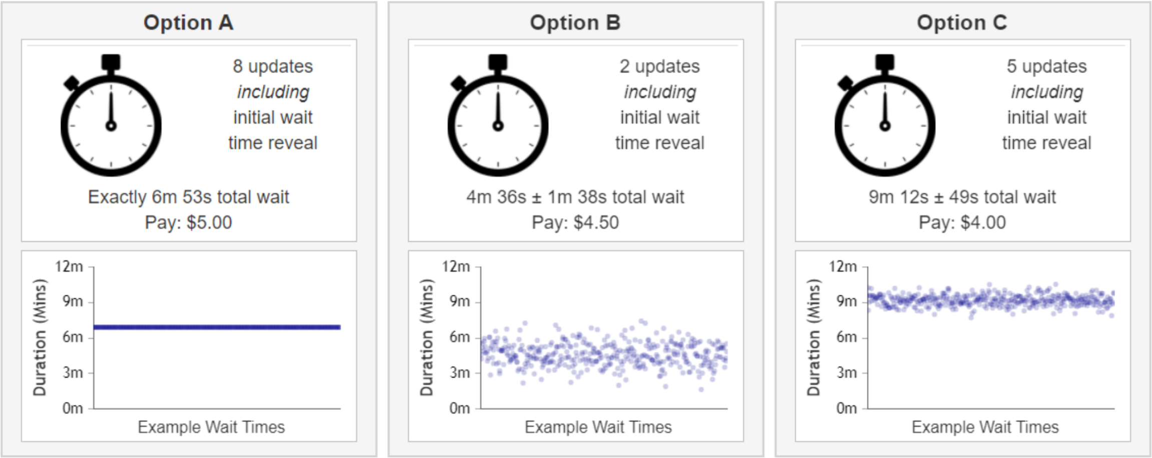
In Section 3 you will experience one waiting scenario. How is this waiting scenario picked?

○ The Section 3 real wait is the same for everyone; it is *not* influenced by my decisions in Section 2.
○ The Section 3 real wait is completely random; it is *not* influenced by my decisions in Section 2.
● The Section 3 real wait is selected from among the scenarios that I chose in Section 2.

Clock:Full+Prior Instructions

Section 2: Introduction

In section 2, you will be faced with a choice like the one below in each of 20 rounds. Your job is to decide whether you most prefer Option A, Option B or Option C.



Option A

8 updates *including* initial wait time reveal

Exactly 6m 53s total wait
Pay: $5.00

Option B

2 updates *including* initial wait time reveal

4m 36s ± 1m 38s total wait
Pay: $4.50

Option C

5 updates *including* initial wait time reveal

9m 12s ± 49s total wait
Pay: $4.00

Section 2: Why your choices matter

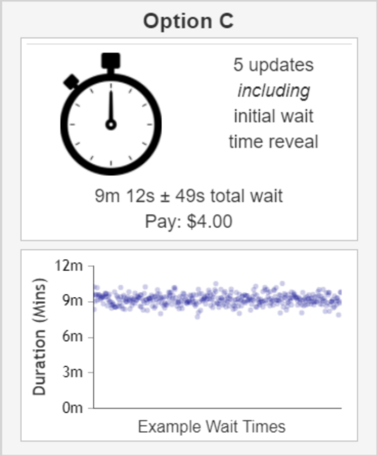
The choices you make in Section 2 will influence the waiting scenario that you will actually face in Section 3. The Section 3 wait gives you an opportunity to earn a large bonus (up to $5.00), so please consider each choice in Section 2 carefully.

After you complete Section 2, the computer will select one of the 20 rounds at random (each round is equally likely to be selected). The waiting scenario that you chose (Option A, Option B or Option C) in that randomly selected round will become the Section 3 real wait scenario.

You should treat every choice you make in Section 2 as if it will decide the Section 3 wait.

Section 2: Detailed look waiting scenarios

Each of the waiting scenarios will require you to wait for a certain amount of time. When presented in Section 2, you may know the exact waiting time or you may only know a range from which the exact time will be selected. Regardless, when you begin the actual wait in Section 3, you will always be told the precise duration of your wait. Throughout the wait, you will receive updates on the amount of time remaining. The number of such updates is given ("5 updates including the initial wait time reveal" in the scenario at right).



How long will the wait take?

Information about how long the wait will take is provided in two ways.

First, information about the wait duration appears beneath the image of the clock ("9m 12s ± 49s total wait" in the example at right). This tells you that wait takes 9m 12s on average, but that your wait will take a little longer or a little shorter than this time. The "± 49s" range means that there is a 90% chance that your total wait time will be within 49 seconds of the average. In other words, 90% of the time, your wait will take between 8m 23s and 10m 1s.

Second, information about the wait duration is given visually in the plot with the blue dots. Each dot represents a possible wait time. You can see that most dots are slightly above 9m (because the average wait time is 9m 12s), and that nearly all fall within the range of about 8m to 10m. Your total wait time will be chosen at random from these dots.

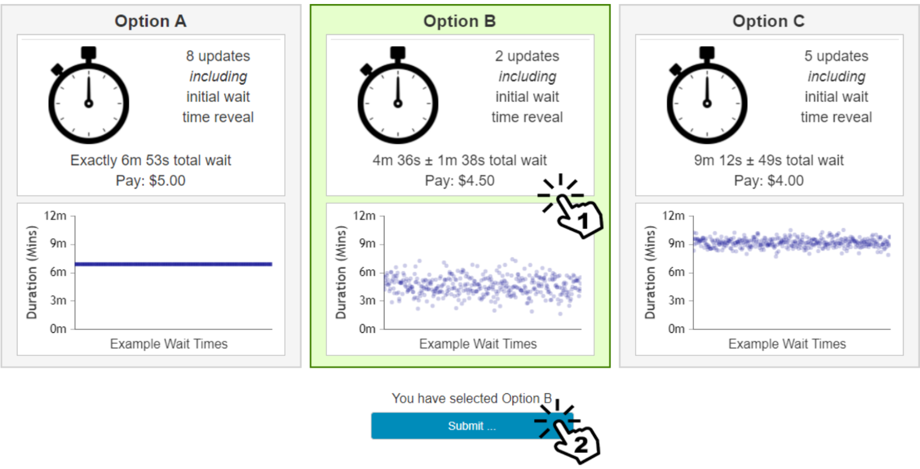
How do I successfully complete a wait?

A wait is considered a success after the actual wait time has elapsed. You will earn a bonus equal to the Pay listed ($4.00 in the scenario at right).

Section 2: Making your selection

In each of the 20 rounds in Section 2, you will be presented with three waiting scenarios. Your job is to carefully consider which wait you would most prefer to actually experience in Section 3.

To indicate the option (A, B or C) that you prefer, simply click anywhere inside of its box (see **1** below). Text will appear at the bottom of the screen indicating which option you have indicated. You may change your selection by clicking on another box. To finalize your choice, click the "Submit" button (see **2** below). You will then move on to the next round.



Section 3: Actual wait

After you complete Section 2, the computer will select one of the 20 rounds at random (each round is equally likely to be selected). The waiting scenario that you chose (Option A, B or C) in that randomly selected round will become the Section 3 real wait scenario.

You can think of this random selection process in another way. You can imagine that each time you choose a waiting scenario as your more preferred option in Section 2, it goes into a "bin" of preferred waits. The real wait in Section 3 will be chosen at random out of this "bin."

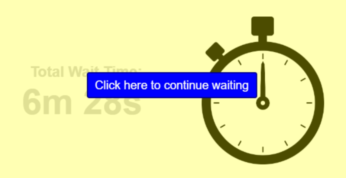
Since your choices in Section 2 directly affect the Section 3 real wait selection, you should carefully consider each decision you make.

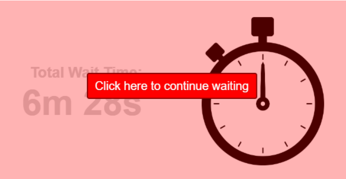
Section 3 attention checks

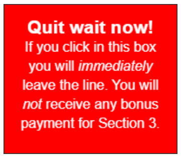
To successfully complete a wait, you must wait until you complete your service. During this time, there will be regular attention checks. Periodically, the screen will begin to flash yellow and a button will appear that says "Click here to continue waiting." You have 15 seconds to click the button. When 5 seconds remain, the screen will turn solid red.

To successfully complete the wait and earn the Section 3 pay, you must pass *all* attention checks.

You may exit the Section 3 wait at any time by clicking on the *Quit wait now!* button (see below). Clicking this button means that you will *not* earn any pay from Section 3. However, whether you successfully complete the Section 3 wait or not, you are still entitled to the HIT reward and your earnings from the quiz on Section 1.






Quit wait now!
If you click in this box you will *immediately* leave the line. You will *not* receive any bonus payment for Section 3.

Section 3: Examples of actual waits

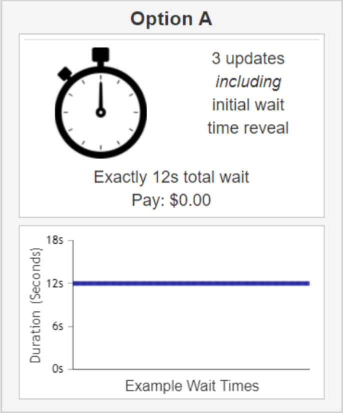
On the next screens you will see examples of two waiting scenarios. On each screen, you will see a description of a short waiting scenario in the way it would appear in Section 2. Additionally, a short animation (about 10 seconds) will illustrate what the wait would look like if it were chosen as the real wait in Section 3.

The duration of each of these example waits is *much* shorter than those you will really be considering in Section 2. However, these animations will give you a good feel for how real waits will proceed, both with and without uncertainty.

Please pay close attention to each animation. The animations will automatically repeat, so you may watch as many times as you like.

Example 1

If you saw this description in Section 2 ...



Option A

3 updates *including* initial wait time reveal

Exactly 12s total wait

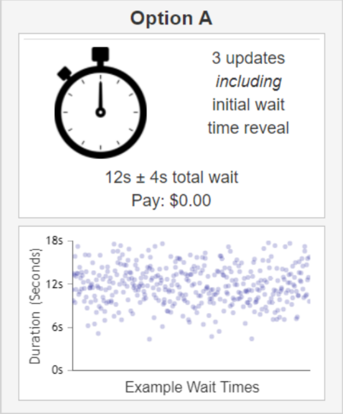
Pay: $0.00

... the wait would look like this in Section 3 if it was picked.

[Animated GIF showing evolution of this wait]

Example 2

If you saw this description in Section 2 ...



Option A

3 updates *including* initial wait time reveal

12s ± 4s total wait

Pay: $0.00

... the wait would look like this in Section 3 if it was picked.

[Animated GIF showing evolution of this wait]

Quiz questions, Clock treatment

Quiz question 1

Which of the following is true about the real wait you will experience in Section 3?

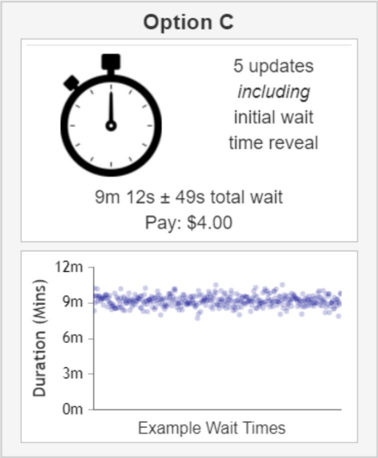
○ There will be frequent attention checks. If you miss one, you will earn no pay for Section 3.

○ There will be a *Quit wait now!* button. If you press it, you will earn no pay for Section 3.

○ You will receive periodic wait time updates that tell you *exactly* how much longer your wait will last.

● **All** of the above are **true**

Quiz question 2



Option C

5 updates *including* initial wait time reveal

9m 12s ± 49s total wait
Pay: $4.00

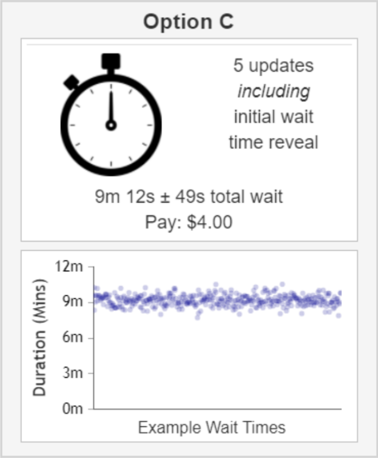
Consider the wait above. Imagine that you just started this waiting scenario in Section 3. Which of the following is true about what you would know about the precise duration of your wait once you started it in Section 3?

○ There is variability in the wait time, so you would never know the precise duration in Section 3.

● You would learn the precise duration of the wait at the very *beginning* of Section 3.

○ You would learn the precise duration of the wait, but not until the *end* of Section 3 (after waiting).

Quiz question 3



Option C

5 updates *including* initial wait time reveal

9m 12s ± 49s total wait
Pay: $4.00

Based on the waiting scenario described above, which of the following is **most likely** to be the actual wait time?

○ 3 minutes
○ 6 minutes
● 9 minutes
○ 12 minutes

Quiz question 4

In Section 3 you will experience one waiting scenario. How is this waiting scenario picked?

○ The Section 3 real wait is the same for everyone; it is *not* influenced by my decisions in Section 2.

○ The Section 3 real wait is completely random; it is *not* influenced by my decisions in Section 2.

● The Section 3 real wait is selected from among the scenarios that I chose in Section 2.

REFERENCES

Agarwal N, Ashlagi I, Rees MA, Somaini P, Waldinger D (2021) Equilibrium allocations under alternative waitlist designs: Evidence from deceased donor kidneys. *Econometrica* 89(1):37–76.

Akşin Z, Ata B, Emadi SM, Su CL (2013) Structural estimation of callers' delay sensitivity in call centers. *Management Science* 59(12):2727–2746.

Akşin Z, Ata B, Emadi SM, Su CL (2017) Impact of delay announcements in call centers: An empirical approach. *Operations Research* 65(1):242–265.

Akşin Z, Gencer B, Gunes ED (2019) How observed queue length and service times drive queue behavior in the lab, SSRN working paper number 3387077.

Alliance CC (2016) A Better Way: How Cook County is reducing its jail population through innovation and reform 2.

Allon G, Federgruen A, Pierson M (2011) How much is a reduction of your customers' wait worth? an empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing & Service Operations Management* 13(4):489–507.

Allon G, Kremer M (2018) Behavioral foundations of queueing systems. Donohue K, Katok E, Leider S, eds., *Handbook of Behavioral Operations*, 325–366 (John Wiley & Sons).

Arrow KJ (1970) Essays in the theory of risk-bearing. Technical report.

Arrow KJ (1971) The theory of risk aversion. *Essays in the theory of risk-bearing* 90–120.

Ata B, Barjesteh N, Kumar S (2019) Spatial pricing: An empirical analysis of taxi rides in new york city. Technical report, Working paper.

Ata B, Friedewald J, Randa AC (2020) Structural estimation of kidney transplant candidates' quality of life scores: Improving national kidney allocation policy under endogenous patient choice and geographical sharing .

Ata B, Glynn PW, Peng X (2017) An equilibrium analysis of a discrete-time markovian queue with endogenous abandonments. *Queueing Systems* 86(1-2):141–212.

Ata B, Peng X (2018) An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. *Operations Research* 66(1):163–183.

Avi-Itzhak B, Shinnar R (1973) Quantitative models in crime control. *Journal of Criminal Justice* 1(3):185–217.

Ayer T, Zhang C, Bonifonte A, Spaulding AC, Chhatwal J (2019) Prioritizing hepatitis c treatment in us prisons. *Operations Research* 67(3):853–873.

Bartkowiak AM (2010) Anomaly, novelty, one-class classification: A short introduction. *2010 International Conference on Computer Information Systems and Industrial Management Applications, CISIM 2010* (November 2010):1–6, URL `http://dx.doi.org/10.1109/CISIM.2010.5643699`.

Batt RJ, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* 61(1):39–59.

Becker GM, DeGroot MH, Marschak J (1964) Measuring utility by a single-response sequential method. *Behavioral Science* 9(3):226–232.

Bekker J, Davis J (2018) Learning from Positive and Unlabeled Data under the Selected At Random Assumption 1–15, URL `http://arxiv.org/abs/1808.08755`.

Bekker J, Davis J (2020) Learning from positive and unlabeled data: a survey. *Machine Learning* 109(4):719–760, ISSN 15730565, URL `http://dx.doi.org/10.1007/s10994-020-05877-5`.

Ben-Akiva M, Lerman SR (1985) *Discrete choice analysis: theory and application to travel demand.*

Berger RJ, Free MD, Searles P (2005) *Crime, justice, and society: An introduction to criminology* (Lynne Rienner Publishers New York).

Berry S, Levinsohn J, Pakes A, Pakes1 A (1995) Automobile Prices in Market Equilibrium. *Econometrica* 63(4):841–890.

Bimpikis K, Elmaghraby WJ, Moon K, Zhang W (2020) Managing market thickness in online business-to-business markets. *Management Science* 66(12):5783–5822.

BJS (2021) The Justice System — Bureau of Justice Statistics. URL `https://bjs.ojp.gov/justice-system#pros_pretrial`.

Blumstein A (2007) An or missionary's visits to the criminal justice system. *Operations Research* 55(1):14–23.

Blumstein A, Larson R (1969) Models of a total criminal justice system. *Operations Research* 17(2):199–232.

Board of Commissioners of Cook County (2021) Board of Commissioners of Cook County January 27 2021 Minutes.

Bogira S (2005) *Courtroom 302 : a year behind the scenes in an American criminal courthouse* (A. Knopf), ISBN 978-0679752066.

Borch K (1969) A note on uncertainty and indifference curves. *The Review of Economic Studies* 36(1):1–4.

Brandão F (2021) Ampl api documentation .

Brantingham PL (1977) *DYNAMIC MODELLING OF THE FELONY COURT SYSTEM.* (The Florida State University).

Bray RL, Coviello D, Ichino A, Persico N (2016) Multitasking, multiarmed bandits, and the italian judiciary. *Manufacturing & Service Operations Management* 18(4):545–558.

Bray RL, Yao Y, Duan Y, Huo J (2019) Ration gaming and the bullwhip effect. *Operations Research* 67(2):453–467.

Buchholz N (2018) Spatial equilibrium, search frictions and dynamic efficiency in the taxi industry. *The Review of Economic Studies* .

Buell RW (2020) Last place aversion in queues, SSRN working paper number 3090591.

Byrd RH, Nocedal J, Waltz RA (2006) K nitro: An integrated package for nonlinear optimization. *Large-scale nonlinear optimization*, 35–59 (Springer).

Cassidy RG (1985) Modelling a criminal justice system. *Prediction in criminology* 193–207.

Chen KD, Hausman WH (2000) Mathematical properties of the optimal product line selection problem using choice-based conjoint analysis. *Management Science* 46(2):327–332.

Claesen M, Davis J, De Smet F, De Moor B (2015) Assessing binary classifiers using only positive and unlabeled data 1–14, URL `http://arxiv.org/abs/1504.06837`.

Clear TR (2009) *Imprisoning communities: How mass incarceration makes disadvantaged neighborhoods worse* (Oxford University Press).

Conte A, Scarsini M, Sürücü O (2016) The impact of time limitation: Insights from a queueing experiment. *Judgment and Decision Making* 11(3):260–274.

Cubitt RP, Starmer C, Sugden R (1998) On the validity of the random lottery incentive system. *Experimental Economics* 1(2):115–131.

Dabbaghian V, Jula P, Borwein P, Fowler E, Giles C, Richardson N, Rutherford AR, van der Waall A (2014) High-level simulation model of a criminal justice system. *Theories and Simulations of Complex Social Systems*, 61–78 (Springer).

Davis A (2018) Biases in individual decision-making. Donohue K, Katok E, Leider S, eds., *Handbook of Behavioral Operations*, 151–198 (John Wiley & Sons).

de Palma A, Picard N (2005) Route choice decision under travel time uncertainty. *Transportation Research Part A: Policy and Practice* 39(4):295–324.

De Vries J, Roy D, De Koster R (2018) Worth the wait? how restaurant waiting time influences customer behavior and revenue. *Journal of Operations Management* 63:59–78.

Delgado CA, van Ackere A, Larsen ER (2011) A queuing system with risk-averse customers: sensitivity analysis of performance. *2011 IEEE International Conference on Industrial Engineering and Engineering Management*, 1720–1724 (IEEE).

Ding M (2007) An incentive-aligned mechanism for conjoint analysis. *Journal of Marketing Research* 44(2):214–223.

Ding M, Park YH, Bradlow ET (2009) Barter markets for conjoint analysis. *Management Science* 55(6):1003–1017.

Divito GL (2001) Chapter 5: Crimes and punishment. *Illinois Sentencing and Disposition Guide.*

Dong J, Yom-Tov E, Yom-Tov GB (2019) The impact of delay announcements on hospital network coordination and waiting times. *Management Science* 65(5):1969–1994.

Dong Y, Song S, Venkataraman S, Yao Y (2020) Mobile money and mobile technologies: A structural estimation. *Information Systems Research* 32(1):18–34.

Duong Q (2017) Skip the line: restaurant wait times on search and maps. URL `blog.google/products/maps/skip-line-restaurant-wait-times-search-and-maps`.

Elkan C, Noto K (2008) Learning Classifiers from Only Positive and Unlabeled Data. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08* URL `http://dx.doi.org/10.1145/1401890`.

Elsner A (2006) *Gates of injustice: The crisis in America's prisons* (Pearson Prentice Hall Upper Saddle River, NJ).

Emadi SM, Staats BR (2020) A structural estimation approach to study agent attrition. *Management Science* 66(9):4071–4095.

Engel C, Weinshall K (2020) Manna from heaven for judges: Judges' reaction to a quasi-random reduction in caseload. *Journal of Empirical Legal Studies* 17(4):722–751.

Epstein LG (1985) Decreasing risk aversion and mean-variance analysis. *Econometrica* 53(4):945–961.

Farrell D, Greig F (2016) Paychecks, paydays, and the online platform economy. *Proceedings. Annual Conference on Taxation*, volume 109, 1–40 (JSTOR).

Foxx K (2018) Cook County State's Attorney 2017 Data Report (February), URL `https://www.cookcountystatesattorney.org/sites/default/files/files/documents/ccsao_2017_data_report_180220.pdf`.

Freeman J (1992) Planning police staffing levels. *Journal of the Operational Research Society* 43(3):187–194.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5(2):79–141.

Gilleron R, Denis F, Gilleron R, Laurent A, Tommasi M (2003) Text classification and co-training from positive and unlabeled examples. *Proceedings of the ICML 2003 workshop: the continuum from labeled to unlabeled data* (May):80–87, URL `http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Text+Classifcation+and+Co-training+from+Positive+and+Unlabeled+Examples#0`.

Gordon BR, Hartmann WR (2016) Advertising competition in presidential elections. *Quantitative Marketing and Economics* 14(1):1–40.

Green LV, Kolesar PJ (2004) Anniversary article: Improving emergency responsiveness with management science. *Management Science* 50(8):1001–1014.

Gregory G, Satyamurty P (1965) A note on the queueing system M/M/1 with balking. *Biometrika* 52(3/4):643–645.

Guo P, Zipkin P (2007) Analysis and comparison of queues with different levels of delay information. *Management Science* 53(6):962–970.

Guo P, Zipkin P (2009) The impacts of customers' delay-risk sensitivities on a queue with balking. *Probability in the engineering and informational sciences* 23(3):409–432.

Haight FA (1957) Queueing with balking. *Biometrika* 44(3/4):360–369.

Hancock PG, Raeside R (2010) Analysing communication in a complex service process: an application of social network analysis in the scottish prison service. *Journal of the Operational Research Society* 61(2):265–274.

Hara K, Adams A, Milland K, Savage S, Callison-Burch C, Bigham JP (2018) A data-driven analysis of workers' earnings on amazon mechanical turk. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.

Harris CM, Moitra SD (1978) On the transfer of some or/ms technology to criminal justice. *Interfaces* 9(1):78–86.

Harrison GW, List JA (2004) Field experiments. *Journal of Economic literature* 42(4):1009–1055.

Hassin R, Haviv M (2003) *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59 (Springer Science & Business Media).

Hendriks A (2012) Sophie-software platform for human interaction experiments. Working paper, University of Osnabrueck.

Horowitz JL (2001) The bootstrap. *Handbook of econometrics*, volume 5, 3159–3228 (Elsevier).

Huang T, Allon G, Bassamboo A (2013) Bounded rationality in service systems. *Manufacturing & Service Operations Management* 15(2):263–279.

Huang T, Chen YJ (2015) Service systems with experience-based anecdotal reasoning customers. *Production and Operations Management* 24(5):778–790.

Hui MK, Tse DK (1996) What to tell consumers in waits of different lengths: An integrative model of service evaluation. *Journal of Marketing* 60(2):81–90.

Initiative PP, Wagner WS, Peter (2020) Mass Incarceration: The Whole Pie 2020. URL `https://www.prisonpolicy.org/reports/pie2020.html`.

Iverson B (2018) Get in line: Chapter 11 restructuring in crowded bankruptcy courts. *Management Science* 64(11):5370–5394.

Johnstone D, Lindley D, et al. (2013) Mean–variance and expected utility: The borch paradox. *Statistical Science* 28(2):223–237.

Jouini O, Akşin Z, Dallery Y (2011) Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management* 13(4):534–548.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–292.

Kato M, Teshima T, Honda J (2019) Learning from positive and unlabeled data with a selection bias. *7th International Conference on Learning Representations, ICLR 2019* 1:1–17.

Katok E (2018) Designing and conducting laboratory experiments. *Handbook of Behavioral Operations* 1–33.

Kaufmann N, Schulze T, Veit D (2011) More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk. *Amcis*, volume 11, 1–11 (Detroit, Michigan, USA).

Khan SS, Madden MG (2014) One-class classification: Taxonomy of study and review of techniques. *Knowledge Engineering Review* 29(3):345–374, ISSN 14698005, URL `http://dx.doi.org/10.1017/S026988891300043X`.

Korporaal R, Ridder A, Kloprogge P, Dekker R (2000) An analytic model for capacity planning of prisons in the netherlands. *Journal of the Operational Research Society* 51(11):1228–1237.

Kremer M, Debo L (2016) Inferring quality from wait time. *Management Science* 62(10):3023–3038.

Kuhfeld WF (2005) *Experimental design and choice modeling macros* (SAS Institute).

Kumar P, Kalwani MU, Dada M (1997) The impact of waiting time guarantees on customers' waiting experiences. *Marketing science* 16(4):295–314.

Leclerc F, Schmitt BH, Dube L (1995) Waiting time and decision making: Is time like money? *Journal of Consumer Research* 22(1):110–119.

Lee WS, Liu B (2003) Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression. *Proceedings, Twentieth International Conference on Machine Learning* 1:448–455.

Li J, Granados N, Netessine S (2014) Are consumers strategic? structural estimation from the air-travel industry. *Management Science* 60(9):2114–2137.

Li X, Liu B (2003) Learning to classify texts using positive and unlabeled data. *IJCAI International Joint Conference on Artificial Intelligence* 587–592, ISSN 10450823.

Lin T (2019) Valuing intrinsic and instrumental preferences for privacy. *Available at SSRN* 3406412.

Lipton ZC, Elkan C, Narayanaswamy B (2014) Thresholding Classifiers to Maximize F1 Score URL `http://arxiv.org/abs/1402.1892`.

List JA (2001) Do explicit warnings eliminate the hypothetical bias in elicitation procedures? evidence from field auctions for sportscards. *American economic review* 91(5):1498–1507.

Liu L (2004) A new foundation for the mean–variance analysis. *European Journal of Operational Research* 158(1):229–242.

Lu Y, Musalem A, Olivares M, Schilkrut A (2013) Measuring the effect of queues on customer purchases. *Management Science* 59(8):1743–1763.

Maister DH (1985) The psychology of waiting lines. URL `davidmaister.com/articles/the-psychology-of-waiting-lines`.

Maister DH (2005) *The Psychology of Waiting Lines.*

Maltz MD (1994) Chapter 7 operations research in studying crime and justice: Its history and accomplishments. *Operations Research and The Public Sector*, volume 6 of *Handbooks in Operations Research and Management Science*, 201–262 (Elsevier), URL `http://dx.doi.org/https://doi.org/10.1016/S0927-0507(05)80088-X`.

Maltz MD (1996) From poisson to the present: applying operations research to problems of crime and justice. *Journal of Quantitative Criminology* 12(1):3–61.

Markowitz H (1952) Portfolio selection. *The Journal of Finance* 7(1):77–91.

Markowitz H (1959) Portfolio selection.

Master N, Reiman MI, Wang C, Wein LM (2018) A continuous-class queueing model with proportional hazards-based routing. *Available at SSRN 3390476* .

Moon K, Bimpikis K, Mendelson H (2018) Randomized markdowns and online monitoring. *Management Science* 64(3):1271–1290.

Morris N, Rothman DJ (1995) *The Oxford history of the prison: The practice of punishment in Western society* (Oxford University Press).

Musalem A, Olivares M, Borle S, Che H, Conlon CT, Girotra K, Gupta S, Misra K, Mortimer JH, Vulcano G, et al. (2017) A review of choice modeling in the marketing-operations management interface. *Kelley School of Business Research Paper* (17-60):17–85.

Musalem A, Olivares M, Bradlow ET, Terwiesch C, Corsten D (2010) Structural estimation of the effect of out-of-stocks. *Management Science* 56(7):1180–1197.

Myers JJ, Lough T (2014) *Illinois's Criminal Justice System* (Carolina Academic Press).

Nagel SS, Neef M (1976) *Operations research methods: as applied to political science and the legal process* (Sage Publications Beverly Hills, CA).

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

National Center for State Courts (2019) Cook County, Illinois Criminal Courtroom Utilization Study Final Report .

Nevo A (2000) A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics and Management Strategy* 9(4), ISSN 10586407, URL `http://dx.doi.org/10.1162/105864000567954`.

Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. *Proceedings of the 22nd international conference on Machine learning*, 625–632.

Olivares M, Terwiesch C, Cassorla L (2008) Structural estimation of the newsvendor model: an application to reserving operating room time. *Management Science* 54(1):41–55.

Pazgal AI, Radas S (2008) Comparison of customer balking and reneging behavior to queueing theory predictions: An experimental study. *Computers & Operations Research* 35(8):2537–2548.

Phillips R, Şimşek AS, van Ryzin G (2021) Predicting transaction outcomes under customized pricing with discretion: A structural estimation approach. *Garrett, Predicting Transaction Outcomes Under Customized Pricing with Discretion: A Structural Estimation Approach (February 14, 2021)* .

Pratt JW (1964) Risk aversion in the small and in the large. *Econometrica* 122–136.

Pratt TC (2014) *Advancing quantitative methods in criminology and criminal justice* (Routledge).

Puterman ML (2014) *Markov decision processes: discrete stochastic dynamic programming* (John Wiley & Sons).

Rao VR (2014) *Applied conjoint analysis* (Springer).

Rapoport A, Stein WE, Parco JE, Seale DA (2004) Equilibrium play in single-server queues with endogenously determined arrival times. *J. of Economic Behavior & Organization* 55(1):67–91.

Reich RB (1973) Operations research and criminal justice. *J. Pub. L.* 22:357.

Rodríguez G (2007) Lecture notes on generalized linear models. URL https://data.princeton.edu/wws509/notes/.

Rust J (1987) Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society* 999–1033.

Sauré D, Vielma JP (2019) Ellipsoidal methods for adaptive choice-based conjoint analysis. *Operations Research* 67(2):315–338.

Savage S, Chiang CW, Saito S, Toxtli C, Bigham J (2020) Becoming the super turker: Increasing wages via a strategy from high earning workers. *Proceedings of The Web Conference*, 1241–1252.

Seabold S, Perktold J (2010) statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Seale DA, Parco JE, Stein WE, Rapoport A (2005) Joining a queue or staying out: effects of information structure and service time on arrival and staying out decisions. *Experimental Economics* 8(2):117–144.

Seepma A (2020) Just integrating or integrating justice?: Understanding integration mechanisms in criminal justice supply chains .

Shen Y, Chan C, Zheng F, Escobar GJ (2020) Structural estimation of intertemporal externalities on icu admission decisions. *Available at SSRN 3564776* .

Shunko M, Niederhoff J, Rosokha Y (2017) Humans are not machines: The behavioral impact of queueing design on service time. *Management Science* 64(1):453–473.

Singh BK (2019) Port authority launches real-time tracking for TSA and taxi wait-times. URL airport-technology.com/news/port-authority-real-time-tracking.

Staudt S (2020) Waiting for Justice: An examination of the Cook County Criminal Court backlog in the age of COVID-19 - Chicago Appleseed Center for Fair Courts. URL https://www.chicagoappleseed.org/2021/01/28/long-waits-for-justice-cook-county-criminal-court-backlog/.

Stone A (2012) Why waiting is torture. URL nytimes.com/2012/08/19/opinion/sunday/why-waiting-in-line-is-torture.html.

Swersey AJ (1994) The deployment of police, fire, and emergency medical units. *Handbooks in operations research and management science* 6:151–200.

Tom G, Lucey S (1997) A field study investigating the effect of waiting time on customer satisfaction. *The Journal of psychology* 131(6):655–660.

Tong J, Feiler D (2017) A behavioral model of forecasting: Naive statistics on mental samples. *Management Science* 63(11):3609–3627.

Ülkü S, Hydock C, Cui S (2020) Making the wait worthwhile: Experiments on the effect of queueing on consumption. *Management Science* 66(3):1149–1171.

US Sentencing Comm'n (2018) Annotated 2018 Chapter 5 — United States Sentencing Commission. URL `https://www.ussc.gov/guidelines/2018-guidelines-manual/annotated-2018-chapter-5#NaN`.

USSC (2019) Criminal History Primer. Technical report, United States Sentencing Commission, URL `https://www.ussc.gov/sites/default/files/pdf/training/primers/2019_Primer_Criminal_History.pdf`.

Usta M, Wein LM (2015) Assessing risk-based policies for pretrial release and split sentencing in los angeles county jails. *Plos one* 10(12):e0144967.

Veeraraghavan S, Debo L (2009) Joining longer queues: Information externalities in queue choice. *Manufacturing & Service Operations Management* 11(4):543–562.

Wang C, Beggs-Cassin M, Wein L (2018) A simple way to maximize the number of hits from ballistic imaging when processing capacity is limited. *AFTE J* 50(3):175–179.

Wang C, Beggs-Cassin M, Wein LM (2017) Optimizing ballistic imaging operations. *Journal of forensic sciences* 62(5):1188–1196.

Wang C, Wein LM (2018) Analyzing approaches to the backlog of untested sexual assault kits in the usa. *Journal of forensic sciences* 63(4):1110–1121.

Wang J, Zhang ZG (2018) Strategic joining in an M/M/1 queue with risk-sensitive customers. *Journal of the Operational Research Society* 69(8):1197–1214.

Wang Z, MacMillan K, Powell M, Wein LM (2020) A cost-effectiveness analysis of the number of samples to collect and test from a sexual assault. *Proceedings of the National Academy of Sciences* 117(24):13421–13427.

Ward AR (2012) Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys in Operations Research and Management Science* 17:1–14.

Weber EU, Milliman RA (1997) Perceived risk attitudes: Relating risk perception to risky choice. *Management science* 43(2):123–144.

Weisburd D (2017) *Quantitative Methods in Criminology* (Routledge).

Winnipeg Regional Health Authority (n.d.) Emergency department and urgent care wait times. `wrha.mb.ca/wait-times`, accessed 7 April, 2020.

Yang G (2019) Yelp Waitlist lets you plan life better. URL `blog.yelp.com/2019/09/yelp-waitlist-new-predictive-wait-time-and-notify-me-features`.

Yang M, Wong SC, Coid J (2010) The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychological bulletin* 136(5):740.

Yu Q, Allon G, Bassamboo A (2017) How do delay announcements shape customer behavior? an empirical study. *Management Science* 63(1):1–20.