# REDEFINING RISK: CONSTRUCTING AND VALIDATING A BAYESIAN TOOL FOR REDUCING RECIDIVISM IN THE ERA OF MASS INCARCERATION

by

PHILIP M. PENDERGAST

BS, Western Washington University, 2010

A dissertation submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Sociology

2018

This dissertation entitled:

Redefining Risk: Constructing and Validating a Bayesian Tool for Reducing Recidivism in the Era of Mass Incarceration

written by Philip M Pendergast

has been approved for the Department of Sociology

_____

Tim Wadsworth, University of Colorado Department of Sociology, Chair

_____

Jason Boardman, University of Colorado Department of Sociology, Committee Member

_____

David Pyrooz, University of Colorado Department of Sociology, Committee Member

_____

James Dykes, University of Colorado Computing and Research Services, Committee Member

_____

Zachary Hamilton, Washington State University Department of Criminal Justice and Criminology, Committee Member

Date: _____

The final copy of this dissertation has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

IRB Protocol # _____

ABSTRACT

Pendergast, Philip M. (Ph.D., Sociology)

Redefining Risk: Constructing and Validating a Bayesian Tool for Reducing Recidivism in the Era of Mass Incarceration

Dissertation directed by Associate Professor Tim Wadsworth

The actuarial assessment of recidivism risk is the foremost way that criminal justice systems balance managing caseloads, prioritizing rehabilitative services, and protecting public safety in the modern era of punishment. As such, many researchers have worked to develop new and better methods for more accurately classifying offender risk, but this has often happened at the expense of making theoretical contributions to the field of criminology. Furthermore, researchers have ignored some innovative methods that hold great promise for improving the accuracy of risk predictions. In this dissertation, I use Washington State Department of Corrections data to develop and validate the first recidivism risk assessment instrument to use Bayesian statistics, which I argue are particularly well suited to the task of prediction on a theoretical level. By comparing my results to those of a similar instrument developed using more common, frequentist regression models, I demonstrate the utility of Bayesian statistics for reducing model uncertainty and classification error. Findings also show that repeat property and drug offenders exhibit patterns of behavior that suggest a great deal of specialization in offending, whereas violent criminals display more versatile criminal behaviors, and highlight a troublesome tendency for risk assessments of all types to systematically overclassify recidivism risk among black men. Implications for life course, racial formation, and intersectional theories are discussed, along with recommendations for practice.

ACKNOWLEDGEMENTS

This project would not have been possible without the help of my committee members, the cooperation of the Washington State Institute of Criminal Justice, and the emotional support of my family and friends. Two very special thanks go to Stefanie Mollborn for her willingness to step in at the last possible moment as a committee member and to Carrie Bagli for her help in getting my committee approved, despite the odds. Thanks also to Alex Kigerl for his R code that helped me to produce some of the validation statistics reported herein. Thank you Jasmin Castillo for your willingness to speak up in my Punishment, Law, and Society course about your experiences as an intern at the Boulder County probation office—this ultimately solidified my desire to focus on risk assessment for this dissertation. The training provided by David Kaplan during the summer program on Bayesian Statistics for the Social Sciences at ICPSR heavily informed my statistical approach, and his notes from the class were a useful resource time and time again during the analysis process. To my advisor Tim Wadsworth, thank you dearly for providing a living lesson of how to best nurture a healthy work-life balance, and for always supporting my escapades during the last months of this project despite the incredibly tight timeline. Last but not least, I cannot express how grateful I am for my parents' undying support and for my wife Katie's encouragement and love when times were particularly hard and it seemed there was no end in sight. This dissertation is dedicated to the memories of Johnny Wahl and Lawrence Pendergast—I'm sorry you didn't make it to see me finish this, but this is for you guys.

CONTENTS

TABLES

FIGURES

# Chapter 1 – BACKGROUND AND OVERVIEW:

## Risk Assessment and Bayesian Analysis

The United States has, in recent years, increasingly embraced a bipartisan spirit of criminal justice reform (Arnold & Arnold, 2015). In an effort to curb some of the moral and economic costs of mass incarceration, sentencing and correctional innovations like drug diversion courts have been able to successfully divert low-level offenders in several states from harsh punishment while maintaining public safety (Gottfredson et al., 2005; 2006; Kearley, 2017; Rempel et al., 2012). For now, the momentum for reform appears to have "redirected the discussion on crime away from the question of how best to punish, to how best to achieve long-term public safety" (Subramanian, Moreno, & Broomhead, 2014, p. 2). Along with efforts to expand and strengthen community-based sanctions, support the reentry of offenders into the community, and adopt evidence-based programming for prisoner rehabilitation (Vera Institute of Justice, 2014), states have increasingly relied on sophisticated risk assessment tools to determine which inmates pose the greatest danger of recidivism (Monahan & Skeem, 2016). There is some evidence that using these instruments to help inform parole release decisions can actually reduce recidivism rates overall by releasing non-violent offenders earlier and incapacitating the most dangerous violent criminals (Berk, 2017). In terms of reforms that balance reducing the social and economic costs of mass incarceration with concerns for public safety (Skeem & Lowenkamp, 2016), risk assessment is among the easiest to implement and most appealing.

The use of assessments to forecast offender behavior is not new. The U.S. correctional system has relied on clinical and actuarial assessments to determine bail amounts, influence sentencing decisions, and help decide who receives probation or parole since at least the 1920s (Burgess, 1928; Zeng, Ustun, & Rudin, 2017). For almost as long, these instruments have been objects of scholarly attention and theoretical critique. Behind most of this attention is a group of statisticians and criminologists (Andrews & Bonta, 1995; Barnoski & Drake, 2007; Berk et al., 2009; Brennan, Dieterich, & Ehret, 2009; Duwe, 2013; Hamilton et al., 2016; Hare, 1991; Latessa et al., 2009; Liu et al., 2011; Schaffer, Kelly & Lieberman, 2011; Skeem & Louden, 2007) who have worked to incorporate new statistical methods and potential predictors in efforts to construct and validate ever more complex and accurate instruments for assessing risk or offender needs (i.e. dynamic factors that may be protective against future offending). Their focus has been largely practical, attempting to develop instruments that will be able to better distinguish between recidivists and non-recidivists in actual criminal justice settings; however, criminology has long been a theoretical discipline, and very few have commented on how well these tools capture existing theories or used the comparison of different assessment strategies as a means of testing theory. A good example of this focus on practicality at the expense of theory is in recent work that tests whether sophisticated machine learning algorithms can help to achieve better predictions than traditional regression-based methods. Even though these comparisons have produced inconclusive results (Liu et al., 2011; Hamilton et al., 2015), scholars have persisted in their interest in these methods despite the fact that they are inherently atheoretical (Kitchin, 2014) and move even further away from the discipline's theoretical origins. In fact, across this whole area of research, few (see Bonta et

al., 1998; Bonta, 2002 for exceptions) have stopped to consider how their findings inform existing theories of offending.

The other major group engaged in this research consists of a variety of critical race theorists, intersectional feminists, and social justice-oriented criminologists that identify largely theoretical issues related to the content of or outcomes related to risk assessment. These scholars have, for instance, critiqued risk assessments for ignoring key gender differences in crime precipitating factors (Belknap & Holsinger, 2006; Blanchette & Brown, 2006; Brennan et al., 2012; Smith et al., 2009) or for applying a statistical veneer to yet another practice that disproportionately impacts the disenfranchised (Alexander, 2010), stating that factors used to measure risk are essentially proxies for race (Harcourt, 2015; Skeem & Lowenkamp, 2016). Because of their critiques, 'fairness' and responsivity have emerged as newfound concerns for some hoping to develop or improve upon risk classification systems, but researchers have found it hard to balance these notions against the goal of improving predicative accuracy (Berk et al., 2017; Chouldechova, 2017). Furthermore, none of the research has taken these critiques to their logical conclusion and examined whether the inputs typically used in risk assessment instruments actually exhibit variation across racial *and* gender subgroups, or attempted to examine the predictive validity of a race and gender-specific instrument.

This dissertation sits at the confluence of these two related bodies of research. In the interests of attempting to improve the predictive validity of assessments, I propose a new statistical method that combines strengths of both traditional and machine-learning approaches to risk estimation. Bayesian logistic regression, like typical frequentist logistic regression, is an intuitive and computationally transparent method of predicting the

probability of an uncertain outcome using multivariate models. However, like machine-learning approaches, the method has the ability to learn from and adapt to empirical patterns in the data, which can lend additional predictive power to the model. Taking the development and validation of a Bayesian Assessment for Recidivism Risk (BARR) as a launching point, I then employ model output to take advantage of rich corrections and recidivism data to investigate some unresolved and under-examined theoretical questions related to offense specialization and intersectionality.

Whether criminals tend to display specialized patterns of offending versus a general proclivity for all types of crime has been a long-standing debate in the field of criminology, with strong implications for whether there can truly be 'general' theories that explain all crime (Piquero et al., 1999). I use offense-specific recidivism data and measures related to Gottfredson and Hirschi's (1990; Hirschi & Gottfredson, 1995; 2008) self-control theories and Moffitt's taxonomy of offending (1993; 2015) to interrogate whether there seems to be evidence of specialization in certain types of repeat offending patterns. Harnessing the added power of the Bayesian statistical approach, I also construct the first race *and* gender-specific risk assessment instruments, and use these to comment on intersectionality, generally, and on the possibility of risk assessment becoming a 'racial project' (Omi & Winant, 1994) that exacerbates racial disparities in punishment. Overall, the project intends to inform both practice and theory in ways that help better identify risky offenders and prioritize public safety while also improving our understanding of risk assessment's implications for racial justice in the era of mass incarceration.

**Background**

*Mass Incarceration and the Need for Risk Assessment*

The United States has the world's largest correctional population. Peaking in 2008 with just over 7.3 million people under some form of supervision by adult correctional systems (BJS, 2014), the nation's correctional population has since been in a period of significant decline. By 2015 the population had fallen to 6.7 million for the first time since 2002, and the national rate had declined to its lowest level (one in 37 adults) in over two decades (since 1994; BJS, 2015). However, despite declines in overall rates, mass incarceration has historically and continues to disproportionately impact people of color (Alexander, 2010). While these disproportionalities have somewhat narrowed since peaking in the early 1990s, substantial gender, racial, and ethnic disparities remain (Subramanian, Riley & Mai, 2018). Blacks in 2015 were incarcerated at a rate of 1,408 per 100,000, which is over 5 times the rate of whites (275 per 100,000), and 3.7 times the rate of Latinos (378 per 100,000; Nellis, 2016). Most of these national trends are driven by men, who make up the vast majority of those currently under correctional supervision, and these general statistics obscure a very different reality for women.

While men's correctional and incarceration rates have fallen in recent years, women are becoming incarcerated at an increasing rate. The vast majority of women in corrections are on probation (Carson, 2015), but the number incarcerated in the nation's jails and prisons is rising. Compared with the previous year, women's community supervision rates in 2015 went essentially unchanged while incarceration rates rose dramatically (from 120 to 160 per 100,000), continuing a long-running trend of increasing women's incarceration that began in 1980 (Bureau of Justice Statistics, 2015). Unlike for men, racial disparities in women's incarceration have closed somewhat despite the overall increase in the female

correctional population. Steadily more white and Latina women have come under correctional control since at least 2000, whereas black women have experienced a considerable drop in incarceration rates. In fact, black women (205 per 100,000) were 6 times more likely than white women (34 per 100,000) and over 3 times more likely than Latinas (60 per 100,000) to be incarcerated in 2000—highlighting racial disparities not unlike men's—but by 2014, black women's incarceration rates (109 per 100,000) were only about twice that of their white (53 per 100,000) and Latina (64 per 100,000) counterparts (Sentencing Project, 2015). Clearly, changes in offending, sentencing, and corrections have not played out the same over the past few decades for men and women, nor for blacks, whites, and Latinos. Ignoring the vast differences in criminal justice outcomes between racial/ethnic groups and across gender gives a misleading perception of who is at greatest risk of committing crimes and receiving corrections in the future. This may play out in risk assessments, as well, where criminal history is a prominent factor. It may be difficult to disaggregate the increased incarceration rates of black men and women from the actual risk presented by a given offender if criminal history is indeed a 'proxy for race', as Harcourt suggests (Harcourt 2015; but see also Skeem & Lowenkamp, 2016; Starr, 2014).

Still, one thing is and has been clear about our system of corrections for quite some time; the immense size and cost of the U.S. correctional system places a heavy burden on state and federal governments tasked with the maintenance of law and order. By any metric, this is an overburdened system. States currently spend approximately $56.9 billion per year on corrections, far out-pacing rates of funding growth in other public sectors, such as for public education, since at least 1980 (U.S. Department of Education, 2016). Numbers

vary widely by state and county, but probation officers in some particularly overburdened parts of the country like Detroit, Michigan report having caseloads of up to 177 clients per month (DeMichele, 2007), and about 73% of county public defenders, nationally, exceed maximum recommended caseload limits of 150 felonies or 400 misdemeanors per year (BJS, 2007).

At the same time, rapid advances in computing technologies and criminological research have dramatically expanded the toolkit available to police, parole agencies, courts, and justice departments for performing their jobs in increasingly efficient ways. Changes in law and procedure, such as the reduction in judicial discretion under sentencing guidelines (Freed, 1992; Nagel, 1990) and the adoption of neoliberal political attitudes toward criminals (Garland, 1990) have also motivated an increasingly 'objective' approach toward crime control and resource allocation. Simultaneously, the discourse around crime has become increasingly concerned with preserving public safety throughout the era of mass incarceration, as evidenced by the frequent use of victim-impact statements during trials, the regular introduction of laws bearing victims' names, and the development of sex-offender registries designed to protect future victims of sexual assault (Garland, 2001). As such, protection of the public from convicted criminals is of upmost importance. All of these changes have increased our desire to find objective ways to classify offenders for the sometimes-conflicting purposes of allocation of rehabilitative resources and societal risk reduction.

I argue that this constellation of structural and cultural factors has brought Data-Driven Practices (DDP) to bear on practically every aspect of crime control and correctional supervision in the age of mass incarceration. Specifically, I define DDPs as practices that

rely on the statistical analysis of vast quantities of administrative data to directly influence action by police, the courts, and parole or probation officers. When it comes to crime prevention, these practices are typified by data-driven policing strategies, such as those derived from IBM Crime Management Centers (IBM Software Group, 2010) and "hot spots" policing (Braga, 2001), on one hand, and by actuarial risk assessment and classification of existing offenders, on the other. Both approaches use extensive data to structure criminal justice process in ways that minimize costs and maximize the limited resources available for police and other practitioners to benefit public safety and maintain law and order. While policing DDPs predict where, when, and under what conditions crime is expected to occur and assign resources accordingly (Braga, 2007; Braga et al., 2014; Cotton, 2015; Mazerolle, Rombouts, & McBroom, 2007; Rosenfeld et al., 2014; U.S. Department of Justice, 2016), risk assessments help determine correctional placement, access to rehabilitative programming, or conditions of parole or probation. Thus, risk assessments explicitly aim to prioritize public safety with an emphasis on minimizing recidivism risk rather than preventing localized crime (Zeng, Ustun, & Rudin, 2017).

**Risk Assessment Overview**

In a relatively brief period of time, risk assessment has asserted itself as one of the defining features of the neoliberal style of corrections that characterizes the U.S. system in the era of mass incarceration (Garland, 2001). While nearly omnipresent in the U.S., the degree to which risk classification impacts criminal justice decisions varies by state and jurisdiction. Most commonly, people newly indicted for a crime are assessed soon after

their arrest and their resulting classification is used to inform decisions regarding bail and the conditions of probation or parole, including the level of community supervision, if found guilty (Barbaree et al., 2001). While these decisions, alone, are hugely impactful, nearly 20 states also use assessments to help determine sentencing or carceral placement (Kehl, Guo, & Kessler, 2017). In practice, this means that nearly every aspect of punishment, at least in some places, is shaped by an offender's score on these instruments (Hamilton et al., 2017).

**Figure 1.1: Risk Assessment Impact on Criminal Legal Process, John Doe in Ohio**

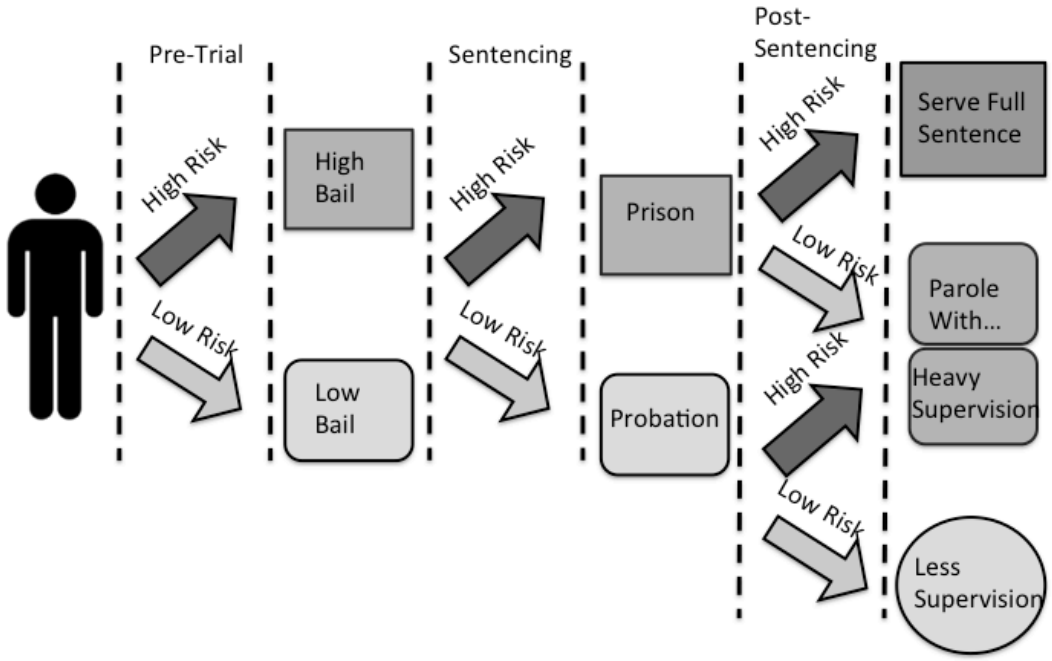



Figure 1.1 demonstrates how assessment scores might impact criminal justice decisions for a John Doe charged with his first felony property offense in Ohio. John will complete an assessment sometime after being indicted, which will impact the amount of bail he owes and thus how much time he spends in jail before having even gone to trial (Lowenkamp, VanNostrand, & Holsinger, 2013). In Ohio, John's score will again come into

play during sentencing, when a "high-risk" classification may make the difference between serving time in community corrections or a minimum to moderate security prison, as well as determining the length of his stay (Kehl, Guo, & Kessler, 2017). If incarcerated, the types of rehabilitative services he has access to, when his case comes up for parole, and the ensuing conditions of his parole will all be determined, to large extent, by his classification.

It is thus possible, depending on the instrument used, for offenders like John with no prior criminal history (in some states) to be rated high enough risk that they will be forced to spend their entire pre-trial period behind bars, will be ineligible for community corrections during sentencing, and will be unlikely to qualify for parole. While humbling on their own, the impacts of each of these consequences of being deemed high-risk are best understood as multiplicative. Otherwise-identical defendants who are detained for their entire pretrial period are much more likely to be sentenced to jail or prison and to serve longer sentences than those who were released at some point prior to trial or case disposition (Lowenkamp, VanNostrand, & Holsinger, 2013). Similarly, offenders serving longer prison terms are more likely to violate parole (Vito, Higgins, & Tewksbury, 2012) and recidivate than their counterparts who committed the same crimes but received community corrections or shorter terms, and this is especially evident among those deemed low-risk (Gendreau & Goggin, 1999). The risk of these multiplicative impacts on harsh punishment are even greater for blacks, and black men in particular, as studies have shown additional impacts of race and gender on each of these outcomes (Freiburger & Hilinski, 2010; Steffensmeier, Davis, & Ulmer, 2017; Steinmetz & Henderson, 2015). In short, risk assessments are hugely important determinants of criminal legal outcomes, and may themselves influence recidivism through their consequences.

As classification has grown to influence an increasing number of decisions in criminal legal processing, the methods used to assess risk have generally become more complex, scientific, and multidimensional. Today's tools often incorporate elements of both risk and needs (which are items used to allocate certain rehabilitative strategies and resources to the offenders most likely to benefit from them) to better meet practitioner and legislative requirements, but these developments have not always improved upon our ability to distinguish between offenders who do and do not recidivate (Baird, 2009). This next section outlines the historical development and adoption of actuarial methods for risk assessment, concluding with a discussion of how new methods like Bayesian analyses may improve our ability to correctly identify recidivists.

*From Life Insurance to Risk Avoidance: A Brief History of Actuarial Risk Assessment*

The actuarial assessment of risk is not a new concept, nor one unique to criminology. Actuarial science was developed in the mid 1800s to justify a scientific basis for the organization and management of life insurance companies, which were seen at the time as a "fraudulent enterprise" (Hickman, 2004, p. 1); so much so, that in 1843 the English writer and social critic Charles Dickens created the Anglo-Bengalee Disinterested Loan and Life Assurance Company as the perfect vehicle for the nefarious villain Jonas Chuzzlewit to swindle the masses in his novel *The Life and Adventures of Martin Chuzzlewit.* The mutual life insurance companies of the time were owned by their shareholders and operated for their benefit, such that profits were paid out to policyholders through dividends on their policies. However, the estimation of profits was not so straightforward given the long-term nature of life insurance policies. Surplus revenue had to be estimated

as a function of the contingent liabilities for future benefit payments minus the expected value of future assets. This required a method for predicting so-called "divisible surplus" from a "combination of analysis of past experience and an informed judgment about future trends" (Hickman, 2004, p. 1); a theory of calculation shared by the criminological instruments used today.

Looking back, the public benefits of the actuarial profession of the late 19th and early 20th centuries have been described as more "a slogan… [than] a reality" (Hickman, 2004, p. 5). It was not until the vast expansion in size, complexity, and economic importance of the insurance industry in the latter half of the 20th century that actuaries became necessary for the purposes of regulation and monitored compliance with all aspects of public interest in financial security systems. Much like early criminology, actuarialism as a discipline was developed out of the need for a scientific underpinning to justify a practice that could be readily interpreted as predatory despite serving a public good. While the science behind both industries helped inform practice, ranging from how much to pay out in dividends to which offenders could be released early to most effectively deter crime, it was not until the scope of these otherwise dissimilar enterprises expanded beyond their capacity for effective practice and management, otherwise, that actuarialism really became an necessary organizing tool, in and of itself, used to increase the efficiency of and regulate the insurance and crime control industries. This is the point where actuarialism became an integral part of criminal justice practice.

Long before this merger, however, scholars had proposed the idea that offender characteristics could be used to predict parole success and failure. The earliest studies by Burgess (1928) and Tibbitts (1931) provided detailed statistical schemes for classifying

parole failure risk among 3,000 Illinois youth that bear much in common with the techniques used today. The studies used a wide range of predictors including personal characteristics like employment (work status prior to arrest, work assignment while institutionalized, and jobs taken while on parole), mental health (intelligence, personality, and psychiatric prognosis), social type, area of residence, age, and nationality, as well as institutional characteristics like criminal record, sentence length, number of criminal associates, and statements of the prosecuting attorney. Factors could either contribute to an individual's risk ("black marks") or mitigate that risk as indicators of parole success ("white marks"), which were then summed, without variable weighting, to arrive at one of twelve possible numerical rankings associated with varying rates of parole failure. The authors urged the institutions from which their data were drawn to adopt these classification schemes and by 1935, the Burgess method (as modified by Tibbitts) became widely used, albeit locally, in the field.

At the time, the Burgess method was the only structured parole-risk prediction tool being used in the U.S. (Harcourt, 2008). However, it wasn't long before scholars offered other alternatives that would ultimately push the science of prediction forward and expand the ways in which the instruments were used. The much-maligned team of Sheldon and Eleanor Glueck, whose work was generally criticized throughout sociology for being too individualistic and for underestimating the importance of social influences on behavior (Laub & Sampson, 1991), almost prophetically embraced the idea that such a tool's utility need not be restricted to one point in the criminal legal system. Their underappreciated alternative (Glueck & Glueck, 1930) addressed many of the critiques of Burgess' method, offering a more parsimonious tool that included factor weighting. It also included four

separate prediction tables, meaning that practitioners could conduct assessments at the time of initial sentencing, for parole, for continued parole supervision, and for sentencing recidivists. This was an important development that would foreshadow the criminal justice system's coming reliance on actuarial assessment for many aspects of legal proceedings in the era of mass incarceration.

*Clinical versus Actuarial Assessment in the Rehabilitative and Modern Eras*

Still, the actuarial instruments of the early and mid 20th century were objects of some skepticism. Perhaps most importantly, the statistical classification of offenders was notably at odds with the largely person-focused, rehabilitative mindset that characterized correctional approaches from approximately 1950 until the social upheaval of the civil rights movement, gradual rise in violent crime, and declaration of the War on Drugs ushered in the era of mass incarceration in the mid 1970s (Alexander, 2010). Historian and criminological theorist David Garland has described this era of "penal-welfarism" as featuring "rehabilitative interventions rather than negative, retributive punishment… the use of social inquiry and psychiatric reports [and] the individualization of treatment based upon expert assessment and classification" (Garland, 2001, p. 34). Following from this psychiatric perspective, the tools designed to assess risk in this period came from the clinical assessment procedures of therapists. Clinicians, typically forensic psychologists or psychiatrists, would conduct interviews and observations using rating schemes developed by other mental health professionals to determine the attitudes, behavioral orientations, mental disabilities, personal histories, and social skills of an offender (John Howard Society, 2000). These personal and interactional characteristics were then used to make judgments

about the potential harm to society posed by an offender, which could change over time as the individual learned to adopt more positive, prosocial attitudes and as they received treatment for psychological issues.

While clinical assessments were well suited to the mindsets and aims of rehabilitative corrections, these labor intensive and typically unstructured processes were too time consuming and cost inefficient (due to the use of highly skilled clinicians) to be scalable under mass incarceration. Moreover, the very nature of the process precluded precise, reproducible predictions of risk (Grove et al., 2000); one of the defining features of assessment under the neoliberal, empirical mindset of punishment in the modern era (Garland, 2001). A bevy of research began demonstrating the superiority of more structured and "mechanical" statistical tools (Sawyer, 1966; Dawes, Faust, & Meehl, 1989; Wiggins, 1981; Grove et al., 2000), which paved the way for the wholesale adoption of statistical risk assessment. With faith in clinical assessment declining, focus returned to actuarial instruments derived from the Burgess method, coinciding with the exponential rise in the correctional population and increases in affordable computing power. The groundwork was thus laid for researchers and statisticians to innovate and improve upon the primitive instruments of the early 20th century.

*Actuarial Risk Assessment in the Modern Era*

Seizing on this opportunity, Andrews, Bonta, and Wormith developed and validated the Level of Service Inventory (LSI) at the start of the incarceral boom (Andrews, 1982), creating a sophisticated regression-based actuarial instrument that could be used to predict recidivism risk according to a static array of characteristics about an offender and

his/her offense history. In its original and revised forms, the LSI family of tools and LSI-R (Revised; Andrews & Bonta, 1995) were and may still be the most widely used instruments of their kind in North America (Jones et al., 1999). In part, the LSI's popularity stems from its ease of use (requiring a relatively brief, structured interview with a trained practitioner), utility at various points in the legal process, and relatively consistent predictive validity for a range of correctional outcomes (Andrews, Dowden, & Gendreau, 1999). But the instrument is also appealing from criminological and theoretical perspectives: many of the 54 items included in the LSI-R are aligned with theoretically relevant factors, such as attitudes/orientations and emotional/interpersonal problems, or are derived from social learning, differential association, and life course theories (e.g. education and employment, finances, family and marital relationships, residential accommodations, leisure and recreational activities, substance use and dependence, and social networks and criminal associations; Andrews & Bonta, 1995). By bridging the gap between clinical and mechanical risk assessment and incorporating factors underlying prominent criminological theories, the instrument has wide appeal and can be used to both predict recidivism and assess whether offender needs are improved by the use of interventions and programming.

Since the introduction of the LSI, a group of risk assessment experts have emerged to help meet the demand for new instruments with greater applicability to specific states, types of offenders, or points in the criminal legal process (Andrews & Bonta, 1995; Barnoski & Drake, 2007; Berk et al., 2009; Brennan, Dieterich, & Ehret, 2009; Duwe, 2013; Hamilton et al., 2016; Hare, 1991; Latessa et al., 2009; Liu et al., 2011; Schaffer, Kelly, & Lieberman, 2011; Skeem & Louden, 2007). Most of these researchers have focused on

developing and validating new instruments, examining the predictive validity of existing instruments among new offender populations, investigating how methodological variation across instruments impacts predictive power, or advocating for the importance of risk *and* needs assessment. In essence, this body of research is practical in focus and intends to produce better performing tools for practitioners to adopt and use in the criminal justice system.

*Improving Assessment*

There are many important decisions that researchers must consider when attempting to improve the predictive validity of risk assessment instruments. Developers must first consider how the instrument is to be used when deciding whether to assess both static (i.e. unchanging, like gender and criminal history) risks and dynamic (i.e. employment, education, attitudes, drug use) needs. While needs assessment is important for some criminal justice agencies that would like to prioritize interventions and other services to offenders demonstrating the highest needs (Hamilton & Wormer, 2015), this type of assessment cannot be automated and may actually reduce the instrument's ability to accurately classify recidivism risk (Barnoski & Drake, 2007). A second issue is how items are selected. At a minimum, characteristics should be predictive of recidivism in bivariate analyses; otherwise they will not add any predictive value (Baird, 2009). In a statistical sense, items that remain significant predictors in multivariate models should provide the best predictions because they help explain more variance in recidivism, while other measures that are no longer associated in multivariate models simply add prediction noise (Hamilton et al., 2016). Once items have been selected, researchers must then decide how

best to weight them in terms of their contribution to recidivism risk. The earliest methods (and many newer ones, such as the LSI-R) employed Burgess weighting schemes, which equally weighted all risk factors and allowed practitioners to easily sum the items to arrive at a risk score. However, some items (e.g. age, criminal history) are much more strongly predictive of recidivism than others (e.g. education, employment) and analytic weighting schemes derived from regression models can result in rankings that yield better performance (Barnoski & Aos, 2003). Ideally, instruments also measure several different types of recidivism risk (violent, sex, property, drug), allowing these analytic weights to vary across each offense type. This tends to best reflect the underlying processes that lead different types of individuals to reoffend, and may result in more accurate classification decisions (Hamilton & Wormer, 2015).

Given that the most accurate predictions tend to come from instruments using multivariate item selection and analytic weighting schemes, the majority of modern assessments use frequentist logistic regression models to get a sense of how strongly various risk factors are related to recidivism outcomes. This method strikes a good balance between achieving prediction accuracy and maintaining the basic interpretability of findings, which may be important for legal and ethical reasons related to upholding defendants' rights to due process (Loomis v. Wisconsin, 2017). However, scholars interested in further improving on the predictive validity of assessment tools have recently questioned whether logistic regression is really the optimal method for prediction purposes. They suggest that it may be better to use methods developed exclusively for prediction, which sacrifice basic interpretability, to better accomplish this goal (Berk, 2017; Berk & Bleich , 2013; Liu et al., 2011). As such, an increasing number of criminologists have

become interested in using machine learning algorithms to determine offender recidivism risk (e.g. Berk et al., 2009; Caulkins et al., 1996; Liu et al., 2011; Neuilly et al., 2011).

Rather than making assumptions about how offender characteristics may be related to recidivism, either through the application of theory, bivariate associations, or multivariate item selection and modeling, machine learning techniques simply use whatever data is available to find an algorithm that best predicts recidivism (Breiman, 2001). Because there is no concern for deriving information about how individual predictors are related to the outcome, and there is no reason to 'vet' variables before entering them into the predictive model, machine learning algorithms can make use of any and all data to reach classification decisions. This allows even variables that share variance with each other to inform decision making through both independent and interactive mechanisms, unlike in regression models, where these effects are diluted (Steadman et al., 2000). Furthermore, through the use of partitioning data into training and validation sets, machine learning algorithms have the ability to 'learn' from patterns in the data, updating classification rules before ultimately being tested for predictive accuracy in a third 'test' set, which generally improves the validity of classifications when predictions are then applied to new data. Classification decisions are also free to vary across subgroups of individuals that exhibit different relationships between predictors and recidivism in the data, which may help better reflect the reality of risks in offender subpopulations (Steadman et al., 2000), such as unique pathways to crime for men and women, and for whites and people of color (Belknap & Holsinger, 2006; Bell, 2013; Potter, 2015). As such,

machine learning approaches derived from neural networks[1] and decision trees/random forests[2] have several considerable statistical advantages over conventional logistic regression models for the purposes of prediction (Breiman, 2001).

Regardless of the statistical theory that suggests machine learning should be ideally suited to prediction, studies directly comparing the predictive validity of machine learning and logistic regression models have been inconclusive. Some have observed small advantages of machine learning approaches (Berk, 2017; Liu et al., 2011; Zeng, Ustun, & Rudin, 2017), while others (Hamilton et al., 2015) have demonstrated that such techniques, at best, perform comparably with traditional logistic regression models. With such mixed evidence about whether machine learning can truly improve the validity of recidivism predictions, the fact that logistic regression provides a good compromise between predictive potential, interpretability and transparency continues to make the method attractive. So far, regression-based assessments have held up as constitutional in the U.S. Supreme Court because the statistical parameters that result in an inmate's classification can be explained to them and these instruments do not solely assign risk to constitutionally impermissible aggravating factors such as race or gender (Loomis v. Wisconsin, 2017). Instruments derived from machine learning techniques may abdicate too much responsibility to hidden decision-making processes (Tashea, 2017) or draw on too much

[1] Neural Networks are adaptive systems of logistic regression models using connectionist approaches to computation. Inputs are randomly tied to a hidden layer of artificial neurons that receive different weights as various algorithms are applied in a learning phase, until models return an optimal function that minimizes model deviance (Breiman, 2001; Hamilton et al., 2015)

[2] Decision trees are question-decision models that split data into statistically homogenous groups defined by the best fitting predictor variables. As more predictors are applied, groups become increasingly homogenous until some pre-defined stopping point is reached (Harper, 2005). Random forests provide a means of summarizing classification decisions made over multiple trees (Berk et al., 2009).

data that is not directly relevant to criminogenic risk (e.g. race) to hold up to constitutional challenge. These reasons provide convincing practical rationale for further investigating the utility of logistic regression for the purpose of risk assessment.

Overall, my review of the risk assessment literature suggests that instruments that employ multivariate item selection strategies and analytic weights taken from logistic regression models achieve the best compromise between predictive accuracy and interpretability. However, there are some distinct advantages of machine learning techniques. The ability for instruments to adapt to empirical patterns displayed in the data over separate samples may help to better capture underlying relationships between criminogenic risks and recidivism while also reducing prediction error in out of sample classifications, and allowing risks to vary across subgroups (such as gender and race) in the data aligns well with pathways and intersectional theories of crime. I argue that Bayesian logistic regression, which has not been previously used for assessing recidivism risk, capitalizes on the strengths of both approaches while generating models that can be used to inform theory.

**A Bayesian Approach to Risk Assessment**

Bayesian analysis is based out of a different statistical paradigm than the frequentist assumptions that underlie both traditional, frequentist regression and machine learning. Both traditional approaches to risk assessment assume that there are true population values of recidivism risk that can be estimated using long-run frequency to approximate the probability of an offender with a certain set of characteristics being arrested or

reconvicted of a crime. The definition of probability in Bayesian statistics is different. Rather than a knowable quantity that we can estimate precisely through repeated sampling, probability to Bayesians is inherently uncertain and simply a reflection of our beliefs about some outcome, conditioned on what some set of observed data tells us about the outcome (Kaplan, 2014). The best we can hope for is to make an educated guess about whether something will occur, see what happens, and then make a more precise estimate next time. We can never know for sure what the outcome will be, but with enough information we can minimize our uncertainty and obtain an accurate assessment of the probability for prediction purposes.

On a purely philosophical level, this is a better reflection of how we intuitively understand probability. After all, the very concept of probability implies uncertainty—there is a chance that something will happen, but also a chance that it will not. The Bayesian concept of estimating probability also aligns well with the reality of how we conduct risk assessment, where we identify characteristics that are theoretically linked to recidivism, see how these measures actually predicted recidivism in some set of data, and then hope that these estimates of probability are correct when we apply them to some new data where the probability of recidivism is unknown. Because human behaviors like reoffending are inherently uncertain, but conditioned on prior observed behaviors (Gottfredson & Moriarty, 2006), Bayesian statistics are more theoretically aligned with the purpose of risk assessment than conventional methods.

The process of conducting Bayesian regression analysis is mostly very similar to that of frequentist regression, thus retaining the strengths of that approach for the purpose of predicting recidivism risk. The main difference is that Bayesian methods use two types

of data to arrive at estimates. Rather than simply assessing statistical relationships within some set of observed data, Bayesian analyses condition estimates on some 'prior' knowledge of how we assume these relationships should look. The amount that this prior information influences estimates and where these priors come from are perhaps the most important and unique considerations that Bayesians face when analyzing data. While some frequentists may see the inclusion of priors as a weakness that makes Bayesian analysis too subjective (Gelman, 2008), it is in fact a strength of the approach that can improve statistical power and reduce the uncertainty in model estimates when well-aligned with empirical patterns in the data (Loh et al., 2015; Stegle et al., 2010). Furthermore, priors provide a mechanism for updating estimates over different samples, much like machine learning algorithms do. The main difference from how this plays out in machine learning is that the researcher maintains compete control over the process. This ability to update estimates over new inmate samples is a major advantage for researchers hoping to norm instruments for new offender populations (Hamilton & Wormer, 2015) or looking to develop state or jurisdiction-specific assessments.

The resulting parameters of Bayesian models are given in terms of posterior probability distributions, which are direct estimates of the anticipated probability of an outcome and the uncertainty associated with the estimate. Posterior distributions can be summarized using credibility intervals that detail a range of values within which the probability of an outcome is expected to occur, with the central tendency reflecting the value with the highest probability (or least uncertainty) of being correct. Compared to the frequentist coefficient and confidence interval, these values are much more informative about how a given characteristic is expected to impact the probability of an outcome.

Because they correspond to the frequentist assumption of repeated sampling, confidence intervals do not give meaningful information about the most likely value and instead demonstrate the range of probable parameter estimates from all possible samples (Spanos, 2012). While this is useful for performing null hypothesis tests, it is not necessarily a useful way of determining the ideal value for parameters in predictive models. The Bayesian posterior distribution, on the other hand, is ideally suited to this purpose and not intended to be used for null hypothesis testing.

It follows that the other major, and perhaps even more disconcerting, departure from traditional frequentist statistics is that Bayesian analyses do not rely on p-values, significance, or any other rigidly defined conventions to determine whether something is substantively important (Gelman et al., 2014). As such, in this dissertation I do not report or discuss anything in terms of being statistically significant. I do, however, note when estimates seem to indicate that some variability is not likely to be due to chance. To arrive at these conclusions, Bayesians typically consult the posterior probability distribution to determine whether estimates lie outside of the range wherein 95% of the probability is expected to occur (Gelman et al., 2014). Ideally, strong priors that are well aligned with the data produce very tight posterior distributions, which make these determinations easy, but sometimes comparison estimates may be just on the verge of the Bayesian credibility interval. In these cases, because the posterior is an explicit probability distribution with very low probabilities associated with the tails, I tend to conclude that differences are still meaningful. Overall, since null hypothesis tests are mostly just used to determine which items to include in multivariate frequentist regression models, this final distinction has

relatively little impact on understanding differences between Bayesian and frequentist models for risk assessment.

Despite success producing risk assessments in the fields of medicine and genomics (Fenton & Neil, 2012; Newcombe et al., 2012; Ogino & Wilson, 2004), and some scholars recently advocating for greater use of Bayesian statistics in the area of criminal justice (Fenton, Neil, & Berger, 2016; Philipse, 2015), very few criminologists have used Bayesian methods in their research (see Berk et al., 1992 for an exception). Even more surprising, given its strong theoretical alignment with the aims of risk assessment and its ability to capitalize on strengths of both regression and machine learning-based instruments, none have developed or examined the predictive validity of Bayesian instruments for recidivism risk classification. While I work to demonstrate the utility of a general Bayesian Assessment for Recidivism Risk (BARR) in this dissertation, I also recognize the ability of regression-based approaches to obtain interpretable results that can be used to help inform theory, and the potential usefulness of machine learning techniques to allow predictors to vary across different subgroups of offenders. The Bayesian approach, in combining advantages of both estimation strategies, affords me the ability to investigate these additional topics of inquiry.

**Toward More Theoretically Informative Risk Assessment**

Risk assessment studies have been primarily driven by practical concerns with improving the predictive validity of instruments. This is certainly important and has driven the field of risk assessment forward, but the constant focus on improving classification

accuracy obscures another potential contribution that these studies may be well poised to make: assessments provide information about how criminogenic risks are associated with recidivism, and these results can have some bearing on larger criminological debates regarding theory. There are many research questions beyond "Can this strategy improve predictive accuracy?" that risk assessment models can answer, with studies typically relying on large datasets of offenders and a host of variables regarding offense histories, demographic factors, and occasionally more dynamic measures related to employment, education, relationships, and views towards figures of authority and crime. Some of these data, such as that used to validate the LSI-R (Andrews & Bonta, 1995), are actually very well suited to approximating major theories of crime, allowing for strong tests of these theories' ability to predict future offending. Despite the richness of data and the presence of measures related to theory, James Bonta is the only prominent risk assessment scholar who has explicitly used data from their validation studies to inform theoretical development. By comparing the LSI-R, which was derived from social learning theory, to other scales representing sociological and psychopathological theories of offending, Bonta and colleagues (Bonta et al., 1998; Bonta, 2002) suggested that the superior predictive validity of the LSI-R was confirmation that social learning was the best-supported theoretical explanation for crime.

While other instruments are typically not based so closely on specific theories of crime, measures included in them, and their associations with recidivism, may be similarly theoretically informative when comparing the validity of different assessment models. For example, examinations of the LSI-R's predictive validity against simpler instruments composed of items derived exclusively from life course theories (age, gender, and criminal

history) may not only demonstrate the superiority of the simpler models for predictive purposes (Barnoski & Drake, 2007) but may also suggest that life course theories are better at explaining patterns of reoffending than social learning theory. Life-course theories provide natural starting points for recidivism classification decisions because they explicitly concern factors that contribute to desistance from criminal lifestyles. Desistance implies that characteristics other than those that contribute to crime, generally, are at play—some offenders commit an offense and then do not recidivate, while others go on offending. By speculating about which factors increase the likelihood of continued, repeat offending, these theories prove useful for predicting recidivism. It also helps that the measures implied in many of these theories such as age, gender, and criminal history are easily measured reliably using archival data, which minimizes alternate sources of error (such as measurement error and low interrater reliability) in the estimation process.

Importantly, while age, gender, and criminal history are good candidates for risk estimation, they are not characteristics unique to one specific theory. In fact, all three characteristics play important roles in many prominent theories of crime, including social learning theory. However, unlike models representing social learning theory, which feature a multitude of measures that increase statistical noise in predictions (Baird, 2009), models representative of life course approaches can be relatively simple. Most notably, key elements of Moffitt's taxonomy of offending (1993; 2015), Gottfredson and Hirschi's general/self-control theories of crime (1990; Hirschi & Gottfredson, 2000), Sampson and Laub's age-graded explanations (Laub & Sampson, 1993, 2006; Sampson & Laub, 1990, 1992, 2005), state-dependence theories (Nagin & Paternoster, 1991; 2000), and various intersectional and pathway theories (Belknap & Holsinger, 2006; Blanchette & Brown,

2006; Brennan et al., 2012; Smith et al., 2009) could be reasonably approximated using simple demographic characteristics and detailed information about criminal histories. This is because, despite having positions of prominence in each of these perspectives, age, gender, and criminal history are expected to differentially impact the maintenance of criminal behavior across offense types in ways related to the central mechanisms implied by each theory, such as through offense specialization and variation in criminogenic risks by race and gender. This expected variation allows me to construct models with relatively few covariates that can still be used to weigh in on important debates in the field arising from these perspectives, while also using offense-specific models ideally suited to studying the impacts of methodological variation on predictive validity (Hamilton et al., 2015).

Specifically, Moffitt's and Gottfredson and Hirschi's theories pertaining to the existence of a separate class of serious, repeat offenders provide the basis for testable hypotheses related to the debate over criminal versatility or specialization. Both speculate that younger, male offenders with longer and more intense criminal histories marked by a greater number of juvenile and adult offenses will tend to exhibit greater variation in their offense patterns, including engagement in violent offending. For Gottfredson and Hirschi, this is because offenders with these characteristics lack self-control and become susceptible to all kinds of offending (1990; Hirschi & Gottfredson, 2000). Moffitt, on the other hand, suggests that offenders with these characteristics can be classified as life-course persistent, meaning that they are more likely to recidivate and engage in many different types of serious crime than others (1993; 2015). Those who do not lack self-control and who are not life-course persistent are expected to engage in much more selective offending without advancing to the most serious, violent types of crime (Piquero

et al., 1999). This presents a problem for most general theories of crime, however, because it implies that the explanations underlying each type of offense are unique, rather than supposing that there is a general propensity for crime that determines all criminal behavior. Because Bayesian analyses can be updated upon the introduction of new data, the unique patterns of offending that predict violent, property, and drug recidivism are given more time to establish themselves. I thus expect that item weights given to criminal history measures will vary across offense-specific Bayesian models in ways that allow me to weigh in on whether offenders tend to specialize.

Intersectional theories, on the other hand, suggest that the lived experiences of people of color, women, and women of color are all unique, and that we cannot understand gender without understanding race, and vice versa. The ways that race and gender influence offending behaviors also vary across the life course; as such, risk factors are expected to vary by race, gender, and age (Potter, 2015). Some have suggested that risk assessment instruments are biased against blacks (e.g. Angwin et al, 2016), especially when instruments are composed primarily of factors related to criminal history (Harcourt, 2015; Skeem & Lowenkamp, 2016), and others demonstrate convincingly that men and women reflect different offender populations (Brennan et al., 2012) with very different risks and needs (Salisbury, Van Voorhis, & Spiropoulos, 2009); as such, existing instruments that are normed with majority male samples are poorly conceptualized to predict recidivism among women (Duwe & Kim, 2016). While some have incorporated gender-specific weighting procedures (Duwe & Kim, 2016; Hamilton et al., 2016) or developed entirely separate instruments for women (Salisbury, Van Voorhis, & Spiropoulos, 2009) none have taken these critiques to their logical conclusion and explored whether better and fairer

classifications can result from risk assessment instruments that vary substantially by both race and gender. Further, none have tested instruments' predictive validity among both racial *and* gender subgroups (such as among white men and women and black men and women). Part of this may be due to the large amount of data necessary to obtain sufficiently large samples of white and black male and female inmates. Because Bayesian methods have greater predictive power when priors are well aligned with data (Loh et al., 2015; Stegle et al., 2010) the method yields a sufficiently large sample for me to consider these hypotheses for the first time.

In conclusion, Bayesian statistics provide a promising solution for aligning risk assessment approaches with the best knowledge of how to improve the validity and reliability of classification, as well as enabling researchers to investigate questions of theoretical interest. Rather than examining large cross-sections of recidivists to obtain item weights based on logistic regression or abdicating some decision making processes to machine learning 'black boxes' (Hamilton et al., 2015), Bayesian statistics model the probabilities of recidivating directly, given a set of factors, update these estimates based on which factors actually influenced reoffending in prior cohorts, and examine how well they return the actual, observed outcome. Because the resulting estimates can then be updated with future cohorts' data, the potential predictive power of these models for classification is greater than with analogous frequentist models. As I will show, this not only improves on existing estimation strategies, but it also gives increased power over frequentist models to conduct analyses on gender and race/ethnic subgroups and investigate theories related to crime specialization in offense-specific models. Ultimately, Bayesian methods help to bridge the gap between regression-based and machine learning techniques while providing

probabilistic assessments of risk that better align with the probabilistic nature of human behavior.

**Aims**

This dissertation advocates for the use of Bayesian analyses in recidivism risk assessment. Chapter 2 outlines the data and method in greater detail, highlighting specific reasons why probabilistic methods improve upon conventional regression-based approaches to risk assessment. Chapter 3 outlines the construction and validation of the first Bayesian Assessment for Recidivism Risk (BARR), demonstrating the validity and power of the method for reducing uncertainty in model estimates and highlighting theoretically relevant risk factors related to crime specialization and versatility. Chapter 4 then applies this same technique and adopts an exploratory lens to examine the possibility of using intersectional race and gender-specific predictors to reduce misclassification, investigating the predictive potential of the BARR among racial and gender subgroups. I conclude in Chapter 5 by highlighting the theoretical, statistical, and practical advantages of Bayesian statistics for assessing recidivism risk, as well as discussing the implications of various theoretical findings for criminology and our understanding of race/ethnic relations, more generally.

Organizationally, the more empirically substantive Chapters 3 and 4 are written as separate journal articles. This helps to give additional space for expanding on the relevant literature and specific theoretical contributions of those two chapters, which are very different from each other but share much of the background outlined in this introduction

and in the data and methods discussed in Chapter 2. I also include further discussion of methods unique to each that help to better address the unique theoretical and practical questions motivating each chapter. While this organizational approach results in some necessary redundancies, I believe that it helps better differentiate the diverse aims of this project; affording each aim the space it needs to clearly link data with hypotheses, and findings with broader contributions to the fields of criminology, risk assessment, and beyond.

# Chapter 2 – DATA AND METHODS

## Setting

I use data from Washington State's Department of Corrections (WADOC) over the period 1986 to 2008 for this dissertation. Washington is the 13th most populous state in the U.S. and represents a relatively unique setting both politically and in terms of its laws and approaches to punishment. The state is widely considered to be among the most liberal in the U.S. and has consistently voted for Democratic senators, governors, and presidents in each election since 1994, matching Delaware for the longest streak on record (Cohn, 2017). Throughout the period for which I have data, Washington has consistently elected Democrats in presidential and gubernatorial races, going longer without electing a Republican governor than any other state in the country (Leip, 2016). However, the Cascade mountain range forms a natural division between the state's urban coastal and rural agricultural regions that represent very different political orientations. While the western half of the state is home to most of its population centers, which are characterized by more liberal voting behaviors, the eastern half of the state and rural parts of the west have long been Republican strongholds. This conflict has at times resulted in notable differences between laws at the local and state levels, which could have impacts for policing, legal process, and recidivism. Unfortunately, my data do not contain geoidentifiers that would enable the examination of geographic heterogeneity in recidivism risk factors. On the whole, then, the state is best understood as relatively liberal and progressive in terms of crime control policy.

Perhaps the clearest example of Washington State's progressive orientation toward lawmaking is in its history of drug reform. Long before Washington joined Colorado as the first two states to legalize the recreational sale of Marijuana, policies indicated a liberal attitude toward low-level drug crimes. Following California's lead, the state legalized medical marijuana in 1998, becoming one of the first to do so. Soon after, Seattle became the first major U.S. city to decriminalize routine marijuana use (Stevenson, 2003). King County, which includes Seattle, established the twelfth drug court in the country in 1994, and continues to be a leader in adult diversion programs. Notably, statewide changes to Washinton's Drug Offender Sentencing Alternative (DOSA) law in 2005 diverted many low-risk drug offenders from prisons into chemical dependency treatment (Aos, Phipps, & Barnoski, 2005).

On the other hand, certain aspects of punishment in Washington fall behind the norm of the most progressive states. Capital punishment remains legal, and Washington is the only state that maintains a gallows and allows inmates to die by hanging, though Governor Jay Inslee declared a moratorium against executions in 2014 (Ganga, 2014) and proposed legislation to abolish the death penalty in 2017 (O'Sullivan, 2017). This latest effort to replace the death penalty with life in prison, however, failed to pass the state's legislature in early 2018 (Wasserman, 2018).

Between 1980 and 2013, WADOC underwent significant changes in its correctional population. In 1980, WADOC had the highest rate of disproportionate minority incarceration in the nation (Christianson, 1980). While disparities fell thereafter, even in 2010 black (2,372 per 100,000) and Latino incarceration rates (601 per 100,000) were much higher than the national average, and some orders of magnitude larger than those of

whites (392 per 100,000). Overall, the state has one of the lowest incarceration rates in the country (475 per 100,000), along with relatively low levels of violent crime (284 per 100,000) and high levels of property crime (3,464 per 100,000) in its cities. Washington also has the lowest ratio of policemen to residents of any state, at about 174 per 100,000 residents (Reaves, 2011).

Washington has long been a leader in making evidence-based crime policy recommendations. The state's legislature created the Washington State Institute for Public Policy (WSIPP) in 1983 to carry out non-partisan research that informs criminal justice, education, child welfare, mental health, substance abuse, and health care policy. In 1999, Washington State passed the Offender Accountability Act (OAA) to fundamentally change the way that the WADOC supervised convicted felons after their release. Aiming to "reduce the risk of reoffending by offenders in the community," (RCW 9.94A.010) lawmakers began requiring WADOC to classify and supervise felons differentially according to a risk classification system. In turn, WADOC adopted the widely used Level of Service Inventory-Revised (LSI-R), a 54-question survey consisting of both "static" and "dynamic" risk factors, as the state's predictive instrument. A subsequent 2003 report (Barnoski & Aos, 2003) on the validity of the LSI-R by WSIPP identified several areas where the tool could be strengthened to improve the accuracy of classification, increase the objectivity of the instrument, and reduce the amount of time spent completing the assessment.

These concerns with the LSI-R compelled the Institute to construct and validate their own instrument in 2006 for assessing recidivism risk using administrative data from all offenders released from prison/jail or placed on community supervision between January 1986 and September 2002 (Barnoski & Drake, 2007). The resulting instrument, the

Static Risk Assessment (SRA), improved substantially on the predictive validity of the LSI-R for Washington's offender population while streamlining the data collection process and condensing the number of risk factors to only the 23 static factors most strongly associated (theoretically and empirically) with recidivism: age, gender, and detained criminal history. During the rollout of this instrument, a risk-only assessment, WADOC also began development of a dynamic needs assessment—the Offender Needs Assessment (ONA). The resulting instrument was implemented in 2008 to assist in case management for program referral after recidivism risk had been established with the SRA.

During this time (2008), significant changes occurred with regards to Washington State's correctional policy that removed many low and moderate risk offenders from community supervision and increased the chances of technical violations. This prompted WADOC to update the SRA (resulting in the SRA-2; Barnoski, 2010). Soon after, as part of an effort to expand its use of evidence-based programming, WADOC sought to construct and validate a system of assessments that could address both criminal risks and needs. The resulting instruments, the STRONG and revised STRONG-R (Hamilton et al., 2014) incorporate both static and dynamic measures and demonstrate very good predictive validity (Hamilton et al., 2016).

Despite the advances made with the STRONG-R, there is far less WADOC data available in the combined static criminal risk and offender needs data than there is the in the sample initially used to construct and validate the SRA. Moreover, trajectories of risk behavior may have changed substantially in the wake of these policy changes, but there is a long timeline of corrections and recidivism data available during a period where policy was relatively constant (January 1986- January 2008) in the static risk data. When trying to

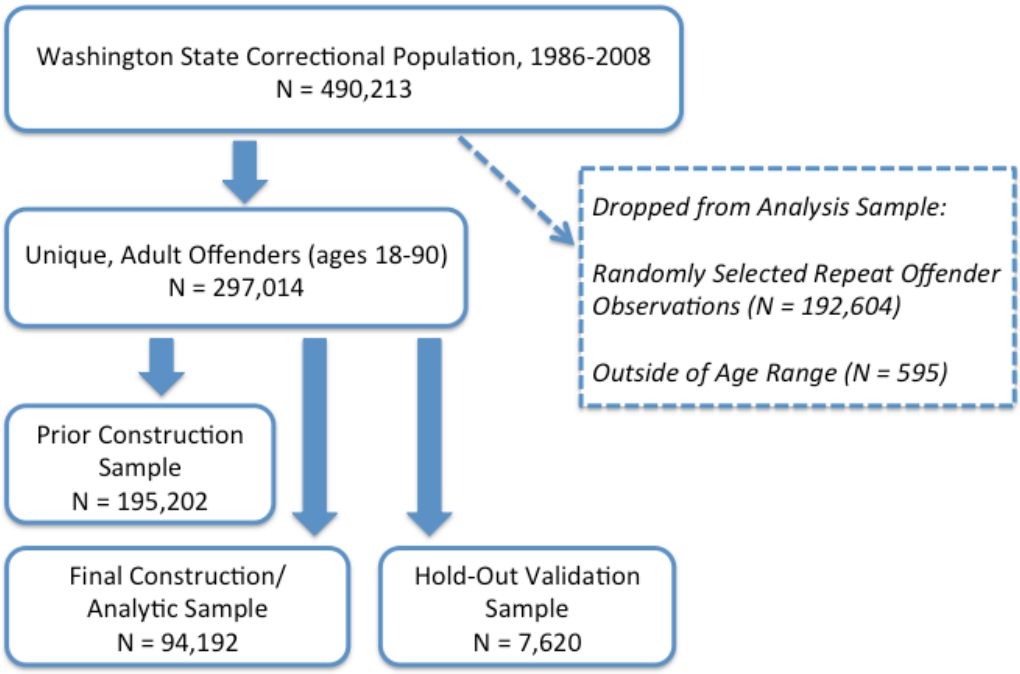
isolate differences in methodological performance of risk assessment instruments, other studies have elected to focus on only static risk items to minimize issues related to variability in interrater reliability and fidelity with dynamic scoring systems (Hamilton et al., 2015). Given that I intend to make use of sufficient historical data to generate and update empirically informative priors within a Bayesian modeling framework, I chose to utilize the same data used by Hamilton and colleagues in their comparison of logistic regression and machine learning techniques, which were also used for constructing and validating the original SRA instrument.

**Data**

Data comes from WADOC administrative records of all convicted felony offenders and gross misdemeanants that completed sentences and reentered the community between July 1st, 1986 and January 1st, 2008 (n=490,213). Recidivating offenders enter and exit the records multiple times during this period, and thus contribute more than one person-record in the data. This could present a problem for analysis, as repeated observations from the same offender have correlated error terms that violate the assumptions of frequentist and Bayesian general linear models. While hierarchical linear models could be used to ameliorate this issue, I instead randomly select one observation from each repeat offender to include in the analysis sample. Because these models only consider static inputs, trajectories of reconviction within offenders would contribute no additional information regarding the association between risk factors and recidivism risk, beyond increasing the count of prior offenses. Since variation in this dimension primarily

occurs between subjects in the data, retaining only a single, random incarceration and possible recidivism period from repeat offenders is a parsimonious solution to this problem, and one that has been used in prior research of this data (Hamilton et al, 2015). I also focus only on adult offenders who were between the ages of 18 and 90 at the time of conviction. Youth offenders serving time in adult corrections tend to be unique, in that they almost universally face long sentences for serious crimes but also tend to have no prior adult convictions. Likewise, very old offenders are extremely uncommon, and likely display a very small risk of recidivism before mortality, if they are eligible for release at all. The inclusion of either youth or very old offenders in the models may thus bias any estimates regarding age, which are central to the models depicted here. In all, these selection criteria reduce the overall analytic sample size to 297,014. Figure 2.1 describes the flow of offenders throughout the sample construction process.

**Figure 2.1: Analytic Sample Construction and Offender Flow**



## *Measures*

### *Dependent Variables*

The primary recidivism outcome of interest is any new felony conviction within the three years following offender reentry into the community. WADOC has utilized this three-year recidivism window for the purposes of risk assessment construction and validation since the introduction of the SRA (Barnoski & Aos, 2003). Other outcomes include new felony convictions for drug, property, and violent crimes over the same time period. Researchers have demonstrated that risk assessments tend to perform better when items take on different weights for specific offense types, rather than just focusing on felony recidivism more generally (Hamilton et al., 2016). In addition, the variation in item weights introduced by the use of offense-specific recidivism outcomes allows me to test whether certain characteristics associated with criminal specialization or generalized criminal behavior better predict some types of recidivism than others. All outcome measures were dichotomously coded (0 or 1).

### *Independent Variables*

Age (but relatedly, also period and cohort) is perhaps the earliest identified correlate of criminal behavior in criminology. Starting with Quetelet's (1984) work in 1831, age has been seen as a key correlate of crime at both the individual and aggregate levels. As people age through their teen years and into early adulthood, there is a decrease in parental controls and an increase in the importance of peer influences that correspond with a greater propensity for crime, which then falls again later in life as family,

community, and work-related considerations gain precedence (Blumstein et al., 1986; Farrington, 1984; Gottfredson & Hirschi, 1990). As such, most of the factors highlighted by the different criminological theories of offending are moderated by age, suggesting it is a fundamental determinant of offending. Evidence at the aggregate level suggests much the same thing (Levitt, 1999; Ulmer & Steffensmeier, 2014). Overall, then, age can be expected to impact the risk of recidivism in much the same way as it is expected to contribute to initial criminality: as people age, they tend to age out of criminal lifestyles. As such, recidivism is expected to peak in late adolescence and early adulthood, dropping off soon after. In the analysis I code respondents into six different ranges (18-19; 20-29; 30-39; 40-49; 50-59; 60+), using offenders who are 60+ as the reference group, based on when they are released from correctional supervision and come at risk of recidivism.

Gender is the single best sociodemographic predictor of crime (Hagan, 2012). Like age and prior behavior, gender is a similarly fundamental determinant of crime, though there is much less consensus about how or why exactly a person's gender so greatly impacts their likelihood of offending. Innate sex and hormonal differences (Wright & Boisvert, 2009), characteristics associated with socialization into feminine or masculine roles (Akers, 2009), differences in opportunities and social controls (Gottfredson & Hirschi, 1990; Broidy & Agnew, 1997), and other explanations have all been posited to explain the much higher rates of offending among men in the U.S. Regardless of the cause of gender disparities in crime, gender is an important contributor to offender risk. Instruments differ, however, in terms of how they deal with this. While there are convincing theoretical reasons to suggest that assessments be separated by gender (Else-Quest et al., 2012; Van Voorhis et al., 2010) and some predictors demonstrate different associations with

recidivism in male and female offenders (Smith, Cullen, & Latessa, 2009), an instrument that allowed for gender-specific item weights with Washington State data did not perform significantly better than one that simply employed gender as a predictor (Hamilton et al., 2016). This does not mean that constructing gender-sensitive risk assessments is not an important endeavor. Perhaps these efforts have not gone far enough, as intersectional theories would suggest (Crenshaw, 1989; Richie, 2012), so Chapter 4 investigates the possibility of constructing a gender *and* race-specific instrument. However, for the primary analysis in Chapter 3, I simply include gender as a covariate in the models.

All other measures included in my models are taken from WADOC's records of past adult and juvenile convictions. This type of data formed the basis for the SRA/SRA-2 and STRONG/STRONG-R assessments and consists of a total of 22 general (the number of prior adult and juvenile felony convictions) and offense-specific (e.g. violent and non-violent felony property offenses, domestic and non-domestic felony assaults, etc.) measures that summarize offenders' criminal histories at the time of release from their most recent stint of corrections. Relatedly, the outcomes of interest are any, violent, property, and drug felony convictions within 3 years of re-entry into the community.

*Descriptive Statistics*

Table 2.1 provides the means and standard deviations for all offender information contained in the SRA data, both overall and for those who did and did not recidivate. These data demonstrate that individuals facing corrections in Washington State are mostly male (79%), white (82%), or black (13%), and average approximately 31.5 years of age at the time of release into the community. Inmates have, on average, 1.4 prior adult felony

convictions and 0.2 juvenile felonies, with the most common prior felony convictions being for property (*m*=0.55), drug (*m*=0.45), or non-domestic violence assaults (*m*=0.13). Misdemeanors are also relatively common, with offenders averaging 0.34 property, 0.16 non-domestic assault, 0.15 domestic assault, 0.15 alcohol-related, and 0.12 drug-related prior convictions. Within the three-year recidivism period covered by the data, 29% of those released are reconvicted. The majority of offenders who recidivate are reconvicted for felony offenses (57%), with 21% of these being for property crimes, 17% being for drug crimes, and 17% for violent felonies. An additional 22% of them are convicted of violent misdemeanors, meaning that overall slightly more than 38% of those who recidivate are for a violent misdemeanor or felony.

In general, inmates who are reconvicted within 3 years of release are slightly more likely to be male, black, young, have a greater number of prior adult and juvenile felonies, and to have committed prior non-domestic assaults, property crimes, and drug offenses than those who did not recidivate.

**Analysis**

As described in Chapter 1, Bayesian statistics offer a new and potentially fruitful way of thinking about and constructing recidivism risk assessment instruments. Like frequentist regression models, Bayesian statistics can produce multivariate estimates of recidivism risk and utilize different link functions to obtain parameter estimates for a variety of outcome distributions. With a binary outcome measure like recidivism, I therefore employ

logistic regression models similar to those fitted with frequentist methods, which are widely used to construct risk assessment instruments.

*Item Selection*

Potential predictors were limited to the array of factors available in the WADOC data that were used to construct the SRA, representing the most basic and well-established correlates of offending: age, sex, and prior criminal records. Items were included in the recidivism prediction models if they demonstrated a positive association with felony recidivism that exceeded the size of their associated posterior standard deviations in bivariate Bayesian linear regression analyses. Those measures that were included in the data but failed to demonstrate positive associations with any felony recidivism in bivariate analyses (sex offenses and murders) were not included in the models, as the inclusion of such unrelated measures can contribute to classification uncertainty and increase statistical noise (Baird, 2009; Hamilton et al, 2016). While much of the inputs and outputs of these models are thus similar to those used in the SRA, there are several notable differences between Bayesian and Frequentist methods that highlight why and how the results of these models may diverge from those of more traditional estimation strategies, and by extension, emphasize the contributions made by this study.

## Table 2.1: Descriptive Statistics for Washington State Prisoners by Recidivism Status, 1986-2008

| | Overall | | Non-Recidivist | | Recidivist | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Male | 0.79 | (0.40) | 0.78 | (0.41) | 0.82 | (0.39) |
| *Race* | | | | | | |
| Other | 0.01 | (0.12) | 0.02 | (0.13) | 0.01 | (0.10) |
| White | 0.82 | (0.38) | 0.84 | (0.37) | 0.79 | (0.41) |
| Asian | 0.03 | (0.17) | 0.03 | (0.17) | 0.03 | (0.16) |
| Native | 0.03 | (0.18) | 0.03 | (0.17) | 0.05 | (0.21) |
| Black | 0.13 | (0.34) | 0.11 | (0.31) | 0.19 | (0.39) |
| *Age* | | | | | | |
| Age | 31.5 | (10.46) | 32.13 | (10.77) | 30 | (9.51) |
| 60+ | 0.01 | (0.12) | 0.02 | (0.14) | 0.01 | (0.07) |
| 50-59 | 0.04 | (0.21) | 0.05 | (0.22) | 0.03 | (0.17) |
| 40-49 | 0.16 | (0.36) | 0.16 | (0.37) | 0.14 | (0.35) |
| 30-39 | 0.28 | (0.45) | 0.29 | (0.45) | 0.28 | (0.45) |
| 20-29 | 0.41 | (0.49) | 0.4 | (0.49) | 0.43 | (0.50) |
| 18-19 | 0.09 | (0.28) | 0.08 | (0.27) | 0.11 | (0.31) |
| *Criminal History (Felony #s)* | | | | | | |
| Adult Felonies | 1.42 | (1.27) | 1.23 | (1.00) | 1.88 | (1.66) |
| Juvenile Felonies | 0.22 | (0.75) | 0.13 | (0.55) | 0.44 | (1.05) |
| Juvenile Violent Felonies | 0.04 | (0.24) | 0.03 | (0.18) | 0.09 | (0.35) |
| Current Commitments | 1.41 | (1.19) | 1.23 | (0.96) | 1.85 | (1.52) |
| Homicides | 0.01 | (0.12) | 0.02 | (0.13) | 0.01 | (0.09) |
| Sex Offenses | 0.07 | (0.27) | 0.09 | (0.30) | 0.04 | (0.20) |
| Violent Property Offenses | 0.07 | (0.28) | 0.06 | (0.26) | 0.09 | (0.31) |
| Non-Domestic Assault | 0.13 | (0.37) | 0.12 | (0.35) | 0.16 | (0.41) |
| Domestic Violence Assault | 0.02 | (0.13) | 0.01 | (0.11) | 0.02 | (0.16) |
| Weapons Offenses | 0.03 | (0.17) | 0.02 | (0.15) | 0.04 | (0.21) |
| Property Crimes | 0.55 | (0.89) | 0.45 | (0.75) | 0.79 | (1.12) |
| Drug Offenses | 0.45 | (0.70) | 0.39 | (0.64) | 0.57 | (0.82) |
| Escapes | 0.05 | (0.22) | 0.04 | (0.20) | 0.07 | (0.26) |
| *Criminal History (Misdem. #s)* | | | | | | |
| Non-Domestic Assault | 0.16 | (0.49) | 0.11 | (0.39) | 0.27 | (0.66) |
| Domestic Violence Assault | 0.15 | (0.46) | 0.11 | (0.39) | 0.27 | (0.60) |
| Sex Offenses | 0.02 | (0.18) | 0.02 | (0.14) | 0.04 | (0.25) |
| Domestic Othert | 0.01 | (0.09) | 0.01 | (0.07) | 0.02 | (0.12) |
| Weapons Offenses | 0.02 | (0.15) | 0.02 | (0.13) | 0.04 | (0.21) |
| Property Crimes | 0.34 | (0.75) | 0.22 | (0.59) | 0.63 | (0.98) |
| Drug Offenses | 0.12 | (0.39) | 0.08 | (0.32) | 0.21 | (0.51) |
| Escapes | 0 | (0.07) | 0 | (0.06) | 0.01 | (0.09) |
| Alcohol Offenses | 0.15 | (0.35) | 0.12 | (0.33) | 0.21 | (0.40) |
| *Recidivism and Risk* | | | | | | |
| SRA Risk Level | 2.16 | (1.33) | 1.87 | (1.15) | 2.84 | (1.47) |
| Some Reconviction | 0.29 | (0.46) | - | - | 1 | (0.00) |
| Felony Reconviction | 0.17 | (0.37) | - | - | 0.57 | (0.49) |
| Felony Property Reconviction | 0.06 | (0.24) | - | - | 0.21 | (0.41) |
| Felony Drug Reconviction | 0.05 | (0.22) | - | - | 0.17 | (0.38) |
| Violent Felony Reconviction | 0.05 | (0.21) | - | - | 0.17 | (0.37) |
| Felony Sex Reconviction | 0 | (0.07) | - | - | 0.01 | (0.12) |
| Violent Misdem. Reconviction | 0.06 | (0.24) | - | - | 0.22 | (0.41) |
| Any Violent Reconviction | 0.11 | (0.32) | - | - | 0.38 | (0.49) |
| N | 297014 | | 209768 | | 87246 | |

Note: Sample represents a single, randomly selected observation for recidivists

*Bayesian Regression Models*

Especially when considering binary outcomes, the most fundamental difference between Bayesian and frequentist approaches is clear: each is based on one of the two theories of probability that underlie all statistics. The *frequentist paradigm* advocated most prominently by R. A. Fisher, Jerzy Neyman, and Egon Pearson defines probability in terms of long-run frequency. That is, given a sample space of possible outcomes, the probability is determined by the proportion of times that a specific outcome occurs, over the total number of trials. The canonical example is that of the coin toss; we obtain a coin and flip it a given number of times (say, 50), and by counting the number of "heads" (30) we estimate that the probability of obtaining "heads" is approximately 0.60. We realize, however, that this one coin may not be representative of all possible coins, so ideally we would need to repeat this experiment as many times as possible. Theoretically, if a new sample is drawn enough times and each sample consists of an adequate number of trials, we can then ascertain the "true" probability of the event occurring from the mean of all of these samples. Since this is not practically possible, we use the standard error of the mean to compute 95% confidence intervals, within which 95% of potential sample means are expected to lay. Importantly, this does *not* imply that a given sample mean at the center of this interval is the most probable value for the population parameter—instead, values throughout the interval have a nearly equal probability of capturing the "true" population value because of post-data degeneracy in factual error probabilities (Spanos, 2012).

The *Bayesian paradigm*, on the other hand, views probability as a subjective (rather than knowable) experience of uncertainty, which we arrive at from some combination of prior knowledge and data. For example, when someone places a bet, they do so with some

idea of what to expect, given their understanding of how the game works and any other information they have available to them, such as standard conventions about how to best play the game. As the outcome of their bet is revealed, the gambler learns from this experience and updates their assumptions about the probability of them winning their next game. Thus, instead of relying on the (unrealistic) notion that we can ascertain a true value for probability by infinitely repeating an event, the Bayesian paradigm sees probability as inherently uncertain and subjective, but something that we can estimate more precisely as we learn from experience, like the gambler. Under this paradigm, we are forced to be explicit about how uncertain our expectations (or "priors") are, which are then evaluated in terms of how well these expectations capture the actual relationships present in some set of data. The results of a Bayesian analysis are given in terms of posterior probability, which summarizes the uncertainty of our prior belief, given the data, in terms of a new probability distribution that we can describe using credibility intervals (Gelman et al., 2014). Interpretation of these summaries of the new probability distribution is relatively straightforward—If we want this range of values to contain 95% of the probability, then we would report the 95% credibility interval. Unlike with frequentist confidence intervals, the central tendency of this credible interval is in fact the *most probable value* for the parameter of interest, since it is the densest point of a probability distribution that directly reflects our uncertainty about the estimate.

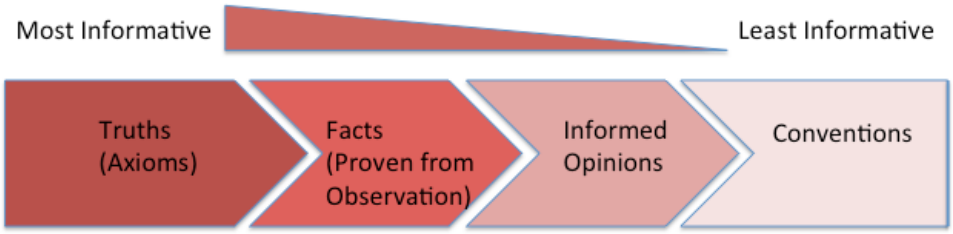
In short, Frequentists view probabilities as inherently knowable, provided that we collect enough data and that certain assumptions about that data are met, whereas Bayesians see probability as uncertain, but estimable given prior knowledge about the event, which can be evaluated and updated after analyzing new data. Despite the

dominance of the frequentist paradigm in scientific research and discourse, the Bayesian notion of probability is probably more closely aligned with how the majority of people (including scientists) think about the concept in their day-to-day lives, and certainly better captures the reality of predicting human behaviors like recidivism. Having findings that are more readily interpretable and intuitively understood by a wider audience can only be a good thing for risk assessment, where the majority of practitioners are likely untrained in complex statistical methods. Finally, the idea that prior beliefs can be updated as they are exposed to new data is potentially useful for agencies that might adopt a Bayesian risk assessment instrument, as new data that they collect can be easily integrated into existing tools. This can allow instruments to adapt to changes in recidivism risk over time or unique patterns of offending across different jurisdictions without requiring the vast amount of data that would be necessary to construct such instruments from scratch.

*Prior Selection*

Good selection of model priors is perhaps the most important aspect of Bayesian inference (Gelman et al., 2014). While the idea that Bayesian methods involve subjective choices that influence how empirical relationships should play out may be concerning to some scientists (Gelman, 2008), it is also a distinct strength of the approach that can reduce model uncertainty and increase statistical power (Loh et al., 2015; Stegle et al., 2010) if priors are well-aligned to data. Furthermore, even frequentist models regularly incorporate this type of subjective information, though it is much less explicit. Leamer (1983) provides a widely cited taxonomy (represented in Figure 2.2) of priors based on their level of confidence in providing useful information about some effect of interest.

**Figure 2.2: Leamer's (1983) Taxonomy of Priors**



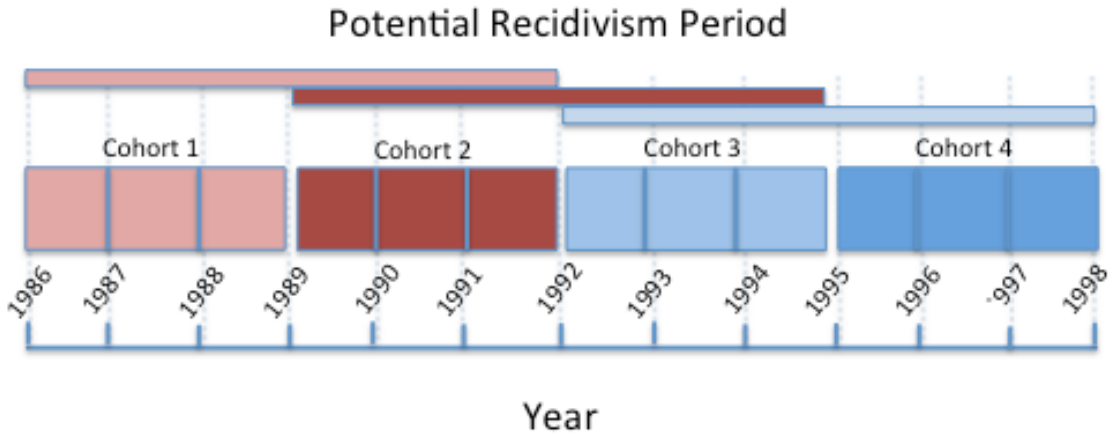

Models using conventional priors can closely resemble frequentist approaches and may simply reflect the inherent assumptions that underlie typical null hypothesis testing. For example, conventional priors are deeply engrained in the basic Ordinary Least Squares model, where we have to assume relationships are linear, that data are normally distributed, and that associations are only significant when p-values exceed specified alpha levels. The beauty of Bayesian approaches is that we can use prior information from near the top of this hierarchy, where we have a means of "systematically incorporating existing human knowledge, quantitative or qualitative, into the statistical specification" of models (Gill, 2014, p. 99). Priors derived from truths or facts tend to closely approximate information that could be gleamed from other potentially observable data, and can thus be used to improve upon the estimates obtained from frequentist models. These types of more deeply informative priors are often derived from empirically substantiated theories, past research, or observed patterns in earlier cohorts of similar data (Kaplan, 2014).

Until recently, informative prior selection has proven relatively difficult. Especially in the face of relatively new, scarce, or qualitative data, it has not always been computationally straightforward how to quantify and synthesize information into suitable priors. However, in the presence of good historical data, the quantification of this type of

information has become much more straightforward, with many studies demonstrating the utility of using historical data to elicitate strong priors in clinical trials (e.g. Berry, 1991; Chen, Ibrahim, & Yiannoutsos, 1999; Ibrahim, Ryan, & Chen, 1998). These so-called 'power priors' (Ibrahim & Chen, 2000; Ibrahim et al., 2016) hold great promise for Bayesian models predicting recidivism because many criminal justice agencies archive historical data. I exploit this fact in the Washington State corrections data to generate strong priors, which should improve the predictive power and reduce the uncertainty of estimates in the resulting risk assessment instrument.

I follow an iterative process to arrive at the power priors ultimately used in the BARR. I begin by first splitting the data into 8 separate cohorts, each determined by the dates that offenders were released (their "at-risk date" for recidivating) and the 36-month recidivism window used by Washington State (Drake, Aos, & Barnoski, 2010). Because I only use one record from each person in the data, each cohort contains all offenders released over a 3-year period, covering a total at-risk period of 6 years starting with the date that offenders at the beginning of the period are released and ending with the closing of the 3-year recidivism window for those released at the end of the third year (see Figure 2.3 for a graphical representation). Cohort sizes range from 30,456 to 50,164 across the study period.

**Figure 2.3. Example Cohort Construction, 1986-1998**

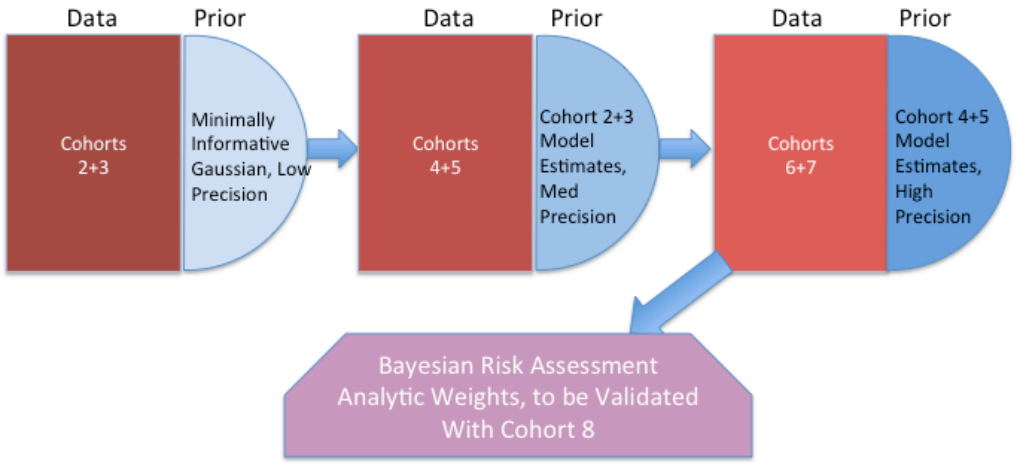



Note: Range of years is truncated for demonstration purposes. Actual data extends from 1986-2008.

The earliest cohort of data (1986-1989; n=30,456) was used to determine what type of distribution best fit the data. Previous instruments used for recidivism risk classification in Washington State (e.g. Barnoski & Drake, 2007; Drake & Barnoski, 2009; Hamilton et al., 2016) have typically used general linear models to obtain risk estimates, with most employing logit link functions for the dichotomous recidivism outcome data. A series of bivariate analyses between the overall felony recidivism outcome and criminal history and demographic predictors confirmed this trend in the first cohort of the Washington State correctional data, leading me to conclude that a Gaussian distribution could be used to accurately characterize the relationships between predictors and outcomes. In Bayesian estimation, priors are defined in terms of *hyperparameters*, which the researcher uses to specify the characteristics of the prior distribution (Gelman et al., 2014). The minimally informative prior that I employ based on these preliminary results has two

hyperparameters that need to be specified: the distributional shape and precision, or variance, of the distribution. Thus, from this cohort I generate a Gaussian prior with very low precision (B0=0.0001), which allows the data in subsequent cohorts to largely dictate the shape of the posterior distribution.

**Figure 2.4: Prior Construction Process**




Note: Lighter colors signify that less weight is given to this element of the model during analysis. Thus, data is weighted more heavily in early cohorts, whereas priors are weighted more heavily in the final models due to increased precision.

From here, prior selection was the result of an iterative process, where I tested a variety of methods for combining cohorts, coding predictor variables, and determining the most suitable level of precision for priors using Bayes Factors (described below) to determine optimal model specifications. The process I describe below (as summarized by Figure 2.4) reflects the decisions made to arrive at this optimal model. To leverage additional statistical power, I combined Cohorts 2 and 3 (n=75,117) to test the initial Bayesian model and begin to estimate more empirically informed priors for use in later

cohorts. In this model, Bayesian logistic regression was used to predict reconvictions as a function of age, gender, and previous offending histories, which were all expected to have a Gaussian relationship with recidivism based on the findings of preliminary analyses in Cohort 1. The very low precision of this prior belief allowed for the relationships present in the data to speak for themselves, yielding results that were nearly identical to a traditional frequentist model fit to the same data (results available upon request).

At this stage, I also recoded each predictor to ensure that variable categories would result in positive model coefficients (increasing the risk of felony recidivism) that increase monotonically with variable risk (e.g. ensuring that the odds of recidivism increase with the number of prior juvenile felonies). This often necessitated top-coding variables, as very high numbers of some types of previous sentences were relatively infrequent in the data and displayed inconsistent relationships with recidivism. In the case of non-domestic assaults, this required dichotomizing the measure so that two or fewer offenses were coded "0" and three or more instances were coded "1" to empirically reflect where the increase in recidivism risk actually appeared to occur. This method of coding variables, which resulted in a mix of continuously coded and categorical variables, is somewhat unique in risk assessment where measures are traditionally left as continuous with some top-coding. However, many of the criminal history characteristics and age displayed non-linear associations with recidivism that were best represented using categorical variables, and all of these coding decisions resulted in notably improved model fit and reduction in model error. The resulting point estimates and standard deviations of the posterior probabilities were then retained and used as the hyperparameters for the power priors employed in the next cohorts of data.

I again combined two cohorts of data (cohorts 4 and 5, N=89,285) for the next round of analyses, which were used to refine the priors with a more recent set of data before making them the basis for the instrument's final item weights. The previous model's point estimates were used as priors for the cohort 4-5 models, with the priors' precision determined by running a series of models using various multipliers (0.5, 1, 2, 3, 5, 10, 20, and 30) of the accompanying posterior precision (1/SE^2) and comparing Bayes Factors from the resulting models. Bayes Factors are used to determine the level of support for one prior over another, as they indicate the relative likelihood of the observed data occurring under one set of prior assumptions versus another. Kass and Raftery (1995) proposed guidelines for the strength of the evidence demonstrated by Bayes Factors (BF), suggesting that there was at least moderate evidence for one prior over another when BF>3, strong evidence when BF>10, and very strong evidence when BF>30. The best precision multiplier and resulting prior specification was chosen from the model demonstrating at least moderate evidence over the other possibilities, which in all cases also demonstrated at least very strong evidence over comparison models using a minimally informative Gaussian prior, that closely approximated traditional logistic regression models. From this optimized model, I again saved the resulting regression coefficients and posterior precisions to serve as the basis for priors in the final instrument construction model.

The final models used for instrument construction were estimated using the data from cohorts 6 and 7 (N= 94,192), with priors taken from the posterior estimates of the cohort 4-5 model. The precision of these priors was again determined by examining the Bayes factors of models estimated using a variety of multipliers of the previous model's precision, and again the best fitting model had strong evidence for better reproducing the

observed data than comparable, minimally informative Bayes models that approximated traditional frequentist logistic regression results. I repeat this entire process for each of 4 different recidivism outcomes, allowing there to be different item weights for any felony recidivism, violent felony recidivism, property felony recidivism, and drug recidivism. The posterior estimates of the log-odds of each type of recidivism and the accompanying 95% Bayesian Credibility Intervals derived from these final sets of models, as well as the log-odds and 95% confidence intervals from frequentist models using the same data and variable coding schemes, are reported in Chapter 3. The results reported in Chapter 4 used a very similar approach to prior construction, with some slight modifications described in further detail in that chapter.

*Model Convergence*

Aside from the existence of model priors, the other significant departure from frequentist statistics is how the models are actually estimated. Unlike frequentist logistic regression, which uses maximum likelihood to arrive at estimates of model parameters, Bayesian model estimates use Monte Carlo Markov Chains (MCMC) to sample directly from the joint posterior distribution of model parameters. While a full description of this process is beyond the scope of this dissertation (but see Gelman, 2014; Siu & Kelly, 1998; Kelly & Smith, 2009), essentially Markov chains maximize the marginal likelihood of a given area in parameter space with starting values determined by random draws from the prior distribution. These chains expand iteratively until they achieve adequate coverage of the underlying posterior distribution, which is itself a function of the prior distribution and some covariance in the observed data. Because the process is inherently stochastic and

parameter spaces defined by multivariate models can be highly multidimensional, the first chains tend to give highly disparate views of the posterior. For the stability of model estimates, these starting chains are thus omitted from the final chains used to summarize the posterior probability by specifying some 'burn-in' period that sets a cut-point wherein chains are stabilized and retained for model fitting. To achieve this in regression models with many covariates and large samples, a proliferation of chains (both burn-in and total) are necessary to ensure adequate coverage of the posterior distribution. It is entirely possible in these cases for models to fail to converge properly, which is why Bayesian analyses should always report not only an extensive description of the prior selection process, but also some evidence of how model convergence was assessed (Van De Schoot et al., 2014).

Given the large number of covariates, it was more difficult than anticipated to achieve adequate coverage of the posterior distribution in the present study, requiring a very large number of total Markov chains (220,000) and a long burn-in time (65,000) before estimates stabilized. The resulting models were very computationally intensive, but showed good convergence properties according to the Gelman-Rubin criterion (Gelman et al., 2014), including careful examination of Geweke and Gelman plots and diagnostics. I inspected all trace plots manually for evidence of whether chains tended to converge on the same target distribution, which resulted in my decision to use such a long burn-in period—it was only after approximately 50,000 iterations that Markov chains became stable, with best results occurring after removing the first 65,000 chains. Attempts to use fewer Markov chains, overall, returned very similar results, though they tended to be somewhat less

stable than those reported in this study, while increasing the number of chains further simply reduced computational efficiency.

*Instrument Validation*

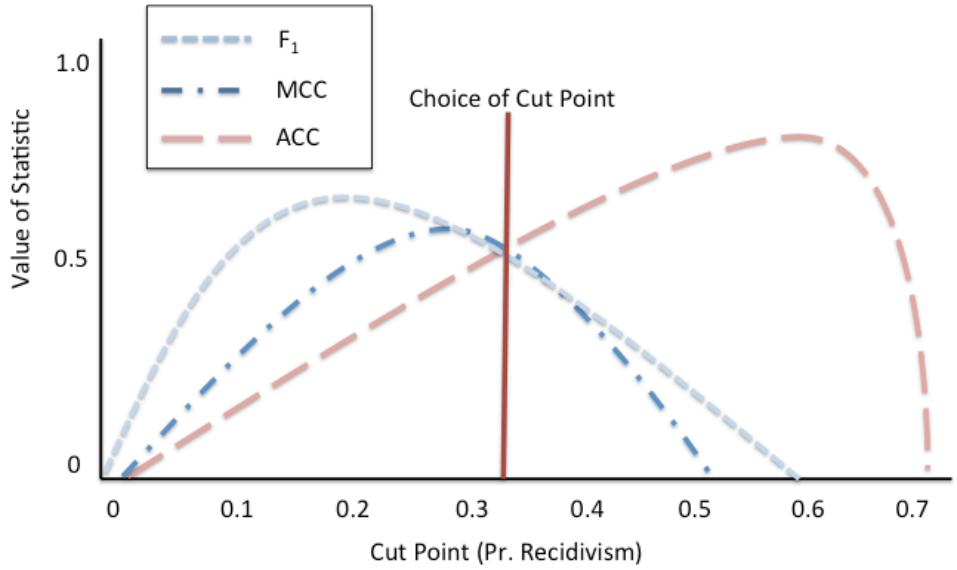
To compare the predictive characteristics of the proposed Bayesian and frequentist classification models, I compute four of the most commonly reported measures of risk assessment validity: Area Under the Receiver Operating Characteristic Curve (AUC), classification accuracy (ACC), the F-score ($F_1$ Score), and the Matthews Correlation Coefficient (MCC). When modeling recidivism risk, AUC is a measure that summarizes a classification system's ability to discriminate between those who do and do not recidivate, overall, across the Receiver Operating Characteristic (ROC) curve[3]. ACC, on the other hand, is perhaps the simplest and most intuitive metric that can be used to assess classification errors, though it is very sensitive to sample imbalance in the outcome and requires that a cut point be defined for classification decisions, unlike AUC. The ACC is the percentage of cases correctly classified by this cut point in a confusion matrix (true positives + true negatives/ positives + negatives), which is easily transformed to obtain the incorrect classification rate (1-ACC). The $F_1$ Score is the harmonic average of precision (or positive predictive value; true positives / true positives + true negatives) and sensitivity (or the true positive rate; true positives / positives). The MCC (or phi coefficient) gives the correlation between observed and predicted binary classifications and is generally regarded as an optimal metric for assessing classification accuracy in the presence of

[3] The ROC curve plots sensitivity (or the *true positive rate*) against specificity (or the *false positive rate*) and the AUC is a way of summarizing classification error at various thresholds that prioritize minimization of one type of error over the other.

imbalanced outcome data (Boughorbel, Jarray, & El-Anbari, 2017; Powers, 2011). Each of these accuracy measures is related (see Figure 2.5 for a depiction of the relationship between these statistics and hypothetical cut point selection) and ranges from 0 to 1, with higher numbers generally indicating better predictive properties.

To determine cut points for computing ACC, $F_1$ Score, and MCC, I started with the base rate of each offense, and then increased the cut point until it optimized each of the three statistics (but especially MCC), allowing these points to vary across the frequentist and Bayesian models. In the event of ties, I chose the lowest cut point that maintained an optimal rating on each of the three statistics. In my mind, this strategy results in the best compromise between optimizing positive and negative classifications, which should minimize the number of classification errors made in practice. Table 2.2 gives these cut points for each outcome. I use the same cut-points to classify offenders as generally high risk, high violent risk, high property risk, high drug risk, or low risk (those not classified as recidivists by any of the models), giving practitioners a simple classification scheme for guiding decision-making according to offenders' risk of recidivating across a variety of outcomes.

**Figure 2.5: Relationship of Accuracy Statistics to Cut Point**




Note: Figure is for demonstration purposes only and does not represent the actual calculated statistics for the reported recidivism prediction models.

**Table 2.2: Classification Cut Points**

| | Predicted Pr. | |
|---|---|---|
| Any Felony | Bayes | 0.28 |
| | Freq | 0.29 |
| Violent Felony | Bayes | 0.15 |
| | Freq | 0.11 |
| Property Felony | Bayes | 0.15 |
| | Freq | 0.29 |
| Drug Felony | Bayes | 0.10 |
| | Freq | 0.10 |

Chapter 3 – EMPIRICAL ANALYSIS:

BARR Validation and Offense Specialization

## Introduction

A large body of research has focused on improving the science of risk assessment (e.g. Andrews, 1982; Berk, 2017; Chouldechova, 2017; Duwe & Kim, 2016; Gottfredson & Moriarty, 2006; Hamilton et al., 2015, 2016; Liu et al., 2011; Zeng, Ustun, & Rudin, 2017), with much of this attention focused on development, validation and refinement of new or existing regression-based instruments (Andrews & Bonta, 1995; Barnoski & Drake, 2007; Brennan & Oliver, 2000; Hare, 1991; Latessa et al., 2009). Given the practical nature of these instruments and importance of risk classification for public safety, rehabilitative resource allocation, and bureaucratic efficiency, it is perhaps unsurprising that many researchers are concerned with identifying new methods that can better separate risky offenders from those who pose less of a threat. For example, there has been a rapid escalation of interest in using machine learning techniques such as random forests and neural networks for recidivism risk assessment (Berk et al., 2009; Liu et al., 2011; Schaffer et al., 2011), though evidence is mixed regarding their effectiveness relative to more conventional methods (Berk, 2017; Hamilton et al., 2015; Zeng, Ustun, & Rudin, 2017). That no one has yet tested a Bayesian risk assessment instrument for recidivism is therefore surprising, given evidence for its predictive potential in the area of genetic risk assessment (Fenton & Neil, 2012; Newcombe et al., 2012; Ogino & Wilson, 2004).

Yet, because of this understandably practical focus on improving predictive validity, researchers developing these instruments have largely ignored discussing the implications of methodological variation on uncertainty (i.e. error) in model estimates. While it is already difficult to compare the potential of different methods because of variation in base offending rates, items selection, jurisdictional laws and practices, and validation procedures (Hamilton et al., 2015), it is even more challenging to determine how precisely models are estimating the item weights ultimately responsible for scoring. Most validation studies omit reporting the standard errors or confidence intervals of point estimates obtained by their models, if these reports are forthcoming with the coefficients used to obtain item weights at all (e.g. Barnoski & Drake, 2007; Colorado Division of Criminal Justice, 2008; Turner, Hess, & Jannetta, 2009). Aside from those developing instruments for specific states like California (Turner, Hess, & Jannetta, 2009), Georgia (Meredith, Speir, & Johnson, 2007), Minnesota (Duwe, 2014), and Washington (Barnoski & Drake, 2007; Hamilton et al., 2016) or locally norming existing instruments (e.g. Wagner et al., 1998), there is no recognition that risk estimates are expected to vary across potential samples, which could dramatically reduce their ability to reliably classify new cohorts or populations of offenders. In fact, instruments like the ORAS (Latessa et al., 2009) and LSI-R (Andrews, Bonta, & Wormith, 2004) claim robustness to a wide variety of populations and are widely used throughout the U.S. despite the fact that they were constructed and validated using small, specialized samples that would yield considerable uncertainty in model estimates. For optimal prediction in the real world, it is critical that models not only demonstrate sufficient validity, but minimize both estimation and prediction error.

The focus on practical concerns of risk assessment has also resulted in a more surprising omission: researchers have regularly failed to highlight potential theoretical contributions to the field of criminology that their findings may make. While many of the predictors used in these instruments have some basis in criminological theories of offending, perhaps best demonstrated by the LSI-R and LS/CMI's reliance on measures explicitly derived from social learning theory (Andrews & Bonta, 1999; Bonta, 2002), few have taken risk assessment validation research as an opportunity to study criminological theory. Coincidently, the most notable exceptions have been examinations of the LS family of instruments. Andrews and colleagues have been some of the few to highlight how the validity of their instruments, compared to others based on sociological or psychopathological measures, suggests support for social learning theory (Bonta et al., 1998; Bonta, 2002). Their instruments and associated theory have also received significant critique from feminist criminologists, who in studying the LSI-R's predictive validity among men and women have highlighted the importance of gendered pathways to crime (Holtfreter & Cupp, 2007; McGee & Mazerolle, 2016; Reisig, Holtfreter, & Morash, 2006; Smith, Cullen, & Latessa, 2009; Van Voorhis et al., 2010). Their contributions have come full circle, with some risk assessments now including gender-specific models to better capture these theoretical developments (Hamilton et al., 2016).

This chapter primarily addresses the predictive validity of a Bayesian risk assessment instrument, relative to a nearly identical classification scheme similar to the SRA that was developed using frequentist methods. But in doing this, I also highlight improvements that these models make for reducing model and prediction uncertainty. The findings also speak to one of the longest-running debates in criminological theory, perhaps

best represented by Gottfredson and Hirschi's claims (1986) of a general explanation for crime and Cloward and Ohlin (1960) and Moffitt's (1993) suggestions that certain offenses have unique explanations: do offenders tend to be 'versatile' in their criminal behavior, or do they specialize in certain types of offenses? I conclude the chapter by discussing implications of the models for classification science, generally, and for criminological theories of offending related to the versatility versus specialization debate.

**Classification Error and Model Validity**

To improve the predictive performance of risk assessment instruments, researchers hope that new predictors or different estimation strategies will yield point estimates (i.e. regression coefficients) that will help to better differentiate between recidivists and non-recidivists. These estimates are not only important for producing the predicted probabilities that are used to assess model accuracy and discrimination but, in assessment tools that employ analytically derived weights, they also form the basis for the scores used in the resulting instruments employed by practitioners (Bonta, 2002). Thus, for determining the statistical validity and practical applicability of these instruments, no statistic is more important than the coefficient.

However, this preoccupation with coefficients ignores another important statistical property that is equally relevant to classification: error. While there are many potential sources of error in any classification scheme (see Bowker & Star, 1999), the most relevant for the statistician hoping to improve risk assessment models are probably the standard error of the mean and prediction error. Both are directly related to the point estimates that

researchers are so often concerned about, but rarely addressed as a strength or weakness of a new method of classification. Perhaps this is because they do not have as direct a bearing on the primary statistic of interest for most researchers (AUC); yet, reducing these errors can give us more confidence in our predictions at the individual level, where they ultimately matter the most in practice.

For out of sample prediction, the standard error of a regression coefficient is arguably as important a statistic as the coefficient itself. Standard error expresses the reliability of a sample statistic relative to its parametric value in the population (McDonald, 2015), and is used to compute confidence intervals that, in frequentist statistics, denote a range of values wherein 95% (or 99%, or 90%) of possible sample means (or regression coefficients) are expected to lie. For example, an odds ratio of 2.12 with a 95% confidence interval ranging from 1.25 to 2.99 shows that a characteristic is expected to increase the odds of some outcome occurring by 112% based on the present sample, but that 95% of other samples from this same population would yield projected increases between 25% and 199%; furthermore, there is a nearly equal probability of the true population value lying anywhere within this interval (Spanos, 2012), contrary to common misperception that the point estimate is the 'most likely value' in the interval. This presents a problem for risk classification, especially when individuals fall near cut points on an assessment instrument. Over the many characteristics that can contribute to a person's risk, this problem is compounded, and there can be a great deal of variability in how offenders could be classified depending on which sample means produced the regression coefficients that were ultimately used to derive item weights. Even worse, the instruments are by their very

nature applied to new samples in practical settings, where the specific sample means that produced the risk estimates may not be the best fit.

The Bayesian approach to modeling helps to alleviate some of this uncertainty. First, well-defined priors (those that align with the patterns present in the data) help contribute additional power to the analysis (Loh et al., 2015; Stegle et al., 2010), generally producing smaller standard errors and tighter confidence (or credibility) intervals than comparable frequentist approaches (Gray et al., 2015). Second, the resulting credible interval, which summarizes the posterior probability distribution (Van De Schoot et al., 2014), actually reflects the likelihood of the outcome (here, recidivism) occurring. This means that while there is still some uncertainty in the point estimates of Bayesian models, the coefficient has the highest probability of all other points in the credible interval of correctly predicting the outcome, and values near the tails of the interval have a much smaller probability of being correct. For the purposes of classifying offenders as potential future recidivists, Bayesian methods should thus produce item weights that are more precise (Gray et al., 2015) and less likely to reflect sampling error, giving us greater confidence in their ability to distinguish between offenders on the cusp of risk categories. While this may not necessarily result in higher model AUCs and ACCs than conventional frequentist models, reductions in the widths of the accompanying confidence intervals would signal support for this hypothesis. This greater specificity in item weights should minimize classification errors when the instrument is then used to make out of sample predictions in the real world, particularly when they are applied to samples whose characteristics differ from those used to develop the instrument.

*Expectations Related to Model Validity*

There are thus two ways that Bayesian methods may potentially improve on the predictive validity of recidivism risk assessment. First, as outlined in earlier chapters, current strategies derived from frequentist statistics or machine learning may simply fail to make the best use of the available data or lack the power of well-specified Bayesian models, resulting in over or underestimates of the recidivism risk associated with a given characteristic. When this is the case, the model-derived predicted probabilities of recidivism in test data will fail to predict observed recidivism, resulting in lower percentages of cases correctly classified. With fewer cases correctly classified, these models will tend to perform worse along the ROC curve and ultimately demonstrate lower AUC and ACCs than models that more accurately capture the risk posed by various offender profiles. Thus, if Bayesian models using the same predictors and construction data can more accurately differentiate between recidivists and non-recidivists than similar frequentist models, then we would expect the Bayesian results to classify a greater percentage of cases correctly and achieve higher AUC and ACCs than existing models. We might consider this *mean* improvement, of sorts, in that the point estimates derived from the models are a truer approximation of offenders' demonstrated reoffending behavior, on average, in the model that demonstrates the best predictive properties.

The other way that Bayesian risk assessment can improve classification, which has received much less attention from criminologists, is in the minimization of estimation and prediction error. The surer we can be of the risk score for a given offender, the fairer our assessment and the less likely there will be error in making predictions. Bayesian models

with well-specified priors can display greater power than similar frequentist ones (Loh et al., 2015; Stegle et al., 2010), which tends to minimize standard error and confidence interval width. Because estimates are drawn from a probability distribution, Bayesian methods also produce point estimates that have the highest likelihood of being correct, meaning that there is less uncertainty involved in the estimation of item weights. For both of these reasons, Bayesian models are anticipated to improve the *error* inherent in classification schemes. These expectations lead to two hypotheses regarding the relative advantages of Bayesian methods for the validity of risk assessment.

**Hypothesis 1:** Because Bayesian logistic regression models take into account a greater array of information about how prior offenses and demographic characteristics relate to recidivism, these models will display higher accuracy (ACC, MCC, and $F_1$ Score) and discrimination (AUC) than frequentist logistic regression models with the same sets of covariates for:

*a.* Felony Recidivism

*b.* Violent Felony Recidivism

*c.* Property Felony Recidivism

*d.* Drug Felony Recidivism

**Hypothesis 2:** By using strong priors, we can reduce the uncertainty in model estimates, which will yield less biased and more accurate estimates of individual risk classification. This will be demonstrated by:

*a.* Narrower confidence intervals in model estimates

*b.* Less biased and lower standard errors of prediction

**Theoretical Considerations**

In many ways, the type of data used to construct and validate risk assessment tools is ideal for studying offending behaviors and affirming or modifying existing theory. Whether an instrument intends to use static (e.g. age, gender, criminal history) or dynamic (e.g. employment, criminogenic attitudes, relationships) predictors, many of these factors reflect the presence or absence of important mechanisms influencing offending behavior in the foremost theories of crime. For example, age, gender, and criminal history play fundamental roles in developmental and life course theories of crime (e.g. Catalano & Hawkins, 1996; Farrington, 1991; Moffitt, 1993; Sampson & Laub, 1995; Thornberry & Krohn, 2001), whereas criminal history, personality characteristics, attitudes, and social support networks are "the Big Four" of social learning theory (Andrews & Bonta, 1998). The outcomes of interest in studies testing these theories have typically been forms of self-reported delinquency, but recidivism (self-reported, arrest, or reconviction) drawn from official sources is another relevant outcome for studying the maintenance and desistance of criminal lifestyles—a phenomenon that lies at the heart of life course (Laub & Sampson, 1993, 2006; Sampson & Laub, 1990, 1992, 2005; Warr, 1998), taxonomic (Moffitt, 1993, 2015; Leaw et al., 2015), and general/self-control theories of crime (Gottfredson & Hirschi, 1990; Grasmick et al., 1993; Hirschi & Gottfredson, 1993; Piquero et al., 2005).

Data that incorporate both static and dynamic characteristics of offenders are especially well suited to lending support to or challenging these theories. Such data may be

equipped to test how the developmental and interactional mechanisms posited by these theories impact recidivism (i.e. persistent offending or the maintenance of criminal careers). Nonetheless, even fully static data like that used here contain measures of age, gender, prior criminal history, and recidivism, which tap into the core tenants of life course and social control theories. Strong, positive relationships between young age, male gender, and extensive criminal backgrounds and recidivism in this data would bolster support for each of these theories, as younger and more active male offenders tend to desist from crime later in life (*á la* Sampson and Laub), have a higher likelihood of belonging to a life-course-persistent group of offenders who tend to chronically reoffend (*á la* Moffitt), and likely possess less self-control and have greater access to criminal opportunities (*á la* Gottfredson and Hirschi) than older, female, and less experienced offenders, generally. These findings are commonplace in criminological research and the assumed connections between these characteristics and future offending underlie nearly all risk assessment instruments; however, an examination of how these factors play out in offense-specific models can be used to test hypotheses derived from more specific theories related to offense specialization or versatility.

While most data used to construct risk assessment instruments can distinguish between offense types (violent, property, drug, and sex felonies), and several instruments contain different item weights for each of these outcomes (e.g. COMPAS, MnSTARR, SRA, STRONG-R), most use this information simply to improve the predictive accuracy of their model (Kroner et al., 2005). To my knowledge, none have taken this as an opportunity to examine one of the original criminological conundrums—namely, whether there are

separate classes of offenders that can be classified according to which crimes they typically commit.

*Are Criminals Versatile or Specialists?*

There is a long history of speculation and debate in criminology about whether criminals tend to exhibit *specialization* in certain areas or *versatility* across different types of offenses over their careers (Armstrong, 2008; Blumsttein et al., 1988; Brennan, Medrick, & Moitra, 1989; Britt, 1996; Healy & Bronner, 1926; Kempf, 1987; Klein, 1984; Sillivan et al., 2006). The bulk of this evidence suggests that specialization is heavily dependent on the age of delinquency onset, but that even offenders who do show evidence of specialization later tend to engage in a wider array of more serious offenses if they fail to desist from criminal lifestyles (Deslauriers-Varin, Lussier, & Tzoumakis, 2016). Interestingly, despite limited evidence for widespread specialization of criminals, the concept remains theoretically attractive for a number of reasons.

The idea that the majority of offenders tend to specialize aligns well with both the popular discourse, in which offenders are often described as "murderers", "sex offenders", or "drug addicts", as well as with risk classification systems that include different item weights for each category of recidivism. Nonetheless, the perspective that offenders tend to mainly commit the same types of crimes throughout their careers is somewhat at odds with many of the foremost theories of offending, which propose unified theories of crime (Farrington, Synder, & Finnegan, 1988; Piquero et al., 1999). Under general strain theory (Agnew, 2006; Broidy & Agnew, 1997; Merton 1938), for example, individuals who struggle to achieve societal goals or face mistreatment experience negative emotions that then

increase their propensities for crime. This propensity for crime then demands some corrective action, of which crime and delinquency are possible responses. Self-control theory is perhaps even more basic, linking the propensity for criminal behavior to individuals' level of control over their actions and emotions (Gottfredson & Hirschi, 1990). Implicitly, these and other generalist explanations suggest that offending patterns should be marked by versatility, and that knowledge of what a specific offender's first crime was would not help to predict the type of their second offense, beyond knowing that they have a history of offending (Piquero et al., 1999). But, if this were true, multivariate risk assessment models would tend to find a positive association between the total count of previous felony sentences and recidivism, with no additional information conveyed by the types of offenses that these sentences consisted of. Many instruments, however, do give various types of prior offenses different item weights on top of assigning risk to an overall number of offenses, implying that this assumption is not typically born out in multivariate models.

If, instead, we assume that different theoretical constructs underlie each type of offending, then there may be some factors that foster a unique tendency toward violent, property, or drug crime (Farrington et al., 1988; Piquero et al., 1999). From this perspective, we would expect some degree of specialization in offending, where individuals tend to engage in one type of offense throughout their career, with minimal variation. Unlike with the generalist perspective, having information about the type of someone's last offense would help predict the type of a future offense to a greater degree than simply knowing that they have a criminal record. The implication is that multivariate models would find counts of specific types of offense to be informative predictors even after

controlling for a count of prior sentences, which seems to reflect the reality implied by the majority of risk assessments that include information on in-depth criminal histories. While the majority of general theories inherently discount the possibility of specialization, and there is limited evidence that offenders are not versatile, some modern theories speak to this possibility and provide the groundwork for a set of testable hypotheses related to offender specialization.

*Theoretical Evidence of Specialization*

Though the majority of general theories of crime suggest there is a single underlying factor that increases individuals' propensities for crime, implying versatility in offending behaviors, there are some theories that have successfully integrated general explanations while allowing for the possibility that specialization does occur. Perhaps the oldest and most explicit examples of this would be cultural deviance perspectives (e.g. Anderson, 2000; Cloward & Ohlin, 1960; Sampson & Groves, 1989; Shaw & McKay, 1942). These theories propose that the local cultures of a place interact with individuals' socioeconomic, racial/ethnic, and gender characteristics in ways that may encourage or discourage offending, overall, but also in ways that determine their cultural embeddedness in a space. When the culture of a place encourages drug use, for instance, the degree of embeddedness may then also foster a propensity for specialization in drug crime. Likewise, the characteristics of a place may encourage violence, white-collar crime, or property offending, if those are culturally normative behaviors; regardless, these perspectives directly specify mechanisms that may help explain both the versatility of some offenders, and the specialization of others. Unfortunately, testing for the presence or absence of

cultural embeddedness is difficult with archival data like that collected by the Washington State Department of Corrections, and would require a great deal more information about the backgrounds of inmates including where they lived and with whom they tended to associate. Luckily, two other leading theories provide hypotheses regarding specialization that are much more easily testable using the types of variables available in the Washington State corrections data. Gottfredson and Hirschi's (1990; Hirschi & Gottfredson, 1995; 2008) general theory of crime and Moffitt's (1993; 2015; Moffitt & Caspi, 2001) taxonomic theory of offending both place emphasis on the roles of age and the types of offenses that a perpetrator commits to understand whether individuals persist in criminogenic lifestyles and to what degree specialization occurs in their criminal careers.

For Gottfredson and Hirschi, self-control is the basic trait that limits most people from engaging in crime. In general, they propose that men tend to have less self-control than women, youth tend to lack the discipline and restraint of adults, and the impulsivity, risk-seeking behaviors, and self-centeredness that come with a lack of self-control makes some people more inclined to commit crime than others. These traits tend to emerge early in life, leading to an early age of onset for offending, more frequent offending while engaged in a criminal lifestyle, and later desistance from crime than those with high self-control. Relatedly, the impulsivity and self-centeredness underlying offending also imply that criminals motivated by low self-control are opportunistic, meaning that the choice of crime likely varies from situation to situation. For those with low self-control, then, Gottfredson and Hirschi's theory predicts that "there will be much versatility among offenders in the criminal acts in which they engage" (1990; pp. 91). In contrast, those with high self-control have greater agency over their emotions and actions. This may result in

less offending, overall, but it also may lead to greater employment of rational thought, causing some to only engage in types of crime that are particularly rewarding and outweigh the potential costs of punishment (Cornish & Clarke, 1987). Later explications of their theory made the connection between offense diversity and self-control even more explicit, contending that an index of the different types of offenses committed by a person could be used as a valid proxy for their level of self-control (Hirschi & Gottfredson 1993; 1995).

Similarly, Moffitt's taxonomy of offending (1993) distinguishes between two categories of criminal offenders who are classified according to when they begin to engage in crime and the frequency and duration of their involvement in antisocial behavior. Inadequate socialization, neurological defects, and destructive parental responses to early problem behavior interact to ensure that some children miss out on the opportunities to practice prosocial behaviors at important stages of their development, which Moffitt suggests increases their chances of committing delinquent acts throughout their lifespan. This relatively small group of offenders who generally begin offending at a young age and continue to engage in problem behavior throughout the life course are labeled "life-course persisters". These persistent offenders make up only about 10% of the population but commit about half of all crime, according to Moffitt (1993, 1994), and engage in a broad variety of offending behaviors including violent and serious property crime. Thus, life-course persistent offenders are expected to have more varied offending histories and to engage in more serious violent and property crime in adulthood than others, whose deviance tends to be limited to adolescence and whose delinquent acts are more specialized and typically restricted to non-violent crimes and status offenses.

Taken together, Moffitt's taxonomy and Gottfredson and Hirschi's theories about social control provide the basis for several expectations about offense diversity, criminal specialization, and what contributes to the risk of certain types of reoffending, which can be examined using the criminal history and basic demographic information included in the Washington State correctional data. From Moffitt's taxonomy, I assume that violent recidivists are more likely than other types of criminals to be life-course persisters, who are expected to be versatile in their offending behaviors. Gottfredson and Hirschi suggest that versatility is a function of low self-control, which is itself associated with having longer histories of offending and being male. These assumptions allow me to test two hypotheses related to offender characteristics and the risk of violent recidivism.

***Hypothesis 3:*** After controlling for current age and gender, the strongest non-demographic predictors of property and drug recidivism will be the number of prior offenses of the same type, indicating specialization among those who persist in these types of criminal lifestyles. For violent recidivism, the strongest predictors will be the overall counts of juvenile and adult felonies, which signify the length and intensity of offenders' criminal careers and are indicative of low self-control and life-course persistent behavior.

***Hypothesis 4:*** Net of their contribution to prior felony counts, histories of drug and non-violent property offenses will be negatively associated with violent reconvictions. This would help to confirm the patterns implied in hypotheses 1 and 2—namely, that offenders with a history of specialization in non-violent property and drug crimes are relatively unlikely to be reconvicted for violent offenses after they are released from correctional

supervision, and that violent recidivists tend to be generalists rather than specialists (even in violent crime).

## Methods

The construction and validation procedures for the Bayesian Assessment of Recidivism Risk (BARR) are described in detail in Chapter 2. The results reported here are taken from model estimates derived from Bayesian and frequentist logistic regression models fit to the final, Cohort 6+7 construction sample (N= 94,192) of Washington State inmates released from July 1, 2001 to June 30, 2007. MCMC was used to estimate 95% credible intervals from the posterior probability distribution in the Bayesian models, rather than the Highest Density Interval, to enable direct comparison with frequentist 95% confidence intervals (Gray et al., 2015). These regression results provide the basis for testing Hypotheses 2a, 3, and 4.

Measures of instrument validity were calculated by generating predicted probabilities from these models, making classification decisions based on the cut-points described in Chapter 2, and testing these predictions against observed recidivism outcomes in a hold-out sample of offenders released on or after July 1, 2007 (N=7,620). While performing validation on a single hold-out sample is generally seen as inferior to k-fold cross-validation (Hamilton & Kigerl, 2016), these more stochastic and automated validation procedures were not possible to implement with the Bayesian model. To correctly perform K-fold cross-validation, every step of the modeling procedure (including variable selection, recoding, and regression modeling) must be subject to the resampling

procedure (Hastie, Tibshirani, & Friedman, 2016). While this is possible with both frequentist models (as model estimates are easily stored and manipulated and item selection decisions can be automated) and machine learning techniques (which are entirely automated) in R, the packages currently available for conducting Bayesian analyses[4] do not store the MCMC estimates in a way that is usable with these procedures[5]. Also, as described in Chapter 2, several aspects of my model building procedure related to prior selection and cut point optimization were the result of subjective, iterative processes that could not be automated.

Nonetheless, split-sample validation procedures are common in the risk assessment literature (e.g. Barnoski & Drake, 2007), and recent statistical experiments have demonstrated that the two methods produce equivalent results in cases where sample sizes are relatively large (Yadav & Shukla, 2016), as is the case in the present study. In fact, some statistical handbooks only recommend cross-validation when a split sample approach would compromise the accuracy of parameters in the construction sample (Hastie, Tibshirani, & Friedman, 2016). If anything, subjecting the instrument to a completely separate validation sample that is at risk of recidivism after the model is fully developed is a close approximation of how the model should be expected to perform in the real world, where patterns of recidivism may change slightly over time. This procedure is thus used to test Hypothesis 1.

---

[4] I used MCMCpack; Martin, Quinn, & Park, 2011

[5] Models are stored as propriety data types that can only be manipulated using the functions built into these software packages. Other efforts to capture and manipulate the necessary data resulted in extremely large matrices that could not be stored locally on the computational servers available at CU Boulder through the Institute of Behavioral Science.

The classifications made among this validation sample are also used to assess evidence for Hypothesis 2b. An ideal test of the prediction error induced by greater uncertainty in point estimates (wider confidence intervals) would use the underlying probability distributions of the parameter estimates in both models as the basis for Monte-Carlo simulations that vary the point estimates used to generate the predicted probabilities of offending within the interval. The distributions of incorrect classifications resulting from these simulations would then be an accurate representation of prediction error incurred by the uncertainty implied in the model parameters. These calculations proved to be impossible in the R software for the same reasons listed in footnote 2, because I could not allocate sufficient memory to perform matrix algebra using the very large MCMC objects. As a more preliminary test of this hypothesis, I instead bootstrapped the entire classification and validation process, after model parameters had already been obtained, to observe how the number of incorrect classifications would vary across repeated samples drawn from the validation data. Using 1,000 bootstraps sampled with replacement allows me to construct a distribution of the expected number of incorrect classifications, had there been repeated samples drawn from the validation data rather than a single sample, which can provide a rough estimate of the standard error of these predictions.

**Results**

Table 3.1 displays the log-odds and 95% confidence and credibility intervals of any felony reconviction among Washington State inmates in the final construction sample (Cohorts 6 and 7; those released from corrections between July 1, 2001 and June 30, 2007), using both

Bayesian and Frequentist techniques. These are the results used to construct the risk assessment instrument weights and evaluate its accuracy for predicting reconvictions in the final, validation sample (cohort 8). The results of models predicting offense-specific recidivism for violent, property, and drug crimes are displayed in Tables 3.3-3.5.

***Felony Recidivism***

For Felony recidivism, overall, differences between frequentist and Bayesian model estimates were minimal. Variation in the log-odds of recidivism associated with demographic factors and criminal history appeared to be well within the realms of chance when comparing the two strategies, with only a few notable exceptions. The frequentist models roundly estimated higher recidivism risk associated with the number of prior felony sentences than did Bayesian models, with these differences large enough to lie outside of the 95% Bayesian Credible Interval. In other words, with more information about how previous felony sentences were associated with recidivism in earlier cohorts, the Bayesian models found there to be slightly less risk associated with these earlier convictions than the frequentist models predict. Given the relatively minor differences between the coefficients generated by the two approaches, it is not surprising that there were no differences in the predictive accuracy of the two strategies for felony recidivism, with no notable improvement (or decline) demonstrated by the Bayesian model. At least for felony recidivism, both estimation strategies yield comparable accuracy (ACC=0.73; $F_1$= 0.50; MCC=0.32) and discrimination (AUC=0.74) when the predictions are applied to the validation sample, even though the methods differ in how much weight they give to the number of prior felony sentences.

## Table 3.1: Bayesian and Frequentist Logistic Regression Models Predicting Felony Recidivism, 2001-2007

| | Bayesian Model | | | Frequentist Model | | Difference in CI width |
|---|---|---|---|---|---|---|
| | Log-Odds | 95% BCA | | Log-Odds | 95% CI | (BCA-CI width) |
| | | | **Felony Recidivism** | | | |
| Age (60+ ref.) | | | | | | |
| 50-59 | 0.245 | (0.042 , 0.451) | | 0.303 | (0.092 , 0.513) | -0.012 |
| 40-49 | 0.539 | (0.365 , 0.712) | | 0.601 | (0.403 , 0.799) | -0.049 |
| 30-39 | 0.724 | (0.563 , 0.878) | | 0.787 | (0.590 , 0.983) | -0.078 |
| 20-29 | 0.981 | (0.818 , 1.125) | | 1.037 | (0.841 , 1.233) | -0.086 |
| 18-19 | 1.550 | (1.373 , 1.726) | | 1.619 | (1.417 , 1.821) | -0.052 |
| Male | 0.169 | (0.124 , 0.206) | | 0.169 | (0.126 , 0.213) | -0.006 |
| Juvenile Felonies | | | | | | |
| 1 | 0.478 | (0.428 , 0.532) | | 0.485 | (0.430 , 0.541) | -0.008 |
| 2 | 0.668 | (0.589 , 0.749) | | 0.661 | (0.584 , 0.739) | 0.005 |
| 3+ | 0.975 | (0.913 , 1.029) | | 0.965 | (0.893 , 1.037) | -0.029 |
| Felony Sentence Count | | | | | | |
| 1 | 0.458 | (0.396 , 0.518) | * | 0.534 | (0.457 , 0.612) | -0.033 |
| 2 | 0.934 | (0.841 , 1.013) | * | 1.041 | (0.948 , 1.133) | -0.014 |
| 3 | 1.115 | (1.020 , 1.214) | * | 1.242 | (1.129 , 1.356) | -0.034 |
| 4+ | 1.146 | (1.031 , 1.273) | * | 1.317 | (1.174 , 1.460) | -0.045 |
| Felony Violent Property # | 0.205 | (0.159 , 0.247) | | 0.177 | (0.118 , 0.236) | -0.030 |
| 3+ Non-Domestic Assaults | 0.339 | -(0.010 , 0.646) | | 0.231 | -(0.170 , 0.632) | -0.146 |
| Felony Domestic Assault # | 0.177 | (0.115 , 0.241) | | 0.149 | (0.066 , 0.231) | -0.040 |
| Felony Weapons # | 0.079 | (0.016 , 0.133) | | 0.066 | -(0.004 , 0.135) | -0.023 |
| Felony Property Offenses | | | | | | |
| 1 | 0.230 | (0.189 , 0.267) | | 0.197 | (0.150 , 0.243) | -0.015 |
| 2 | 0.420 | (0.342 , 0.488) | | 0.366 | (0.284 , 0.448) | -0.019 |
| 3 | 0.602 | (0.471 , 0.720) | | 0.530 | (0.409 , 0.651) | 0.006 |
| 4 | 0.701 | (0.542 , 0.848) | | 0.606 | (0.444 , 0.769) | -0.019 |
| 5+ | 0.833 | (0.667 , 0.949) | | 0.725 | (0.571 , 0.880) | -0.027 |
| Felony Drug Offenses | | | | | | |
| 1 | 0.309 | (0.264 , 0.345) | | 0.279 | (0.235 , 0.323) | -0.007 |
| 2 | 0.467 | (0.391 , 0.534) | | 0.402 | (0.323 , 0.481) | -0.014 |
| 3+ | 0.536 | (0.440 , 0.618) | | 0.453 | (0.343 , 0.563) | -0.042 |
| Felony Escapes # | 0.117 | (0.058 , 0.187) | | 0.082 | (0.009 , 0.155) | -0.016 |
| Misd. Non-Domestic Assaults | | | | | | |
| 1 | 0.206 | (0.161 , 0.260) | | 0.204 | (0.154 , 0.253) | 0.001 |
| 2+ | 0.260 | (0.168 , 0.337) | | 0.264 | (0.184 , 0.343) | 0.009 |
| Misd. Domestic Assaults | | | | | | |
| 1 | 0.251 | (0.213 , 0.301) | | 0.247 | (0.194 , 0.300) | -0.018 |
| 2+ | 0.441 | (0.386 , 0.508) | | 0.440 | (0.379 , 0.502) | -0.002 |
| Misd. Weapons # | 0.213 | (0.115 , 0.303) | | 0.232 | (0.149 , 0.315) | 0.022 |
| Misd. Property Offenses | | | | | | |
| 1 | 0.354 | (0.313 , 0.383) | | 0.348 | (0.305 , 0.392) | -0.017 |
| 2 | 0.489 | (0.434 , 0.544) | | 0.482 | (0.418 , 0.547) | -0.020 |
| 3+ | 0.648 | (0.598 , 0.717) | | 0.651 | (0.587 , 0.715) | -0.009 |
| Misd. Drug Offenses | | | | | | |
| 1 | 0.228 | (0.188 , 0.276) | | 0.236 | (0.187 , 0.284) | -0.009 |
| 2+ | 0.273 | (0.203 , 0.347) | | 0.270 | (0.198 , 0.343) | -0.001 |
| Misd. Escapes # | 0.041 | -(0.086 , 0.172) | | 0.039 | -(0.124 , 0.202) | -0.067 |
| Misd. Alcohol Offenses # | 0.041 | (0.008 , 0.081) | | 0.044 | (0.005 , 0.083) | -0.006 |
| Intercept | -3.688 | -(3.894 , -3.500) | | -3.797 | -(4.006 , -3.588) | -0.024 |

* = Frequentist estimate lies outside of the 95% Bayesian Credible Interval

**Table 3.2: Comparison of Bayesian and Frequentist Model Predictive Properties**

| | Any Felony | | Violent Felony | | Property Felony | | Drug Felony | |
|---|---|---|---|---|---|---|---|---|
| | Bayes | Freq | Bayes | Freq | Bayes | Freq | Bayes | Freq |
| AUC | 0.74 | 0.74 | 0.76 | 0.75 | 0.76 | 0.73 | 0.72 | 0.72 |
| F1 | 0.50 | 0.50 | 0.29 | 0.30 | 0.30 | 0.26 | 0.24 | 0.24 |
| MCC | 0.32 | 0.32 | 0.23 | 0.23 | 0.23 | 0.21 | 0.18 | 0.18 |
| ACC | 0.73 | 0.73 | 0.87 | 0.86 | 0.84 | 0.86 | 0.78 | 0.78 |

AUC= Area Under the Receiver Operating Characteristic (ROC) Curve
F1= Harmonic mean of precision and sensitivity
MCC= Matthews Correlation Coefficient between observed and predicted classification
ACC= Accuracy or percentage of cases correctly classified
Note: Results shown were drawn from a separate validation sample of prisoners released after June 30, 2007 (N=7,620).

On the other hand, an examination of the width of the confidence intervals associated with these estimates reveals a strength of the Bayesian approach. As expected, of the 39 total parameters in the model, 35 (90%) produce tighter confidence intervals using the Bayesian vs. frequentist method. While these differences are generally small in magnitude (in most cases, around half the size of the frequentist standard error), any reduction in the uncertainty of model estimates is worth noting because it may increase our confidence in the resulting analytic weights for classification purposes.

### *Offense-Specific Models*

Generally, advantages of the Bayesian approach to modeling recidivism risk were somewhat more pronounced in offense-specific models, though the general pattern of findings mirrors those of the overall felony recidivism models. Bayesian models tended to yield notably different point estimates for a few key predictors of each type of recidivism and produced tighter confidence intervals for the majority of these estimates, compared to frequentist approaches. For two of the three outcomes, it appears that these differences

resulted in slightly better discrimination (AUC; see Table 3.2), which may indicate that the additional information employed by Bayesian models helps to better capture the actual risk associated with some offender characteristics when the outcome of interest is relatively uncommon. This additional power sheds light on some findings of particular theoretical interest: patterns suggest property and drug recidivists tend to engage in more specialized offending behaviors, whereas more general mechanisms appear to underlie the versatile offending of violent recidivists.

*Violent Recidivism*

There are several notable differences in the predicted log-odds of violent recidivism across modeling strategies. Perhaps most importantly, the number of prior felony sentences is associated with greater odds of reconviction in Bayesian than in frequentist models. The frequentist estimates for 2, 3, and 4+ previous felonies all fall just outside of the lower bounds of the accompanying Bayesian credible intervals when carried out to the fifth decimal place (see Table 3.3), suggesting that frequentist methods may somewhat underestimate the risk of prior felony sentences on violent recidivism. Substantial differences also existed for age, with younger ages associated with somewhat lower recidivism risk in the Bayesian models than in frequentist ones, and for prior felony property offenses, where Bayesian estimates suggested a stronger protective effect of 3 and 5+ prior offenses than frequentist models. As expected, estimates in the Bayesian models again had less uncertainty associated with them, resulting in tighter credible intervals for 32 of 39 measures (82%) compared to the corresponding frequentist intervals. Despite variability in the anticipated risk of predictors across models, overall discrimination was

not substantially different, though slightly improved, when classifying offenders using the Bayesian (AUC=0.76) and frequentist-derived (AUC=0.75) estimates. Accuracy was also comparable across each metric, suggesting that the models generally tended to classify violent offenders similarly despite differences in item weights.

## Table 3.3: Bayesian and Frequentist Logistic Regression Models Predicting Violent Felony Recidivism, 2001-2007

| | Bayesian Model | | | Frequentist Model | | Difference in CI width |
|---|---|---|---|---|---|---|
| | Log-Odds | 95% BCA | | Log-Odds | 95% CI | (BCA-CI width) |
| | | | **Violent Felony Recidivism** | | | |
| Age (60+ ref.) | | | | | | |
| 50-59 | 0.277 | (0.010 , 0.442) | * | 0.442 | (0.058 , 0.826) | -0.336 |
| 40-49 | 0.487 | (0.265 , 0.667) | * | 0.667 | (0.305 , 1.029) | -0.322 |
| 30-39 | 0.619 | (0.386 , 0.801) | * | 0.801 | (0.441 , 1.160) | -0.304 |
| 20-29 | 1.041 | (0.822 , 1.223) | * | 1.223 | (0.866 , 1.581) | -0.314 |
| 18-19 | 1.619 | (1.378 , 1.788) | * | 1.788 | (1.423 , 2.153) | -0.320 |
| Male | 1.079 | (0.979 , 1.157) | | 1.094 | (0.988 , 1.200) | -0.033 |
| Juvenile Felonies | | | | | | |
| 1 | 0.388 | (0.298 , 0.465) | | 0.378 | (0.294 , 0.462) | 0.000 |
| 2 | 0.656 | (0.559 , 0.756) | | 0.636 | (0.530 , 0.742) | -0.016 |
| 3+ | 0.894 | (0.823 , 0.979) | | 0.872 | (0.779 , 0.966) | -0.030 |
| Felony Sentence Count | | | | | | |
| 1 | 0.541 | (0.468 , 0.637) | | 0.482 | (0.368 , 0.596) | -0.059 |
| 2 | 1.119 | (1.019 , 1.224) | * | 1.019 | (0.882 , 1.156) | -0.069 |
| 3 | 1.465 | (1.330 , 1.601) | * | 1.330 | (1.162 , 1.499) | -0.066 |
| 4+ | 1.629 | (1.429 , 1.827) | * | 1.429 | (1.215 , 1.642) | -0.028 |
| Felony Violent Property # | 0.210 | (0.117 , 0.254) | | 0.242 | (0.161 , 0.323) | -0.026 |
| 3+ Non-Domestic Assaults | 0.290 | -(0.108 , 0.719) | | 0.409 | -(0.042 , 0.860) | -0.075 |
| Felony Domestic Assault # | 0.353 | (0.266 , 0.444) | | 0.396 | (0.297 , 0.496) | -0.021 |
| Felony Weapons # | 0.129 | (0.029 , 0.203) | | 0.159 | (0.065 , 0.254) | -0.015 |
| Felony Property Offenses | | | | | | |
| 1 | -0.241 | -(0.315 , -0.185) | | -0.191 | -(0.266 , -0.116) | -0.020 |
| 2 | -0.449 | -(0.571 , -0.343) | | -0.349 | -(0.479 , -0.219) | -0.032 |
| 3 | -0.523 | -(0.705 , -0.351) | * | -0.351 | -(0.539 , -0.163) | -0.022 |
| 4 | -0.648 | -(0.941 , -0.427) | | -0.466 | -(0.720 , -0.212) | 0.006 |
| 5+ | -0.528 | -(0.793 , -0.372) | * | -0.372 | -(0.617 , -0.127) | -0.070 |
| Felony Drug Offenses | | | | | | |
| 1 | -0.269 | -(0.332 , -0.204) | | -0.234 | -(0.305 , -0.163) | -0.013 |
| 2 | -0.632 | -(0.771 , -0.550) | | -0.564 | -(0.696 , -0.431) | -0.045 |
| 3+ | -0.870 | -(1.071 , -0.727) | | -0.752 | -(0.936 , -0.567) | -0.024 |
| Felony Escapes # | -0.138 | -(0.228 , -0.061) | | -0.084 | -(0.195 , 0.027) | -0.054 |
| Misd. Non-Domestic Assaults | | | | | | |
| 1 | 0.356 | (0.301 , 0.425) | | 0.345 | (0.272 , 0.418) | -0.022 |
| 2+ | 0.607 | (0.475 , 0.692) | | 0.611 | (0.506 , 0.715) | 0.008 |
| Misd. Domestic Assaults | | | | | | |
| 1 | 0.519 | (0.449 , 0.589) | | 0.523 | (0.446 , 0.600) | -0.015 |
| 2+ | 0.778 | (0.697 , 0.891) | | 0.769 | (0.685 , 0.853) | 0.026 |
| Misd. Weapons # | 0.236 | (0.136 , 0.323) | | 0.221 | (0.103 , 0.340) | -0.050 |
| Misd. Property Offenses | | | | | | |
| 1 | 0.182 | (0.113 , 0.237) | | 0.175 | (0.104 , 0.245) | -0.017 |
| 2 | 0.273 | (0.207 , 0.368) | | 0.265 | (0.162 , 0.367) | -0.043 |
| 3+ | 0.324 | (0.233 , 0.433) | | 0.309 | (0.207 , 0.411) | -0.004 |
| Misd. Drug Offenses | | | | | | |
| 1 | 0.038 | -(0.028 , 0.118) | | 0.038 | -(0.041 , 0.117) | -0.012 |
| 2+ | 0.010 | -(0.096 , 0.115) | | 0.010 | -(0.107 , 0.128) | -0.024 |
| Misd. Escapes # | -0.035 | -(0.255 , 0.245) | | -0.039 | -(0.286 , 0.209) | 0.005 |
| Misd. Alcohol Offenses # | 0.090 | (0.005 , 0.151) | | 0.096 | (0.035 , 0.157) | 0.024 |
| Intercept | -5.565 | -(5.736 , -5.295) | | -5.719 | -(6.101 , -5.338) | -0.321 |

* = Frequentist estimate lies outside of the 95% Bayesian Credible Interval

Overall, both sets of estimates tell a cohesive story about how offender characteristics impact the odds of engaging in future violent crime. Comparing these results across offense types (see Tables 3.4 and 3.5 for models predicting property and drug recidivism) makes this especially clear. Gender, young age, and juvenile and adult sentence counts have strong, positive relationships with violent recidivism, especially when compared to the magnitude of these associations in the drug and (to a much lesser extent) property reconviction models. This pattern of findings confirms expectations derived from Gottfredson and Hirshi's and Moffitt's theories of offending, which together suggest that violent recidivists are most likely of all offender types to engage in a versatile array of criminal behaviors over the life course. If violent recidivists are more likely to be young, male, and have long histories of juvenile and adult offending, then it is probably true that they tend to lack self-control and have committed a wide variety of crimes rather than focusing on just a single type. The strongest evidence of this is perhaps the fact that prior adult and juvenile felony counts are very strongly associated with violent recidivism compared to property and drug recidivism, whereas prior non-violent property and drug felonies are actually associated with reduced odds of receiving a violent reconviction. Even previous violent felonies (violent property, non-domestic assaults, domestic assaults, and weapons charges) contribute only minimally to risk, over and above general counts of felony offenses, highlighting that violent recidivists even appear to reject specializing in violent offenses. Taken together, it is clear that violent recidivism is characterized by behaviors that suggest such offenders are versatile and persistent in their criminal careers, as expected.

*Property and Drug Recidivism*

As with violent recidivism, the primary differences between Bayesian and frequentist estimates of felony property reconvictions concern the amount of elevated risk associated with prior felony and property offenses. For new property offenses, however, the counts of prior felonies in the Bayesian models are associated with much lower log-odds of recidivism than in the frequentist models. The opposite is true of previous felony property offenses, which are expected to increase the odds of a new property conviction within 3 years by a substantially larger amount using Bayesian rather than frequentist methods. In both cases, the frequentist estimates lie outside of the 95% Bayesian credible intervals (see Table 3.4). Like with violent felony recidivism, the Bayesian models again tended to produce tighter credible intervals (for 82% of predictors) than their frequentist counterparts, which helped result in notably higher model discrimination (AUC=0.76 vs. 0.73) and improvement in two of three[6] measures of accuracy ($F_1$= 0.30 vs. 0.26; MCC=0.23 vs. 0.21). These improvements suggest that the higher weights given to prior property felonies and the lower weights given to overall felony offense counts ultimately helped the Bayesian model to better capture the dimensions of risk that predict felony property recidivism, more accurately reflecting the characteristics of these types of recidivists.

[6] While ACC was somewhat lower in the Bayesian model (0.84 vs. 0.86), this is explainable by the selection of a much lower cut point for the predicted probability, as ACC is inherently higher with cut-points closer to 1 when outcomes are uncommon.

## Table 3.4: Bayesian and Frequentist Logistic Regression Models Predicting Property Felony Recidivism, 2001-2007

| | Bayesian Model | | | Frequentist Model | | Difference in CI width |
|---|---|---|---|---|---|---|
| | Log-Odds | 95% BCA | | Log-Odds | 95% CI | (BCA-CI width) |
| | | | **Property Felony Recidivism** | | | |
| Age (60+ ref.) | | | | | | |
| 50-59 | 0.077 | -(0.135 , 0.357) | | 0.116 | -(0.241 , 0.474) | -0.223 |
| 40-49 | 0.438 | (0.200 , 0.646) | | 0.485 | (0.151 , 0.818) | -0.221 |
| 30-39 | 0.749 | (0.512 , 0.984) | | 0.794 | (0.463 , 1.124) | -0.189 |
| 20-29 | 1.013 | (0.773 , 1.245) | | 1.071 | (0.741 , 1.400) | -0.187 |
| 18-19 | 1.531 | (1.270 , 1.752) | | 1.602 | (1.265 , 1.938) | -0.190 |
| Male | -0.108 | -(0.163 , -0.061) | | -0.110 | -(0.171 , -0.049) | -0.019 |
| Juvenile Felonies | | | | | | |
| 1 | 0.330 | (0.261 , 0.406) | | 0.304 | (0.226 , 0.383) | -0.012 |
| 2 | 0.424 | (0.338 , 0.533) | | 0.388 | (0.282 , 0.495) | -0.017 |
| 3+ | 0.579 | (0.479 , 0.653) | | 0.530 | (0.435 , 0.624) | -0.015 |
| Felony Sentence Count | | | | | | |
| 1 | 0.156 | (0.041 , 0.262) | * | 0.311 | (0.183 , 0.438) | -0.034 |
| 2 | 0.317 | (0.181 , 0.476) | * | 0.520 | (0.372 , 0.669) | -0.002 |
| 3 | 0.278 | (0.137 , 0.451) | * | 0.539 | (0.364 , 0.714) | -0.035 |
| 4+ | 0.008 | -(0.154 , 0.189) | * | 0.336 | (0.121 , 0.550) | -0.087 |
| Felony Violent Property # | 0.232 | (0.162 , 0.297) | | 0.180 | (0.095 , 0.265) | -0.035 |
| 3+ Non-Domestic Assaults | 0.192 | -(0.366 , 0.714) | | 0.205 | -(0.389 , 0.799) | -0.109 |
| Felony Domestic Assault # | -0.031 | -(0.208 , 0.124) | | -0.093 | -(0.239 , 0.053) | 0.040 |
| Felony Weapons # | 0.035 | -(0.061 , 0.131) | | 0.009 | -(0.094 , 0.112) | -0.015 |
| Felony Property Offenses | | | | | | |
| 1 | 0.844 | (0.771 , 0.904) | | 0.772 | (0.702 , 0.841) | -0.006 |
| 2 | 1.361 | (1.251 , 1.477) | * | 1.248 | (1.136 , 1.361) | 0.000 |
| 3 | 1.777 | (1.655 , 1.893) | * | 1.622 | (1.466 , 1.778) | -0.073 |
| 4 | 2.020 | (1.815 , 2.200) | | 1.864 | (1.665 , 2.063) | -0.013 |
| 5+ | 2.208 | (2.048 , 2.395) | * | 2.040 | (1.853 , 2.228) | -0.028 |
| Felony Drug Offenses | | | | | | |
| 1 | 0.285 | (0.229 , 0.340) | | 0.243 | (0.177 , 0.309) | -0.021 |
| 2 | 0.451 | (0.343 , 0.570) | | 0.371 | (0.257 , 0.485) | 0.000 |
| 3+ | 0.352 | (0.218 , 0.483) | | 0.233 | (0.080 , 0.386) | -0.041 |
| Felony Escapes # | 0.186 | (0.093 , 0.281) | | 0.134 | (0.034 , 0.235) | -0.013 |
| Misd. Non-Domestic Assaults | | | | | | |
| 1 | 0.016 | -(0.049 , 0.086) | | 0.025 | -(0.049 , 0.098) | -0.013 |
| 2+ | -0.097 | -(0.232 , 0.027) | | -0.098 | -(0.217 , 0.021) | 0.021 |
| Misd. Domestic Assaults | | | | | | |
| 1 | 0.084 | (0.005 , 0.159) | | 0.090 | (0.011 , 0.168) | -0.003 |
| 2+ | 0.120 | (0.035 , 0.227) | | 0.119 | (0.026 , 0.213) | 0.006 |
| Misd. Weapons # | 0.124 | (0.014 , 0.230) | | 0.133 | (0.017 , 0.249) | -0.016 |
| Misd. Property Offenses | | | | | | |
| 1 | 0.458 | (0.396 , 0.506) | | 0.440 | (0.376 , 0.503) | -0.017 |
| 2 | 0.652 | (0.567 , 0.729) | | 0.639 | (0.552 , 0.726) | -0.011 |
| 3+ | 0.828 | (0.762 , 0.895) | | 0.811 | (0.728 , 0.894) | -0.033 |
| Misd. Drug Offenses | | | | | | |
| 1 | 0.196 | (0.127 , 0.259) | | 0.197 | (0.128 , 0.266) | -0.005 |
| 2+ | 0.180 | (0.083 , 0.268) | | 0.181 | (0.081 , 0.280) | -0.014 |
| Misd. Escapes # | 0.061 | -(0.165 , 0.277) | | 0.068 | -(0.144 , 0.280) | 0.018 |
| Misd. Alcohol Offenses # | -0.003 | -(0.059 , 0.059) | | 0.001 | -(0.058 , 0.059) | 0.002 |
| Intercept | -4.472 | -(4.700 , -4.239) | | -4.611 | -(4.960 , -4.261) | -0.238 |

* = Frequentist estimate lies outside of the 95% Bayesian Credible Interval

## Table 3.5: Bayesian and Frequentist Logistic Regression Models Predicting Drug Felony Recidivism, 2001-2007

| | Bayesian Model | | | Frequentist Model | | Difference in CI width |
|---|---|---|---|---|---|---|
| | Log-Odds | 95% BCA | | Log-Odds | 95% CI | (BCA-CI width) |
| | | | **Violent Felony Recidivism** | | | |
| Age (60+ ref.) | | | | | | |
| 50-59 | 0.262 | (0.034 , 0.465) | | 0.355 | (0.004 , 0.705) | -0.270 |
| 40-49 | 0.495 | (0.252 , 0.673) | | 0.590 | (0.256 , 0.923) | -0.247 |
| 30-39 | 0.531 | (0.295 , 0.725) | | 0.621 | (0.289 , 0.954) | -0.235 |
| 20-29 | 0.465 | (0.240 , 0.654) | | 0.551 | (0.219 , 0.883) | -0.251 |
| 18-19 | 0.762 | (0.534 , 0.969) | | 0.867 | (0.521 , 1.214) | -0.259 |
| Male | -0.097 | -(0.142 , -0.050) | | -0.131 | -(0.198 , -0.064) | -0.042 |
| Juvenile Felonies | | | | | | |
| 1 | 0.310 | (0.229 , 0.391) | | 0.328 | (0.233 , 0.423) | -0.028 |
| 2 | 0.401 | (0.284 , 0.511) | | 0.393 | (0.262 , 0.523) | -0.035 |
| 3+ | 0.387 | (0.290 , 0.489) | | 0.407 | (0.284 , 0.529) | -0.046 |
| Felony Sentence Count | | | | | | |
| 1 | 0.221 | (0.084 , 0.354) | * | 0.528 | (0.357 , 0.699) | -0.072 |
| 2 | 0.419 | (0.288 , 0.547) | * | 0.809 | (0.619 , 1.000) | -0.122 |
| 3 | 0.268 | (0.131 , 0.426) | * | 0.731 | (0.514 , 0.949) | -0.140 |
| 4+ | 0.197 | (0.034 , 0.393) | * | 0.774 | (0.519 , 1.028) | -0.150 |
| Felony Violent Property # | 0.201 | (0.100 , 0.285) | | 0.129 | (0.026 , 0.232) | -0.021 |
| 3+ Non-Domestic Assaults | 0.069 | -(0.623 , 0.880) | | -0.130 | -(0.923 , 0.663) | -0.083 |
| Felony Domestic Assault # | -0.113 | -(0.256 , 0.027) | | -0.209 | -(0.388 , -0.029) | -0.075 |
| Felony Weapons # | 0.238 | (0.116 , 0.331) | | 0.164 | (0.056 , 0.273) | -0.002 |
| Felony Property Offenses | | | | | | |
| 1 | 0.153 | (0.086 , 0.208) | * | 0.068 | -(0.013 , 0.148) | -0.039 |
| 2 | 0.263 | (0.147 , 0.366) | * | 0.116 | -(0.019 , 0.252) | -0.052 |
| 3 | 0.204 | (0.030 , 0.370) | * | 0.001 | -(0.193 , 0.195) | -0.048 |
| 4 | 0.163 | -(0.050 , 0.337) | | -0.044 | -(0.301 , 0.213) | -0.128 |
| 5+ | 0.023 | -(0.151 , 0.240) | * | -0.171 | -(0.412 , 0.071) | -0.091 |
| Felony Drug Offenses | | | | | | |
| 1 | 1.208 | (1.135 , 1.268) | * | 1.083 | (1.006 , 1.160) | -0.021 |
| 2 | 1.762 | (1.654 , 1.864) | * | 1.562 | (1.441 , 1.683) | -0.031 |
| 3+ | 2.197 | (2.053 , 2.319) | * | 1.926 | (1.765 , 2.087) | -0.055 |
| Felony Escapes # | 0.239 | (0.145 , 0.340) | | 0.169 | (0.057 , 0.281) | -0.029 |
| Misd. Non-Domestic Assaults | | | | | | |
| 1 | 0.111 | (0.036 , 0.182) | | 0.118 | (0.036 , 0.200) | -0.018 |
| 2+ | 0.060 | -(0.063 , 0.158) | | 0.046 | -(0.084 , 0.176) | -0.040 |
| Misd. Domestic Assaults | | | | | | |
| 1 | -0.019 | -(0.100 , 0.068) | | -0.032 | -(0.122 , 0.058) | -0.012 |
| 2+ | 0.081 | (0.001 , 0.167) | | 0.089 | -(0.014 , 0.192) | -0.040 |
| Misd. Weapons # | 0.075 | -(0.024 , 0.168) | | 0.057 | -(0.074 , 0.188) | -0.069 |
| Misd. Property Offenses | | | | | | |
| 1 | 0.235 | (0.164 , 0.296) | | 0.228 | (0.156 , 0.301) | -0.012 |
| 2 | 0.176 | (0.089 , 0.277) | | 0.148 | (0.042 , 0.255) | -0.025 |
| 3+ | 0.294 | (0.206 , 0.381) | | 0.254 | (0.153 , 0.354) | -0.026 |
| Misd. Drug Offenses | | | | | | |
| 1 | 0.330 | (0.242 , 0.417) | | 0.320 | (0.247 , 0.394) | 0.028 |
| 2+ | 0.407 | (0.308 , 0.504) | | 0.410 | (0.309 , 0.511) | -0.007 |
| Misd. Escapes # | -0.136 | -(0.369 , 0.079) | | -0.115 | -(0.367 , 0.138) | -0.057 |
| Misd. Alcohol Offenses # | 0.013 | -(0.037 , 0.066) | | 0.029 | -(0.035 , 0.094) | -0.025 |
| Intercept | -4.590 | -(4.834 , -4.381) | * | -4.851 | -(5.217 , -4.484) | -0.281 |

* = Frequentist estimate lies outside of the 95% Bayesian Credible Interval

The patterns of findings for drug recidivism are very similar. General felony sentence counts contribute much less to recidivism risk in the Bayesian models compared to frequentist models, whereas the number of prior drug felonies is associated with greater risk in the Bayesian models (see Table 3.5). Interestingly, the number of prior felony property offenses, which essentially had a null effect on drug recidivism in the frequentist model, contributes to a slightly elevated risk of reconviction in the Bayesian model. The Bayesian approach led to almost universally tighter confidence intervals, with only one predictor demonstrating lower standard errors in the frequentist model. Given the numerous differences in point estimates and uncertainty demonstrated between the two modeling strategies, it is surprising that discrimination and accuracy did not differ. Both models examining drug recidivism displayed relatively lower predictive potential compared to those predicting any felony, violent felony, and property felony recidivism, especially for the more stringent measures of accuracy ($F_1$=0.24; MCC =0.18). Nonetheless, these results are comparable to those reported elsewhere for logistic regression, neural network, and random forest-derived assessments of drug felony recidivism (Hamilton et al., 2015). With no notable improvements in the ability to predict drug reconvictions using the Bayesian model, it is difficult to know whether the different item weights and reduced uncertainty of estimates actually better captured the characteristics of those who recidivated and did not for drug offenses.

Regardless, the estimates of both the Bayesian and frequentist models indicate a pattern of offending among drug and property recidivists that differs considerably from that of violent recidivists. Extensive histories of prior property and drug felonies are strong predictors of who will be reconvicted of these same types of offense within 3-years,

whereas general counts of prior felonies contribute relatively little to our ability to identify either type of recidivist. This suggests that property and drug recidivists are best characterized as specialists. Compared with violent recidivists, past offending behaviors of the same type are strongly associated with future criminality among these offenders, implying that generalist theories of crime may not explain the deviant behavior of this group as well as the proponents of these theories have assumed. In fact, in models predicting violent recidivism, established histories of property and drug offending actually reduce the odds of reconviction.

### *Bootstrapped Estimates of Prediction Error*

Table 3.6 and Appendix Figures A.1-A.4 summarize the results of my bootstrapped analysis of prediction error among the resampled validation data. As expected, Bayesian models generally reduced the mean number, bias, and standard error of incorrect classifications across recidivism outcomes. While differences were insubstantial for the outcomes that displayed similar accuracy and discrimination across models (felony recidivism and drug recidivism), these analyses highlight the gains made by the Bayesian models in predicting property and violent recidivism. The differences in the mean number of incorrect classifications for violent and especially property recidivism are much larger than appears would be due to chance. It is also worth noting that, with the exception of felony recidivism, the Bayesian models yielded smaller standard errors and bias in these predictions than frequentist models did. Overall, these results indicate a preference for the Bayesian models in reducing prediction error, and these improvements are particularly notable for the instruments predicting violent and property felony recidivism.

**Table 3.6: Comparison of Bayesian and Frequentist Model Bootstrapped Prediction Error**

| | Any Felony | | Violent Felony | | Property Felony | | Drug Felony | |
|---|---|---|---|---|---|---|---|---|
| | Bayes | Freq | Bayes | Freq | Bayes | Freq | Bayes | Freq |
| # Incorrectly Classified | 2050 | 2053 | 993 | 1375 | 1257 | 2089 | 1875 | 1875 |
| Boot Std. Err. | 40.22 | 38.67 | 28.73 | 34.87 | 32.93 | 38.75 | 35.82 | 38.45 |
| Bias | -1.56 | -0.82 | 0.01 | -0.50 | 0.32 | -0.98 | 0.08 | -1.17 |

Note: Used 1000 bootstraps to compute classification errors in validation sample

## Discussion

Across several of the most common metrics used to compare and validate risk assessment strategies, I found that Bayesian logistic regression models performed as well or somewhat better than more common frequentist models, with all model AUCs demonstrating large effect sizes (Rice and Harris, 2005). Even when Bayesian methods did not noticeably improve accuracy or discrimination over a nearly identical frequentist model, as was the case with overall felony and drug felony recidivism, they did yield almost universally tighter confidence intervals, increasing the precision of item weights and reducing the uncertainty contained in these estimates. Findings were even more favorable for offense-specific outcomes. Here, Bayesian methods often yielded different item weights for critical predictors that reduced standard and prediction error and improved the accuracy of classifications compared to frequentist models with similar specifications. Specifically, I found that the Bayesian model placed greater emphasis on the total number of prior felony sentences for violent recidivists, which helped reduce the rate of incorrect

classification for this critical outcome. For property recidivism, this estimation strategy better highlighted the importance of prior same-offense-type patterns, which improved classification accuracy and resulted in a notable increase in discrimination (AUC). In out-of-sample predictions, there were substantial reductions in incorrect classifications made using the Bayesian violent and property recidivism models. However, in general felony and drug recidivism models, performance did not differ.

These findings demonstrate that Bayesian methods are not only a valid statistical tool for predicting recidivism, but that they hold the potential to be especially useful for real world application, where they may reduce classification errors in practical settings that involve out-of-sample prediction. Perhaps most intriguing, though not explored here, the statistical framework used to construct these models allows for expert judgment to be easily integrated into the empirical relationships that typically determine risk estimates (Fenton & Neil, 2011), which could prove helpful for practitioners interested in making use of qualitative knowledge about the nature of criminogenic risks in their local jurisdiction. I also suggest that these results reinforce the need for risk assessment instruments to include different item weights for each type of recidivism. For violent felony recidivism, long histories of drug or non-violent property offenses were actually protective against future offending, but these reductions in risk were offset by increases in the risk of same-offense-type recidivism in my models. This could be useful for practitioners, who can more readily identify those offenders who pose the greatest societal threat (violent criminals) while targeting specific interventions aimed toward encouraging desistance for those who exhibit repeated property and drug offending.

Aside from these important, practical findings, this research also weighs in on the debate over whether offenders are versatile, committing many different types of crime over the life course, or whether they are specialists, preferring a single type of crime throughout their criminal careers (Deslauriers-Varin, Lussier, & Tzoumakis, 2016). Deriving certain expectations from two major theories that speak to this topic, Gottfredson and Hirschi's general theory of crime and Moffitt's taxonomy of offending, I found that characteristics indicative of versatility were associated with violent recidivism, whereas property and drug recidivists appeared to be much more specialized, though there was some comorbidity of property crime among drug offenders (Pernanen et al., 2002). These findings speak to potential shortcomings of any generalist theories of crime that propose unified explanations for all types of offending. While such theories could be used to help understand the varied offending behaviors of violent recidivists in this sample, the presence of specialization among property and drug offenders requires an explanation that sheds light on how certain individuals develop a proclivity for some types of crime over others.

Gottfredson and Hirshi's and Moffitt's theories are valuable starting points, but more work needs to be done to integrate these and other promising perspectives (perhaps cultural deviance, state dependence, and other life course theories) if we hope to better understand the process of specialization. Future developers of risk assessment instruments should keep this in mind, as their work is not just well-poised to make valuable practical contributions, but important theoretical ones too. I encourage researchers to make use of the types of data often employed for risk assessment construction and validation, especially when such data is longitudinal and contains information about both static and

dynamic offender characteristics, to weigh in on this claim. Qualitative research of offenders with long histories of drug and property offenses would also be very useful to uncover the mechanisms leading to this kind of offense specialization, which may help answer the question of why these types of offenders tend to not engage in violent offenses.

This study is not without limitations, however. While this is one of the largest samples ever used to construct and validate a new risk assessment strategy (Hamilton et al., 2015), the validation sample was relatively small (N=7,620) compared to the size of the construction sample (N= 94,192) and a non-optimal holdout method was used for validation rather than preferred k-fold cross validation methods. However, the size of the sample should provide more than adequate power for using this method (Pang & Jung, 2013), and given the temporal ordering of the data, this strategy closely approximates real world application where instruments are developed and then applied to later cohorts of offenders. Another limitation is the lack of dynamic predictors, which some suggest can improve the predictive power of risk assessment strategies (e.g. Andrews, Bonta, & Wormith, 2006; Bonta, 2002; Hamilton, 2016). This additional set of predictors would be useful for testing other theoretically relevant mechanisms, though there may be issues with the construct validity and reliability of such measures (Cording, Christofferson, & Grace, 2016). Regardless, it would be informative to test Bayesian models using this data, as the benefits demonstrated by the method may be amplified with an even larger array of inputs. This may be particularly true of dynamic measures, which experts expect will have theoretically relevant associations with recidivism. These expectations could be directly nested within models and tested by quantifying them as priors and comparing changes to posterior predictive properties. Such tests are beyond the scope of the data used in the

present study, but would be informative directions for future research with data from Washington State's Offender Needs Guide—for Recidivism (ONG).

*Conclusion*

Despite recent advocating for greater use of Bayesian statistical methods in the area of criminal justice (Fenton, Neil, & Berger, 2016; Philipse, 2015), and the success of Bayesian models in the areas of genetics (Fenton and Neil, 2012; Ogino & Wilson, 2004) and cancer (Newcombe et al., 2012) risk projection, this is the first attempt to develop and validate a Bayesian risk assessment instrument for predicting recidivism. While results were promising, especially for predicting violent and property crime, and theoretically informative, with regards to crime specialization and versatility, model performance of the BARR was about on par with the best machine learning and sophisticated logistic regression-based strategies reported elsewhere (Barnoski & Drake, 2007; Berk & Bleich, 2013; Duwe & Kim, 2016; Hamilton et al., 2016; Liu et al., 2011).

This tendency for a variety of estimation strategies, across several different samples and sets of predictors, to coalesce around similar maximum values of AUC (~0.75) seems to suggest that there is a ceiling to how well these types of models can predict felony recidivism (Liu et al., 2011). Because of the disparate approaches that have resulted in similar levels of predictive accuracy across samples, it seems that the remaining error in classification systems may come from sources other than modeling strategies and choices of predictors. In their classic treatise on classification, *Sorting Things Out,* Bowker and Star (1999) challenge the basic idea that classification systems like those produced by risk assessment instruments can ever perfectly succeed in distinguishing between different

types of people, because the definitions of categories such as "criminal" and "recidivist" rely on some degree of human interpretation. Even if researchers identify common definitions, like using reconvictions within 3 years of release as an agreed-upon marker of recidivism, there is a great deal of error in the categorization and measurement of the concept that goes largely unacknowledged.

For example, those who are ultimately classified as reconvicted felons in risk assessment instruments need to meet a variety of conditions to qualify. First, the offender needs to engage in a criminal action that will result in (at a minimum) their arrest. This, of course, requires that their action is considered a legally punishable offense of a sufficient magnitude to warrant an arrest, but also that a police officer capable of making the arrest is made aware of the crime. Every aspect of this is process is subject to human interpretation, leaving open the possibility for a great deal of error in the measurement of recidivism: some actions are merely deviant in one place, but criminal in another; the police officer may exercise discretion in whether to make an arrest or which charges to press based on their interpretation of a situation; and relatively few crimes that victims report are ever actually reported to the police. This is just the beginning—on top of all of these potential sources of error, a great deal of judgment goes into the odds of offender reconviction, which results in even greater uncertainty about how the recidivist is ultimately defined.

While Bowker and Star's point is at least partially philosophical, risk assessment scholars have largely ignored the implications for their research. The present study is simply another example that there is likely a ceiling to how well mathematical models can be expected to predict something like recidivism, because human behavior is inherently uncertain and there is error in the categorizations of characteristics on both sides of the

prediction equation. The Bayesian approach does not help in reducing these sources of categorization error, but the resulting estimates *are* inherently more explicit about this uncertainty. Bayesian models almost universally led to more precise risk estimates in this sample, which reduced some metrics of prediction error. This helps to establish Bayesian risk assessment tools as viable alternatives to ones derived from frequentist logistic regression and machine learning techniques. Their interpretability and ability to adapt to data patterns over time makes them a potentially good compromise between these other, more thoroughly researched strategies. Future researchers should further investigate the potential for Bayesian methods of risk classification.

Chapter 4 – EMPIRICAL ANALYSIS:

Consequences of Intersectional Risk Assessment

## Introduction

Feminist and critical race scholars have long suggested that gender and race fundamentally structure how individuals interact with the criminal legal system. Many of these scholarly efforts have adopted cultural lenses in attempts to explain why men engage in higher rates of criminal behavior (Doude, 2014; Hayslett-McCall & Bernard, 2002; Messerschmidt, 1993; Steffensmeier & Allan, 1996; Whitehead, 2005), or why rates of correctional supervision are so high for black men (Alexander, 2010; Bureau of Justice Statistics, 2014; 2016), but these differential rates of offending and control are only part of the story. There exists a complex feedback loop in the U.S. criminal justice system wherein corrections have become an important setting of the creation and reinforcement of racial and gender boundaries (Alexander, 2010; Omi & Winant, 1994; Richie, 2012; Walker, 2016), which both instigates and responds to the various raced, classed, and gendered pathways that men and women take toward and away from criminality (Belknap, 2007; Brennan et al., 2012; Chesney-Lind, 1997; Daly, 1992, 1994; Potter, 2015; Rios, 2011; Salisbury and Van Voorhis, 2009).

Risk assessments likely play an important role in this process. While these tools usually pay little attention to gender, race, or ethnicity (Maurutto & Hannah-Moffat, 2005), there exists a growing body of evidence that men, women, and people of color have unique sets of risks and needs (e.g. Belknap & Holsinger, 2006; Blanchette & Brown, 2006; Gavazzi,

Yarcheck & Lim, 2006; Hudson & Bramhall, 2005). Furthermore, practitioners tend to manage offenders in a gendered and racialized manner. In perhaps the most obvious examples, men and women are confined to separate facilities and inmates are frequently forced to identify with and be segregated by racial or ethnic group in carceral settings (Brennan et al., 2004; Goodman, 2008; Petersilia, 2006). Yet, risk assessment, which structures action at nearly every step in the criminal legal process, almost always proceeds in a gender and race-neutral manner despite some evidence that certain instruments like the LSI-R perform poorly for women and people of color (Desmaris & Singh, 2013).

While researchers have decried risk assessment strategies for being unresponsive to gender-specific risk factors (Brennan et al., 2012) and relying heavily on measures that may act as proxies for race (Harcourt, 2015), few have integrated the two critiques and none have constructed or evaluated risk assessment tools from an intersectional perspective. Intersectionality (Crenshaw, 1989) dictates that the selection of recidivism risk factors and the amount of weight devoted to each should vary considerably across both race *and* gender, because the lived experiences of offenders are determined so strongly by both characteristics. If so, instruments tailored to each population would perform better than a general tool that incorporates neither gender nor race/ethnicity-specific items, but these types of adaptations have not yet been tested. That is the object of this chapter, which explores whether race and gender-specific predictors constructs can be used to increase the predictive validity of separate Bayesian models for black and white men and women released from Washington State correctional supervision.

**Racial Formation and Pathways to Offending**

The U.S. criminal justice system may well be the paramount institutional 'racial project' of the 21st century. A key element of Omi and Winant's racial formation theory, racial projects refer to situations that "connect what race *means* in a particular discursive practice and the ways in which both social structures and everyday experiences are racially *organized,* based upon that meaning" (pp. 56; 1994). The proliferation of black criminality in various forms of media (Gilliam et al., 1996; Oliver, 2003; Smiley & Fakunle, 2016; Welch, 2007) and the perceived inmate "preference" for segregation of blacks, whites, and Latinos in many jails and prisons (Goodman, 2008; Walker, 2016), for instance, impact and structure how the justice system treats people of color.

When media-enhanced perceptions of African Americans as dangerous impact our cultural understanding of blackness, the ability for practitioners to disaggregate race from risk may be compromised. Researchers have demonstrated that non-white race and darker skin color increase the odds of harsher criminal justice outcomes at each stage of the legal process (Bennett & Plaut, 2018), influencing everything from the odds of arrest (Finkeldey, 2014) and detention/bail amounts (Demuth, 2003) to the perceived risk of violent recidivism in a recent death penalty case overturned by the U.S. Supreme Court (Buck v. Davis, 2016). Likewise, racial segregation in prisons is rationalized by claims of violence reduction and inmate preference, but the process often involves coercing inmates into making "appropriate" binary classifications or having correctional officers make their own racial determination without further input from the inmate (Goodman, 2008). If anything, racial segregation appears to engender more, not less, inmate-on-inmate violence (Walker, 2016). The de-segregated Texas prison system (Trulson et al., 2008) illustrates this best: in

the 10 years since desegregation, only 45% of all violent incidents occurred inter-racially, and of these, only 20% took place in the context of desegregated two-man cells (Trulson & Marquart, 2002). In all of these cases, the legal system engages in "race making" (Wacquant, 2001) that produces poorer outcomes for people of color and fundamentally restructures the nature of criminogenic risk among these populations.

These and other instances of race making do not, however, impact men and women of color in the same ways. The above examples come from studies either mostly or entirely composed of men, and probably best summarize how racial formation structures black men's interactions with the criminal justice system. For black women and other men (e.g. Latino boys in Rios, 2011) and women of color, gender is just as important a factor as race in shaping cultural perceptions of criminality and the pathways that lead offenders into and out of criminal lifestyles (Potter, 2015). Richie (2012), for instance, gives several examples of how black women's experiences differ from those of both white women and black men in ways that make them particularly susceptible to victimization, which may put them at risk of psychological disorders, homelessness, intimate partner violence, and drug use. Overall, the factors that precipitate offending and the needs that encourage desistance among female criminals vary considerably from men's (Crenshaw, 1991; Finzen, 2005; Johnson, 1995; Vaughn et al., 2007), but the theories, risks, and needs factors that inform most risk assessment procedures have typically been determined using majority male samples. This strategy fails to address the more complex realities of female risk (Alarid, Burton, & Cullen, 2000; Brennan et al., 2012; Dehart, 2008; Salisbury & Van Voorhis, 2009), though it does capture traditional antisocial pathways that are more general and appear to cut across gender (Jones et al., 2014; McCoy & Miller, 2013).

*Intersectionality and Crime*

The researchers investigating these unique pathways to crime among female offenders often recognize the salience of race as a fundamental determining factor (Bell, 2013; Potter, 2015; Richie, 2012). Thus, criminologists have adopted feminist theories of intersectionality as a way of helping to explain these findings (Belknap 2007; Burgess-Proctor, 2006; Daly & Tonry, 1997; Katz, 2000). Intersectional perspectives suggest that it is impossible to disaggregate gender and race (and class, age, sexual orientation, and other marginalizing statuses) because they do not act in isolation in social life (McCall 2005). The overlapping systems of oppression comprised by gender inequality and institutional racism interact in ways that play out differently than either social classifier, alone (Hill-Collins, 1998a).

This is especially true in criminal justice settings (Hill-Collins 1998b). Beyond what we know about offending (e.g. Baskin-Sommers et al., 2013; Brown, 2006; Haynie & Armstrong, 2006; Simpson, 1991), there is evidence of interactive impacts of gender and race in studies of probation failure (Steinmetz & Henderson, 2015), pretrial outcomes (Freiburger & Hilinski, 2010) and in sentencing, where harsher racial disparities tend to exist between men but are more attenuated between women (Steen, Engen, & Gainey, 2005; Steffensmeier & Demuth, 2006). The intersections of age, race, and gender also appear particularly important. A body of work (Doerner & Demuth, 2010; Kramer & Ulmer, 2009; Steffensmeier et al., 1998; Warren et al., 2012) demonstrates that race and gender impacts on sentencing are strongly graded by age, with substantial disparities between young black and Latino males under the age of 30 and similar women and white men, but

few racial or gender differences among older groups. Particularly noteworthy is the added leniency afforded young teenage-adult women and white men, whose odds of harsh sentencing are hugely disparate to men of color in the same age range; in fact, these interactive effects appear larger than the main effects of race or gender on sentencing severity (Steffensmeier, Davis, & Ulmer, 2017).

While these findings confirm the importance of intersectionality in shaping criminal justice outcomes, very few scholars have applied this same lens to studying the accuracy of risk assessment from an intersectional perspective, and none have explored whether predictors that respond to differences in risk factors by both gender *and* race can be used to improve the accuracy of risk assessment models. Researchers have investigated prediction error by gender and race, separately, sometimes even in the same studies (e.g. Desmarais & Singh, 2013), but they have not compared how well these instruments perform for white men and women and men and women of color, more specifically. Without investigating these patterns of recidivism, the possibility remains that disparities in sentencing and risk classification merely reflect reactions to heightened risk among certain demographic groups. For example, young black men may simply be sentenced more harshly because they are, all else being equal, more likely than others to recidivate[7]. A gender and race-specific instrument composed of item weights from completely different models for each subgroup could result in better classifications and help adjust for biases that are introduced into other assessments that ignore how risks differ across these social categories. On the other hand, separate models for each race/gender combination may

[7] This of course runs the risk of becoming a self-fulfilling prophecy, because studies show that harsher punishment itself is predictive of recidivism (Mears, Cochran, & Bales, 2012), especially among men (Mitchell et al., 2017)

simply reflect the empirical realities of higher base rates of reoffending among some groups, resulting in systematic overclassification of risk among groups like black men. Regardless, given the scholarly attention devoted to incorporating gender-sensitive items on risk assessment instruments and the body of research that suggests pathways to offending are both raced *and* gendered, it is worth exploring whether methods arising from this intersectional framework hold any promise for improving risk assessment.

*Potential issues with fitting race and gender-specific models*

There are at least three reasons why researchers have not previously attempted to develop and validate risk assessment instruments that incorporate gender *and* race-specific estimates. First, sample sizes for women in these development models have tended to be relatively small, and further disaggregation by race would likely result in small cell counts in many regression models, particularly those with many predictors. The WADOC data is ideal for this purpose because it is among the largest sources of data to ever be used for risk assessment development and validation (Hamilton et al., 2016), and the number of predictors is relatively small, opening up more degrees of freedom for model estimation. While Washington State's correctional population contains a very high proportion of white offenders (82%), there is still a sufficient percentage who are black (13%) to enable subgroup analysis by both gender and race. Unfortunately, the data do not identify respondent ethnicity, preventing the additional analysis of a possible Latino subgroup, and Asians (3%) and Native Americans (3%), while present, make up a very small share of the correctional population.

Second, traditional frequentist methods may be underpowered for developing and validating gender and race-specific models. This is especially true if the developer hopes to only use the most temporally relevant (i.e. most recent) data for the purposes of prediction. While such an analysis may be possible with the large sample included in the WADOC data, Bayesian approaches with precise and well-defined priors have a power advantage over comparable frequentist models (Loh et al., 2015; Stegle et al., 2010), which should yield estimates with less uncertainty and better predictive potential. It is therefore unsurprising that these types of models have not been tested before, because Chapter 3 provided the first test of a Bayesian instrument for assessing recidivism risk.

Third, and perhaps most important, the application of any form of race-specific instrument will be problematic in a real-world setting. Developers of risk assessments have traditionally been motivated by producing tools for practical use rather than academic curiosity, and race-specific instruments run the risk of failing to exhibit sufficient external validity to attract practitioner attention. If, for example, prior felony assault convictions are weighted more heavily for men than women, practitioners are still likely to view the instrument as useful because this reflects well-documented realities about gender differentials in crime (e.g. Choy et al., 2017; Heimer, Lauritsen, & Lynch, 2009; Schwartz et al., 2009), the greater burden is not placed on the more marginalized individual, and because offenders are already processed in a gendered manner. With race, the same scenario could be seen as discriminatory (if black offenders are rated higher risk for certain characteristics) or overly lenient (if risk factors are given less serious item weights despite higher rates of some crimes among black offenders). Relatedly, properly enacting a race-specific method of assessment would require practitioners and clients to agree on rigid

racial classifications, which may be contentious when the decision could impact an offenders' risk level. Ultimately this process may just further engrain risk assessment procedures as a racial project (Omi & Winant, 1994) that has very real consequences for marginalized groups of people who are already overrepresented in the criminal justice system. Regardless, given the attention garnered by gender-specific assessments and intersectional critiques of risk assessment, it is both theoretically and practically informative to attempt to model risk this way to determine what benefits, if any, there may be to intersectional actuarial assessment strategies.

*Expectations Regarding Risk Assessment and Race/Gender*

It is reasonable to assume that risk assessment classifications will tend to reflect race and gender differences in recidivism. Indeed, research on the Wisconsin risk needs assessment, at least, shows that black offenders have the greatest likelihood of being classified as high risk (Eisenburg, Bryle, & Fabelo, 2009; Henderson, 2006; Henderson et al., 2007). However, the research on whether these more severe classifications are warranted or not is unclear. Some have found minority offenders to be systematically overclassified on these instruments (with higher rates of false positives; Rembert et al., 2014; Whiteacre, 2006), while several others have found no evidence of bias that is not warranted by higher rates of enacted recidivism among black offenders (e.g. Flores, Bechtel, & Lowenkamp, 2016). Likely, these findings differ by how instruments are constructed, and with what populations they are normed and validated. Those instruments that only use static predictors, for instance, appear to be the worst offenders in terms of

racial overclassification (Skeem & Lowenkamp, 2016) because criminal history is correlated with race (Harcourt, 2015).

As such, an instrument constructed and validated using the static risk WADOC data, like the BARR that I developed in Chapter 3, should exhibit some degree of variation by race and gender. Variation in which criminal history factors predict recidivism and to what degree age predicts recidivism would lend support for intersectional theories that suggest black and white men and women engage in different criminal pathways. The direction and manner of these differences are more difficult to predict, however, given the complex ways that intersectionality is expected to play out in terms of offending and desistance from crime. I would also expect that model intercepts will vary by race and gender in response to different rates of offending and reconviction by demographic group, with black men experiencing the largest baseline risk of reoffending. If this is the case, it suggests that gender and race-sensitive assessments (or at least those that include only static criminal history measures) would have the opposite effect of what intersectional theorists would hope for. Rather than resulting in more equitable risk classifications through better recognition of their unique risks, it may operate like other aspects of the justice system that ensure harsher punishment of black men, contributing to the racial project of mass incarceration.

Comparing race and gender-specific risk assessments to a previously validated gender and race-neutral instrument that was developed using the same data and statistical techniques will also be informative. First, examining the predictive validity of new instruments across racial and gender subgroups should be common practice, with some arguing that predictive parity should be an explicit goal of assessment systems (Angwin et

al., 2016; Citron, 2016). Because the BARR exhibited good performance in Chapter 3, I expect that it will continue to perform well among all racial and gender subgroups in the present study. However, the use of race and gender-specific item weights and variable selection processes in the group-specific models should result in better predictive validity for those groups who differ the most from the largely white and male sample used to construct and validate the BARR. If intersectional theories are correct that risk factors among black women are most different from white men, then I would expect the black women's models to outperform the general instrument for predicting recidivism risk among that group. Altogether, the expectations derived from intersectional theory and a review of the prior literature yield three testable hypotheses.

**Hypothesis 1:** The estimated risks associated with demographics and criminal history will vary by both race and gender. Even the most well established predictors of recidivism like prior felony sentences and age will exhibit variation in their associations with recidivism because of the different pathways that black and white men and women take through their criminal careers. Evidence of this will provide quantitative support for one of the core tenants of intersectionality, that marginalized groups of people exhibit unique criminogenic risks, which differ from those typically identified in largely white and male samples.

**Hypothesis 2:** The BARR instrument developed in chapter 3 will exhibit good predictive validity across racial and gender subgroups, with similar rates of correct classification and discrimination for all Washington State inmates. Predictive parity would demonstrate

further the utility of the instrument for balancing fairness and accuracy, which I expect will be the case because Bayesian methods appeared to adapt to empirical patterns in the data and minimize prediction error across the sensitivity-specificity curve. This should prevent excessive rates of false positives or negatives, which would instead indicate bias in the instrument.

**Hypothesis 3:** Overall, the BARR instrument will better predict recidivism among the groups that are most prevalent in the data, and that share the most in common with white men (gender for black men, and race for white women). For black women, who would exhibit the greatest differences in risk factors according to intersectional and pathways theories, the gender and race-specific models will better capture these risks and result in more accurate classifications than the neutral BARR.

**Methods**

Procedures for constructing and validating the Bayesian models used to predict recidivism rates by race and gender subgroups were similar to those described in Chapter 2, with a few notable differences. Cohorts were combined in different ways to harness large enough samples for analysis in each subgroup. Given that I already established that a minimally informative Gaussian prior provided a good fit for the data, I began by fitting models with this prior for all four subgroups (black and white men and black and white women). For women, these models were fitted using data from cohorts 1, 2, and 3 (1986-

1995). For men, where there was much more data available, I fit this minimally informative model to cohorts 1 and 2 (1986-1992).

Using these models, I optimized predictor inclusion and coding by inspecting predictor log-odds and posterior standard deviations. Since I wanted models to be optimized for the specific risks of each racial/gender subgroup, parameter inclusion was determined by the presence of positive log-odds (in the expected direction) of larger magnitude than their associated standard deviations. This resulted in slightly different combinations of variables being included for each subgroup, which sends some support to *Hypothesis 1.* Next, I coded these substantively important variables such that coefficients would increase monotonically over the categories of each predictor (i.e. lower age=higher recidivism, more prior felonies=higher recidivism). This was particularly important with these reduced sample sizes, as often there would be small cell sizes among female offenders that would result in illogical coding patterns. Changes to both variable inclusion and coding were assessed via model AICs and Bayes Factors. All changes resulted in improved model fits within each racial/gender subgroup.

The large range of sample sizes and numbers of parameters across models required a bit more fine tuning for the Monte Carlo Markov Chains (MCMCs) to converge than in the overall models. For white women, stability occurred for most measures with a burn-in of 10,000 chains, but some variables (most notably, age) required a longer run for shrinkage factors to converge around 1 (which is preferred in the Gelman and Rubin convergence diagnostics; Gelman et al., 2014). As a result, I increased burn-in to 40,000 and the overall number of MCMCs to 150,000. For black women, MCMCs had difficulty converging around the intercept, prior drug felonies, and prior violent property offenses. After increasing the

burn-in period to 100,000 with a total number of MCMCs of 200,000, estimates stabilized considerably. The models for white men were also difficult to fit, requiring a burn-in period of 150,000. To ensure that there were adequate chains to perform estimation on, I increased the total number of MCMCs to 250,000. Black men's models stabilized much more quickly, however, with a burn-in of 80,000 and 180,000 MCMCs.

For black and white women, the final models used to construct the assessment were fitted to cohorts 4 and 5 (1996-2001) using the posterior means of the cohort 1-3 model as priors. Their precision was determined by fitting various multipliers of the posterior precision (1/SE^2) and testing the different models against each other using the Bayes Factor to decide on the optimal model. Because the men's samples were substantially larger, the process of obtaining and updating priors was more similar to the strategy described in Chapter 2. I used cohorts 3 and 4 (1993-1998) to update the priors obtained from cohorts 1 and 2, and then fitted the final models in cohorts 5 and 6 (1999-2004).

Assessments for each group were validated using the newest data available (that yielded sufficient sample size) and strategies described in Chapter 3. Table 4.1 displays the cut points used to make classification decisions for each group's risk of felony recidivism. The validation samples consisted of 2,688 black women, 6,034 black men, 18,943 white women, and 33,187 white men. I compare the resulting accuracy and discrimination of each of these subgroup models against the predictions made by the overall instrument generated in Chapter 3 to test *Hypothesis 3*. To test *Hypothesis 2*, I also re-examined the predictive validity of the BARR separately for each of the race and gender subgroups. Because of the relatively small sample sizes for some groups, all of these results were confirmed using 1000 bootstraps of the classification error following the same procedure

described in Chapter 3, though I present the results of the hold-out validation sample here (but results available upon request) for brevity.

**Table 4.1: Classification Cut Points for Gender and Race-Specific Models**

| | Predicted Pr. |
|---|---|
| White Males | 0.25 |
| White Females | 0.16 |
| Black Males | 0.30 |
| Black Females | 0.26 |

## Results

*Women's models*

Table 4.2 reports the Log-Odds and 95% Bayesian Credible Intervals (BCAs) for the variables and coding schemes that made it into the final risk assessment models for black and white women. Among women, overall, felony recidivism risk was comprised of just 12 of the 18 criminogenic domains employed by the BARR. Of these, only 7 were shared across race. Even among the shared factors, optimal coding schemes differed considerably and several estimates of risk differed enough to lie outside of the 95% BCA for white women. For example, a large number of prior felony sentences did not contribute nearly as much to felony recidivism risk in black women as in white women, but the number of prior felony property offenses contributed much more. Likewise, the number of felony drug sentences was a very important predictor of black women's felony recidivism risk, whereas the measure was only a moderate predictor for whites. Perhaps most important, the difference

in the intercepts of the two groups was well outside of either group's estimated BCA, indicating much higher assumed baseline risks of black women being reconvicted for a felony offense than white women.

**Table 4.2: Bayesian Logistic Regression Models Predicting Female Felony Recidivism**

| White Women | | | Black Women | | |
|---|---|---|---|---|---|
| Log-Odds | 95% BCA | | Log-Odds | 95% BCA | |
| **Felony Recidivism** | | | | | |
| Age (50+ ref.) | | | - | | |
| 30-49 | 0.717 | (0.472 , 0.971) | - | | |
| 18-29 | 0.810 | (0.565 , 1.079) | - | | |
| Juvenile Felonies | | | Juvenile Felonies | | |
| 1 | 0.566 | (0.393 , 0.744) | 1+ | 0.723 | (0.505 , 0.930) |
| 2+ | 1.022 | (0.851 , 1.183) | - | | |
| Felony Sentence Count | | | Felony Sentence Count | | |
| 1 | 0.231 | (0.075 , 0.387) | 1 | 0.308 | (0.130 , 0.500) |
| 2 | 0.942 | (0.753 , 1.115) * | 2 | 0.541 | (0.291 , 0.782) |
| 3 | 1.361 | (1.146 , 1.585) * | 3+ | 0.635 | (0.350 , 0.917) |
| 4+ | 1.717 | (1.468 , 1.967) | - | | |
| - | | | Felony Violent Property # | 0.293 | (0.070 , 0.507) |
| Felony Property Offenses | | | Felony Property Offenses | | |
| 1 | 0.157 | (0.053 , 0.263) | 1 | 0.182 | (0.040 , 0.330) |
| 2 | 0.172 | -(0.018 , 0.364) * | 2 | 0.831 | (0.582 , 1.075) |
| 3+ | 0.234 | -(0.011 , 0.475) * | 3+ | 0.924 | (0.589 , 1.262) |
| Felony Drug Offenses | | | Felony Drug Offenses | | |
| 1+ | 0.339 | (0.237 , 0.442) * | 1 | 0.646 | (0.500 , 0.792) |
| - | | | 2+ | 1.649 | (1.243 , 2.049) |
| Felony Escapes # | 0.115 | -(0.070 , 0.301) * | Felony Escapes # | 0.504 | (0.253 , 0.749) |
| - | | | Felony Non-Domestic Assaults | | |
| - | | | 1+ | 0.369 | (0.198 , 0.537) |
| Misd. Weapons # | 0.400 | (0.154 , 0.645) | Misd. Weapons # | 0.380 | (0.033 , 0.719) |
| Misd. Property Offenses | | | Misd. Property Offenses | | |
| 1 | 0.480 | (0.370 , 0.598) | 1 | 0.417 | (0.268 , 0.555) |
| 2 | 0.803 | (0.631 , 0.975) * | 2 | 0.577 | (0.358 , 0.804) |
| 3+ | 0.966 | (0.799 , 1.130) | 3+ | 1.088 | (0.882 , 1.294) |
| - | | | Misd. Drug Offenses | | |
| - | | | 1 | 0.313 | (0.057 , 0.582) |
| - | | | 2+ | 0.699 | (0.192 , 1.193) |
| - | | | Misd. Alcohol # | 0.175 | -(0.049 , 0.391) |
| Intercept | -3.572 | -(3.847 , -3.299) * | Intercept | -2.988 | -(3.163 , -2.815) |

* Estimate of characteristic lies outside of the 95% BCA of White women

The way that these differences in risk play out over multiple offender characteristics is informative. Figure 4.1 presents the predicted probabilities of felony reconviction among two very different groups of potential offenders under correctional supervision: the first,

minimally risky group has one prior felony conviction for a drug charge while the other, very high risk group has an extensive criminal history consisting of multiple juvenile and adult felonies and misdemeanors for property, drug, and weapons crimes. I choose to focus on these characteristics because they were important in both the white and black female models and because I could choose equivalent categories across both racial groups for comparability, though I also vary age among white women because it is a fundamental characteristic of all individuals in the sample, though it was not a substantive predictor for black women. The results are illustrative of an important outcome of the race and gender-specific models—despite the varying weights given to each of these offending characteristics across race, black women with identical histories are always going to be expected to recidivate at a somewhat higher rate. In fact, not even the large increase in the odds associated with belonging to the youngest age range among white women can offset this difference. Overall, while the types and magnitudes of risk factors vary between black and white women, as intersectionality would suggest, the ways that they play out in terms of classification would likely tend to disadvantage black women more than it would result in fairer risk estimates.

*Men's Models*

The Log-Odds and 95% BCAs of the final white and black men's models are displayed in Table 4.3. Overall, the factors identified as salient predictors of felony recidivism in both groups of men were similar to those utilized in the BARR. Fourteen of the 18 criminogenic features used in the gender-neutral model were informative in distinguishing between recidivists and non-recidivists, with 11 shared across both racial

groups. Misdemeanor drug offenses were the only item used in the white men's model that failed to identify recidivists among black men, while misdemeanor escapes and felony weapons charges were only predictive of reconviction among black inmates. Optimal coding schemes tended to be similar, but black men between the ages of 50 and 59 showed no evidence of heightened recidivism, and recidivism risk seemed to peak after two adult and juvenile felony convictions among black inmates. There were, however, some distinct differences in the predicted log-odds of these items across race. Violent property felonies and felony drug offenses increased the odds of reconviction more for black than white male offenders, while misdemeanor domestic assaults were a stronger predictor of recidivism for white men, and all of these distinctions were beyond what would be expected by chance (outside of the 95% BCAs for the white offenders). The intercept also varied systematically between the two groups, with black males displaying a much higher baseline risk of recidivating than white men in the sample.

**Figure 4.1: Predicted Probabilities of Felony Reconviction Among Women by Race and Offender Characteristics**

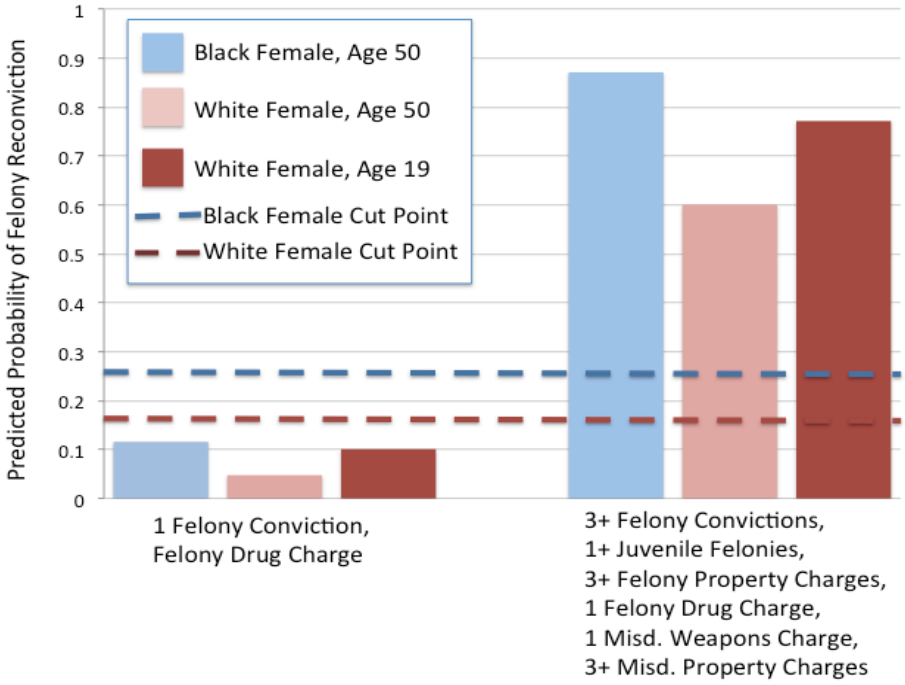
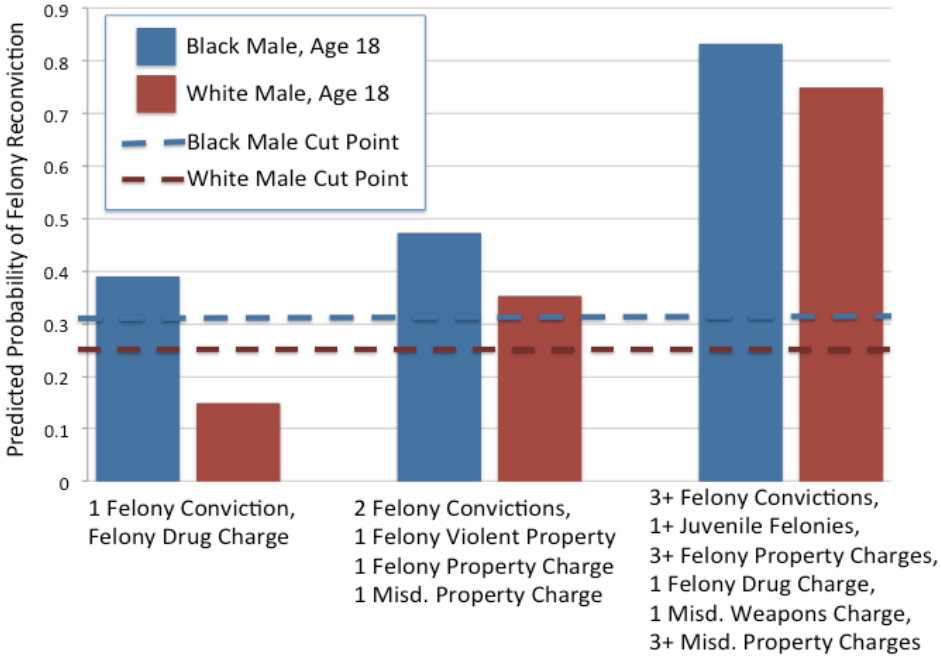



**Figure 4.2: Predicted Probabilities of Felony Reconviction Among Men by Race and Offender Characteristics**



## Table 4.3 Bayesian Logistic Regression Models Predicting Male Felony Recidivism

| | White Men | | | | Black Men | |
|---|---|---|---|---|---|---|
| | Log-Odds | 95% BCA | | | Log-Odds | 95% BCA |
| | | | | **Felony Recidivism** | | |
| Age (60+ ref.) | | | | Age (50+ ref.) | | |
| 50-59 | 0.209 | (0.000 , 0.427) | | - | | |
| 40-49 | 0.540 | (0.349 , 0.719) | | 40-49 | 0.395 | (0.242 , 0.554) |
| 30-39 | 0.729 | (0.537 , 0.913) | | 30-39 | 0.558 | (0.409 , 0.694) |
| 20-29 | 0.793 | (0.612 , 0.981) | | 20-29 | 0.730 | (0.608 , 0.868) |
| 18-19 | 1.265 | (1.074 , 1.460) | | 18-19 | 1.280 | (1.104 , 1.465) |
| Juvenile Felonies | | | | Juvenile Felonies | | |
| 1 | 0.503 | (0.432 , 0.578) | | 1 | 0.442 | (0.294 , 0.599) |
| 2 | 0.634 | (0.543 , 0.734) | * | 2+ | 1.033 | (0.906 , 1.162) |
| 3+ | 1.038 | (0.954 , 1.132) | | - | | |
| Felony Sentence Count | | | | Felony Sentence Count | | |
| 1 | 0.336 | (0.262 , 0.417) | | 1 | 0.409 | (0.300 , 0.519) |
| 2 | 0.822 | (0.731 , 0.920) | * | 2+ | 0.468 | (0.316 , 0.615) |
| 3 | 1.057 | (0.945 , 1.162) | | - | | |
| 4+ | 1.080 | (0.941 , 1.223) | | - | | |
| Felony Violent Property # | 0.258 | (0.184 , 0.317) | * | Felony Violent Property # | 0.336 | (0.236 , 0.430) |
| - | | | | Felony Weapons # | 0.053 | -(0.109 , 0.200) |
| Felony Property Offenses | | | | Felony Property Offenses | | |
| 1 | 0.364 | (0.313 , 0.413) | * | 1 | 0.120 | (0.020 , 0.220) |
| 2 | 0.543 | (0.456 , 0.624) | * | 2 | 0.341 | (0.157 , 0.514) |
| 3 | 0.695 | (0.575 , 0.829) | | 3 | 0.639 | (0.402 , 0.909) |
| 4 | 0.962 | (0.779 , 1.135) | * | 4 | 0.576 | (0.236 , 0.915) |
| 5+ | 1.066 | (0.866 , 1.245) | | 5+ | 1.216 | (0.878 , 1.593) |
| Felony Drug Offenses | | | | Felony Drug Offenses | | |
| 1 | 0.347 | (0.296 , 0.398) | * | 1 | 0.487 | (0.393 , 0.586) |
| 2 | 0.548 | (0.458 , 0.651) | | 2 | 0.644 | (0.489 , 0.802) |
| 3+ | 0.802 | (0.662 , 0.937) | * | 3+ | 1.229 | (1.038 , 1.425) |
| Felony Escapes # | 0.165 | (0.082 , 0.240) | * | Felony Escapes # | 0.413 | (0.272 , 0.567) |
| Misd. Non-Domestic Assaults | | | | Misd. Non-Domestic Assaults | | |
| 1 | 0.204 | (0.140 , 0.265) | | 1 | 0.144 | (0.027 , 0.252) |
| 2+ | 0.309 | (0.212 , 0.422) | * | 2+ | 0.211 | (0.039 , 0.365) |
| Misd. Domestic Assaults | | | | Misd. Domestic Assaults | | |
| 1+ | 0.478 | (0.416 , 0.536) | * | 1+ | 0.263 | (0.090 , 0.406) |
| Misd. Weapons # | 0.266 | (0.163 , 0.369) | | Misd. Weapons # | 0.219 | (0.029 , 0.394) |
| Misd. Property Offenses | | | | Misd. Property Offenses | | |
| 1 | 0.376 | (0.323 , 0.426) | * | 1 | 0.313 | (0.206 , 0.412) |
| 2 | 0.483 | (0.398 , 0.574) | | 2 | 0.562 | (0.389 , 0.719) |
| 3+ | 0.654 | (0.566 , 0.732) | | 3+ | 0.694 | (0.522 , 0.856) |
| Misd. Drug Offenses | | | | - | | |
| 1 | 0.319 | (0.262 , 0.382) | | - | | |
| 2+ | 0.411 | (0.314 , 0.511) | | - | | |
| - | | | | Misd. Escapes # | 0.272 | -(0.226 , 0.756) |
| Intercept | -3.690 | -(3.876 , -3.493) | * | Intercept | -2.623 | -(2.760 , -2.492) |

* Estimate of characteristic lies outside of the 95% BCA of White men

The baseline difference in the anticipated odds of reoffending from these race-specific models exerts a huge influence when we examine the predicted probabilities of recidivism for hypothetical offenders in the sample. Figure 4.2 shows the predicted probabilities of felony reconviction for 3 sets of hypothetical, 18 year old offenders in the sample—a low-risk offender with a single drug felony, a violent offender with a relatively short criminal history, and a very high-risk offender with several convictions for property, drug, and weapons crimes—while varying only offender race. This depiction makes clear that even with few criminogenic risks, young black men will almost always be expected to recidivate. In fact, the only felony conviction that an 18-19 year old black man could have and not result in their being classified as a recidivist is a single weapons charge. Even a single felony property offense (not pictured), which was the next lowest item weight in the model (LO=0.120), results in a predicted probability high enough to be classified as a recidivist on the strength of the black male intercept and age coefficient, alone. While this racial gap seems to close somewhat with more extensive criminal histories, it does highlight a potential drawback of an empirical classification system based on an intersectional perspective. When gender and race-specific risks and item weights are used to predict recidivism, models highlight the heightened rates of offending among young black men in the sample. In turn, the resulting risk classifications may ensure harsher punishment for young black men, which may in turn result in increased recidivism.

**Table 4.4: Gender and Race-Specific Model Predictive Properties vs. Gender and Race-Neutral Model, Felony Recidivism**

| | White Women | | Black Women | | White Men | | Black Men | |
|---|---|---|---|---|---|---|---|---|
| | Specific | Neutral | Specific | Neutral | Specific | Neutral | Specific | Neutral |
| AUC | 0.66 | 0.73 | 0.71 | 0.73 | 0.74 | 0.74 | 0.69 | 0.70 |
| F1 | 0.38 | 0.43 | 0.45 | 0.42 | 0.47 | 0.49 | 0.55 | 0.55 |
| MCC | 0.24 | 0.29 | 0.29 | 0.28 | 0.30 | 0.32 | 0.28 | 0.28 |
| ACC | 0.77 | 0.78 | 0.72 | 0.77 | 0.74 | 0.74 | 0.66 | 0.66 |
| NPV | 0.85 | 0.86 | 0.87 | 0.85 | 0.84 | 0.85 | 0.76 | 0.75 |
| PPV | 0.39 | 0.43 | 0.37 | 0.43 | 0.46 | 0.46 | 0.51 | 0.52 |

AUC= Area Under the Receiver Operating Characteristic (ROC) Curve
F1= Harmonic mean of precision and sensitivity
MCC= Matthews Correlation Coefficient between observed and predicted classification
ACC= Accuracy or percentage of cases correctly classified
NPV= Negative Predictive Value, or the probability of correctly identifying someone who does not recidivate
PPV= Positive Predictive Value, or the probability of correctly identifying someone who does recidivate

*Predictive Validity*

Comparisons of the validity of the race and gender-specific model predictions to the gender-neutral BARR reveal further disadvantages of intersectionality-based assessment models. Table 4.4 displays the AUCs, ACCs, $F_1$ Scores, and MCCs of the race and gender-specific model classifications, as well as the same figures assessed using the gender-neutral model on the same validation sample. I also include here the positive (PPV) and negative predictive values (NPV), which are included in the calculations of ACC, $F_1$, and MCC, to show which type of errors are most likely to be made by the models. Positive prediction errors occur when offenders recidivate when they were *not* expected to, whereas negative errors occur when offenders no not recidivate when they *are* expected to. While the black women's model demonstrates good accuracy (ACC=0.72, $F_1$ =0.45, MCC=0.29) and discrimination (AUC=0.71), the white women's model performs notably worse on each metric aside from the ACC (AUC=0.66, $F_1$ =0.38, MCC=0.24, ACC=0.77). Both race-specific models were also relatively poor at correctly identifying recidivists (PPV= 0.37-0.39). In comparison, the gender-neutral model performs better in nearly every respect, with almost

identical validity across both groups of women (AUC=0.73, ACC=0.77-0.78, $F_1$ =0.42-0.43, MCC=0.29, PPV=0.43). So, while the black women's model displayed comparable predictive validity, the white women's model performed somewhat poorly and there was no notable improvement over the gender-neutral model in either case.

There were fewer differences in the predictive validity of the gender/race-specific and neutral models among men. In fact, AUCs, $F_1$ scores, MCCs, and ACCs were nearly identical regardless of whether specific item weights were employed. There was one notable difference between black men and all other groups, however; negative predictive values were much lower among black men than all other groups, regardless of model type, while positive predictive values were somewhat higher. In other words, while it resulted in a slightly better ability to correctly identify those black men who actually did recidivate (PPV=0.51-0.52 vs. 0.37-0.46 in other groups), both risk assessment strategies tended to overstate the risk posed by black men, making it more difficult to correctly identify those who did not recidivate (0.75-0.76 vs. 0.84-0.87 in other groups). This helps to confirm the problem demonstrated by Figure 4.2 with a gender and race-neutral classification method. Overall, it appears that black men are so likely to be drawn back into the correctional system that it is difficult for static risk assessment tools to correctly identify non-recidivists, which may result in some degree of systematic overclassification of black male risk when using such tools, even if they do not explicitly use risk in the classification process.

**Discussion**

Researchers have argued that women have their own set of criminogenic risks and needs, using pathways theory to highlight the unique factors that precipitate entry into and desistence from criminal careers (e.g. Belknap 2007; Daly & Tonry, 1997; Hill-Collins, 1998b; Katz, 2000). In response, some instruments like the STRONG-R and Women's Risk Needs Assessment (WRNA) have adopted separate item scorings or entire systems of classification specifically for women in efforts to improve classification accuracy and better reflect women's pathways to crime (Hamilton et al., 2016; Salisbury, Van Voorhis, & Spiropoulos, 2009). However, these adaptations have not always resulted in better predictions (Hamilton et al., 2016). This may be because these same theories of gendered pathways often incorporate an intersectional framework, positing that both race *and* gender are central to understanding the behaviors of doubly marginalized groups like black women (Potter, 2015). It is possible that gender-sensitive instruments are simply not enough, and that risk assessment needs to reflect this intersectional understanding of offending. Yet, so far this hypothesis has remained untested.

In this chapter I took this argument to its logical conclusion, examining the validity of Bayesian race and gender-specific models for predicting recidivism among black and white men and women using a large population of individuals under correctional supervision in Washington State. As expected, there were distinct racial and gender differences in which risk factors were most predictive of felony recidivism. For example, violent offenses were not substantive predictors of reconviction among white women, whereas violent property felonies were associated with recidivism among black women and both groups of men. Likewise, age was not predictive of recidivism among black women, though recidivism was strongly age-graded among all other groups, while felony

property and drug offenses were much more strongly associated with future offenses for black than white women. Perhaps most surprisingly, there was not a strong gradation of risk associated with the count of prior felony convictions for black men (and to a somewhat lesser extent, black women), despite this being one of the most important predictors in gender-neutral instruments and in the white male and female models. The presence of so many differing item weights and characteristics across racial and gender subgroups lends considerable support for intersectional theories and the idea that both race and gender contribute to the unique pathways that push and pull people from criminal lifestyles throughout the life course.

However, findings regarding the validity of classification decisions made using race and gender-specific predictors indicate some considerable statistical and practical consequences of separately modeling recidivism risk for each of these groups. The intersectionality-based models roundly failed to outperform the gender and race-neutral BARR instrument that I developed in Chapter 3. While white and black men's models displayed predictive potential very similar to that demonstrated by the neutral model, both black and white women were classified more accurately using the neutral strategy. This was contrary to expectations derived from intersectional theories, and instead suggests that gender specific pathways co-exist with more general explanations for crime (Jones et al., 2014). At least for the purposes of predicting recidivism risk, these general explanations appear to result in more accurate risk classifications. It thus appears that the set of characteristics predictive of recidivism in each subgroup model does not provide a more comprehensive scheme for distinguishing between recidivists and non-recidivists than the fuller array included in the BARR. This may be because the actual risks that would help to

best identify possible recidivists for these groups are related to dynamic or other outside factors that are not measured here, which would be more consistent with intersectional explanations of crime. There is some support for this in the fact that the gender and race-specific models perform best for white men, who make up the largest share of the WADOC data and whose comprehensive set of risk factors may be best represented by the present data. Regardless, it is necessary to perform a similar test using risk *and* needs measures to determine more precisely why the BARR performed best for all groups.

More concerning, the gender and race-specific instruments' differing intercepts for black men and women ensured that these groups would be systematically rated as higher risk than their white counterparts. This was especially consequential for young black men, who would almost automatically be rated as likely recidivists, regardless of their criminal histories. One hypothetical illustration of this was especially impactful: for a single prior drug felony conviction, 18-19 year old black men were rated as 25% more likely to recidivate than white men with the same characteristics. Models that result in this kind of imbalance purely as a result of race or gender run the risk of exacerbating existing disparities rather than reducing them. Given that intersectional criminologists have tended to publicly mix social activist and scholarly roles (Arrigo, 2016; Belknap, 2015; Potter, 2013), this would be an undesirable outcome and one that should impact our understanding of both how theoretical perspectives can or should inform practice, and what we can do to improve the fairness of risk assessment. It is possible that incorporating racially responsive dynamic needs factors into these models may help to alleviate some of this problem, given that static characteristics tend to exacerbate racial differences in recidivism risk (Skeem & Lowenkamp, 2016). This could be a promising area of future

research, particularly since the use of racially specific needs that *reduce* the predicted risk of recidivating are likely to be less contentious in practice than assigning different scores for criminal history characteristics that *increase* the expected risk of recidivism by race and gender.

*Implications for Race Making and Theoretically informed Practice*

My results indicate that when theory meets practice, intersectional, gender *and* race-specific risk assessments may have unintended consequences for black men and women. In fact, by accepting the racial classifications made by inmates themselves and/or correctional staff, race becomes "real" and runs the risk of reifying social truths like the overrepresentation of black men in the U.S. correctional system. When race itself becomes a risk factor, as represented by the varying intercepts of the race and gender-specific instruments, risk assessment becomes a concrete example of a racial project (Omi & Winant, 1994). Not unlike racial and ethnic segregation in prisons (Walker, 2016), the use of any form of race-specific statistical assessment will likely act as a way of empirically rationalizing harsher treatment and increased surveillance against black men, which may then result in a self-fulfilling prophecy of increased recidivism (Mears, Cochran, & Bales, 2012; Mitchell et al., 2017). What is especially concerning is the fact that this gender and race-specific tool ultimately yielded the same predictive characteristics as the gender and race-neutral BARR, with both demonstrating good validity overall but somewhat poorer ability to correctly identify non-recidivists among black men than other groups. This highlights the importance of researchers explicitly reporting these more in-depth

classification statistics for racial and gender subgroups in their data, as overall depictions of accuracy and discrimination may obscure potentially worrisome errors being made.

Despite the centrality of risk assessment to the criminal justice process, research on the implications of risk assessment for racial disparities in punishment is still somewhat scarce. Previous work has demonstrated that static risk assessment scores on the Post-Conviction Risk Assessment (PCRA) were correlated with race (Skeem & Lowenkamp, 2016), and some examinations of the COMPAS instrument have suggested that there is slight racial bias in its predictive validity (Angwin et al., 2016; Hall & Gill, 2017), while others have argued that there are alternative explanations that better account for the bias (Flores, Lowenkamp, & Bechtel, 2017). It is therefore still somewhat unclear if race-neutral assessment strategies result in biased classifications, but it is likely that if risk and race are related, it is primarily through criminal history that this association occurs (Skeem & Lowenkamp, 2016). After all, numerous studies have demonstrated that black men tend to face greater odds of being arrested for crimes (Alexander, 2010; Frase, 2009; 2014; Ulmer, Painter-Davis, & Tinik, 2014) and receive jail or prison sentences more often for the same crimes (Chiricos & Crawford, 1995; Freiburger & Hilinski, 2013; Zatz, 2000), while also having higher rates of probation and parole revocations for new offenses and technical violations than white men, white women, or black women (Huebner & Bynum, 2008; Olson & Lurigio, 2000; Schlesinger, 2005; Vito, Higgins, & Tewksbury, 2012).

Given that the WADOC data employed here consists almost entirely of factors related to criminal history, it is perhaps unsurprising that I found race and gender-specific models to predict much higher odds of black men recidivating. It is somewhat disheartening to know that assessment tools adopting an intersectional perception of risk

could further reify race and exacerbate racial disparities in rates of correctional supervision, but even more neutral assessment practices appear to overclassify black men compared to other groups. However, there may still be hope for needs assessments to benefit from these kinds of specific models, as gender and race-specific programming could be useful for mitigating risk. Unfortunately, an examination of the impact of including these types of dynamic measures in the models was not possible with the present data. Future research should use a similar strategy to investigate whether criminogenic needs do, in fact, vary by race and gender, and whether the inclusion of race and gender-specific weights in needs assessment has any utility.

In so doing, researchers should be aware of the potential for any race-based treatment in the criminal justice system to become a racial project of its own (Omi & Winant, 1994), wherein the imagined differences between racial groups take on some form of reality through institutional response. That racial classification may not even reflect the ones that inmates would ideally make for themselves is particularly problematic in this regard, notwithstanding issues with 'accurately' classifying something as socially constructed as race, more generally (Bowker & Star, 1999). Therefore, adopting an intersectional framework for addressing the risks and needs of reentry is likely considerably more difficult than it sounds.

Unique intersectional risks and needs extend far beyond the simple race and gender ones investigated here to dimensions like class, victimization histories, psychological problems, substance addiction, family demands, and queer identities. Social service systems are often inadequate as is, and these "multiple needs" deeply complicate the kinds of demands placed on this system (Bunn, 2018). From a statistical perspective, capturing

the interaction of all of these complex factors would require the measurement of hundreds of factors, and very large data sets upon which to construct and validate instruments. Rather than focusing on quantitative applications of intersectional perspectives in the area of assessment, perhaps the greatest promise of the intersectional perspective is in its ability to qualitatively inform practitioners of ways to better understand client motivations and constraints that affect their ability and willingness to desist from criminal lifestyles. These understandings could be used to identify programs that instill the proper "hooks for change" (Giordano, Cernovich & Rudolph, 2002) necessary for some offenders to desist from crime. Unfortunately, the lack of dynamic measures in the WADOC data made it impossible to investigate how these types of factors varied in their relationships with recidivism across race and gender. Nonetheless, this would be another fruitful area of future research.

*Limitations and Conclusion*

No research is without its limitations, and this study is no exception. The lack of data on offender ethnicity and small samples of other racial and ethnic groups made it impossible to investigate the impacts of race and gender-specific assessment strategies among groups other than blacks and whites. This is especially important because the current analyses likely group together Latinos and Latinas with white men and women. It is difficult to infer how this may have impacted findings because information regarding the ethnic makeup of the Washington State correctional population by ethnicity and gender is scarce. If many of the women classified as white were actually Latina, and Latina women have very different risk factors for reoffending than white women, then the conflation of

risks among the two groups could help explain why the white women's model demonstrated such poor predictive validity in the gender and race-specific instrument. More in-depth research with data that contains good information on inmate ethnicity would be necessary to determine if this is the case. I was also unable to include dynamic measures of offender risks and needs, which may have highlighted best the myriad mechanisms that underlie the various intersectional pathways to crime. Relatedly, the data do not include information about sexual orientation, victimization histories, socioeconomic background, or other characteristics that help to flush out the many sources of oppression that intersectional theories emphasize as very important determinants of behaviors like crime.

While all of these constraints limited my ability to fully test how risk assessments might be improved by adopting an intersectional focus, I did make use of a very large dataset to provide the first examination of how risk factors play out across the 4 most prominent race and gender subgroups in the U.S. The results, as expected, highlighted that predictors of recidivism risk differ widely between black and white men and women, with some factors common to all existing risk assessment strategies showing no association with recidivism among select groups (like age for black women). However, none of these deeper understandings of the risk associated with each group actually resulted in better or fairer predictions, and in the case of black men, race and gender-responsive predictors led to systematic overclassification of risks that could exacerbate existing disparities in punishment. Thus, while the principles of intersectionality were in some ways supported, the implication that recognizing the unique risks contributing to criminality in each group would result in better or more equitable classifications was not.

These findings demonstrate that any form of risk assessment that directly or indirectly relies on race to arrive at classification decisions, however well intentioned, runs the risk of furthering the process of racial formation (Omi & Winant, 1994) and race making (Wacquant, 2001) taking place already in the criminal justice system. Because criminal history is so strongly related to race, I propose that risk assessment itself already constitutes a racial project, wherein even more conventional gender and race-neutral instruments have difficulty in correctly identifying young black men, especially, as non-recidivists relative to other race/gender subgroups. Developers of such instruments should keep this in mind and examine the predictive validity of their prediction models across racial/ethnic and gender groups to determine what kinds of disparate impacts their tools will have on different offender populations. Regular investigation of this previously underexplored domain may perhaps yield methods that can be used to mitigate these unintended consequences, resulting in greater fairness and accuracy of predictions across all groups.

# Chapter 5 - CONCLUSION

## Implications for Theory and Practice

Risk assessment is a defining feature of the modern system of punishment in the U.S. (Garland, 2001). Much like other data-driven practices adopted by the criminal justice system (e.g. COMPSTAT, Mazerolle, Rombouts, & McBroom, 2007; IBM Crime Management Systems, IBM Software Group, 2010 ; "Hot Spots" policing, Braga, 2001), actuarial assessment aims to increase efficiency in the midst of an overburdened system. In many ways, the practice is not dissimilar to other changes in law and procedure associated with adopting neoliberal attitudes toward criminals (Garland, 1990), signaling an increasingly "objective" approach toward crime control and resource allocation. But, in other ways, risk assessment is a successful evidence-based reformative practice (Vera Institute of Justice, 2014) that shows promise for reducing recidivism rates by releasing low-risk offenders earlier and incapacitating those criminals who pose the greatest threat to public safety (Berk, 2017). Thus, the actuarial assessment of risk lies at the confluence of necessity and reform, shaping offender outcomes at the pre-trail, sentencing, program provision, probation, and parole stages of processing.

Given the number of ways that the practice influences offender outcomes, it is no surprise that a deep body of research has emerged to investigate its impacts. The majority of researchers have adopted a practical focus, examining the predictive validity of existing instruments, developing and validating new instruments, investigating how methodological variation across instruments impacts predictive power, or advocating for the importance of risk *and* needs assessment (Andrews & Bonta, 1995; Barnoski & Drake, 2007; Berk et al.,

2009; Brennan, Dieterich, & Ehret, 2009; Duwe, 2013; Hamilton et al., 2016; Hare, 1991; Latessa et al., 2009; Liu et al., 2011; Schaffer, Kelly & Lieberman, 2011; Skeem & Louden, 2007). These areas of inquiry are certainly important, with their findings holding direct relevance for which instruments are adopted in real world settings and for the internal and external validity of the practice, as a whole, in its ability to increase efficiency and boost public safety. But, a largely separate group of scholars have tended to question risk assessment's basic ability to accomplish these goals, levying theoretical critiques against the tools for being unresponsive to gendered and raced pathways to crime, failing to establish predictive validity amongst racial/ethnic and gender subgroups, and utilizing predictors that are strongly related to, and may in fact be proxies of, race (Angwin et al., 2016; Belknap & Holsinger, 2006; Blanchette & Brown, 2006; Desemarais & Singh, 2013; Harcourt, 2015; Hudson & Bramhall, 2005; Salisbury, van Voorhis, & Spiropoulos, 2009; Skeem & Lowenkamp, 2016). These concerns are easily justified—as part of the primary system of institutional racism in the modern era (Alexander, 2010), we should be concerned about how all criminal justice practices differentially impact people of color and marginalized women.

However, neither line of research goes far enough. Those studying risk assessment with a primarily practical focus have regularly ignored the potential theoretical contributions of their work (though see Andrews & Bonta, 1999; Bonta, 2002 for exceptions), and relatively few of the critical theorists have used quantitative analysis of risk classifications or developed assessments of their own to investigate their claims and ameliorate the problems they work to identify (though see Salisbury, Van Voorhis, & Spiropoulos, 2009; Van Voorhis & Presser, 2001 for exceptions). My dissertation exists at

the confluence of both bodies of research, and attempts to fill these gaps. I used the development and validation of a Bayesian Risk Assessment instrument (the BARR) as the jumping off point to weigh in on the debate over offense specialization or versatility, and to test whether assessment strategies that feature race and gender-specific items, as suggested by intersectional theories, outperform more traditional, general instruments. Altogether, my findings have clear ramifications for both theory and practice.

**The Utility of Bayesian Methods for Risk Assessment**

Methodological innovation has played an important role in the development of risk assessment, as a discipline, from its earliest days using bivariate associations (e.g. Burgess, 1928) to a new era of instruments that rely on machine learning algorithms to predict recidivism (e.g. Berk & Bleich, 2013). However, an entire branch of statistics has been overlooked in this constant search for new methods: Bayesian analysis. While some have recently advocated for greater use of Bayesian statistics in the area of criminal justice (Fenton, Neil, & Berger, 2016; Philipse, 2015), and criminologists have briefly flirted with Bayesian methods in the past (Berk et al., 1992), none have employed this type of analysis for the purposes of recidivism risk assessment; though it has been used fruitfully in the medical and genomics fields for this purpose (Fenton and Neil, 2012; Newcombe et al., 2012; Ogino & Wilson, 2004). I argue that Bayesian statistics, in incorporating prior information and directly estimating the underlying probability and uncertainty of recidivism, is an ideal middle ground between theory and practice-derived regression models and empirical machine learning techniques.

I found strong evidence for the utility of Bayesian risk assessment in this study. The BARR performed as well or slightly better than a carefully specified frequentist regression model, with improvements especially evident in rates of incorrect classification in offense-specific models. Predictive validity was comparable with the best regression and machine learning instruments reported elsewhere (Desemarais & Singh, 2013; Hamilton et al., 2015; 2016), but this average improvement only tells part of the story. The BARR also seemed to reduce uncertainty in model estimates, which should reap additional benefits for recidivism prevention when applied in practical settings. This would be especially evident if the technique was applied to samples with less data, where the use of empirically derived priors can lend additional power to the analysis (Loh et al., 2015; Stegle et al., 2010).

Overall, it seems that the BARR and other sophisticated regression and machine-learning risk assessment instruments may have closed in on a ceiling for how accurately we can predict recidivism using statistical methods. The fact that nearly all of these tools attain AUCs in the range of 0.75, despite substantial variation in statistical techniques, datasets, outcomes, and predictors across studies seems to confirm this. Bowker and Star (1999) would suggest that this is because different modeling strategies can only reduce some of the myriad sources of error that go into human classifications[8]. This is why the Bayesian strategy is especially useful, from a philosophical standpoint—by directly estimating the probability of recidivism and the uncertainty involved in this estimate, it better aligns with the reality of predicting human action and implicitly reflects some of these sources of error. If we hope to continue improving risk classifications in the future, it will be important to

[8] Other sources that may be important to this issue include measurement error, category definitions, and shifting cultural meanings of crime, race, and gender.

develop new strategies for accounting for these other sources of error. Bayesian statistics may hold great promise for doing this.

## Implications for Theory

### *Specialization vs. Versatility*

Taking the development of the BARR as an opportunity to use a large dataset of convicted offenders to study patterns of offense-specific recidivism, I found evidence of crime specialization among repeat property and drug offenders. These findings support some of the basic tenants of self-control theory (Gottfredson & Hirschi, 1990; Hirschi & Gottfredson, 1995; 2008) and the taxonomic theory of offending (Moffitt, 1993; 2015; Moffitt & Caspi, 2001), which together imply that the most serious violent repeat offenders tend to have longer histories of criminality and engage in versatile offending behaviors. Indeed, I found that counts of juvenile and adult felony convictions were strongly predictive of violent recidivism, whereas counts of specific crimes (violent, property, and drug) added little additional value to these predictions. In fact, histories of property and drug convictions were associated with *reductions* in the odds of violent re-offense, suggesting that those who commit these types of crimes tend to not engage in violent recidivism. Conversely, previous convictions for the same offense type were strongly predictive of drug and property recidivism, with total counts of juvenile and adult felonies contributing little. This pattern of findings suggests that theories of offending that posit universal mechanisms for crime are flawed, as these theories would predict versatility of criminal behaviors (Piquero et al., 1999). In other words, it doesn't appear that all

individuals have an underlying propensity for crime that makes the type of behavior exhibited inconsequential. Instead, these types of theories may be more helpful for understanding violent crime, but less useful for explaining repeated property and drug offending. Future research should focus on integrating taxonomic and self-control theories of offending with state dependence and other, more cultural perspectives, to attempt to better understand the roots of criminal careers devoted to property and drug crime.

*Intersectionality and Racial Formation*

Using the additional power inherent in Bayesian methods, I also sought to integrate the critiques of intersectional and pathways theorists (Belknap, 2007; Brennan et al., 2012; Chesney-Lind, 1997; Daly, 1992, 1994; Potter, 2015; Rios, 2011; Salisbury & Van Voorhis, 2009) by developing risk assessment models that examine whether race and gender-specific predictors might result in better classification than gender and race-neutral approaches. Overall, these models revealed a complex and disheartening reality of risk assessment and race, while confirming some basic tenants of intersectional theory. As expected, there was a great deal of variation in which factors predicted recidivism across black and white men and women. Even some of the most fundamental characteristics invoked by risk assessment demonstrated very different relationships among black men and women, with age not differentiating recidivists at all in black women's models and prior felony convictions not showing a strong gradation of risk in black men's models. Despite these differences, gender and race-specific models were roundly outperformed by the neutral BARR. However, both modeling strategies revealed that instruments composed

entirely of static measures have a difficult time correctly identifying black men who do not recidivate. This was more explicit in the race and gender-specific models, where nearly all young black men were classified as likely recidivists, but the BARR also exhibited somewhat lower negative predictive values for black men than other groups.

This has considerable implications for how we view risk assessment in the context of the greater criminal justice system. The correctional system has become an important setting for the creation and reinforcement of our cultural understandings of race and gender (Alexander, 2010; Omi & Winant, 1994; Richie, 2012). While these understandings are certainly stoked by media depictions of black criminality (Gilliam et al., 1996; Oliver, 2003; Smiley & Fakunle, 2016; Welch, 2007), the criminal justice system has become the quintessential racial project, or place that connects "what race *means* in a particular discursive practice and the ways in which both social structures and everyday experiences are racially *organized,* based on that meaning" (Omi & Winant, 1994; pp. 56). As an organizational tool of this system, I argue that risk assessment instruments may constitute racial projects in their own right. Because these tools exhibit more difficulty in correctly identifying black male non-recidivists, a greater share of black men will be subject to somewhat harsher punishment and increased surveillance, both of which cause increased recidivism (Mears, Cochran, & Bales, 2012; Mitchell et al., 2017). This may lead to a self-fulfilling prophecy, wherein cultural expectations of black men being more dangerous are reinforced by higher risk scores and harsher punishment, which in turn result in greater criminal involvement.

While the race and gender-specific assessments were much more likely to instill these detrimental impacts, the BARR also displayed some evidence of this tendency to

overclassify black male risk[9]. Others have investigated predictive properties of some widely used risk assessments by race or by gender (e.g. Desemarais & Singh, 2013), but this was the first study to examine the validity of classifications by intersections of both. I also reported additional accuracy statistics (positive and negative predictive values) that many studies fail to display, which directly correspond to these points of concern. Investigations of overall predictive validity may miss these important sources of error[10], just as failing to disaggregate findings by race and gender obscures this issue of overestimating black male risk. It is important that developers of new instruments check to see how their models perform over any and all prominent racial/ethnic and gender subgroups in their data, including on underreported dimensions like positive and negative predictive values, so that we become aware of the potential implications of using these tools in the real world.

***Limitations***

Many risk assessment systems now include static and dynamic characteristics to allow practitioners the ability to use assessment for both risk classification and program allocation (e.g. COMPAS, LSI-R, STRONG-R). While I focused on static factors, which enables simpler comparison of differences in predictive validity due to methodological variation (Hamilton et al., 2015), there is greater practical applicability to combined instruments. Furthermore, by introducing more potential sources of differentiation between offenders, it is possible that Bayesian methods may display even more promise for risk *and* needs

[9] As did the frequentist instrument that I compared with the BARR in Chapter 3 (results available upon request).
[10] For example, overall AUCs for black men were good in both the race and gender-specific and neutral models.

assessment than for risk classification, alone. This may be particularly true when it comes to reducing racial overclassification of risk, as instruments composed exclusively of static factors tend to run the greatest risk of exacerbating black male risk (Skeem & Lowenkamp, 2016). I intend to apply the same methods used here to WADOC's combined SRA and ONG data to re-investigate many of the same hypotheses examined in this dissertation. Dynamic factors better capture many of the theoretical mechanisms related to intersectionality and pathways, including possible "hooks for change" (Giordano, Cernovich, & Rudolph, 2002) that I was unable to tap into in this study. After all, desistance from crime is certainly related to recidivism and an important concept in gendered and raced pathways. As such, it is reasonable to assume information about dynamic risks and needs could be useful for improving the BARR, overall, but these added measures are especially important for more thoroughly investigating race and gender-specific assessment strategies.

WADOC data only contain information about inmate race, which precluded me from assessing predictive validity among Latino/as, who are likely included in the present data as white men and women. This is a significant limitation, particularly for my examination of intersectional assessments, because pathways to crime and desistance also differ by ethnicity (Harris & Feldmeyer, 2015; Lopez & Chesney-Lind, 2014; Lopez & Nuño, 2016). As a result, I would expect for the ideal choice of predictors and item weights to vary further for Latinos and Latinas, as they did for black and white men and women. However, it is unclear whether this would result in the same kind of overclassification that I observed for young black men. Latino boys and young men do seem to be subject to much of the same over-policing and surveillance as black men (Rios, 2011). Hispanic men are also more likely than whites to experience parole failure (Steinmetz & Henderson, 2015), and

Hispanic defendants tend to receive harsher prison sentences more like black defendants than whites (Demuth & Steffensmeier, 2004). This would seem to indicate that issues with classifying black men would also extend to Latinos, but a more thorough investigation is necessary with a sufficient sample.

Another possible limitation of this study relates to the use of correctional data from Washington State. While this is one of the largest samples ever used to construct or validate a risk assessment instrument (Hamilton et al., 2015), Washington is a unique criminological setting with a strong history of progressive politics and liberal drug policy. Changes in drug policies throughout the final construction cohort (roughly 2001-2007) and during the recidivism window of the validation sample (2008-2011) were minimal, with medical (1998) and recreational marijuana (2012) laws passed just outside of this period. However, changes to Washington's Drug Offender Sentencing Alternative (DOSA) law in 2005 diverted many low-risk drug offenders from prisons into chemical dependency treatment (Aos, Phipps, & Barnoski, 2005). It is possible that this important change may have made it more difficult to predict drug recidivism in the validation sample, where punishments may not have been as severe, given that most offender characteristics used to construct the instrument were taken from the proceeding period. However, the validity of drug recidivism predictions compared to felony, violent, and property ones was broadly comparable, if somewhat lower, suggesting that these changes did not make a huge difference.

Perhaps more important is Washington's relatively small black population and low police-to-citizen ratio (Reaves, 2011), and the state's history of highly disproportionate minority incarceration (Christianson, 1980). These factors limit the generalizability of the

BARR's item weights for other offender populations, but an important strength of the Bayesian method is the ability to adapt to new data. By applying these item weights as priors to another state's data, the instrument should be easily normed to a new population without the need to repeat the entire development process. A next step for the continuation of this project would be to see how well the instrument adapts to new settings, or additional cohorts of the Washington State data, as adaptability is a huge potential benefit of the BARR.

Finally, it is important to note that in mixing practical and theoretical aims, the modeling strategy and discussion surrounding the BARR's construction and validation fails to optimize either the ability to explain (i.e. theoretically explain) or predict (i.e. best classify offender risk in an applied, criminal justice setting) recidivism. While I contend that the bridging of these two bodies of literature was an important exercise that was missing from the literature, it is true that this approach limits the ability of my findings to deeply inform theories of specialization/versatility and intersectionality as well as preventing the BARR from being constructed in such a way that it can be used immediately in a practical setting. Integrating dynamic factors into the BARR models is an important next step for addressing both of these concerns. Dynamic factors may help to better test the wide range of explanatory mechanisms that potentially underlie the specialization observed among property and drug offenders in this sample, which could also help better distinguish these types of recidivists from those who exhibited more versatile offending behaviors and engaged in violent recidivism. For example, I only had the ability to test explanations representing population heterogeneity theories (Moffitt's taxonomy and Gottfredson and Hirschi's self control theories), despite the fact that state dependence theories (e.g. Nagin &

Paternoster, 1991; 2000) may be better suited to this task. These dynamic measures would also likely increase the predictive potential of the BARR, and some aspects of the construction process that I followed here (such as mixing continuous and categorical predictors), while useful for better representing theoretical concepts and empirical patterns in the data, are uncommon in risk assessment validation studies and may prove confusing to practitioners. As such, more work is necessary to adapt the BARR before policymakers are likely to use it in applied settings.

## Conclusion

Risk assessment, with its ability to reduce recidivism and correctional costs (when effective), is one of the most promising avenues we currently have for large-scale bipartisan criminal justice reform. I used Bayesian statistics to develop and validate a new risk assessment instrument for Washington State, the BARR, which performs as well as the best existing strategies and requires only administrative data to complete. This analytic strategy harnesses strengths of both traditional regression-based and machine learning models to arrive at direct estimates of the probability and uncertainty we have of offenders recidivating across four outcomes. By being interpretable, easy to complete, and accurate, the BARR holds great promise for use in practical settings.

However, as recent studies have made clear (Berk et al., 2017; Harcourt, 2015; Skeen & Lowenkamp, 2016) there is an urgent need to balance predictive accuracy with fairness in these models. By failing to investigate how risk assessment tools perform across all measurable racial/ethnic and gender subgroups, current practices may simply be

recreating existing racial disparities in biased policing practices (Chambliss, 1994; Rios, 2011), pretrial rulings (Demuth, 2003; Freiburger & Hilinski, 2010; Lowenkamp, VanNostrand, & Holsinger, 2013), and sentencing (Freiburger & Hilinski, 2013; Steen, Engen, & Gainey, 2005; Steffensmeier & Demuth, 2006; Ulmer, Painter-Davis & Tinik, 2014). As the first to examine the predictive validity of separate race and gender-specific instruments and investigate the accuracy of a gender and race-neutral tool for predicting recidivism at the intersections of race and gender, I was surprised to find that the gender and race-neutral BARR best optimized accuracy *and* fairness. Despite this, concerns still remain about the ability of risk assessment, in general, to correctly identify non-recidivists among groups that exhibit high rates of recidivism, like young black men. It may be that risk assessment is a racial project that ultimately perpetuates cultural understandings of young black men as criminal, but further investigation in other data is necessary to confirm this. Nonetheless, researchers should be aware of this possibility, and work with practitioners to reduce the chances that young black men will automatically be labeled as likely recidivists, through the incorporation of mitigating circumstances or otherwise.

These results also highlight the importance of assigning different item weights for offense-specific outcomes. For violent crime, a kitchen-sink approach to past offending behavior seems to work well, supporting the idea that the most important predictors of recidivism are the length and magnitude of criminal involvement. However, for property and drug crimes, models indicated that histories of prior offending of the same type were much more important, suggesting that these criminals tend to exhibit some specialization in offending behavior. More work should be devoted to integrating generalist theories with

explanations that allow for the development of offense specialization to better capture this reality.

Most of all, researchers interested in practical issues surrounding risk assessment development should not remain so siloed from those who apply theoretical critiques to their methods. Risk assessment development and validation studies hold great promise for making important theoretical contributions, especially when models are disaggregated by offense type, race/ethnicity, and gender. Likewise, I encourage the intersectional and pathways theorists and critical criminologists who have consistently pushed the field forward to collaborate with those scholars who develop these instruments. This collaboration would likely result in much more thorough tests of intersectional theories than I conducted here, and researchers may have new ideas for improving the usefulness and fairness of risk and needs assessment instruments. Ultimately, this kind of work will help ensure that risk assessment succeeds as criminal justice reform for everyone, rather than simply benefitting some and disadvantaging those whom the system has traditionally maligned.

BIBLIOGRAPHY

Agnew, R. (2006). *Pressured into Crime: An Overview of General Strain Theory.* Los Angeles, CA: Anderson.

Akers, R. L. (2009). *Social Learning and Social Structure: A General Theory of Crime and Deviance.* New Brunswick, NJ: Transaction.

Alarid, L. F., Burton, V. S. Jr., & Cullen, F. T. (2000). Gender and crime among felony offenders: Assessing the generality of social control and differential association theories. *Journal of Research in Crime and Delinquency, 37*(2), 171-199.

Alexander, M. (2010). *The New Jim Crow: Mass Incarceration in the Age of Colorblindness.* New York, NY: New Press.

Alpert, G. P., MacDonald, J. M., & Dunham, R. G. (2005). Police suspicion and discretionary decision making during citizen stops. *Criminology, 43*(2), 407-434.

Anderson, E. (2000). *Code of the Street: Decency, Violence and the Moral Life of the Inner City.* W. W. New York, NY: Norton & Company.

Andrews, D. A. (1982). *The Level of Supervision Inventory (LSI): The First Follow-Up.* Ontario Ministry of Correctional Services, Toronto.

Andrews, D. A., & Bonta, J. (1995). *The Level of Service Inventory- Revised.* Multi-Health Systems, Toronto.

Andrews, D. A., & Bonta, J. (1998). *The Psychology of Criminal Conduct (2nd Edition).* Cincinnati, OH: Anderson.

Andrews, D. A., Bonta, J., & Wormith, S. J. (2004). *The Level of Service/Case Management Inventory (LS/CMI).* Toronto, Canada: Mutli-Health Systems.

Andrews, D. A., Dowden, C., & Gendreau, P. (1999). *Clinically Relevant and Psychologically Informed Approaches to Reduced Re-Offending: A Meta-Analytic Study of Human Service, Risk, Need, Responsivity, and Other Concerns in Justice Contexts*. Calreton University, Ottawa.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica.* Retreived from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Aos, S., Phipps, P., & Barnoski, R. (2005). *Washington's Drug Offender Sentencing Alternative: An Evaluation of Benefits and Costs.* Olympia, WA: Washington State Institute for Public Policy.

Armstrong, T. A. (2008). Are trends in specialization across arrests explained by changes in specialization occurring with age? *Justice Quarterly, 25*(1), 201-222.

Arnold, J., & Arnold, L. (2015). Fixing justice in America. *Politico Magazine.* Retreived from https://www.politico.com/magazine/story/2015/03/criminal-justice-reform-coalition-for-public-safety-116057.

Arrigo, B. A. (2016). Critical criminology as academic activism: On praxis and pedagogy, resistance and revolution. *Critical Criminology, 24*, 469-471.

Austin, J. (2004). The proper and improper use of risk assessment in corrections. *Federal Sentencing Reporter FSR, 16*(3), 194-199.

Austin, J., & Allen, R. (2016). *Development of the Nevada Pretrial Risk Assessment System: Final Report.* Carson City, Nevada: Supreme Court of Nevada Judicial Council Committee to Study Evidence-based Pretrail Release.

Baird, C. (2009). *A Question of Evidence: A Critique of Risk Assessment Models Used in the Justice System*. Washington, D.C.: Focus- Views from the National Council on Crime and Delinquency. Retreived from http://cjjr.georgetown.edu/pdfs/ebp/baird2009_QuestionOfEvidence.pdf.

Baird, S., Heinz, R. C., & Bemus, B. J. (1979). *The Wisconsin Case Classification/Staff Deployment Project: Two-Year Follow-Up Report.* Madison, WI: Wisconsin Division of Corrections.

Barbaree, H. E., Seto, M. C., Langton, C. M., & Peacock, E. J. (2001). Evaluating the predictive accuracy of six risk assessment instruments for adult sex offenders. *Criminal Justice and Behavior, 28*(4), 490-521.

Barnoski, R. & Aos, S. (2003). *Washington's Offender Accountability Act: An Analysis of the Department of Corrections' Risk Assessment*. Olympia, WA: Washington State Institute for Public Policy, Document No. 03-12-1202.

Barnoski, R., & Drake, E. K. (2007). *Washington's Offender Accountability Act: Department of Corrections' Static Risk Assessment*. Olympia, WA: Washington State Institute for Public Policy.

Baskin-Sommers, A., Baskin, D. R., Sommers, I. B., & Newman, J. P. (2013). The intersectionality of sex, race, and psychopathology in predicting violent crimes. *Criminal Justice and Behavior, 40*(10), 1068-1091.

Belknap, J. (2007). *The Invisible Woman: Gender, Crime, and Justice (3rd Ed.).* Belmont, CA: Thompson Wadsworth.

Belknap, J., & Holsinger, K. (2006). The gendered nature of risk factors for delinquency. *Feminist Criminology, 1*(1), 48-71.

Belknap, J. (2015). Criminologists' responsibility to advocate for social and legal justice. *Criminology, 53*(1), 1-22.

Bell, K. E. (2013). Young adult offending: Intersectionality of gender and race. *Critical Criminology, 21*, 103-121.

Bennett, M. W., & Plaut, V. C. (2018). Looking criminal and the presumption of dangerousness: Afrocentric facial features, skin tone, and criminal justice. *UC Davis Law Review, 51*(3), 745-803.

Berk, R. (2017). An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology, 13,* 193-216.

Berk, R., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior. *Criminal Justice Policy Review*, *12*, 513-544.

Berk, R. A., Campbell, A., Klap, R., & Western, B. (1992). A Bayesian analysis of the Colorado Springs spouse abuse experiemnt. *The Journal of Criminal Law and Criminology, 83*(1), 170-200.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art. Working paper 2017-1.0, University of Pennsylvania, Department of Criminology. Retrieved from https://arxiv.org/abs/1703.09207.

Berk, R., Sherman, L., Barnes, G., Kurtz, E., & Ahlman, L. (2009). Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. *Journal of the Royal Statistical Society, 172*(1), 191–211.

Berry, D. A. (1991). Bayesian methods in phase III trials. *Drug Information Journal, 25*, 345-368.

Blanchette, K., & Brown, S. L. (2006). *The Assessment and Treatment of Women Offenders: An integrative Perspective.* New York, NY: Wiley and Sons.

Blumstein, A., Cohen, J., Das, S., & Moitra, S. (1988). Specialization and seriousness during adult criminal careers." *Journal of Quantitative Criminology, 4,* 303-345.

Bonta, J. (2002). Offender risk assessment: Guidelines for selection and use. *Criminal Justice and Behavior, 29*(4), 355-379.

Bonta, J., Law, M., & Hanson, R. K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders. Psychological Bulletin, *123*, 123-142.

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient. *PLoS One, 12*(6), e0177678.

Bowker, G. C., & Star, S. L. (1999). *Sorting Things Out: Classification and Its Consequences.* Cambrige, MA: MIT Press.

Braga, A. A. (2001). The effects of Hot Spots policing on crime. *The Annals of the American Academy of Political and Social Science, 578*(1), 104-125.

Braga, A. A. (2007). Effects of hot spots policing on crime. *Campbell Systematic Reviews.* Retrieved from http://www.aic.gov.au/campbellcj/reviews/titles.html.

Braga, A. A., Papachristos, A. V., & Hureau, D. M. (2014). The effects of Hot Spots policing on crime: An updated systematic review and meta-analysis. *Justice Quarterly, 31*(4), 633-663.

Brambila, N. C. (2016, September 28). Recidivism risk assessment as sentencing tool is controversial. *Reading Eagle*. Retreived from http://www.readingeagle.com/news/article/recidivism-risk-assessment-as-sentencing-tool-is-controversial.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199-231.

Brennan, P., Mednick, S., & John, R. (1989). Specialization in violence: Evidence of a criminal subgroup. *Criminology, 27*, 437-453.

Brennan, T., & Oliver, W. (2000). *Evaluation of Reliability and Validity of COMPAS Scales: National Aggregate Sample*. Traverse City, MI: Northpointe Institute for Public Management.

Brennan, T., Breitenbach, M., Dieterich, W., Salisbury, E. J., & Van Voorhis, P. (2012). Women's pathways to serious and habitual crime: A person-centered analysis incorporating gender responsive factors. *Criminal Justice and Behavior, 39*(11), 1481-1508.

Brennan, T., Wells, D., & Demory, R. (2004). *Classification Implementation Manual for Smaller Jails.* Traverse City, MI: Northpointe Institute for Public Management. Retrieved from http://www.northpointeinc.com/files/publications/Smalls_Jails_Manual_Complete.pdf.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluation of the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior, 36*(1), 21-40.

Britt, C. L. (1996). The measurement of specialization and escalation in the criminal career: An alternative modeling strategy. *Journal of Quantitative Criminology, 12,* 193-222.

Broidy, L. M., & Agnew, R. (1997). Gender and crime: A general strain theory perspective. *Research in Crime and Delinquency, 34*, 275-306.

Buck v. Davis, 137 S. Ct. 759 (2017).

Bunn, R. (2018). Intersectional needs and reentry: Re-conceptualizing 'multiple and complex needs' post-release. *Criminology and Criminal Justice*. Retrieved from http://journals.sagepub.com/doi/abs/10.1177/1748895817751828.

Bureau of Justice Statistics. (2007). County-based and local public defender offices, 2007. Prepared by Donald J. Farole Jr. and Lynn Langton, BJS Statisticians. Retreived from https://www.bjs.gov/content/pub/pdf/clpdo07.pdf.

Bureau of Justice Statistics. (2014). Correctional populations in the United States, 2013. Prepared by Lauren E. Glaze and Danielle Kaeble, BJS statisticians. Retrieved from https://www.bjs.gov/index.cfm?ty=pbdetail&iid=5177.

Bureau of Justice Statistics. (2016). Correctional populations in the United States, 2015. Prepared by Danielle Kaeble and Lauren E. Glaze, BJS statisticians. Retrieved from https://www.bjs.gov/content/pub/pdf/cpus15.pdf.

Burgess, E. W. (1928). Factors determining success or failure on parole. In A. A. Bruce, E. W. Burgess, J. Landesco & A. J. Harno (Eds.), *The Workings of the Indeterminate Sentence Law and the Parole System in Illinois.* Springfield, IL: Illinois State Board of Parole.

Burgess-Proctor, A. (2006). Intersectionality of race, class, gender, and crime. *Feminist Criminology, 1*(1), 27–47.

Catalano, R. F., & Hawkins, J. D. (1996). The social development model: A theory of antisocial behavior. In J. D. Hawkins (Ed.), *Delinquency and Crime: Current Theories.* Cambridge, MA; Cambridge University Press.

Caulkins, J., Cohen, J., Gorr, W., & Wei, J. (1996). Predicting criminal recidivism: A comparison of neural networks with statistical models. *Journal of Criminal Justice, 24*(3), 227-240.

Cernkovich, S. A., Giordano, P. C. & Rudolph, J. L. (2000). Race, crime, and the American dream. *Journal of Research in Crime and Delinquency, 37*(2), 131-170.

Chambliss, W. J. (1994). Policing the ghetto underclass: The politics of law and law enforcement. *Social Problems, 41*(2), 177-194.

Chen, M. H., Manatunga, A. K., & Williams, C. J. (1998). Heritability estimates from human twin data by incorporating historical prior information. *Biometrics, 54*, 1348-1362.

Chesney-Lind, M. (1997). *The Female Offender: Girls, Women, and Crime*. Thousand Oaks, CA: Sage.

Chiricos, T. G., & Crawford, C. (1995). Race and imprisonment: A contextual assessment of the evidence. In D. F. Hawkins (Ed.), *Ethnicity, Race, and Crime: Perspectives Across Time and Place*. Albany, NY: State University of New York Press.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Working paper. Retrieved from https://arxiv.org/pdf/1703.00056.pdf.

Choy, O., Raine, A., Venables, P. H., & Farrington, D. P. (2017). Explaining the gender gap in crime: The role of heart rate. *Criminology, 55*(2), 465-487.

Christianson, Scott. (1980). Corrections law developments: Racial disparities and prison confinement—A follow-up. *Criminal Law Bulletin, 16*(6), 616-621.

Citron, D. (2016). (Un)fairness of risk scores in criminal sentencing. *Forbes.* Retrieved from https: //www.forbes.com/sites/daniellecitron/ 2016/07/13/unfairness-of-risk-scores- in-criminal-sentencing/.

Cloward, R. A., & Ohlin, L. E. (1960). *Delinquency and Opportunity: A Study of Delinquent Gangs.* Abringdon-on-Thames, United Kingdom: Routledge.

Cohn, Nate. (2017-06-19). The 15 best-educated districts in the U.S., and why It matters in the Georgia race. *The New York Times.* Retrieved from https://www.nytimes.com/2017/06/19/upshot/this-list-of-well-educated-districts-explains-why-georgias-election-is-close.html.

Collins, P. H. (2002). *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment.* Abringdon-on-Thames, United Kingdom: Routledge.

Colorado Division of Criminal Justice. (2008). Colorado Actuarial Risk Assessment Scale (CARAS Version 5, 2008) parole guidelines. C.R.S. 12-22.50404.5(b)(d)(e)(f). Retrieved from https://www.colorado.gov/pacific/sites/default/files/CARAS_V5_1.pdf.

Cording, J. R., Christofferson, S. M. B., & Grace, R. C. (2016). Challenges for the theory and application of dynamic risk factors. *Psychology, Crime & Law, 22*(1), 84-103.

Cornish, D. B., & Clarke, R. V. (1987). Understanding crime displacement: An application of rational choice theory. *Criminology, 25*(4), 933-948.

Cotton, A. (2015, April 30). Denver police launching 'game changer' in data-driven crime fighting. *The Denver Post.* Retrieved from http://www.denverpost.com/2015/04/30/denver-police-launching-game-changer-in-data-driven-crime-fighting/.

Crenshaw, K. (1989). Antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum, 8*(1), 139-167.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics and violence against women of color. *Stanford Law Review, 43*, 1241-1245.

Daly, K. (1992). Women's pathway to felony court: Feminist theories of law-breaking and problems of representation. *Review of Law and Women's Studies, 2*(1), 11-52.

Daly, K. (1994). *Gender, Crime, and Punishment.* New Haven, CT: Yale University Press.

Daly, K., & Tonry, M. (1997). Gender, race, and sentencing. In M. Tonry (Ed.), *Crime and Justice: An Annual Review of Research*. Chicago, IL: University of Chicago.

Dehart, D. D. (2008). Pathways to prison: Impact of victimization in the lives of incarcerated women. *Violence Against Women, 14*(12), 1362-1381.

DeMichele, M. T. (2007). Probation and parole's growing caseloads and workload allocation: Strategies for managerial decision making. The American Probation & Parole Association. Retrieved from https://www.appa-net.org/eweb/docs/appa/pubs/SMDM.pdf.

Demuth, S. (2003). Racial and ethnic differences in pretrial release decisions and outcomes: A comparison of Hispanic, black, and white felony arrestees. *Criminology, 41*, 873-895.

Demuth, S., & Steffensmeier, D. (2004). Ethnicity effects on sentence outcomes in large urban courts: Comparisons among white, black, and Hispanic defendants. *Social Science Quarterly, 85*(4), 994-1011.

Desemarais, S. L., & Singh, J. P. (2013). Risk assessment instruments validated and implemented in correctional settings in the United States. Report for the Council of State Governments Justice Center. Retrieved from https://csgjusticecenter.org/wp-content/uploads/2014/07/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf.

Deslauriers-Varin, N., Lussier, P., & Tzoumakis, S. (2016). Crime specialization. *Oxford Handbooks Online.* Retrieved from http://www.oxford handbooks.com/view/10.1093/oxfordhb/9780199935383.001.0001/oxfordhb-9780199935383-e-27.

Doude, S. B. (2014). Masculinity and crime. In *The Encyclopedia of Theoretical Criminology*, J. M. Miller (Ed.).

Drake, E. K., & Barnoski, R. (2009). *New Risk Instrument for Offenders Improves Classification Decisions.* Olympia, WA: Washington State Institute for Public Policy, Document No. 09-03-1201.

Drake, E. K., Aos, S., & Barnoski, R. (2010). *Washington's Offender Accountability Act: Final Report on Recidivism Outcomes.* Olympia, WA: Washington State Institute for Public Policy, Document No. 10-01-1201.

Duwe, G. (2014). The development, validity, and reliability of the Minnesota screening tool for assessing recidivism risk (MnSTARR). *Criminal Justice Policy Review, 25*(5), 579-613.

Duwe, G., & Kim, K. (2016). Sacrificing accuracy for transparency in recidivism risk assessment: The impact of classification method on predictive performance. *Corrections, 1*(3), 155-176.

Eisenburg, M., Bryle, J., & Fabelo, T. (2009). *Validation of the Wisconsin Department of Correction Risk Assessment Instrument.* New York, NY: Council of State Governments Justice Center.

Else-Quest, N. M., Higgins, A., Allison, C., & Morton, L. (2012). Gender differences in self-conscious emotional experience: A meta-analysis. *Psychological Bulletin*, *138*, 947-981.

Fagan, J., & Davies, G. (2000). Street stops and broken windows: Terry, race, and disorder in New York City. *Fordham Urban Law Review, 28*, 457-504.

Farrington, D. P. (1986). Age and crime. *Crime and Justice, 7,* 189-250.

Farrington, D. P., Snyder, H. N., & Finnegan, T. A. (1988). Specialization in Juvenile Court Careers. *Criminology, 26*, 461-487.

Farrington, D. P. (1991). Antisocial Personality from Childhood to Adulthood. *The Psychologist, 4*, 389–394.

Fazel, S., Singh, J. P., Helen, D., Martin, G. (2012). Use of risk assessment instruments to predict violence and antisocial behavior in 73 samples involving 24,827 people: Systematic review and meta-analysis. *BMJ, 345*, e4692.

Fenton, N., & Neil, M. (2011). The use of Bayes and causal modeling in decision making, uncertainty and risk." *CEPIS UPgrade, 12*(5), 10-21.

Fenton, N., & Neil, M. (2012). *Risk Assessment and Decision Analysis with Bayesian Networks.* Boca Raton, FL: CRC Press.

Fenton, N., Neil, M., & Berger, D. (2016). Bayes and the Law. *Annual Review of Statistical Applications, 3*(1), 51-77.

Finkeldey, J. G. (2014). The influence of skin color on the likelihood of experiencing arrest in adulthood. Unpublished masters thesis, Bowling Green University. Retrieved from https://etd.ohiolink.edu/!etd.send_file?accession=bgsu1403293558&disposition=inline.

Finzen, M. E. (2005). Systems of oppressions: The collateral consequences of incarceration and their effects on black communities. *Georgetown Journal on Poverty Law and Policy, 12*(2)*,* 299-324.

Flores, A., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoiner to 'Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks.' *Federal Probation, 80*(2), 38-46.

Frase, R. (2009). What explains persistent racial disproportionality in Minnesota's prison and jail populations? In Tonry, M (Ed.) *Crime and Justice, Vol. 38.* Chicago, IL: University of Chicago Press.

Frase, R. (2014). Recurring policy issues of guidelines (and non-guidelines) sentencing: Risk assessments, criminal history enhancements, and the enforcement of release conditions. *Federal Sentencing Reporter, 26*, 145-157.

Freed, D. J. (1992). Federal sentencing in the wake of guidelines: Unacceptable limits on the discretion of sentencers. *The Yale Law Journal, 101*(8), 1681-1754.

Freiburger, T. L., & Hilinski, C. M. (2013). A examination of the interactions of race and gender on sentencing decisions using a trichotomous dependent variable. *Crime and Delinquency, 59*(1), 59-86.

Garland, D. (1990). *Punishment and Modern Society.* Chicago, IL: University of Chicago Press.

Garland, D. (2001). *The Culture of Control: Crime and Social Order in Contemporary Society.* Chicago, IL: Univeristy of Chicago Press.

Gavazzi, S. M., Yarcheck, C. M., & Lim, J. Y. (2006). Ethnicity, gender, and global risk indicators in the lives of status offenders coming to the attention of the juvenile court. *International Journal of Offender Therapy and Comparative Criminology, 49*, 696-710.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis, 3*(3), 445-450.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis (3rd Edition).* Boca Raton, FL: Chapman and Hall/CRC Press.

Gill, J. (2014). *Bayesian Methods: A Social and Behavioral Sciences Approach, Third Edition*. Boca Raton, FL: CRC Press.

Giordano, P. C., Cernkovich, S. A., & Rudolph, J. L (2002). Gender, crime, and desistance: Toward a theory of cognitive transformation. *American Journal of Sociology, 107*(4), 990-1064.

Goffman, A. (2014). *On the Run: Fugitive Life in an American City (Fieldwork Encounters and Discoveries).* Chicago, IL: University of Chicago Press.

Goodman, P. (2008). 'It's just black, white, or Hispanic': An observational study of racializing moves in California's segregated prison reception centers. *Law and Society, 42*(4), 735-770.

Gottfredson, M., & Hirschi, T. (1986). The true value of lambda would appear to be sero: An essay on career criminals, criminal careers, selective incapacitation, cohort studies, and related topics. *Criminology, 24,* 213-233.

Gottfredson, M., & Hirschi, T. (1990). *A General Theory of Crime.* Palo Alto, CA: Stanford University Press.

Gottfredson, D. C., Kearley, B., Najaka, S. S., & Rocha, C. M. (2005). The Baltimore City drug treatment court: 3-year self-report outcome study. *Evaluation Review, 29*(1), 42-64.

Gottfredson, S. D. & Moriarty, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime and Delinquency, 52*(1), 178-200.

Gottfredson, D.C., Najaka, S.S., Kearley, B., & Rocha, C.M. (2006). Long-term effects of participation in the Baltimore City drug treatment court: Results from an experimental study. *Journal of Experimental Criminology, 2*, 67-98.

Grasmick, H. G., Tittle, C. R., Bursik, R. J., & Arneklev, B. J. (1993). Testing the core empirical implications of Gottfredson and Hirschi's general theory of crime. *Journal of Research in Crime and Delinquency, 30*(1), 5-29.

Gray, K., Hampton, B., Silveti-Falls, T., McConnell, A., & Bausell, C. (2015). Comparison of Bayesian credible intervals to frequentist confidence intervals. *Journal of Modern Applied Statistical Methods, 14*(1), 43-52.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19-30.

Haas, H., & Cusson, M. (2015). Comparing theories' performance in predicting violence. *International Journal of Law and Psychiatry, 38*(1), 75-83.

Hagan, F. E. (2012). *Introduction to Criminology: Theories, Methods, and Criminal Behavior.* Thousand Oaks, CA: SAGE Publishing.

Hall, P., & Gill, N. (2017). Debugging the black-box COMPAS risk assessment instrument to diagnose and remediate bias. Paper presented at the Workshop on Human Interpretability in Machine Learning, Sydney, Australia. Retrieved from https://openreview.net/forum?id=r1iWHVJ7Z&noteId=r1iWHVJ7Z.

Hamilton, Z. & Kigerl, A. (2016). Development and validation of the Nebraska Department of Correctional Services Prison Classification System. *Unviersity of Nebraska, Omaha Reports,* 14. Retreived from https://digitalcommons.unomaha.edu/cgi/viewcontent.cgi?article=1013&context=ncjrreports.

Hamilton, Z., Kigerl, A., Champagna, M., Barnoski, R., Lee, S., Wormer, J. V., & Block, L. (2016). Designed to fit: The development and validation of the STRONG-R recidivism risk assessment. *Criminal Justice and Behavior, 43*(2), 230-263.

Hamilton, Z., Campagna, M., Tollefsbol, E., Wormer, J .V., & Barnoski, R. (2017). A more consistent application of the RNR model: The STRONG-R needs assessment. *Criminal Justice and Behavior, 44*(2), 261-292.

Hamilton, Z., Malanie-Angela, N., Lee, S., & Barnoski, R. (2015). Isolating modeling effects in offender risk assessment. *Journal of Experimental Criminology, 11*, 299-318.

Hamilton, Z., & Wormer, J. (2015). *Customizing Offender Assessment.* Spokane, WA: Washington State Institute for Criminal Justice.

Hamilton, Z., Wormer, J., Kigerl, K., Campagna, M., Barnoski, R., Block, L., & Lee, S. (2014). *The STRONG-R Recidivism Risk Assessment*. Tumwater, WA: Washington State Institute for Public Policy.

Harcourt, B. E. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter, 27*, 237-248.

Hare, R. (1991). *The Revised Psychopathy Checklist*. Toronto, Ontario, Canada: Multi-Health Systems.

Harper, P. R. (2005). A review and comparison of classification algorithms for medical decision making. *Health Policy, 7*(3), 315-331.

Harris, C. T., & Feldmeyer, B. (2015). A shot of morality? Hispanic immigration, religious contextual characteristics, and violence. *Sociological Spectrum, 35*(3), 229-253.

Hayslett-McCall, K. L., & Bernard, T. J. (2002). Attachment, masculinity, and self control: A theory of male crime rates. *Theoretical Criminology, 6*(1), 5-33.

Haynie, D., & Armstrong, D. (2006). Race and gender-disaggregated homicide offending rates. *Homicide Studies, 10*(1), 3-32.

Healy, W., & Bronner, A. (1926). *Delinquents and Criminals: Their Making and Unmaking.* New York, NY: McMillan.

Heimer, K., Lauritsen, J. L., & Lynch, J. P. (2009). The national crime victimization survey and the gender gap in offending: Redux. *Criminology, 47*, 427-438.

Henderson, H. (2006). The predictive utility of the Wisconsin Risk Needs Assessment instrument in a sample of Texas probationers. *Dissertation Abstracts International, 68*(1), 1-95.

Henderson, H., Daniel, A., Adams, T., & Rembert, D. (2007). The predictive validity of the Wisconsin Risk Needs Assessment instrument in post-probation success. *International Journal of Crime, Criminal Justice, and Law, 2*, 95-103.

Hill-Collins, P. (1998a). Intersectionality of race, class, gender, and nation: Some implications for black family studies. *Journal of Comparative Family Studies, 29*(1), 27-36.

Hill-Collins, P. (1998b). *Fighting Words: Black Women and the Search for Justice.* Minneapolis, MN: University of Minnesota Press.

Hirschi, T., & Gottfredson, M. (1993). Commentary: Testing the general theory of crime. *Journal of Research in Crime and Delinquency, 16*(1), 35-38.

Hirschi, T., & Gottfredson, M. R. (2000). In defense of self-control. *Theoretical Criminology, 4*(1), 55-69.

Hirschi, T., & Gottfredson, M. R. (2008). Critiquing the critics: The authors respond. In E. Goode (Ed.), *Out of Control: Assessing the General Theory of Crime*. Palo Alto, CA: Stanford University Press.

Holder, E. (2014). Attorney General Eric Holder speaks at the National Association of Criminal Defense Lawyers 57th annual meeting. Retrieved from http://www.justice.gov/opa/speech/ attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers- 57th.

Holtfreter, K., & Cupp, R. (2007). Gender and risk assessment: The empirical status of the LSI-R for women." *Journal of Contemporary Criminal Justice, 23*(4), 363-382.

Horney, J., Osgood, W. D., & Marshall, H. I. (1995). Criminal careers in the short-term: Intra-individual variability in crime and its relation to local life circumstances. *American Sociological Review, 60,* 655-673.

Hudson, B., & Bramhall, G. (2005). Assessing the 'other': Constructions of 'Asianness' in risk assessments by probation officers. *British Journal of Criminology, 45*, 721-740.

Huebner, B. M., & Bynum, T. S. (2008). The role of race and ethnicity in parole decisions. *Criminology, 46*(4), 907-938.

IBM Software Group. (2010). Crime prediction and prevention: A safer public through advanced analytics. *IBM White Paper Series.* Retrieved from ftp://ftp.software.ibm.com/software/data/sw-library/cognos/pdfs/whitepapers/wp_crime_prediction_and_prevention.pdf.

Ibrahim, J. G., & Chen, F. (2000). Power prior disctributions for regression models. *Statistical Science, 15,* 46-60.

Ibrahim, J. G., Chen, M. H., Gwon, Y., & Chen, F. (2015). The power prior: Theory and applications. *Statistics in Medicine, 34*(28), 3724-3749.

Ibrahim, J. G., Ryan, L. M., & Chen, M. H. (1998). Use of historical controls to adjust for covariates in trend tests for binary data. *Journal of the American Statistical Association, 93*, 1282-1293.

John Howard Society of Alberta. (2000). Offender risk assessment. Retrieved from http://www.johnhoward.ab.ca/pub/old/C21.htm.

Johnson, P. C. (1995). At the intersection of injustice: Experiences of African American women in crime and sentencing. *American University Journal of Gender, Social Policy and the Law, 4,* 1-6.

Jones, N. J., Brown, S. L., Wanamaker, K. A., & Greiner, L. E. (2014). A quantitative exploration of gendered pathways to crime in a sample of male and female juvenile offenders. *Feminist Criminology, 9*(2), 113-136.

Jones, D. A., Johnson, S., Latessa, E. J., & Travis, L. F. (1999). Case classification in community corrections: Preliminary findings from a national survey. Topics in Community Corrections, National Institute of Corrections, U.S. Department of Justice. Washington, D.C.

Kaplan, D. (2014). *Bayesian Statistics for the Social Sciences.* New York, NY: Guilford Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773-795.

Katz, R. S. (2000). Explaining girls' and women's crime and desistance in the context of their victimization experiences. *Violence Against Women, 6,* 633-660.

Kaufman, J. M., Rebellon, C., Thaxton, S., & Agnew, R. (2008). A general strain theory of racial differences in criminal offending. *Australian & New Zealand Journal of Criminology, 41*(3), 421-437.

Kearley, B. W. (2017) Long-term effects of drug court participation: Evidence from a 15-year follow-up of a randomized controlled trial (Unpublished doctoral disseration). University of Maryland, College Park.

Kehl, D. L., Guo, P., & Kessler, S. (2017). Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. Responsive Communities Initiative, Berkman Klien Center for Internet & Society, Harvard Law School. Retrieved from https://dash.harvard.edu/handle/1/33746041.

Kelly, D. L., & Smith, C. L. (2009). Bayesian inference in probabilistic risk assessment—The current state of the art. *Reliability Engineering and System Safety, 94*(1), 628-643.

Kempf, K. L. (1987). Specialization and the criminal career. *Criminology, 25,* 399-417.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 1-12.

Klein, M. W. (1984). Offense specialization and versatility among juveniles. *British Journal of Criminology, 24*, 185-194.

Kroner, D., Mills, J., & Reddon, J. (2005). A coffee can, factor analysis, and prediction of antisocial behavior: The structure of criminal risk. *International Journal of Law and Psychiatry*, *28*, 360-374.

Latessa, E. J., Smith, P., Lemke, R., Makarios, M., & Lowenkamp, C. (2009). *Creation and Validation of the Ohio Risk Assessment System: Final Report*. Cincinnati, OH: Ohio Department of Rehabilitation and Correction.

Laub, J. H., & Sampson, R. J. (1991). The Sutherland-Glueck debate: On the sociology of criminological knowledge. *American Journal of Sociology, 96*(6), 1402-1440.

Laub, J. H., & Sampson, R. J (1993). Turning points in the life course: Why change matters to the study of crime. *Criminology, 31*, 301-325.

Laub, J. H, & Sampson, R. J. (2006). *Shared Beginnings, Divergent Lives: Delinquent Boys to Age 70.* Cambridge, MA: Harvard University Press.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review, 73*(1), 31-43.

Leaw, J. N., Ang, R. P., Huan, V. S., Chan, W. T., & Cheong, S. A. (2015). Re-examining Moffitt's theory of delinquency through agent based modeling. *PLoS One, 10*(6), e0126752.

Leip, David. (2017). General Election Results – Washington. United States Election Atlas. Retrieved from https://uselectionatlas.org/RESULTS/.

Levitt, S. D. (1999). The limited role of changing age structure in explaining aggregate crime rates. *Criminology, 37*(3), 581-598.

Liu, Y. Y., Yang, M., Ramsey, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology, 27*(4), 547-573.

Loh, P. R., Tucker, G. Bulik-Sullivan, B. K., Vilhjálmsson, B. J., … & Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics, 47*, 284-290.

Loomis v. Wisconsin, 16-6387 U.S. Supreme Court (2017).

Lopez, V., & Chesney-Lind, M. (2014). Latina adolescent girls speak out: Stereotypes, traditional gender roles, and relationship dynamics. *Latino Studies, 12*, 527-549.

Lopez, V., & Nuño, L. (2016). Latina and African-American girls in the juvenile justice system: Needs, problems, and solutions. *Sociology Compass, 10*(1), 24-37.

Lowenkamp, C. T., VanNostrand, M., & Holsinger, A. (2013). Investigating the impact of pretrial detention on sentencing outcomes. The Arnold Foundation. Retrieved from http://www.arnoldfoundation.org/wp-content/uploads/2014/02/LJAF_Report_state-sentencing_FNL.pdf.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software, 42*(9), 1-21.

Maurutto, P., & Hannah-Moffat, K. (2005). Assembling risk and the restructuring of penal control. *British Journal of Criminology, 45*(1), 1-17.

Mazerolle, L., Rombouts, S., & McBroom, J. (2007). The impact of COMPSTAT on reported crime in Queensland. *Policing: An International Journal, 30*(2), 237-256.

McCall, L. (2005). The complexity of intersectionality. *Signs, 40*(1), 1771-1800.

McCoy, L. A., & Miller, H. A. (2013). Comparing gender across risk and recidivism in nonviolent offenders. *Women and Criminal Justice, 23*, 143-162.

McDonald, J. H. (2014). *Handbook of Biological Statistics (3rd Ed.).* Baltimore, MD: Sparky House Publishing.

McGee, T. R., & Mazarolle, P. (2016). Gendered experiences in developmental pathways to crime: Editorial ntroduction. *Journal of Development and Life-Course Criminology, 2*(3), 257-261.

Mears, D. P., Cochran, J. C., & Bales, W. D. (2012). Gender differences in the effects of prison on recidivism. *Journal of Criminal Justice, 40*(5), 370-378.

Meredith, T., Speir, J. C., & Johnson, S. (2007). Developing and implementing automated risk assessments in parole. *Justice Research and Policy, 9*(1), 1-24.

Messerschmidt, J. W. (1993). *Masculinities and Crime: Critique and Reconceptualization of Theory.* Boulder, CO: Rowman and Littlefield.

Mitchell, O., Cochran, J. C., Mears, D. P., & Bales, W. D. (2017). Examining prison effects on recidivism: A regression discontinuity approach. *Justice Quarterly, 34*(4), 571-596.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy." *Psychological Review, 100*(4), 674-701.

Moffitt, T. E. (1994). Natural histories of delinquency. In Weitekamp, E. and Kerner, H. (Eds.), *Cross-National Longitudinal Research on Human Development and Criminal Behavior*. Dordrecht, Netherlands: Kluwer.

Moffitt, T. E. (2015). Life-course-persistent versus adolescence-limited antisocial behavior. In *Developmental Psychopathology* (eds. D. Cicchetti and D. J. Cohen).

Moffitt, T. E., and Caspi, A. (2001). Childhood predictors differentiate life-course persistent and adolescence-limited antisocial pathways among males and females. *Development and Psychopathology, 13*(2), 355-375.

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology, 12*, 489-513.

Nagel, I. H. (1990). Structuring sentencing discretion: The new federal sentencing guidelines. *The Journal of Criminal Law and Criminology, 80*(4), 883-943.

Nagin, D. S., & Paternoster, R. (1991). On the relationship of past and future participation in delinquency. *Criminology, 29*(2), 163-189.

Nagin, D. S., & Paternoster, R. (2000). Population heterogeneity and state dependence: State of the evidence and directions for future research. *Journal of Quantitative Criminology, 16*(2), 117-144.

Nellis, A. (2016). The color of justice: Racial and ethnic disparity in state prisons. *The Sentencing Project.* Retrieved from http://www.sentencingproject.org/publications/color-of-justice-racial-and-ethnic-disparity-in-state-prisons/.

Neuilly, M. A., Zgoba, K. M., Tita, G. E., & Lee, S. (2011). Predicting recidivism in homicide offenders using classification tree analysis. *Homicide Studies, 34*(4), 271-284.

Newcombe, P. J., Reck, B. H., Sun, J., Platek, G. T., Verzilli, C., … & Xu, J. (2012). A comparison of Bayesian and frequentist approaches to incorporating external information for the prediction of prostate cancer risk. *Genetic Epidemiology, 36*(1), 71-83.

Ogino, S., & Wilson, R. B. (2004). Bayesian analysis and risk assessment in genetic counseling and testing. *Journal of Molecular Diagnostics, 6*(1), 1-9.

Olson, D. E., & Lurigio, A. J. (2000). Predicting probation outcomes: Factors associated with probation rearrest, revocations, and technical violations during supervision. *Justice Research and Policy, 2*(1), 73-86.

Omi, M., & Winant, H. (2014). *Racial Formation in the United States.* Abingdon, United Kingdom: Routledge.

Onifade, E., Davidson, W., Campbell, C., Turke, G., Malinowski, J., & Turner, K. (2008). Predicting recidivism in probationers with the Youth Level of Service/Case Management Inventory (YLS/CMI). *Criminal Justice and Behavior, 35*(4), 474-483.

O'Sullivan, Joseph. (2017-01-25). Abolish Washington's death penalty? Deep divide remains. *The Seattle Times.* Retrieved from https://www.seattletimes.com/seattle-news/politics/abolish-washingtons-death-penalty-deep-divide-remains/.

Pang, H., & Jung, S. H. (2013). Sample size considerations of prediction-validation methods in high-dimensional data for survival outcomes. *Genetic Epidemiology, 37*(3), 276-282.

Peguero, A. A., Popp, A. M., Latimore, T. L., Shekarkhar, Z., & Koo, D. J. (2010). Social control theory and school misbehavior: Examining the role of race and ethnicity. *Youth Violence and Juvenile Justice, 9*(3), 259-275.

Pernanen, K., Cousineau, M. M., Brochu, S., & Sun, F. (2002). *Proportions of Crimes Associated with Alcohol and Other Drugs in Canada.* Toronto, Canada: Canadian Centre on Substance Abuse.

Petersilia, J. (2006). *Understanding California Corrections.* Berkeley, CA: California Policy Research Center.

Pettit, B., & Western, B. (2004). Mass incarceration and the life course: Race and class inequality in U.S. incarceration. *American Sociological Review, 69*, 151-169.

Piquero, A. R., Paternoster, R., Mazerolle, P., Brame, R., & Dean, C. W. (1999). Onset age and offense specialization. *Journal of Research in Crime and Delinquency, 36*(3), 275-299.

Piquero, A. R., MacDonald, J., Dobrin, A., Daigle, L. E., & Cullen, F. T. (2005). Self-control, violent offending, and homicide victimization: Assessing the general theory of crime. *Journal of Quantitative Criminology, 21*(1), 55-71.

Philipse, H. (2015). Probability arguments in criminal law – Illustrated by the case of Lucia de Berk. *Utrecht Law Review, 11*(1), 19-32.

Potter, H. (2013). Intersectional criminology: Interrogating identity and power in criminological research and theory. *Critical Criminology, 21*(3), 305-318.

Potter, H. (2015). *Intersectionality and Criminology: Disrupting and Revolutionizing Studies of Crime.* Routledge: New York, New York.

Powers, D. M. (2011). Evaluation: From precision, recall, and F-Measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63.

Quetelet, A. (1984). *Research on the Propensity for Crime at Different Ages.* Cincinnati, OH: Anderson. Originally published in 1831.

Reaves, Brian A. (2011). Census of state and local law enforcement agencies, 2008. *Bureau of Justice Statistics.* Retrieved from https://www.bjs.gov/content/pub/pdf/csllea08.pdf.

Reisig, M. D., Holtfreter, K., & Morash, M. (2006). Assessing recidivism risk across female pathways to crime. *Justice Quarterly, 23*, 384-405.

Rempel, M., Green, M., & Kralstein, D. (2012). The impact of adult drug courts on crime and incarceration: findings from a multi-site quasi-experimental design. *Journal of Experimental Criminology,* 8(2): 165-192.

Rettinger, L. J., & Andrews, D. A. (2009). General risk and need, gender specificity, and the recidivism of female offenders. *Criminal Justice and Behavior, 37*(1), 29-46.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior, 29*(5), 615-620.

Richie, B. (2012). *Arrested Justice: Black Women, Violence, and America's Prison Nation.* New York, NY: NYU Press.

Rios, V. (2011). *Punished: Policing the Lives of Black and Latino Boys.* New York, NY: NYU Press.

Rosenfeld, R., Deckard, M. J., & Blackburn, E. (2014). The effects of directed patrol and self-initiated enforcement on firearm violence: A randomized controlled study of hot spot policing. *Criminology, 52*(3), 428-449.

Salisbury, E. J., & Van Voorhis, P. (2009). Gendered pathways: A quantitative investigation of women probationers' paths to incarceration. *Criminal Justice and Behavior, 36*(6), 541-566.

Salisbury, E. J., Van Voorhis, P., & Spiropoulos, G. (2009). The predictive validity of a gender-responsive needs assessment. *Crime & Delinquency*, *55*, 550-585.

Sampson, R. J., & Graves, W. B. (1989). Community structure and crimes: Testing social-disorganization theory. *American Journal of Sociology, 94*, 774-802.

Sampson, R. J., & Laub, J. H. (1990). Crime and deviance over the life course: The salience of adult social bonds. *American Sociological Review, 55*, 609-627.

Sampson, R. J., & Laub, J. H. (1992). Crime and deviance in the life course. *Annual Review of Sociology, 18*, 63-84.

Sampson, R. J., & Laub, J. H. (2005). A life course view of the development of crime. *The ANNALS of the American Academy of Political and Social Science, 601*(1), 12-45.

Schaffer, D., Kelly, B., & Lieberman, J. (2011). An exemplar-based approach to risk assessment: Validating the risk management systems instrument. *Criminal Justice Policy Review, 22*(2), 167-186.

Schwartz, J., Steffensmeier, D. J., Zhong, H., & Ackerman, J. (2009). Trends in the gender gap in violence: Reevaluating NCVS and other evidence. *Criminology, 47*, 401-425.

Shaw, C. R., & McKay, H. H. (1942). *Juvenile Delinquency in Urban Areas.* Chicago, IL: University of Chicago Press.

Simpson, S. (1991). Caste, class, and violent crime: Explaining differences in female offending. *Criminology, 29*, 115-135.

Silver, E., & Miller, L. (2002). A cautionary note on the use of actuarial risk assessment tools for social control. *Crime and Delinquency, 48*, 138-161.

Siu, N. O., & Kelly, D. L. (1998). Bayesian parameter estimation in probabilistic risk assessment. *Reliability Engineering and System Safety, 62*(1), 89-116.

Skeem, J., & Louden, J. (2007). *Assessment of Evidence on the Quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS).* Prepared for the California department of corrections and rehabilitation. Davis, CA: University of California Davis Press.

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology, 54*(4), 680-712.

Smith, P., Cullen, F. T., & Latessa, E. J. (2009). Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders." *Criminology and Public Policy, 8,* 183-208.

Spanos, A. (2012). Philosophy of economics. In Maki, U., Gabbay, D. M., Thagard, P., & Woods, J. (Eds.), *Philosophy of Economics.* Amsterdam, Netherlands: North Holland Publishing.

Starr, S. B. (2015). The new profiling: Why punishing based on poverty and identity is unconstitutional and wrong. *Federal Sentencing Reporter, 27*, 229-240.

Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Robbins, P. C., & Mulvey, E. P. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior, 24*, 83-100.

Steen, S., Engen, R., & Gainey, R. (2005). Images of danger and culpability: Racial stereotyping, case processing, and criminal sentencing. *Criminology, 43*, 435-468.

Steffensmeier, D., & Allan, E. (1996). Gender and crime: Toward a gendered theory of female offending. *Annual Review of Sociology, 22*, 459-487.

Steffensmeier, D., & Demuth, S. (2006). Does gender modify the effects of race-ethnicity on criminal sanctioning? Sentences for male and female, white, black, and Hispanic defendants. *Journal of Quantitative Criminology, 22,* 241-261.

Steffensmeier, D., Painter-Davis, N., & Ulmer, J. (2017). Intersectionality of race, ethnicity, gender, and age on criminal punishment. *Sociological Perspectives, 60*(4), 810-833.

Stegle, O., Parts, L., Durbin, R., & Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology.* Retrieved from http://journals.plos.org/ploscompbiol/article/related?id=10.1371/journal.pcbi.1000770.

Steinmetz, K. F., & Henderson, H. (2015). On the precipice of intersectionality: The influence of race, gender, and offense severity interactions on probation outcomes. *Criminal Justice Review, 40*(3), 361-377.

Stevenson, Reed. (2003-09-17). Seattle votes to make marijuana crime low priority. *Reuters.* Retrieved from http://www.yeson75.org/media/2003-09-17-reuters.php.

Subramanian, R., Moreno, R., & Broomhead, S. (2014). *Recalibrating Justice: A Review of 2013 State Sentencing and Corrections Trends*. New York, NY: Vera Institute of Justice. Retrieved from http://www.ajc.state.ak.us/acjc/sentencing%20reform/trends.pdf.

Subramanian, R.,Riley, K., & Mai, C. (2018). *Divided Justice: Trends in Black and White Jail Incarceration, 1990-2013*. New York, NY: Vera Institute of Justice. Retrieved from https://storage.googleapis.com/vera-web-assets/downloads/Publications/divided-justice-black-white-jail-incarceration/legacy_downloads/Divided-Justice-full-report.pdf.

Sullivan, C. J., McGloin, J. M., Pratt, T. C., & Piquero, A. R. (2006). Rethinking the 'norm' of offender generality: Investigating specialization in the short-term. *Criminology, 44,* 199-233.

Tashea, J. (2017). Courts are using AI to sentence criminals: That must stop now. *Wired.* Retrieved from https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/.

Tibbitts, C. (1931). Success or failure on parole can be predicted: A study of the records of 3,000 youth paroled from the Illinois State Reformatory. *American Institute of Criminal Law and Criminology, 22(*11), 11-50.

Tonry, M. H. (1995). *Malign Neglect: Race, Crime, and Punishment in America.* New York, NY: Oxford University Press.

Trulson, C., & Marquart, J. W. (2002). Racial desegregation and violence in the Texas prison system. *Criminal Justice Review, 27*(2), 233-255.

Trulson, C. R., Marquart, J. W., Hemmens, C., & Carroll, L. (2008). Racial desegregation in prisons. *Prison Journal, 88*(2), 270-299.

Turner, S., Hess, J., & Jannetta, J. (2009). Development of the California Static Risk Assessment Instrument (CSRA). *CEBC Working Paper.* Retrieved from http://ucicorrections.seweb.uci.edu/files/2009/11/CSRA-Working-Paper.pdf.

Ulmer, J. T., Painter-Davis, N., & Tinik, L. (2014). Disproportional imprisonment of black and Hispanic males: Sentencing discretion, processing outcomes, and policy structures. *Justice Quarterly, 33*, 642-681.

Ulmer, J. T., & Steffensmeier, D. (2014). The age and crime relationship: Social variation, social explanations. In Beaver, K. M., Barnes, J. C., & Boutwell, B. B. (Eds), *The Nurture Versus Biosocial Debate in Criminology: On the Origins of Criminal Behavior and Criminality* (pp. 377-396). Thousand Oaks, CA: SAGE Publishing.

U.S. Department of Education. (2016). State and local expenditures on corrections and education. Prepared by Stullich, St., Morgan, I., & Schak, O. Retrieved from https://www2.ed.gov/rschstat/eval/other/expenditures-corrections-education/brief.pdf.

U.S. Department of Justice. (2016). Growing number of communities are using data to improve policing and criminal justice. Prepared by Davis, R. L., Austin, R. L., & Patil D. Retrieved from https://www.justice.gov/archives/opa/blog/growing-number-communities-are-using-data-improve-policing-and-criminal-justice.

Van De Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. G. (2014). A gentle introduction to Bayesian analysis: Applications to developmenal research. *Child Development, 85*(3), 842-860.

Van Voorhis, P., & Presser, L. (2001). *Classification of Women Offenders: A National Assessment of Current Practices*. Washington, DC: National Institute of Corrections.

Van Voorhis, P., Wright, E. M., Salisbury, E., & Bauman, A. (2010). Women's risk factors and their contributions to the existing risk/needs assessment: The current status of a gender-responsive supplement. *Criminal Justice and Behavior, 37,* 261-288.

Vaughn, M. G., Wallace, J. M. Jr., Davis, L. E., Fernandes, G. T., & Howard, M. O. (2008). Variations in mental health problems, substance use, and delinquency between African American and Caucasian juvenile offenders: Implications for reentry services. *International Journal of Offender Therapy and Comparative Criminology, 52*(3), 311-329.

Vera Institute of Justice. (2014). *Recalibrating Justice: A Review of 2013 State Sentencing and Corrections Trends—Report Summary.* Retrieved from http://archive.vera.org/sites/default/files/resources/downloads/state-sentencing-and-corrections-trends-2013-summary.pdf.

Vito, G. F., Higgins, G. E., & Tewksbury, R. (2012). Characteristics of parole violators in Kentucky. *Federal Probation, 76*, 19-23.

Wacquant, L. (2001). Deadly Symbiosis: When ghetto and prison meet and mesh. *Punishment and Society, 3*(1), 95-133.

Wagner, D., Quigley, P., Ehrlich, J., & Baird, C. (1998). *Nevada Probation and Parole Risk Assessment Findings.* Madison, WI: National Council on Crime and Delinquency.

Walker, M. L. (2016). Race making in a penal institution. *American Journal of Sociology, 121*(4), 1051-1078.

Warr, M. (1998). Life course transitions and desistance from crime. *Criminology, 36*(2), 183-216.

Wasserman, M. (March 08, 2018). Efforts to ban the death penalty fizzle out in Legislature. *The News Tribune.* Retrieved from http://www.thenewstribune.com/news/politics-government/article204222859.html.

Weinberg, S. K. (1995). Theories of criminality and problems of prediction. *Journal of Criminal Law and Criminology, 45(*4*),* 412-424.

Whitehead, A. (2005). Man to man violence: How masculinity may work as a dynamic risk factor. *The Howard Journal of Crime and Justice, 44*(4), 411-422.

Wisconson v. Loomis. 881 N.W.2d 749. (2016).

Wright, J. P., & Boisvert, D. (2009). What biosocial criminology offers criminology. *Criminal Justice and Behavior, 36*(11), 1228-1240.

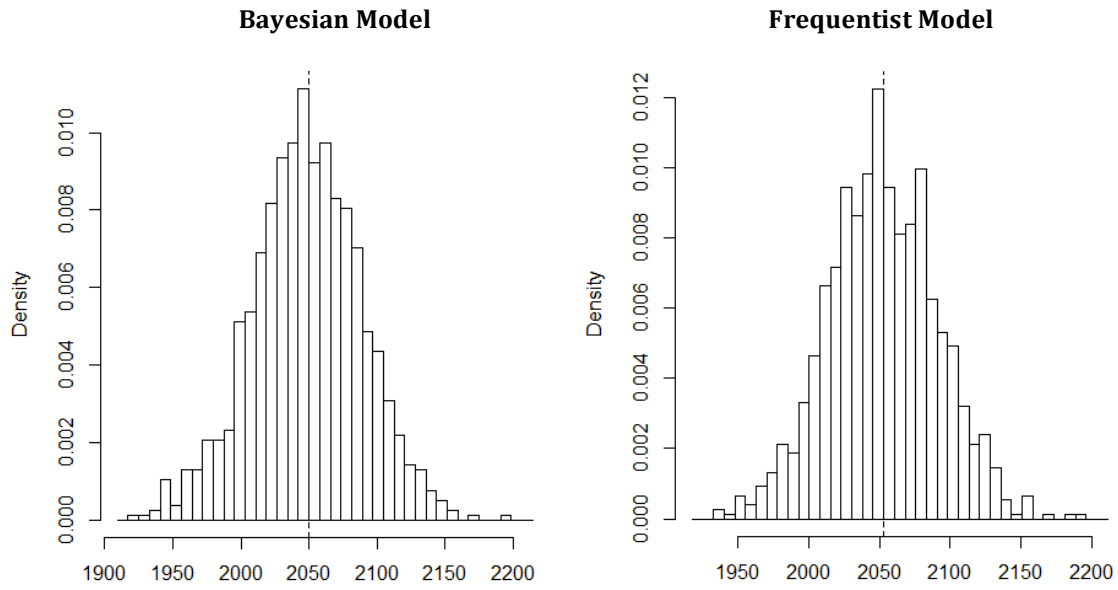
Yesberg, J. A., Scanlan, J. M., Hanby, L. J., Serin, R. C., & Polaschek, L. L. (2015). Predicting women's recidivism: Validating a dynamic community-based 'gender-neutral' tool. *Probation Journal, 62*(1)*,* 33-48.

Zatz, M. S. (2000). The convergence of race, ethnicity, gender, and class in court decision making: Looking toward the 21st century. In *Criminal Justice 2000* (Vol. 3). Washington DC: U.S. Department of Justice.

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Statistics in Society, 180*(3), 689-722.

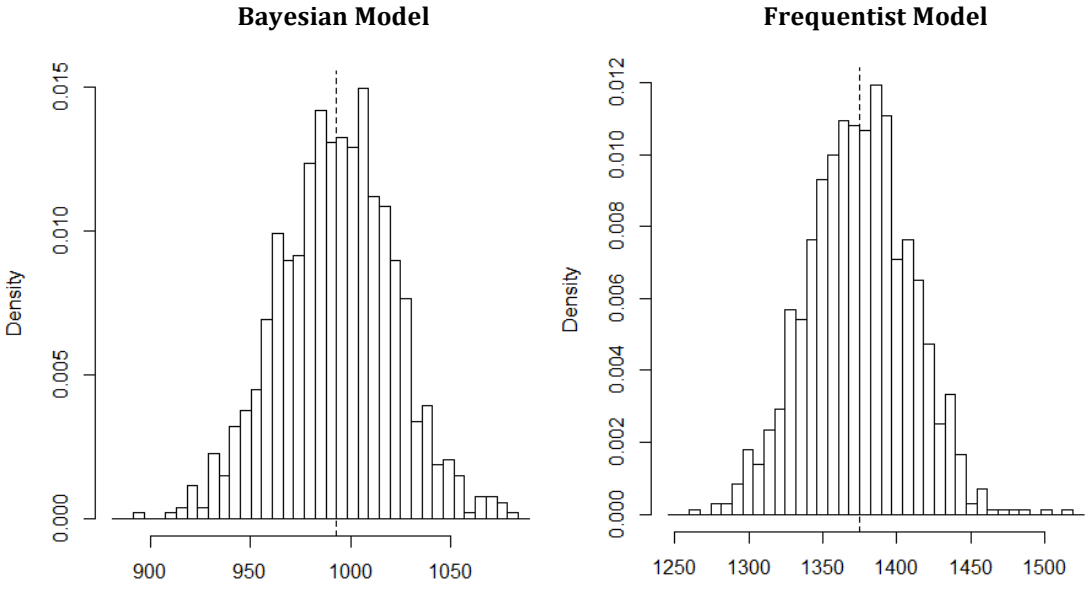
APPENDIX

**Figure A.1: Bootstrapped Estimates of Felony Recidivism Prediction Error**



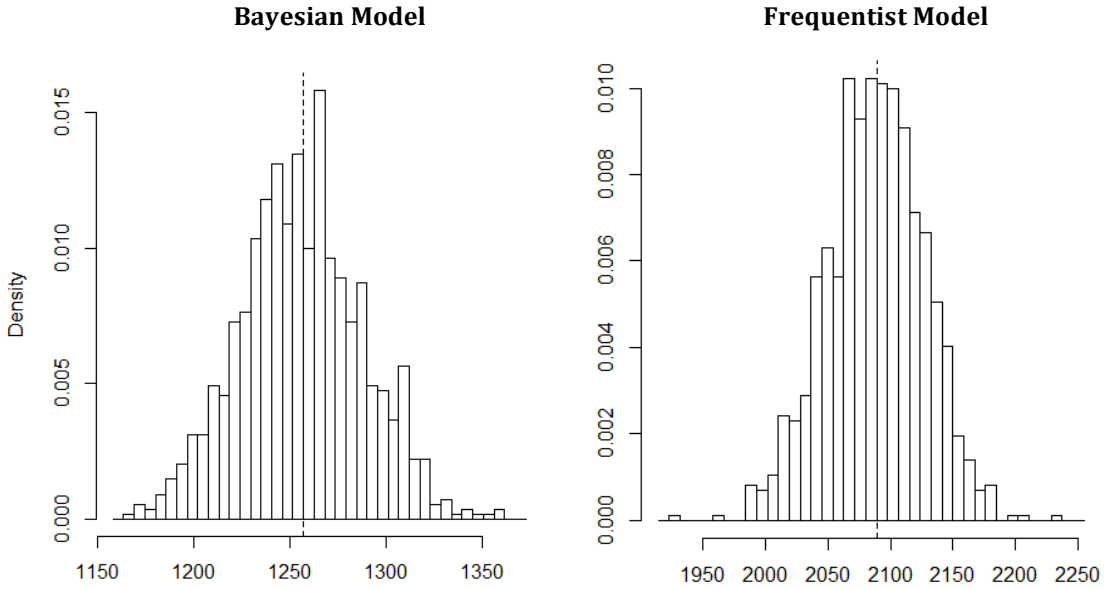

Number of Incorrect Felony Recidivism Classifications

**Figure A.2: Bootstrapped Estimates of Violent Recidivism Prediction Error**



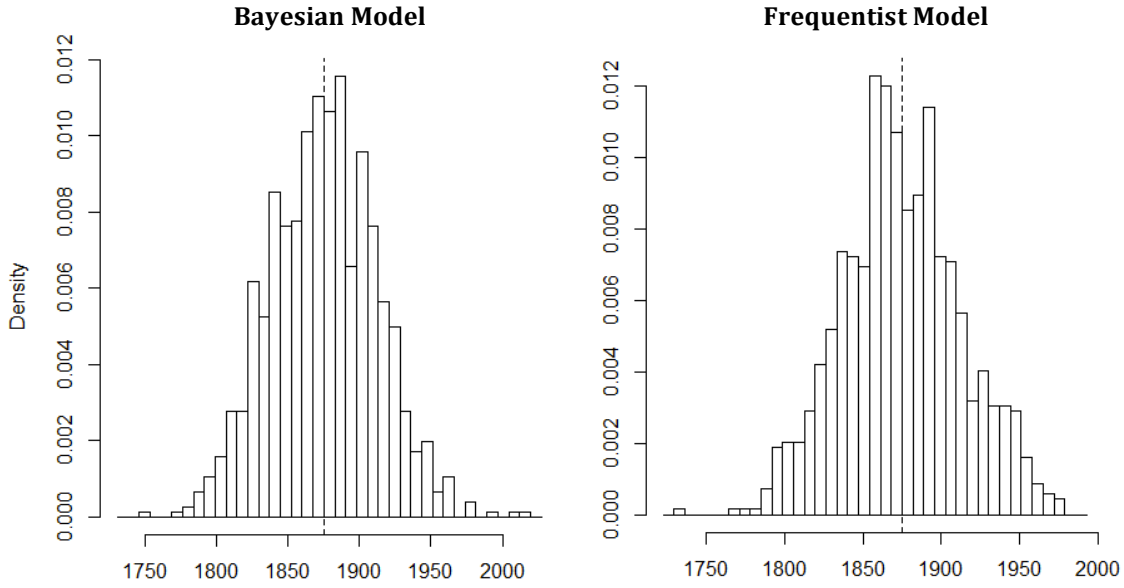

Number of Incorrect Violent Recidivism Classifications

**Figure A.3: Bootstrapped Estimates of Property Recidivism Prediction Error**




Number of Incorrect Property Recidivism Classifications

**Figure A.4: Bootstrapped Estimates of Drug Felony Recidivism Prediction Error**




Number of Incorrect Drug Recidivism Classifications